Author:
Watkins, Sarah H

Title:
Leveraging the correlation structure in DNA methylation data to identify stable and persistent regulatory networks in the human methylome

Leveraging the correlation structure in DNA methylation data to identify stable and persistent regulatory networks in the human methylome

Sarah Holmes Watkins

A dissertation submitted to the University of Bristol in accordance with the requirements for award of the degree of Doctor of Philosophy in the Faculty of Health Sciences

Bristol Medical School: Population Health Sciences

November 2019

Word count: 56,218

Abstract

DNA methylation (DNAm) is an epigenetic modification that influences genetic function, which can be altered by environmental and genetic factors. Relationships between DNAm sites are important because genome biology functions as a system, and as such it is unlikely that DNAm sites act in isolation. Identifying key relationships between DNAm sites may uncover systems or pathways which regulate, or are regulated by, DNAm. This could lead to a better understanding of regulatory mechanisms involved in disease susceptibility, which may lead to new therapeutic targets being identified.

Previous studies have shown that DNAm sites in close proximity often have correlated DNAm states. However the drivers of this correlation have so far only been established for selected DNAm sites. This thesis provides a comprehensive description of correlation structure across the entire 450k array in cis and trans, and how this is preserved in the same individuals over time, as well as across datasets and ethnicities.

I illustrate that across the genome (as measured by the 450k), cis correlation structure is consistent and replicable, both across cohorts and across ethnicities. I show that genetic influences on DNAm correlation structure are present, but that they do not seem to reflect LD structure. I show that highly correlated trans DNAm sites are enriched for active transcription start sites, promotors, and transcription regulation. Network visualisation shows that these highly correlated trans sites are interconnected.

I finally create DNAm networks using WGCNA, to ascertain whether correlated DNAm sites form pathways that associate with biological functions and phenotypes related to development. I find that DNAm modules are strongly preserved over time, between datasets, and between ethnicities.

This work shows that correlations between DNAm sites are replicable, stable, and biologically meaningful, and can be leveraged to gain novel insights about genome function.

Dedicated to the memory of my mum, Diana.

You were awesome. I miss you

♥

Acknowledgements

Academic

A lot of people have been very helpful during my PhD, and I have been lucky to work in a department with a very friendly and helpful culture.

My supervisors Tom, Josine, and Nic have been brilliant and I'm really grateful for all their help and support over the past 4 years.

Particular thanks also go to Matt Suderman, Kim Burrows, Ryan Arathimos, Hannah Elliott, and Nabila Kazmi for helping me with code or analyses. Their contributions are noted in the main text.

I would like to thanks the Wellcome Trust, who have funded my PhD.

I would also like to extend a huge thank you to the participants of ALSPAC and Born in Bradford, without whom this project would not have been possible.

And a special mention goes to all my friends at work, particularly everyone in BF8.

Emotional

I also would not have made it through the last four years without all the emotional support I've been lucky enough to have.

Firstly my partner Graeme, who has always been completely supportive and encouraging, and most importantly makes me laugh no matter what ☺

My family, Dad, Roger and Dave. We've had some tough things to deal with over the past 4 years but being so close with you guys has made it much easier than it might have been. Plus my niece and nephew, who keep me entertained in countless videos, photos and stories.

My grandma, who passed away recently. You were always up for a chat and you had some excellent stories. And my grandpa, who is probably the most kind and intelligent person I've ever met. You always have wise words and unwavering faith in me.

And all my friends, particularly Kate, Anne, and Beki, who have had some very wise and encouraging words that have helped me through the PhD.

Author's declaration

I declare that the work in this dissertation was carried out in accordance with the requirements of the University's *Regulations and Code of Practice for Research Degree Programmes* and that it has not been submitted for any other academic award. Except where indicated by specific reference in the text, the work is the candidate's own work. Work done in collaboration with, or with the assistance of, others, is indicated as such. Any views expressed in the dissertation are those of the author.

SIGNED: .... DATE:..........................

Table of contents

List of tables

List of figures

List of appendices

List of abbreviations

450k - HumanMethylation450 beadchip array

ALSPAC - The Avon Longitudinal Study of Parents and Children

ARIES - Accessible Resource for Integrated Epigenomic Studies

BiB – Born in Bradford

BMI – body mass index

DMR – differentially methylated region

DNAm – DNA methylation

DNMT - DNA methyltransferase

EPIC - MethylationEPIC beadchip array

EWAS – Epigenome wide association study

GO – gene ontology

GWAS - genome-wide association studies

IgG – Immunoglobulin G

ISVA – independent surrogate variable analysis

KEGG – Kyoto Encycolpedia of Genes and Genomes

kME – DNA methylation site connectivity score

LD – linkage disequilibrium

mQTL – methylation quantitative trait loci

Pol2 – RNA polymerase II

Pol3 – RNA polymerase III

SNP – single nucleotide polymorphism

SV – surrogate variable

SVA - Surrogate Variable Analysis

TET - ten-eleven translocation enzymes

TF – Transcription factor

TFBS – transcription factor binding site

VMR – variably methylated region

WGCNA – Weighted Gene Correlation Network Analysis

1 Introduction

1.1 Epigenetics

Epigenetics is a field of study concerned with changes to the genome that are inherited through mitosis and meiosis, and which affect DNA function, but do not alter the DNA sequence (A. Bird, 2007; Felsenfeld, 2014; Probst, Dunleavy, & Almouzni, 2009). Epigenetics is of particular interest to epidemiologists because although it is stable throughout the cell cycle and heritable through cellular division, it can be altered by the environment, illustrating a mechanism by which phenotypic differences can arise between individuals in the absence of variation in genetic sequence. Epigenetic features discussed in this thesis are illustrated in Figure 1, and are discussed in more detail in the sections below.

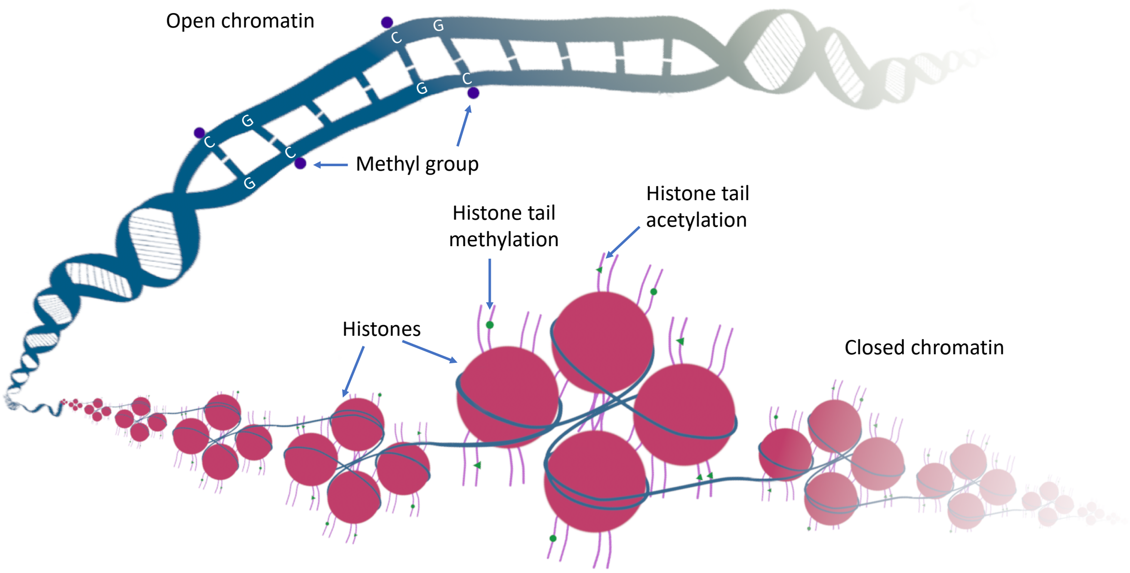


Figure 1: Illustration of epigenetic features along a strand of DNA

1.2 DNA methylation

1.1.1 Brief introduction to DNA methylation

DNA methylation (DNAm) refers to the addition of a methyl group to DNA bases (Teschendorff & Relton, 2018). It is one of the most studied epigenetic processes, because it is relatively stable (Eckhardt et al., 2006) and easy to measure. Both cytosine and adenine can be methylated, but because the methylation of adenine was only conclusively established in human cells last year (Xiao et al., 2018), discussion of DNA methylation in this thesis will refer solely to cytosine methylation.

In humans, methylation of cytosines was initially thought to be almost exclusively as part of a cytosine-guanine dinucleotide (known as a CpG) (Sinsheimer, 1955). However, in the last 10 years DNAm in non-CpG contexts (known as CpH) has been described in most human tissues (Schultz et al., 2015) and is widespread in the brain (Price et al., 2019). As the data used in this thesis predominantly targets CpG methylation, that will be the focus in this chapter. In a CpG, on the forward strand the cytosine base is always 5' to the guanine, and the cytosines on both strands are symmetrically methylated (A. P. Bird, 1978; Cedar, Solage, Glaser, & Razin, 1979).

The methylation of cytosines is catalysed by a group of enzymes known as DNA methyltransferases (DNMTs), of which there are three in humans that catalytically active. DNMT1 is responsible for maintaining DNAm patterns, as it is selective for sites with symmetrical CpGs (Bestor & Ingram, 1983; Lyko, 2018). DNMT3A and DNMT3B are responsible for de novo methylation of unmethylated DNA, and are not selective about the sequence to be methylated (Okano, Bell, Haber, & Li, 1999) - they are directed by binding to unmethylated H3K4 (a histone mark, see section 1.2.2.2)(Ooi et al., 2007). However there is evidence to suggest a slightly more complex interplay of functions between these three DNMTs (Liao et al., 2015; Lyko, 2018; Tiedemann et al., 2014). DNMTs are regulated by molecular interactions (such as sequestration by non-coding RNA), post-translational modifications, and alternative splicing; this in turn regulates methylation of the genome (Lyko, 2018).

DNA is de-methylated by three ten-eleven translocation (TET) enzymes, TET1, TET2, and TET3 (He et al., 2011; Tahiliani et al., 2009). These enzymes oxidise and decarboxylate methylated cytosines (Ito et al., 2011), changing the cytosine from 5-methylcytosine (5mC) to other modified forms (5-hydroxymethylcytosine (5hmC), 5-formylcytosine (5fC), or 5-carboxylcytosine (5acC)); for a comprehensive review of this please see (Wu & Zhang, 2017). De-methylation can then occur in one of two ways; either during replication, where the new strand will have an unmethylated cytosine instead of the modified version, as DNMT1 will not recognise the modified cytosine (Kohli & Zhang, 2013; Lio & Rao, 2019; Otani et al., 2013); or alternatively cytosines converted to 5fC and 5acC can be actively excised and replaced by an unmethylated cytosine by thymine-DNA glycosylase (TDG), a glycosylase known to excise DNA bases (He et al., 2011; Maiti & Drohat, 2011).

1.2.1 Distribution of DNA methylation across the genome

There are around 28 million CpG dinucleotides on the human genome, which occur at around 25% of their expected frequency, based on the GC content of the human genome (Lister et al., 2009; Saxonov, Berg, & Brutlag, 2006). CpGs are sparsely distributed across the genome, aside from CpG-dense regions known as CpG islands, which are around 1kb long, with high GC and CpG content. Most CpG islands are located near transcription start sites, and these tend to be unmethylated (Deaton & Bird, 2011; Eckhardt et al., 2006; Saxonov et al., 2006). CpG islands that reside in gene bodies are more likely to be methylated (Jeziorska et al., 2017), and DNAm is often intermediate at enhancers (Stadler et al., 2011), suggesting differing functions for CpG islands outside promotor regions.

1.2.2 How DNA methylation affects genome function

1.2.2.1 How DNA methylation affects gene expression

The relationship between DNAm and gene expression is complex and context dependent. Where DNAm is associated with gene expression levels, it tends to be associated with reduced gene expression in *cis,* and positively associated with gene expression in *trans* (Siegfried et al., 1999; van Eijk et al., 2012). DNAm in gene bodies is associated with increased gene expression (Lister et al., 2009), whereas it is negatively associated with gene expression at promotors (Bell et al., 2011; Gutierrez-Arcelus et al., 2013; Y. Y. Zhang et al., 2009), enhancers and CTCF transcription factor binding sites (Gutierrez-Arcelus et al., 2013).

The causal direction of this relationship is similarly complex. DNAm has been shown to causally affect gene expression (Gutierrez-Arcelus et al., 2013; Maeder et al., 2013; Saunderson et al., 2017), where the mechanism of action is through DNAm affecting the ability of transcription factors to bind (Prendergast & Ziff, 1991; Razin & Riggs, 1980), or methylated cytosines initiating the recruitment of methyl-CpG-binding proteins, which initiate chromatin remodelling and thus block access of transcriptional machinery (Hendrich & Bird, 1998; Nan et al., 1998). DNAm can also be a consequence of gene expression; it can be causally influenced by transcribed microRNAs influencing DNMTs (Sinkkonen et al., 2008), the binding of transcription factors (Stadler et al., 2011), and the expression of genes encoding DNMTs and TETs. It is not currently known whether DNAm at one locus can influence DNAm at another. As a consequence, DNAm levels might be expected to be similar at sites which bind the same transcription factors, have correlated genetic variants, or have similar chromatin states.

It seems that DNAm can be either be a cause or a consequence of altered gene expression; DNAm can be passively influenced though genetic variants, transcription factors and gene expression (Gutierrez-Arcelus et al., 2013; Stadler et al., 2011), but it can also itself affect the binding of transcription factors (Prendergast & Ziff, 1991; Razin & Riggs, 1980), and causally affect gene expression (Maeder et al., 2013; Saunderson et al., 2017).

1.2.2.2 DNAm interactions with chromatin

DNAm is not the only epigenetic modification to impact DNA function; histone modifications (which form chromatin structure) and non-coding RNA also have important roles (Groom, Elliott, Embleton, & Relton, 2011). The role of chromatin structure is relevant to this thesis, and so consideration of its interaction with DNAm is needed.

DNA is wound up into tight structures by groups of proteins called histones, and the structure this forms is known as chromatin (B. Li, Carey, & Workman, 2007). If DNA is required for a particular function, an appropriate length must be unwound and made accessible. Specific chromatin modifications (such as methylation or acetylation of the histones) perform specific functions; for example there are specific chromatin states at enhancers and transcription start sites, because each will require a specific length of DNA to be accessible (Ernst et al., 2011). Many of these states have been mapped, giving us the ability to identify which areas of the genome are associated with particular chromatin states, and thus activity at these sites such as transcription, repression, or enhancer activity (Ernst & Kellis, 2010; Heintzman et al., 2007; Hon, Wang, & Ren, 2009). Most relevant to this thesis is the Roadmap Epigenomic consortium's map of 25 chromatin states across the genome (Roadmap Epigenomics et al., 2015).

Chromatin states and DNAm have an interacting relationship. Methylated DNA sites tend to be associated with closed chromatin, and unmethylated sites with open chromatin; this influences gene transcription as when chromatin is closed transcription factors (TFs) cannot access the DNA strand; and when DNA is methylated TFs cannot bind to DNA (Fuks, 2005).

Some specific chromatin states influence establishment DNAm (including at imprinting sites (Ciccone et al., 2009; Z. D. Smith & Meissner, 2013)), and de novo methylation is strongly related to the methylation status of the lysine 4 on the H3 histone (known as H3K4) (Ooi et al., 2007). DNAm patterns are used to re-create chromatin conformation after cell division (Cedar & Bergman, 2009), and can initiate methylation of histones (Matsumura et al., 2015).

Chromatin also organises DNA into a 3D structure. This was comprehensively mapped by (Rao et al., 2014), who demonstrated the importance of chromatin loops for distant regulatory elements to come into physical contact with areas of the genome that they regulate. They showed that the CTCF TF is almost always present at the contact point of the loop, and the CTCF motifs are almost always oriented in convergent directions. Almost 10,000 loops were identified in humans, demonstrating the potential importance of *trans* interactions across the genome. DNAm at CTCF sites has been shown to correlate negatively with gene expression (Gutierrez-Arcelus et al., 2013). Using Hi-C in hematopoietic cell lines, (Javierre et al., 2016) demonstrated that around 10% of over 30,000 promotors that they studied have long-range interactions (including *trans*-chromosomal), which are enriched for contacts with active enhancers and genetic variants that control gene expression.

1.2.3 The role of DNA methylation in pre and post natal development, and age

1.2.3.1 Embryonic and pre-natal development

DNAm is dynamic during embryonic development, with substantial global demethylation shortly after fertilisation. DNAm is then widely reprogrammed at the blastocyst stage, so that the cells can become pluripotent (Feng, Jacobsen, & Reik, 2010; Smallwood & Kelsey, 2012). Some DNAm marks then become part of cell lineage differentiation, and become fixed for life; because of this these marks can be used to estimate cell type proportions (Houseman et al., 2012; Zaimi et al., 2018). DNAm controls the expression of imprinted genes through methylation at imprinting control regions, something which is established in gametes and is not subject to reprogramming in the early embryo (Messerschmidt, Knowles, & Solter, 2014). In human studies, DNAm has been shown to have dynamic involvement in fetal brain development (Spiers et al., 2015), as well as relevant tissue-specific patterns in multiple other tissues (Slieker et al., 2015).

1.2.3.2 Post-natal development

DNAm continues to change after birth, and across the lifecourse, although development-related functions of DNAm from birth through to adulthood have been less well described than the pre-natal period. A small study of 7 individuals showed that DNAm is also dynamic at selected sites between birth and 5 years, particularly at immune-related sites (with the changes not entirely attributable to changes in cell counts, which were directly measured) (D. J. Martino et al., 2011). This was followed up by a study of 15 twin pairs which showed that sites differentially methylated over time were related to development- and morphogenesis-related ontologies (D. Martino et al., 2013). A larger study confirmed dynamic changes in DNAm across childhood that were replicable and enriched for immune- and development- related ontologies, although the overlap of these sites with adult age-related sites and the lack of control for cell counts (see discussion of DNAm age below) could suggest these loci were related to cellular proportions (Alisch et al., 2012).

1.2.3.3 Age

DNAm changes as we age. This is in part due to a phenomenon known as epigenetic drift, where errors in DNAm maintenance lead to stochastic variation (Fraga et al., 2005; Teschendorff, West, & Beck, 2013), although this is limited at many sites by genetic or environmental influences (Shah et al., 2014). Caution is needed when examining epigenetic changes with age in tissues with heterogeneous cell types, because cell type proportions can change with age, and so changing DNAm may simply reflect the changing cell proportions; something that is particularly applicable to blood (Houseman et al., 2012; Langevin et al., 2011; Teschendorff et al., 2013).

Age-related changes in DNAm are well documented, and numerous ‘epigenetic clock’ methods have been developed to predict an individual’s age from blood methylation (Bocklandt et al., 2011; Hannum et al., 2013; Horvath, 2013; Horvath et al., 2012; Levine et al., 2018). Having a DNAm age greater than chronological age has been associated with higher risk of mortality (Christiansen et al., 2016; Marioni et al., 2015); however a more recent study indicated that using a large enough sample size to train the predictor, and adjusting for cell counts, led to an age prediction that was more accurate, and an attenuation of the association of increased epigenetic age with mortality (Q. Zhang et al., 2019), underscoring the importance of cell types in the examination of DNAm and age.

1.2.4 DNA methylation and ethnicity

Ethnicity and ancestry are two different constructs that are both important for genomics research, although the terms have not always been used appropriately (Ali-Khan, Krakowski, Tahir, & Daar, 2011; Yudell, Roberts, DeSalle, & Tishkoff, 2016). Ancestry is a construct derived solely from an individual's genetic heritage. Ethnicity is a social construct that overlaps ancestry, but also encompasses shared experiences and exposures that may be missed by genetic ancestry, and so is particularly relevant for DNAm studies (Galanter et al., 2017). Where discussing previous literature, for consistency I have used the term used in the original research article.

Ethnicity is of interest to studies of DNAm because rates of non-communicable diseases often differ between ethnic groups, such as type 2 diabetes (Anand et al., 2000; McWilliams, Meara, Zaslavsky, & Ayanian, 2009; Tillin, Hughes, Godsland, et al., 2013), cardiovascular disease (McWilliams et al., 2009; Tillin, Hughes, Mayet, et al., 2013), and asthma (Centers for Disease & Prevention, 2004; Keet et al., 2015). Environmental and social factors such as racism, social inequalities, and resulting exposure to environmental risk factors differ between ethnicities, which adds to disease risk disparities (Cooper, 2001; Nguyen, Moser, & Chou, 2014). Different ethnicities will also generally be associated with different genetic backgrounds (Galanter et al., 2017; Genomes Project et al., 2015; International HapMap et al., 2010) which has demonstrable effects on DNAm (Rahmani et al., 2017). As DNAm is affected by both environmental factors and genetic variants, it provides a potential mechanism by which disease rates may differ between ethnic groups.

In addition to looking to explain the disparities between rates of disease, multi-ethnic studies also provide an opportunity to examine the role and consistency of DNAm in traits and diseases in a wider context, which provides stronger evidence for the functions DNAm might perform (Tang, 2006). If DNAm is consistent between ethnicities, this might point to basic functional roles that would confer greater understanding of the role DNAm plays in genome function – if DNAm is consistent in the presence of differing genetic and environmental influences.

Substantial differences in DNA methylation across the genome have been described between different ethnicities (Fraser, Lam, Neumann, & Kobor, 2012), including at birth (Adkins, Krushkal, Tylavsky, & Thomas, 2011). Some of these differences have been shown to be driven by genetic factors (Husquin et al., 2018; Moen et al., 2013), as allele frequencies differ between human populations (and so often ethnicities) (Genomes Project et al., 2015; International HapMap et al., 2010), and so methylation can differ because it is being influenced by different genotypes (C. Tian, Gregersen, & Seldin, 2008). Evidence from a trans-ancestry GWAS that disease-related SNPs are mQTLs (Kato et al., 2015), and ethnic differences in methylation correlate with gene expression differences (Husquin et al., 2018; Moen et al., 2013), suggesting that DNAm could be part of the regulatory pathway that leads from genetic variants to differences in phenotypes between ethnic groups. This is demonstrably relevant to DNAm research, as studies have identified that differences in DNAm exist between different ethnic groups in type 2 diabetes susceptibility genes (Chambers et al., 2015; Elliott et al., 2013), smoking loci (Elliott et al., 2014; Park et al., 2018), and loci related to autoimmunity (Coit et al., 2015). In contrast to differences found between ethnicities, one study found that methylation profiles were similar between ethnicities in normal tissue but differed in tumour samples (Chu & Yang, 2017), and another found that *cis* correlation structure between ethnicities looks to be the same (Saffari et al., 2018).

1.3 DNA methylation and epidemiology

DNAm has become very interesting to epidemiology in the last decade, particularly so because it is influenced by both environmental (such as smoking and adversity; discussed in detail below in section 1.3.4) and genetic factors. DNAm is an epigenetic modification suited to epidemiological studies because it is relatively stable, over cell divisions (Stein, Gruenbaum, Pollack, Razin, & Cedar, 1982; Wigler, Levy, & Perucho, 1981), and over time (Shah et al., 2014). It has become relatively easy to measure in large epidemiological studies with the introduction of array technologies (Mansell et al., 2019; Teschendorff & Relton, 2018), and the reasonable stability makes it easier to associate with environmental factors (which are often measured around the time of the biological sample, but not necessarily at the exact time). The resulting field of study is known as epigenetic epidemiology.

1.3.1 Measuring DNAm in epidemiological studies

1.3.1.1 Array technology

Epidemiological studies most often use array technology to measure DNAm, because this is the most cost-effective way to do so in large samples. The most frequently used arrays are the Illumina HumanMethylation27 (27k), HumanMethylation450 (450k), and Methylation EPIC (EPIC) beadchip arrays. These measure DNAm across the genome at 25,578, 485,577 and 866,836 sites, respectively. The 450k array measures about 2% of DNAm sites, and the EPIC around 4% (Michels et al., 2013; Pidsley et al., 2016). Beadchip arrays measure the proportion of sites that are methylated, and the beta value measures the proportion methylated from 0 to 1 (Bibikova et al., 2011).

The 450k is quite biased towards promotors and TSS, and to coding transcripts (Sandoval et al., 2011), and as such any enrichment analyses need to take this in to consideration. It also does not measure DNAm in many regulatory regions, which have been shown to be important; and so the larger EPIC array was designed to target more of these regions, as well as over 90% of the sites targeted by the 450k array (Pidsley et al., 2016). The arrays are gene-centric, which likely misses additional regulatory regions. Probes on the arrays are also unevenly distributed, with some regions having much denser coverage. This can cause problems for enrichment analyses that do not account for coverage. It can also cause problems for studies investigating relationships between DNAm sites, such as correlation or regional analysis, because many regions will not support the resolution to investigate whether neighbouring DNAm sites exhibit similar methylation levels.

Not all probes on the 450k array work optimally. If they contain a SNP near the 3’ end of the probe; if they map to multiple locations after bisulfite conversion; or if a considerable proportion of a unique probe consists of non-unique sequence, the signal on the array will be impacted (Dedeurwaerder et al., 2014; Naeem et al., 2014; Zhou, Laird, & Shen, 2017). A number of groups have investigated which probes are affected by these issues, and have provided lists of probes suggested for removal.

1.3.2 Considerations stemming from the frequent use of blood as the tissue in epigenetic epidemiology

1.3.2.1 DNA methylation is tissue-specific

The majority of epidemiological studies use blood to measure DNAm, as it is relatively non-invasive and can easily be collected longitudinally. However using blood raises a number of issues. The first is that blood may well not be the tissue of interest to many research questions. This can be a problem for DNAm research because, as discussed above, DNAm is cell lineage specific, which means it is generally tissue specific. This has been quantified by an analysis of the consistency of DNAm across the Illumina 450k between blood and brain, which showed that DNAm in blood is generally quite poorly related to DNAm in the brain. Some sites were found to be concordantly methylated, and so blood may be useful for identifying biomarkers of disease or phenotype in other tissues (Hannon, Lunnon, Schalkwyk, & Mill, 2015). Tissue-specific DNAm has also been shown comparing adipose tissue to blood (Allum et al., 2019), in relation to BMI in adipose, skin and blood (Dick et al., 2014), and in relation to type 2 diabetes in liver, adipose and blood (Barajas-Olmos et al., 2018).

1.3.2.2 DNA methylation is cell-type specific

The second issue is that blood is made up of a number of heterogeneous white cell types, which each have a distinct methylation profile (see discussion of lineage-specific methylation in section 1.2.3). The difficulty this creates is being able to separate whether differences in DNAm are due to real differences in methylation level, or to differences in cell type proportions (Teschendorff & Relton, 2018). Because cell type proportions are rarely measured, algorithms have been developed to estimate relative proportions of cell types in each sample based on the methylation level of cell type-specific sites (Bakulski et al., 2016; Houseman et al., 2012). These estimates can then be adjusted for, and in theory leave DNAm values that do not reflect differences in cell type composition. In practice, these are just estimates, and as such some residual cell type effects are inevitable.

1.3.2.3 Removing variation from unknown sources in DNAm data

To try and remove batch effects from unknown or unconsidered sources, and attempt to remove more of the variation from the known sources (particularly cellular proportions), methods have been developed to statistically identify and adjust for sources of variation in DNAm data. These methods construct surrogate variables to represent all the variation in the data that is not accounted for by specified known covariates and any trait(s) of interest. Surrogate Variable Analysis (SVA) (Leek & Storey, 2007) and Independent Surrogate Variable Analysis (ISVA) (Teschendorff, Zhuang, & Widschwendter, 2011) are the most frequently used methods for DNA methylation data. Although these techniques can be particularly useful for studies associating DNAm with a specific phenotype, they may not be so appropriate for more exploratory analyses because there is little indication of what the surrogate variables represent, and these methods risk removing biologically interesting sources of variation in the data.

1.3.3 Common methods in epigenetic epidemiology and their limitations

1.3.3.1 EWAS

Epigenome-wide association studies (EWAS) were modelled on genome-wide association studies (GWAS). The method generally uses linear regression to test the association of every individual DNAm site with a trait of interest, whilst controlling for batch effects and relevant covariates (Mansell et al., 2019). As genome-wide testing requires rigorous correction for multiple testing, there is a high significance threshold sites must pass (Mansell et al., 2019; Saffari et al., 2018). Although necessary to avoid type I errors, this may also cause true associations with phenotype-associated DNAm sites to be missed. This is likely to be especially true in epigenetic studies, which have sample sizes orders of magnitude smaller than GWAS, and so are likely to be underpowered to find effects (Suderman et al., 2018).

1.3.3.2 Differentially methylated regions

As a solution to the issue of sites being unlikely to act in isolation, a number of methods have been developed to identify regions of co-acting DNAm sites (known as differentially methylated regions; DMRs) (Butcher & Beck, 2015; Gomez et al., 2019; Jaffe et al., 2012; Pedersen, Schwartz, Yang, & Kechris, 2012; Peters et al., 2015). These methods tend to identify proximal regions of DNAm sites which are differentially methylated between two groups of participants. They use various methods to group DNAm sites into regions, as there is no clear consensus about functional genomic divisions between co-acting DNAm sites.

1.3.3.3 Limitations of these methods

The limitation of both EWAS and regional associations is that they miss the big picture; biology functions as an interconnected system, and most biological functions depend on multiple interactions rather than the effects of a single site or a single region (Barabasi & Oltvai, 2004). As EWAS identifies sites in isolation, it is difficult to identify whether the sites may be part of a biological pathway, or if they are single-site changes what sort of function they might have (Lappalainen & Greally, 2017). This makes the results hard to interpret in a functional context. EWAS also do not control for correlations between DNAm sites (as GWAS do for LD); this is because DNAm data does not have a thoroughly defined correlation structure that necessarily applies throughout life (see section 1.5 below), but in these association analyses accounting for correlation between sites is important (Saffari et al., 2018). The same applies to many of the DMR methods in terms of *cis* correlations, and none account for *trans* correlations between DNAm sites. In addition, many of the methods for detecting differentially methylated regions have been shown to have problems with false positive rates and power (Suderman et al., 2018).

1.3.4 DNA methylation and environmental exposures

Exposure to a multitude of environmental factors has been shown to associate with differences in DNAm. Tobacco smoke (Ambatipudi et al., 2016; Breitling, Yang, Korn, Burwinkel, & Brenner, 2011; Shenker et al., 2013), neighbourhood disadvantage (J. A. Smith et al., 2017), radon exposure (de Vocht et al., 2019), alcohol consumption (Christensen et al., 2009; Wilson et al., 2019), dietary intake of nutrients involved in one-carbon metabolism such as folate and vitamin B12 (Caramaschi et al., 2017; Mandaviya et al., 2019; Perrier et al., 2019), healthy diet and exercise (Hibler, Huang, Andrade, & Spring, 2019), dietary fat content (Perfilyev et al., 2017), childhood adversity (Dunn et al., 2019; Houtepen et al., 2018), socioeconomic position (Hughes et al., 2018; McGuinness et al., 2012), economic hardship (Simons et al., 2016), and air pollution (Baccarelli et al., 2009) have all been associated with changes in an individual’s DNAm.

Environment-induced changes in DNAm are not limited to direct exposure; pre-natal environmental exposures are also associated with DNAm differences in children at birth, which, given the essential involvement of DNAm in fetal development, has the potential to cause health consequences in later life. Tobacco smoke (Hannon et al., 2019; JoubertFelix, et al., 2016; Richmond et al., 2015), maternal antidepressant use (Cardenas et al., 2019), maternal age (Markunas et al., 2016), maternal folate levels (Joubert, den Dekker, et al., 2016) and maternal exposure to air pollution (Gruzieva et al., 2019) have all been shown to affect DNAm in newborns. These changes can persist throughout early life; as an example, exposure to maternal smoking is associated with differences in DNAm throughout childhood and into adolescence (Richmond et al., 2015).

Currently it is not entirely clear how environmental exposures might result in changes to DNAm, although it is likely there are multiple mechanisms. It may be through alteration of DNMT or TET expression, non-coding RNA activity, or histone modifications (Aluru, 2017). It may be due to the environmental influence altering transcription factor binding to DNA (Martin & Fry, 2016). These mechanisms are just starting to be elucidated for specific exposures, for example with exposure to diesel exhaust altering the methylation and expression of TET1 (Somineni et al., 2016).

1.3.5 DNA methylation and disease

Changes in DNAm have also been associated with numerous diseases. Some DNAm sites have been found to be early markers of diabetes (J. Liu et al., 2019); DNAm sites associated with asthma have been found in both blood and airway-related tissue, in children (Arathimos et al., 2017; Xu et al., 2018) and in newborns, some of which correlate with gene expression (Reese et al., 2019). DNAm has multiple roles in cancer (Feinberg, Koldobskiy, & Gondor, 2016), is associated with psychotic experiences (Roberts et al., 2019) and schizophrenia (Jaffe et al., 2016).

1.3.5.1 Anthropometric traits

1.3.5.1.1 Birthweight

In a large meta-analysis, birthweight has been shown to be associated with DNAm at a large number of sites across the genome; some of which overlap sites associated with maternal smoking (Kupers et al., 2019). The identified sites did not overlap gene expression loci, and so it is unclear how these sites may be associated with birthweight. Another study found a smaller number of sites associated with birthweight in whole blood in newborns, and that DNAm mediates some of the relationship between maternal smoking and lower infant birthweight (Hannon et al., 2019). A longitudinal study found associations of DNAm with birthweight that did not persist into childhood and adolescence (Simpkin et al., 2015).

1.3.5.1.2 BMI

BMI has been associated with differences in DNAm in peripheral blood in multiple studies, where the DNAm sites are often related to lipid metabolism and inflammation (Demerath et al., 2015; Dick et al., 2014; S. Wahl et al., 2017), as well as being associated with expression of genes related to lipid metabolism (Mendelson et al., 2017). DNAm at BMI-associated sites predicts variance in BMI independently of BMI-associated genetic variants (Shah et al., 2015). There is evidence to suggest that DNAm is a consequence of being overweight (Richmond et al., 2016; S. Wahl et al., 2017), although as DNAm differences related to BMI have been shown to explain variance in adverse health outcomes additional to that of phenotypic BMI, this suggests that BMI-induced differences in DNAm might still be important (Hamilton et al., 2019). Maternal obesity, and to a larger extent maternal underweight, have also been found to be related to multiple DNAm sites in their offspring; maternal BMI was associated with only 2 sites, which persisted to 7 years of age (Sharp et al., 2015). It is not clear whether these BMI-related changes in DNAm are part of a concerted pathway, or whether they are independent, site-specific changes.

1.3.5.2 Gestational age

Gestational age has been associated with DNAm in a number of studies. (Simpkin et al., 2015) found many DNAm sites associated with gestational age in ARIES (a cohort detailed in chapter 2), although these differences became non-significant in childhood and adolescence. Another study found that DNAm sites associated with gestational age were situated in genes related to labor and delivery, and so DNAm related to gestational age could be reflective of proteins that play a role in delivery (Schroeder et al., 2011). It is not clear how these proteins relate to DNAm and whether there might be a regulatory network behind this association. A study using whole blood from newborns found over 4,000 DNAm sites associated with gestational age, which replicated across cohorts (Hannon et al., 2019).

1.3.6 Genetic influence on DNA methylation

It is well established that genetic variants affect DNAm at many sites across the genome (Bell et al., 2011; McRae et al., 2014), and this provides a mechanism by which genetic variants can affect gene function outside of transcription. DNAm sites that are affected by genetic variants are known as methylation quantitative trait loci, or mQTL. *Cis*-genetic regulation has been shown to be stable from birth to adolescence in a longitudinal cohort, with the contribution of genetic variants decreasing slightly over time due to increasing effects of the environment and stochastic changes (Gaunt et al., 2016). *Trans*-genetic regulation has been a more tricky aspect to measure, because of the polygenic nature of *trans* mQTLs. *Trans* mQTLS are influenced by a large number of variants with small effects, requiring large sample sizes; this is important because the majority of DNAm variance has been estimated to come from *trans* genetic variants (Gaunt et al., 2016), and this is an aim of the Genetics of DNA methylation consortium (GoDMC). mQTLs have been shown to be functionally relevant - DNAm sites which influence gene expression tend to be under *cis* genetic control (Bell et al., 2011; van Eijk et al., 2012), and t*rans* mQTL regulation of DNAm has been shown to be due, in part, due to the regulation of transcription factors (Bonder et al., 2017). DNAm has also been shown to be under stronger genetic control in particular genomic regions (Gutierrez-Arcelus et al., 2013).

1.3.6.1 Heritability

Heritability refers to the amount of variation in a trait that is due to genetic factors (Vinkhuyzen, Wray, Yang, Goddard, & Visscher, 2013). In terms of DNAm, heritability simply means the amount of variation in methylation level that is due to genetic influences. Heritability is a useful concept because it allows the quantification of the contribution of genetic variants to variation in DNAm levels. Estimates of heritability and environmental influences for most sites on the 450k array have been created and made available by two groups (Hannon et al., 2018; van Dongen et al., 2016). These authors used data from MZ and DZ twins to estimate how much DNAm variability is due to the influences of additive genetics, common environment, and unique environment (which includes measurement error); van Dongen et al also estimated non-additive genetic contributions.

Hannon et al found that the contribution of heritability to methylation variability was generally low, but was strongest in sites which were more variable and had intermediate methylation levels. They also found that DNAm sites which were associated with environmental exposures such as smoking were influenced by heritability, suggesting that actually genetic effects may account for the differences we see in DNAm relative to environmental exposures. This finding underlines the importance of identifying whether relationships between DNAm sites are also influenced by genetic effects. Van Dongen et al also found that for most sites heritability contributed the smallest amount of variance, and most variance was attributed to unique environment (which includes measurement error). As they used a dataset with a larger age range, they identified that the contribution of environment to the variability of some sites increased with age.

1.4 Systems biology: a solution to some of the single-site problems

An alternative to these single site or region methods is to look at DNAm as an interconnected system. Systems biology approaches are less focused on individual constituents and instead more interested in the dynamics and the structure of that biological system (Ideker, Galitski, & Hood, 2001; Kitano, 2002). Looking at DNAm in terms of an interconnected system makes sense for a number of reasons. If DNAm sites can be grouped into highly related modules, this could be a powerful mechanism of feature reduction, which would reduce the large burden of multiple testing that is inherent within EWAS studies. If DNAm networks can be detected, then by their nature they will provide insight into pathways and mechanism that cannot be inferred from single site analyses (Lappalainen & Greally, 2017). And finally, grouping DNAm sites into networks will provide greater weight and confidence in their association with phenotypes of interest, because the association does not depend on a single site.

1.5 Relationships between DNAm sites

To identify whether two DNAm sites might be part of a pathway in a population study, we can look at whether the sites have correlated methylation states across all study participants. If the two sites vary to the same degree in each person, it is likely that their variation is being regulated by the same factors, and as such they might be in the same biological pathway. Systems biology in non-experimental settings uses correlations between DNAm sites to infer regulatory pathways (more on these methods in section 1.6 below). Because of this, it is important to understand the basic correlation structure in DNAm data that would underlie these networks.

1.5.1 Cis correlation structure

Previous research has shown that DNA methylation forms local correlation structures, with DNAm sites within 1-2kb often having correlated methylation states. This has been demonstrated using both whole genome bisulfite sequencing and array technology (Eckhardt et al., 2006; Kuan & Chiang, 2012; Y. Liu et al., 2014; Ong & Holbrook, 2014; Saffari et al., 2018; Shoemaker, Deng, Wang, & Zhang, 2010; W. Zhang, Spector, Deloukas, Bell, & Engelhardt, 2015). Bisulfite sequencing, which has the greatest resolution, shows that immediately adjacent sites almost always have the same methylation state (Eckhardt et al., 2006).

The functions and locations of these *cis* correlations have not had particularly consistent answers. (Y. Liu et al., 2014) found that highly correlated local clusters of DNAm sites were not related to genomic annotation such as CpG islands, although it should be noted that they used only the 25% most variable sites on the 450k array, which were grouped into correlated regions. One analysis has demonstrated that in fact genomic context may influence the decay of correlation; (W. Zhang et al., 2015) found differences in the decay of correlation between DNAm sites in CpG islands, and DNAm sites on the shelves and shores of CpG islands. (Garg, Joshi, Watson, & Sharp, 2018) took a slightly different approach, grouping local DNAm sites which were all variably methylated (>95th percentile of standard deviation) – not sites which were necessarily correlated. They found that these variably methylated regions (VMRs) were enriched for location in 3’UTR and Introns, and for chromatin states associated with enhancers, and deficient in chromatin states associated with promotors.

DNAm correlation structure has been discussed as being like the genetic phenomenon of LD (Linkage Disequilibrium, where genetic variants at different loci are inherited together (Slatkin, 2008)) as it can form correlated blocks, as LD does (Shoemaker et al., 2010). Some studies suggest that DNAm correlation structure is not related to genetic LD, as it does not mirror the large blocks that LD forms (Y. Liu et al., 2014), and it looks like it may be consistent across ethnic groups (Saffari et al., 2018); however there has not been a test of this across the whole genome. Correlations between DNAm sites may often be driven by genetic variants, as regions of highly correlating DNAm sites have been associated with nearby SNPs (Y. Liu et al., 2014; Shoemaker et al., 2010); although there has been no illustration to date of the extent of impact of these SNPs on the correlation structure. It has also been suggested that correlations between DNAm sites may be driven by environmental exposures (Garg et al., 2018).

1.5.2 Trans correlation structure

Trans correlation between DNAm sites is much less well defined. (Garg et al., 2018) found that when they correlated the most variable probe in each VMR they identified, genome-wide, there were strong *trans*-chromosomal correlations that corresponded to DNAm sites within the HOX gene clusters. These *trans*-correlating DNAm sites grouped together in a module when analysed using the network method WGCNA (details in section 1.6.1 below; (B. Zhang & Horvath, 2005)), suggesting that *trans*-chromosomal correlations between DNAm sites exist and are likely to be functional. These modules were enriched for cell-type specific gene ontology terms and transcription factor binding sites, and the trans-correlating sites had reduced heritability (although it should be noted this analysis this relates only to some of the most variable DNAm sites on the 450k). Another recent paper has shown that DNAm sites around inter- and intra-chromosomal chromatin contact points have correlated methylation states, with those within the same topologically associating domains and with the same chromatin states having more correlated methylation states (G. Li et al., 2019). This gives a substantial functional insight as to the reasons for correlation between distant DNAm sites.

1.6 Methods for DNAm network construction

There are many methods available for network construction from biological data. Network-based analysis methods for gene expression data have been developed for almost 20 years (Butte & Kohane, 2000; Carter, Brechbuhler, Griffin, & Bond, 2004; Stuart, Segal, Koller, & Kim, 2003; B. Zhang & Horvath, 2005). High throughput technologies to measure DNA methylation were developed almost 10 years ago; however for the most part gene expression network methods had been adopted or adapted for use with methylation data.

1.6.1 WGCNA

Perhaps the most widely used co-expression/co-methylation network method is WGCNA (Langfelder & Horvath, 2008; B. Zhang & Horvath, 2005). Originally developed for gene expression data, it has been used successfully for DNA methylation data in recent years, both by the developers and independent groups (Busch et al., 2016; Horvath et al., 2016; Langfelder et al., 2016; Spiers et al., 2015). The R package is regularly maintained, has extensive documentation, and comes with a number of easy-to-use tutorials on the group's UCLA website (Langfelder & Horvath, 2016). The method has numerous quite complex parameters, but is thoroughly explained both in the papers and on online resources. A distinct advantage of WGCNA is that it does not require a comparison group, so can be used on a single cohort; or it can be used to create consensus networks between two or more groups.

WGCNA constructs DNAm networks by conducting pairwise correlations between nodes (in this case, DNAm sites). Initially a correlation matrix is constructed between all possible pairs of nodes. To weight the network, the 'adjacency' of the nodes in the network is calculated by raising all correlation coefficients to a soft threshold power. Raising the correlation coefficient to a power means that high correlations will be emphasised, and small correlations will be minimised, without the need for an arbitrary hard threshold. The soft threshold power is based on scale-free topology because scale-free networks are thought to be more biologically relevant – scale free networks assume non-random connections between nodes, and that there are key hub nodes with many connections, which differs from other network theories (Albert, 2005; Barabasi, 2009; Barabasi & Albert, 1999). Finally, the topological overlap between the nodes is calculated. This represents the number of shared connections that nodes have, and is a more robust and biologically meaningful measure of network connectedness than the correlation between nodes (Xue et al., 2013).

WGCNA has been successfully used to generate many novel biological insights that would not have been possible with single site methods. Among many studies, WGCNA has been used to reveal how gene expression is organised in the human brain (Oldham et al., 2008); how networks of DNAm sites map fetal brain development; (Spiers et al., 2015); map the development of human and mouse embryos using RNA-seq data (Xue et al., 2013); identify novel, druggable targets in frontotemporal dementia (Swarup et al., 2019); and identify molecular networks affected by CAG repeat length in Huntington's (Langfelder et al., 2016). It has shown that DNAm sites associated with age can be detected across different platforms and tissues (Horvath et al., 2012); that networks of DNAm activity associate with functionally relevant pathways in brain tissue in autism spectrum disorder (Wong et al., 2019); and identify pathways by which childhood trauma is related to increased stress reactivity (Houtepen et al., 2016).

1.7 Thesis aims and objectives

The work I have discussed in this chapter leads to two strands of questioning. The first is delineating the actions and utility of pathways of DNAm after birth, their stability, and their relation to phenotypes that have been shown to be relevant to development. Network analysis has the potential to provide mechanistic insights about DNAm, but this has not yet been investigated relevant to normal post-natal development. The second strand relates to obtaining clarity on the underlying structure and meaning behind correlations between DNAm sites across the genome, and how this might change over time. To answer the first question fully, I felt it important to answer the second question first.

A great deal of work has identified individual DNAm sites associated with specific traits and diseases, particularly through the use of EWAS. However there is no clear picture of the biological pathways that DNAm might be involved in as part of normal development, and how this might change as we progress from birth, through childhood to adolescence. Network analysis provides an opportunity to answer these questions, as it can identify DNAm sites which may be co-regulated, and therefore whether DNAm might be involved in pathways of biological relevance. If we have a clear picture of any normative networks of DNAm activity, this would provide a valuable background for the associations DNAm may have with a phenotype of interest, or when it is perturbed by an exposure. To address this, I construct robust DNA methylation networks in two large, independent datasets. The first dataset, the Avon Longitudinal Study of Parents and Children (ALSPAC) (Boyd et al., 2013), is a longitudinal cohort with DNAm measured at three timepoints, from birth to adolescence, in around 900 children. This longitudinal analysis is presented in Chapter 4. To assess the reproducibility of these networks, and the preservation of these networks between different ethnicities, I use the birth cohort Born in Bradford (BiB) (Wright et al., 2013). I present this analysis in Chapter 6.

I have selected a small number of phenotypes to test for association with the DNAm modules, that have previously been associated with DNAm in single site analyses. I selected some to do with physical growth (gestational age, birthweight, BMI); two exposures (maternal smoking, and either socio-economic position or deprivation; deprivation measures were only available in BiB and this was closer to the exposure I am interested in); and a childhood disease that changes over development (asthma).

To understand the basis of these co-methylation networks more fully, however, some clarity is needed on the drivers of DNAm correlation. *Cis* correlation has been relatively well delineated, but to date no study has interrogated *cis* correlation structure across the whole Illumina 450k array. There has also been no analysis of how this correlation structure might change over time, which is a key question that this thesis will address. Further work is also needed on the genomic features these sites are enriched for. *Trans* correlation structure has not been well studied across the genome, and this is quite important to understanding WGCNA networks, because most of the correlations going into the network will be in *trans*. If network modules comprise DNAm sites on different chromosomes, it is important to know what these represent. Currently one study in human cells suggests they may represent biologically meaningful regulation, but this only looked at a small subset of sites on the 450k array (Garg et al., 2018). Mouse cell lines show that it is possible that these trans correlations to represent inter-chromosomal chromatin contacts (G. Li et al., 2019), but this has yet to be demonstrated in humans. To address these questions, I conduct an analysis of DNAm correlations across the genome (as measured by the 450k array), first in ALSPAC (Chapter 3), and then in BiB (Chapter 5).

I aim to:

1. Describe the full correlation structure over the 450k array, in *cis* and *trans* (Chapters 3 and 5).
2. Identify the main biological features of highly correlating sites (Chapters 3 and 5).
3. Identify whether this correlation structure changes over time, between birth and adolescence (Chapter 3).
4. Validate this correlation structure in a separate cohort, to identify whether it is likely to be stable (Chapter 5).
5. Identify whether this correlation structure is preserved across ethnicities (Chapter 5).
6. Construct DNAm networks, to identify whether biological networks that might be related to normal development can be detected in DNAm data (Chapters 4 and 6).

7. Use the insights about correlation structure to interpret the network results (Chapters 4 and 6).

Analysis chapter structure

Chapter 2: General methods

Chapter 3: Drivers of correlation between DNAm sites

Chapter 4: Systems biology network analysis

Chapter 5: Assessing the stability and reproducibility of DNAm correlation structure

Chapter 6: Validation of systems biology networks

Chapter 7: Discussion

2 General methods

2.1 Summary

This chapter describes the cohorts and data used in this thesis, and all methods which are applicable to two or more chapters.

2.2 Cohort descriptions

2.2.1 ALSPAC

The Avon Longitudinal Study of Parents and Children (ALSPAC) is a multi-generational cohort study based in Bristol. ALSPAC was set up with the aim to investigate factors that influence child health and development, and has collected data about a great number of exposures and outcomes related to this, including genetic and epigenetic data.

All pregnant women in the former Avon region with a delivery date between April 1991 and December 1992 were eligible to take part in the study, and 14,541 pregnancies were initially recruited, with 14,062 live births (Boyd et al., 2013). Recruitment numbers were increased some years later, where individuals who were eligible at the start of the study but did not join were contacted (as long as they had not declined to take part in the study initially). 913 more children were recruited, giving a total of 15,454 pregnancies recruited, and 14,901 participants alive at one year old (Northstone et al., 2019).

Data has been collected frequently from ALSPAC participants, from pregnancy through to the present day, through questionnaires, clinics, biological samples and linkage data (Boyd et al., 2013). The ALSPAC website contains a searchable data dictionary, containing all available data on the cohort, available at http://www.bristol.ac.uk/alspac/researchers/our-data/

Ethical approval for the study was obtained from the ALSPAC Ethics and Law Committee and the Local Research Ethics Committees.
http://www.bristol.ac.uk/alspac/researchers/research-ethics/ , proposal number B2808.

2.2.1.1 ARIES

ARIES (Accessible Resource for Integrated Epigenomic Studies) is a sub-sample of the ALSPAC study, described in detail by (Relton et al., 2015). This sub-sample has DNA methylation (DNAm) data, and so this thesis uses only this sub-sample of ALSPAC.

ARIES consists of 1022 mother-child pairs from ALSPAC, who were selected because they had appropriate DNA samples for DNAm profiling from specified timepoints. Participants included in ARIES are reasonably representative of those in ALSPAC as a whole; however ARIES mothers were slightly older, less likely to smoke during pregnancy, and more likely to have a non-manual occupation (Relton et al., 2015).

DNA methylation was profiled at two timepoints for the mothers: during pregnancy, and at follow-up approximately 15-17 years later. In the children, it was profiled at birth (using a sample of umbilical cord blood), during childhood (mean age 7.5) and during adolescence (mean age 17.1 years) (Relton et al., 2015). This resulted in 5469 DNAm profiles.

2.2.1.2 DNA methylation data generation: ARIES

Consent for biological samples was collected in accordance with the Human Tissue Act (2004). Standard procedures were used to collect blood samples. At birth, all samples were cord blood. All other timepoints were whole blood or buffy coat. DNA was extracted from samples and bisulfite converted with the Zymo EZ DNA MethylationTM kit (Zymo, Irvine, CA). DNA methylation was then profiled using the Illumina Infinium® HumanMethylation450K BeadChip (subsequently referred to as the Illumina 450k array), using the standard protocol. An Illumina iScan was used to scan the arrays, and GenomeStudio (version 2011.1) was used for initial data quality review (Relton et al., 2015). The Beta-value statistic was used to represent methylation levels. Low quality profiles were removed, leaving 4593 for further analysis.

All ARIES samples were run at the same time, so samples from all timepoints were semi-randomised across Beadarrays (referred to from here as slide) , to minimise the likelihood of batch effects inducing differences between timepoints. A semi-randomisation procedure was used rather than a fully random one, in order that that every array contained samples from every timepoint.

2.2.1.3 Removal of outlying samples: ARIES

Before normalisation, 615 outlying samples were removed from the dataset. The details of outlier removal can be found in (Min, Hemani, Davey Smith, Relton, & Suderman, 2018). I also removed 21 individuals from the full ARIES dataset as they were the only sample on a slide, and single observations in a category cause problems when running regression in R.

After these samples were removed, the following number remained for analysis:

Timepoint	**Number of participants**
Birth	849
7 years	910
15-17 years	921
Number of participants with data at all 3 timepoints	788

Table 1: Number of ARIES participants after removal of outlying samples and single samples on a slide

2.2.1.4 DNA methylation data normalisation: ARIES

All ALSPAC timepoints were normalised together, as they were all run at the same time. Timepoints were randomised over slides to minimise the likelihood of batch effects inducing differences between timepoints.

Methylation data were normalised using the Functional Normalization algorithm (Fortin et al., 2014) implemented in `meffil`, using the top 10 principal components from the control probes; full details of this have been published in the `meffil` paper (Min et al., 2018). This detects unwanted technical variation by identifying variability that is explained by the control probes on the 450k array, and regressing it out from the data. It is applied separately to the type I and type II probes, and separately to the Meth and Unmeth intensities separately.

Slide effects were notable even after normalisation, so slide was regressed out from the raw methylation betas before normalisation. This was to remove some of the slide effects so that the technical artefacts could be better captured by the control probes. Slide row has also been shown to affect DNAm data; however this is picked up by the staining control probes, and was regressed out during the normalisation.

2.2.1.5 Tissue sample type: ARIES

Blood samples in ARIES were a mix of blood spots, white cells, peripheral blood lymphocytes (PBL), and whole blood. Unfortunately for the ARIES cord blood samples, sample type (either blood spots or white cells) was confounded by slide; slides only contained cord samples of either blood spots or white cells. The numbers of tissue sample types are shown in Table 2. It is important to note that the differences in cell types and their proportions at the three timepoints in ARIES are likely to impact the results and their interpretation. This is because having different cell populations present will inevitably lead to differences in measured methylation, and cell type correction panels are unlikely to completely adjust for this. Possible impacts of this are discussed where appropriate in the results and the discussion.

Timepoint	Bloodspots	White cells	Whole blood
Birth	154	695	0
7 years	0	57	853
15-17 years	0	921	0

Table 2: Table of sample type numbers in ARIES

2.2.1.6 Generating cellular composition estimates: ARIES

Cellular composition was estimated from the normalised beta matrix by Josine Min, using `meffil` (Min et al., 2018) which implements the houseman algorithm (Houseman et al., 2012). Cell counts for the birth timepoint were estimated using the Gervin reference panel (Gervin et al., 2016). This estimates proportions of B cells, CD4+ T cells, CD8+ T cells, Natural Killer cells, CD14+ Monocytes, and Granulocytes. Cell counts for the 7 and 15-17 year-olds were estimated using the (Reinius et al., 2012) panel, which also estimates proportions of B cells, CD4+ T cells, CD8+ T cells, CD14+ Monocytes, Natural Killer cells, and Granulocytes.

To investigate the relationship between cell count proportions and DNA methylation network modules, I estimated cell counts for all timepoints from the normalised betas using the (Reinius et al., 2012) complete reference panel, because this separates the Granulocytes to Eosinophils and Monocytes, and eosinophils are relevant to the phenotypes that were tested.

2.2.1.7 Study design: ARIES

2.2.1.7.1 Filtering DNAm sites

I used the (Zhou et al., 2017) list of probes to exclude sub-optimal DNAm sites from the analysis (the need for which is discussed in introduction section 1.3.1.1). This list excludes all probes with a SNP with MAF >1% within 5 bases of the 3' end of the probe, probes where the 30bp 3'-subsequence of the probe is not unique, probes where there is a mapping issue, probes where there is a SNP in the extension base, and probes which have an extension base inconsistent with their mapping.

I also removed DNAm sites which were identified by GoDMC and TwinsUK (Josine Min, personal communication), as multi-mapping probes (bisulfite converted sequences allowing two mismatches at any position mapped to the hg19 primary assembly, i.e. no haplotypes included as in Naeem), probes with variants (MAF >5%, UK10K) at the CpG dinucleotide or the extension base (for type I probes), and any probes targeting non CpG sites that failed liftover to hg19.

2.2.1.7.2 Removing outlying methylation values

Removing outlying measurements from DNA methylation data is often performed as extreme outliers can skew analyses. For each DNAm site I removed observations that were more than 10 standard deviations from the mean, repeating this process three times to remove sufficient outliers. Where outliers were removed, they were replaced by the mean for that probe. There is no clear consensus about whether outliers should be removed from DNAm data. Recent work suggests that outlying DNAm values represent rare genetic variants (Chundru et al., 2020), and so they are likely to be informative. However the aim of the present analysis is to aggregate DNAm measurements across nearly 1000 samples, and so outliers caused by rare genetic variants will simply serve to skew the estimates without contributing information useful to the analysis. The outlying values were replaced with the mean for that probe for practical reasons, as missing values would cause problems in downstream analyses. As Table 3 shows, there were a small number of outliers in comparison to the size of the methylation matrix (which is around 900 x 394,842); equivalent to approximately one quarter of the DNAm sites having one outlying value.

Timepoint	Outlier count
Birth	115,971
7 years	110,742
15-17 years	104,458

Table 3: Count of outlying values for each timepoint in ARIES

2.2.1.7.3 Adjusting DNAm data for known covariates

DNAm data are known to be affected by a number of covariates, and so it is important to adjust for these. All known variables were adjusted for by entering them into a linear regression model with the methylation data in R, using the `lm` function (as shown below). The residuals from this regression were taken forward for all analyses.

Even though slide was regressed out before normalisation, slide effects were detectable after normalisation, and so slide was regressed out from the normalised methylation betas:

```
residuals(lm(norm.beta ~ slide))
```

Sample type was regressed in each timepoint that used more than one sample type (birth and 7 years; this is coded as `cellType` below). Sex was regressed out as a categorical variable at all ARIES timepoints. Age was regressed out at each individual timepoint (apart from at birth), to remove any effects that might be due to age rather than differences in methylation between individuals. Estimated blood cell counts (estimated as detailed in section 2.2.1.6) were also adjusted for in the model:

```
cordData.residuals <- residuals(lm(cordData ~ sex + cellType +
Bcell + CD4T + CD8T + CD14 + NK + Gran))
F7Data.residuals <- residuals(lm(F7Data ~ sex + age + cellType
+ Bcell + CD4T + CD8T + Mono + NK + Gran))
I5upData.residuals <- residuals(lm(I5upData ~ sex + age +
Bcell + CD4T + CD8T + Mono + NK + Gran))
```

2.2.1.8 Genotype data generation: ARIES

ARIES participants were genotyped as part of the main ALSPAC study. All ALSPAC child participants were genotyped with the Illumina HumanHap550 quad genome-wide SNP array (Illumina Inc., San Diego, CA) by the Laboratory Corporation of America (LCA, Burlington, NC, USA) and the Wellcome Trust Sanger Institute (WTSI, Cambridge, UK), supported by 23andMe (Relton et al., 2015). Participants were excluded if they had the incorrect gender assigned, if there was abnormal heterozygosity (defined as <0.310 or >0.330 for the LCA data, and <0.320 or >0.345 for the WTSI data), high missingness (>3 %), if there was cryptic relatedness (>10 % identity by descent), and if the individual was of non-European ancestry (which was detected by multidimensional scaling analysis). After QC, the dataset consisted of 500,527 directly genotyped SNP loci.

SNP data were then imputed to increase SNP density. They were imputed using the 1000 Genomes reference panel (phase 1, version 3, phased using SHAPEIT (version 2, December 2013) (Delaneau, Marchini, Genomes Project, & Genomes Project, 2014), using all populations (Genomes Project et al., 2015)), using IMPUTE (v2.2.2) (B. Howie, Marchini, & Stephens, 2011; B. N. Howie, Donnelly, & Marchini, 2009). Genotypes were retained if they had Hardy Weinberg equilibrium $p > 5e{-7}$, a minor allele frequency of more than 1%, and an imputation info score over 0.8.

2.2.1.9 Phenotype data: ARIES

A range of phenotypes in ARIES were selected to test for their association with network modules. I selected phenotypes which are known to have an impact on DNAm in children, to ascertain whether additional insight can be gained into the role of DNAm (i.e., might DNAm be part of a regulatory or affected pathway that cannot be detected by single site analyses?). I selected phenotypes which represent growth, to ascertain whether pathways of DNAm might have a role in physical growth. For this I selected birthweight, gestational age, and BMI. I also selected phenotypes which are exposures known to have an impact on DNAm in children, to ascertain whether these exposures might affect development as part of a biological pathway. I selected exposure to maternal smoking in pregnancy, own smoking in the teenagers, maternal pre-pregnancy BMI, maternal age at delivery, and household class. I also chose to test whether asthma affects DNAm in a concerted manner.

Summary measures of phenotypes in ARIES that were tested for association with network modules over the three timepoints are identified in Table 4. Measures which were either a composite of a number of variables, transformed before analysis, or had covariates added in to the analysis, are described in detail below in section 2.2.1.9.1.

Measure	Birth	7 years	15-17 years
Number of participants	820	859	887
% female	51.6%	50.8%	51.4%
Age (years)	-	7.45 (0.15)	17.7 (0.4)
Maternal age	28.4 (5)	-	-
Gestational age (weeks)	39.6 (1.4)	-	-
Birthweight (g)	3500.7 (467.1)	-	-
Maternal BMI (kg/m2)	22.9 (3.8)	-	-
BMI (kg/m2)	-	16.2 (2.1)	22.6 (3.9)
Sustained maternal smoking (% yes)	10.1%	10.4%	10.3%
Own smoking (% yes)	-	-	8.7%
Household class (high)	61%	60%	61%
Household class (middle)	31%	31%	30%
Household class (low)	4%	4%	4%
Asthma (% yes)	11.2% (at 7 years)	12.5%	11%

Table 4: Table summarising the phenotypes in ARIES for each of the three timepoints. Measurements are mean (SD) unless otherwise stated.

2.2.1.9.1 Composite phenotypes

2.2.1.9.1.1 Sustained maternal smoking

Sustained maternal smoking during pregnancy was assessed using a previously developed composite measure provided by Rebecca Richmond which has been used in a number of publications (JoubertFelix, et al., 2016; Richmond et al., 2015). As (Richmond et al., 2015) showed that maternal smoking has a lasting impact on DNAm at certain sites, it was tested in all age groups. Sustained maternal smoking was defined as the mother smoking for at least two of the three trimesters of pregnancy. Covariates used, as in the PACE analysis, were maternal age at birth and maternal SES measured by the household class variable (grouped to low, middle, and high; see section 2.2.1.9.1.4). At 15-17 years, own smoking was also used as a covariate.

2.2.1.9.1.2 Own smoking

Own smoking was assessed using a previously reported measure (Prince et al., 2019). 15-17 year olds were classified as smokers if they reported smoking more than one cigarette per week at the TF3 clinic, and classified as non-smokers if they reported that they had never tried a cigarette. The covariate used for this analysis was sustained maternal smoking.

2.2.1.9.1.3 Asthma

2.2.1.9.1.3.1 7 years

The asthma measurement at 7 years was derived from data from the ALSPAC Focus at 7 clinic visit. The measure was a composite used in the ARIES contribution to the PACE consortium EWAS of school-age asthma (Reese et al., 2019). It comprised measures of doctor diagnosed asthma, any asthma medication in the last 12 months, and asthma in the last 12 months. Participants were classed as having asthma only if they answered yes to doctor diagnosed asthma plus yes to either asthma medication in the last 12 months or to asthma in the last 12 months. Code to extract this variable was provided by Kimberly Burrows. Covariates used in asthma associations were those used in the PACE analysis. They were maternal age at child's birth (years), maternal socio-economic status as measured by the household class variable (grouped to low, middle, and high; see section 2.2.1.9.1.4), maternal asthma (yes, no), and sustained maternal smoking (yes/no).

2.2.1.9.1.3.2 15-17 years

As there were more limited asthma measures at the 15-17 year old clinics, the exact measure used at 7 years could not be replicated in the adolescents. I derived the measure from the ALSPAC TF3 clinic, and classed participants as having asthma if they answered yes to both suffering from asthma and having doctor diagnosed asthma. Covariates used in asthma associations were those used in the PACE analysis. They were maternal age at child's birth (years), maternal socio-economic status as measured by the household class variable (grouped to low, middle, and high; see section 2.2.1.9.1.4), maternal asthma (yes, no), and sustained maternal smoking (yes/no).

2.2.1.9.1.4 Household socio-economic status

Household socio-economic status comprised a measure derived from self-report of the mother's occupation, using the Registrar General's Social Classes (based upon SOC 2000 codes). As in (JoubertFelix, et al., 2016), the 6 classes were condensed to high (1 and 2), middle (3 and 4) and low (5 and 6).

2.2.1.9.2 Transforming BMI phenotypes

Both the BMI and maternal BMI phenotypes had a skewed distribution, and so before analysis they were transformed using the rank-based inverse normal transformation. To do this I used the R package `RNOmni` (McCaw, 2019), removed NAs from the dataset, and ran the code:

```
datTraits$bmi_transf = rankNorm(datTraits$bmi)
```

2.2.1.9.3 Simple phenotypes with covariates

2.2.1.9.3.1 Birthweight

For the birthweight phenotype, all individuals who were born preterm (defined as <35 weeks) were removed from the analysis. This comprised 6 individuals. Birthweight was adjusted for gestational age, maternal age, maternal smoking during pregnancy, maternal BMI, and socio-economic status (household class) as in (Kupers et al., 2019).

2.2.1.9.3.2 Gestational age

Gestational age was transformed using the rank-based inverse normal transformation. To do this I used the R package `RNOmni` (McCaw, 2019), removed NAs from the dataset, and ran the code:

```
datTraits$gestage = rankNorm(datTraits$gestage)
```

2.2.2 Born in Bradford

Born In Bradford (BiB) is a longitudinal, multi-ethnic cohort study based in Bradford, UK. Like ALSPAC it was set up to investigate factors which influence child health and development, but with a particular focus on child morbidity and mortality, as rates of these have been higher in Bradford than the rest of the UK (Wright et al., 2013). Bradford has a high rate of economic deprivation – one third of the neighbourhoods in Bradford are in the most deprived 10% of neighbourhoods in England (Department for Communities and Local Government, 2015; T. Smith et al., 2015). Around 20% of the population are of South Asian descent (Wright et al., 2013), so this cohort also provides the opportunity to study correlation between DNAm sites in a non-European ethnic group.

All women booking an oral glucose tolerance test (OGTT) at the Bradford Royal Infirmary at 26-28 weeks' gestation (around 80% of expectant mothers in Bradford) were invited to take part in the study. At the OGTT appointment, consent was obtained from those who were willing to take part (about 80% of those attending), and those who did not attend the OGTT were contacted at other hospital appointments. In total 12,453 women with 13,776 pregnancies were recruited, leading to 13,740 live births.

Ethical approval for this study was granted by the Bradford Research Ethics Committee (Ref 07/H1302/112). Written informed consent was obtained from the mothers (for themselves and their children) when they registered for the study.

2.2.2.1 Subsample of BiB with DNAm data

A subsample of 1000 mothers and their children in BiB had DNAm data generated from blood samples during pregnancy in the mothers, and from cord blood in the children. Eligibility for this subsample was defined as mothers who had both completed the OGTT, and had genetic data available (which comprised around 65% of the individuals with the completed OGTT). This subsample was specifically designed to be multi-ethnic, and so of those eligible, 500 White British and 500 Pakistani mothers were selected to have DNAm generated for themselves and their children.

2.2.2.2 Ethnicity

Born in Bradford was designed as a multi-ethnic study. As mentioned in the previous section, the DNAm subsample was designed to have equal representation of two ethnicities. In this thesis I use the term ethnicity to describe the white British and Pakistani participants, because ethnicity was assessed through self-report questionnaires completed by the mother. The response contained 9 ethnic groups, of which all participants with DNAm data were either white British or Pakistani. Chapter 1, section 1.2.4 details the motivations for studying ethnicity in DNAm studies, and Chapters 5 and 6 discuss the specific motivations for the analyses in this thesis using individuals of different ethnicities.

2.2.2.3 DNA methylation generation: BiB

To generate DNAm data, the EZ-96 DNA methylation kit (Zymo Research, Orange, CA, USA) was used to bisulfite-convert 500ng of high molecular weight DNA. DNAm was assessed using the Illumina Infinium MethylationEPIC beadchip arrays (Illumina, San Diego, CA, USA). Batch variables were recorded using a laboratory information management system (LIMS). The Beta-value statistic was used to represent methylation levels. 2010 samples (mothers, children and control samples) were generated together. 98 failed genotype concordance checks, and 13 failed other QC checks, and as some of these individuals overlapped a total of 100 individuals removed during QC. This left 1910 samples for normalisation, comprising 25 controls, 934 mother and 951 child participants. Please see Figure 3 for a summary of BiB participant numbers.

2.2.2.4 Normalising methylation data: BiB

BiB DNAm betas were generated from idat files by Nabila Kazmi (senior research associate, Bristol IEU; see acknowledgements), using `meffil` (Min et al., 2018). As the DNAm betas were originally normalised with lab control samples and outlying samples. I used the meffil QC object generated by Nabila to identify a total of 29 unique outlying samples (which overlapped with 1 control sample) which mapped to 34 issues (as detailed in Table 5 below). I removed these samples, along with the lab control samples, from the QC object. This left 922 mothers and 935 children. The child samples 451 white British children and 484 Pakistani children. I calculated 15 PCs were required to re-normalise the samples, and I then used `meffil.normalize.samples` to create a normalised beta matrix using the idat files. `Meffil` uses functional normalization (Fortin et al., 2014) to normalise samples.

Issue	**Number**
X-Y ratio outliers	6
Methylated vs unmethylated intensity	18
Control probe dye bias	9
Bisulfite conversion control probe issue	1

Table 5: Summary of outlying samples in BiB

The normalisation report showed that the first 3 DNAm PCs largely described ethnicity and whether the sample was a mother or a child (shown in Figure 2. There were still associations with slide and a number of the PCs even after normalisation, and so slide was regressed out as part of adjusting for covariates (section 2.2.2.7.3).

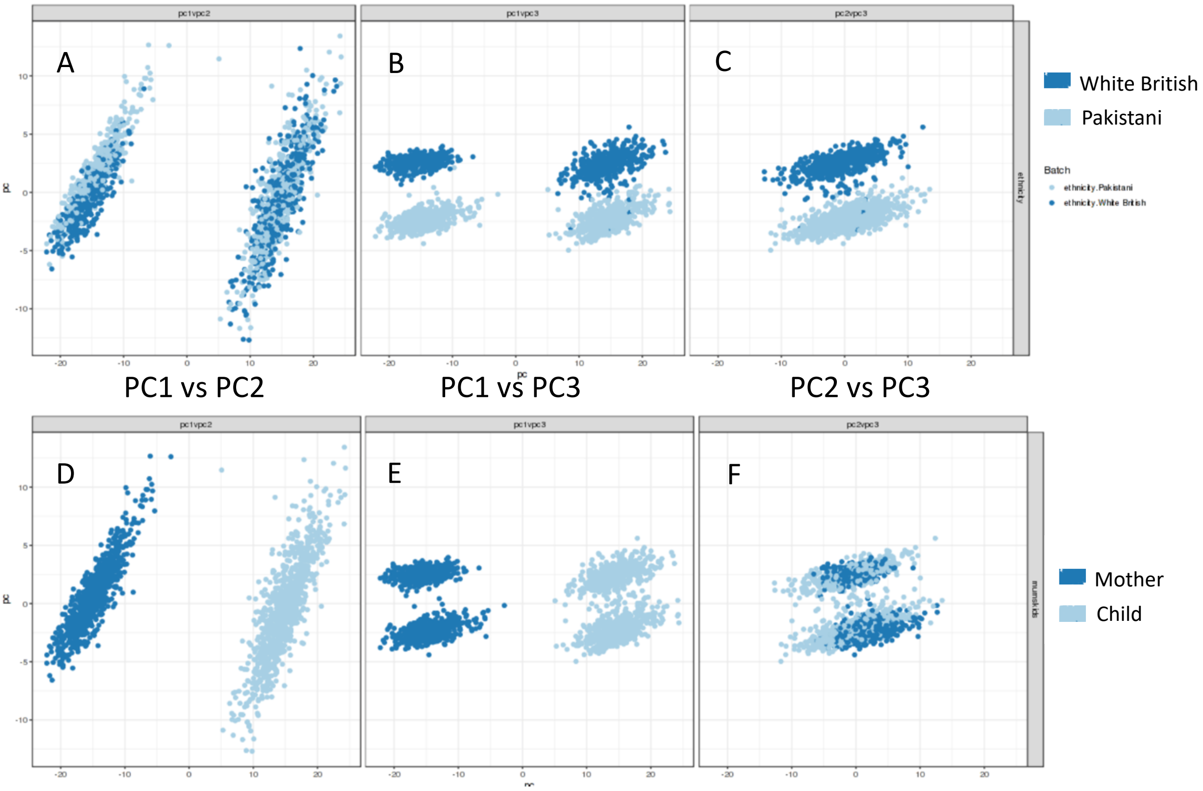


Figure 2: Plots of the first three normalisation PCs. ***A-C****: PCs 1-3 coloured by ethnicity, dark blue = white British, light blue = Pakistani.* ***D-F****: PCs 1-3, coloured by mothers/children, dark blue = mothers, light blue = children.*

2.2.2.5 Removing related individuals

Relatedness can bias both methylation and genetic analyses. To avoid this I removed all individuals identified as being related >12.5% from all analyses. The genetic data is described in detail in section 2.2.2.7. To identify related individuals, I first split the genetic data by ancestry, and by mothers and children. This resulted in 4 datasets, each of which was subset to the hapmap 3 SNPs. Using `plink` version 1.90 (Purcell et al., 2007), I created a genetic relatedness matrix (GRM) for each dataset, with a minor allele frequency cutoff of 0.01. I then used the GRM to remove all individuals related above a cutoff of 0.125 from the genetic data. For all analyses, the individuals related above 0.125 were also removed from the DNAm data. This left 424 white British children and 439 Pakistani children for further analysis. This is illustrated in Figure 3.

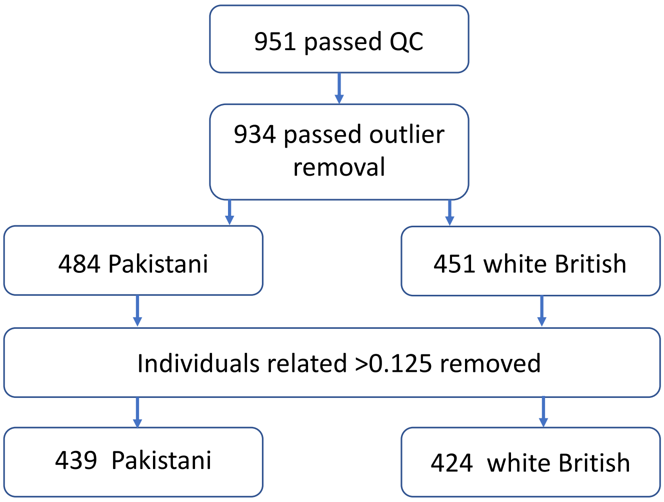


Figure 3: Flow diagram of BiB child participant numbers

2.2.2.5.1 Regressing out population stratification

It is important to regress out population stratification, because population stratification can lead to false positive results in DNAm data, just as it can do in genetic studies (Barfield et al 2014). Once the related individuals (identified from genetic data as detailed in section 2.2.2.5 above) were removed from the DNAm data, I split the data by ethnicity and by mothers and children (resulting in 4 DNAm datasets). In each dataset, using `plink` hapmap3 SNPs were extracted and pruned to be in linkage equilibrium using `--indep-pairwise`, with a 10,000 variant window, a 5 variant window shift, and a pairwise r^2 threshold of 0.1, and a MAF of 0.2. I then calculated the first 20 genetic PCs of each dataset separately using `plink`.

Figure 4 shows the plots of the first 3 genetic PCs for each of the groups. From these we can see that population stratification is more pronounced in the Pakistani group, but it is also present to a degree in the white British group, so I decided to regress out population stratification from all BiB datasets. To remove population stratification from the methylation data, I regressed out the first 20 PCs of the genetic data using `lm` in R, and took the residuals forward for analysis.

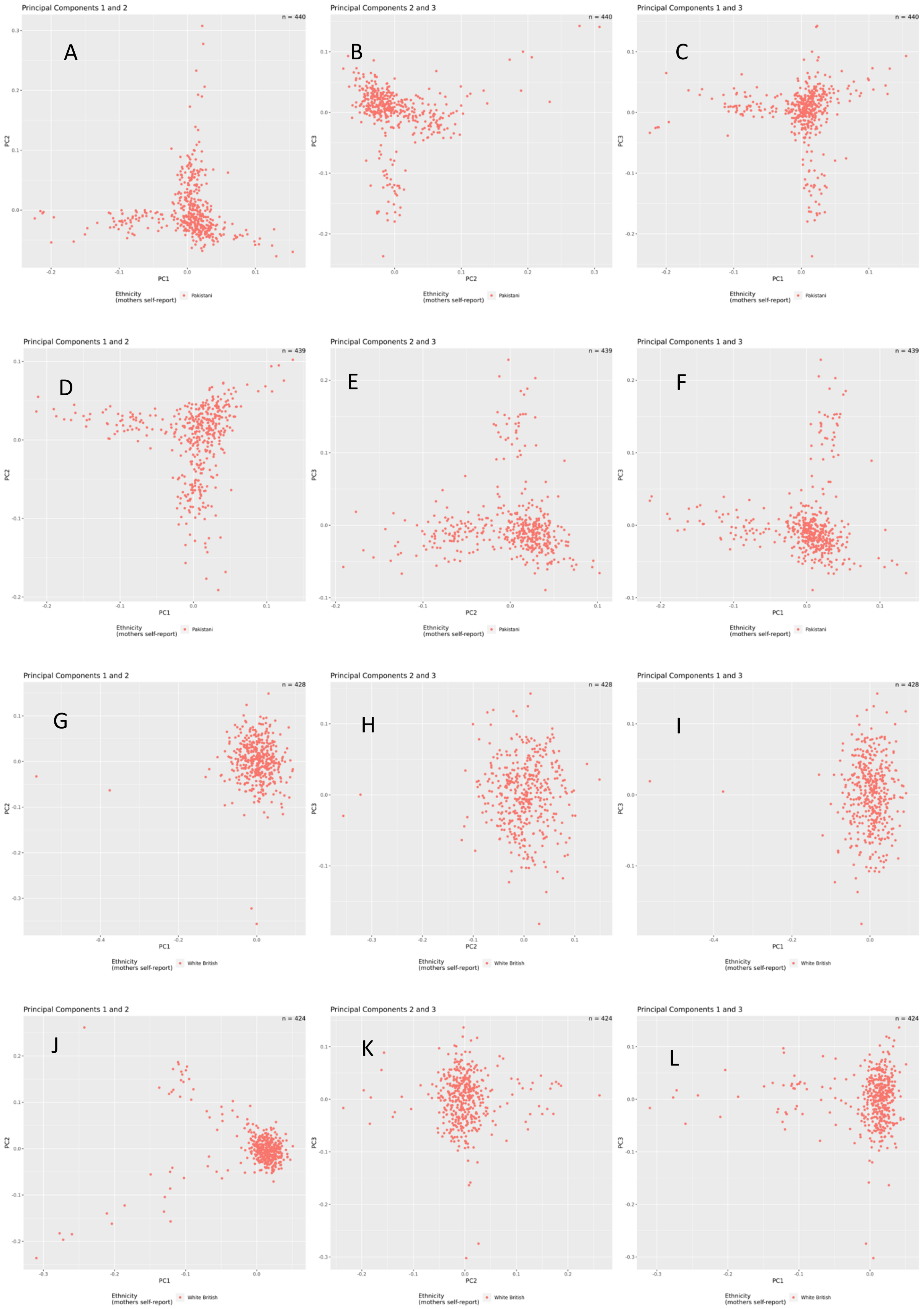


Figure 4: First three genetic PCs of the Born in Bradford data, illustrating population stratification. A-C Pakistani mums, D-F Pakistani children, G-I white British mums, J-L white British children

To confirm the ethnicity of the BiB genetic samples, genetic principal components were overlaid with samples from the 1000 Genomes project. The analysis followed parts 1 and 2 of the tutorial by (Marees et al., 2018), with scripts and data available through the GitHub repository linked to in the paper. The plots show that the white British group overlap with the European population, as expected; however the Pakistani group are clustered very close to the European and American groups, which one would not expect (please see Figure 5). This is clearly highlighting an issue with one of the processing steps taken in this work, and so further work needs to be done to ascertain the reasons for this; this may well provide the reason for the decay plot in BiB not working (as the issue could have affected the genetic data for both ethnic groups).

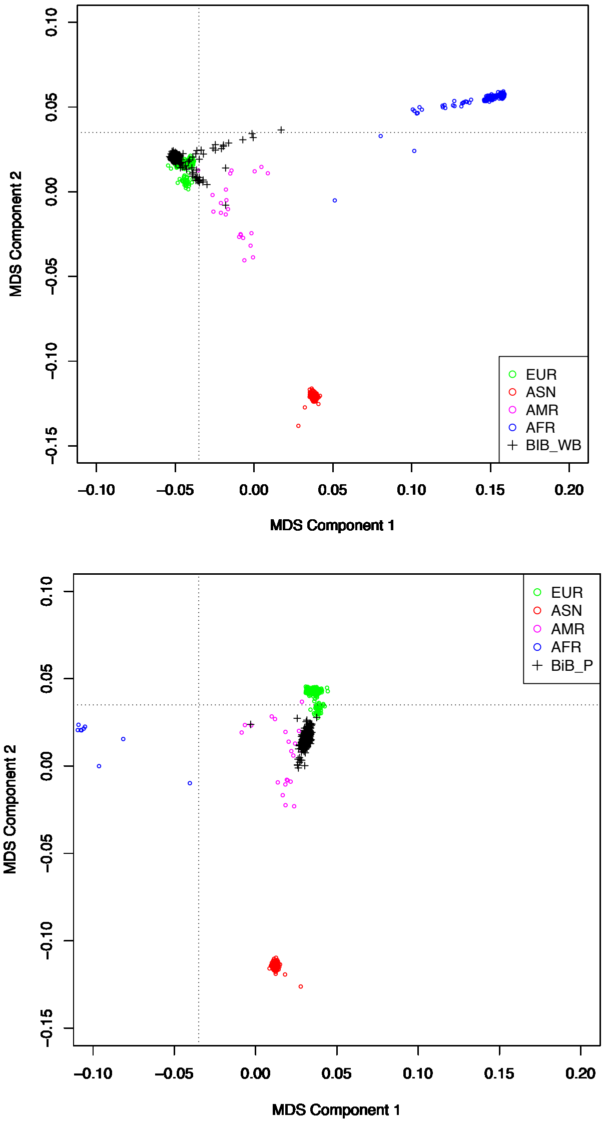


Figure 5: Plots overlaying genetic principal components for BiB white British (top) and BiB Pakistani (bottom) participants with genetic principal components from four 1000 Genomes ethnic groups

2.2.2.6 Generating cellular composition estimates: BiB

I estimated cellular composition from the normalised beta matrix, using `meffil's` (Min et al., 2018) function `meffil.estimate.cell.counts.from.betas`. The houseman algorithm (Houseman et al., 2012) is used to estimate cell counts with this function. I used the Andrews and Bakulski cord blood reference panel (Bakulski et al., 2016) to create the cord blood cell count estimates.

2.2.2.7 Study design: BiB

2.2.2.7.1 Filtering DNAm sites

I used the (Zhou et al., 2017) list of probes to exclude sub-optimal DNAm sites from the analysis (the need for which is discussed in introduction section 1.3.1.1). This list excludes all probes with a SNP with MAF >1% within 5 bases of the 3' end of the probe, probes where the 30bp 3'-subsequence of the probe is not unique, probes where there is a mapping issue, probes where there is a SNP in the extension base, and probes which have an extension base inconsistent with their mapping. In using the (Zhou et al., 2017) list, which was designed for the 450k array, I also excluded probes on the EPIC array that were not on the 450k.

I also removed DNAm sites which were identified by GoDMC and TwinsUK (Josine Min, personal communication), as multi-mapping probes (bisulfite converted sequences allowing two mismatches at any position mapped to the hg19 primary assembly, i.e. no haplotypes included as in Naeem), probes with variants (MAF >5%, UK10K) at the CpG dinucleotide or the extension base (for type I probes), and any probes targeting non CpG sites that failed liftover to hg19.

2.2.2.7.2 Removing outlying methylation values

Removing outlying measurements from DNA methylation data is often performed as extreme outliers can skew analyses. For each DNAm site I removed observations that were more than 10 standard deviations from the mean, repeating this process three times to remove sufficient outliers. Where outliers were removed, they were replaced by the mean for that probe.

2.2.2.7.3 Adjusting for known covariates

DNAm data are known to be affected by a number of covariates, and so it is important to adjust for these. All known variables were adjusted for by entering them into a linear regression model with the methylation data in R, using the `lm` function. The residuals from this regression were taken forward for all analyses.

Even though slide was regressed out before normalisation, slide effects were detectable after normalisation, and so slide was regressed out from the normalised methylation betas. Sex was regressed out as a categorical variable at all ARIES timepoints. Blood cell counts (estimated as detailed in section 2.2.2.6) were also adjusted for in the model:

```
BiB.residuals <- residuals(lm(BiB.m ~ sex + Bcell + CD4T +
CD8T + Mono + NK + Gran + nRBC))
```

2.2.2.8 Genotype data: BiB

Samples in BiB were genotyped using either the Illumina HumanCoreExome Exome-24 v1.1 microarray, or the Infinium global screen-24+v1.0 array. GenomeStudio 2011.1 was used to pre-process samples. If samples had a call rate of <0.95, they were excluded. Poorly performing SNPs were removed. Most multi-allelic SNPs were discarded. 459,340 SNPs remained, and these were imputed by the Sanger Impute Service using the 1000genomes and UK10K reference panels (as 1000 genomes contains a number of different ethnicities).

2.2.2.9 Phenotype data: BiB

As far as was possible I selected the same phenotypes in BiB as I did for ARIES (relevant to birth), so that I could directly compare the relationships of the DNAm networks to the phenotypes across two cohorts, and across ethnicities. I selected gestational age, exposure to maternal smoking in pregnancy, own smoking in the teenagers, maternal pre-pregnancy BMI, maternal age at delivery, and index of multiple deprivation (which is related to, but not the same as, socioeconomic status).

Summary measures of phenotypes in BiB that were tested for association with network modules are identified in Table 6, and are presented by ethnicity. Measures which were either a composite of a number of variables, transformed before analysis, or had covariates added in to the analysis, are described in detail below.

Measure	White British	Pakistani
Number of participants (N)	416	429
% female	48%	48%
Maternal age	26.6 (6.2)	27.9 (5.3)
Gestational age (weeks)	39.7 (1.8)	39.6 (1.4)
Maternal BMI (kg/m2)	27 (6.2)	25.8 (5.2)
Maternal smoking (% yes)	34%	0%
Index of Multiple Deprivation	36.9 (19.7)	46.9 (14.8)

Table 6: Table summarising the phenotypes in BiB for each of the ethnicities separately. Measurements are mean (SD) unless otherwise stated.

2.2.2.9.1 Maternal smoking

Maternal smoking during pregnancy was assessed using a yes/no response to the question of whether they had smoked during pregnancy.

2.2.2.9.2 Maternal BMI

Maternal BMI was transformed using the rank-based inverse normal transformation. To do this I used the R package `RNOmni` (McCaw, 2019), removed NAs from the dataset, and ran the code:

```
datTraits$matbmi_transf = rankNorm(datTraits$matbmi)
```

2.2.2.9.3 Gestational age

Gestational age was transformed using the rank-based inverse normal transformation. To do this I used the R package `RNOmni` (McCaw, 2019), removed NAs from the dataset, and ran the code:

```
datTraits$gestage = rankNorm(datTraits$gestage)
```

2.2.3 GoDMC

The Genetics of DNA Methylation Consortium (GoDMC, www.godmc.org.uk) was set up with the aim of delineating the genetic basis of variation in DNAm. It was designed to bring experts in epigenetics and their resources together, to carry out a large meta-GWAS of DNAm. Cohorts included in the consortium had DNAm measured using the Illumina 450k or

EPIC Beadchip arrays. The GoDMC analyses are available and can be requested from http://www.godmc.org.uk/projects.html.

For this thesis I used the results of the meta-GWAS. This included 36 cohorts (with at least 100 individuals in each cohort) with 27,750 participants of European ancestry. The cohorts are not described in this thesis, but they are described in detail on the GoDMC website. The analysis is described briefly below.

2.2.3.1 Genotype data: GoDMC

All autosomes, and chromosome X if it was available, were imputed to the 1000genomes reference panel, using hg19/build37. Filters of a minor allele frequency of 0.01 and info score of 0.8 were applied. The genotype data was converted to bestguess data, without using a probability cut-off. QC steps included sex check, removing samples with over 5% missingness, removing samples which were identified as ethnicity outliers, and in the non-family cohorts removing individuals related over 12.5%.

2.2.3.2 DNA methylation data: GoDMC

The Illumina 450k or EPIC beadchip arrays were used to measure DNAm in either cord blood or whole blood. Methylation beta values were used, and were preferentially normalised with `meffil` (Min et al., 2018), using the functional normalisation protocol that can be found on the meffil github wiki (https://github.com/perishky/meffil/wiki). Outlying observations more than 10 standard deviations from the mean were removed over three iterations, and then replaced with the probe mean. Cell count estimates, age, sex, predicted smoking status, genetic principal components and methylation principal components were then regressed out of the methylation data. In the cohorts containing families, relatedness was regressed out of the methylation betas using the `GRAMMAR` approach (Aulchenko, Ripke, Isaacs, & van Duijn, 2007).

2.2.3.3 GoDMC analysis

Each study created a list of candidate mQTLs below the threshold of p<1e-5. These were combined for all studies, giving a unique candidate list of 102,965,711 potential *cis* (within 1Mb of the DNAm site) and 710,638,230 *trans* mQTLs. mQTLs in *trans* were reduced to those found in at least 2 cohorts, and combined with the cis candidates, and a total of 120,212,413 mQTLs were returned for association testing in all cohorts. The results of the second association test were meta-analysed using a modified version of `METAL` (Willer, Li, & Abecasis, 2010), with a fixed effects model, in 36 cohorts of European ancestry. The data used in this thesis was the association of DNAm sites with a SNP at the threshold p<10-8 for cis mQTLs and p<10-14 for trans mQTLs.

2.2.3.4 Adjusting for unknown factors

I considered the use of ISVA for the WGCNA network chapters. There tended to be over 100 surrogate variables that were detected in the data, and correlations with measures of blood cell type were no more than moderate (>0.5). It is not clear when running SVA or ISVA how many surrogate variables one should regress out from the data, or which ones to pick. This can result in quite arbitrary decisions. The authors of SVA explain that when removing surrogate variables, one risks removing biological signals in the data (Leek J, 2011). As my analyses were quite exploratory, particularly the analysis of the correlation structure on the 450k, I decided not to use SVA or ISVA and instead adjust for the known covariates in the data, as I did not want to risk regressing out biological signals.

2.3 Choice of correlation statistic

It is not clear which correlation method would be best to use for methylation data. Pearson correlation is mean based and susceptible to outliers, and as such may be unsuitable for DNAm data, particularly because DNAm levels may be influenced by genotype and form clusters. Spearman correlation is a rank based correlation, which transforms all values into ranks, so it will not be affected by outlying values. It also assesses . However because it uses ranks, it loses some information about the data and this can cause a loss of power, and it is also more computationally expensive. The biweight mid-correlation is another option; it is a median based method developed to be similar to Pearson, but which is robust to outliers (Hardin 2007;Langfelder 2012; Song 2012). Because it is similar to Pearson, it has more power than Spearman rank-based correlation and so may be advantageous.

However, I did not manage to find a published assessment of the comparison between Spearman and Bicor with regards to DNA methylation data, so I have made a comparison between the two methods. I correlated all DNAm sites on the Illumina 450k in ARIES 7 year olds using both Spearman and Bicor correlation. For comparison, I extracted all correlations >=0.9 using Spearman, and all correlations >=0.9 using Bicor. I also extracted all correlations that were >=0.9 using Spearman from the Bicor data, and vice-versa. I found that there is a fairly high correlation between the two methods, and that Spearman tends to have a lower value than Bicor. This is probably not surprising, as Spearman will have less power as a rank-based method. This is shown in Figure 6.

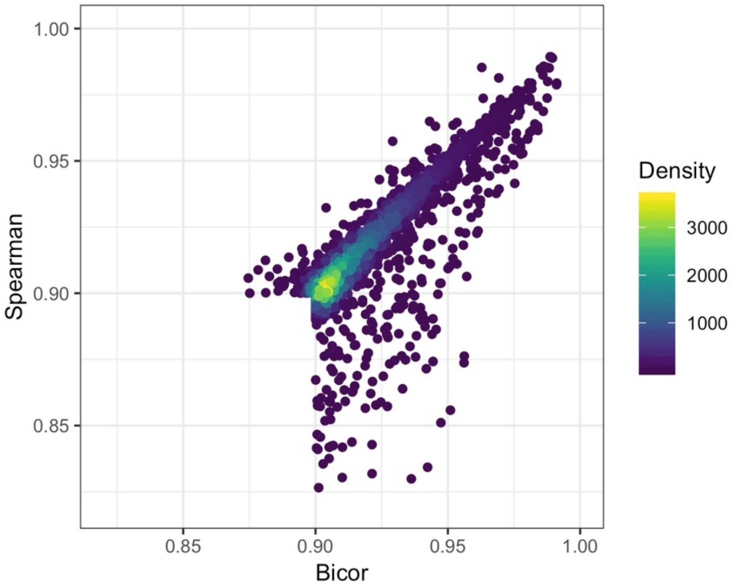


Figure 6: Correlation between DNAm site correlations using Spearman and Bicor correlations >0.9. correlation between the methods =0.88

To check that this was not just the case for the high correlations, I extracted the correlations between 9904 randomly selected probes from both the Spearman and the Bicor datasets. When I plotted the correlation values for these two methods against each other, they are highly correlated (r=0.998); see Figure 7. As the methods seem equivalent, I selected Bicor as the method to use for the 450k correlation. It has more power than Spearman, is computationally much more efficient, and it is the method we are using for WGCNA so it will make the two strands of analysis more comparable.

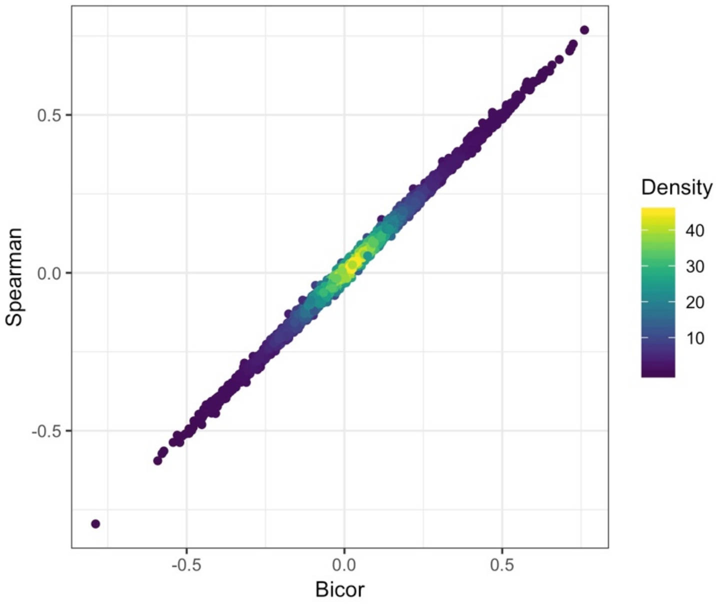


Figure 7: Scatter plot comparing the correlations between 9904 pairs of randomly selected DNAm sites, using Spearman and Bicor correlations, in ARIES 7 year olds. The correlation between the methods is r=0.99.

2.4 Reproducibility

For reproducibility, all scripts I have developed to run these analyses are publicly available on GitHub, under shwatkins/PhD. At the end of each section, there is a link to the relevant script. For most of the scripts I have run, I have uploaded the ARIES 7 year olds analysis scripts (which are applicable to all the other datasets). Where I have used or adapted a script from another source that will become publicly available (such as GoDMC), I have referenced the original script. Where I have used a script provided to me by another researcher, I have included a reproducible description of the script. I state where I was not the sole developer of analysis scripts.

2.5 Methods for correlation structure analysis (chapters 3 and 5)

2.5.1 Creating the correlation matrix and extracting pairwise correlations for analysis

Please see Figure 8 for an overview of this process.

In ARIES, a correlation matrix of 394,842 x 394,842 gives 77,949,905,061 unique correlations (when we remove the diagonal of the matrix). The simplest method of analysing this matrix would be to create a single matrix containing all pairwise correlations; however because of data frame size limitations when using R this was not possible.

As a solution to running this analysis in R, the DNAm sites were split into blocks of 25,000, and all blocks were correlated against each other. This was done using R code written by Matt Suderman (MRC IEU lecturer; see acknowledgements), as no current R package can perform this function. This resulted in the correlation matrix being distributed over 136 files. Where blocks were correlated against themselves, the portion below the diagonal was set to NA to avoid counting any correlation twice. The code to perform this is found under GitHub link 1 at the end of this section.

To assess the features of correlating pairs, the correlations were then split by correlation value, from -1 to 1, in increments of 0.1. This allowed me to conduct analyses of the features of correlations of different strengths (ie, do high correlations differ from low correlations?).

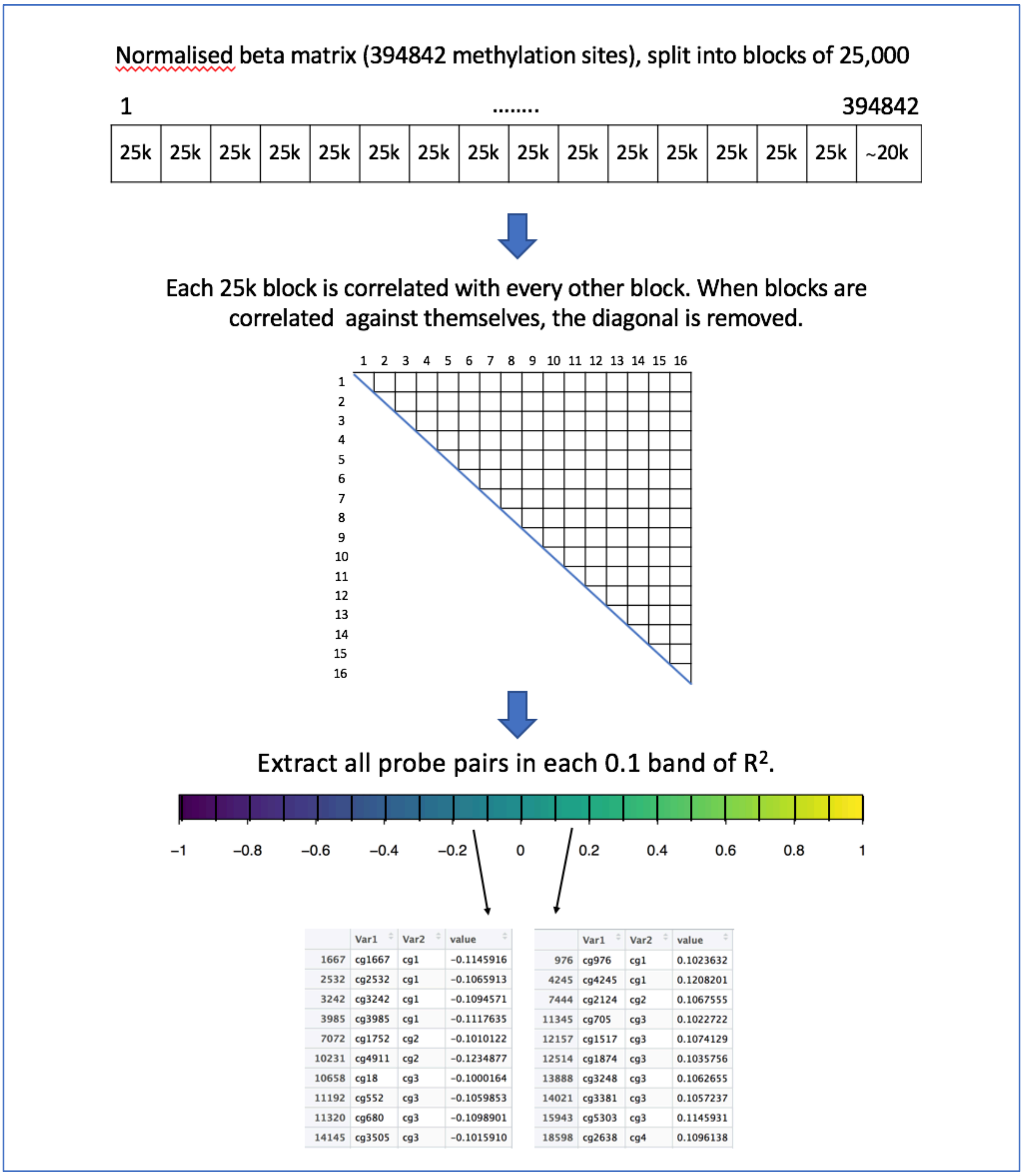
To do this, I wrote a function that looped through each of the correlation matrix files (136 in ARIES) in turn. From each correlation matrix, correlations were subset to one of the 20 ranges (eg ≥-1 & <-0.9). Using the function `melt` from the package `reshape2` (Wickham, 2007) the subset matrix was transformed into a dataframe, with two columns containing the DNAm site IDs, and a third column containing the correlation value between those two sites. The results from each of the 136 correlation files were added to a single dataframe, that at the end of the function contained all correlations in one of the correlation ranges (e.g. ≥-1 & <-0.9). This was saved for further analysis. This function was repeated for all 20 ranges of correlation value. The code to do this is found under GitHub file path 2 below.

The distribution of correlations (see Figure 10 in Chapter 3) show that around 85% of correlations fell between -0.2 and 0.2. Because there were so many correlations, dataframes summarising pairwise correlations in ranges between -0.3 and 0.3 were too large for R to handle as one file. To work around this, the 136 correlation matrices were split into smaller groups, and the function was run for each of these groups separately. This resulted in there being between 3 and 16 files for each correlation range between -0.3 and 0.3.

GitHub file paths

1: shwatkins/PhD/450k_correlation_analysis/01_F7-450k-singlefiles-bicor.R

2: shwatkins/PhD/450k_correlation_analysis/02_F7_full450k_extract.R



Normalised beta matrix (394842 methylation sites), split into blocks of 25,000

Each 25k block is correlated with every other block. When blocks are correlated against themselves, the diagonal is removed.

Extract all probe pairs in each 0.1 band of R^2.

	Var1	Var2	value
1667	cg1667	cg1	−0.1145916
2532	cg2532	cg1	−0.1065913
3242	cg3242	cg1	−0.1094571
3985	cg3985	cg1	−0.1117635
7072	cg1752	cg2	−0.1010122
10231	cg4911	cg2	−0.1234877
10658	cg18	cg3	−0.1000164
11192	cg552	cg3	−0.1059853
11320	cg680	cg3	−0.1098901
14145	cg3505	cg3	−0.1015910

	Var1	Var2	value
976	cg976	cg1	0.1023632
4245	cg4245	cg1	0.1208201
7444	cg2124	cg2	0.1067555
11345	cg705	cg3	0.1022722
12157	cg1517	cg3	0.1074129
12514	cg1874	cg3	0.1035756
13888	cg3248	cg3	0.1062655
14021	cg3381	cg3	0.1057237
15943	cg5303	cg3	0.1145931
18598	cg2638	cg4	0.1096138

Figure 8: Overview of creating the correlation matrix and extracting pairwise correlations (numbers are representative of the correlation matrix in ARIES).

2.5.2 Plotting the full distribution of correlation values

To plot the distribution of correlation values between all DNAm sites on the Illumina 450k, I wrote a script which looped through the file(s) containing all correlations in a correlation range (for example, all correlations >0.1 and <0.2). The function recorded the dimensions of each of the files. For correlation bands between -0.3 and 0.3 with more than one file, I then summed the dimensions of all files for that range of correlation. These dimensions were then plotted as a bar plot using `ggplot2` (Wickham, 2016). The script to do this can be found in GitHub link 1 below.

2.5.2.1 Proportions of cis and trans correlations

To plot the distribution of *cis* vs *trans* sites, I used a very similar script to output the *cis* and *trans* correlation dimensions for every correlation file. *Cis* was defined as both DNAm sites being within 1Mb on the same chromosome. The script to do this can be found in GitHub link 2 below. Percentages were plotted rather than absolute numbers, because the difference in the overall number of cis and trans correlations means that percentages make a much more interpretable plot.

GitHub links

1: shwatkins/PhD/450k_correlation_analysis/03_F7_correlation_structure_plot.R

2: shwatkins/PhD/450k_correlation_analysis/04_cis_trans_proportions.R

2.5.3 Illustration of cis correlation structure across the genome

To assess cis correlation structure, I produced decay plots of cis correlations on each chromosome. I based this on the method used in (Saffari et al., 2018).

I firstly wrote a function which identified and extracted all *cis* correlations (<1Mb apart) from each of the correlation band files. Genomic location for each probe was attained using the `meffil` (Min et al., 2018) `meffil.featureset` function. Distance between probe pairs was calculated as the absolute value of subtracting the location of probe 2 from probe 1, and only probes on the same chromosome, less than 1Mb apart, were retained. The resulting files contained only cis correlations. I then looped through each cis correlation file, and reorganised the cis correlations into separate files for each chromosome.

I then wrote another function which firstly subset each chromosome to correlations within 10kb. Positive and negative correlations were then separated so they could be plotted separately, as we hypothesised that negative correlations may not have the same structure as positive *cis* correlations. The pairwise correlations were then binned, using `cut2` from R package `Hmisc` (Harrell & others., 2019), where I specified at least 100 pairwise correlations per bin. Once the correlations were grouped into bins, I calculated the mean pairwise correlation for each bin, the standard deviation of the correlation values in each bin, and the median pairwise distance between the correlating pairs of DNAm sites per bin. These values were used to create the cis decay plots.

Line plots were created using `ggplot2`, with the x-axis representing the median distance of the bin and the y-axis representing the mean correlation of the bin. The bin standard deviations were plotted as vertical errorbars either side of the line. Positive and negative correlations were included on the same plot, although they were calculated separately.

The information about the highest correlation values that we attain from the decay plots can be somewhat limited, because averages are plotted, so I also created a histogram of all correlation values that were within 1kb of each other, using `ggplot2`. I did so for each chromosome separately.

GitHub link: shwatkins/PhD/450k_correlation_analysis/05_cis_decay_plots.R

2.5.4 Genetic influences on correlations between DNAm sites

I used three complementary approaches to investigate the influences of genetics on correlations:

2.5.4.1 Influences of heritability estimates on DNAm sites

Heritability has been shown to influence DNAm at a fairly low proportion of sites, but it has a high influence on sites that are generally of interest in epigenetic epidemiology – those which are most variable, and those which are associated with traits, such as obesity and smoking (Hannon et al., 2018).

To assess the impact of heritability on DNAm correlations, I took the estimates of heritability and environmental influences on DNAm created by (Hannon et al., 2018) (kindly provided to me as a full dataset), and used it to estimate the proportion of sites in each correlation band that were influenced by genetic, unique environmental, and shared environmental factors.

To do this, I wrote a function which extracted a unique list of all DNAm sites which featured in each correlation range (-1 to 1, in increments of 0.1). For each of these 20 lists, I then created a dataframe of the contributions of estimated heritability, unique environment, and common environment, to the variation in each DNAm site.

For each timepoint, I plotted each influence separately as a ridgeline plot using `ggplot2`, including the command `geom_density_ridges_gradient` from the `ggridges` package (Wilke, 2018) to create multiple density ridges on a single plot. This assessed whether the contribution of the three influences to variation in DNAm sites changed over time, and between strengths of correlation. The script to run this analysis can be found under the GitHub link below.

GitHub link:
shwatkins/PhD/450k_correlation_analysis/06_heritability_ridge_plots.R

2.5.4.2 Influence of mQTLs on correlations between DNAm sites

Another way to assess the influence of genetics on DNAm is to identify whether the level of DNAm is associated with a genetic variant (mQTL). The most comprehensive analysis identifying mQTLs is the GoDMC consortium's analysis (described above in section 0). I used the GoDMC data in two analyses:

2.5.4.2.1 Plotting the proportion of correlating DNAm sites associated with mQTLs

To initially illustrate how mQTLs might drive correlations between DNAm sites, I identified whether, for each correlating pair, neither, one or both of the DNAm sites were associated with an mQTL. I did this for the DNAm sites in each correlation range, to illustrate the distribution of mQTLs across values of correlation. Please note that this does not identify whether both DNAm sites are associated with the same mQTL.

To do this I took each file containing a range of correlating pairs (eg, all pairs correlating $r > 0.9$ & $r< 1$). For each DNAm site in a pair, I allocated a 0 if the DNAm site was not associated with any SNPs in the GoDMC dataset, and a 1 if there were one or more SNP associations. I then simply summed the two columns, which resulted in 0 if neither DNAm site was associated with a SNP, 1 if only one of the DNAm sites was associated with a SNP, and 2 if both DNAm sites were associated with a SNP. This was done for both cis and trans correlations. This returned a list, with the number of cis and trans correlations featuring 0, 1 or 2 DNAm sites associated with an mQTL, for every correlation file. The script to do this is on github under file path 1 below.

To plot the results, I first merged the list outputs where there was more than one file per correlation range; this was necessary for all correlations between -0.3 and 0.3 (see section 2.5.1 above for more details). I then calculated the percentage of correlations in each correlation range that featured 0, 1 or 2 DNAm sites associated with an mQTL, and used `ggplot2` to create a bar plot of these percentages, for cis and trans correlations separately. The code to do so is on github under file path 2.

GitHub file paths

1: shwatkins/PhD/450k_correlation_analysis /07_godmc_mqtls_percorrelation.R

2: shwatkins/PhD/450k_correlation_analysis /08_extractingmQTLsforgraph.R

2.5.4.2.2 Removing genetic influence from cis correlation plots

To then illustrate some of the impact these mQTLs actually have on correlations between DNAm sites, I adjusted a cis correlation decay plot for the strongest *cis* mQTL associated with each DNAm site, thereby removing the strongest single genetic influence on DNAm correlations.

For this analysis, I used chromosome 20 as an example. The analysis was performed using parts of the GoDMC analysis scripts (referenced in the box at the end of this section). I firstly created an additive genetic file (which indicates whether each individual has 0, 1 or 2 copies of the minor allele) from the ARIES genotype data, using the `plink` (Purcell et al., 2007) `--recode A` option. I then estimated the allele frequencies from the ARIES data using the `plink --freq gz` option.

I then adapted the GoDMC script (github file path 1, below) to identify the strongest cis mQTL for each DNAm site on chromosome 20. This merges the additive genetic file and the SNP frequencies. Where there was a mismatch between the effect alleles in these files (which occurred where effect allele frequencies were very close to 0.5), those SNPs were removed from the analysis. The cis-SNP most associated with each DNAm site (with the lowest p-value) was then identified. The DNAm value for each individual was adjusted for their genotype for the relevant SNP using the additive data, using the code `lm(DNAm~snp)`. The residuals were then taken forward to plot. The DNAm values which did not have an associated SNP were not adjusted; these were still included in the cis decay plot and were merged with the residuals dataframe.

I then ran the cis decay plots in a very similar way as in section 2.5.3 above, with this new adjusted set of DNAm values. The decay plot included a comparison with the original plotted line, and so the standard deviation errorbars were left off the plot for a clearer comparison. I also plotted the mean change in correlation per bin, between the original and adjusted bin correlation values. The script for this plot can be found on github under file path 2, below.

GitHub file paths

1: MRCIEU/godmc/resources/phase2/run_analysis.R

2: shwatkins/PhD/450k_correlation_analysis /09_regressoutcis_chr20.R

2.5.4.3 Influence of LD on correlation structure

To try to assess the influence of LD on correlation structure on a larger scale than regional plots would allow, assessed the relationship between linkage disequilibrium (LD) and DNAm connectivity. To do this, I used chromosome 20 as an example, as chromosome 20 is relatively small and so is computationally inexpensive.

I calculated DNAm connectivity using the formula for the `kTotal` measure used in the R package `WGCNA` (Langfelder & Horvath, 2008). This is simply, for each DNAm site, a sum of its correlations with all other DNAm sites. I calculated LD score using the `ldsc` package (Bulik-Sullivan et al., 2015). LD score was calculated for distances of 10,000kb (as LD becomes vanishingly small over large distances).

To compare LD score and DNAm kTotal score, I identified DNAm sites that were associated with a cis SNP in the GoDMC consortium data, and paired the DNAm site kTotal scare with the SNP LD score. Pairing DNAm sites with their strongest cis SNP, rather than the SNP at the closest genomic position, makes more sense as if DNAm correlation is related to LD, the strongest cis SNP will have the greatest relation to the DNAm correlation structure at that locus. Once the LD score and kTotal score were paired up for each of the DNAm sites, I plotted them as a scatterplot using `ggplot2`, adding a smoothed regression line to indicate whether LD score is related to DNAm connectivity. The code to run this analysis can be found in the link below.

GitHub file path:
shwatkins/PhD/450k_correlation_analysis/10_regressoutcis_chr20.R

2.5.5 Analysis of strong correlations

2.5.5.1 Genomic region enrichment

To identify whether DNAm sites which form strong correlations overlap with genomic sites of interest, I used the locus overlap package `LOLA` (Sheffield & Bock, 2016). `LOLA` assesses enrichment based on genomic regions rather than genes, which is advantageous for DNAm analyses.

I based the overlap assessment on a script provided by Josine Min. I loaded the file containing all correlations r>0.9, and from this created a list of unique DNAm sites in this correlation file. These sites formed the test dataset. I used the list of all sites in the analysis as the background. I used region sets created by the LOLA team, available through http://lolaweb.databio.org. I used the ENCODE transcription factor binding sites, chromHMM imputed 25 chromatin states from Roadmap Epigenomics (Ernst & Kellis, 2015; Roadmap Epigenomics et al., 2015), and Cistrome histone marks (Q. Wang et al., 2014).

I obtained the genomic locations of the DNAm sites using annotations in the `IlluminaHumanMethylation450kanno.ilmn12.hg19` R package (Hansen, 2016). Start and end sites were computed as -500bp and +500bp from the DNAm site position, respectively. A radius of 1kb was thought to be appropriate overlap because DNAm sites within 1-2kb are highly correlated. Overlapping sites were collapsed to avoid inflating the results. Because binding of transcription factors is enriched in GC-rich areas of the genome, the content of the background set was reduced and matched to the GC content of the test set using frequency quantiles. The test set and background were converted to `GRanges` objects (Lawrence et al., 2013). I then ran the LOLA overlap enrichment analysis using the `LOLA` command `runLOLA`.

`LOLA` results were plotted on a scatterplot using `ggplot2`, with genomic regions on the x-axis and the log odds ratio of the overlap on the y-axis. Points were scaled by the log10 P-value of the overlap. Points were coloured by tissue or cell type, and grouped into genomic regions (eg transcription factors). The R code I used to assess the overlap and produce the plots is on github under the below file path. The example is for transcription factor binding sites, but any of the LOLA databases can be substituted in this code.

GitHub file path: shwatkins/PhD/450k_correlation_analysis /11_F7_LOLA_tfbs.R

2.5.6 Trans correlation structure

2.5.6.1 Visualising trans correlation structure

2.5.6.1.1 Circos plots

Circos plots were generated to visualise the distribution of strong trans correlations across the genome. I used the R package `circlize` (Gu, Gu, Eils, Schlesner, & Brors, 2014) to generate the plots. As input data I used all trans correlating DNAm sites with r>0.9. The code to create the plots can be found in the GitHub link below.

GitHub file path: shwatkins/PhD/450k_correlation_analysis/12_trans_circosplot.R

2.5.6.1.2 Cytoscape

Cytoscape (version 3.6.1) plots were generated to visualise the connectedness of strong trans correlations (Shannon et al., 2003). To generate the plots, I generated a text file containing the correlations between DNAm sites r>0.9. These were imported to cytoscape. I ran a network analysis using *Tools>NetworkAnalyzer>Network Analysis>Analyze Network*, and used the undirected edges option. I coloured the nodes by the number of connections they have, which was indicated by the *NumberOfDirectedEdges* column.

2.6 Methods for WGCNA analysis

All WGCNA analyses were run using the WGCNA R package (Langfelder & Horvath, 2008). All R scripts used to run the WGCNA analysis were adapted from the tutorials available on the UCLA WGCNA website (Langfelder & Horvath, 2016). The description of methods in this section were taken from papers from the group that developed the method (Horvath et al., 2012; Langfelder & Horvath, 2007, 2008; Langfelder, Luo, Oldham, & Horvath, 2011; B. Zhang & Horvath, 2005) and from the R package documentation.

2.6.1 Removing outlying samples

The first step in network construction was clustering samples using the whole DNAm dataset. Doing so helps to identify outlying samples that may skew the clustering of DNAm sites, and therefore affect the modules it can detect. Anecdotally, this made a noticeable difference to whether the network would reach scale free topology (described in the next section). To create a hierarchical cluster dendrogram, DNAm data was transformed to a distance matrix which was clustered using `hclust` (which is implemented from the package `fastcluster` (Mullner, 2013). The code I used to do so was as follows:

```
hclust(dist(data), method = "average")
```

Where data is the DNAm dataframe, dist transforms data into a distance matrix, and the option "average" is the agglomeration method to use (which is the option used in the WGCNA tutorials).

Outlying samples were identified from the resulting dendrogram, and removed from the DNAm matrix, which was then re-clustered to check for further outliers. An example of the samples cut from the tree can be seen in Figure 9 – all samples above the red line were considered to be outliers and removed from the dataset. The code can be found in the GitHub file path below.

GitHub file path: shwatkins/PhD/WGCNA_analysis/WGCNA_1.R

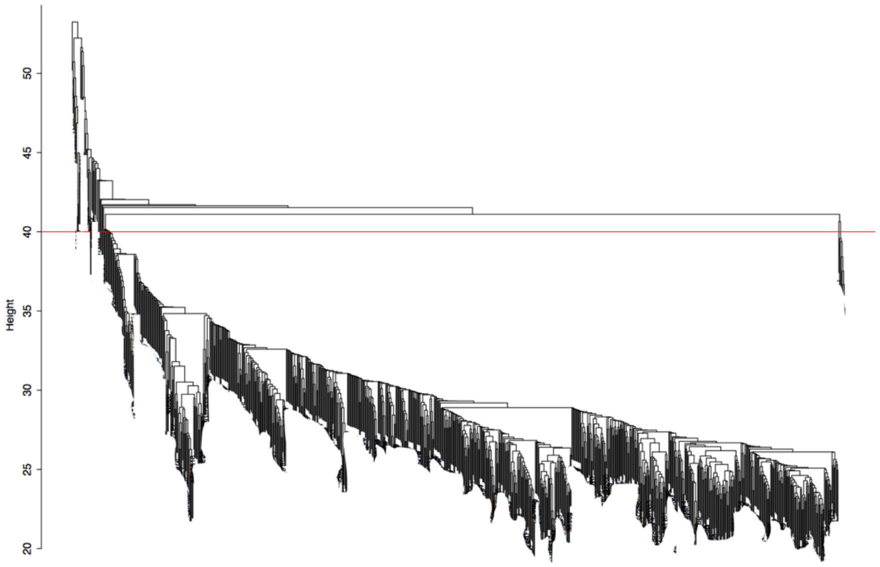


Figure 9: Hierarchical clustering of samples at 7 years in ARIES. Samples above the red line were removed from the analysis.

2.6.2 Calculating soft threshold power

Once outlying samples were removed, the power at which the data reached scale free topology was calculated. This power formed a soft threshold, to emphasise high correlations, rather than removing lower correlations from the network using a hard threshold. This is done by transforming the network to a scale free topology as scale free networks are a network theory that assumes non-random connections and highly connected hub nodes, and is more representative of biological networks than other network theories which assume random connections (Barabasi, 2009; Barabasi & Albert, 1999). If there is a major driver in DNAm data that causes some samples to have very different DNAm profiles to others, the network may not exhibit scale free topology, and so this analysis also forms a check that there are not unexpected systematic differences between some of the samples in the analysis (Langfelder & Horvath, 2017).

The power to raise the network to in order to reach scale free topology is using the `pickSoftThreshold` function in the WGCNA R package. This function creates a model to predict at which power the network reaches scale free topology. I calculated this for powers 1-20, using a signed network. The code I used was as follows (and can also be found under the GitHub link in section 2.6.1):

```
powers = c(c(1:20))
pickSoftThreshold(data, powerVector = powers, networkType =
"signed", verbose = 5)
```

2.6.3 Blockwise network construction

WGCNA networks can be constructed in two ways - either in one go, as a single network, or in multiple blocks (from which modules are then merged together if they are similar enough).

I used the blockwise method, because it is not computationally possible in R to create a correlation matrix large enough for close to 400,000 nodes. Although the blockwise method sacrifices some accuracy, the modules it creates are demonstrably approximate to the single block method (an analysis that can be run using tutorial 1 on the WGCNA tutorial website (Langfelder & Horvath, 2016)). The GitHub link to my code is at the end of this section.

The blockwise network is created using the `blockwiseModules` command. This command has a wealth of complex parameters that can be adjusted, depending on the needs of the user. The parameters I selected are displayed and discussed in detail below:

```
blockwiseModules(df, maxBlockSize = 45000,
                 corType = "bicor", power = 5,
                 networkType = "signed", TOMType = "signed",
                 minModuleSize = 30, deepSplit = 2,
                 maxPOutliers = 0.05,
                 reassignThreshold = 0, mergeCutHeight = 0.25,
                 numericLabels = TRUE, saveTOMs = TRUE,
                 saveTOMFileBase = "path/to/file/base",
                 verbose = 3)
```

This function initially performs a rough k-means cluster of the DNAm data (`df`, a dataframe with DNAm sites as columns and participants as rows), to split it into blocks of a pre-specified number of DNAm sites which are similar to each other. I specified blocks of 45,000 DNAm sites because there was a cap at around 46,000 nodes. In each block, all sites are correlated against each other using a specified correlation statistic; I selected the biweight mid-correlation, known as `bicor` (see section 2.3 for details on this choice). When using bicor it is important to set `maxPOutliers` to 0.05 or 0.1, because this limits the maximum percentile of observations which can be considered outliers. This is particularly important when data has a bi-modal distribution, which is relevant for DNAm data (as advised by the package developers).

Two network types can be created; signed or unsigned. Unsigned networks work with the absolute value of correlation, and so lose the sign of negative correlations. This can raise issues when calculating shared connections (topological overlap; discussed below) and so it is generally recommended to use signed networks, which essentially set the value of negative correlations to 0 (Langfelder, 2013). Because of these considerations, and because I have demonstrated that negative correlations are different to positive correlations (Chapter 3), I constructed signed networks.

This correlation matrix is then raised to the soft threshold power (the calculation of which is discussed in section 2.6.2) in order to emphasise strong correlations, and de-emphasise weak correlations. In the above example, `power` is 5, although this varied between datasets and is reported in the relevant chapter. Once the correlations have been raised to this power, a topological overlap matrix (TOM) is calculated. The TOM represents relationships between nodes as a result of their shared connections, rather than the direct correlation between them. I selected a signed `TOMType`, although in practice when working with a signed network it does not make much difference whether the TOM is signed or unsigned (Langfelder, 2013).

With regard to other parameters, a minimum module size of 30 has been used frequently for DNAm data, as modules smaller than this might be too likely to represent noise. I selected a relaxed `reassignThreshold` so that DNAm sites would be assigned to the module they are most closely correlated with, and a higher `mergeCutHeight` than the default because the heights of the dendrograms for DNAm data are much closer to 1 than for gene expression. When experimenting with the parameters, I found the impact of adjusting them tends to be quite minimal on the modules that are created.

To define network modules, the TOM is clustered using average linkage hierarchical clustering, and DNAm sites are exclusively assigned to a module. A representation of each module is then created – this is known as a module eigengene, and is essentially equivalent to the first principal component of the module members. Because biological features may well not be restricted to only one pathway, correlations between every DNAm site and every module eigengene are calculated, and DNAm sites are reassigned to the module which they have the strongest relationship with. This measure of DNAm site – module relationships is termed kME, and as it is a correlation the measure ranges from 1 to -1. Finally, modules with highly correlated module eigengenes are merged at `mergeCutHeight`.

GitHub file path: shwatkins/PhD/WGCNA_analysis/WGCNA_2.R

2.6.4 Associating WGCNA modules with phenotypes and cell counts

One of the advantages of using WGCNA is that the modules can be represented by eigengenes, and so it is easy to associate traits of interest with the module eigengenes, to see if they might be associated. I ran this analysis separately for phenotypes (as discussed in section 2.2.1.9), and for estimated cell counts (as discussed in section 2.2.1.6). Code for this analysis was based on the WGCNA single network tutorials available on the WGCNA website (Langfelder & Horvath, 2016). I altered the code to use a linear regression model to test each trait, rather than just a correlation, as a regression model is a more powerful tool which allows for adjustment of covariates. The code to create the regression model is below. The "MEs" matrix is a matrix of module eigengene scores, where each participant in the dataset has a score for each eigengene. These scores are created using the WGCNA function `moduleEigengenes`, which calculates the first principal component of the methylation matrix for each module. The trait is regressed on each eigengene, so that for each eigengene a regression model is fitted separately. For cell counts, each cell type was modelled separately. This assesses how well the trait predicts each module eigengene score. Covariates were added to the model as specified in section 2.2.1.9.

```
lm(as.matrix(ModuleEigengenes) ~ trait + covariates)
```

The outputs of the regression models were combined into a single heatmap for each timepoint separately, to easily visualise the relationship between all network modules and all traits. To assess whether modules associated with traits were robustly associated, p-values were adjusted for multiple testing using the Bonferroni correction. As there are established issues with using p-values do declare associations 'significant' (Sterne & Davey Smith, 2001) a threshold was not used, but the corrected p-values were used to contribute to the strength of evidence. The code to run this analysis can be found in the link below.

GitHub file path: shwatkins/PhD/WGCNA_analysis/WGCNA_3.R

2.6.4.1 Preservation of network modules

The preservation of modules between single networks was assessed using a number of module preservation statistics, developed for the WGCNA package (Langfelder et al., 2011), and implemented using the `modulePreservation` function in the WGCNA R package. I adapted the code from the WGCNA tutorials website (Langfelder & Horvath, 2016). Multiple module preservation statistics are calculated, assessing preservation of the density, connectivity, and separability of the network created in a reference dataset (e.g., at birth) against that created in a test dataset (e.g., at age 7).

Although many preservation statistics are calculated, the authors recommended to utilise the composite measures, Zsummary and median rank, which they show to be effective at identifying preserved modules (Langfelder et al., 2011). The preservation Zsummary statistic is composed of 4 density measures and 3 connectivity measures, and a Zsummary > 10 indicates strong module preservation; however (Langfelder et al., 2011) show this statistic is influenced by module size. The preservation median rank measure ranks the observed connectivity and density module statistics, and gives a summary of these, enabling comparative module preservation. The authors show that this method is not biased by module size as Zsummary is; therefore use of both measures can give a better picture of module preservation. The code I used is in the link below.

GitHub file path: shwatkins/PhD/WGCNA_analysis/WGCNA_4.R

2.7 Gene ontology and KEGG pathway analysis

To assess whether DNAm sites of interest were associated with genes which were enriched for specific gene ontologies or molecular pathways, I used the R package `missMethyl` (Phipson, Maksimovic, & Oshlack, 2016). Enrichments are assessed in `missMethyl` by using the `goana` and `kegga` commands from the `limma` R package (Ritchie et al., 2015). Genes are assigned to DNAm sites using the annotations from the `IlluminaHumanMethylation450kanno.ilmn12.hg19` package (Hansen, 2016), with gene ontology annotations assessed using the `org.Hs.eg.db` R package (Carlson, 2019), and the KEGG (Kyoto Encyclopedia of Genes and Genomes) pathways from the KEGG website (Kanehisa & Goto, 2000; Kanehisa, Sato, Kawashima, Furumichi, & Tanabe, 2016). To select the most appropriate DNAm sites for each WGCNA module, I took all probes with high similarity to the module eigengene (defined as kME > 0.7; kME is defined in section 2.6.3), and ran the gene ontology and KEGG analysis on these sets of probes. A kME of 0.7 is a commonly used threshold for strong module membership. The top 20 enriched terms were returned for all modules; associations close to or under FDR threshold of 0.05 were considered, as it is not recommended to use p-values a s a hard threshold (Sterne & Davey Smith, 2001). Code for the analysis can be found in the GitHub link at the end of the section.

I used `missMethyl` because it uses a user defined background of DNAm sites (which in this case was the 394,842 DNAm sites used in the network analysis in ARIES; and the 369,796 sites in BiB). This is important because the DNAm sites on the 450k array are not randomly distributed across the genome – they are very biased to locations in promotors, transcription start sites and protein coding transcripts (Sandoval et al., 2011), so enrichments will be likely to be inflated without the appropriate background. Another feature of `missMethyl` is that it corrects for the number of DNAm sites on the 450k array that tag each gene. This is important, because otherwise genes which have more probes associated with them may appear enriched when in fact they are simply more represented on the 450K array.

GitHub file path: shwatkins/PhD/WGCNA_analysis/GOplusKEGG.R

3 Chapter 3

Drivers of correlation between DNAm sites

3.1 Summary

3.1.1 The importance of relationships between DNAm sites

Relationships between DNAm sites are important, because DNAm is one part of an interconnected biological system. As such, it is unlikely that DNAm sites act in isolation. If we can identify key relationships between DNAm sites, we may be able to uncover systems or pathways which regulate, or are regulated by, DNAm. In turn this could lead to better understanding of traits or diseases which may lead to new therapeutic targets being identified. This is the central tenant of systems biology, which is discussed in chapter 1 section 1.4.

3.1.2 Current knowledge about relationships between DNAm sites

As detailed in chapter 1, section 1.5, a number of studies have addressed the question of correlation between DNAm sites. It is clear that there are cis relationships between DNAm sites (Eckhardt et al., 2006; Y. Liu et al., 2014; Ong & Holbrook, 2014; Saffari et al., 2018), but there is not a complete picture of this across the whole 450k array, and it is not clear whether it changes over time. Although correlation between DNAm sites has been shown to not relate directly to LD (Y. Liu et al., 2014; Saffari et al., 2018), it is not clear whether smaller blocks of *cis* correlation structure might still relate in some way to LD. Published examples of genetic or environmental influence often focus on only the most variable sites; they illustrate their arguments with a limited selection of cases; and there has been no separation of positive and negative correlations, and the differences there might be between them. There has been rather limited characterisation of trans relationships between DNAm sites, but studies that have point to their being functional (Garg et al., 2018; J. Liu et al., 2019).

3.1.3 What the current project will address

In this chapter, I am going to answer the gaps in the literature as described in the above section. These questions are important to answer, as although network analyses (which exploit the relationships between DNAm sites to inform us about biological pathways) are valuable, if we do not know what the drivers of these relationships are, there is some uncertainty about what DNAm networks might be illustrating. this chapter sets out to address what I see as five main questions about relationships between DNAm sites, to better inform the interpretation of DNAm networks constructed in later chapters.

3.1.4 Hypotheses

H.3.1. Based on work on epigenetic drift (Shah et al., 2014; Teschendorff et al., 2013), I hypothesise that strong correlations between DNAm sites will decrease for a proportion of sites between birth and adolescence.

H.3.2. I hypothesise that if DNAm has regulatory functions, correlations between DNAm sites in *cis* will form in functional locations, such as gene promotors and transcription start sites, as these are locations where DNAm could exert its strongest regulatory functions (eg permitting transcription of a gene).

H.3.3. Based on previous work, I hypothesise that highly correlating sites will have greater enrichment for genetic influences such as heritability and mQTLs. I expect that correlations will not related to LD structure.

H.3.4. I hypothesise that there will be an enrichment of chromatin contacts between strongly *trans*-correlating DNAm sites, and that these sites are likely to be located in enhancers and promotors, based on work on long-range chromatin contacts.

H.3.5. If trans correlations between DNAm sites are functional, I hypothesise they will form discrete networks with each other; in which case one ought to be able to map a group of relationships and map this to either a biological pathway, or a master regulator gene.

3.1.5 Aims

I aim to answer these hypotheses in the following ways:

A.3.1. Provide an illustration of the distribution of all correlation values across the whole 450k array. This is a very basic description of DNAm correlation structure, but it has not been published anywhere to date. I will use ARIES data to illustrate whether this broad spectrum of correlations changes over time; one might expect there to be fewer high correlations as we age based on work on epigenetic drift.

A.3.2. Create a comprehensive description of cis correlation structure, as there are a number of gaps remaining in the literature. I will illustrate cis correlation structure for all chromosomes, provide measures of heterogeneity, separate out the positive and negative correlations, and illustrate whether this changes in the same individuals over time.

A.3.3. Conduct a comprehensive analysis of the extent of genetic control over relationships between DNAm sites. I will use a number of sources for this: I will assess the influence of heritability on correlations between DNAm sites; I will illustrate whether highly correlating sites are more likely to have their methylation level associated with a genetic variant (mQTL – methylation Quantitative Trait Loci); I will assess if genetic variants in high LD are related to highly correlating DNAm sites; and I will directly assess the impact of genetic variants on *cis* correlation structure.

A.3.4. Provide an analysis of trans correlations between DNAm sites across the genome (as measured by the 450k). I will assess whether these correlations are spurious, or whether they may, as hypothesised, form functional regulatory relationships.

A.3.5. Identify whether highly correlating sites have identifiable relevance to genetic function, using enrichment analyses for chromatin states, transcription factor binding sites, and histone marks.

A.3.6. Identify whether strong trans chromosomal correlations overlap chromatin contacts.

3.2 Methods

3.2.1 Data: ARIES

The DNAm datasets used in this chapter were the 3 ARIES young person DNAm datasets. ARIES is a subsample of ALSPAC, where 1,000 mothers and their children were selected to have DNA methylation profiled. This was at birth (cord blood), 7 years and 15-17 years (Relton et al., 2015). The ARIES cohort is described in detail in chapter 2 section 2.2.1.1. DNAm was measured in blood (either whole blood, white cells, PBS or blood spots), using the Illumina 450k array.

3.2.1.1 Adjustments

I filtered sites on the 450k array, as detailed in chapter 2, section 2.2.1.7. Briefly, all non-autosomal probes, probes in the HLA region, poorly functioning probes and probes containing SNPs were removed from the dataset. This left 394,842 probes for analysis. As all timepoints were normalised together, the same probes were included for all timepoints. All outlying values (> 10 standard deviations from the mean) were replaced with the mean for that probe, over three iterations.

As discussed in chapter 2 section 2.2.1.7.3, all datasets had the appropriate combination of age, sex, and sample type, as well as estimated cell counts, regressed out using a linear model. Slide was also regressed out, to account for batch effects, using a linear model. A random effects model was not used because batch effects were removed as a random effect during normalisation; slide was regressed out after normalisation because the normalisation did not completely remove associations with slide. All individuals with methylation data were included, unless they were the only individual on a slide (as detailed in chapter 2, section 2.2.1.3).

3.2.2 Data : GoDMC

The mQTL data used in this section was generated by the GoDMC consortium. It is described in detail in chapter 2, section 2.2.3. Briefly, the consortium data consisted of 27,750 participants of European ancestry from 36 cohorts. Each cohort identified mQTLs below the threshold of p<1e-5 individually, and mQTLs from all cohorts were combined to make a unique list. This unique list was then tested for association in every cohort, and the results were meta-analysed. The data used in this thesis was the association of DNAm sites with a SNP at the threshold $p<10^{-8}$ for *cis* mQTLs and $p<10^{-14}$ for *trans* mQTLs.

3.2.3 Data: Heritability and environmental influences on DNAm

The relative contributions of heritability, common environment, and unique environment have been estimated for all DNAm sites on the 450k array by (Hannon et al., 2018). These estimates were constructed using twin data, and have been made available as a searchable dataset online (http://www.epigenomicslab.com/online-data-resources). The full dataset was kindly provided to me by the authors.

3.2.4 Compute resources

The following analyses were run on a high memory server, running Ubuntu 18.04, using R version 3.6.0. Relevant R packages are detailed, as appropriate, throughout the thesis.

3.2.5 Correlation statistic

I used the biweight mid-correlation from the R package `WGCNA` (Langfelder & Horvath, 2008) as a correlation statistic, as Pearson correlation is not suitable for bimodal distributions (which can be the case for some DNA methylation sites). I have demonstrated that the biweight mid-correlation is a suitable, more efficient alternative to spearman in chapter 2 (section 2.3).

3.2.6 Creating the correlation matrix and extracting pairwise correlations for analysis

ARIES had 394,842 probes retained for analysis. A correlation matrix of 394,842 x 394,842 gives 77,949,905,061 unique correlations (when we remove the diagonal of the matrix). The process of creating the correlation matrix is discussed in detail in chapter 2 section 2.5.1; briefly, DNAm sites were split into 16 blocks of 25,000 and all blocks were correlated against each other. Dataframes of correlating pairs were then created for each 0.1 band of correlation (20 bands between -1 and 1), so I could conduct analyses of the features of correlations of different strengths (i.e., do high correlations differ from low correlations?).

3.2.7 Plotting the full distribution of correlation values

To plot the distribution of correlation values between all DNAm sites on the Illumina 450k, I ran the analysis described in detail in chapter 2 section 2.5.2. I ran this for each of the ARIES timepoints.

3.2.7.1 Proportions of cis and trans correlations

To plot the distribution of *cis* and *trans* correlations I plotted percentages in each range of correlation, because there are so many more *trans* correlations than *cis* that it is hard to interpret plots with absolute numbers. To do this, I ran the analysis described in chapter 2, section 2.5.2.1, for each ARIES timepoint.

3.2.8 Illustration of cis correlation structure across the genome

To assess *cis* correlation structure, I produced decay plots of *cis* correlations on each chromosome separately. For each plot, I used correlations that were within 10kb of each other, based on previous literature. I also produced a histogram of correlations within 1kb, for a clearer illustration of the correlation values in close proximity. I did this for each of the three ARIES timepoint. The details of creating these plots can be found in chapter 2, section 2.5.3. To create the cis decay plot for all cis correlations genome-wide, the same method was used, where all cis correlations (within 10kb) and their distances were combined to a single dataframe. As this was larger by a factor of around 10, the minimum bin size was changed from 100 to 1,000. This was run for all ARIES timepoints.

3.2.9 Genetic influences on correlations between DNAm sites

I used the following three complementary approaches to investigate the influences of genetics on correlating sites:

3.2.9.1 Influences of heritability estimates on DNAm sites

To assess the impact of heritability on DNAm correlations, I took the estimates of heritability and environmental influences on DNAm created by (Hannon et al., 2018), and used it to estimate the proportion of sites in each correlation band that were influenced by genetic, unique environmental, and shared environmental factors. I used the methods described in chapter 2 section 2.5.4.1 to produce ridgeline plots to illustrate this.

3.2.9.2 Influence of mQTLs on correlations between DNAm sites

Another way to assess the influence of genetics on DNAm is to identify whether the level of DNAm is associated with a genetic variant (mQTL). The most comprehensive analysis identifying mQTLs is the GoDMC consortium's analysis (described above in section 3.2.2).

3.2.9.2.1 Plotting the proportion of correlating DNAm sites associated with mQTLs

To initially illustrate how mQTLs might drive correlations between DNAm sites, I identified whether, for each correlating pair, neither, one or both of the DNAm sites were associated with an mQTL. I did this for the DNAm sites in each correlation range, to illustrate the distribution of mQTLs across values of correlation. Please note that this does not identify whether both DNAm sites are associated with the same mQTL. The code I used to do this and generate plots illustrating the association of *cis*- and *trans*-correlating DNAm sites is found in chapter 2 section 2.5.4.2.1.

3.2.9.2.2 Removing genetic influence from cis correlation plots

To then illustrate some of the impact these mQTLs actually have on correlations between DNAm sites, I adjusted the *cis* correlation decay plots for the strongest *cis* mQTL associated with each DNAm site, thereby removing some of the genetic influence on DNAm correlations. For this analysis, I used chromosome 20 as an example, in the ARIES 7 year olds. The details of this analysis can be found in chapter 2, section 2.5.4.2.2.

3.2.9.3 Influence of LD on correlation structure

To try to assess the influence of LD on correlation structure on a larger scale than regional plots would allow, I assessed the relationship between linkage disequilibrium (LD) and DNAm connectivity. To do this, I used chromosome 20 as an example, as chromosome 20 is relatively small and so is computationally inexpensive. The details of this analysis can be found in chapter 2, section 2.5.4.3.

3.2.10 Analysis of strong correlations

3.2.10.1Genomic region enrichment

To identify whether DNAm sites which form strong correlations overlap with genomic sites of interest, I used the locus overlap R package `LOLA` (Sheffield & Bock, 2016). `LOLA` assesses enrichment based on genomic regions rather than genes, which is advantageous because DNAm sites are not necessarily functionally linked to their nearest gene. I tested *cis* and *trans* correlations r>0.9 separately for genomic region enrichment, using the list of all

394,842 sites in the analysis as the background. I used two region sets created by the LOLA team, available through http://lolaweb.databio.org: the ENCODE transcription factor binding sites (J. Wang et al., 2012), and Cistrome histone marks (Q. Wang et al., 2014). I also used a region set generated by Josine Min: the chromHMM imputed 25 chromatin states from the Roadmap Epigenomics consortium (Ernst & Kellis, 2015; Roadmap Epigenomics et al., 2015). The details for these analyses are in chapter 2, section 2.5.5.1.

3.2.11 Trans correlation structure

3.2.11.1 Detection p-values

To check whether trans correlations represent array noise, I assessed the detection p-values of the DNAm sites correlating r>0.9 at all ARIES timepoints. Where the detection p-values were 0, I added 1e-100 so I could plot -log10 transformed p-values.

3.2.11.2 Visualising trans correlation structure

3.2.11.2.1 Circos plots

Circos plots were generated to visualise the distribution of strong trans correlations across the genome. The details of producing these plots can be found in chapter 2, section 2.5.6.1.1.

3.2.11.2.2 Cytoscape

Cytoscape plots were generated to visualise the connectivity of strong trans correlations, using Cytoscape version 3.6.1 (Shannon et al., 2003). The detail of this analysis is in chapter 2, section 2.5.6.1.2.

3.2.11.3 Assessing trans correlations for chromatin contacts

(G. Li et al., 2019) have illustrated that inter-chromosomal regions which connect have correlated DNAm states. To test whether highly correlating regions are enriched for chromatin contacts, I adapted the GoDMC analysis pipeline to test for chromatin contact enrichment. This uses the chromatin contacts map produced by (Rao et al., 2014). The original scripts were developed by Kimberly Burrows (Research Associate, University of Bristol) and can be found on the GoDMC GitHub pages.

For each of the pairwise contacts in the (Rao et al., 2014) dataset the pipeline creates a 1kb region for the two contacting areas of the genome. These are split into files containing all interactions for all possible pairs of chromosomes (for example all contacts between chromosome 1 and chromosome 18). This resulted in 231 files containing all inter-chromosomal contact on the autosomes. Next, I took all the inter-chromosomal correlations r>0.9 in the ARIES 7 year olds and defined a 500bp region either side of these DNAm sites. The pipeline then identified which correlating pairs overlapped with the (Rao et al., 2014) contact regions.

To ascertain whether there were more contacts in my data than expected by chance the pipeline creates a permuted dataset, where from the original data the second DNAm site in the highly correlating pair is replaced randomly with another in the dataset. Broken pairs that match a pair from the original dataset are removed, as are duplicates to avoid double counting. Then overlaps with the (Rao et al., 2014) data were calculated for the permuted data. This process was repeated 1,000 times, the permuted datasets were merged together, and a distribution of overlap counts was created for the permuted data. This distribution was used to create a p value for the overlap of the permuted distribution chromatin contact overlaps with the number of overlaps in the real data. The permuted distribution of overlaps and the real number of overlaps were plotted using `ggplot2`.

Variance sensitivity analysis

For a correlation between two variables to be detected, there needs to be variation present. If clear variance is not demonstrated then high correlations could reflect technical noise; for this reason, previous studies sometimes remove DNAm sites with low variance. As a sensitivity analysis, I assessed the variance of highly correlating DNAm sites, for cis and trans correlations separately. I approached this in a number of ways.

Firstly, I created a density plot of the standard deviation for all DNAm sites that have a correlation >0.9. This was done for a unique list of these DNAm sites so the variance of each site would be counted only once, regardless of how many correlations >0.9 the site had. The code to produce this plot can be found in GitHub filepath below. I then plotted the mean of highly correlating DNAm sites against each other, with errorbars showing the standard deviation of each site. This was to illustrate the relationship between DNAm variance and the mean methylation of that site. I plotted cis and trans correlating sites in opposing colours, to show the contrast in the mean and variance between cis and trans correlating sites.

To illustrate the variance of highly correlating sites in detail, I selected examples of cis and trans correlations with high and low variance from the correlations >0.9. The normalised betas (unadjusted for covariates) were plotted for the correlating pair of DNAm sites. This was done with normalised (but unadjusted) betas to give a better illustration of the variance of the sites. Finally, I repeated these analyses for the sites that were identified as chromatin contact sites, to ascertain whether removing low variance sites from analyses might remove biologically interesting signals.

GitHub file path

shwatkins/PhD/450k_correlation_analysis/assessing_variance.R

3.3 Results

3.3.1 Plotting the full distribution of correlation values

To illustrate the full distribution of pairwise correlations between all sites on the 450k array, I plotted the distribution of correlations in bins of 0.1. I have done so for all three ARIES timepoints; at birth, 7 years and 15-17 years.

For all timepoints, there is a positively skewed distribution of pairwise correlation values. The vast majority of correlations between DNAm sites are between -0.2 and 0.2; the percentage ranges between 85% at birth, 83% at 7 years, and 87% at 15-17 years. There are low numbers of high negative correlations (-1 to -0.8) in all timepoints. There are over 1.6 and 2.3 times more high correlations in the 7 year olds than at birth and 15-17 years, respectively. The distribution is illustrated in Figure 10, and the numbers are in Table 7 as the figure does not illustrate the numbers of high correlations. It is notable that there are many more correlations in the 7 year olds compared to the other two timepoints; this is addressed in the discussion.

Ideally to assess whether the strong correlations identified in this thesis are not due to chance, a distribution of correlations that would be expected by chance would be generated, so that we could compare it to the distribution here. Then it would be clearer whether the tails of the distribution (i.e. the strong correlations) would be expected by chance, or whether they are larger than expected by chance (and therefore representative of biology). It was not within the scope of this thesis to develop a method to do this, but it would be an important part of future work; perhaps through the use of carefully planned permutations.

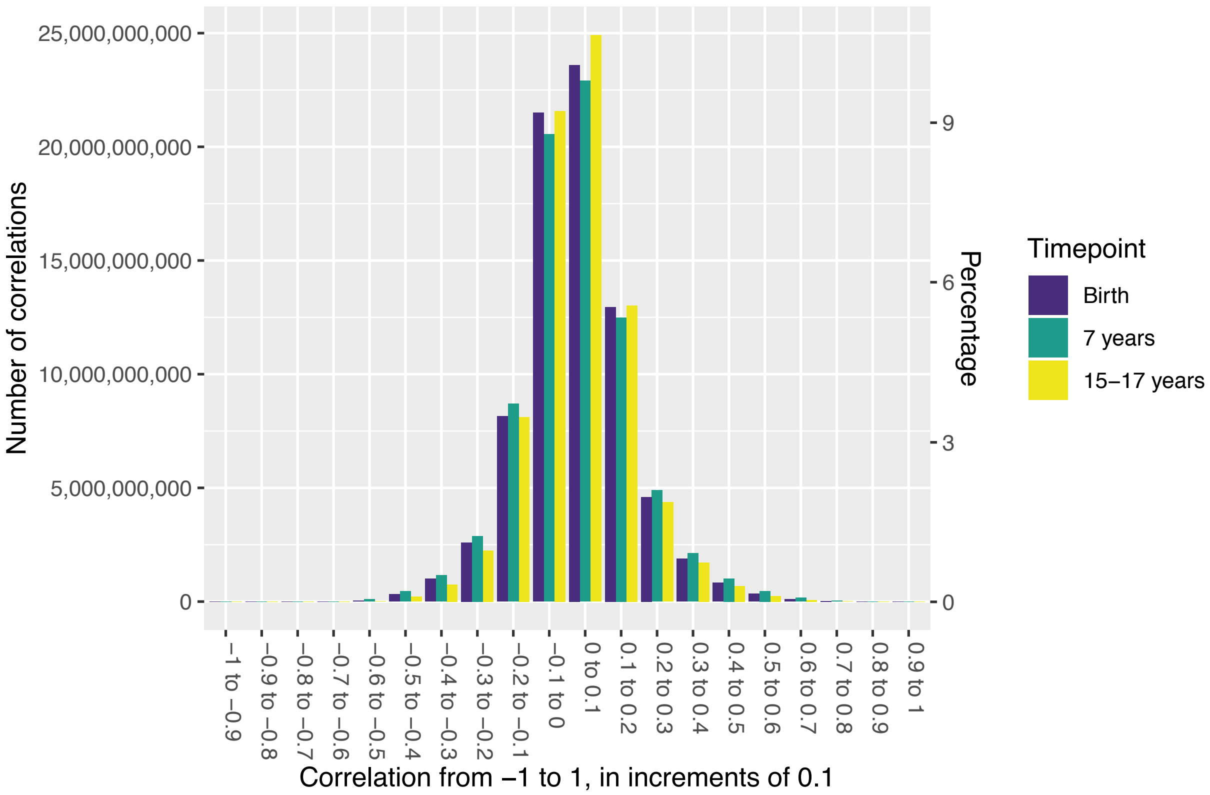


Figure 10: Distribution of pairwise correlations in ARIES

Correlation band	Birth	Percentage of total	7 years	Percentage of total	15-17 years	Percentage of total
-1 to -0.9	3	3.8E-09	7	9E-09	7	9E-09
-0.9 to -0.8	404	5.2E-07	394	5.1E-07	212	2.7E-07
-0.8 to -0.7	4513	5.8E-06	39213	5E-05	1135	1.5E-06
-0.7 to -0.6	923216	0.001	4800488	0.006	183919	0.0002
-0.6 to -0.5	43011431	0.06	95858975	0.1	19430632	0.02
-0.5 to -0.4	317547894	0.4	451211732	0.6	206232315	0.3
-0.4 to -0.3	1002028058	1.3	1151972983	1.5	753955066	1
-0.3 to -0.2	2596753505	3.3	2884113110	3.7	2244916189	2.9
-0.2 to -0.1	8157470531	10.5	8704949753	11.2	8117289019	10.4
-0.1 to 0	21506999111	27.6	20553076285	26.4	21573493322	27.7
0 to 0.1	23598290881	30.3	22904154911	29.4	24911766427	32
0.1 to 0.2	12955550222	16.6	12494857422	16	13021237413	16.7
0.2 to 0.3	4602968925	5.9	4899506543	6.3	4366189865	5.6
0.3 to 0.4	1878135882	2.4	2123259243	2.7	1718676755	2.2
0.4 to 0.5	822139439	1.1	1012645522	1.3	686477261	0.9
0.5 to 0.6	337068287	0.4	468499930	0.6	241908722	0.3
0.6 to 0.7	107032881	0.1	165702180	0.2	70491980	0.09
0.7 to 0.8	22579159	0.03	32935311	0.04	16686137	0.02
0.8 to 0.9	1400404	0.002	2319624	0.003	968239	0.001
0.9 to 1	315	4E-07	1435	1.8E-06	446	5.7E-07

Table 7: Numbers of correlations in each band from -1 to 1 in ARIES

3.3.1.1 Proportions of cis and trans correlations

Examining the proportions of *cis* and *trans* correlations in each correlation band can help to illustrate whether physical proximity between DNAm sites has an effect on the likelihood of them having correlated methylation states. I have defined *cis* as within 1Mb. In ARIES, there are relatively equivalent proportions of *cis* and *trans* correlations for low to moderate correlations, between -0.5 and 0.6. There are marginally more negative *trans* correlations; and there are slightly more *cis* correlations between 0 and 0.2. These patterns are maintained from birth to adolescence, and are illustrated in Figure 11. Because such increasingly small proportions of correlations are <-0.5 and >0.5, they are best viewed in a table; Table 8 shows that a much smaller percentage of trans correlations tend to be found between 0.9 and 1, and between -1 and -0.8, across all age groups. There also appears to be a slight reduction in percentages of correlations r>0.4 in the adolescent group, with a larger percentage of correlations between 0 and 0.1.

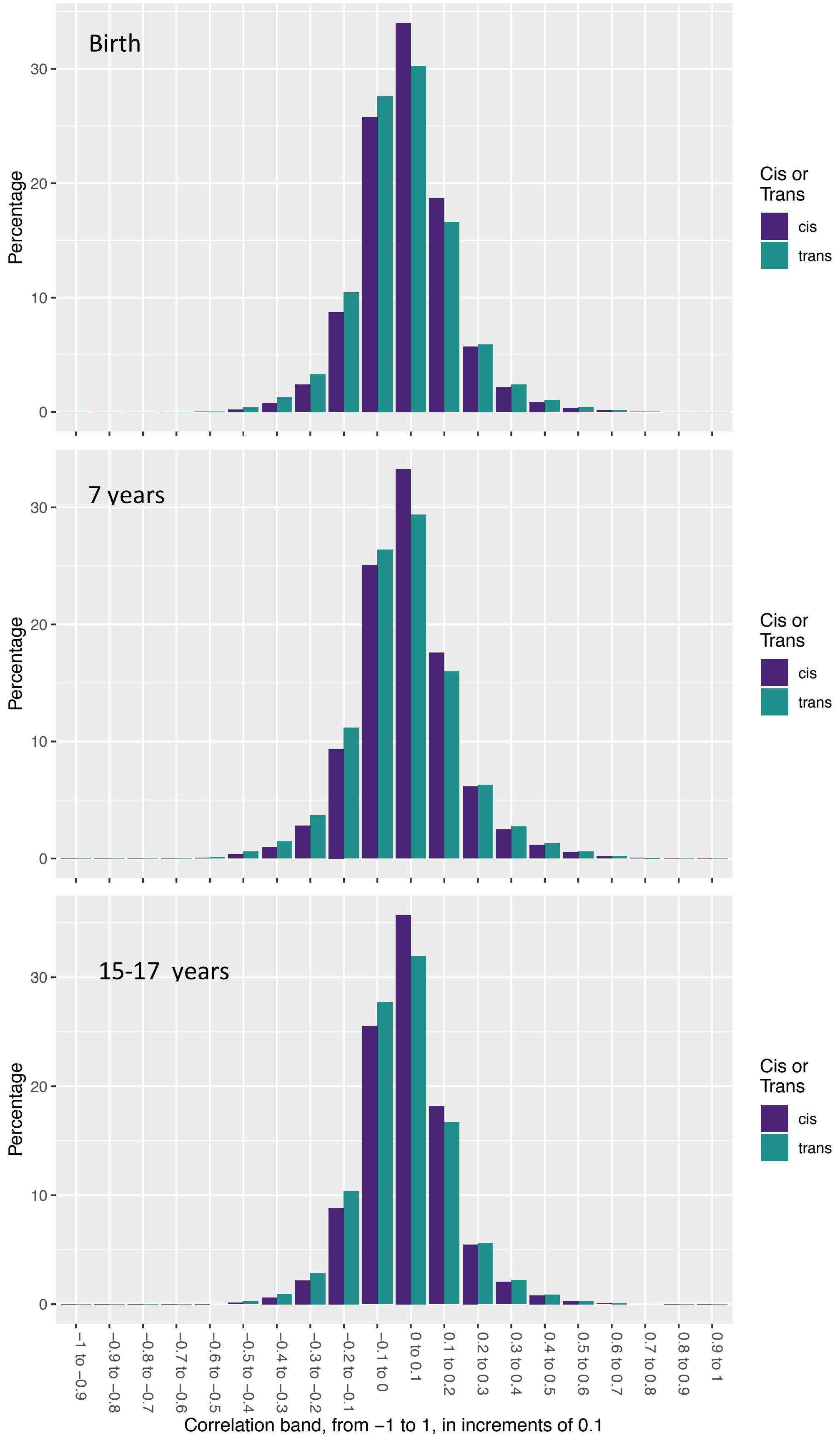


Figure 11: Percentages of the total numbers of cis and trans correlations that sit within each 0.1 correlation band, in ARIES at birth (top), 7 years (middle) and 15-17 years (bottom).

	Birth		7 years		15-17 years	
Correlation band	Cis	Trans	Cis	Trans	Cis	Trans
-1 to -0.9	0	3.85E-09	5.8E-06	0	5.8E-06	0
-0.9 to -0.8	2.5E-06	5.2E-07	2.9E-05	4.6E-07	2.9E-05	2.3E-07
-0.8 to -0.7	3.6E-05	5.7E-06	9.7E-05	5E-05	5E-05	1.4E-06
-0.7 to -0.6	0.0008	0.001	0.004	0.006	0.0003	0.0002
-0.6 to -0.5	0.03	0.06	0.07	0.12	0.01	0.02
-0.5 to -0.4	0.23	0.41	0.34	0.58	0.15	0.26
-0.4 to -0.3	0.8	1.29	0.98	1.5	0.63	0.97
-0.3 to -0.2	2.4	3.3	2.8	3.7	2.2	2.9
-0.2 to -0.1	8.7	10.5	9.3	11.2	8.8	10.4
-0.1 to 0	25.9	27.6	25.1	26.4	25.5	27.7
0 to 0.1	34	30.3	33.2	29.4	35.7	32
0.1 to 0.2	18.7	16.6	17.6	16	18.2	16.7
0.2 to 0.3	5.7	5.9	6.1	6.3	5.5	5.6
0.3 to 0.4	2.1	2.4	2.5	2.7	2	2.2
0.4 to 0.5	0.89	1.1	1.1	1.3	0.81	0.88
0.5 to 0.6	0.38	0.43	0.52	0.6	0.31	0.31
0.6 to 0.7	0.14	0.14	0.2	0.21	0.1	0.09
0.7 to 0.8	0.04	0.03	0.05	0.04	0.03	0.02
0.8 to 0.9	0.003	0.002	0.006	0.003	0.003	0.001
0.9 to 1	0.0002	9.9E-08	0.0004	1.2E-06	0.0003	4E-08

Table 8: Percentage of cis and trans correlations in each correlation band, for each ARIES timepoint

3.3.2 Illustration of cis correlation structure across the genome

To illustrate *cis* correlation structure, I created a decay plot for all cis-correlating sites genome-wide, and decay plots for each chromosome separately. I separated out positive and negative correlations, to identify whether they differ in terms of structure; and I added variance to the plot to demonstrate the uncertainty around the binned estimates.

The correlation structure is virtually identical between timepoints, and across chromosomes. The genome-wide plot shows a much less noisy and therefore more accurate curve as compared to the individual chromosomes. The mean positive correlation at immediately adjacent sites is 0.4, and reduces to a constant of around 0.125 by around 3kb. This is similar to what was found in (Y. Liu et al., 2014) and (Saffari et al., 2018), and it is possible that this correlation level is simply reflecting the baseline level of technical noise in the array data, or the average strength of correlation between DNAm sites that are not functionally related.

My work illustrates that there is substantial heterogeneity in the decay of cis correlations, and that the separation of positive and negative correlations is likely to give a more accurate picture. The decay of DNAm correlations to a constant by 3kb supports the idea that DNAm correlation structure is unlikely to be driven by LD, because the decay is over a vastly smaller distance (see LD decay plot in (Genomes Project et al., 2015) for comparison). The mean negative correlation is fairly constant at around -0.1, and does not change much with genomic distance; however negative correlations are at their lowest when sites are immediately adjacent, and have a slight peak between 2.5 and 4kb. It is possible that this illustrates that close sites tend to be positively correlated, and the peak may suggest negative correlations between transcription start sites or promotors and sites in gene bodies, which one might expect to be negatively correlated. Plots of the whole genome and chromosome 1 at 7 years are shown below in Figure 12; for plots of chromosomes 1:5 and 15:19 (for a broader representation of the genome), at each of the three ARIES timepoints, please see Appendix 1.

The histograms of *cis* correlation values within 1kb illustrate the presence of high correlations a bit more clearly (see Figure 13 below for chromosome 1 and 20 examples at 7 years; plots for chromosomes 1:5 and 15:19 are in Appendix 1 with the *cis* decay plots). Although there are small numbers of high (>0.8) correlations (around 0.27%), there is a greater proportion of them within 1kb compared to the distributions of correlations genome-wide (between 0.001 - 0.003%).

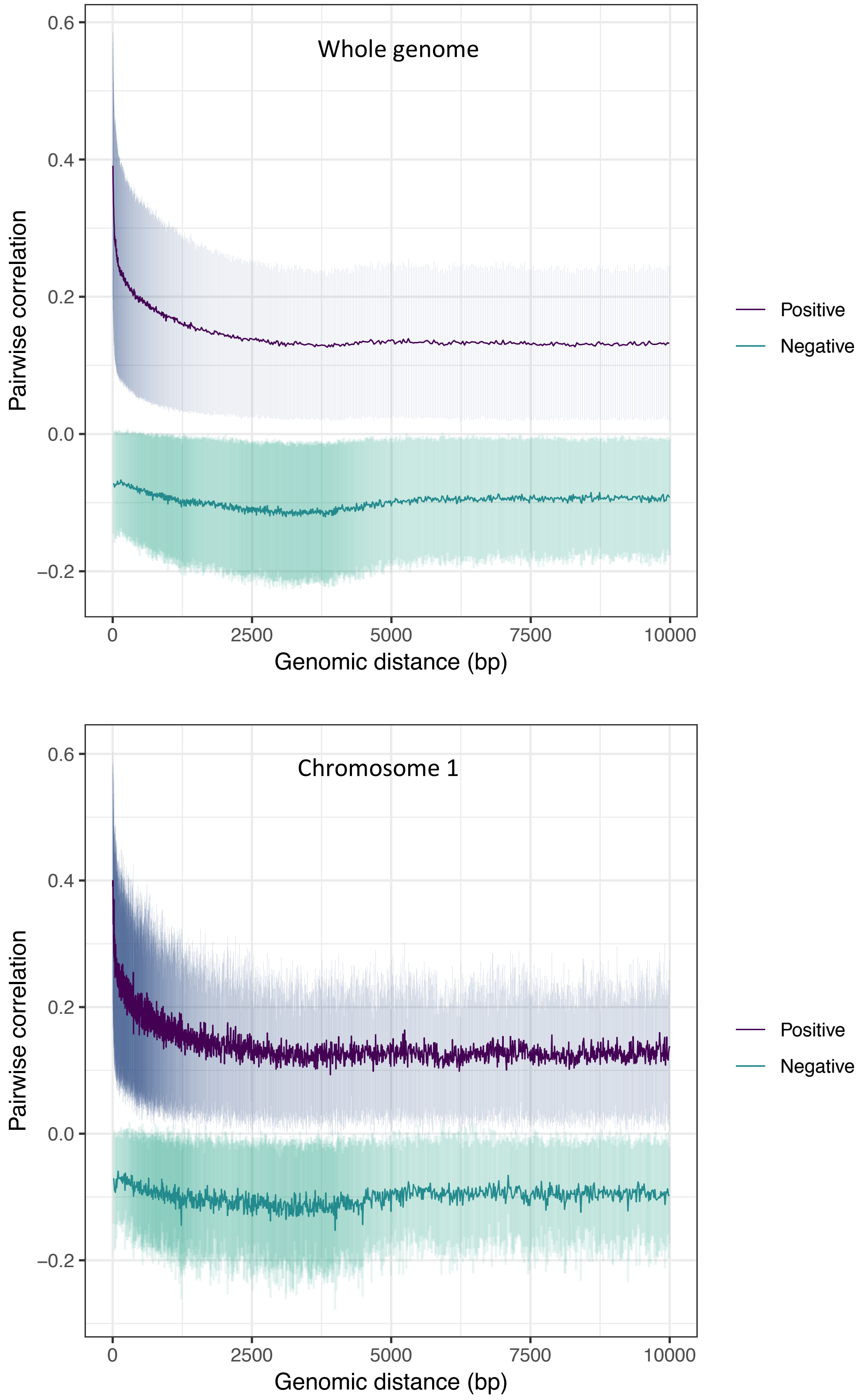


Figure 12: Decay plots of cis correlations, for the whole genome (top) and chromosome 1 (bottom), at 7 years old in ARIES.

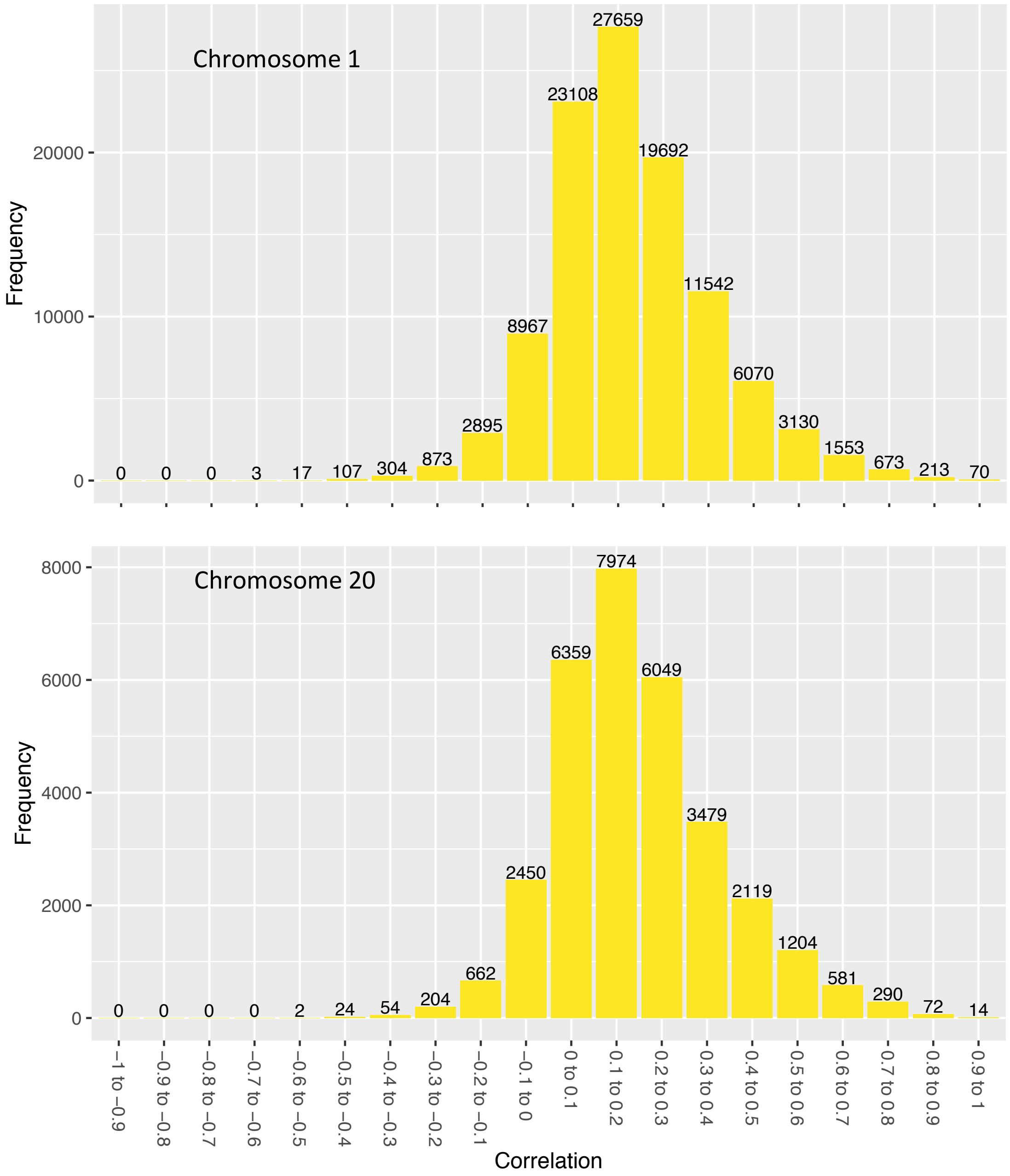


Figure 13: Histogram of correlation values between all probes within 1kb of each other, on chromosomes 1 and 20, in ARIES 7 year olds.

To further illustrate the correlations in the cis decay plot, I plotted examples of pairwise correlations at different distances. We can see from the plots in Figure 14 that immediately adjacent cis correlations are strong clear correlations. At a slightly greater distance 80-250bp) we see some examples of DNAm sites clearly driven by genotype, with clear clusters. Cis sites that are around a distance of 3kb (where the decay plot reduces to a constant) are poorly correlated, supporting the idea that the constant may reflect background noise in the data.

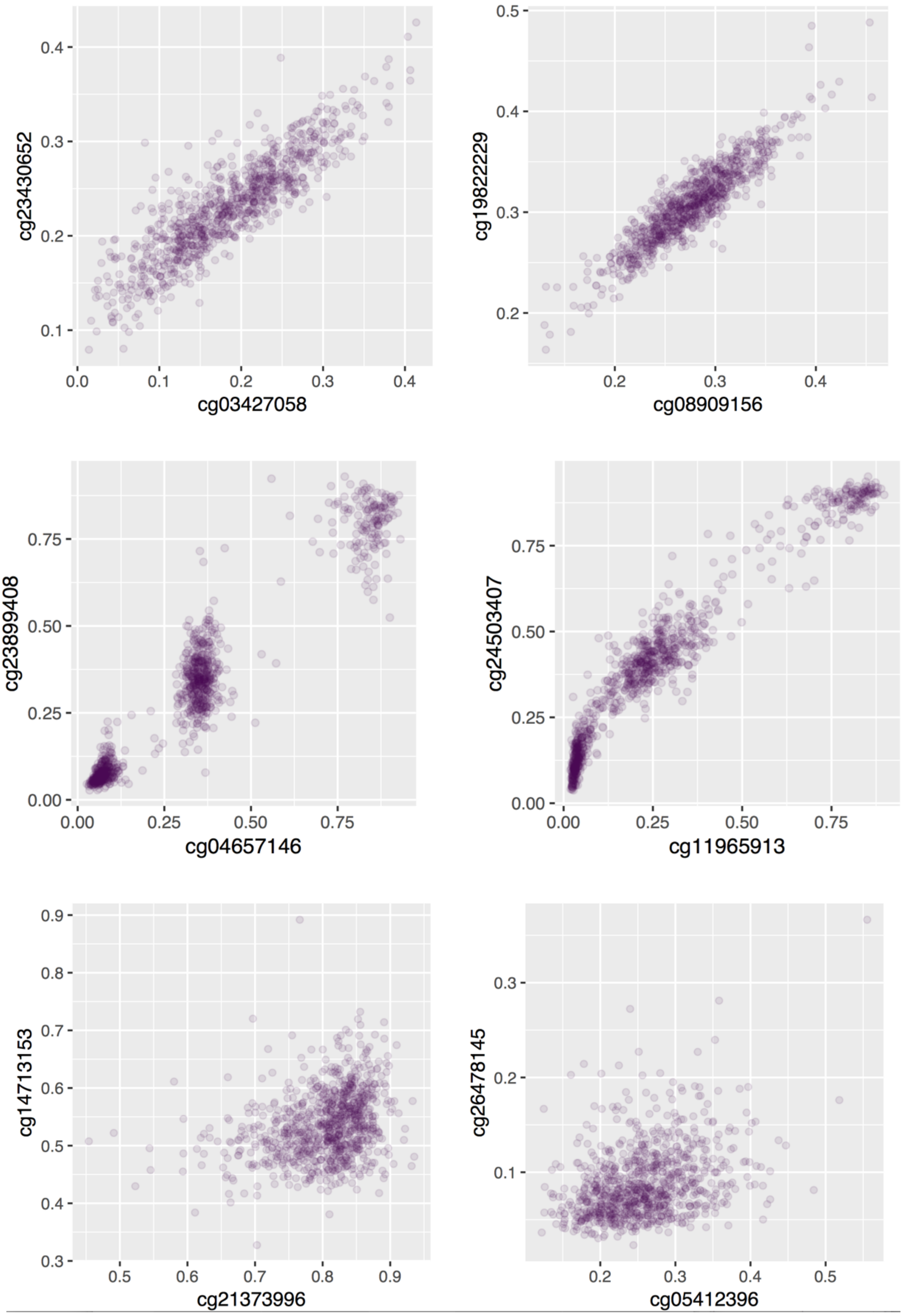


Figure 14: Scatter plots of the mean unadjusted beta value of correlating probe pairs in cis. Top panel shows examples of immediately adjacent sites; middle panel shows examples of sites 80-250bp apart that are clearly influenced by genotype; and bottom panel shows sites around 3kb apart that reflect the baseline level of correlation.

3.3.3 Genetic influences on correlations between DNAm sites

3.3.3.1 Influences of heritability estimates on DNAm sites

To estimate the contributions of heritability and environment to the variance of DNAm sites that feature in the 20 ranges of correlation (from -1 to 1), I utilised the estimates of these contributions for DNAm sites across the 450k created by (Hannon et al., 2018). I find that at birth, variation in DNAm for the strongest positive correlations (0.9 to 1) is mostly attributable to genetic influence. There is some contribution of heritability to variation in sites with correlations between -0.8 and -1, and for the vast majority of DNAm sites heritability makes a very low contribution to variability. Common environment makes a low contribution to DNAm variation, with the highest influence being to the strongly negatively correlated sites (-0.8 to -1). Unique environment has little influence over variation at sites which correlate between 0.9 and 1; for all other correlation bands it is the source of most of the variability in DNAm level. In the 7 year olds, there are two peaks for the influence of heritability, around 0 and 1, suggesting there may be increasing environmental influences driving variation of strongly positively correlated DNAm sites between birth and 7 years. The strong negative correlating sites have most of their variability explained by heritability at 7 years, another change from birth. Common environment contributes little to the variability of the DNAm sites, and unique environment contributes to all of the correlation bands aside from the two extremes (0.9 to 1 and -0.9 to -1). In the 15-17 year olds, the strong positive and negative correlating DNAm sites have large contributions to variability from heritability; and these highly correlated sites have little contribution from common or unique environment. This is illustrated in Figure 15 for birth, Figure 16 for the 7 year olds and Figure 17 for the 15-17 year olds.

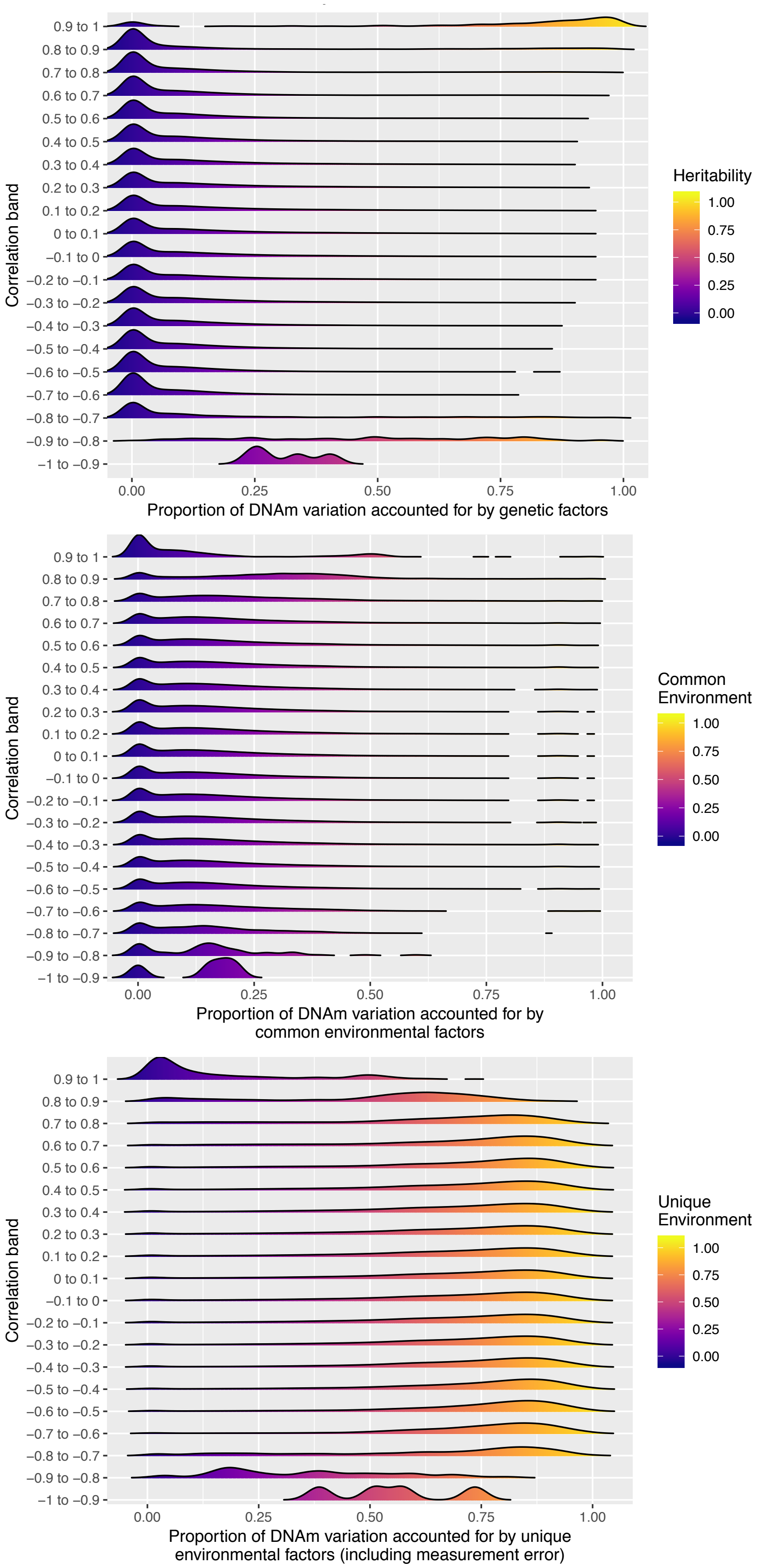


Figure 15: Ridgeline plots illustrating the estimated contributions of genetic and environmental factors to variation in DNAm sites which feature in 20 ranges of correlation strength, at birth in ARIES.

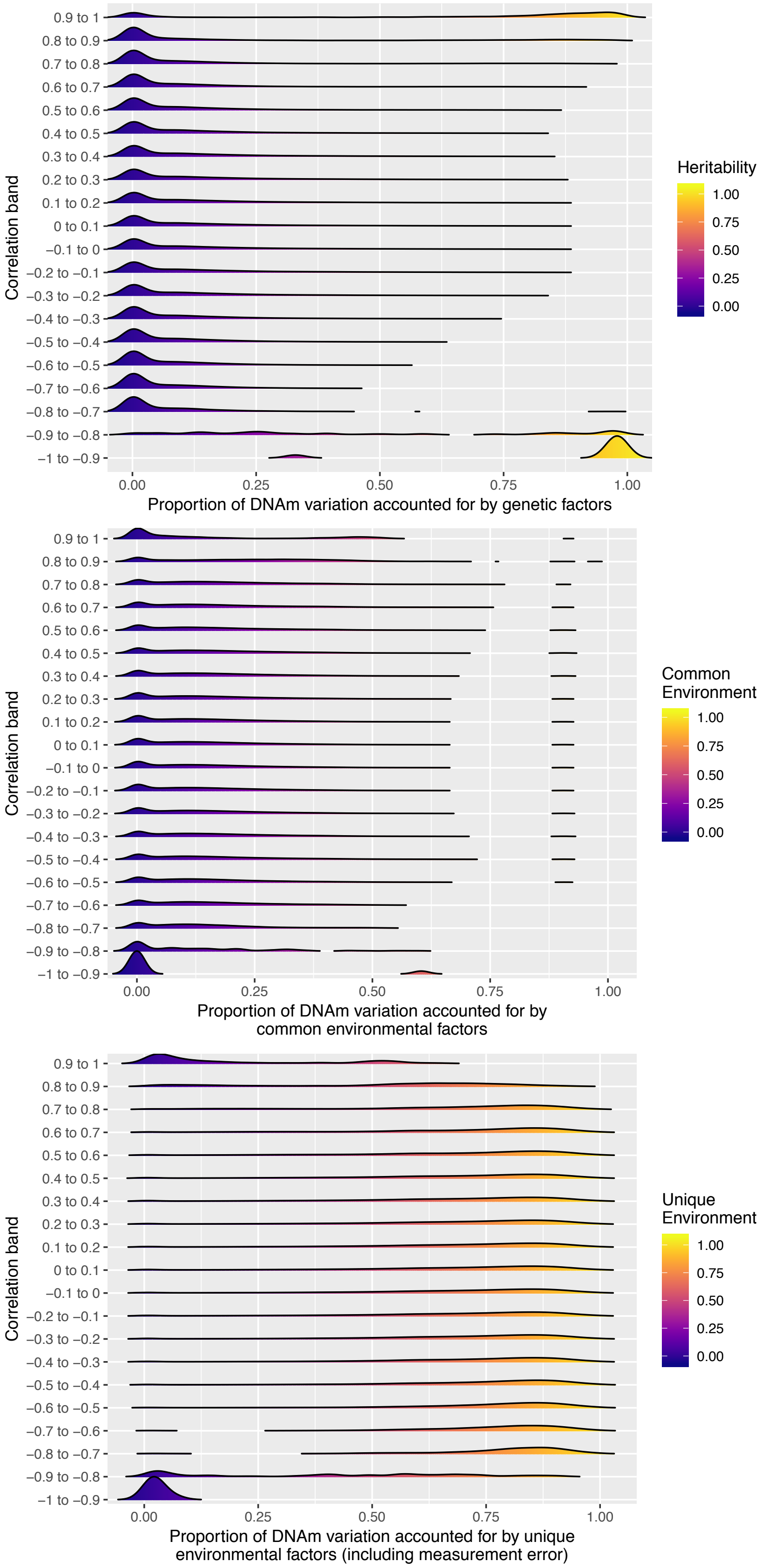


Figure 16: Ridgeline plots illustrating the estimated contributions of genetic and environmental factors to variation in DNAm sites which feature in 20 ranges of correlation strength, in ARIES 7 year olds.

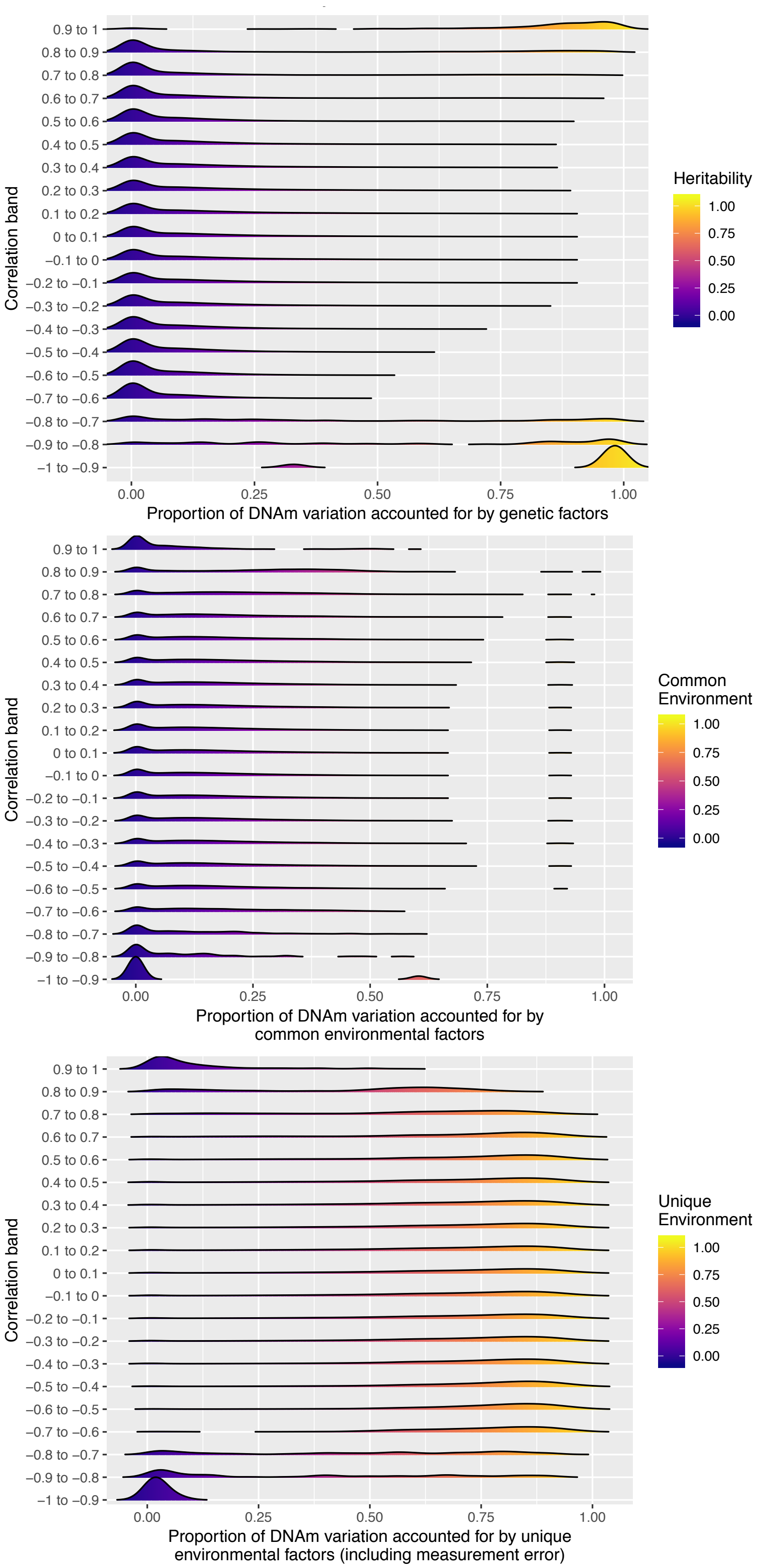


Figure 17: Ridgeline plots illustrating the estimated contributions of genetic and environmental factors to variation in DNAm sites which feature in 20 ranges of correlation strength, in ARIES 15-17 year olds.

This analysis has identified the heritability of DNAm sites involved in the correlation in isolation. Another way to approach this would be to look at the heritability of the correlations themselves - we could look at the heritability of both DNAm sites, and whether those with high heritability are driven by the same mQTL; or we could use a twin dataset to examine the differences in correlations between MZ and DZ twins. As this thesis did not utilise twin data, I assessed whether both DNAm sites in a strongly correlated (>0.9) pair were influenced to the same degree by heritability. Results show that the majority of highly correlated DNAm sites are influenced by heritability to the same degree (Pearson correlation = 0.98). Splitting the sites by whether they are in cis or trans shows very clearly that cis correlations >0.9 tend to be strongly influenced by heritability, and trans correlations tend to be very weakly influenced by heritability (shown in Figure 18).

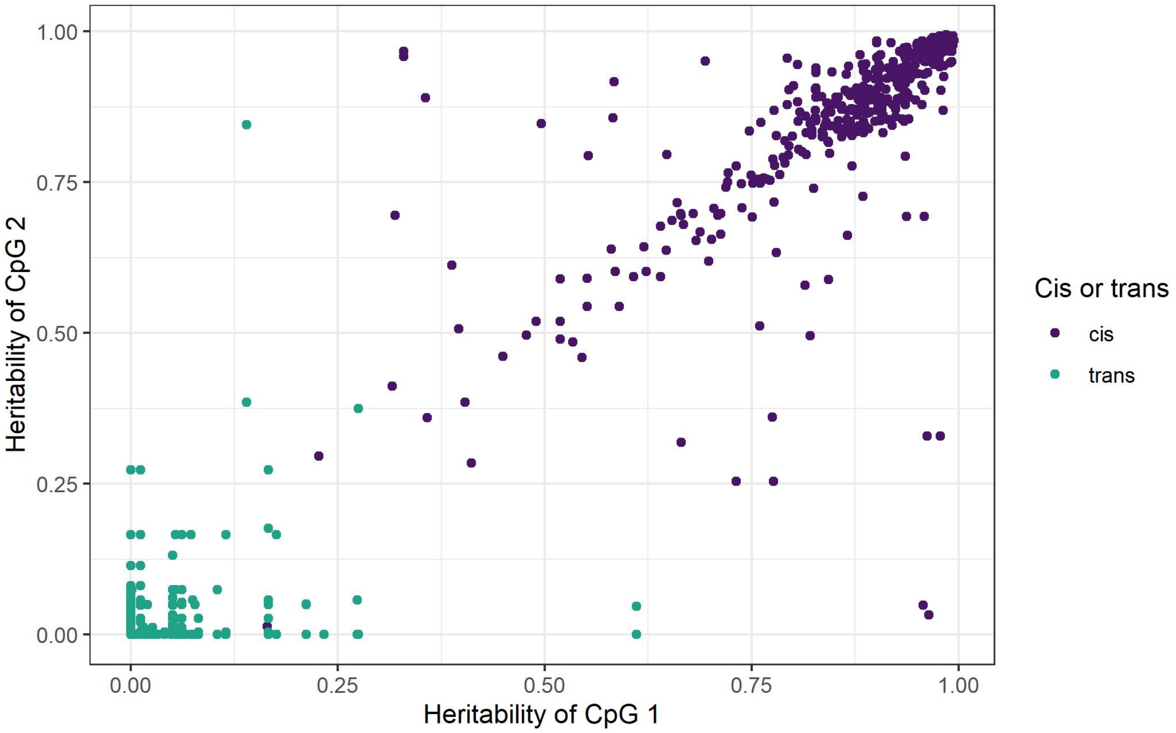


Figure 18: Scatter plot showing the relationship between the heritability estimates for each DNAm site in a highly correlating pair. Results are coloured by cis (purple) and trans (green) pairs. Heritability was taken from the estimations by Hannon et al (2018).

3.3.3.2 Influence of mQTLs on correlations between DNAm sites

3.3.3.2.1 Plotting the proportion of correlating DNAm sites associated with mQTLs

To identify whether strong correlations might be driven by known mQTL associations, and whether this might differ for cis and trans correlations, I plotted the percentage of correlations with 0, 1 or 2 of the DNAm sites associated with an mQTL, stratified by correlation strength. I plotted this for each of the ARIES timepoints to illustrate whether this might change over time.

I find that strong *cis* correlations (-1 to -0.7 and 0.9 to 1) have around 100% of DNAm sites in each correlating pair associated with an mQTL. The percentages are quite similar over time, with a very slight increase in the percentage of strong correlations being associated with 2 mQTLs at 15-17 years, particularly 0.8 to 0.9 and -0.7 to -0.8. The strong negative *trans* correlations (-1 to -0.8) are also most likely to have both DNAm sites in a correlating pair associated with an mQTL, although the percentages are somewhat lower than for *cis* correlations, at around 75%. In contrast, the strong positive *trans* correlations are most likely to have neither of the DNAm sites associated with an mQTL. These are illustrated in Figure 19. This shows slight increases in the proportion of correlations that have both DNAm sites associated with an mQTL as ARIES participants age. This is noticeable for cis correlations between -0.8 and -0.6, and between 0.8 and 0.9 in the adolescents. The proportion of trans correlations 0.9 to 1 that are associated with no mQTLs decreases gradually between birth and adolescence.

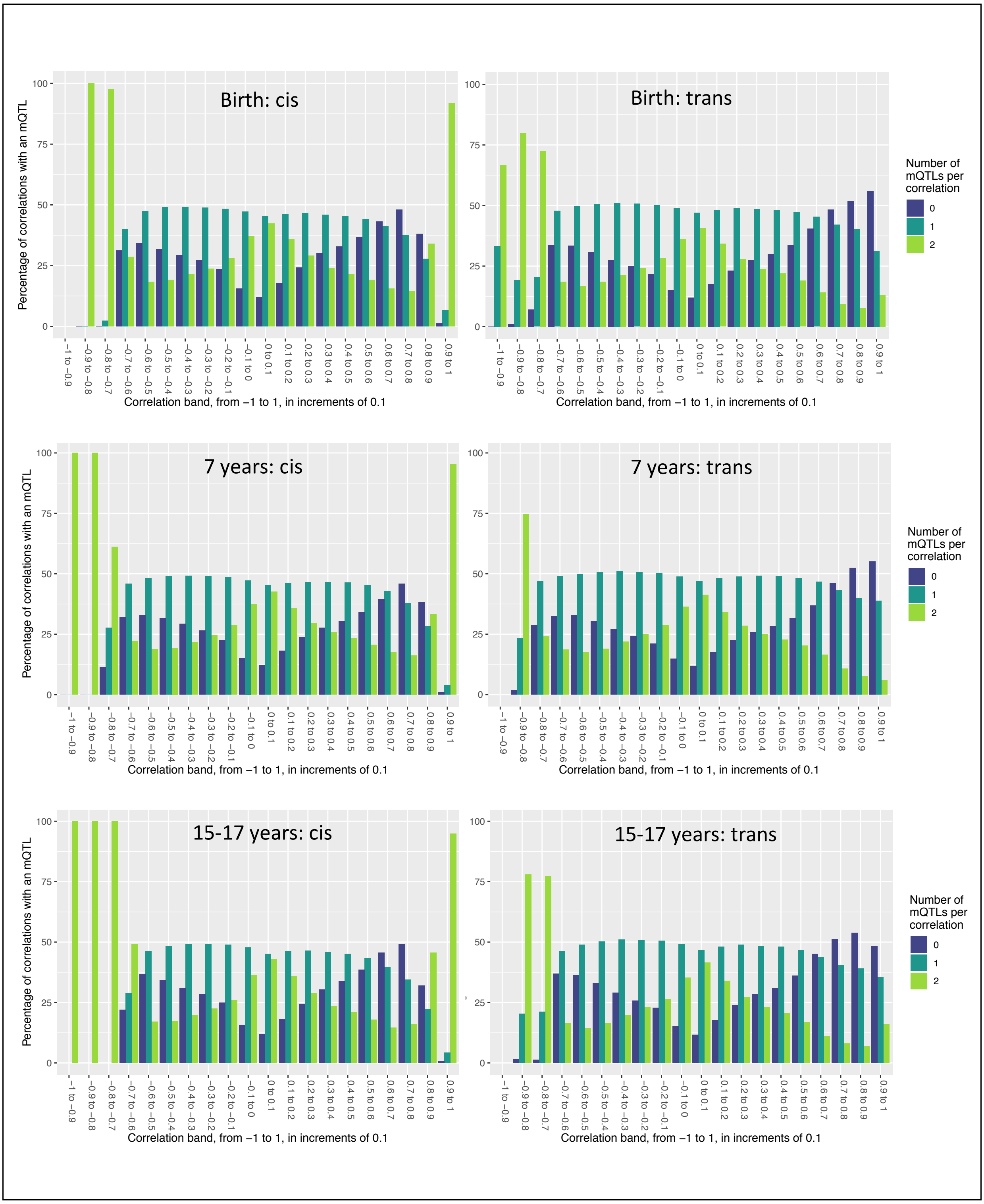


Figure 19: Bar plots of the percentage of pairwise correlations in each correlation range that have 0, 1 or 2 DNAm sites associated with an mQTL identified by the GoDMC consortium. Split by cis (left) and trans (right) correlating pairs, at birth (top), 7 years (middle) and 15-17 years (bottom),

3.3.3.2.2 Removing genetic influence from cis correlation plots

To assess the extent of genetic influence on cis DNAm correlations, I regressed out the effect of the strongest *cis* mQTL from the DNAm data in the ARIES 7 year olds. Correlating

DNAm site pairs were not necessarily adjusted for the same SNP; they were adjusted individually for their strongest mQTL. The SNPs adjusted for were in cis with the DNAm site (defined by the GoDMC consortium as within 1Mb of the DNAm site), although an analysis of the impact of distance between SNP and DNAm site was was not carried out, to ascertain whether proximity to a SNP might be a factor in the impact on correlation. I re-plotted the *cis* decay plot, with both the adjusted and unadjusted values, and without the standard deviation, for a clear comparison. I find that mean *cis* correlations between DNAm sites in immediate proximity reduce from r ≈ 0.45 to r ≈ 0.27 when the strongest *cis* mQTL is regressed out of the DNAm data. By a distance of 2.5kb, the adjusted correlations reduces to a constant mean of r ≈ 0.05. This is illustrated in Figure 20. I also assessed the mean reduction in correlation per bin, between DNAm data adjusted and unadjusted for cis mQTLs. This clearly shows that correlations between immediately adjacent DNAm sites are more strongly affected by mQTLs than more distant DNAm sites. This is shown in Figure 21.

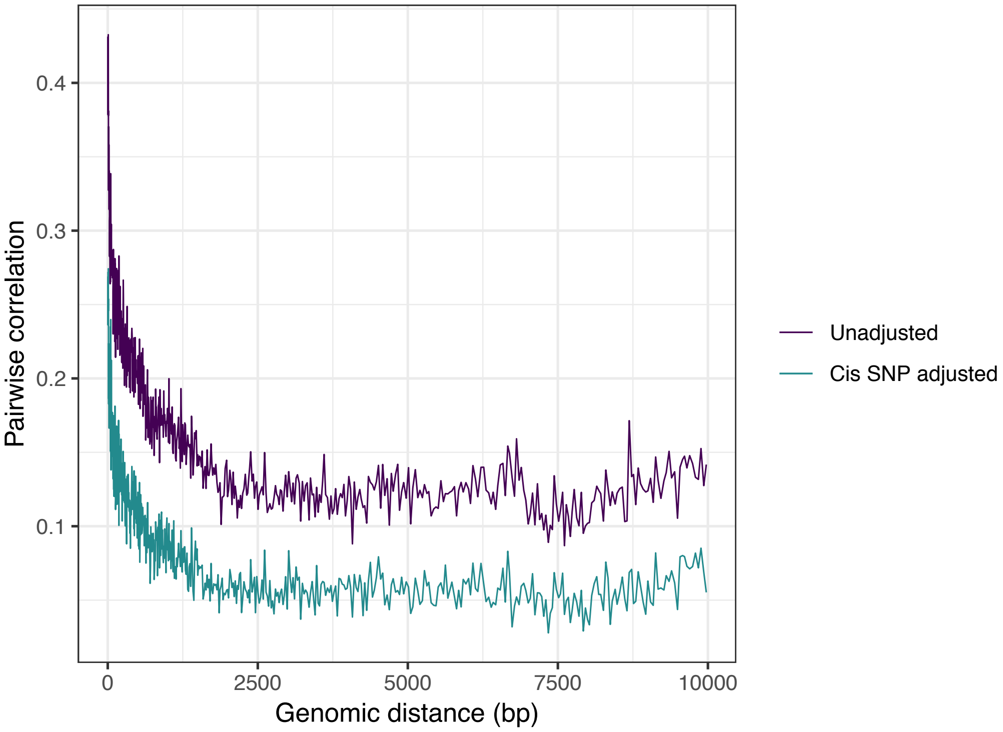


Figure 20: Cis decay plot on chromosome 20 in ARIES 7 year olds, showing the unadjusted binned decay of correlation over distance (purple) and the decay of correlation over distance when adjusting DNAm values for the strongest associated cis SNP.

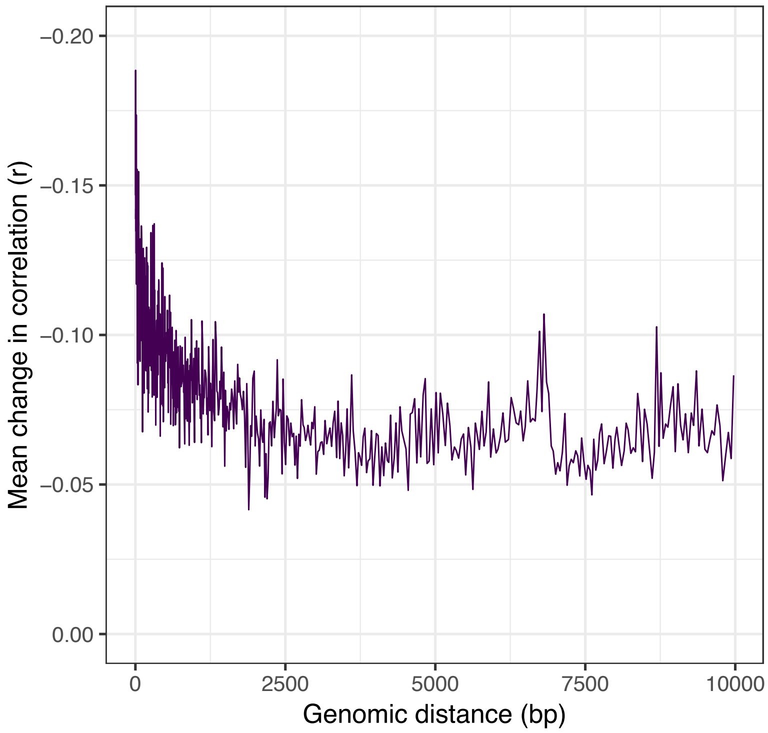


Figure 21: Decay plot of the mean change in correlation binned over genomic distance, between DNAm data adjusted for the strongest mQTL, and unadjusted for mQTL associations.

3.3.3.3 Influence of LD on correlation structure

To take a slightly different approach to the question of the influence of LD on DNAm correlation structure, I paired up DNAm sites with their strongest *cis*-SNP, theorising that *cis* mQTLs might be the most likely source of LD influence on DNAm correlations, and that these might not always be in immediate proximity to each other. It is possible that LD may have some influence over DNAm correlation structure even though it is in smaller portions. If LD were driving the correlation structure, one would expect that in regions of high LD the connectivity (kTotal) of DNAm sites would also be high, and so form a clear positive correlation between LD score and kTotal.

I assessed this across chromosome 20, because it is possible that plotting specific regions does not give a complete answer. I found that in the ARIES 7 year olds across chromosome 20, the connectivity of DNAm sites bears no relation to the LD scores of their strongest *cis* mQTL. This suggests correlations between DNAm sites are not simply a reflection of the relationships between SNPs, because if LD drove DNAm correlation structure, one would expect DNAm sites to have connectivity proportional to the LD score of associated mQTLs. The plot of this relationship can be found in Figure 22.

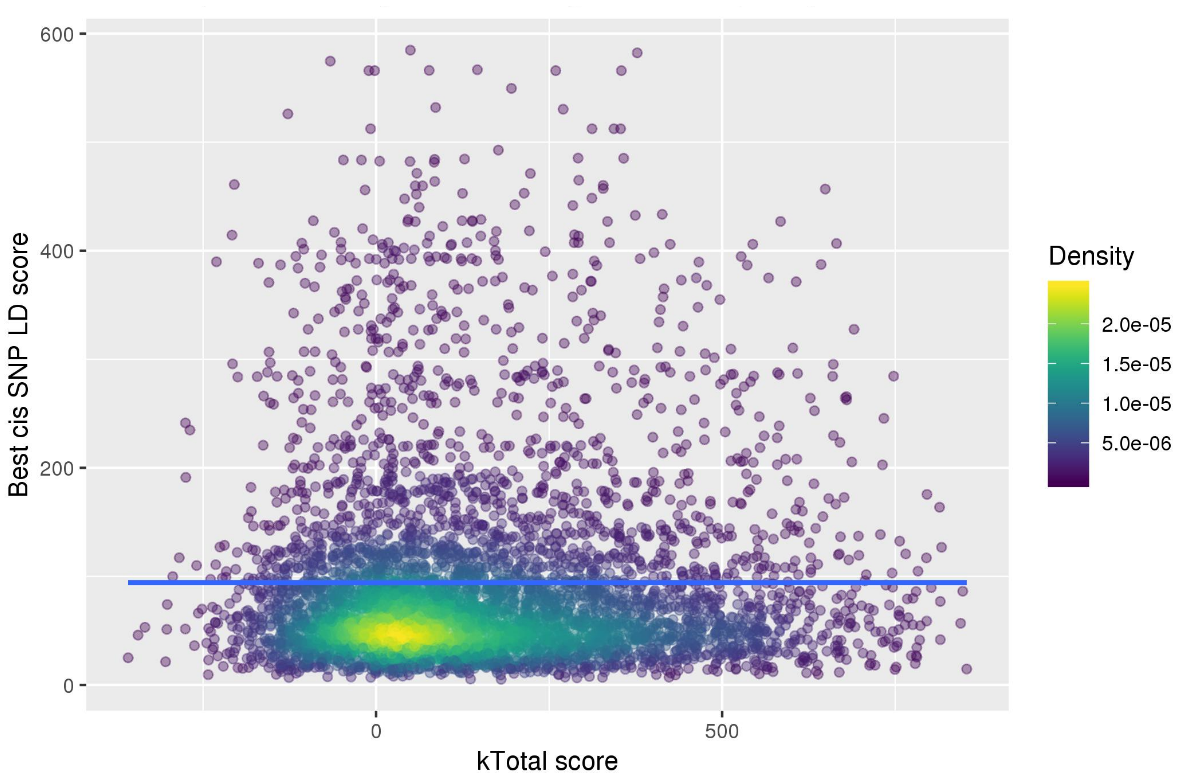


Figure 22: Scatter plot of DNAm connectivity and LD of the most strongly associated cis SNP, in ARIES 7 year olds. The blue line is a smoothed regression line; points are coloured by density.

3.3.4 Analysis of strong correlations

Strong correlations (r>0.9) were analysed for enrichments. Table 9 shows the number of probes in this analysis at each ARIES timepoint. There is a substantially larger number of trans correlations in the ARIES 7 year olds. There are a few possible explanations for this difference. It may be because of the differing sample types used for the ARIES timepoints, specifically the use of whole blood for 94% of the 7 year olds. The use of whole blood may have introduced more variance into the data and thus the correlations in the 7 year olds may be slightly higher. An alternative explanation is that higher correlations were missed at birth due to the confounding of slide and sample type, leading to correlations being reduced by batch effects that could not be disentangled. Under this model it may be possible that correlations between DNAm sites reduce with age, but this study could not detect that because of this confounding.

Timepoint	Cis correlations	Trans correlations
Birth	238	77
7 years	524	911
15-17 years	415	31

Table 9: Numbers of cis and trans correlations r>0.9 at the three ARIES timepoints.

3.3.4.1 Genomic region enrichment

3.3.4.1.1 Chromatin states

3.3.4.1.1.1 Cis correlations

I find that across all three ARIES timepoints *cis*-correlating sites r>0.9 are strongly enriched for locations at poised promoters (PromP), with a slight enrichment for weak enhancers in blood (EnhW1 and, in some samples, EnhW2). Poised promoters are bivalent chromatin states, which can be either activating or repressing, and tend to be located in promoter regions of developmentally important genes (Bernhart et al., 2016). This is illustrated for the ARIES 7 year olds in Figure 23; as the result is very similar across timepoints the figures for birth and 15-17 years are in Appendix 2.

3.3.4.1.1.2 Trans correlations

At all three ARIES timepoints, *trans*-correlating sites r>0.9 are strongly enriched for locations at promotors downstream of transcription start sites 1 (PromD1), and active transcription start sites (TssA). In the 7 year-olds only, they are weakly enriched for locations at promotors upstream of transcription start sites (PromU), and transcription regulatory (TxReg). It is possible this is due to the larger number of *trans* DNAm correlations in the 7 year old data. The plots for the 7 year-old enrichments are in Figure 24; the plots for birth and 15-17 years are in Appendix 2.

3.3.4.1.2 Histone modifications

3.3.4.1.2.1 Cis correlations

Enrichment for histone modifications for *cis* correlating DNAm sites is found only for H3R17me2 at birth. As this is only tested in two breast tissue cell lines, and the enrichment is quite weak, the conclusions we can draw from this are fairly limited. Relating this back to the chromatin states associated with cis-correlating sites in section 3.3.4.1.1.1, it has been shown that poised promoters are enriched for H3K4me1, H3K4me3 and H3k27me3 (Bernhart et al., 2016); H3K27me3 has measures in hematopoietic stem cells in the data I have tested, but these histone marks were not enriched at the *cis* correlating sites. The plot for birth is in Figure 25; the plots for 7 and 15-17 year olds are in Appendix 2.

3.3.4.1.2.2 Trans correlations

For all ARIES timepoints there is clear enrichment for H3K9K14ac for the strongly correlated *trans* DNAm sites, in experiments using blood relevant tissue (hematopoietic stem cells). There is enrichment for H3K9me3 in the sites identified in some of the experiments, but not others, and so it is unclear whether there really is enrichment for H3K9me3. H3K27me3 does not show enrichment at these *trans*-correlating sites in blood. The plot for the 7 year olds can be found in Figure 26; the plots for birth and the 15-17 year olds are in Appendix 2.

3.3.4.1.3 Transcription factor binding sites

3.3.4.1.3.1 Cis correlations

The transcription factor binding site (TFBS) enrichment analysis shows limited enrichments for TFBS in cis-correlating sites; however there are a few clear enrichments that are consistent across the three ARIES timepoints. The TFs that are enriched in blood (and therefore most relevant to this study) are RNA Polymerase III (Pol3), BRF1, and BDP1. BRF1 and BDP1 are two subunits of the transcription factor TFIIIB, which is required for Pol3-mediated transcription (Abascal-Palacios, Ramsay, Beuron, Morris, & Vannini, 2018), and so their co-occurrence points to Pol3-mediated transcription having functional relevance to strong *cis* DNAm correlations. This is illustrated for the 7 year olds in Figure 27; the plots for birth and 15-17 year olds are in Appendix 2.

3.3.4.1.3.2 Trans correlations

The TFBS enrichment analysis for *trans*-correlating DNAm sites shows overlap with many of the TFBS. Perhaps notably CTCF (which aids trans-chromosomal interactions (S. Kim, Yu, & Kaang, 2015)) is only enriched in the 7 year olds. In blood there appears to be particularly strong enrichment for POL2, PHF8, YY1, TAF1, ZBTB7A, MAX, MAZ, ELF1, EGR1, and CMYC. This is illustrated in Figure 28 for ARIES 7 year olds, and in appendix x for birth and 15-17 year olds.

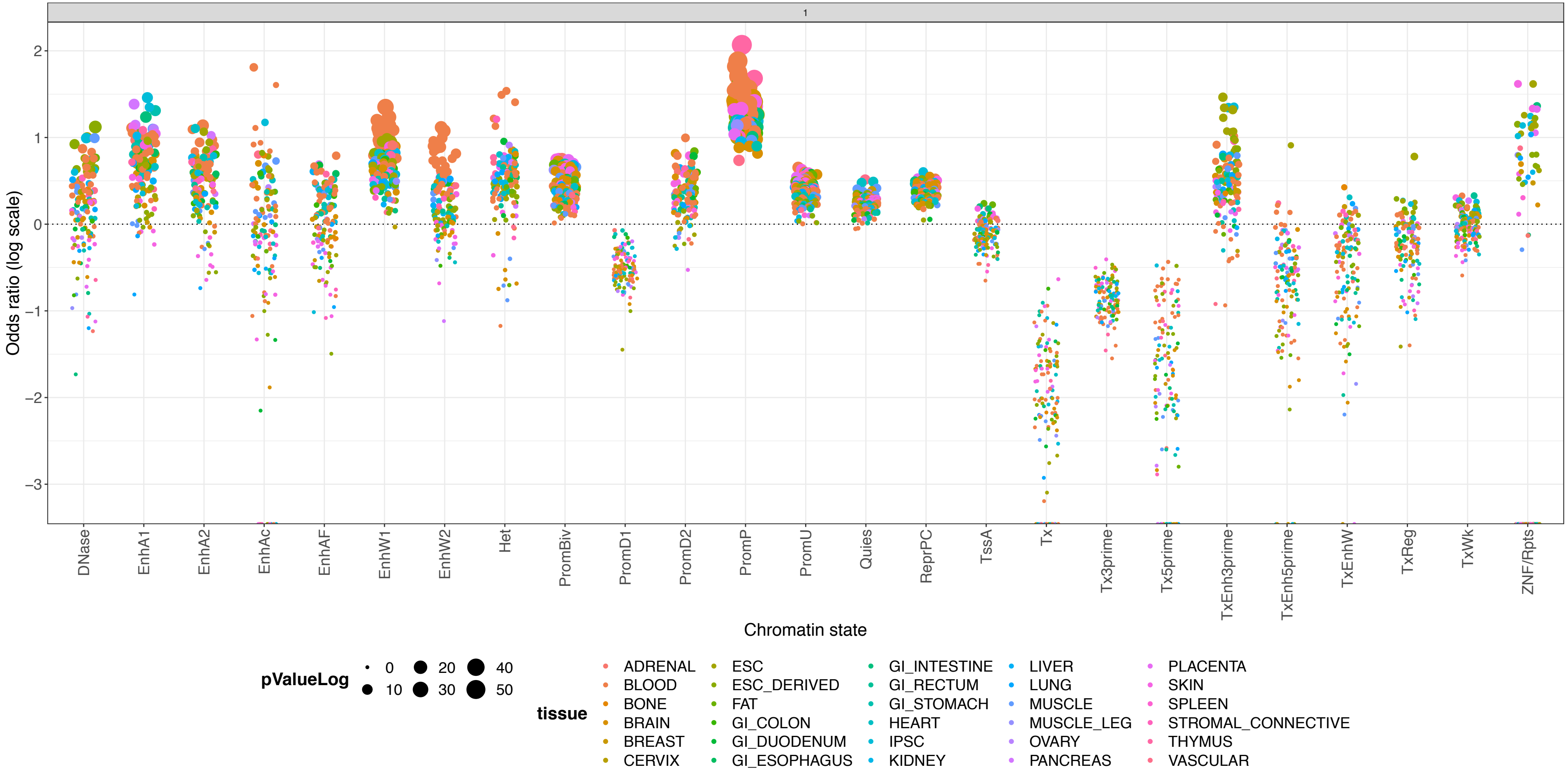


Figure 23: Bubble plot showing the enrichment for the Roadmap Epigenomics 25 chromatin states, for cis-correlating DNAm sites r>0.9, in ARIES 7 year olds

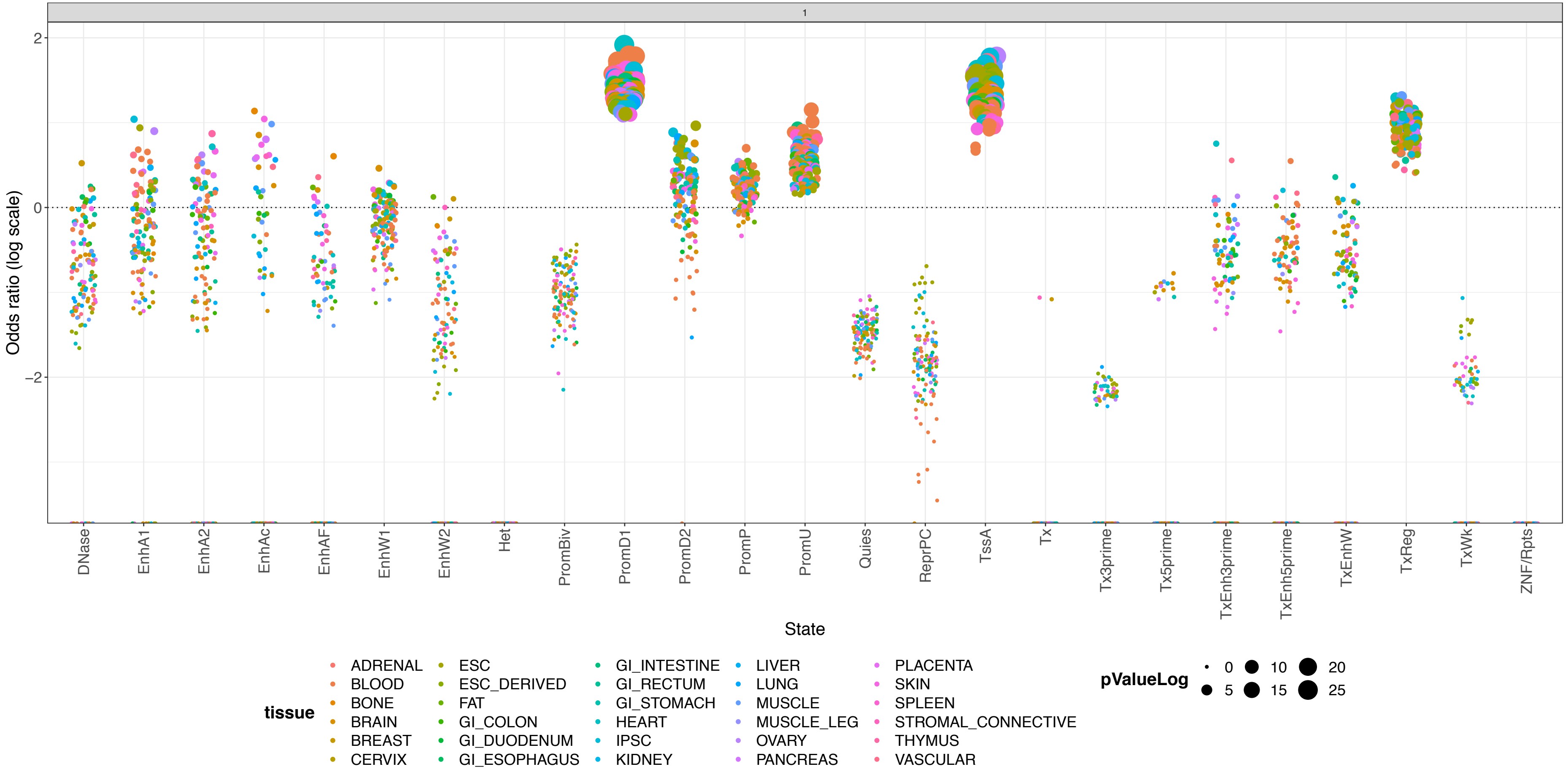


Figure 24: Bubble plot showing the enrichment for the Roadmap Epigenomics 25 chromatin states, for trans-correlating DNAm sites r>0.9, in ARIES 7 year olds

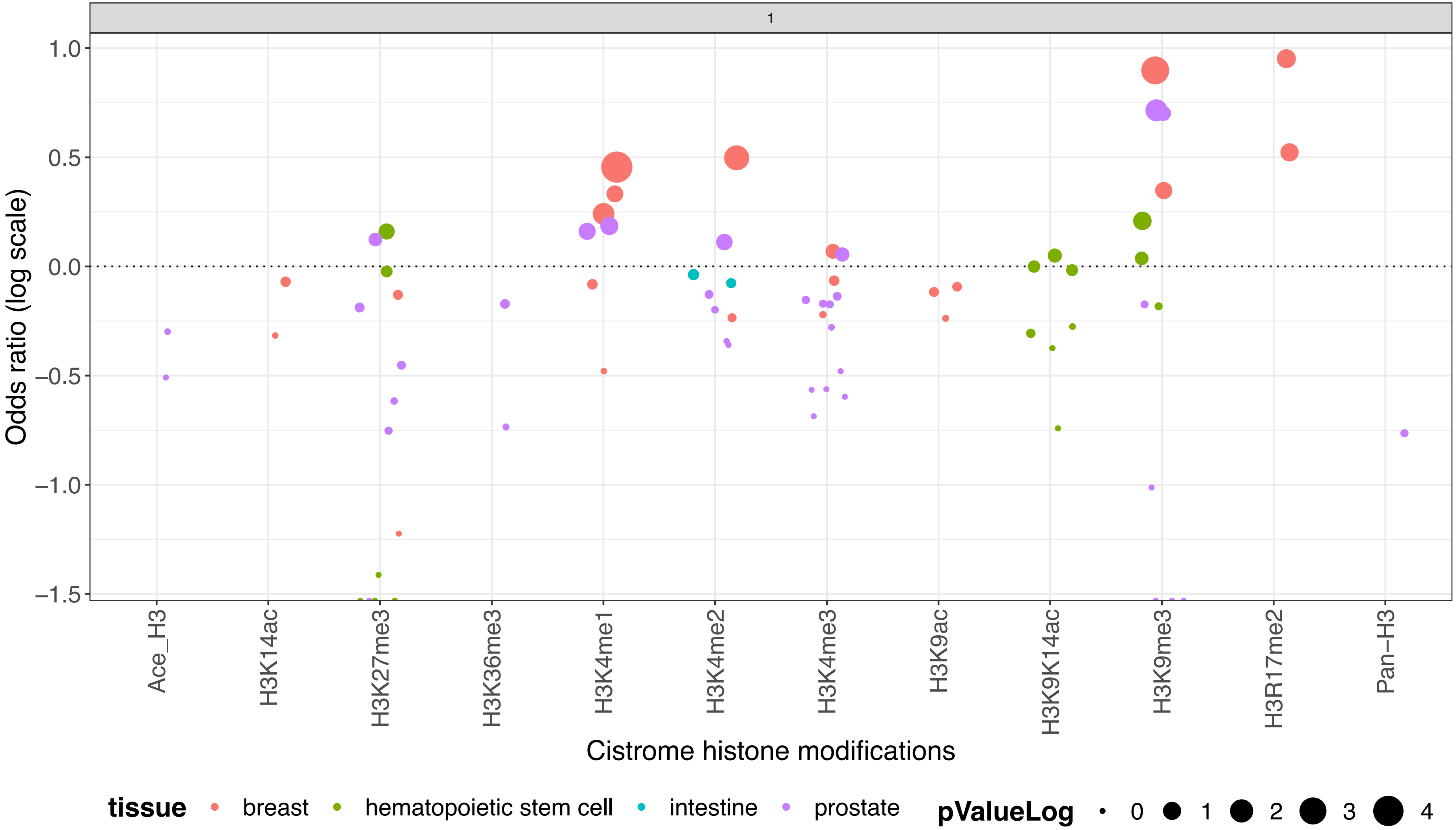


Figure 25: Bubble plot showing the enrichment for Cistrome histone modifications, for cis-correlating DNAm sites r>0.9, in ARIES at birth

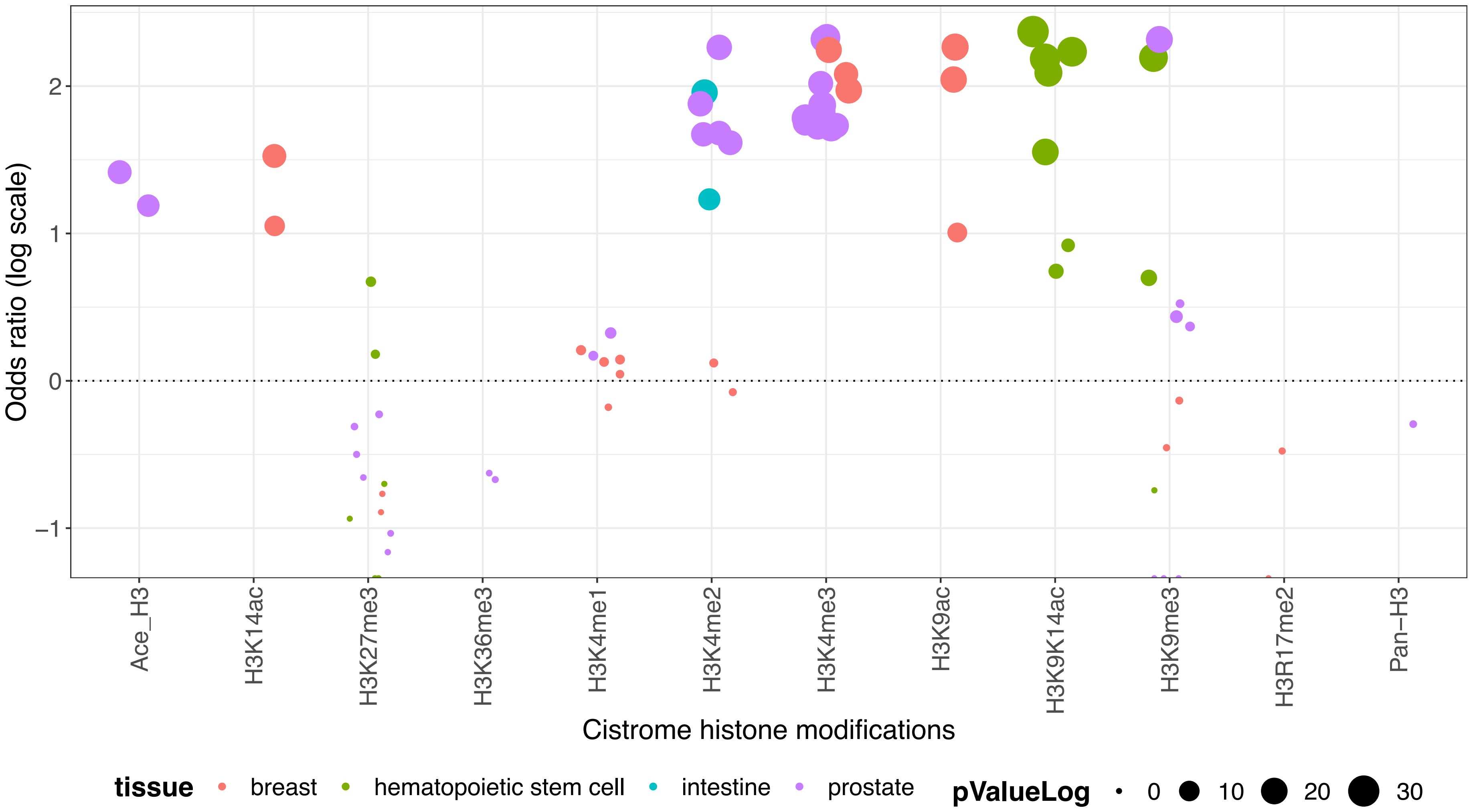


Figure 26: Bubble plot showing the enrichment for Cistrome histone modifications, for trans-correlating DNAm sites r>0.9, in ARIES 7 year olds

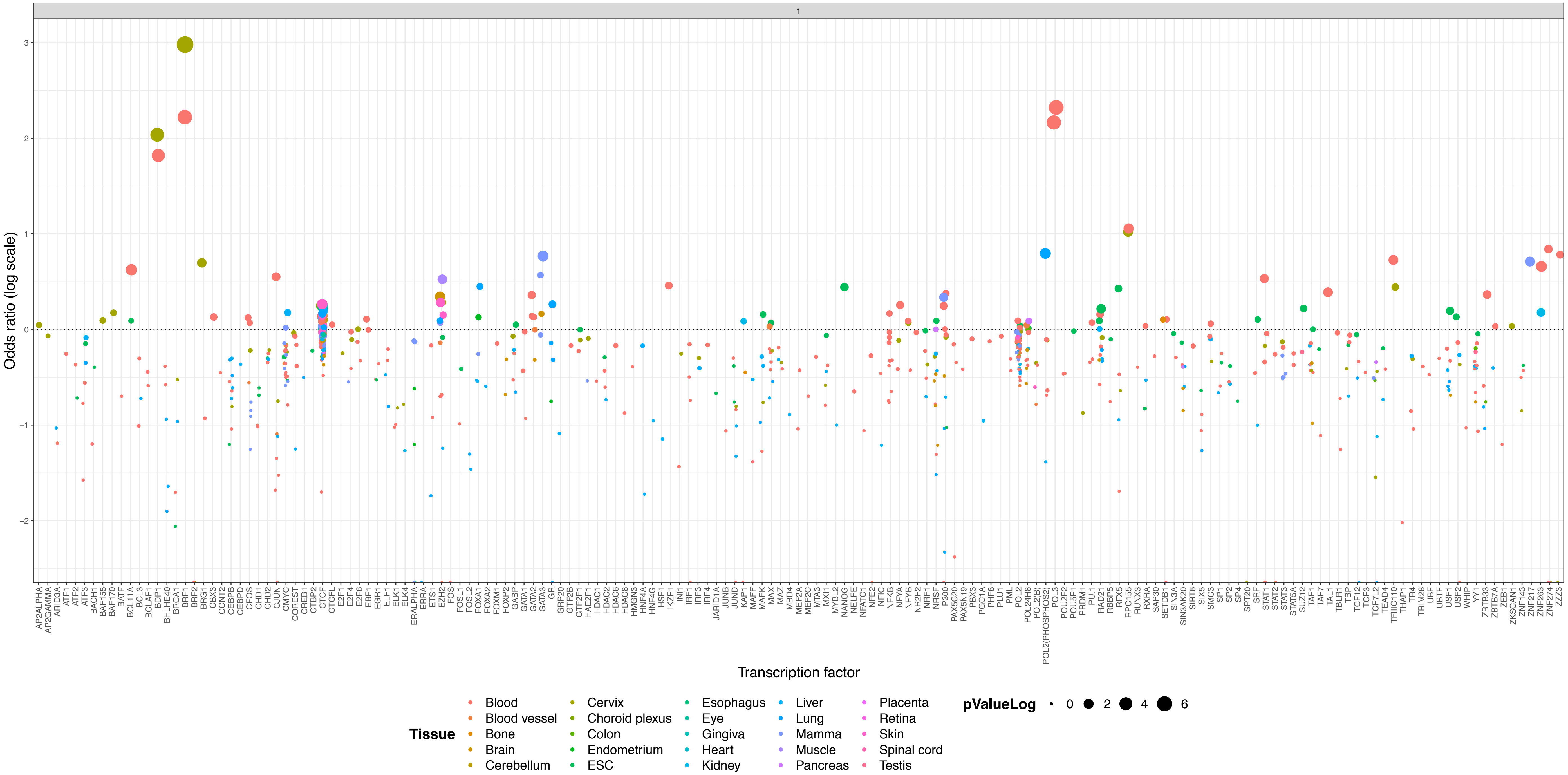


Figure 27: Bubble plot showing the enrichment for the ENCODE transcription factor binding sites, for cis-correlating DNAm sites r>0.9, in ARIES 7 year olds

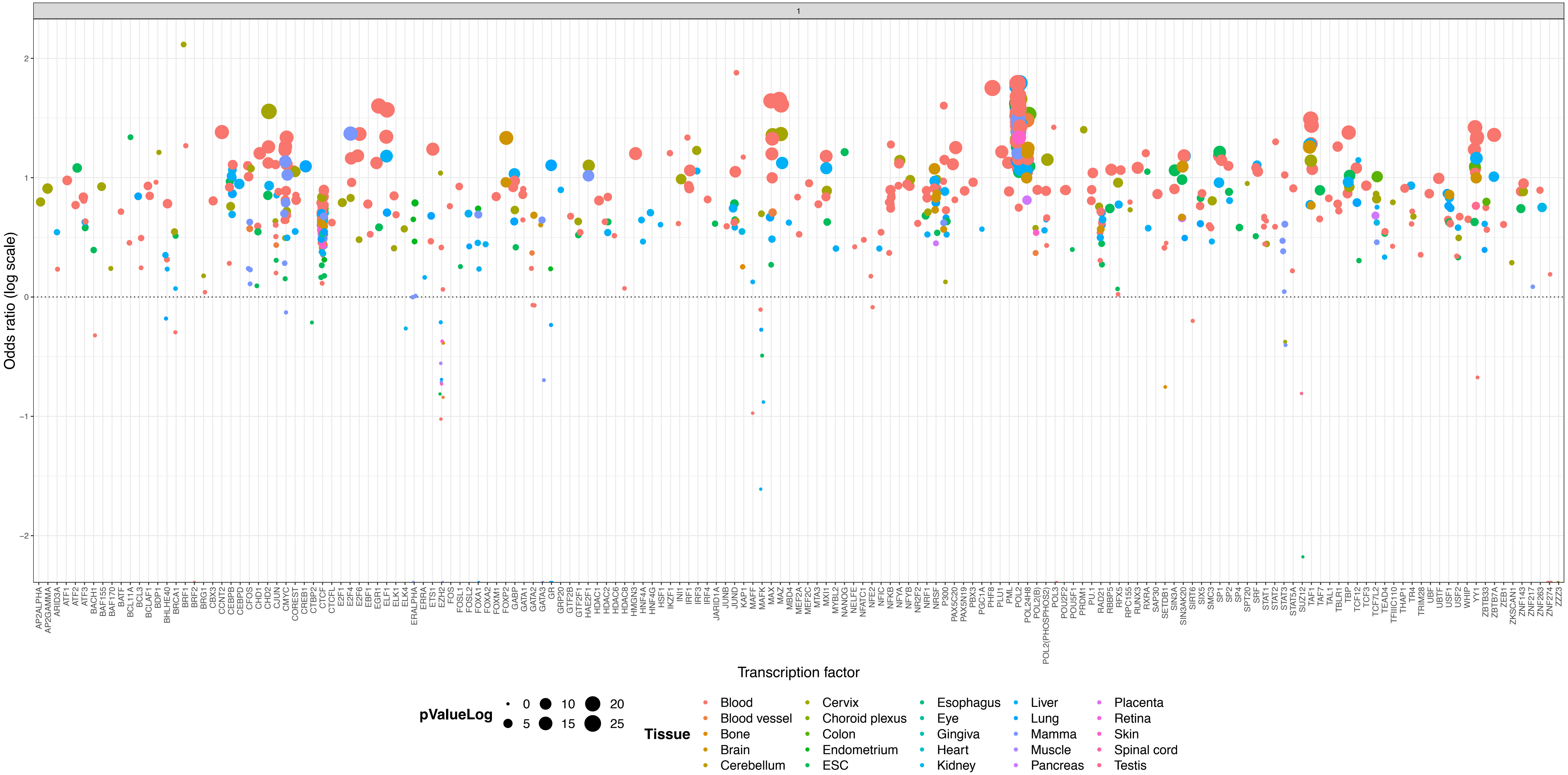


Figure 28: Bubble plot showing the enrichment for the ENCODE transcription factor binding sites, for trans-correlating DNAm sites r>0.9, in ARIES 7 year olds

3.3.4.2 Trans correlation illustration

To illustrate trans correlations, I plotted examples of pairs correlating >0.9 which were chromatin contact sites, and those which were not (for more results of the chromatin contact analysis, please see section 3.2.11.3 below). The chromatin contact sites (panels A and B) show clear and strong correlation in the unadjusted data, as do the sites that were not identified as chromatin contacts. The sites are hypomethylated and have low variance, and are illustrated in Figure 29.

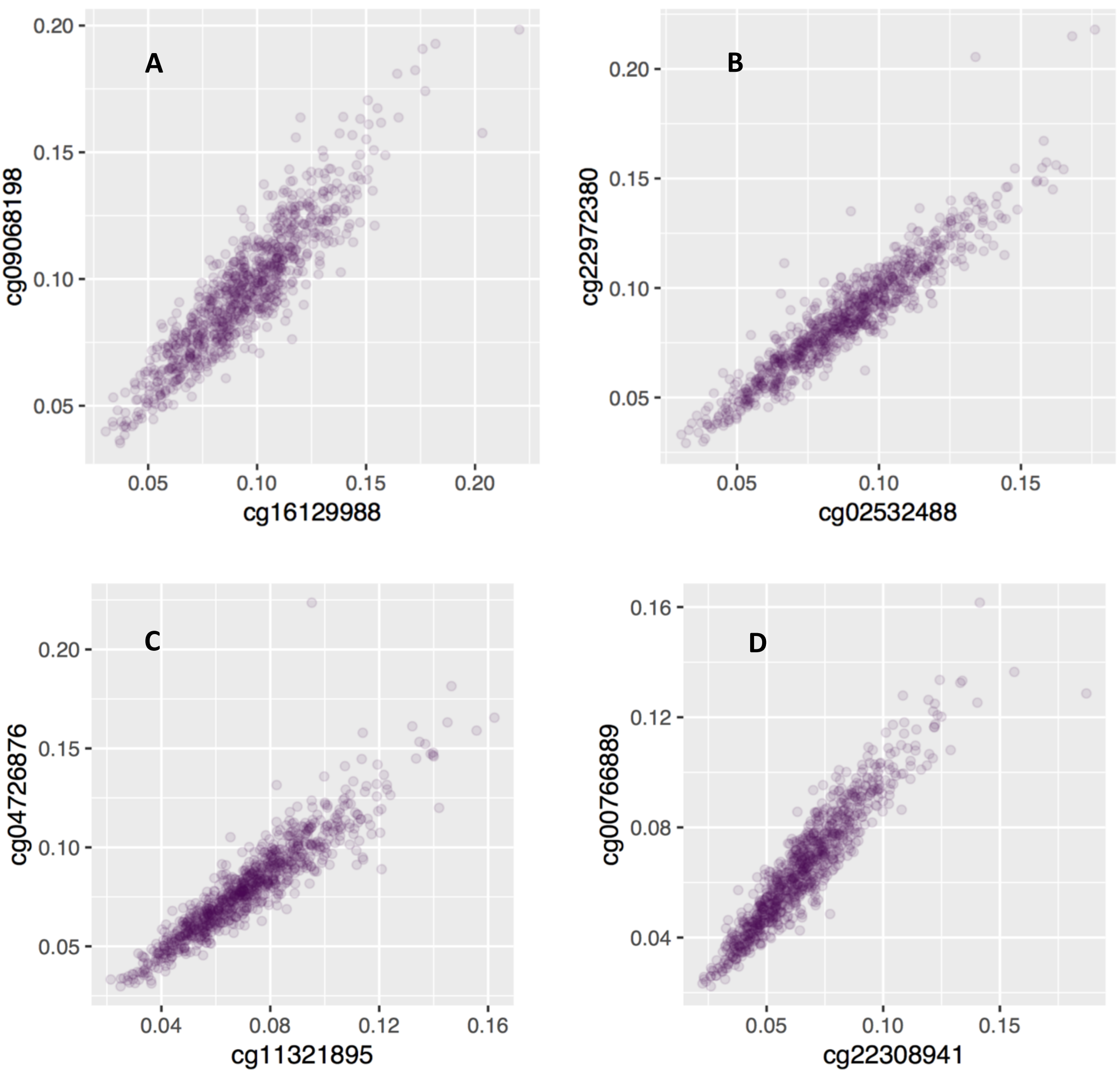


Figure 29: Scatter plot examples of the correlation between trans sites >0.9. Panels A and B are pairs of chromatin contact sites; panels C and D were not identified as chromatin contact sites.

3.3.4.2.1 Detection p-values

The histogram in Figure 30 shows that the detection p-values for DNAm sites which feature in *trans* correlations r>0.9 are well below even the stringent recommendations of (Heiss & Just, 2019). Trans correlations are therefore unlikely to be induced because of noise on the 450k array.

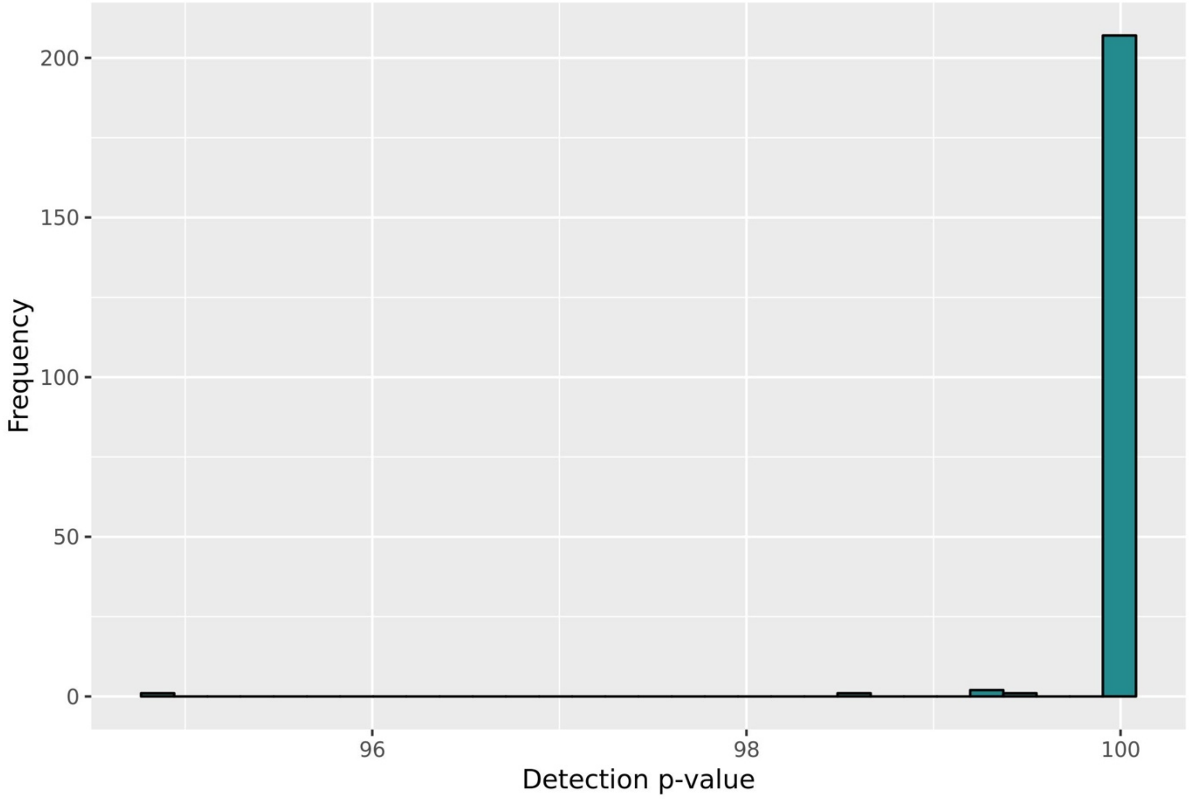


Figure 30: Histogram of -log10 transformed detection p-values for all DNAm sites which have trans correlations r>0.9

3.3.4.2.2 Circos plots

Trans correlations r>0.9 are distributed across the genome, as can be seen in the circos plots in Figure 31. The plots clearly show that there are some loci with many connections to multiple locations across the genome. This plot gives some indication that these sites are all interconnected; for a clearer assessment of this, please see the cytoscape plots in the following section.

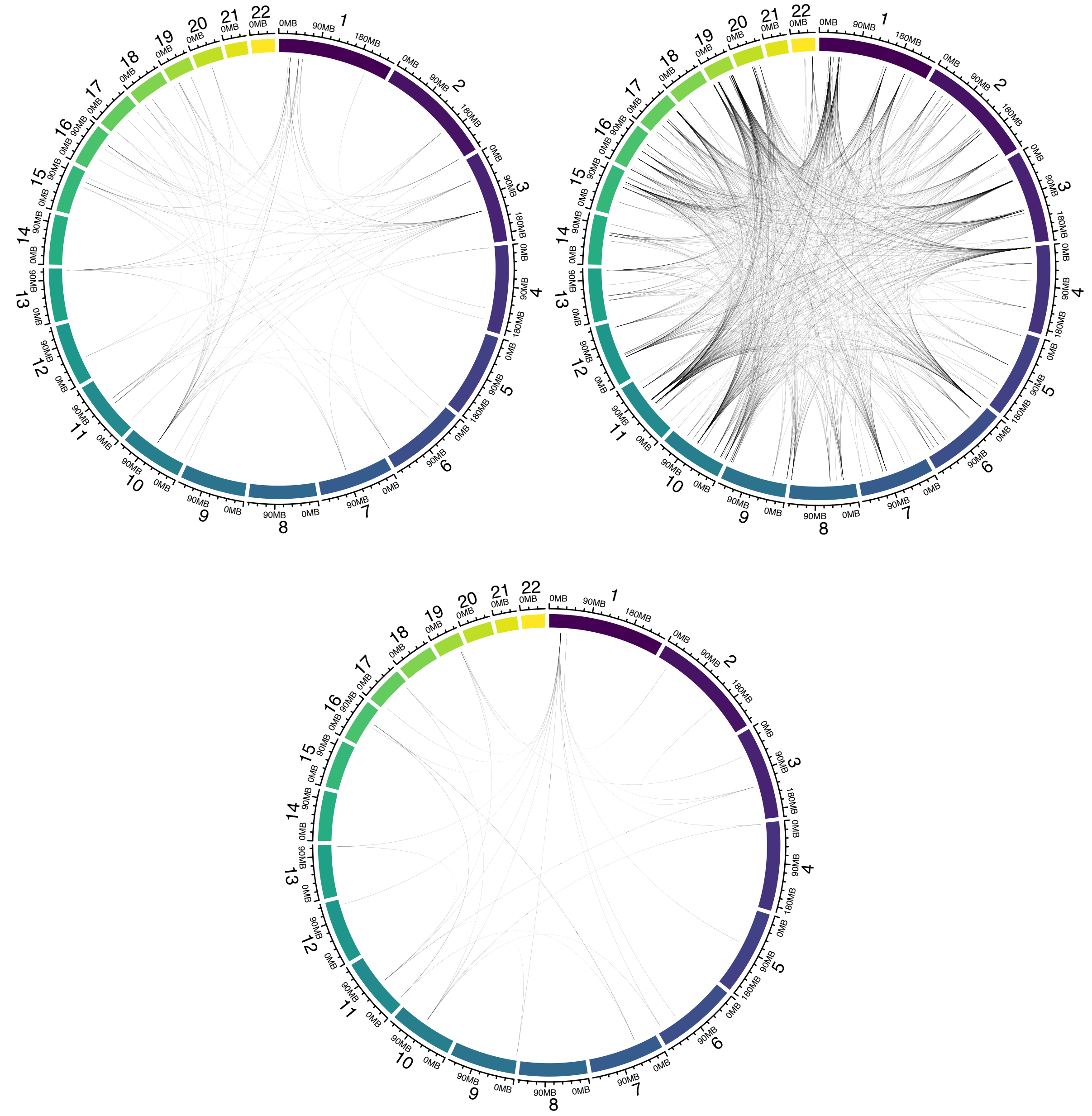


Figure 31: Circos plots visualising trans correlations r>0.9 in ARIES at birth (top left), 7 years (top right) and 15-17 years (bottom).

Across the three ARIES timepoints, 203 cis and 9 trans correlations are preserved above 0.9. As there were a maximum of 238 cis correlations that could be preserved, this represents the vast majority and suggests that cis correlations may be very stable. This is perhaps not surprising given that the cis correlations are highly heritable. 9 out of a possible 31 trans correlations are preserved across ages, 29%, suggesting that some trans correlations are stable. Future work might look into the characteristics of stable vs unstable trans correlations, as unstable one may have developmental relevance, whereas stable ones may form part of essential biological processes.

This can be further investigated using mean difference plots, which can illustrate the extent of differences between the ages without the hard threshold of 0.9. They show no evidence of a difference in trans correlations between birth and 7 years, or between 7 and 15-17 years; mean changes of 0.006 and 0.028, respectively, suggest there may be a slight drop in correlation with age, which increases in the adolescents. What is notable from these plots (found in Figure 32), however, is the strong changes in a small number of trans correlations between birth and 7 years; changes of this magnitude do not appear between 7 and 15-17 years, suggesting that some strong trans correlations are development-specific.

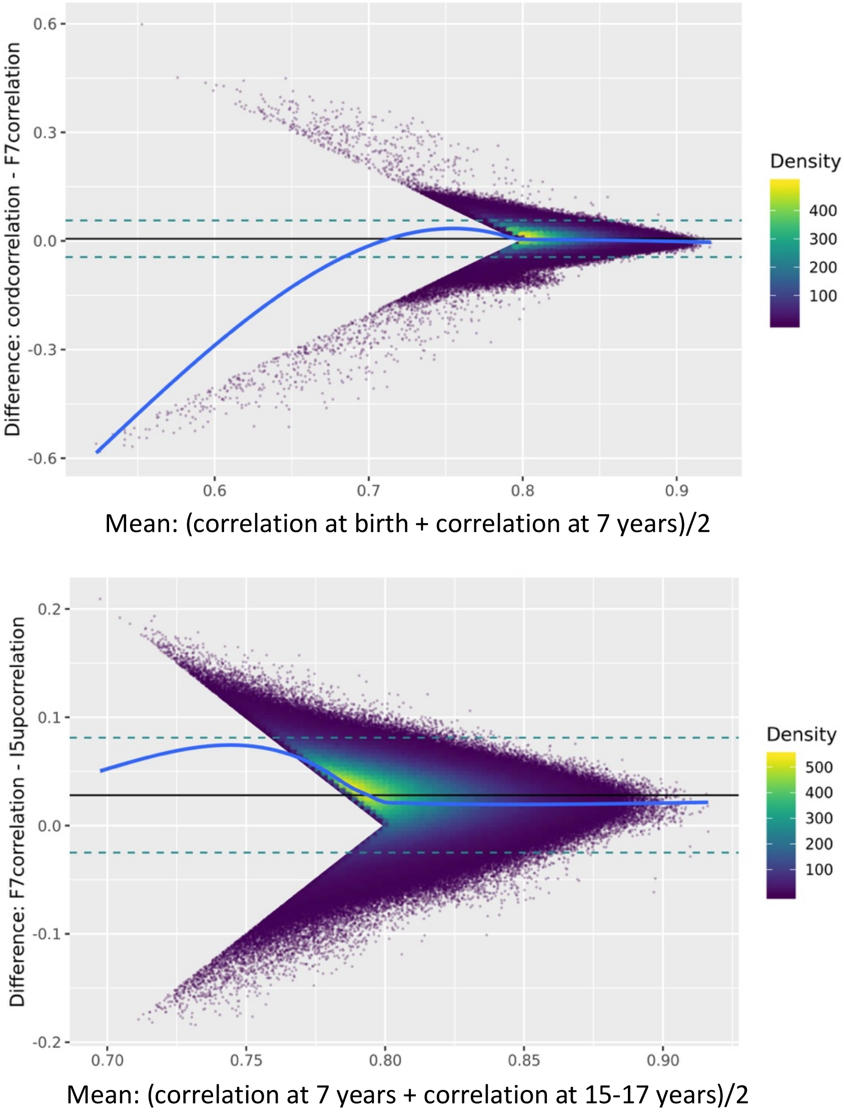


Figure 32: Trans correlation mean difference plots, showing the mean against the difference in correlation between birth and 7 years (top) and 7 years and 15-17 years (bottom), for trans correlations >0.8. the solid black line represents the mean difference in correlation; the dashed green lines are the 95% confidence intervals.

3.3.4.2.3 Cytoscape plots

The cytoscape network plots in Figure 33 illustrate the interconnections between the trans correlating sites r>0.9. Clearly the sites are strongly interconnected, with a small number of hub nodes with large numbers of connections. This is in line with the scale-free topology theory of biological networks.

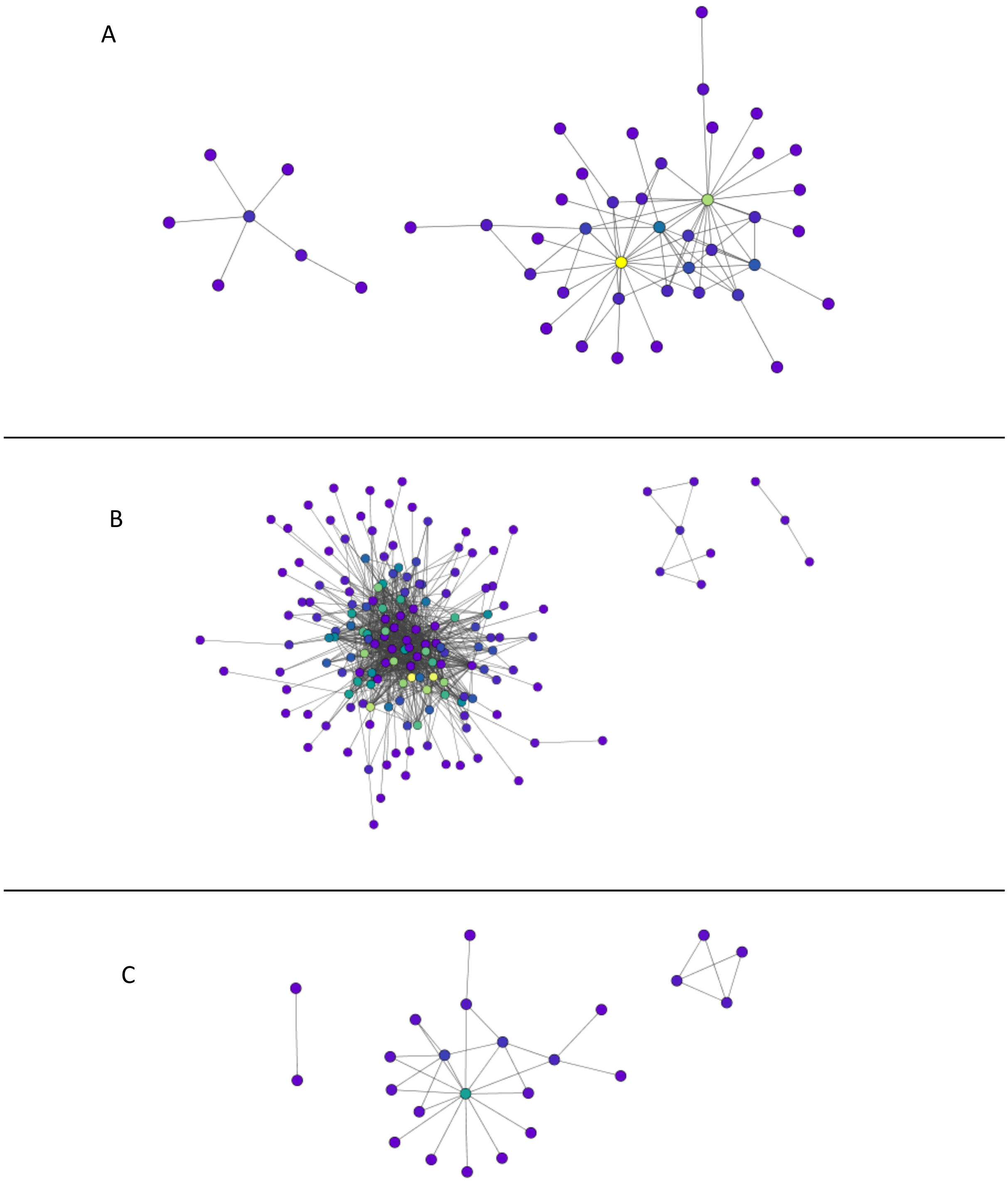


Figure 33: Cytoscape plots illustrating trans correlations r>0.9 in ARIES ***A*** *at birth* ***B*** *at 7 years* ***C*** *at 15-17 years*

3.3.4.3 Assessing trans correlations for chromatin contacts

To identify whether strong *trans* correlations are at sites where chromatin contacts are formed, I assessed the overlap of the *trans* correlations r>0.9 with the (Rao et al., 2014) Hi-C data. I did so with the 860 inter-chromosomal *trans* correlations in the 7 year olds. I found that there is a strong enrichment of Hi-C contacts in the real correlation data as compared to 1000 permutations of the data, with no permutation set having a higher count of overlaps than the real data (p=0). This is illustrated in Figure 34; the overlap in the permuted datasets is never more than 1, whereas there are 46 overlaps in the real data. Because this analysis was run in the 7 year olds, it may have been driven by greater cell type heterogeneity and so further work might clarify whether this can also be detected in the other age groups.

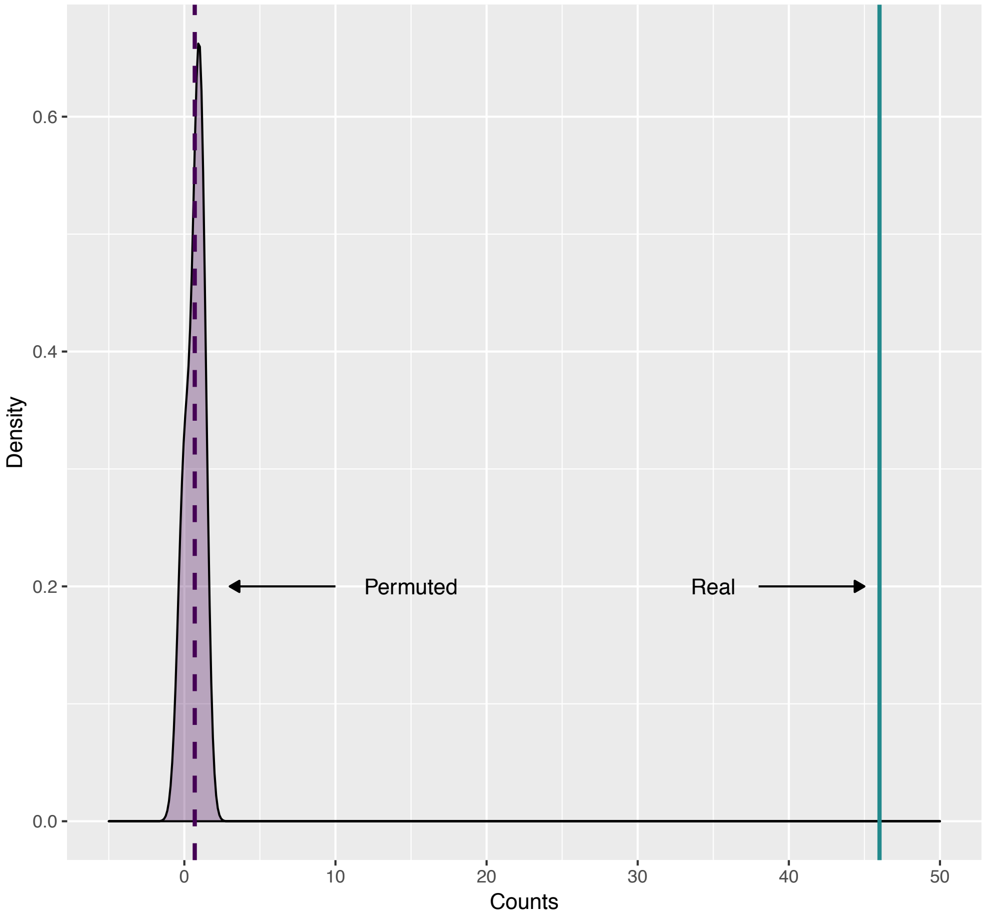


Figure 34: Plot illustrating the overlap counts in the real correlation data (green solid line, "Real") compared to the distribution of a permuted dataset (purple dashed line = mean permuted overlap, "permuted"; purple distribution = distribution of 1000 permuted overlaps).

Variance sensitivity analysis

Cis and trans correlating sites display very different patterns of variance; cis sites have more variable SD, whereas trans highly correlating sites have low variance. Plotting the mean and standard deviations of the correlating pairs illustrates that trans-correlating sites are almost all hypomethylated, and all have low variance. Cis sites are more variable in both mean methylation level and variance, although there is a tendency for hypo and hyper methylated sites to have smaller variance. As one might expect DNAm sites under trans-acting influences such as transcription factors to be hypomethylated (Domcke et al., 2015; Lienert et al., 2011), this would fit the profile of the highly trans-correlated sites here. This is illustrated by the density plot and the plot of mean and standard deviations of correlating DNAm site pairs in Figure 35. Repeating this analysis for the chromosomal contact sites, we see clear hypomethylation and small variance, in Figure 36. The plots of the individual correlating pairs show clear correlations between normalised methylation betas.

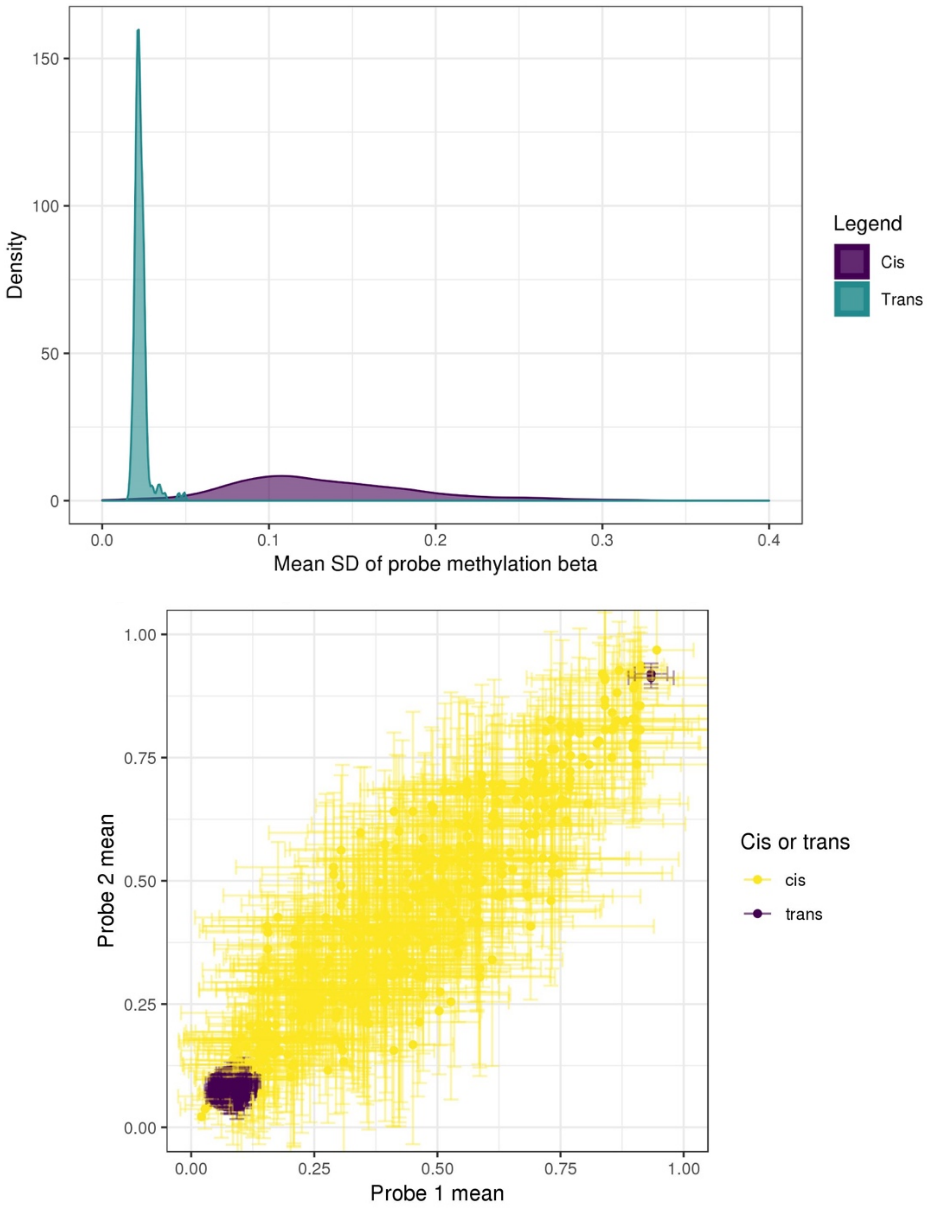
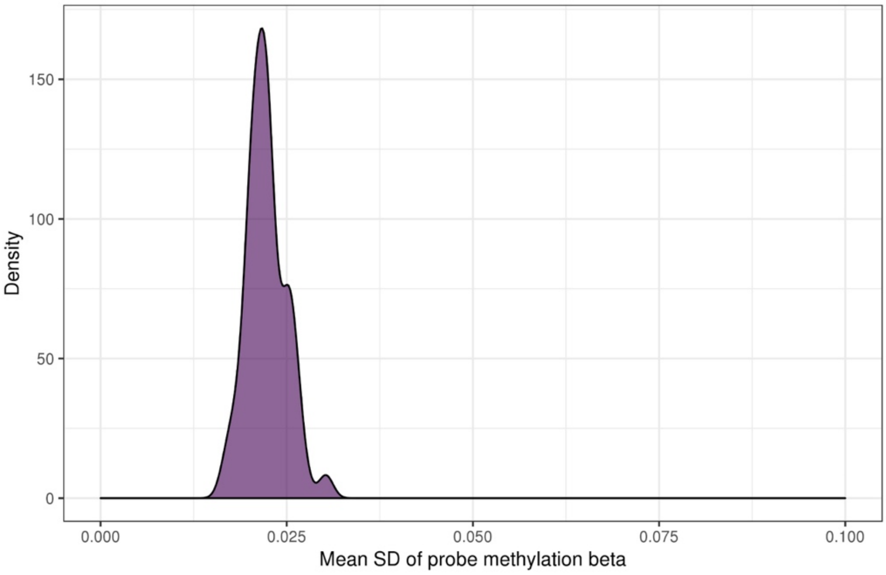


Figure 35: Top: density plot of the mean variance for all DNAm sites correlating >0.9 in ARIES 7 year olds. Bottom: scatter plot of mean values of correlating DNAm sites, with vertical errorbars representing the SD of probe 1 and horizontal errorbars representing SD of probe 2.



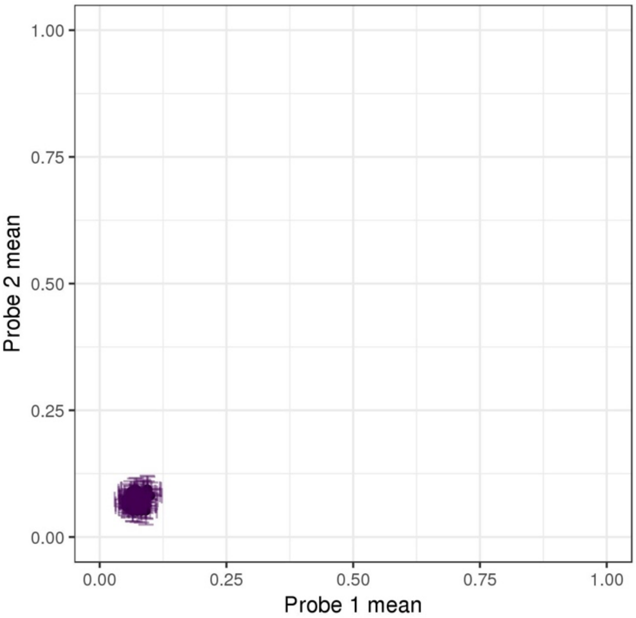


Figure 36: Top: density plot of the mean variance for all DNAm sites correlating >0.9 in ARIES 7 year olds that were identified as chromatin contact regions. Bottom: scatter plot of mean values of chromatin contact DNAm sites, with vertical errorbars representing the SD of probe 1 and horizontal errorbars representing SD of probe 2.

3.4 Discussion

3.4.1 Summary of findings

The work in this chapter has extended understanding of correlation structure across the genome. I have shown that gross *cis* correlation structure is stable from birth to adolescence, emphasising the need to account for this in single site and regional analyses. I show that *cis* correlation structure is strongly influenced by *cis* genetic variants, and strongly correlated *cis* sites are associated with functional enrichments related to transcription. I have provided the first genome-wide (across the 450k) illustration of *trans* correlations between DNAm sites, confirming that these sites are likely to represent biological functions that DNAm is involved in. I illustrate the interconnected nature of the trans correlations, and finally I show for the first time in humans that *trans*-correlating DNAm sites may, at least in part, be involved in inter-chromosomal chromatin contacts.

It is notable how many more correlations >0.9 there are in the 7 year old DNAm data compared to the other two timepoints; it is possible that the greater number of correlations at age 7 are due to the greater proportion of whole blood samples at that timepoint. Whole blood contains a greater number of cell types than the other sample types used in ARIES, which is likely to lead to a greater amount of variation in the data; and greater variation will make it more likely that stronger correlations will be detected because it creates greater spread in the data.

3.4.2 Cis correlation structure

I have illustrated that *cis* correlation structure is consistent across the genome, having the same rate of decay across all autosomal chromosomes. This rate of decay mirrors that found in previous work in adults (Eckhardt et al., 2006; Y. Liu et al., 2014; Saffari et al., 2018), and it also matches the decay pattern illustrated by Zhang et al 2015, as there is a steep drop in correlation up until about 400bp. This, combined with the persistence from birth to adolescence, suggests that this gross correlation structure is persistent throughout life. To my knowledge this is the first demonstration of longitudinal correlation structure. Importantly I have also illustrated that there is a consistent and fairly wide margin of error around the distance-based decay of *cis* correlations, showing that correlation between DNAm sites is quite heterogeneous within 1-2kb. Some interesting further work might examine whether correlation decay over distance is related to genomic regions (such as CpG islands) or units such as promotors and enhancers. This is likely to show an impact, as (W. Zhang et al., 2015) found that genomic location influenced cis correlations between DNAm sites.

As I show that *cis* correlation structure is notably reduced by adjusting for the strongest *cis* mQTL, part of the stability in *cis* correlation decay may be likely to come from the control exerted by associated genetic variants. If further work were to adjust DNAm data for all associated mQTLs, we might well see a further reduction in cis correlation structure. There is most impact of adjusting for *cis* mQTLs on the highest correlations, which fits with the finding that higher *cis* correlations tend to have both DNAm sites under the influence of identified mQTLs. I have illustrated that the connectivity of DNAm sites is not related to the LD score of the strongest associated mQTL, which adds more weight to the argument that *cis* correlation structure is not related to LD. Based on this, and the work by (Y. Liu et al., 2014) and (Saffari et al., 2018), I would not be inclined to investigate the connection between *cis* correlation structure and LD further.

3.4.3 Negative DNAm correlation structure

The separation of negative and positive correlations is, to my knowledge, the first demonstration of negative co-methylation. As the strong negative correlations (-1 to -0.7) almost exclusively have both DNAm sites associated with mQTLs, I would expect that strong negative correlations are likely to be under genetic control. As negative correlations between DNAm sites have not really been dealt with in the literature it is hard to say what their function might be, although one might imagine they could represent inhibitory biological pathways. Inhibitory pathways could represent situations where DNAm has a role in inhibitory pathways; so decreased methylation in a promotor region allows the expression of a gene, which then inhibits the expression of another gene with a concomitant increase of methylation of the second gene’s promotor.

However negative correlations between DNAm sites could also arise from the differing actions of DNAm in different genomic contexts. It is possible that negative correlations represent DNAm sites in CGIs and promotors acting toward the same outcome (because they have opposing effects on gene expression). The relationship is not likely to be quite as simple as this, because DNAm array data represents proportions of sites methylated rather than counts (as in gene expression data). Although I did not examine these negative correlations in detail in this thesis, this is a clear direction for future work.

3.4.4 Heritability

(Hannon et al., 2018) found that the DNAm sites most influenced by heritability were often associated with environmental exposures. Taken with the results at 7 and 15-17, this suggests that these highly heritable sites also tend to be highly correlated with other DNAm sites, meaning they may also be representing shared regulation. Further work could take these sites and ascertain whether they associate with traits such as those identified in (Hannon et al., 2018); or whether they are enriched for associations with traits in EWAS studies. One potential caveat to bear in mind is that the Hannon dataset was generated in individuals who were 18 years old, and so it is possible the estimates would be different in a dataset generated from individuals at birth. Another consideration to bear in mind is that the birth timepoint in ARIES had the sample type (blood spots or white cells) confounded by slide. It is quite possible that this impacted DNAm values and the relationships between them; although it is not clear to what extent this might be the case.

The mQTL analysis shows that the influence of genetic variants on correlations between DNAm sites may change over time. Specifically, the changes I identified were to the stronger correlations (both positive and negative), where they were more likley to have both DNAm sites associated with an mQTL in the adolescents. This may reflect that over time correlations that are influenced by genetic variants are relatively static, and those that are not are more prone to epigenetic drift, which would be very likely to reduce correlations between DNAm sites.

3.4.5 Genomic enrichment in cis correlating sites

I found a number of interesting enrichments for strong *cis* correlations. I demonstrate that they are consistently enriched for genomic locations associated with poised promotor chromatin states. These chromatin states represent loci that are poised for gene expression, and these often have key roles in development, particularly well studied in germ cells (Choate & Danko, 2016; Lesch, Silber, McCarrey, & Page, 2016; Mikkelsen et al., 2007), suggesting that strong *cis* correlations may have roles in transcription activation and repression. This also suggests correlated *cis* DNAm sites may have regional relevance, in contrast to the conclusions of (Y. Liu et al., 2014). My findings may be different from theirs as I used LOLA to assess regional enrichment, which accounts for some key biases in enrichment analyses, especially for DNAm data, as it does not have to assign the DNAm site to a gene or to a pre-defined genomic feature. It may be because I looked at all of the sites on the 450k, instead of those with high variance; or it may be because I used a larger sample.

Cis correlations were enriched for transcription factor binding sites for RNA Polymerase III (Pol3), and two subunits of the transcription factor TFIIIB, which is required for Pol3-

mediated transcription (Abascal-Palacios et al., 2018). This was consistent across all ARIES timepoints, which suggests that very high *cis* correlations may have a role in the transcription of non-coding RNAs (which Pol3 transcribes) (Turowski & Tollervey, 2016).

3.4.6 Limitations of enrichment analyses

There are a number of caveats in the regional overlap enrichment analysis that we should consider alongside these results. In these analyses it is assumed that the test set is equivalent to the background set; although I matched on CG content, enrichments could be confounded by genomic features we don't yet understand. Regional overlap with LOLA doesn't only count genes once; however using regions and deduplicating them in my analyses ought to limit the effects of this. LOLA also doesn't take correlations between the DNAm sites being tested into account; the extent of strong correlations between DNAm sites I have demonstrated illustrates that this may well be an issue. To rectify this, future methods assessing the overlap of DNAm sites with genomic features might use the correlation structure I have described to adjust for this issue to avoid inflating the enrichment analysis.

The absence of the null distribution

The null distribution is difficult to simulate and was beyond the scope of this thesis. Under a null distribution, one might not expect to see strong functional enrichments, or the preservation of these enrichments across datasets. A simpler way to test whether the enrichments found here are due to chance, the correlation matrix could be permuted and each of those permutations could be assessed for functional annotations. This would test, for example, whether enrichment for poised promotors in cis-correlating sites is simply due to chance or overrepresentation that was not accounted for in the analysis.

Variance sensitivity analysis

The variance sensitivity analysis has shown that cis and trans correlating sites have very different profiles in terms of mean and variance of methylation betas. Cis correlating sites have variable mean and standard deviation of the beta value, with some of the variable correlation being quite clearly based on genotype (as can be seen in Figure 14 earlier in the chapter, showing examples of cis correlations). Almost all trans sites are hypomethylated with low variance. The trans-correlating sites in chromatin contact regions are exclusively hypomethylated with low variance. The biological enrichments of the trans-correlating sites, coupled with observations that trans-influences on DNAm sites are often linked with hypomethylation, are suggestive that these are meaningful associations; however one cannot conclusively discount that the results are due to technical noise. Further work might address this through the use of a different technology (such as bisulfite sequencing) to confirm whether these are true associations. This is particularly important as it would provide more conclusive guidance as to whether low-variance DNAm sites should be included in such studies, as my work suggests they may be biologically meaningful.

3.4.7 Trans correlation structure

To my knowledge this is the first illustration of *trans* correlations across the whole genome (as measured by the 450k). I have shown that strong *trans* correlations are unlikely to be artefacts of poor probe detection. I show that they are less likely to be associated with mQTLs, raising the possibility than *trans* correlations between DNAm sites are less driven by genetic variants than *cis* correlations. Of course it could be that they are associated with genetic variants, and these associations have not yet been discovered, so I cannot definitively say that most strongly trans-correlating DNAm sites are not associated with mQTLs. But using the GoDMC data, which is the largest study identifying mQTLs to date, gave me the best existing resource to assess this. Further work could move to quantify the relationship between trans correlating DNAm site sand genetic variants, by regressing out the strongest mQTL variant; however this would have to encompass a broader range of *trans* correlations than I did in this thesis so that sufficient trans-correlating pairs with associated mQTLs could be assessed.

I have shown that *trans* correlations are enriched for chromatin states associated with active transcription start sites (which are associated with the expression of genes (Roadmap Epigenomics et al., 2015)) and promoters downstream of transcription start sites. These enrichments suggest that *trans*-correlating DNAm sites are likely to be functional, and they are likely to have regulatory roles in, or be markers of, gene expression. This is further supported by the clear overlap enrichment of Pol2 (RNA polymerase II) TFBS in strong *trans* correlations in all timepoints in ARIES, as Pol2 transcribes protein-coding genes (Sainsbury, Bernecky, & Cramer, 2015). The enrichment of many other transcription factors suggests that these trans-correlating DNAm sites may have a variety of roles in genomic function.

This supports the work of (G. Li et al., 2019), who found that trans correlating DNAm sites that were found at chromatin contact sites were enriched for other factors in addition to CTCF.

The enrichment of CTCF TFBS in the ARIES 7 year olds suggests that some of these trans-correlating sites form chromosomal loops. The lack of enrichment at the other timepoints may suggest changing *trans* correlation functions, or it may simply be because of the smaller number of *trans* correlations relative to the 7 year olds. It would be interesting to look further into this relationship and dissect whether it is the intra-chromosomal trans correlations that are enriched for CTCF binding sites, which would likely illustrate formation of chromosomal loops. In such an analysis I would likely take a slightly larger range of correlations so that this theory could be tested in all the age groups.

3.4.8 Enrichment for inter-chromosomal chromatin contacts

The overlap of the Hi-C contacts suggests that the purpose of these highly correlated inter-chromosomal DNAm sites is coordinated activity related to the trans-chromosomal contacts. This extends the work of (G. Li et al., 2019), as they demonstrated coordinated DNAm at chromatin contacts in mouse cells, and I have shown that this is also evident in a large human cohort. I have shown that probable DNAm coordination between chromatin contact regions can be detected in array data as well as bisulfite sequencing data, which raises the possibility of using these areas of *trans*-chromosomal contact to investigate the possibility of the alteration of these contacts in datasets with rich phenotype data. So for example, is a particular exposure associated with altered inter-chromosomal chromatin contacts? Something important to bear in mind here is that because trans-chromosomal promoter contacts have been shown to be cell-type specific (Javierre et al., 2016), future work may benefit from utilisation of less heterogeneous cell type populations than blood to investigate the functions of trans-chromosomal co-methylation.

3.4.9 Conclusions

This Chapter has further delineated the correlation structure between DNAm sites, and illustrated that strongly correlated sites are enriched for functional annotations that suggest key roles in genome function. The stability from birth to adolescence is an important discovery, as this will aid development of methods to adjust single site and regional analyses for correlation between DNAm sites.

4 Chapter Four

Systems biology network analysis

4.1 Introduction

Activity of DNAm across the genome has been extensively described in the developing embryo and across pre-natal development (Feng et al., 2010; Smallwood & Kelsey, 2012; Spiers et al., 2015). Widespread changes continue after birth; in brain most DNAm sites on the 450k are differentially methylated between pre- and post-natal life (Jaffe et al., 2016). These changes continue throughout childhood, with dynamic changes reported at DNAm sites across the epigenome (Alisch et al., 2012; D. Martino et al., 2013; D. J. Martino et al., 2011; Xu et al., 2017). There is also extensive research describing differential DNAm in the post-natal period with respect to many exposures and phenotypes related to pregnancy and childhood, particularly through the PACE consortium (Felix et al., 2018). However, as yet there has been little published on the co-methylation of DNAm sites as part of normal post-natal development – which is important as co-methylation analyses have the potential to detect regulatory networks of DNAm sites that may be acting together.

4.1.1 Network analysis

As described in chapter 1, section 1.6, co-methylation network analysis is a systems biology approach which uses relationships between biological features to identify potential biological pathways. I selected WGCNA as the method to use in this chapter because in addition to extensive use for gene expression data, it has been successfully applied to DNAm data for almost a decade (Bocklandt et al., 2011; Horvath et al., 2012; van Eijk et al., 2012). There have been numerous WGCNA studies identifying DNAm pathways that differ between cases and controls, in many different diseases (Busch et al., 2016; de Jong et al., 2012; Horvath et al., 2016; Langfelder et al., 2016; Nicodemus-Johnson et al., 2016) and environmental factors (Houtepen et al., 2018; Maertens, Tran, Kleensang, & Hartung, 2018).

Some studies have applied WGCNA to a single 'normal' population to identify DNAm regulatory functions. (Horvath et al., 2012) used 16 datasets to identify an age-related DNAm pathway that was preserved across both blood and brain (although the module was much better preserved in brain samples). (Spiers et al., 2015) found that WGCNA could identify modules of DNAm related to specific neurodevelopmental functions in fetal brain tissue. These studies illustrate that WGCNA can be used to identify biologically relevant groups of highly co-methylated DNAm sites in a population context.

An additional advantage of network analyses is that they present an opportunity for feature reduction, because they group features into highly related modules. This is advantageous for DNAm data, because there is a high multiple testing burden when using data from the 450k. Reducing the numbers of features to test will increase the power to detect relationships between DNAm and traits of interest, and includes DNAm sites which may not pass genome-wide significance thresholds, but may still contribute to the relationship with the trait.

4.1.1.1 Chapter motivation

In chapter 3 I demonstrated that highly correlating *cis* and *trans* DNAm sites are biologically relevant, as they are associated with chromatin states and transcription factor binding sites suggesting they have roles in transcriptional regulation. In this chapter I extend the analysis to all DNAm sites on the 450k, by using a network analysis which can distil the vast number of correlations into a small number of modules of highly related sites. This functions as a powerful dimension reduction technique, and enables me to assess whether the groups of highly co-methylated sites might be related to traits and exposures related to development, as well as the preservation of these modules over time. I also assess whether the modules are enriched for gene ontologies or pathways that might be relevant to normal human development.

4.2 Hypotheses

H.4.1. I hypothesise that DNAm is involved in regulatory pathways that contribute to normal development, and that network analysis can yield novel information about these pathways.

H.4.2. I hypothesise that DNAm network analysis will reveal whether exposures relevant to development might affect DNAm as part of a pathway, thus increasing our knowledge of how exposures affect DNAm beyond single-site analyses.

H.4.3. I hypothesise that pathways related to DNAm will change over childhood development, from birth to adolescence.

4.3 Aims

A.4.1. To construct putative DNAm regulatory networks for ARIES children.

A.4.2. To identify whether network modules are associated with developmental traits and environmental exposures.

A.4.3. To evaluate the extent to which these modules are influenced by blood cell types.

A.4.4. To identify whether there are network modules that are associated with developmental functions/pathways.

A.4.5. To identify whether there are any consistent network modules over time.

4.4 Methods

4.4.1 Data

4.4.1.1 DNAm data

The datasets used in this chapter were the 3 ARIES young person DNAm datasets. ARIES is a subsample of ALSPAC, where 1,000 mothers and their children were selected to have DNA methylation profiled. This was at birth (cord blood), 7 years and 15-17 years. ARIES data is described in detail in chapter 2 section 2.2.1.1. DNAm was measured in blood (either whole blood, white cells, PBS or blood spots), using the Illumina 450k array.

4.4.1.1.1 Adjustments

I filtered sites on the 450k array, as detailed in chapter 2, section 2.2.1.7). Briefly, all non-autosomal probes, probes in the HLA region, poorly functioning probes and probes containing SNPs were removed from the dataset. This left 394,842 probes for analysis. As all timepoints were normalised together, the same probes were included for all timepoints. All outlying values (> 10 standard deviations from the mean) were replaced with the mean for that probe, over three iterations.

As discussed in chapter 2, all datasets had the appropriate combination of age, sex, and sample type, as well as estimated cell counts, regressed out using a linear model. Slide was also regressed out, to account for batch effects, using a linear model. A random effects model was not used because batch effects were removed as a random effect during normalisation; slide was regressed out after normalisation because the normalisation did not completely remove associations with slide. All individuals with methylation data were included, unless they were the only individual on a slide (as detailed in chapter 2, section 2.2.1.3).

4.4.1.1.2 Estimating cell counts

Blood cell type proportions were estimated and adjusted for as detailed in chapter 2 section 2.2.1.6. For investigation of cell count proportions and their relation to asthma modules, I estimated cell counts from the normalised betas using the 'blood gse35069 complete' reference, which provides estimates eosinophils and neutrophils separately.

4.4.1.2 Phenotype data

Phenotype data that were assessed for relationships with WGCNA modules are described in chapter 2, section 2.2.1.9 and summarised in Table 4. Covariates in the analyses are also detailed in chapter 2, section 2.2.1.9.

4.4.2 WGCNA

WGCNA analyses were run using the protocol described in chapter 2, section 2.6. The networks were created using the same specification for the three ARIES timepoints; the only parameter that varied was the soft threshold power, which was calculated for each timepoint separately. The powers I used can be found in results section 4.5.2.

4.4.2.1 Removing outlying samples

Outlying samples may skew the clustering of DNAm sites, and therefore affect the modules it can detect. As a result it is recommended to cluster samples based on DNAm values, using a hierarchical cluster dendrogram. This method is described in chapter 2, section 2.6.1, and was done for each of the ARIES timepoints separately.

4.4.2.2 Calculating soft threshold power

As WGCNA raises the correlations between DNAm sites to a power to de-emphasise the weak correlations (known as soft thresholding because it doesn't impose an exclusionary threshold), a power must be selected. The selection is based on the power for which the network would reach scale-free topology, a network theory that assumes non-random connections and highly connected hub nodes (Barabasi, 2009; Barabasi & Albert, 1999). The method for calculating this power is in chapter 2, section 2.6.2.

4.4.2.3 Blockwise network construction

The WGCNA networks were constructed using the blockwise method because I used all ~400,000 DNAm sites in the network, and it is not computationally possible to construct a single-block WGCNA network of that size. This was done separately for all ARIES timepoints, and the method is described in chapter 2, section 2.6.3.

4.4.2.4 Association of WGCNA modules with phenotypes

To identify whether WGCNA modules are associated with traits of interest, they can be associated with the module eigengenes (which are essentially the first principal components of each module). The details of the analysis can be found in chapter 2 section 2.6.4; the details of the phenotypes and exposures can be found in chapter 2, section 2.2.1.9 (with a summary in Table 4)

4.4.3 Gene ontology and KEGG pathway analysis

To assess whether the WGCNA modules were associated with identifiable biological functions, I ran the gene ontology and KEGG pathway enrichment analyses as detailed in chapter 2 section 2.7. This was run for each ARIES timepoint separately.

4.4.3.1 Preservation of network modules

To assess how well the network modules were preserved between birth and adolescence, I ran the network preservation analysis described in chapter 2, section 2.6.4.1 between birth and 7 years and between 7 years and adolescence.

4.5 Results

4.5.1 Hierarchical clustering

The number of samples removed after being identified as outliers using hierarchical clustering are detailed in Table 10. The cluster trees with the cut heights are shown in Figure 37, Figure 38, and Figure 39. The hierarchical cluster trees were cut based on visual inspection, to remove the smallest number of individuals possible whilst ensuring the data would reach scale-free topology at a reasonable power. This was trialled for each timepoint (data not presented).

Timepoint	Number of participants before clustering	Number of participants after clustering
Birth	849	820
7 years	910	859
15-17 years	921	887

Table 10: Number of participants removed during hierarchical clustering of DNAm samples

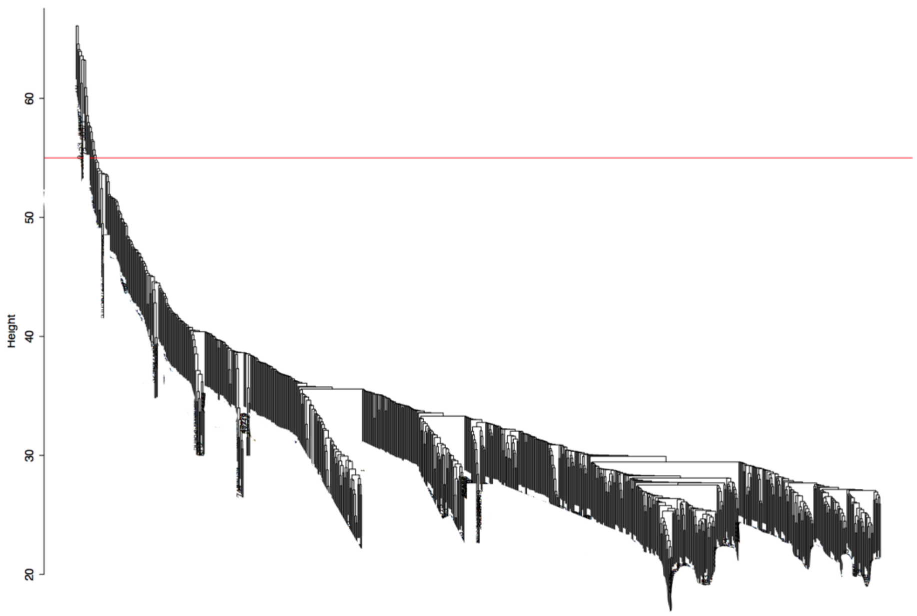


Figure 37: Hierarchical clustering of samples, in ARIES at birth. Red line indicates height at which samples were cut.

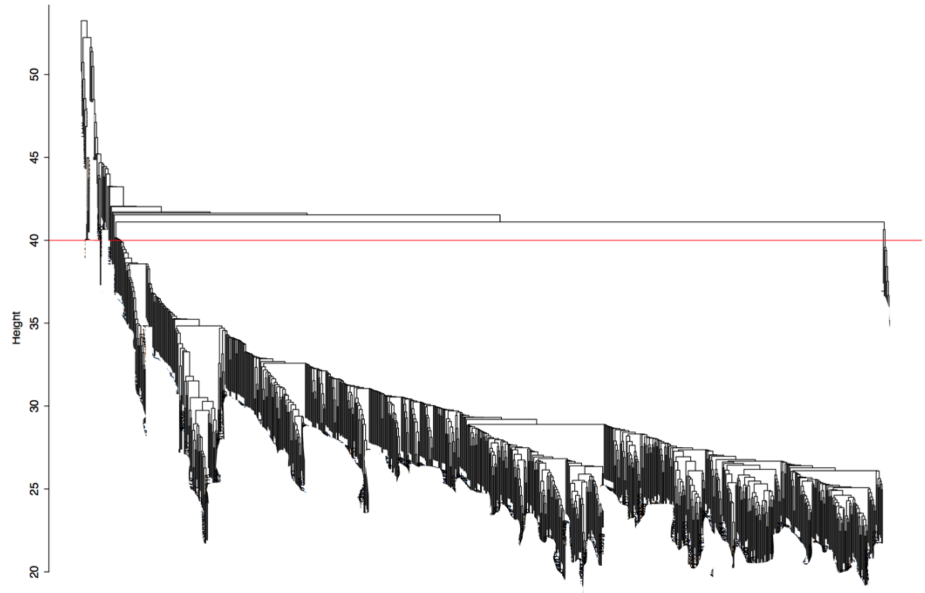


Figure 38: Hierarchical clustering of samples, in ARIES at 7 years. Red line indicates height at which samples were cut.

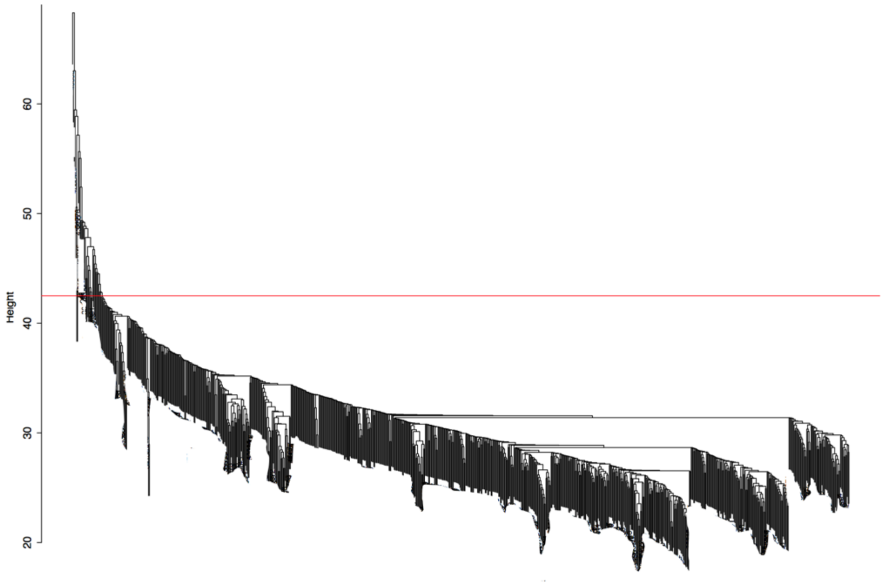


Figure 39: Hierarchical clustering of samples, in ARIES at 15-17 years. Red line indicates height at which samples were cut.

4.5.2 Soft power threshold

Pre-processed DNAm data (adjusted according to section 4.4.1.1.1) was used to calculate the power at which the data would reach scale-free topology. For the ARIES child timepoints, scale free topology was reached at a power of 8 for the birth and 15-17 years timepoints, and at a power of 7 for the 7 year olds (see Figure 40 below). Reaching a scale free topology fit of 0.9 for a reasonably low soft threshold power (under 20) should indicate that there is no major driver in the data; which can be a useful assessment of strong batch effects or a strong phenotype.

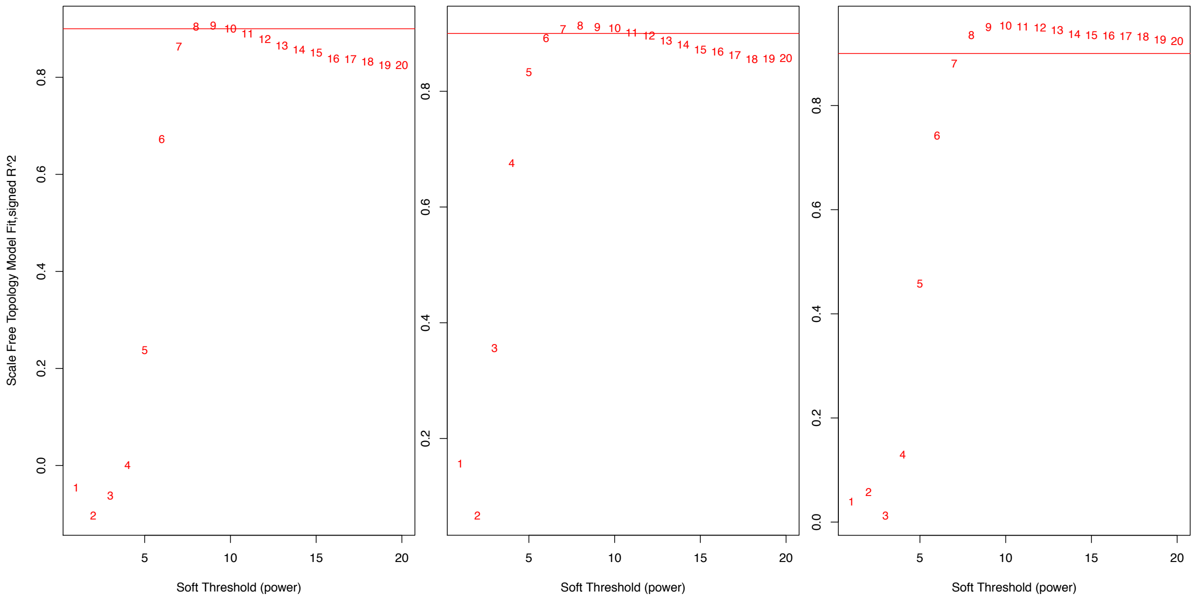


Figure 40: Soft threshold power graphs for ARIES at birth (left), 7 years (middle), and 15-17 years (right). The red line indicates the scale free model fit of 0.9; the power selected is the smallest power that reaches 0.9. Please note the differing scale on the x-axis.

4.5.3 Blockwise network construction

The modules and the number of DNAm sites assigned to them at the three ARIES timepoints are detailed in Table 11, Table 12, and Table 13 below.

4.5.3.1 Birth

Module	Number of DNAm sites	Module	Number of DNAm sites	Module	Number of DNAm sites
Black	4944	Green yellow	3457	Pink	4911
Blue	72648	Grey (unassigned)	108012	Purple	4000
Brown	47960	Grey60	297	Red	7945
Cyan	684	Light cyan	379	Royal blue	172
Dark green	125	Light green	188	Salmon	700
Dark red	127	Light yellow	174	Tan	1427
Dark turquoise	34	Magenta	4557	Turquoise	82705
Green	16399	Midnight blue	489	Yellow	32508

Table 11: Module colours and sizes in the WGCNA network in ARIES at birth. The grey module contains DNAm sites which were not assigned to any module.

4.5.3.2 7 year olds

Module	Number of DNAm sites	Module	Number of DNAm sites	Module	Number of DNAm sites
Black	12063	Grey60	518	Red	13080
Blue	55942	Light cyan	576	Royal blue	81
Brown	30479	Light green	453	Salmon	1907
Cyan	1011	Light yellow	279	Tan	2610
Dark red	43	Magenta	5468	Turquoise	100810
Green	17677	Midnight blue	920	Yellow	27905
Green yellow	3849	Pink	6617		
Grey (unassigned)	108310	Purple	4244		

Table 12: Module colours and sizes in the WGCNA network in ARIES at 7 years. The grey module contains DNAm sites which were not assigned to any module.

4.5.3.3 15-17 year olds

Module	Number of DNAm sites	Module	Number of DNAm sites	Module	Number of DNAm sites
Black	10417	Grey60	525	Red	14533
Blue	64585	Light cyan	740	Royal blue	67
Brown	34903	Light green	480	Salmon	1810
Cyan	1581	Light yellow	67	Tan	4042
Dark red	38	Magenta	6522	Turquoise	76465
Green	18786	Midnight blue	1178	Yellow	25959
Green yellow	5687	Pink	6655		
Grey (unassigned)	114085	Purple	5717		

Table 13: Module colours and sizes in the WGCNA network in ARIES at 15-17 years. The grey module contains DNAm sites which were not assigned to any module.

4.5.4 Association of network modules with traits

To assess whether the DNAm modules were associated with traits related to physical development and development-related exposures, I assessed their relationship using regression analyses. As the eigengenes in WGCNA are module representations that are essentially principal components, every individual is WGCNA calculates a membership score of every module for every individual, which lies between -1 and 1.

4.5.4.1 Birth

At birth, there are relatively weak associations with the traits tested for association with the module eigengenes. Testing 7 traits for 23 modules (which excludes the grey ‘unassigned’ module) gives a Bonferroni adjusted p-value threshold of 0.0003. None of the adjusted models survive Bonferroni correction for multiple testing; the association of the MEtan eigengene marginally misses the threshold for association with birthweight (p=0.0004) (where the model is adjusted for previously specified covariates gestational age, maternal age, maternal smoking, maternal BMI and socio-economic status. In the model birthweight is not a particularly strong predictor of methylation of the eigengene, as the effect size decreases -1e-05 for every gram, which means membership of the MEtan module decreases by 0.01 for every kg of birthweight. The standard error is 3e-06 per gram, which translates to 0.003 per kg, and the full model explains only 2% of the variance in the MEtan eigengene. The heatmap can be found in Figure 41; description of the models and covariates used can be found above in section 2.2.1.9. There were only weak associations of some modules with maternal smoking, suggesting that although maternal smoking is well established as affecting DNAm (JoubertFelix, et al., 2016; Richmond et al., 2015), it may not do so in a concerted way on a group of sites (at least, sites that are measured on the 450k).

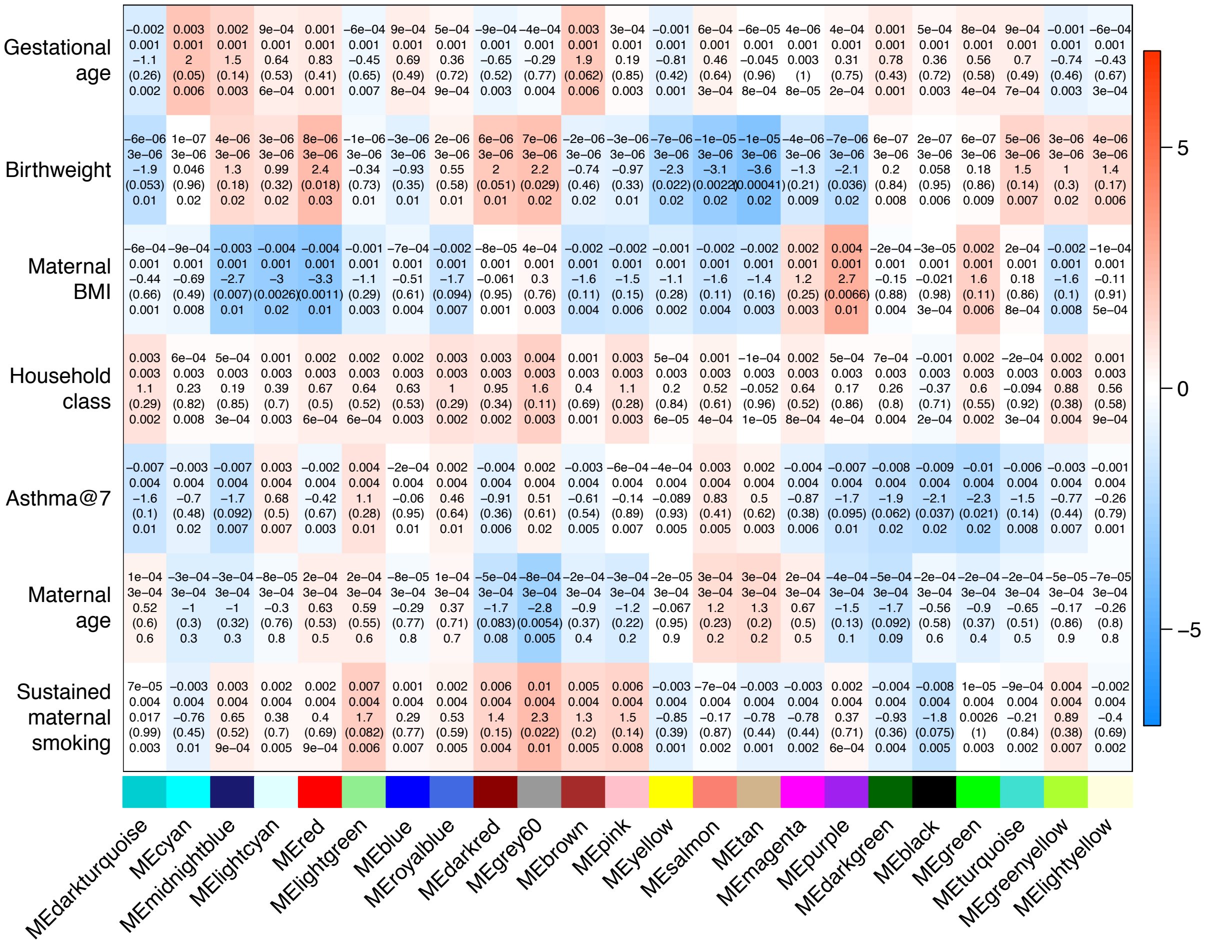


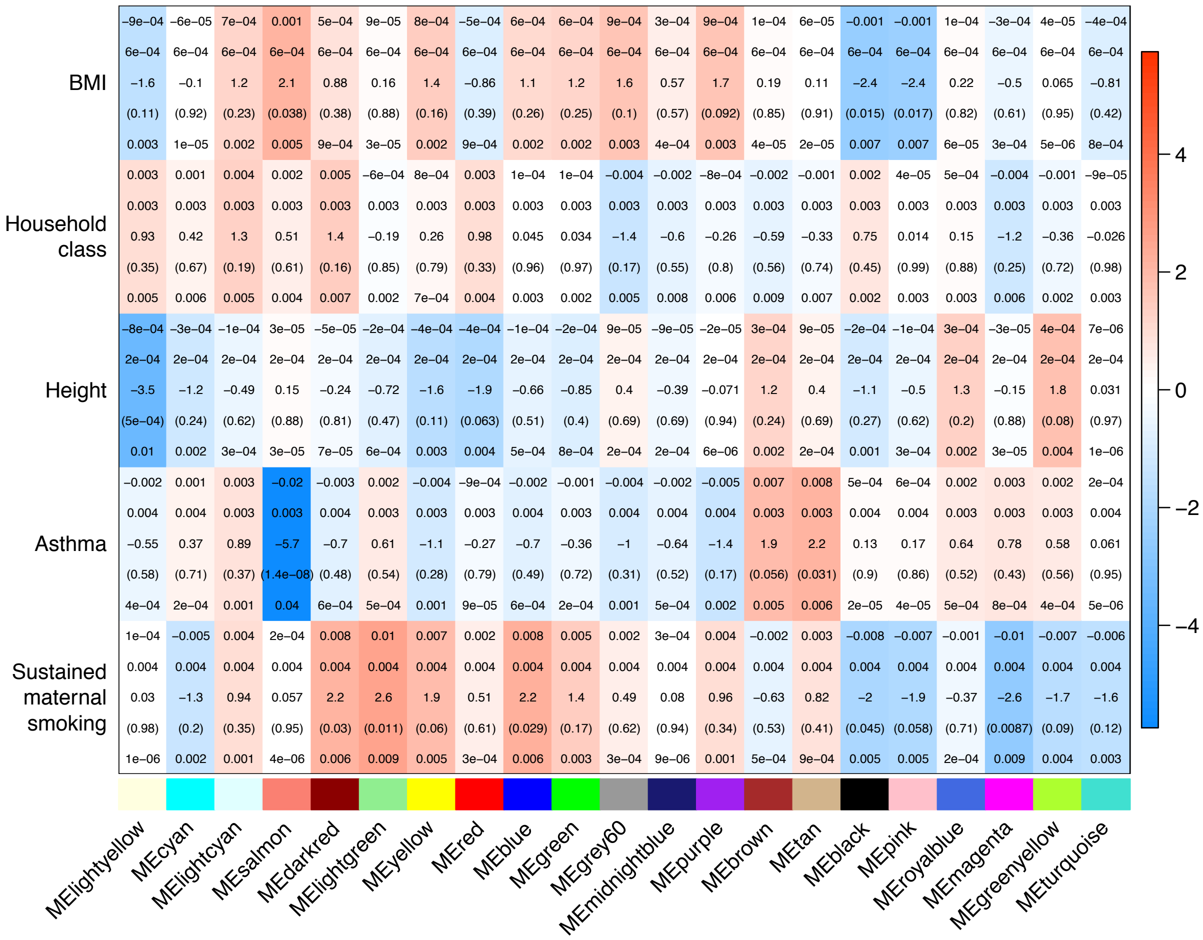
Figure 41: Heatmap of module eigengene-trait relationships at birth in ARIES. Data displayed are ***beta; standard error; t score; p-value; r-squared*** *from the regression model. Heatmap is coloured by* ***t score****.*

4.5.4.2 7 years

At 7 years, there is only one notable module-trait association, as illustrated by the heatmap in Figure 42. The Bonferroni corrected threshold at 7 years is 0.00048. The MEsalmon module is clearly associated with asthma (p= 2.9e-07, effect size = -0.02, se=0.004), showing that the model (adjusted for maternal age, socio-economic status, maternal asthma and maternal smoking). Although the t value (-5.2) and standard error (0.004) demonstrate confidence in this finding, in this model asthma only explains 5% of the MEsalmon eigengene variance.

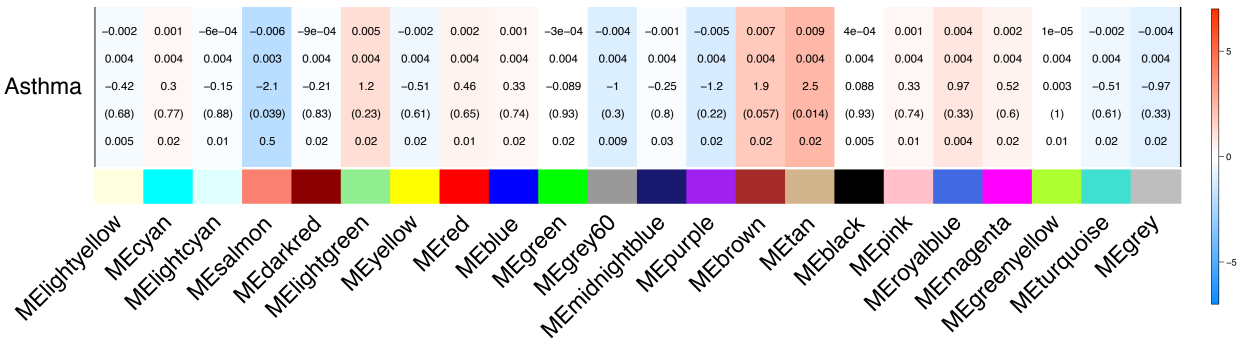
Because eosinophil and neutrophil proportions have been shown to affect the association of DNAm with asthma (Arathimos et al., 2017), I also ran the model adjusting for eosinophil and neutrophil proportions in addition to the original covariates. I found that adding eosinophil and neutrophil counts into the model removes the association of the MEsalmon eigengene and asthma (p=0.039, effect size=-0.006, se=0.003). Adding eosinophils and neutrophils into the model then predicts 50% of the variance in the MEsalmon eigengene; eosinophils are by far the greatest contribution to this (p<2e-16, effect size= -0.56, se=0.02) as compared to neutrophils (p<2e-16, effect size= -0.09, se=0.01), showing that this module is very driven by eosinophil proportions (as illustrated in Figure 43).

None of the other modules approach Bonferroni corrected significance, or have models which are notably associated with the eigengenes. As at birth, there is no module clearly associated with maternal smoking, again suggesting that maternal smoking may not affect DNAm in a concerted way at multiple sites, at least for those measured by the 450k.



	MElightyellow	MEcyan	MElightcyan	MEsalmon	MEdarkred	MElightgreen	MEyellow	MEred	MEblue	MEgreen	MEgrey60	MEmidnightblue	MEpurple	MEbrown	MEtan	MEblack	MEpink	MEroyalblue	MEmagenta	MEgreenyellow	MEturquoise
BMI	−9e−04	−6e−05	7e−04	0.001	5e−04	9e−05	8e−04	−5e−04	6e−04	6e−04	9e−04	3e−04	9e−04	1e−04	6e−05	−0.001	−0.001	1e−04	−3e−04	4e−05	−4e−04
	6e−04	6e−04	6e−04	6e−04	6e−04	6e−04	6e−04	6e−04	6e−04	6e−04	6e−04	6e−04	6e−04	6e−04	6e−04	6e−04	6e−04	6e−04	6e−04	6e−04	6e−04
	−1.6	−0.1	1.2	2.1	0.88	0.16	1.4	−0.86	1.1	1.2	1.6	0.57	1.7	0.19	0.11	−2.4	−2.4	0.22	−0.5	0.065	−0.81
	(0.11)	(0.92)	(0.23)	(0.038)	(0.38)	(0.88)	(0.16)	(0.39)	(0.26)	(0.25)	(0.1)	(0.57)	(0.092)	(0.85)	(0.91)	(0.015)	(0.017)	(0.82)	(0.61)	(0.95)	(0.42)
	0.003	1e−05	0.002	0.005	9e−04	3e−05	0.002	9e−04	0.002	0.002	0.003	4e−04	0.003	4e−05	2e−05	0.007	0.007	6e−05	3e−04	5e−06	8e−04
Household class	0.003	0.001	0.004	0.002	0.005	−6e−04	8e−04	0.003	1e−04	1e−04	−0.004	−0.002	−8e−04	−0.002	−0.001	0.002	4e−05	5e−04	−0.004	−0.001	−9e−05
	0.003	0.003	0.003	0.003	0.003	0.003	0.003	0.003	0.003	0.003	0.003	0.003	0.003	0.003	0.003	0.003	0.003	0.003	0.003	0.003	0.003
	0.93	0.42	1.3	0.51	1.4	−0.19	0.26	0.98	0.045	0.034	−1.4	−0.6	−0.26	−0.59	−0.33	0.75	0.014	0.15	−1.2	−0.36	−0.026
	(0.35)	(0.67)	(0.19)	(0.61)	(0.16)	(0.85)	(0.79)	(0.33)	(0.96)	(0.97)	(0.17)	(0.55)	(0.8)	(0.56)	(0.74)	(0.45)	(0.99)	(0.88)	(0.25)	(0.72)	(0.98)
	0.005	0.006	0.005	0.004	0.007	0.002	7e−04	0.004	0.003	0.002	0.005	0.008	0.006	0.009	0.007	0.002	0.003	0.003	0.006	0.002	0.003
Height	−8e−04	−3e−04	−1e−04	3e−05	−5e−05	−2e−04	−4e−04	−4e−04	−1e−04	−2e−04	9e−05	−9e−05	−2e−05	3e−04	9e−05	−2e−04	−1e−04	3e−04	−3e−05	4e−04	7e−06
	2e−04	2e−04	2e−04	2e−04	2e−04	2e−04	2e−04	2e−04	2e−04	2e−04	2e−04	2e−04	2e−04	2e−04	2e−04	2e−04	2e−04	2e−04	2e−04	2e−04	2e−04
	−3.5	−1.2	−0.49	0.15	−0.24	−0.72	−1.6	−1.9	−0.66	−0.85	0.4	−0.39	−0.071	1.2	0.4	−1.1	−0.5	1.3	−0.15	1.8	0.031
	(5e−04)	(0.24)	(0.62)	(0.88)	(0.81)	(0.47)	(0.11)	(0.063)	(0.51)	(0.4)	(0.69)	(0.69)	(0.94)	(0.24)	(0.69)	(0.27)	(0.62)	(0.2)	(0.88)	(0.08)	(0.97)
	0.01	0.002	3e−04	3e−05	7e−05	6e−04	0.003	0.004	5e−04	8e−04	2e−04	2e−04	6e−06	0.002	2e−04	0.001	3e−04	0.002	3e−05	0.004	1e−06
Asthma	−0.002	0.001	0.003	−0.02	−0.003	0.002	−0.004	−9e−04	−0.002	−0.001	−0.004	−0.002	−0.005	0.007	0.008	5e−04	6e−04	0.002	0.003	0.002	2e−04
	0.004	0.004	0.003	0.003	0.004	0.003	0.003	0.003	0.004	0.003	0.004	0.004	0.004	0.003	0.003	0.004	0.004	0.003	0.003	0.003	0.004
	−0.55	0.37	0.89	−5.7	−0.7	0.61	−1.1	−0.27	−0.7	−0.36	−1	−0.64	−1.4	1.9	2.2	0.13	0.17	0.64	0.78	0.58	0.061
	(0.58)	(0.71)	(0.37)	(1.4e−08)	(0.48)	(0.54)	(0.28)	(0.79)	(0.49)	(0.72)	(0.31)	(0.52)	(0.17)	(0.056)	(0.031)	(0.9)	(0.86)	(0.52)	(0.43)	(0.56)	(0.95)
	4e−04	2e−04	0.001	0.04	6e−04	5e−04	0.001	9e−05	6e−04	2e−04	0.001	5e−04	0.002	0.005	0.006	2e−05	4e−05	5e−04	8e−04	4e−04	5e−06
Sustained maternal smoking	1e−04	−0.005	0.004	2e−04	0.008	0.01	0.007	0.002	0.008	0.005	0.002	3e−04	0.004	−0.002	0.003	−0.008	−0.007	−0.001	−0.01	−0.007	−0.006
	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004
	0.03	−1.3	0.94	0.057	2.2	2.6	1.9	0.51	2.2	1.4	0.49	0.08	0.96	−0.63	0.82	−2	−1.9	−0.37	−2.6	−1.7	−1.6
	(0.98)	(0.2)	(0.35)	(0.95)	(0.03)	(0.011)	(0.06)	(0.61)	(0.029)	(0.17)	(0.62)	(0.94)	(0.34)	(0.53)	(0.41)	(0.045)	(0.058)	(0.71)	(0.0087)	(0.09)	(0.12)
	1e−06	0.002	0.001	4e−06	0.006	0.009	0.005	3e−04	0.006	0.003	3e−04	9e−06	0.001	5e−04	9e−04	0.005	0.005	2e−04	0.009	0.004	0.003

*Figure 42: Heatmap of module eigengene-trait relationships at 7 years in ARIES. Data displayed are **beta; standard error; t score; p-value; r-squared** from the regression model. Heatmap is coloured by **t score**.*



	MElightyellow	MEcyan	MElightcyan	MEsalmon	MEdarkred	MElightgreen	MEyellow	MEred	MEblue	MEgreen	MEgrey60	MEmidnightblue	MEpurple	MEbrown	MEtan	MEblack	MEpink	MEroyalblue	MEmagenta	MEgreenyellow	MEturquoise	MEgrey
Asthma	−0.002	0.001	−6e−04	−0.006	−9e−04	0.005	−0.002	0.002	0.001	−3e−04	−0.004	−0.001	−0.005	0.007	0.009	4e−04	0.001	0.004	0.002	1e−05	−0.002	−0.004
	0.004	0.004	0.004	0.003	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004
	−0.42	0.3	−0.15	−2.1	−0.21	1.2	−0.51	0.46	0.33	−0.089	−1	−0.25	−1.2	1.9	2.5	0.088	0.33	0.97	0.52	0.003	−0.51	−0.97
	(0.68)	(0.77)	(0.88)	(0.039)	(0.83)	(0.23)	(0.61)	(0.65)	(0.74)	(0.93)	(0.3)	(0.8)	(0.22)	(0.057)	(0.014)	(0.93)	(0.74)	(0.33)	(0.6)	(1)	(0.61)	(0.33)
	0.005	0.02	0.01	0.5	0.02	0.02	0.02	0.01	0.02	0.02	0.009	0.03	0.02	0.02	0.02	0.005	0.01	0.004	0.02	0.01	0.02	0.02

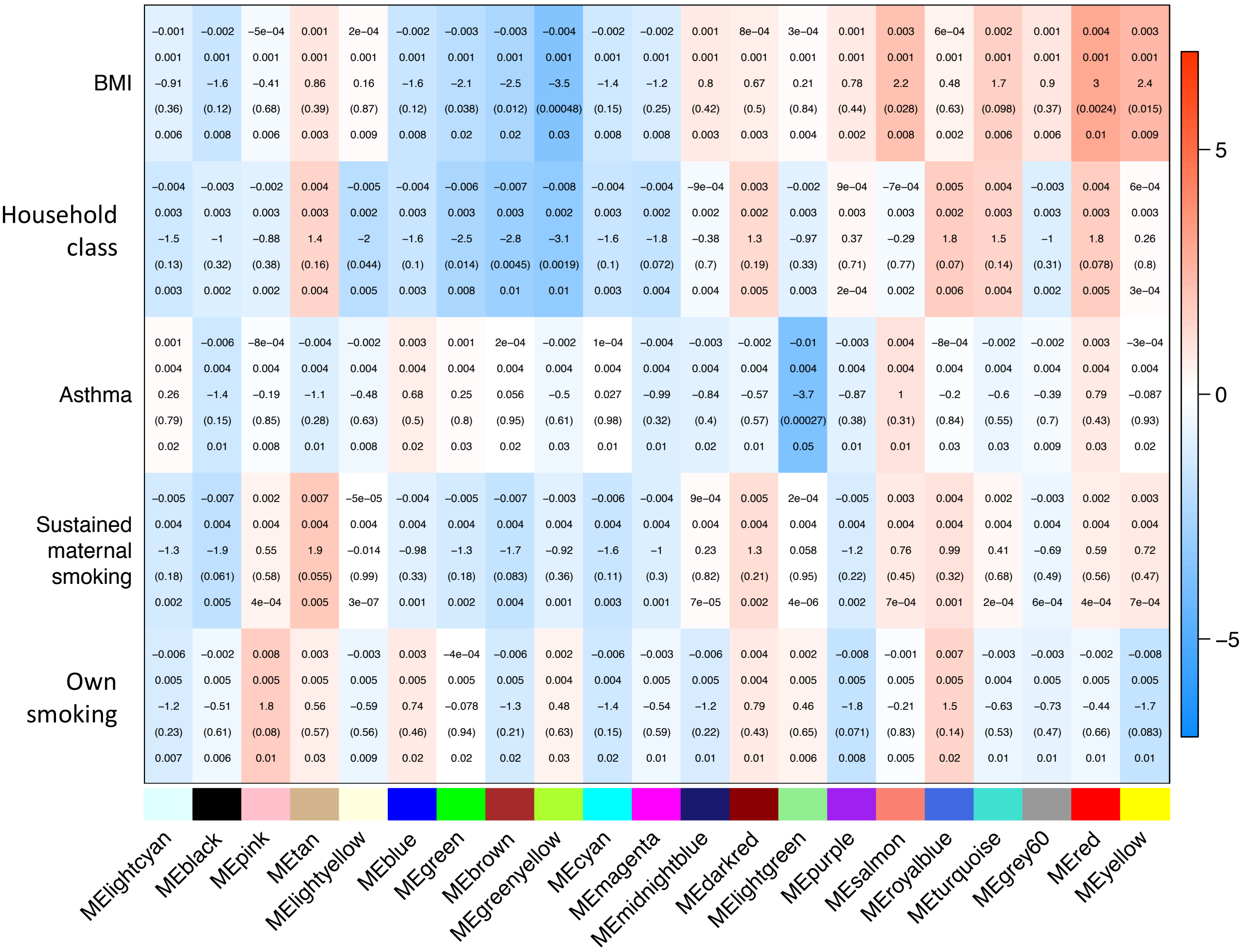
*Figure 43: Heatmap of module eigengene-asthma relationships, **adjusted for eosinophil and neutrophil counts**, at 7 years in ARIES. Data displayed are **beta; standard error; t score; p-value; r-squared** from the regression model. Heatmap is coloured by **t score**.*

4.5.4.3 15-17 years

In the 15-17 year olds, we see two module eigengene-trait associations that survive multiple testing (Bonferroni corrected threshold of 0.00048). The MElightgreen module is associated with asthma (p=0.00027, effect size=-0.01, se=0.004), however the small effect on eigengene methylation, and the small amount of variation in the eigengene asthma explains (5%) suggests that this module is not primarily related to asthma. The MEgreenyellow module is marginally associated with BMI (p=0.00048, effect size=0.004, se=0.001). A heatmap of the module eigengene-trait relationships is shown in Figure 44.

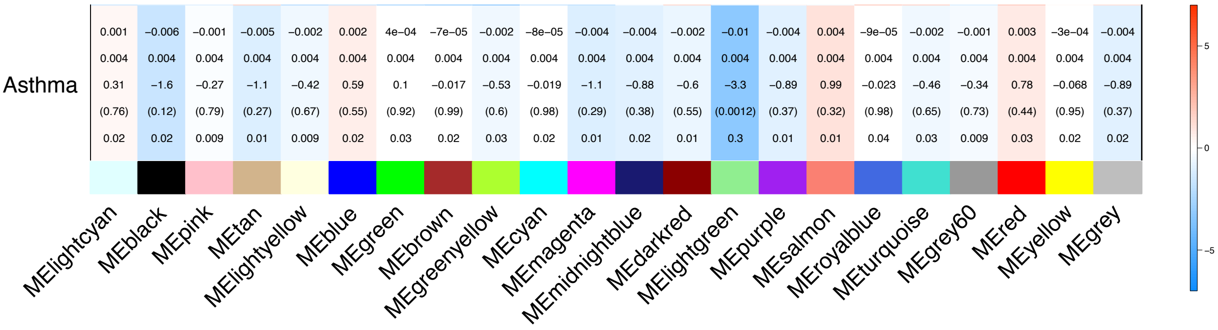
Because eosinophil and neutrophil proportions have been shown to affect the association of DNAm with asthma (Arathimos et al., 2017), I also ran the model adjusting for eosinophil and neutrophil proportions in addition to the original covariates. I found that adding eosinophil and neutrophil counts into the model did not actually change the effect size of the association of the MElightgreen eigengene with asthma (p=0.001, effect size=-0.01, se=0.004), it just reduced the p-value associated with the finding. Adding eosinophils and neutrophils into the model then predicts 30% of the variance in the MElightgreen eigengene; eosinophils have a larger effect size than at 7 years (p<2e-16, effect size= -0.62, se=0.04), and neutrophils have a smaller effect (p<=3.8e-13, effect size= -0.08, se=0.01), showing that this module is very driven by eosinophil proportions (as illustrated in Figure 45).

There is no association with maternal smoking, which would not really be expected from the literature (Richmond et al., 2015). Perhaps surprisingly, we do not see a module associated with own smoking, as own smoking in ARIES has been associated with changes at a number of DNAm sites (Prince et al., 2019). 8.7% of the sample were classed as smokers, so it is possible this number is not large enough to see a module emerge. Alternatively it may again suggest that smoking does not act on DNAm sites as part of a single pathway, that can be detected on the 450k.



	MElightcyan	MEblack	MEpink	MEtan	MElightyellow	MEblue	MEgreen	MEbrown	MEgreenyellow	MEcyan	MEmagenta	MEmidnightblue	MEdarkred	MElightgreen	MEpurple	MEsalmon	MEroyalblue	MEturquoise	MEgrey60	MEred	MEyellow
BMI	−0.001 0.001 −0.91 (0.36) 0.006	−0.002 0.001 −1.6 (0.12) 0.008	−5e−04 0.001 −0.41 (0.68) 0.006	0.001 0.001 0.86 (0.39) 0.003	2e−04 0.001 0.16 (0.87) 0.009	−0.002 0.001 −1.6 (0.12) 0.008	−0.003 0.001 −2.1 (0.038) 0.02	−0.003 0.001 −2.5 (0.012) 0.02	−0.004 0.001 −3.5 (0.00048) 0.03	−0.002 0.001 −1.4 (0.15) 0.008	−0.002 0.001 −1.2 (0.25) 0.008	0.001 0.001 0.8 (0.42) 0.003	8e−04 0.001 0.67 (0.5) 0.003	3e−04 0.001 0.21 (0.84) 0.004	0.001 0.001 0.78 (0.44) 0.002	0.003 0.001 2.2 (0.028) 0.008	6e−04 0.001 0.48 (0.63) 0.002	0.002 0.001 1.7 (0.098) 0.006	0.001 0.001 0.9 (0.37) 0.006	0.004 0.001 3 (0.0024) 0.01	0.003 0.001 2.4 (0.015) 0.009
Household class	−0.004 0.003 −1.5 (0.13) 0.003	−0.003 0.003 −1 (0.32) 0.002	−0.002 0.003 −0.88 (0.38) 0.002	0.004 0.003 1.4 (0.16) 0.004	−0.005 0.002 −2 (0.044) 0.005	−0.004 0.003 −1.6 (0.1) 0.003	−0.006 0.003 −2.5 (0.014) 0.008	−0.007 0.003 −2.8 (0.0045) 0.01	−0.008 0.002 −3.1 (0.0019) 0.01	−0.004 0.003 −1.6 (0.1) 0.003	−0.004 0.002 −1.8 (0.072) 0.004	−9e−04 0.002 −0.38 (0.7) 0.004	0.003 0.002 1.3 (0.19) 0.005	−0.002 0.003 −0.97 (0.33) 0.003	9e−04 0.003 0.37 (0.71) 2e−04	−7e−04 0.003 −0.29 (0.77) 0.002	0.005 0.002 1.8 (0.07) 0.006	0.004 0.003 1.5 (0.14) 0.004	−0.003 0.003 −1 (0.31) 0.002	0.004 0.003 1.8 (0.078) 0.005	6e−04 0.003 0.26 (0.8) 3e−04
Asthma	0.001 0.004 0.26 (0.79) 0.02	−0.006 0.004 −1.4 (0.15) 0.01	−8e−04 0.004 −0.19 (0.85) 0.008	−0.004 0.004 −1.1 (0.28) 0.01	−0.002 0.004 −0.48 (0.63) 0.008	0.003 0.004 0.68 (0.5) 0.02	0.001 0.004 0.25 (0.8) 0.03	2e−04 0.004 0.056 (0.95) 0.02	−0.002 0.004 −0.5 (0.61) 0.03	1e−04 0.004 0.027 (0.98) 0.01	−0.004 0.004 −0.99 (0.32) 0.01	−0.003 0.004 −0.84 (0.4) 0.02	−0.002 0.004 −0.57 (0.57) 0.01	−0.01 0.004 −3.7 (0.00027) 0.05	−0.003 0.004 −0.87 (0.38) 0.01	0.004 0.004 1 (0.31) 0.01	−8e−04 0.004 −0.2 (0.84) 0.03	−0.002 0.004 −0.6 (0.55) 0.03	−0.002 0.004 −0.39 (0.7) 0.009	0.003 0.004 0.79 (0.43) 0.03	−3e−04 0.004 −0.087 (0.93) 0.02
Sustained maternal smoking	−0.005 0.004 −1.3 (0.18) 0.002	−0.007 0.004 −1.9 (0.061) 0.005	0.002 0.004 0.55 (0.58) 4e−04	0.007 0.004 1.9 (0.055) 0.005	−5e−05 0.004 −0.014 (0.99) 3e−07	−0.004 0.004 −0.98 (0.33) 0.001	−0.005 0.004 −1.3 (0.18) 0.002	−0.007 0.004 −1.7 (0.083) 0.004	−0.003 0.004 −0.92 (0.36) 0.001	−0.006 0.004 −1.6 (0.11) 0.003	−0.004 0.004 −1 (0.3) 0.001	9e−04 0.004 0.23 (0.82) 7e−05	0.005 0.004 1.3 (0.21) 0.002	2e−04 0.004 0.058 (0.95) 4e−06	−0.005 0.004 −1.2 (0.22) 0.002	0.003 0.004 0.76 (0.45) 7e−04	0.004 0.004 0.99 (0.32) 0.001	0.002 0.004 0.41 (0.68) 2e−04	−0.003 0.004 −0.69 (0.49) 6e−04	0.002 0.004 0.59 (0.56) 4e−04	0.003 0.004 0.72 (0.47) 7e−04
Own smoking	−0.006 0.005 −1.2 (0.23) 0.007	−0.002 0.005 −0.51 (0.61) 0.006	0.008 0.005 1.8 (0.08) 0.01	0.003 0.005 0.56 (0.57) 0.03	−0.003 0.005 −0.59 (0.56) 0.009	0.003 0.005 0.74 (0.46) 0.02	−4e−04 0.005 −0.078 (0.94) 0.02	−0.006 0.005 −1.3 (0.21) 0.02	0.002 0.004 0.48 (0.63) 0.03	−0.006 0.004 −1.4 (0.15) 0.02	−0.003 0.005 −0.54 (0.59) 0.01	−0.006 0.005 −1.2 (0.22) 0.01	0.004 0.004 0.79 (0.43) 0.01	0.002 0.005 0.46 (0.65) 0.006	−0.008 0.005 −1.8 (0.071) 0.008	−0.001 0.005 −0.21 (0.83) 0.005	0.007 0.005 1.5 (0.14) 0.02	−0.003 0.004 −0.63 (0.53) 0.01	−0.003 0.005 −0.73 (0.47) 0.01	−0.002 0.005 −0.44 (0.66) 0.01	−0.008 0.005 −1.7 (0.083) 0.01

Figure 44: Heatmap of module eigengene-trait relationships at 15-17 years in ARIES. Data displayed are ***beta; standard error; t score; p-value; r-squared*** *from the regression model. Heatmap is coloured by* ***t score****.*



	MElightcyan	MEblack	MEpink	MEtan	MElightyellow	MEblue	MEgreen	MEbrown	MEgreenyellow	MEcyan	MEmagenta	MEmidnightblue	MEdarkred	MElightgreen	MEpurple	MEsalmon	MEroyalblue	MEturquoise	MEgrey60	MEred	MEyellow	MEgrey
Asthma	0.001 0.004 0.31 (0.76) 0.02	−0.006 0.004 −1.6 (0.12) 0.02	−0.001 0.004 −0.27 (0.79) 0.009	−0.005 0.004 −1.1 (0.27) 0.01	−0.002 0.004 −0.42 (0.67) 0.009	0.002 0.004 0.59 (0.55) 0.02	4e−04 0.004 0.1 (0.92) 0.03	−7e−05 0.004 −0.017 (0.99) 0.02	−0.002 0.004 −0.53 (0.6) 0.03	−8e−05 0.004 −0.019 (0.98) 0.02	−0.004 0.004 −1.1 (0.29) 0.01	−0.004 0.004 −0.88 (0.38) 0.02	−0.002 0.004 −0.6 (0.55) 0.01	−0.01 0.004 −3.3 (0.0012) 0.3	−0.004 0.004 −0.89 (0.37) 0.01	0.004 0.004 0.99 (0.32) 0.01	−9e−05 0.004 −0.023 (0.98) 0.04	−0.002 0.004 −0.46 (0.65) 0.03	−0.001 0.004 −0.34 (0.73) 0.009	0.003 0.004 0.78 (0.44) 0.03	−3e−04 0.004 −0.068 (0.95) 0.02	−0.004 0.004 −0.89 (0.37) 0.02

Figure 45: Heatmap of module eigengene-asthma relationships, ***adjusted for eosinophil and neutrophil counts****, at 15-17 years in ARIES. Data displayed are* ***beta; standard error; t score; p-value; r-squared*** *from the regression model. Heatmap is coloured by* ***t score****.*

4.5.4.4 Asthma

4.5.4.4.1 Overall observations

At birth, we do not observe an association with any DNAm module and asthma at 7 years old (which is the first timepoint at which ALSPAC has an asthma measurement). This may suggest that any DNAm involvement in asthma as a network is not seen from birth; alternatively it may mean that future asthma is not an appropriate phenotype. At 7 and 15-17 years, there is a module that is clearly associated with asthma. The stronger association at 7 years may reflect the more refined phenotype; this was not available at 15-17 years due to fewer questions being asked about asthma at that timepoint (please see chapter 2, section 2.2.1.9.1.3 for more detail).

4.5.4.4.2 7 years

When assessing the 1,000 DNAm sites that are most closely related to the MEsalmon eigengene in the 7 year olds (association is known as kME, which is a correlation between each DNAm site and the module eigengene), we see a clear relationship between the membership of the module, and the strength of the relationship between the DNAm site and asthma (Figure 46). The DNAm sites that are most closely related to the MEsalmon eigengene (denoted as kME >0.7) have the strongest relationships with asthma of all the ~400,000 probes in the analysis. The strongest correlations with asthma are negative; a table containing the top 20 asthma related DNAm sites is found in Appendix 3.

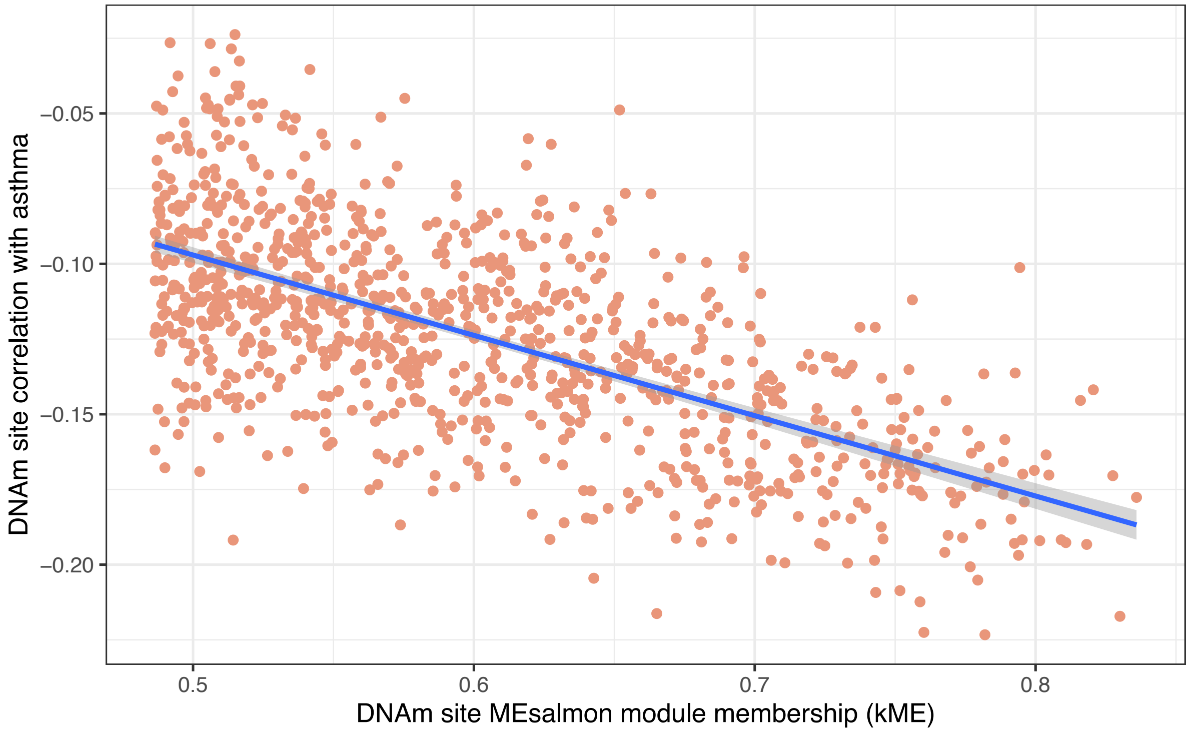


Figure 46: Correlation between module membership of the top 1,000 MEsalmon associated DNAm sites and their correlation with asthma status, in ARIES 7 year olds. Correlation = -0.65.

DNAm sites are conventionally assumed to be strongly associated with a module eigengene with a kME >0.7. There were 161 DNAm sites with kME >0.7 with the MEsalmon eigengene. When testing for associations for gene ontology terms (detail is below in section 4.5.5.1.2 and Table 16), the MEsalmon module is weakly associated with GO terms for *carbohydrate binding* and *monosaccharide binding*. When testing enrichment for KEGG pathways, the only pathway that survives correction for multiple testing is the asthma KEGG pathway (shown below in Table 17).

4.5.4.4.3 15-17 years

When we plot the top 1,000 probes related to the light green eigengene in the 15-17 year olds, we see a weaker relationship than in the 7 year olds (shown in Figure 47). In GO analysis the 21 probes most closely resembling the light green module eigengene were significantly enriched for terms relating to *regulation of interleukin 4 production* (shown in Table 18). Again, the only KEGG pathway that survives correction for multiple testing is *asthma* (Table 19).

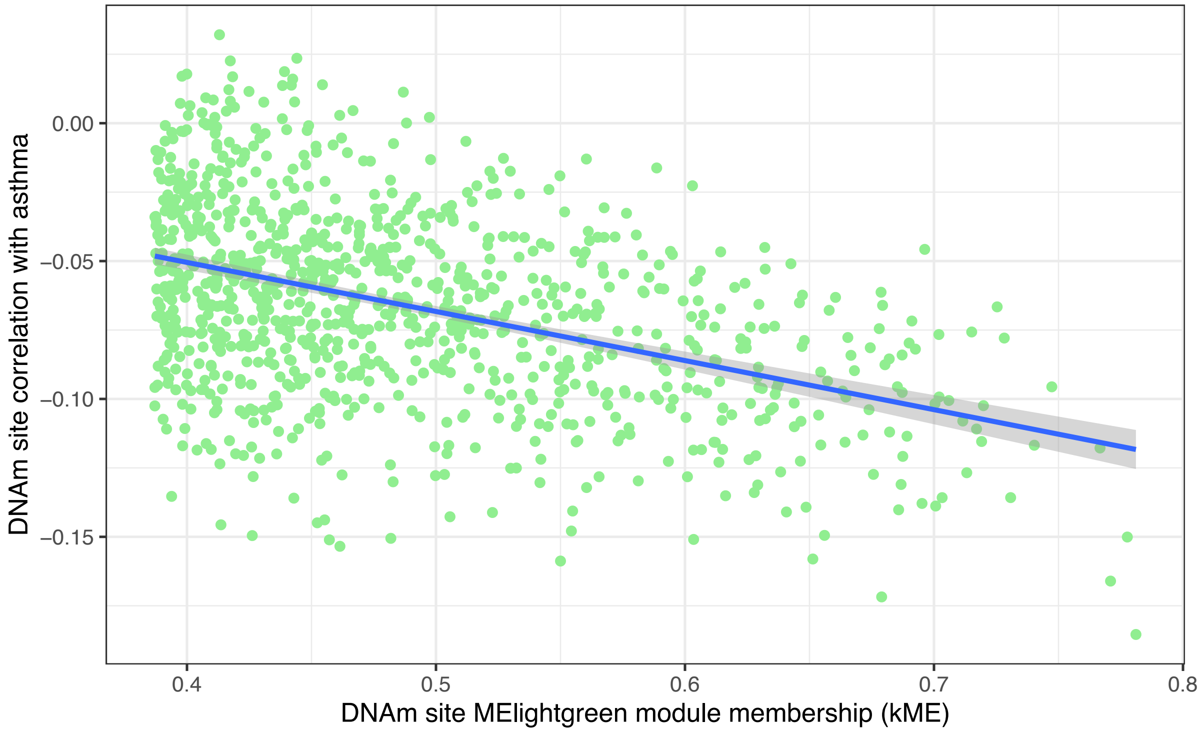


Figure 47: Correlation between module membership of the top 1,000 MElightgreen associated DNAm sites and their correlation with asthma status, in ARIES 15-17 year olds. Correlation = -0.43.

4.5.5 Functional annotation of the modules

4.5.5.1 Gene ontology and KEGG pathway analysis

I conducted a gene ontology analysis to identify whether any of the WGCNA modules were associated with clear biological functions. For each of the three ARIES timepoints separately, module membership of all modules was calculated for all DNAm sites. This means module membership is not exclusive, which is more representative of biology. For each group, the top module members (kME>0.7) were assessed for gene ontology as compared to a background of all probes used in the analysis. Modules which had ontologies associated close to or below FDR p<0.05 are presented in the tables below; tables of the top 20 gene ontology terms for every module with enrichments close to the FDR threshold can be found in Appendix 4.

4.5.5.1.1 Birth

At birth there are multiple modules associated with GO terms related to intracellular processes. Of particular relevance to development is the MEbrown module, which has weak associations with *anatomical structure morphogenesis*, *regulation of developmental process*, and *regulation of developmental growth*. The MEgreen module is associated with *RNA polymerase II (Pol2) activity* and *transcription factor binding*, which is likely to be a reflection of the finding in chapter 3 that strongly correlated *trans* DNAm sites are enriched for Pol2 binding sites, as well as many other TFBS. The MEblack module is enriched for GO terms relating to the nucleus, and the KEGG *pyruvate metabolism* pathway.

The MEturquoise module is very strongly enriched for GO terms relating to *intracellular organelles*, and is enriched for numerous and slightly diverse KEGG pathways. As this is a large module (12,927 DNAm sites are kME>0.7) it is possible that this group of co-methylated sites are involved in multiple pathways. The MElightgreen module is weakly enriched for genes relating to the KEGG pathways for *nucleotide excision repair* and *basal transcription factors,* suggesting this module could be involved in the repair of DNA damage; again this would fit with the strongly correlated DNAm sites being enriched for TFBS in chapter 3.

Module eigengene	n DNAm sites kME>0.7	n GO terms FDR p<0.05	Gene ontology	FDR p-value
Black	1,608	4	Nuclear part Intracellular membrane-bounded organelle Cell cycle	0.003 0.03 0.13
Blue	11,375	25	Cellular components and adhesion Biological adhesion Plasma membrane	1.03E-05 4.85E-05 0.0001
Brown	88	0	Regulation of anatomical structure morphogenesis Regulation of developmental process Negative regulation of developmental growth	0.06 0.08 0.11
Green	290	83	RNA polymerase II transcription factor activity Sequence-specific DNA binding DNA binding transcription factor activity	4.28E-10 1.75E-08 4.70E-08
Purple	497	17	Homophilic cell adhesion via plasma membrane adhesion molecules Integral component of plasma membrane Cell-cell signaling	1.13E-08 6.11E-05 0.0001
Turquoise	12,927	729	Intracellular membrane-bounded organelle Nucleus Cellular metabolic process	1.71E-76 8.87E-51 1.85E-42
Yellow	1,500	99	Intracellular organelle Nucleus Protein-containing complex	6.32E-11 4.22E-07 8.41E-07

Table 14: Gene ontologies associated with the top module eigengene-associated DNAm sites, in ARIES at birth.

Module eigengene	n DNAm sites kME>0.7	n KEGG terms FDR p<0.05	KEGG pathway	FDR p-value
Black	1,608	0	Pyruvate metabolism	0.07
Blue	11,375	0	PI3K-Akt signaling pathway Adherens junction	0.16 0.16
Light green	20	0	Nucleotide excision repair Basal transcription factors	0.1 0.1
Turquoise	12,927	48	Protein processing in endoplasmic reticulum Chronic myeloid leukemia Autophagy - animal	5.16E-07 0.0008 0.0008

Table 15: KEGG pathways associated with the top module eigengene-associated DNAm sites, in ARIES at birth

4.5.5.1.2 7 years

At 7 years, two of the modules (MElightcyan and MEpink) are enriched for immune-related GO terms, which may indicate they are related to cell count proportions. The MElightcyan module is enriched for the KEGG pathway for *Th1 and Th2 cell differentiation*, and for KEGG pathways associated with various infectious diseases, and so is likely to represent a module of DNAm sites that are co-ordinately regulated in T cells, and may be involved in response to infectious disease. The MEpink module is associated with the *Amoebiasis* KEGG pathway, as well as the *pyruvate metabolism* pathway.

The MEyellow module is enriched for the GO term regulation of anatomical structure morphogenesis, which is also an enrichment found at birth. The MEgrey60 module is enriched for *intracellular organelle* related GO terms, and the KEGG pathway enrichment for *Ribosome* may indicate this is the organelle these DNAm sites are involved with. The MEsalmon module has already been discussed in section 4.5.4.4.2 above. The ME turquoise module is very strongly enriched for GO terms relating to intracellular organelles, and is moderately strongly enriched for the *protein processing in endoplasmic reticulum* KEGG pathway, again suggesting the ER is the organelle this relates to.

Module eigengene	n DNAm sites kME>0.7	n GO terms FDR p<0.05	Gene ontology	FDR p-value
Blue	16,274	34	Cellular component Transferase activity Catalytic activity, acting on a protein	1.40E-07 1.39E-05 3.65E-05
Grey 60	457	67	Intracellular organelle part Membrane-enclosed lumen Nuclear part	9.12E-05 0.0001 0.0002
Light cyan	43	25	Alpha-beta T cell receptor complex Immune response Antigen receptor-mediated signaling pathway	0.001 0.001 0.005
Magenta	1,076	0	Cellular nitrogen compound metabolic process	0.06
Pink	718	93	Neutrophil degranulation Neutrophil activation involved in immune response Leukocyte degranulation	8.38E-11 8.38E-11 8.38E-11
Salmon	161	0	Monosaccharide binding Secretion by tissue Carbohydrate binding	0.06 0.06 0.06
Turquoise	13,870	791	Intracellular part Membrane-bounded organelle Cellular metabolic process	2.20E-83 9.14E-65 8.4E-50
Yellow	116	1	Regulation of anatomical structure morphogenesis	0.002

Table 16: Gene ontologies associated with the top module eigengene-associated DNAm sites, in ARIES 7 year olds.

Module eigengene	n DNAm sites kME>0.7	n KEGG terms FDR p<0.05	KEGG pathway	FDR p-value
Grey 60	457	1	Ribosome	0.0004
Light cyan	43	8	Th1 and Th2 cell differentiation Chagas disease (American trypanosomiasis) Epstein-Barr virus infection	0.003 0.003 0.02
Pink	718	1	Amoebiasis Pyruvate metabolism	0.04 0.06
Salmon	161	1	Asthma	0.007
Turquoise	13,870	791	Protein processing in endoplasmic reticulum Chronic myeloid leukemia Viral carcinogenesis	8.43E-05 0.0003 0.0004

Table 17: KEGG pathways associated with the top module eigengene-associated DNAm sites, in ARIES 7 year olds.

4.5.5.1.3 15-17 years

At 15-17 years there are three modules that have GO terms that are related to immune function; MEblack, MElightgreen, and MEsalmon. The MElightgreen module is associated with asthma, and it is associated with the *asthma* KEGG pathway. Interestingly the MEblack module is enriched for the KEGG pathway *adrenergic signaling in cardiomyocytes*, *cholinergic synapse*, and *platelet activation*. Acetylcholine (released from cholinergic synapses) causes the release of nitric oxide from the vascular epithelium, which changes platelet activation (Andrews, Husain, Dakak, & Quyyumi, 2001).

Module eigengene	n DNAm sites kME>0.7	n GO terms FDR p<0.05	Gene ontology	FDR p-value
Black	304	36	Immune response Neutrophil degranulation Myeloid leukocyte activation	3.04E-06 4.06E-06 4.06E-06
Blue	12,812	843	Intracellular membrane-bounded organelle Nuclear lumen Cellular metabolic process	2.01E-93 2.11E-64 3.68E-45
Brown	750	41	Intracellular organelle part Nuclear lumen Protein-containing complex	1.11E-07 3.52E-05 0.0003
Cyan	183	1	2'-5'-oligoadenylate synthetase activity Regulation of ribonuclease activity	0.004 0.09
Green	252	30	Intracellular organelle part Ribonucleoprotein complex Nucleoplasm	0.01 0.01 0.01
Light cyan	209	19	RNA binding Organelle part Cellular aromatic compound metabolic process	4.31E-05 0.002 0.01
Light green	21	6	Regulation of interleukin-4 production Defense response to nematode Regulation of interleukin-10 production	7.26E-05 0.008 0.01
Magenta	71	1	Nuclear part	0.02
Salmon	27	50	Lymphocyte activation T cell activation Positive regulation of immune response	0.0002 0.0002 0.0002
Turquoise	5899	20	Membrane part Cell periphery Cell adhesion	0.0002 0.0002 0.0008
Yellow	219	1	Regulation of anatomical structure morphogenesis	0.0002

Table 18: Gene ontologies associated with the top module eigengene-associated DNAm sites, in ARIES 15-17 year olds.

Module eigengene	n DNAm sites kME>0.7	n KEGG terms FDR p<0.05	KEGG pathway	FDR p-value
Black	304	1	Adrenergic signaling in cardiomyocytes Platelet activation Cholinergic synapse	0.03 0.07 0.07
Blue	12,812	42	Autophagy - animal Protein processing in endoplasmic reticulum mRNA surveillance pathway	8.76E-05 8.76E-05 0.0002
Cyan	183	1	Influenza A	0.03
Green	252	0	Kaposi sarcoma-associated herpesvirus infection Huntington disease Alzheimer disease	0.06 0.07 0.09
Light green	21	1	Asthma	0.007
Salmon	27	10	Th1 and Th2 cell differentiation PD-L1 expression and PD-1 checkpoint pathway in cancer Th17 cell differentiation	1.89E-05 0.0004 0.0004
Turquoise	5899	0	ECM-receptor interaction Olfactory transduction PI3K-Akt signaling pathway	0.06 0.2 0.2

Table 19: KEGG pathways associated with the top module eigengene-associated DNAm sites, in ARIES 15-17 year olds.

4.5.6 Preservation of network modules over time

4.5.6.1 Birth to 7 years

All but one of the modules found at birth are well preserved (Z summary score >10) in the 7 year-olds. Bearing in mind the bias of module size on Z summary score, the MEroyalblue module seems particularly well preserved, taking in the relatively small module size with the highest median rank of the modules, indicating it is the best preserved (see Figure 48 for illustration). The MEroyalblue module is not associated with any GO or KEGG terms, or any of the traits tested, and so it is not entirely clear why this module is well preserved. This is the case for the 9 best preserved modules, according to the median rank. The MEbrown module, which has a Z summary score of around 70, is enriched for the GO term *Regulation of anatomical structure morphogenesis,* and 93% of the module members kME>0.7 are preserved in the MEyellow module in the 7 year olds, which is also enriched for this GO term.

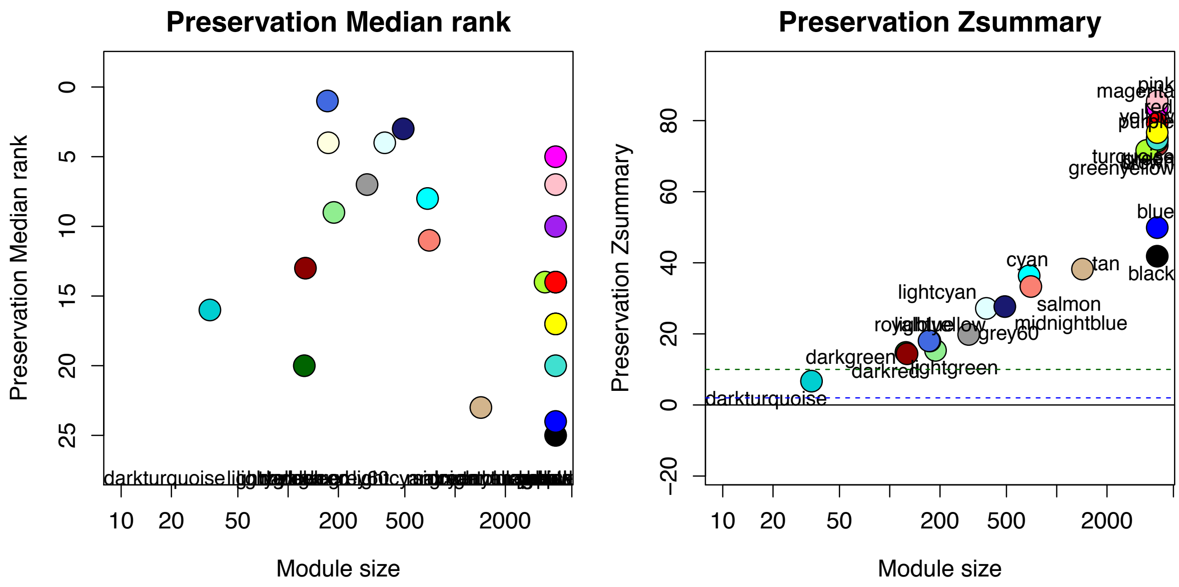


Figure 48: Z summary and median rank statistics plots illustrating the preservation of modules found at birth in the 7 year olds.

4.5.6.2 7 years to adolescence

All but two of the WGCNA modules found at 7 years are well preserved in adolescence, as illustrated by a preservation Z-summary score >10 (shown in Figure 49). The MEsalmon module, which is associated with asthma at 7 years, is strongly preserved in the adolescents, with 96% of the MEsalmon module members kME>0.7 being allocated to the MElightgreen module in the adolescents. This indicates the relationship between asthma-related DNAm sites may change slightly over time, which could relate to the change in gene ontology terms between 7 and 15-17 years; however this means the network relationship is largely the same.

The pink module, which is associated with *immune response*-related GO terms, is very strongly preserved, and has the joint top median rank for preservation. 100% of the MEpink module members kME >0.7 overlap the MEblack module at age 15-17. Inspection of the top 20 GO and KEGG terms for both the timepoints (Appendix 4) reveals that these modules have 20 very similar *immune response*-related GO terms at FDR <0.05, and a number of the same KEGG terms (which reach FDR significance in the adolescents but not the 7 year olds), including *cholinergic synapse* and *adrenergic cardiomyocyte signalling*. As these modules have such similar enrichments, it is likely this represents an immune system-related module that persists in structure and function between 7 and adolescence.

The MEgrey60 module is also strongly preserved and joint top median rank. 49% of the MEgrey60 module members kME>0.7 in the 7 year olds overlap the MElightcyan module in the 15-17 year olds, and 41% overlap the MEblue module. The GO terms for *intracellular organelles* are very similar in both these modules, raising the possibility that the MEgrey60 module branches sufficiently by adolescence that it becomes a second module.

The MEyellow is also very well preserved. 99% of the MEyellow module in the 7 year olds overlap the MEyellow module in the adolescents. Both of these modules are enriched for the GO term *Regulation of anatomical structure morphogenesis*, and as the MEyellow module in the 7 year olds is preserved from birth with the same GO term enrichment, this may suggest this is a module of DNAm sites with strongly preserved co-methylation across development, that may have a role in the regulation of the development of anatomical structures.

99% of the MEturquoise members are found in the MEblue module at 15-17. Both modules are strongly enriched for terms relating to intracellular organelles, and both are enriched for the *Protein processing in endoplasmic reticulum* KEGG pathway.

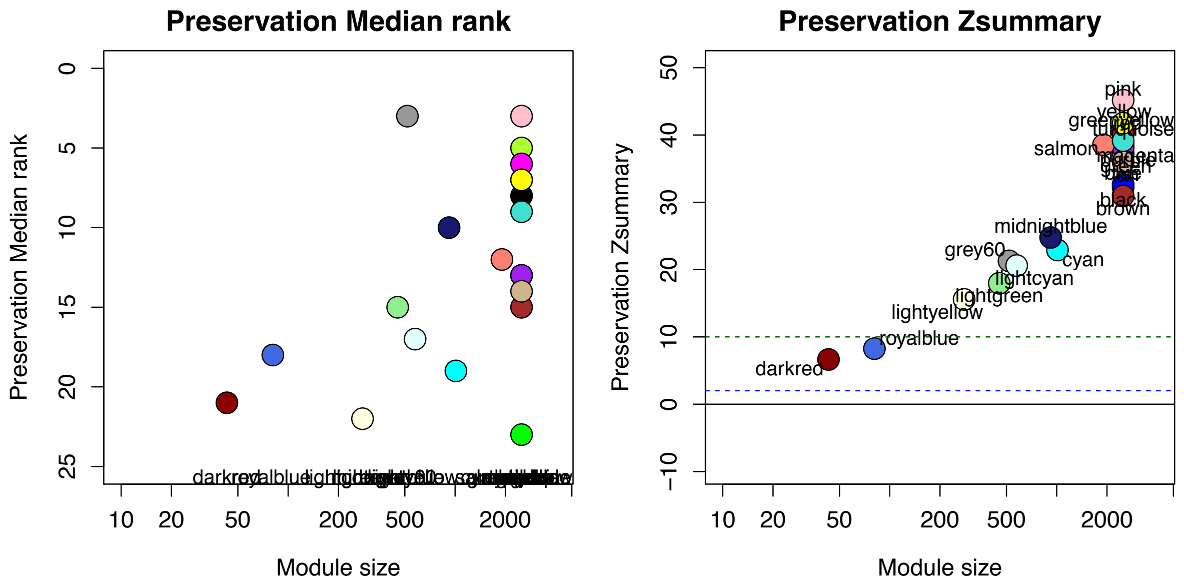


Figure 49: Z summary and median rank statistics plots illustrating the preservation of modules found at 7 years in the adolescents.

4.6 Discussion

4.6.1 Summary

This chapter has demonstrated that networks of highly co-methylated sites can be detected in a large population sample of children. Many of the modules have functional annotations, illustrating that these are likely to be modules of highly co-methylated DNAm sites that represent biological pathways that involve, or result in, DNAm. The preservation analysis shows that most of the network modules are stable over time, hinting that these co-methylation modules may represent persistent biological functions. The only phenotype robustly associated with the modules was asthma, at both 7 and 15-17 years (although most of this association was explained through cell type proportions). BMI was marginally associated with a module in adolescence, but the lack of functional annotations for this module make it hard to infer whether this module might be biologically meaningful. This may be due to the use of blood rather than potentially more relevant tissues for development (such as muscle or bone); it may be because there is no co-methylation involvement in the phenotypes; or it may be that exposures such as smoking are difficult to pick up when a small proportion of the population have been exposed.

4.6.2 Limitations of the study

There are a number of limitations that should be considered as part of this chapter. The first is that a co-methylation analysis is a ‘guilt by association’ method, which assumes connection between sites if they vary together. This is a fairly large assumption, although I feel that the work in chapter 3 demonstrated that strong correlations between DNAm sites are fairly stable and biologically meaningful. The WGCNA package has such a multitude of options that it is very possible I did not use the optimal settings for the network, and so did not identify the true co-methylation structure in the data. My tests (not included in this thesis) indicated that the networks were fairly robust to the parameters I selected, but I cannot discount the possibility that they were not optimal. This study also did not consider negative correlations in the network, which may be of importance to DNAm networks; the work in chapter 3 indicates that these correlations are likely to be different, and so these would certainly warrant further investigation, perhaps in a separate study. The association of some of the modules with cell counts (particularly the asthma-related modules) suggests that cell counts were likely imperfectly adjusted for; however I weighed this against potential over-adjustment of the data and removal of biologically important signals that may be related to the cell types. Indeed the asthma modules point to a potential functional DNAm pathway related to blood cell types.

An alternative approach to this analysis would be to use all three timepoints to build a single network, to interrogate modules that may be stable from childhood to adolescence, and identify those which may associate with a particular timepoint. This would have the advantage of clarity regarding the preservation of specific modules over time. However, as the majority of the individuals in ARIES are in all three timepoint datasets, and the repeated measures would undoubtedly be highly correlated for each individual, and it is possible this may cause problems with the clustering algorithm. Creating modules for each timepoint separately has the advantage of clearly defining all age-specific modules, and module preservation analyses are powerful tools to assess module preservation; they have the advantage of assessing how well connectivity is preserved. A WGCNA consensus module analysis would be an alternative tool to investigate the preservation of network modules over time.

A question which has not been addressed in either the correlation analysis in chapter 3 or the WGCNA analysis in this chapter is the role of negative correlations within the DNAm correlation structure. WGCNA signed networks exclude negative correlations, which was important in the context of investigating positive correlations in chapter 3. Further work examining the properties of negative correlations between DNAm sites could repeat the WGCNA analysis – either together with positive correlations in an unsigned network, or alone. The most appropriate WGCNA method would be determined by whether negative correlations display the same or different characteristics as positive ones – where they would be combined in an unsigned network if they have the same characteristics, and would be analysed in a separate network if they display different characteristics.

4.6.3 The impact of cell type proportions

Any DNAm analysis using tissues with multiple cell types is beset by questions of whether the effects seen are simply a consequence of cell type proportions. This seems particularly evident in studies using blood cells; however blood is an easily accessible tissue amenable to population studies, and is a tissue that would be likely to be used should useful biomarkers be detected. Thus, in my opinion blood remains a useful tissue to study. The impact of cell type proportions can be seen in the DNAm modules, most clearly in the modules associated with asthma at 7 and 15-17 years. However, as discussed below in section 4.6.4, this does not discount a mechanistic role of DNAm with these cell types.

4.6.4 Asthma associated modules

4.6.4.1 7 years

The MEsalmon module at 7 years is associated with asthma, and very strongly with eosinophil count. The GO terms for *carbohydrate binding* and *monosaccharide binding*, as well as the KEGG *asthma* pathway, are enriched in the strongest MEsalmon module members. DNAm has been shown to have a role in the control of glycosylation (which is the term for carbohydrate binding) (Klasic et al., 2016; Trinchera, Zulueta, Caretti, & Dall'Olio, 2014). Glycosylation of immunoglobulins (antibodies) has a role in allergic diseases (Epp, Sullivan, Herr, & Strait, 2016), and DNAm has been shown to have a role in the glycosylation of Immunoglobulin G (IgG), where glycosylation affects inflammatory state (A. Wahl et al., 2018). Nasal DNAm in glycosylation-related genes has also been associated with steroid treatment response in asthmatic patients (X. Zhang et al., 2017). This evidence suggests that the GO enrichments in the ME salmon module for *carbohydrate binding* and *monosaccharide binding* may highlight a mechanism by which this group of co-regulated DNAm sites could either have a role in, or be a biomarker of, the pathogenesis of asthma, or the response to steroid treatment. This could fit with the enrichment of the asthma KEGG pathway, where B cells form Immunoglobulin E (IgE) as part of the allergic response, and IgE stimulates mast cells to recruit eosinophils, which proliferate (Okuda et al., 2008). Thus it is possible DNAm is part of the mechanism of increased eosinophil counts, rather than just differential methylation due to increased cell counts; and even if it is not a regulatory factor, it suggests it could be used as a biomarker.

4.6.4.2 15-17 years

In the 15-17 year olds, the MElightgreen module is associated with asthma and strongly associated with eosinophils. It is enriched for GO terms relating to *interleukin-4 production*, and the KEGG *asthma* pathway. Interleukin-4 (IL-4) is a cytokine that increases the production of IgE, and increases eosinophilic inflammation in asthma (Chung, 2015). There is evidence that DNAm of the IL-4 gene is involved in the regulation of IL-4 production (Kwon, Kim, Lee, Oh, & Choi, 2008; Lee, Agarwal, & Rao, 2002; Tykocinski et al., 2005), which illustrates that this module may have a causal role within the asthma pathway.

4.6.4.3 Implications for asthma

It is unclear why the enrichments for the asthma-associated modules might be different at these two timepoints. It is possible it reflects changes in asthma, or the role of DNAm within asthma, with development. This might be reflected in the adolescent module having its effect size relatively unchanged by adjustment for eosinophils and neutrophils, whereas the module at 7 years was more affected by this. It may reflect the fact that these were slightly different phenotypes, with the adolescent measure having a different composite to that of the 7 year olds. It could reflect the fact that asthma has changing trajectories across life (Granell, Henderson, & Sterne, 2016; Panico, Stuart, Bartley, & Kelly, 2014; Sbihi, Koehoorn, Tamburic, & Brauer, 2017). Or it may represent batch effects that were in some way unaccounted for. The strength of this study is that the same individuals were used at both timepoints, so we can be confident it is not just an issue of sampling. An interesting future research question would be whether this module is specific to blood, or whether a more airway-adjacent tissue also demonstrates the same effect; the replication of some blood-based EWAS hits for asthma found in respiratory epithelial cells suggest it could be the case (Reese et al., 2019; Xu et al., 2018).

The negative association of both the MEsalmon and MElightgreen modules reflects consistent EWAS findings that DNAm at differentially methylated sites is lower in asthmatics than in controls (Xu et al., 2018). Differentially methylated sites have been demonstrated for asthma in numerous studies, but to my knowledge this is the first analysis to demonstrate that these sites may act as part of a pathway. Specifically altered DNAm in eosinophils has been shown in asthma (Xu et al., 2018), adding weight to the argument that there are specific methylation changes in asthma, rather than just different proportions of cell types.

4.6.5 Birthweight

A weak association of the MEtan module with birthweight leaves open the possibility that a pathway of co-methylation is somehow related to birthweight. However the lack of association between this module and functional annotations make it difficult to imagine what this might be; and so it may be that this module does not robustly associate with birthweight. It may be that blood is not an appropriate tissue to identify such a pathway related to physical development; however with such a large number of DNAm sites in blood associated with birthweight detected by (Kupers et al., 2019), it seems unlikely that WGCNA would not be able to identify a module of these sites if they did work as a pathway. It is possible that DNAm sites related to birthweight do not form part of a cohesive pathway, and are instead independently associated, or associated in multiple small independent pathways.

4.6.6 Notable functional annotations

At birth the MEblack module is enriched for GO terms relating to the nucleus, and the KEGG *pyruvate metabolism* pathway. As pyruvate has been shown to be metabolised to acetyl-CoA in the nucleus, and this is used for the acetylation of histones (Sutendra et al., 2014), there is a possibility this illustrates a pathway of histone modification in which DNAm is involved. This would be an interesting functional insight if it were investigated further;

however this does not seem to be well preserved across timepoints, so it could be a newborn-specific module, or it could be a module that is not stable.

The enrichment for the KEGG term *protein processing in the Endoplasmic Reticulum* (ER) in a module in all of the ARIES timepoints (MEturquoise at birth and 7, and MEblue at 15-17 years) is interesting because a previous study found that differential methylation at DNAm sites in genes associated with these GO terms were correlated with BMI, and to a lesser extent insulin resistance and fat mass, which they suggest may be to do with ER stress and a role of DNAm in genes which regulate the ER (Ramos-Lopez, Riezu-Boj, Milagro, Martinez, & Project, 2018).

The preservation of a module related to the GO term *Regulation of anatomical structure morphogenesis* through all three ARIES timepoints is interesting (MEturquoise at birth and 7 years; MEblue at 15-17 years), as it is possible this module reflects preservation of DNAm co-regulation in developmentally relevant genes. However it is not clear what their function would be in blood cells, or what sort of anatomical structures they might represent. They certainly do not appear to be related to anatomical development relevant to birthweight, gestational age, or BMI, as tested in this chapter. As the module seems to be strongly preserved, it would be interesting to follow up the genomic locations and functions of the hub module members, and map the network structure between them.

There are a number of immune related modules which may represent residual cell type effects that have not been corrected for by the Houseman algorithm. This makes sense because cell counts are predicted, and presumably there is some degree of residual left in the methylation data. These immune related modules actually give us the chance to investigate what residual there might be, and whether there is a module of residual cell type effects. If there is, this could be useful for future work looking to remove these residual cell type proportion effects.

4.6.7 Conclusions

To my knowledge this is the first study that has investigated pathways of DNA methylation from birth to adolescence. I have demonstrated DNAm modules that are persistent, with a number having biologically relevant functional annotations. I have demonstrated that asthma, which is found in around 10% of the sample, associates with an identifiable and biologically relevant module; I have also demonstrated that both own and maternal smoking, and gestational age, which have well established effects on DNAm, do not appear in this sample to impact DNAm as part of a concerted network. As many of the network modules indicate biologically relevant pathways, they are likely to lead to more informative and reliable biomarkers than single site analyses.

5 Chapter 5

Assessing the stability and reproducibility of DNAm correlation structure

5.1 Introduction

5.1.1 Summary

DNAm is a dynamic modification of the genome which is influenced by genetic factors, environmental exposures, and disease. It is also vulnerable to batch effects. As such, it is not certain that the correlation structure I have described in chapter 3 would be present in a different dataset. As the aim of this thesis is to identify stable and persistent relationships between DNAm sites, that may have a role in development, this chapter set out to address whether the correlation structures I have defined in Chapter 3 are preserved in a different cohort. As part of this aim, I also assess whether the correlation structure is preserved in another ethnicity, as trans-ethnic preservation of correlation structure would provide additional evidence of preservation.

5.1.2 Importance of assessing correlation structure in a different ethnicity

To date, the majority of epigenetic studies focus on individuals of European heritage. This is problematic, because there are differing rates of disease between ethnic populations (Anand et al., 2000; Keet et al., 2015; McWilliams et al., 2009; Tillin, Hughes, Godsland, et al., 2013), different ethnicities have differing genetic backgrounds (Genomes Project et al., 2015; International HapMap, 2005; International HapMap et al., 2010), and different ethnicities may also experience different environmental exposures and circumstances (Galanter et al., 2017; Nguyen et al., 2014; Tang, 2006). All of these factors contribute to variation in DNAm, and indeed many differences in DNAm exist between different ethnicities (as discussed in chapter 1, section 1.2.4). This is important because if there are disease mechanisms or biomarkers related to DNAm, it is possible they differ between ethnicities, and if we do not identify that it could result in healthcare inequalities. As I am attempting to establish whether correlations between DNAm sites can illuminate normative regulatory pathways DNAm might be involved in, it is important to assess whether these pathways might be the same for non-European populations; and using a multi-ethnic cohort can strengthen inference about the role of DNAm under differing environmental circumstances (Tang, 2006).

5.1.3 DNAm correlation structure and ethnicity

There has been limited prior work on how DNAm correlation structure might vary by ethnicity. Work by (Saffari et al., 2018) illustrated that *cis* DNAm correlation structure on chromosome 1 is consistent in both European and African populations. This suggests that *cis* correlation between DNAm sites is not primarily related to LD, as LD has been shown to decay over distances that vary distinctly between the ethnic groups included in the 1000 Genomes Project (Genomes Project et al., 2015). In this chapter I will extend the analysis of DNAm correlation structure to a comprehensive comparison between white British and Pakistani populations, running the analysis for all chromosomes, including measures of variance, the impact of *cis* genetic influence on the correlation structure, as well as an analysis of the differences in *trans* correlation structure between the two ethnicities.

5.1.4 Born in Bradford

I used the Born in Bradford (BiB) cohort to identify whether DNAm correlation structure is stable between different datasets. In the context of finding reproducible and fundamental connections between DNAm sites, validation in a second population that does not match the first population will increase generalisability of the results. I used Born in Bradford for the following reasons:

First, BiB is, like the Avon Longitudinal Study of Parents and Children (ALSPAC), a birth cohort. DNAm data is available for 1,000 children at birth, which allows a comparison of DNAm between the exact same time in life for ARIES (Accessible Resource of Integrated Epigenomic Studies) and BiB. This avoids confounding by age (Langevin et al., 2011; Teschendorff et al., 2013), and also avoids differences in DNAm being due to dynamic changes such as those that have recently been described in the brain across early childhood (Price et al., 2019).

Second, BiB is based in Bradford, a city with high levels of deprivation (Department for Communities and Local Government, 2015; Wright et al., 2013). Environmental exposures and lifestyle factors differ with deprivation, and as many of these have been shown to be associated with differences in DNAm (see introduction section 1.3.4 for a detailed discussion), it is likely the rates of some of these exposures will differ between ARIES and BiB. If correlation structure is replicable between these two cohorts, it might suggest persistent regulatory functions of DNAm co-methylation.

Third, BiB is a multi-ethnic cohort, as defined in chapter 2 section 2.2.2.2. As all of ARIES and half of BiB participants are white British, this enables me to identify whether correlation structure is stable in participants from the same ethnic group in slightly different environmental circumstances. Comparisons between the white British and Pakistani participants of BiB provides the opportunity to assess whether DNAm correlation structure persists in the context of differing cultural and genetic backgrounds. This will be interesting given the impact of genotype on DNAm correlations described in chapter 3 section 3.2.9.2.2.

5.1.5 Replication considerations

Replication in another study will reduce the likelihood of the results being due to batch effects, as these should not replicate across two separate studies. This is particularly important for DNAm array data because batch effects are prevalent (Leek et al., 2010; Teschendorff et al., 2011). It should be noted that DNAm data for ARIES and BiB were generated in the same laboratory, although years apart; so it may be that batch effects are more similar in these cohorts than those generated in different labs. Different array types were used (ARIES used the Illumina 450k array whilst BiB used the Illumina EPIC array); these arrays use the same technology, and so any array issues affecting DNAm (SNP in probe effects, probes with non-unique mapping, and probes with off-target hybridization (Zhou et al., 2017)) will replicate across cohorts. SNP in probe effects may also induce differences between ethnic groups due to differing genotypes.

To minimise these array issues, all DNAm sites which have these issues identified have been removed from the data used in this thesis; however this does not discount the possibility that other DNAm sites with these issues have not been identified and remain in the dataset. This could only be validated with a different technology, such as whole genome bisulfite sequencing, or reduced representation bisulfite sequencing; however the cost of these methods is relatively prohibitive for cohort studies.

5.1.6 Hypotheses

H.5.1. I hypothesise that correlation structure will replicate between the white British group and ARIES, as chapter 3 demonstrates that correlation structure looks relatively stable over time in ARIES.

H.5.2. I hypothesise that correlation structure will also replicate between the white British and Pakistani groups, based on results from (Saffari et al., 2018), where cis correlation decay looks similar for individuals of African and European ethnicities.

H.5.3. I hypothesise that correlations between DNAm sites represent meaningful co-regulation, and that therefore enrichment analyses will identify the same functions in BiB as they did in ARIES. If these are related to basic regulatory functions, they ought to be conserved across ethnicities.

1.1.1 Aims

A.5.1. To assess whether the distribution of correlations reported for ARIES are replicable, and whether there are differences in this structure in two ethnic groups.

A.5.2. To assess whether cis correlations are primarily driven by LD; if they are not, *cis* correlation decay will not differ between ethnic groups.

A.5.3. To replicate enrichment analyses from ARIES, to identify whether correlations between DNAm sites are likely to be driven by real biological pathways.

A.5.4. To identify whether strong correlations in ARIES can be replicated in both ethnicities in BiB.

5.2 Methods

5.2.1 Data: BiB

Born in Bradford (BiB) is a birth cohort following over 12,000 mothers and their children. The cohort is described in detail in chapter 2 (section 2.2.2). Briefly, a subsample of 1,000 mothers and their children were selected to have DNAm data generated from blood samples provided by the mothers during pregnancy, and from cord blood from the children. The subsample was specifically selected to have 500 participants of white British origin, and 500 participants of Pakistani origin. DNAm was measured using the Illumina Infinium MethylationEPIC BeadChip array. Because I used BiB as a validation for ARIES (which was measured by the Illumina 450k array), and also because of computing considerations, I subset the EPIC data to DNAm sites which are also present on the 450k array.

5.2.1.1 Adjusting for population stratification and relatedness

After normalisation, using genetic data I removed individuals related above 0.125 and regressed out population stratification (using 20 genetic PCs) in the white British and Pakistani groups separately. These steps are described in detail in chapter 2 section 2.2.2.5, and the number of participants remaining after removing related individuals can be found in Figure 50.

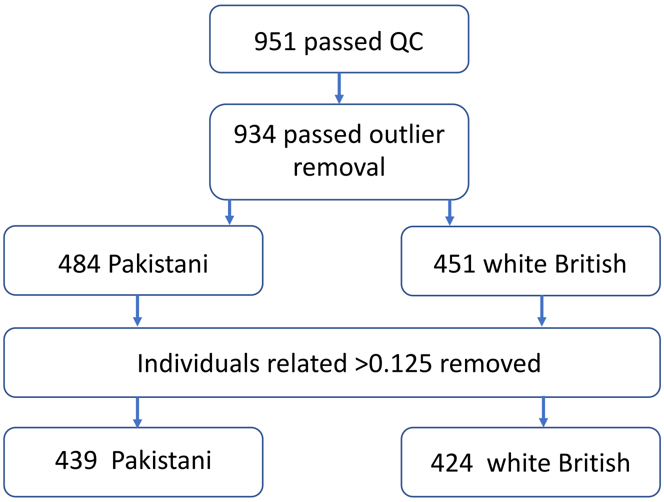


Figure 50: Flow diagram of BiB child participant numbers

5.2.1.2 Adjustments

I filtered the sites that were also present on the 450k array, as detailed in chapter 2, section 2.2.2.7.1. Briefly, all non-autosomal probes, probes in the HLA region, poorly functioning probes and probes containing SNPs were removed from the dataset. This left 369,796 probes for analysis. As both ethnicities were normalised together, the same probes were included for both groups. All outlying values (> 10 standard deviations from the mean) were replaced with the mean for that probe, over three iterations.

As discussed in chapter 2 section 2.2.2.7.3, sex, estimated cell counts, and the BeadArray on which a sample was run were regressed out using a linear model. A random effects model was not used for BeadArray because batch effects were removed as a random effect during normalisation, but BeadArray was still associated with methylation PCs after normalisation.

5.2.2 Data : GoDMC

The mQTL data used in this section was generated by the GoDMC consortium. It is described in detail in chapter 2, section x. Briefly, the consortium data consisted of 27,750 participants of European ancestry from 36 cohorts. Each cohort identified mQTLs below the threshold of p<1e-5 individually, and mQTLs from all cohorts were combined to make a unique list. This unique list was then tested for association in every cohort, and the results were meta-analysed. The data used in this thesis was the association of DNAm sites with a SNP at the threshold $p<10^{-8}$ for cis mQTLs and $p<10^{-14}$ for trans mQTLs.

5.2.3 Data: Heritability and environmental influences on DNAm

The relative contributions of heritability, common environment, and unique environment have been estimated for all DNAm sites on the 450k array by (Hannon et al., 2018). These estimates were constructed using twin data, and have been made available as a searchable dataset online (http://www.epigenomicslab.com/online-data-resources). The full dataset was kindly provided to me by the authors.

5.2.4 Compute resources

The following analyses were run on a high memory server, running Ubuntu 18.04, using R version 3.6.0. Relevant R packages are detailed, as appropriate, throughout the thesis.

5.2.5 Correlation statistic

I used the biweight mid-correlation from the R package `WGCNA` (Langfelder & Horvath, 2008) as a correlation statistic, as pearson correlation is not suitable for bimodal distributions (which can be the case for some DNA methylation sites). I have demonstrated that the biweight mid-correlation is a suitable, more efficient alternative to spearman in chapter 2 (section 2.3).

5.2.6 Creating the correlation matrix and extracting pairwise correlations for analysis

As some probes on the 450k array were not carried over to the EPIC array, the matrix of 369,796 x 369,796 gave 68,374,355,910 unique correlations (when the diagonal is removed). To create the correlation matrix, I followed the same methods as in chapter 2 section 2.5.1. Briefly, DNAm sites were split into blocks of 25,000 and all blocks were correlated against each other. Dataframes of correlating pairs were then created for each 0.1 band of correlation (20 bands between -1 and 1), so I could conduct analyses of the features of correlations of different strengths (i.e., do high correlations differ from low correlations?).

5.2.7 Plotting the full distribution of correlation values

To plot the distribution of correlation values between all DNAm sites on the EPIC array which are also present on the 450k array, I ran the analysis described in detail in chapter 2 section 2.5.2. I ran this for both BiB ethnic groups.

5.2.7.1 Proportions of cis and trans correlations

To plot the distribution of *cis* and *trans* correlations I plotted percentages in each range of correlation, because there are so many more *trans* correlations than *cis* that it is hard to interpret plots with absolute numbers. To do this, I ran the analysis described in chapter 2, section 2.5.2.1, for each of the BiB ethnic groups.

5.2.8 Illustration of cis correlation structure across the genome

To assess *cis* correlation structure, I produced decay plots of *cis* correlations on each chromosome separately. For each plot, I used correlations that were within 10kb of each other, based on previous literature. I also produced a histogram of correlations within 1kb, for a clearer illustration of the correlation values in close proximity. I did this for each of the two BiB groups. The details of creating these plots can be found in chapter 2, section 2.5.3.

5.2.9 Genetic influences on correlations between DNAm sites

I used three complementary approaches to investigate the influences of genetics on correlating sites.

5.2.9.1 Influences of heritability estimates on DNAm sites

To assess the impact of heritability on DNAm correlations, I took the estimates of heritability and environmental influences on DNAm created by (Hannon et al., 2018), and used it to estimate the proportion of sites in each correlation band that were influenced by genetic, unique environmental, and shared environmental factors. I used the methods described in chapter 2 section 2.5.4.1 to produce ridgeline plots to illustrate this.

5.2.9.2 Influence of mQTLs on correlations between DNAm sites

Another way to assess the influence of genetics on DNAm is to identify whether the level of DNAm is associated with a genetic variant (mQTL). The most comprehensive analysis identifying mQTLs is the GoDMC consortium’s analysis (described above in section 5.2.2).

One limitation to bear in mind is that the mQTLs were generated in individuals of European ancestry, and so some mQTLs may not be appropriate for the Pakistani group in this analysis.

5.2.9.2.1 Plotting the proportion of correlating DNAm sites associated with mQTLs

To initially illustrate how mQTLs might drive correlations between DNAm sites, I identified whether, for each correlating pair, neither, one or both of the DNAm sites were associated with an mQTL. I did this for the DNAm sites in each correlation range, to illustrate the distribution of mQTLs across values of correlation. Please note that this does not identify whether both DNAm sites are associated with the same mQTL. The code I used to do this and generate plots illustrating the association of *cis*- and *trans*-correlating DNAm sites is found in chapter 2 section 2.5.4.2.1.

5.2.9.2.2 Removing genetic influence from cis correlation plots

To then illustrate some of the impact these mQTLs actually have on correlations between DNAm sites, I adjusted the cis correlation decay plots for the strongest *cis* mQTL associated with each DNAm site, thereby removing some of the genetic influence on DNAm correlations. For this analysis, I used chromosome 20 as an example. The details of this analysis can be found in chapter 2, section 2.5.4.2.2.

5.2.10 Analysis of strong correlations

5.2.10.1Genomic region enrichment

To identify whether DNAm sites which form strong correlations overlap with genomic sites of interest, I used the locus overlap R package `LOLA` (Sheffield & Bock, 2016). `LOLA` assesses enrichment based on genomic regions rather than genes, which is advantageous for DNAm analyses because DNAm sites are not necessarily functionally linked to their nearest gene. I tested *cis* and *trans* correlations r>0.9 separately for genomic region enrichment, using the list of all 369,796 sites in the analysis as the background. I used two region sets created by the LOLA team, available through http://lolaweb.databio.org: the ENCODE transcription factor binding sites (J. Wang et al., 2012), and Cistrome histone marks (Q. Wang et al., 2014). I also a region set generated by Josine Min containing the chromHMM imputed 25 chromatin states from Roadmap Epigenomics (Ernst & Kellis, 2015;

Roadmap Epigenomics et al., 2015). The details for this analysis are in chapter 2, section 2.5.5.1.

5.2.11 Trans correlation structure

5.2.11.1Visualising trans correlation structure

5.2.11.1.1 Circos plots

Circos plots were generated to visualise the distribution of strong trans correlations across the genome. The details of producing these plots can be found in chapter 2, section 2.5.6.1.1.

5.2.11.1.2 Cytoscape

Cytoscape plots were generated to visualise the connectedness of strong trans correlations, using Cytoscape version 3.6.1 (Shannon et al., 2003). The detail of this analysis is in chapter 2 section 2.5.6.1.2.

5.2.12 Consistency in correlations between ARIES and BiB

I finally conducted analyses to assess the preservation of the strong correlations between ARIES and BiB, and between the two ehtnicities in BiB. This final step of the analysis is critical because it will illustrate whether strong correlations are replicable. If they are not, it would not be clear whether strong correlations are spurious, cohort-specific, or related to batch effects. To assess the preservation of high correlations between ARIES and BiB, and between the two ethnic groups in BiB, I created mean difference plots. Mean difference plots simply take two measurements to be compared, such as the correlations for probe pairs in ARIES and in BiB. For each pair of DNAm sites, the mean of the correlation of the two groups is plotted against the difference in correlation between the two groups. 95% confidence intervals are calculated using the difference, illustrating whether the correlation differs more between the two groups than expected by chance (Ritchie et al., 2015).

5.3 Results

5.3.1 Plotting the full distribution of correlation values

To illustrate the full distribution of pairwise correlations between all sites on the EPIC array which also feature on the 450k array, I plotted the distribution of correlations in bins of 0.1. I have done so for both BiB ethnic groups, and I compare them to ARIES at birth.

The distribution of correlations in BiB looks very similar between the two ethnic groups. When we compare the distributions to ARIES, it is broadly similar, with a positive skew, and the greatest proportion of correlations lie between -0.2 and 0.2; 85.8% in the white British group and 87% in the Pakistani group. This is a marginally higher proportion than the birth timepoint in ARIES (85%). This is illustrated in Figure 51. The other notable feature of this distribution as compared to ARIES is the higher number of strong correlations, >0.8 and <-0.8. Although a very small difference proportional to the absolute number of correlations, there are 676% more correlations >0.8 in the BiB white British group than ARIES at birth, and 755% more in the BiB Pakistani group than in ARIES at birth. There are 6141% more negative correlations <-0.8 in the white British group, and 2310% more in the Pakistani group. The number of correlations in each band in the two BiB groups, and their comparison to ARIES at birth, are detailed in Table 20.

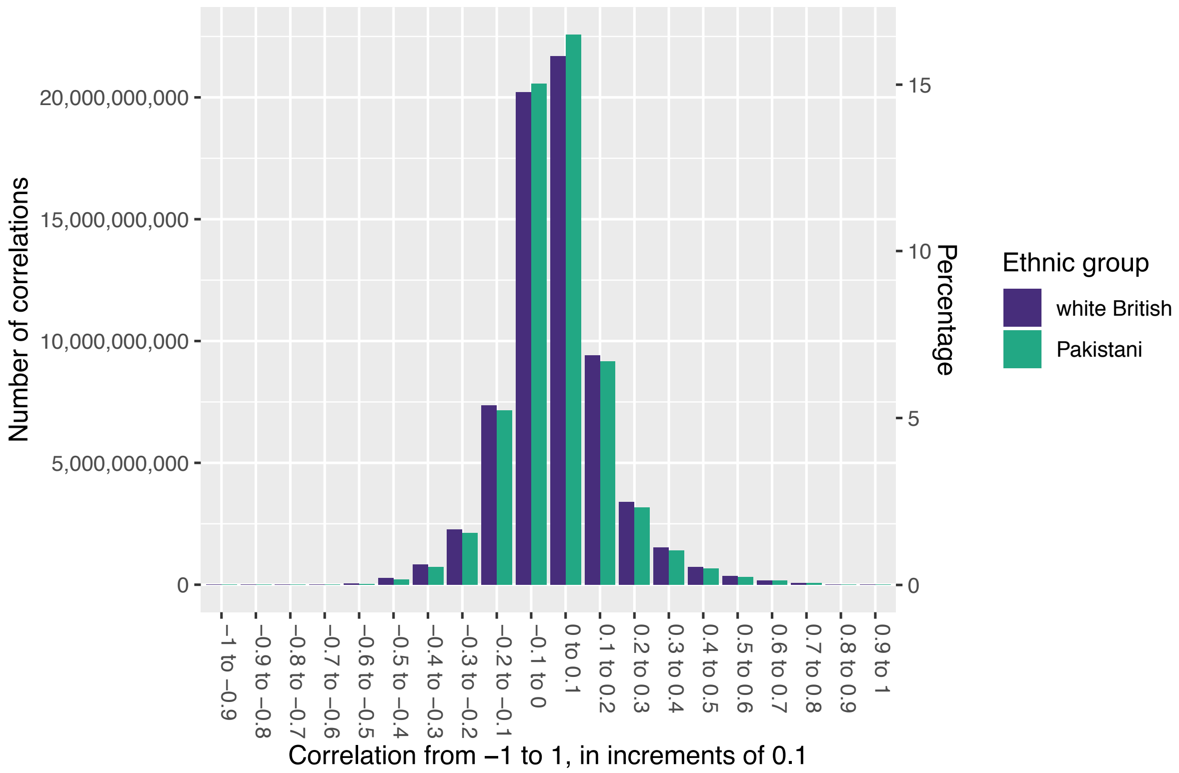


Figure 51: Distribution of pairwise correlations in BiB white British (purple) and Pakistani (green) children at birth.

Correlation band	ARIES (birth)	Percentage of total	BiB white British	Percentage of total	BiB Pakistani	Percentage of total
-1 to -0.9	3	3.8E-09	861	1.1E-06	298	3.8E-07
-0.9 to -0.8	404	5.2E-07	24133	3.1E-05	9102	1.2E-05
-0.8 to -0.7	4513	5.8E-06	176251	0.0002	81175	0.0001
-0.7 to -0.6	923216	0.001	2732748	0.004	882784	0.001
-0.6 to -0.5	43011431	0.06	53197120	0.07	26774426	0.03
-0.5 to -0.4	317547894	0.4	282312607	0.4	208404726	0.3
-0.4 to -0.3	1002028058	1.3	828914090	1.1	729821233	0.9
-0.3 to -0.2	2596753505	3.3	2273035752	2.9	2131484707	2.7
-0.2 to -0.1	8157470531	10.5	7365616520	9.4	7147647197	9.2
-0.1 to 0	21506999111	27.6	20208436245	25.9	20574743480	26.4
0 to 0.1	23598290881	30.3	21692606987	27.8	22571786071	29
0.1 to 0.2	12955550222	16.6	9406166398	12.1	9173828159	11.8
0.2 to 0.3	4602968925	5.9	3399251998	4.4	3168006248	4.1
0.3 to 0.4	1878135882	2.4	1523517117	2	1398634386	1.8
0.4 to 0.5	822139439	1.1	723366575	0.9	666847937	0.9
0.5 to 0.6	337068287	0.4	350423850	0.4	323252927	0.4
0.6 to 0.7	107032881	0.1	179616536	0.2	168087119	0.2
0.7 to 0.8	22579159	0.03	75497920	0.1	73493301	0.09
0.8 to 0.9	1400404	0.002	9458343	0.01	10565416	0.01
0.9 to 1	315	4E-07	3859	5E-06	5218	6.7E-06

Table 20: Numbers of correlations in each band from -1 to 1 in, in ARIES at birth, and the two ethnic groups in BiB at birth

5.3.1.1 Proportions of cis and trans correlations

Examining the proportions of *cis* and *trans* correlations in each correlation band can help to illustrate whether physical proximity between DNAm sites has an effect on the likelihood of them having correlated methylation states. I have defined *cis* as within 1Mb. In BiB, The pattern is almost identical for both ethnicities. There are relatively equivalent proportions of *cis* and *trans* correlations for low to moderate correlations, between -0.5 and 0.6, with just a slight increase in *cis* correlations 0 to 0.1. There are marginally more negative *trans* correlations; and there are slightly more *cis* correlations between 0 and 0.2. These patterns are illustrated in Figure 52. Because such increasingly small proportions of correlations are <-0.5 and >0.5, they are best viewed in a table; Table 21 shows that a much smaller percentage of *trans* correlations tend to be found between 0.9 and 1. At birth there are no *cis* correlations between -1 and -0.9. Table 21 also shows that percentages are generally similar to ARIES, except they illustrate that the substantial increase in correlations >0.9 (discussed in section 5.3.1 above) is due to an increased number of high *trans* correlations in BiB.

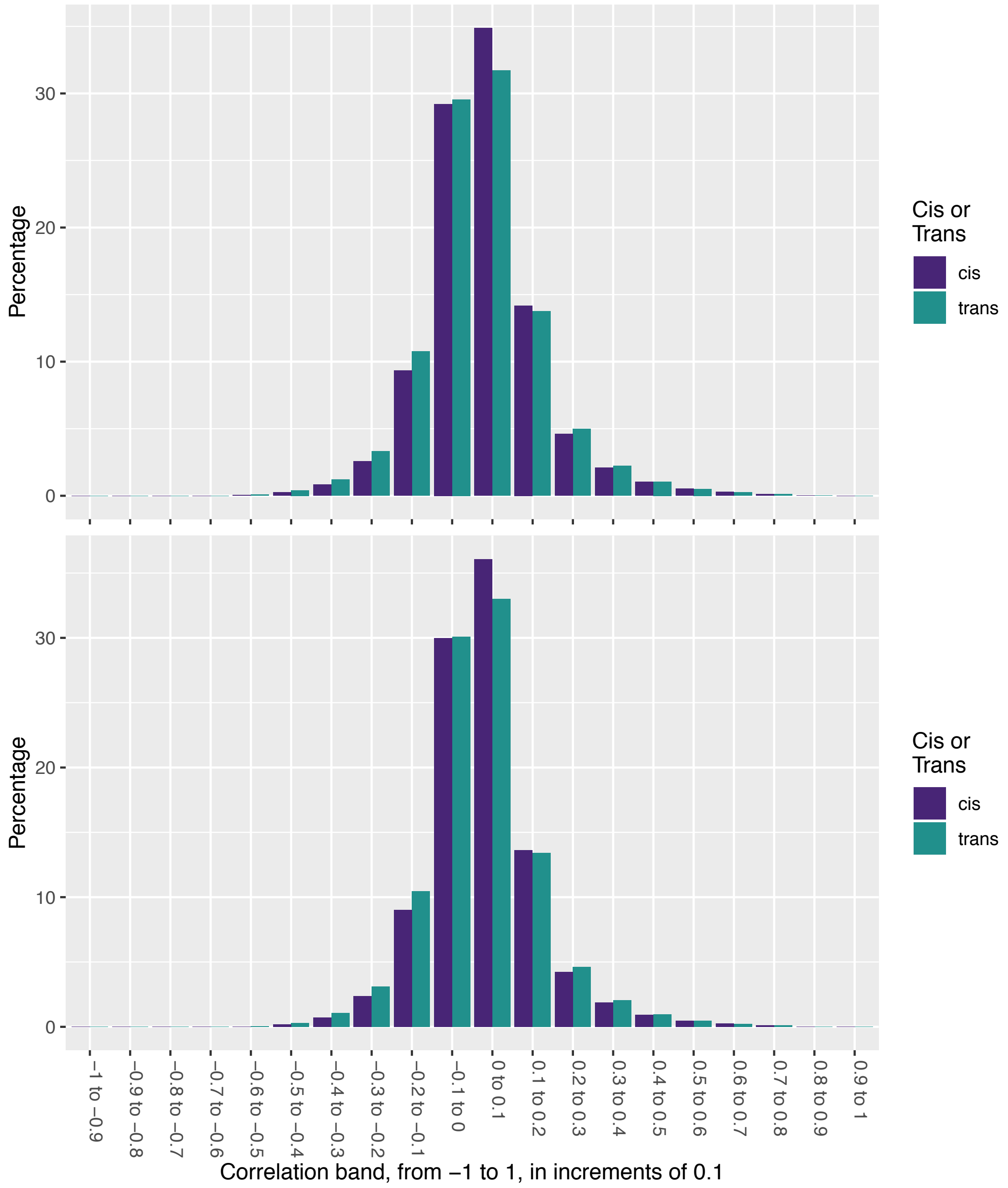


Figure 52: Percentages of cis and trans correlations in each correlation band, in the BiB white British (top) and Pakistani (bottom) ethnic groups.

	ARIES (Birth)		BiB white British		BiB Pakistani	
Correlation band	Cis	Trans	Cis	Trans	Cis	Trans
-1 to -0.9	0	3.85E-09	0	1.3E-06	0	4.4E-07
-0.9 to -0.8	2.5E-06	5.2E-07	4.3E-05	3.5E-05	1.5E-05	1.3E-05
-0.8 to -0.7	3.6E-05	5.7E-06	0.0004	0.0003	0.0002	0.0001
-0.7 to -0.6	0.0008	0.001	0.003	0.004	0.001	0.001
-0.6 to -0.5	0.03	0.06	0.05	0.08	0.02	0.04
-0.5 to -0.4	0.23	0.41	0.26	0.41	0.18	0.3
-0.4 to -0.3	0.8	1.29	0.83	1.2	0.7	1.1
-0.3 to -0.2	2.4	3.3	2.6	3.3	2.4	3.1
-0.2 to -0.1	8.7	10.5	9.3	10.8	9	10.5
-0.1 to 0	25.9	27.6	29.2	29.6	30	30
0 to 0.1	34	30.3	34.9	31.8	36.1	33
0.1 to 0.2	18.7	16.6	14.2	13.8	13.6	13.4
0.2 to 0.3	5.7	5.9	4.6	5	4.3	4.6
0.3 to 0.4	2.1	2.4	2.1	2.2	1.9	2
0.4 to 0.5	0.89	1.1	1	1.1	0.9	1
0.5 to 0.6	0.38	0.43	0.52	0.51	0.5	0.47
0.6 to 0.7	0.14	0.14	0.28	0.26	0.27	0.25
0.7 to 0.8	0.04	0.03	0.13	0.11	0.13	0.1
0.8 to 0.9	0.003	0.002	0.02	0.01	0.02	0.02
0.9 to 1	0.0002	9.9E-08	0.0003	5.2E-06	0.0003	7.3E-06

Table 21: Table of the percentages of cis and trans correlations in each correlation band, comparing ARIES at birth to both BiB ethnic groups.

There are a number of possible reasons for the higher number of correlations >0.9 in BiB compared to ARIES. It is possible that all samples being whole blood may mean there is greater variation in the data, and consequently a greater ability to detect higher correlations. It may also be that because BiB is a smaller sample, correlations are less precise and so erroneously high. It is also possible that this difference is due to the different array platforms. Although the 450k and the EPIC use the same array technology and they measured methylation at the same sites, there are some differences between the arrays; all highly correlating sites in BiB used the same dye and same probe type on both the arrays, but other differences in the arrays may remain.

5.3.2 Illustration of cis correlation structure across the genome

To illustrate cis correlation structure, I created decay plots for each chromosome. I separated out positive and negative correlations, to identify whether they differ in terms of structure; and I added variance to the plot to demonstrate the uncertainty around the binned estimates.

There is no discernible difference in correlation structure between the two ethnic groups in BiB, as is clear from the example of chromosome 1 in Figure 53. This is the same finding for all autosomal chromosomes; for plots of chromosomes 1:5 and 15:19, for both ethnic groups in BiB, please see Appendix 5. The plots also look identical to those in ARIES, which is included in Figure 53 for comparison. As in ARIES, the mean positive correlation at immediately adjacent sites is around 0.4, and reduces to a constant of around 0.125 by around 1kb. As discussed in chapter 3 section 3.3.2, this is similar to what was found in (Y. Liu et al., 2014) and (Saffari et al., 2018). In addition to the literature, my work illustrates the variance in these cis correlations, and that this does not differ between ethnicities; it provides a validation of correlation structure in a moderate sample size of Europeans of the same age; and shows that this structure is stably repeated over all autosomal chromosomes.

The decay of negative correlations, as in ARIES, is fairly constant at around -0.1, and does not seem to change with genomic distance. This indicates that negative correlations are not based on proximity in the DNA sequence the same way that positive correlations are; although Table 20 shows that there are many strong negative cis correlations within 1Mb, these clearly do not function in the same way that positive correlations do.

The histograms of *cis* correlation values within 1kb illustrate the presence of high correlations a bit more clearly (see Appendix 5 with the *cis* decay plots; also see Figure 54 and Figure 55 below for chromosome 1 examples for both ethnicities). Although there are small numbers of high (>0.8) correlations (around 0.3%), there is a greater proportion of them within 1kb compared to the distributions of correlations r2>0.8 genome-wide (around

0.01%) (as illustrated in Figure 51 and Figure 52). Both *cis* and genome-wide proportions of correlations r>0.8 are higher than in ARIES (see chapter 3, section 2.5.2).

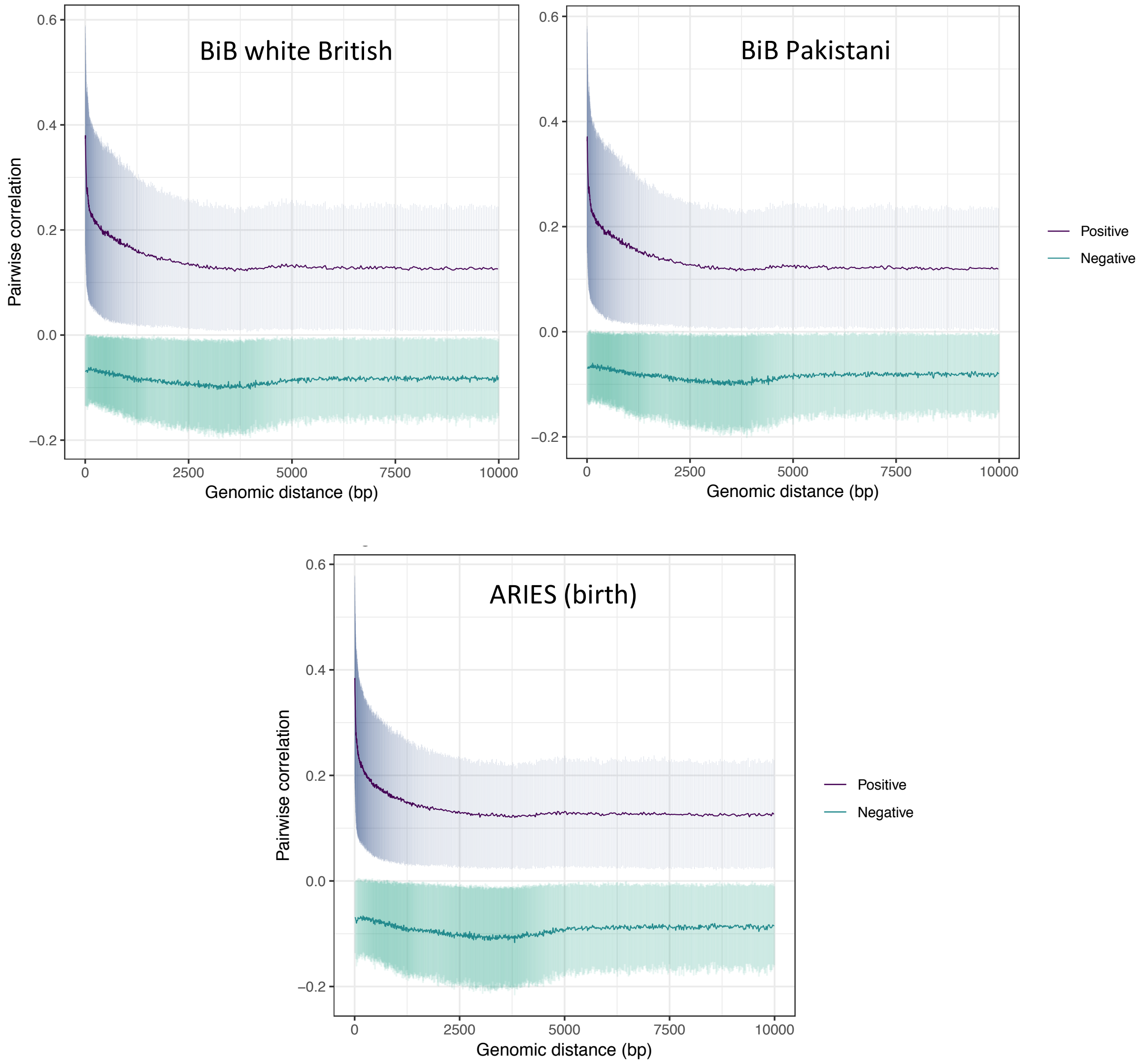


Figure 53: Decay plots of cis correlations, genome-wide. In BiB white British (top left) and BiB Pakistani (top right) ethnic groups, and in ARIES at birth (bottom)

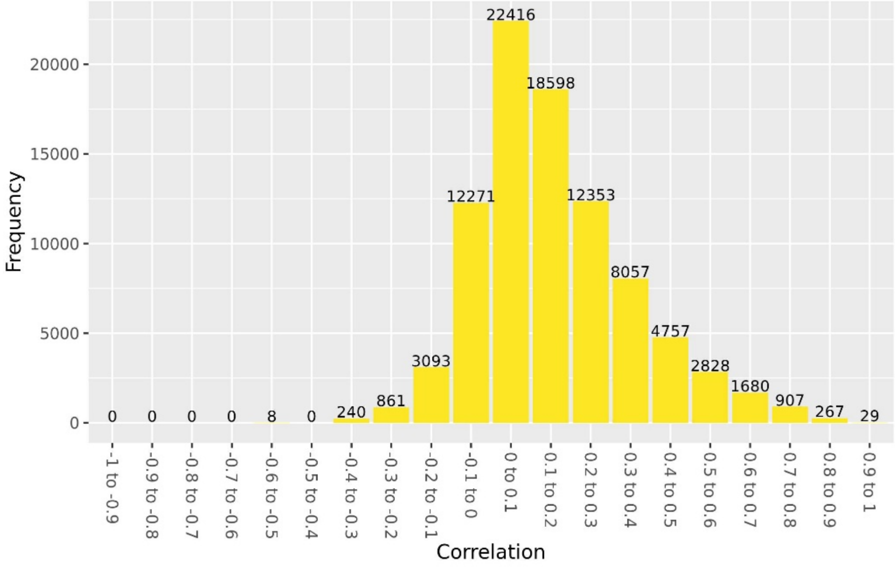


Figure 54: Histogram of correlation values between all probes within 1kb of each other, on chromosome 1 in the BiB white British group.

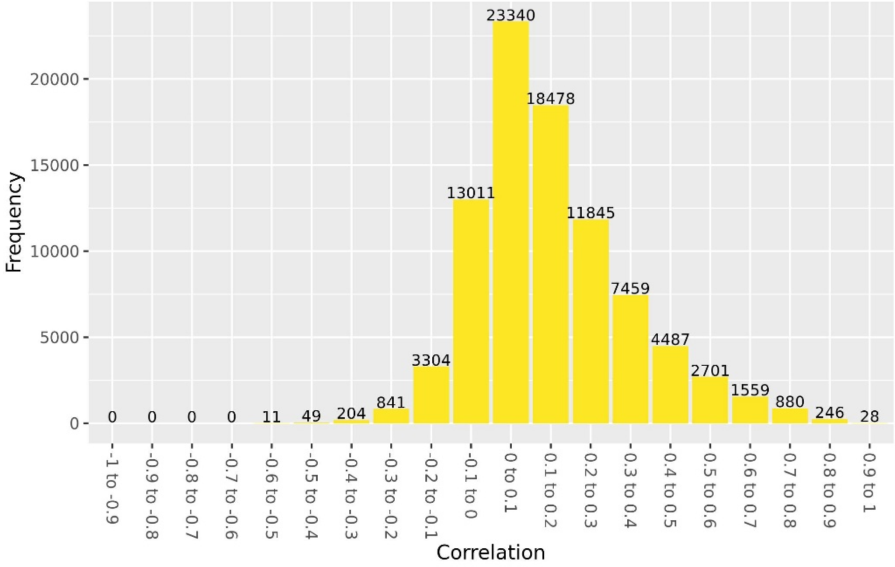


Figure 55: Histogram of correlation values between all probes within 1kb of each other, on chromosome 1 in the BiB Pakistani group.

5.3.3 Genetic influences on correlations between DNAm sites

5.3.3.1 Influences of heritability estimates on DNAm sites

Heritability has been shown to explain a smaller proportion of the variation in DNAm levels than environmental influences at most DNAm sites, and sites at which heritability contributes to more of the variation tend to be those associated with traits of interest to epidemiological studies (Hannon et al., 2018) and tend to be stable over time (Shah et al., 2014). To assess whether heritability is also the main driver of variation in highly correlated DNAm sites, I assessed the relative contributions of heritability and environmental influences on variation of the DNAm sites that feature in the 20 bands of correlation that range from -1 to 1.

I found that the patterns of contribution of heritability and environmental influences to DNAm variation were the same in both BiB ethnic groups. In contrast to ARIES, probes with correlations >0.9 predominantly had a very low contribution of heritability to their variance. There is also a small peak in heritability around 1 for these highly correlated sites, so it is possible that due to the far greater number of probes which have correlations >0.9 in BiB, there are simply also sites with low heritability included in this range of correlations. As the ARIES 7 year olds also had a peak around 0 for heritability in this band of correlation, and the 7 year olds had many more high correlations than the other ARIES timepoints, this may explain the result. Sites with correlations -0.8 to -1 have contributions of heritability quite evenly distributed from 0 to 1.

There is also a greater contribution of common environment for a subset of DNAm sites with correlations >0.9, with a peak around a contribution of 0.5. DNAm sites with correlations in the other ranges are also spread between 0 and 0.5. This suggests that correlations between DNAM sites >0.9 may be under slightly more environmental influence than we see in ARIES. Unique environment is the major source of variance for all ranges of correlation, but accounts for the least variability in DNAm sites with strong correlations, >0.9 and -0.7 to -1. This is illustrated in Figure 56 for the white British group and Figure 57 for the Pakistani group.

It seems there is much lower heritability of probes with correlations >0.9 in BiB than there is in ARIES. As there is a peak of the distribution around 1, it may be that those are the sites that are also highly heritable in ARIES; which corresponds to the cis-correlating sites, as is evident from Figure 18 (when comparing heritability of highly correlating probes in ARIES). As there are far more trans-correlating sites >0.9 in BiB, it seems likely that these sites form the peak of heritability around 0.

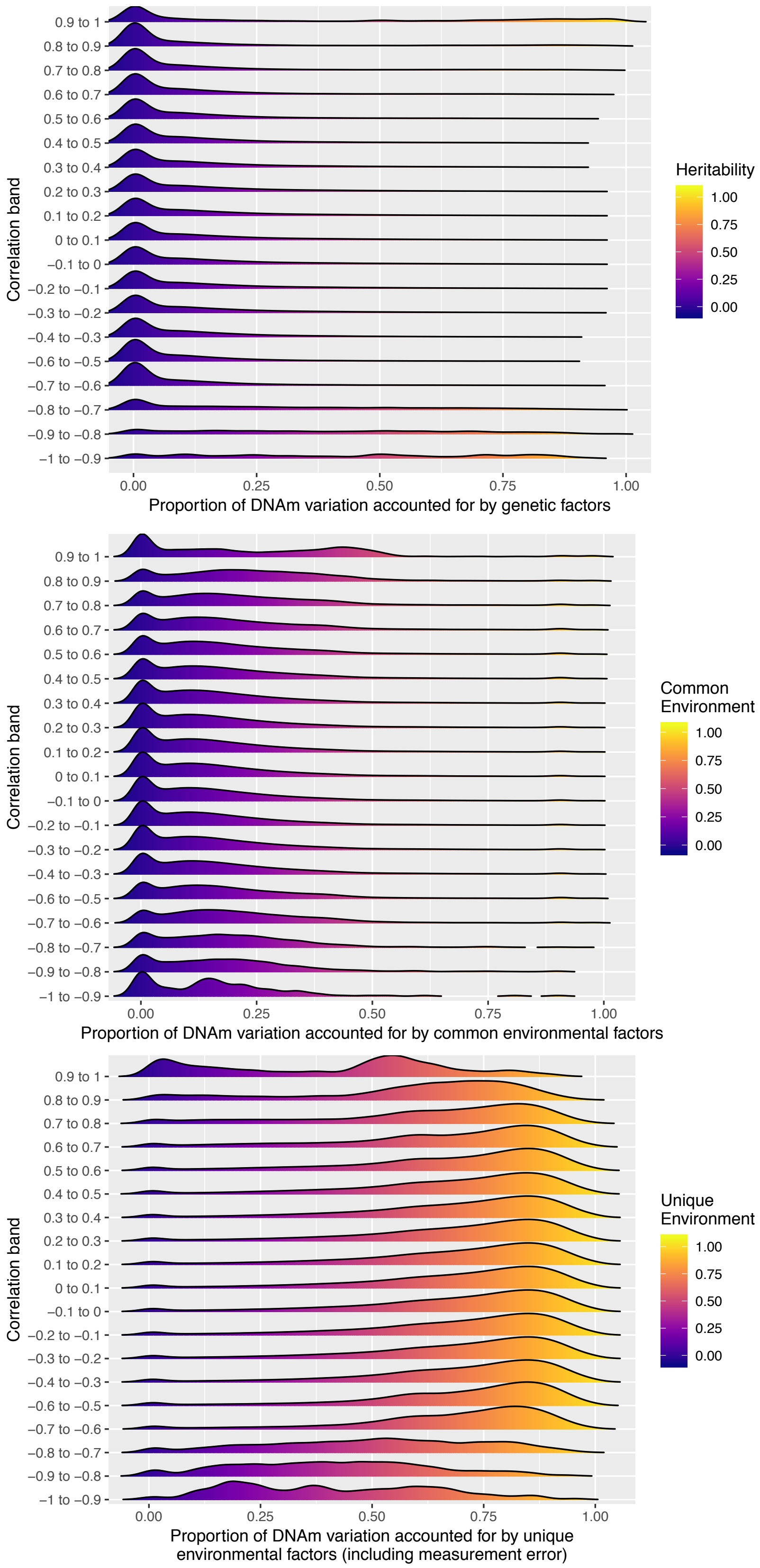


Figure 56: Ridgeline plots illustrating the estimated contributions of genetic and environmental factors to variation in DNAm sites which feature in 20 ranges of correlation strength, at birth in BiB white British participants at birth.

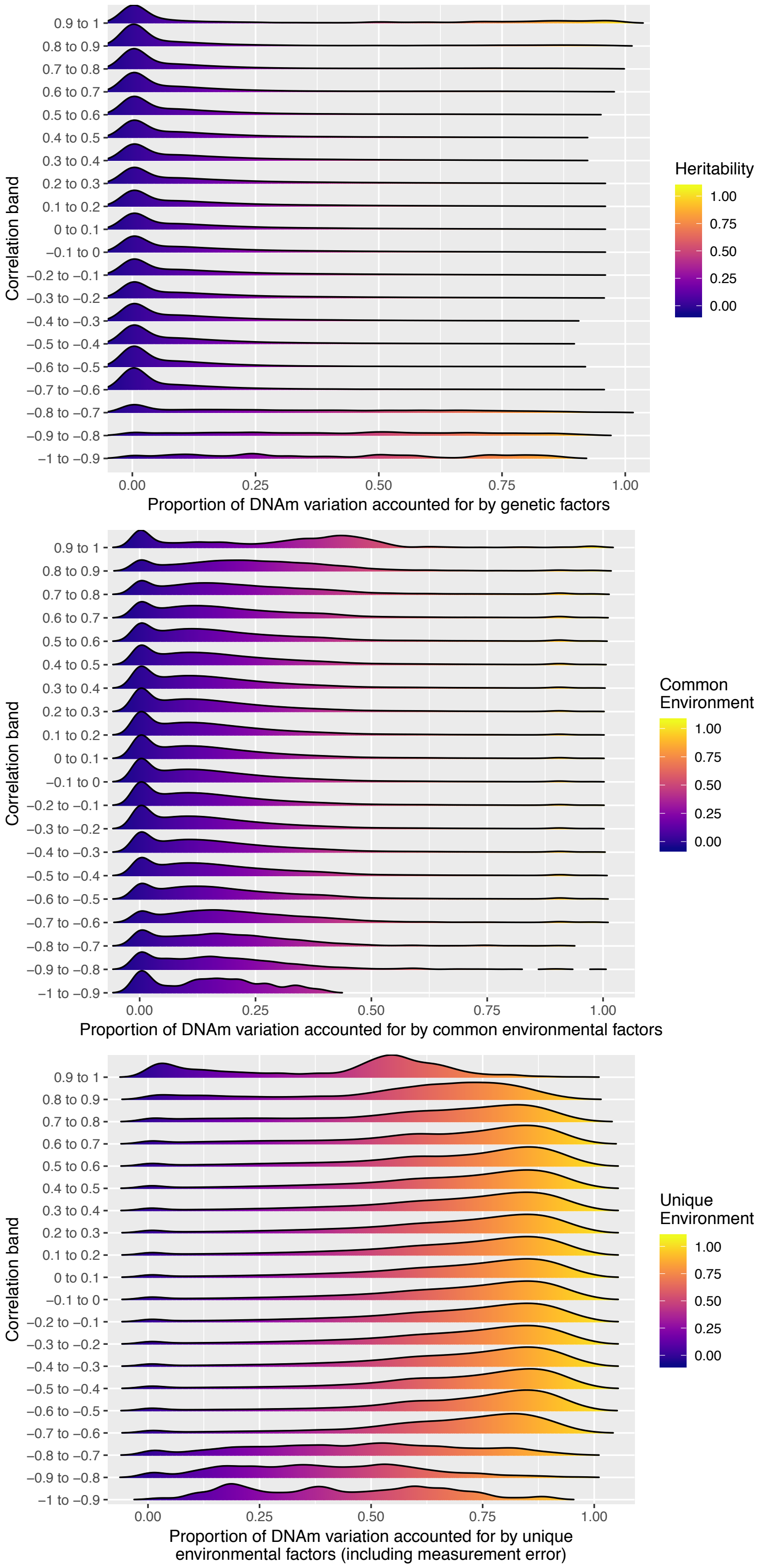


Figure 57: Ridgeline plots illustrating the estimated contributions of genetic and environmental factors to variation in DNAm sites which feature in 20 ranges of correlation strength, at birth in BiB Pakistani participants at birth.

5.3.3.2 Plotting the proportion of correlating DNAm sites associated with mQTLs

To identify whether strong correlations might be driven by known mQTL associations, and whether this might differ for cis and trans correlations, I plotted the percentage of correlations with 0, 1 or 2 of the DNAm sites associated with an mQTL, stratified by correlation strength. I plotted this for both ethnic groups of BiB to illustrate whether this might differ between ethnicities; although it should be noted that mQTLs generated from data from European individuals may not be appropriate for other ethnicities.

I find that strong *cis* correlations (-1 to -0.7 and 0.9 to 1) have around 100% of DNAm sites in each correlating pair associated with an mQTL. The percentages are very similar between the ethnicities, aside from an absence of cis correlations -0.1 to -0.9 in the white British group. The strong negative *trans* correlations (-1 to -0.7) are also most likely to have both DNAm sites in a correlating pair associated with an mQTL. The strong positive *trans* correlations are more likely to have associated mQTLs than in ARIES in the white British group, but not the Pakistani group; this may be induced by the greater number of strong positive correlations in BiB. This is illustrated in Figure 58.

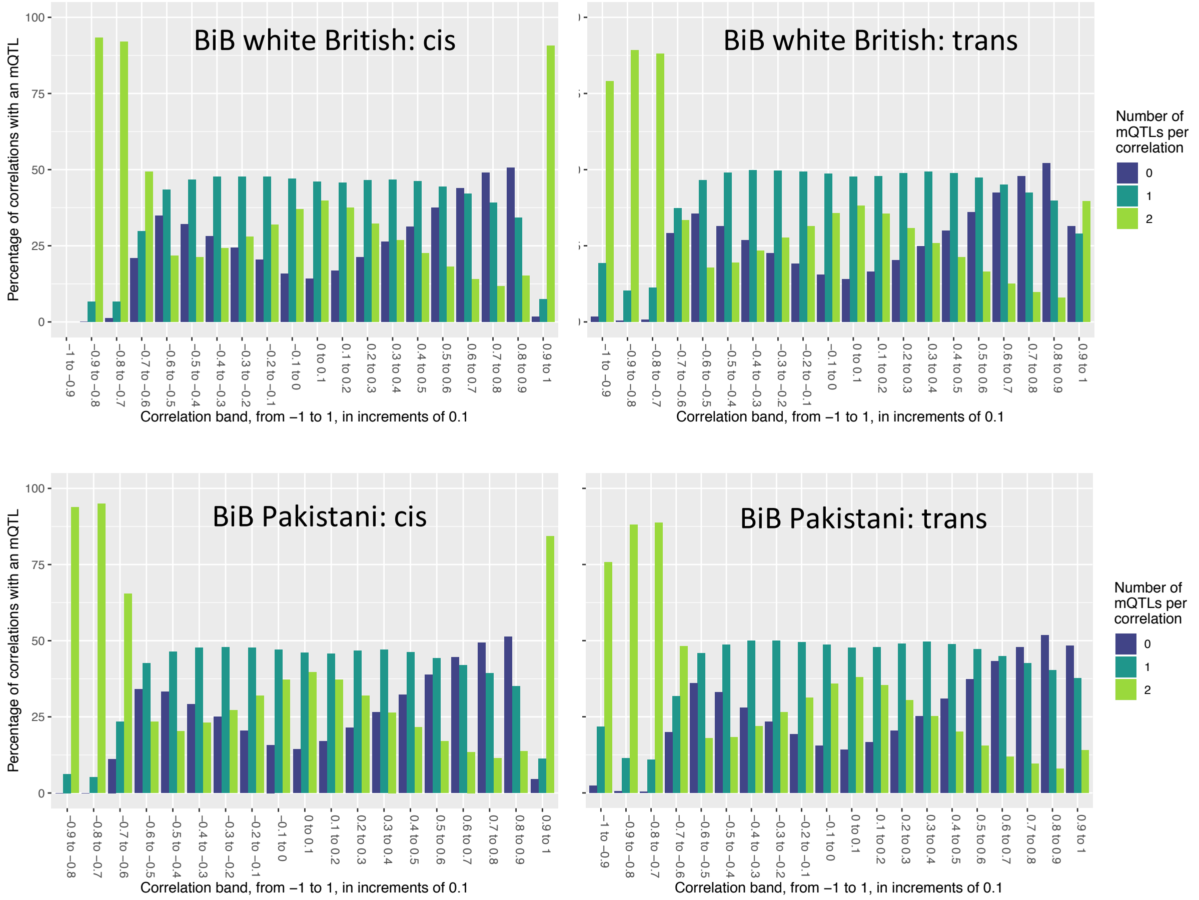
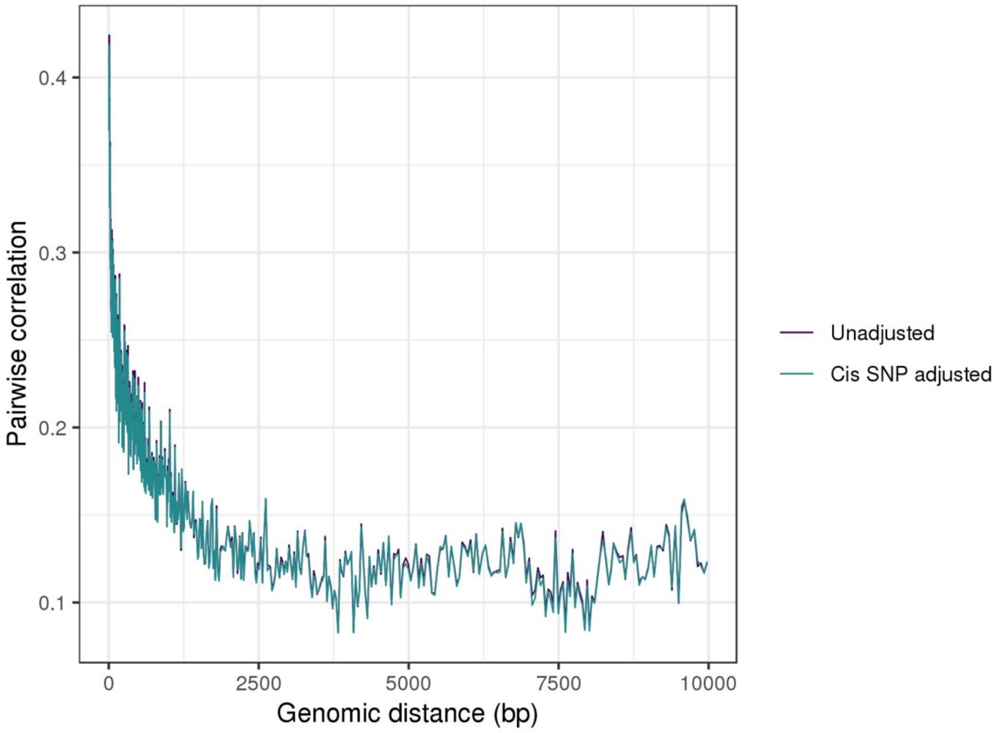


Figure 58: Bar plots of the percentage of pairwise correlations in each correlation range that have 0, 1 or 2 DNAm sites associated with an mQTL identified by the GoDMC consortium. Split by cis (left) and trans (right) correlating pairs, in the white British (top) and Pakistani (bottom) groups in BiB.

5.3.3.2.1 Removing genetic influence from cis correlation plots

To assess the extent of genetic influence on cis DNAm correlations, I regressed out the effect of the strongest *cis* mQTL from the DNAm data in each BiB ethnic group. I re-plotted the *cis* decay plot, with both the adjusted and unadjusted values, and without the standard deviation, for a clear comparison. In contrast to ARIES, there is very little reduction in *cis* correlation in either ethnic group; please see Figure 59. I investigated whether this may be due to the DNAm sites that did not get translated to the EPIC array being more influenced by genetic factors; however removing the DNAm sites exclusive to the 450k from the ARIES data did not notably change the decay of the correlations in ARIES. Substantially more mQTLs $p<10^{-8}$ are found to associate with DNAm sites in ARIES (n= 859) than in BiB (n= 341), which suggests the reduction in correlation may be due to fewer adjustments. However this finding is still under investigation; it may be due to an issue with the processing of genetic data, as the ethnicity PC plots that are overlaid with the 1000 Genomes data show, in Figure 5 in Chapter 2.



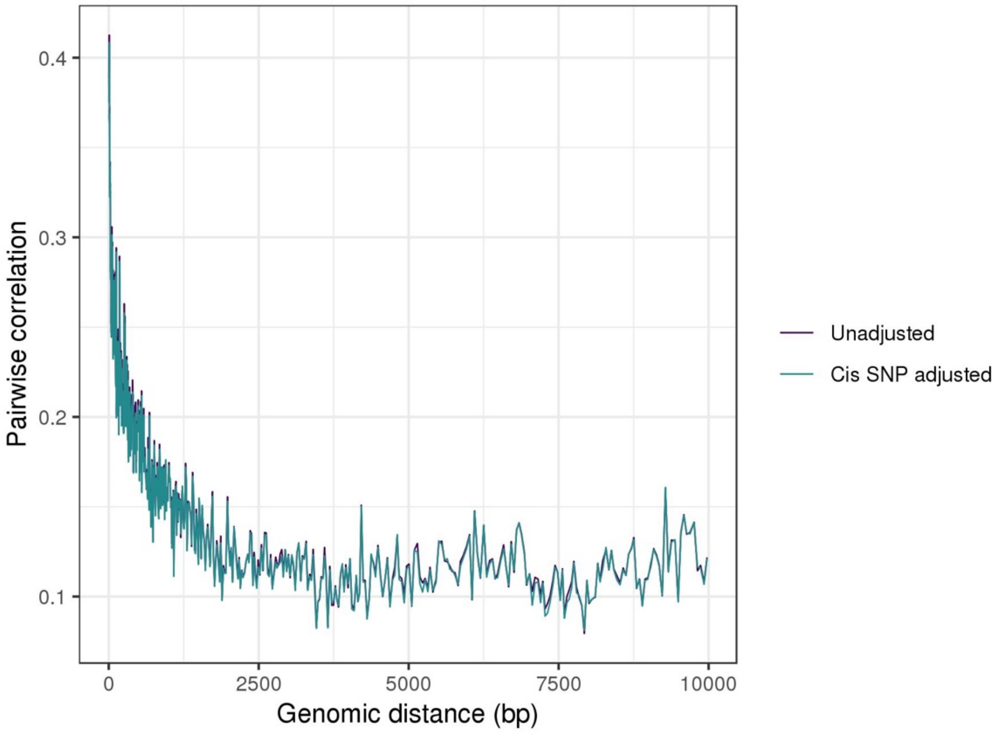


Figure 59: Plot adjusting for cis genetic influence on chromosome 20, in the BiB white British (top) and Pakistani (bottom) groups.

5.3.3.3 Influence of LD on correlation structure

5.3.4 Analysis of strong correlations

Strong correlations (r>0.9) were analysed for enrichments. Table 22 shows the number of probes in this analysis in each of the BiB ethnic groups. It also shows the numbers for ARIES at birth, for comparative purposes. What is striking about this is the almost identical numbers of *cis* correlations as compared to ARIES, and the vastly larger number of *trans* correlations in BiB, particularly in the Pakistani group.

Dataset	Cis correlations	Trans correlations
BiB white British	282	3577
BiB Pakistani	267	4951
ARIES (birth)	238	77

Table 22: Numbers of cis and trans correlations r>0.9 in each of the BiB ethnic groups, and ARIES at birth for comparative purposes.

5.3.4.1 Genomic region enrichment

5.3.4.1.1 Chromatin states

5.3.4.1.1.1 Cis correlations

I find that *cis*-correlating sites r>0.9 are strongly enriched for locations at poised promotors (PromP) in both ethnicities in BiB, with a slight enrichment for bivalent promotors (PromBiv) and ZNF genes & repeats (ZNF/Rpts). The strong enrichment for poised promotors matches the strong enrichment seen in ARIES at all three timepoints. The BiB plots are below in Figure 60 and Figure 61; for the ARIES plot please see Figure 23 in chapter 3.

5.3.4.1.1.2 Trans correlations

In both the BiB ethnic groups, *trans*-correlating sites r>0.9 are strongly enriched for locations at promotors downstream of transcription start sites 1 (PromD1), and active transcription start sites (TssA). They are weakly enriched for locations at promotors upstream of transcription start sites (PromU), and transcription regulatory (TxReg). These are the same enrichments as found in ARIES (see chapter 3 section 3.3.4.1.1.2). The BiB plots can be found in Figure 62 and Figure 63.

5.3.4.1.2 Histone modifications

5.3.4.1.2.1 Cis correlations

The histone modification plots in Figure 64 and Figure 65 show no enrichment for histone marks in the *cis* correlating DNAm sites. It has been shown that poised promoters are enriched for H3K4me1, H3K4me3 and H3k27me3 (Bernhart et al., 2016); H3K27me3 has measures in hematopoietic stem cells, but these histone marks were not enriched at the *cis* correlating sites. This mirrors the absence of enrichment found in all the ARIES timepoints.

5.3.4.1.2.2 Trans correlations

The histone modification plots in Figure 66 and Figure 67 show an enrichment for association with a number of histone methylation and acetylation modifications. The only blood relevant tissue (hematopoietic stem cells) had data for only 3 of the histone modifications. This is shown in Figure 66 and Figure 67. As such we might be able to assume that H3K9K14ac may be enriched at these DNAm sites in blood, H3K9me3 shows some evidence of enrichment, and H3K27me3 does not show enrichment at these sites in blood. This mirrors the enrichment in ARIES.

5.3.4.1.3 Transcription factor binding sites

5.3.4.1.3.1 Cis correlations

The transcription factor binding site (TFBS) enrichment analyses show limited enrichments for TFBS in *cis* correlating pairs for both BiB ethnicities. The TFs that are enriched and were assayed in blood are RNA Polymerase III (Pol3), BRF1, and BDP1. BRF1 and BDP1 are two subunits of the transcription factor TFIIIB, which is required for Pol3-mediated transcription (Abascal-Palacios et al., 2018). These are the same TFBS for which we see enrichment in ARIES *cis*-correlated sites, and can be seen in Figure 68 and Figure 69.

5.3.4.1.3.2 Trans correlations

The TFBS enrichment analysis for highly correlating *trans* DNAm sites show an enrichment for almost all transcription factor binding sites in the dataset. This is the case for both ethnicities, and more TFBS are enriched in BiB than in ARIES; see Figure 70 and Figure 71. The strongest enrichments that were assayed in blood include Pol2, ELF1, YY1, CMYC, MAX and MAZ, and CTCF enrichment is also seen.

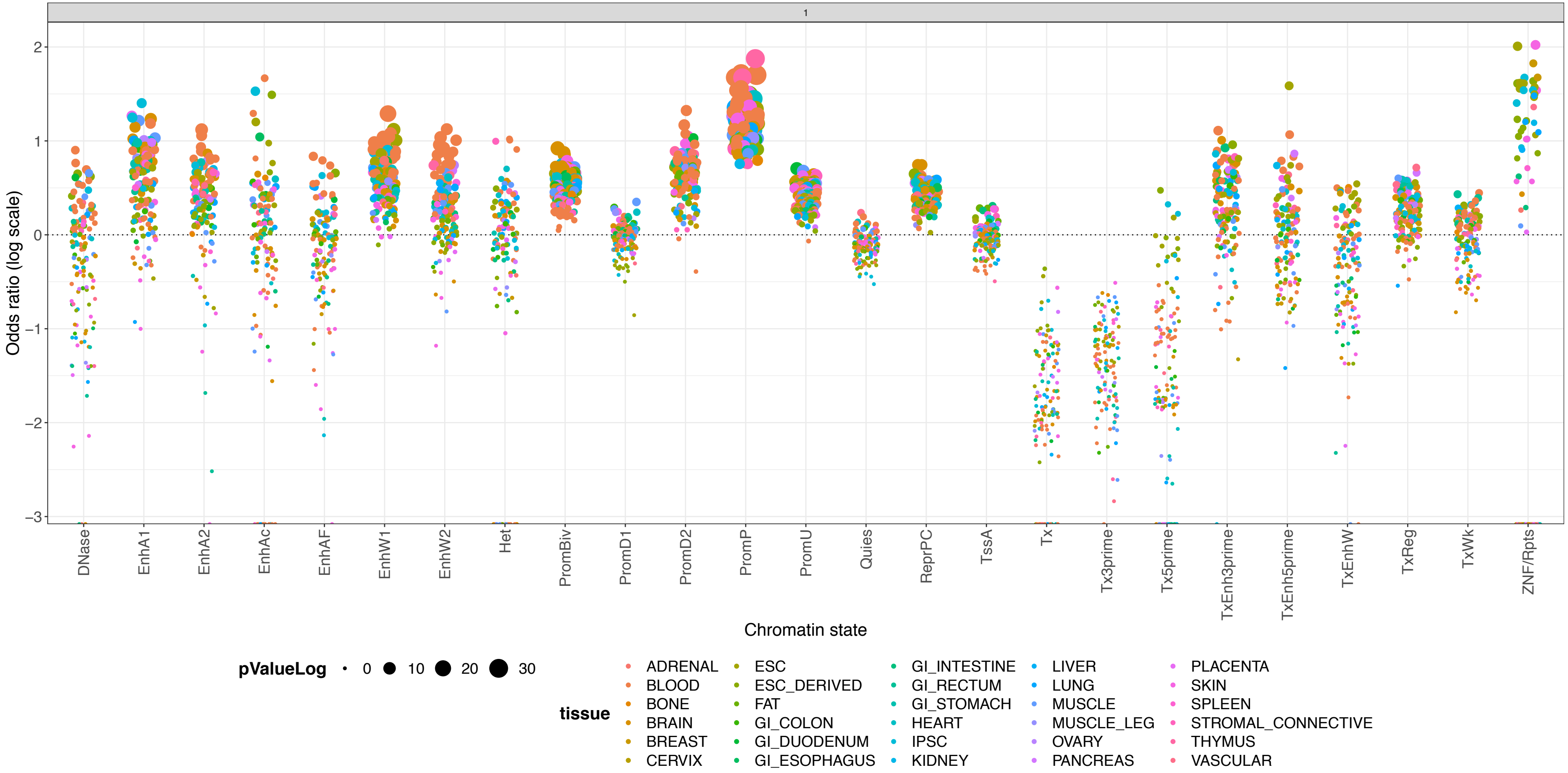


Figure 60: Bubble plot showing the enrichment for the Roadmap Epigenomics 25 chromatin states, for cis-correlating DNAm sites r>0.9, in the BiB white British group

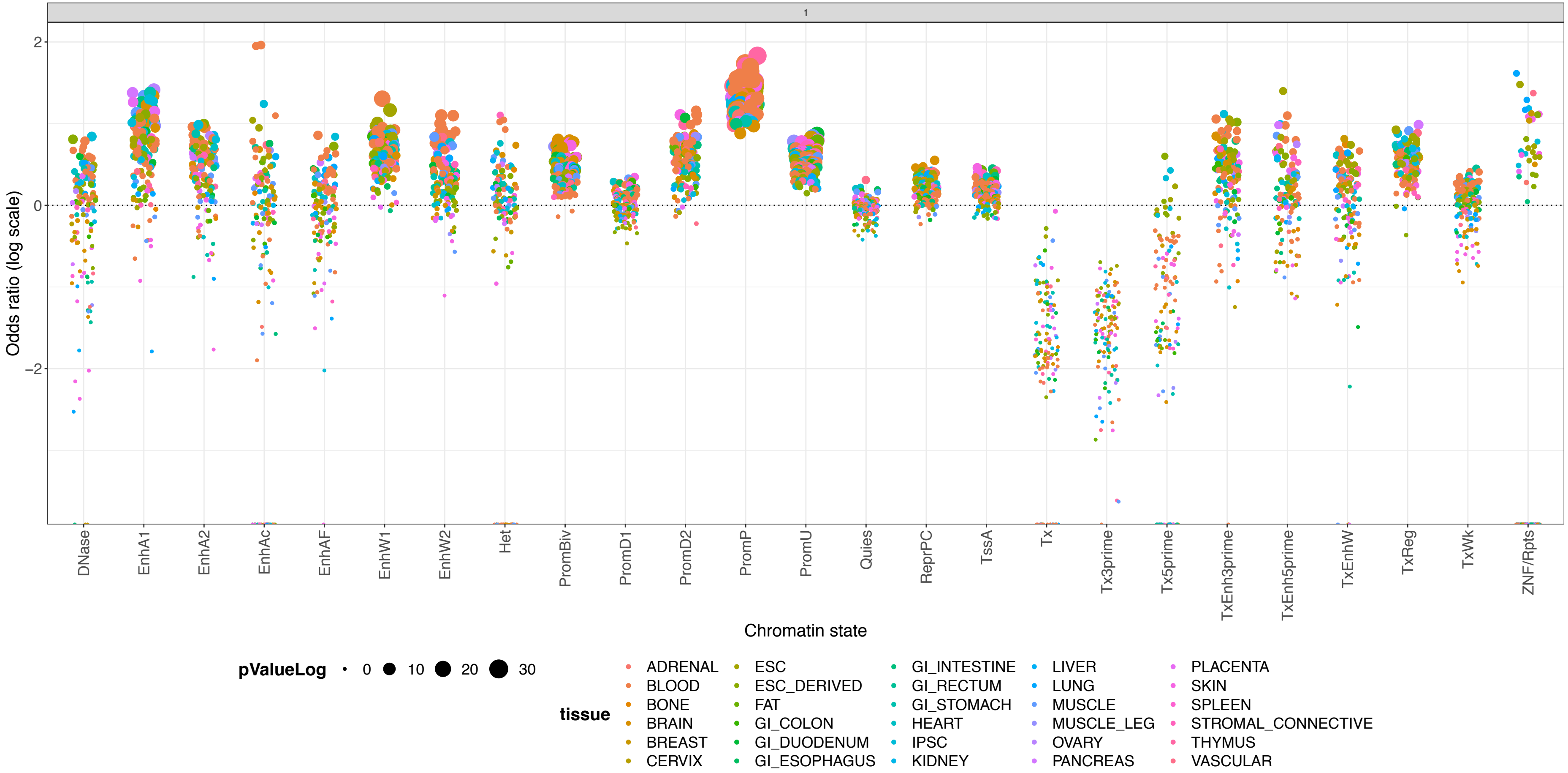


Figure 61: Bubble plot showing the enrichment for the Roadmap Epigenomics 25 chromatin states, for cis-correlating DNAm sites r>0.9, in the BiB Pakistani group

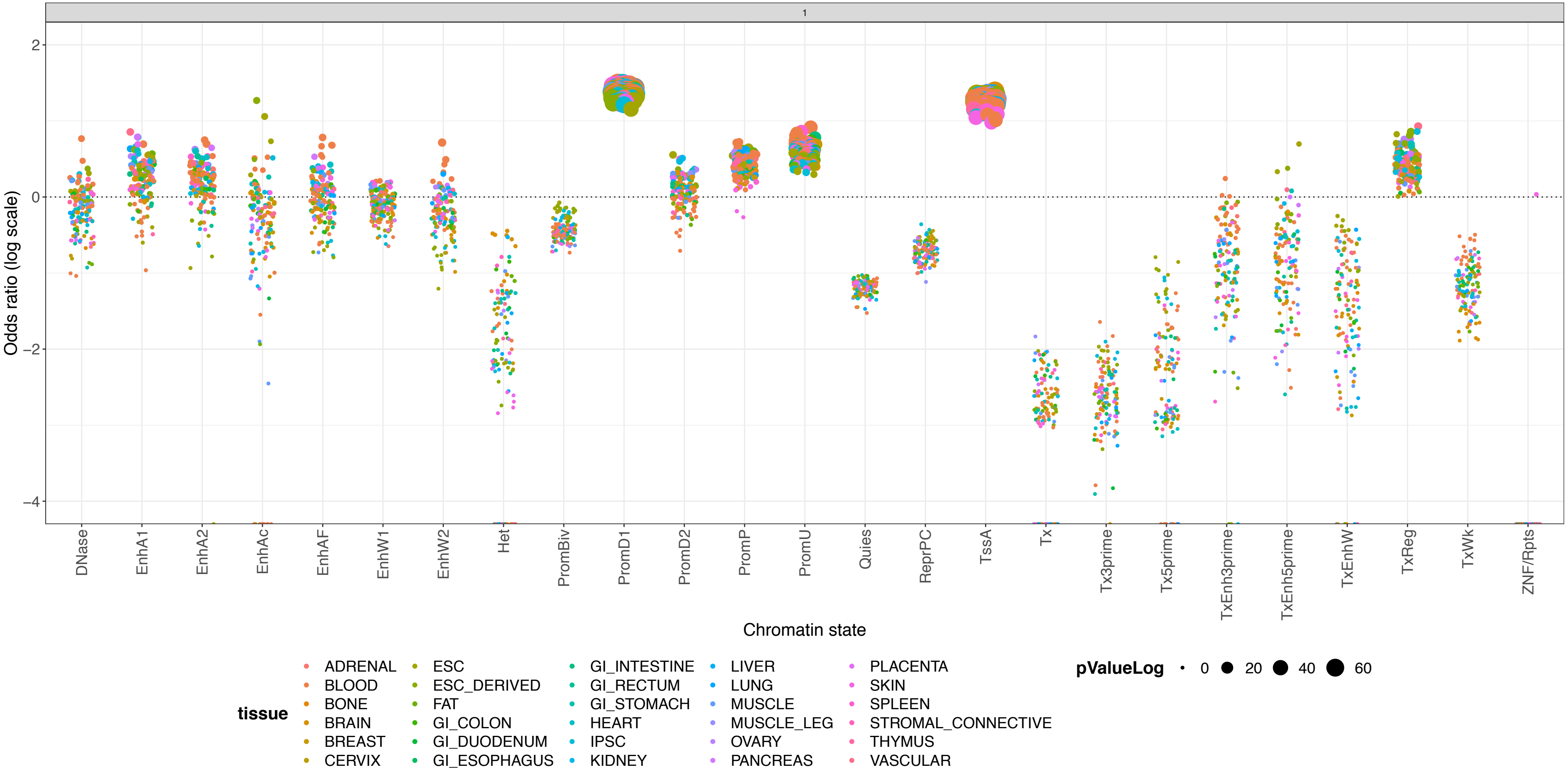


Figure 62: Bubble plot showing the enrichment for the Roadmap Epigenomics 25 chromatin states, for trans-correlating DNAm sites r>0.9, in the BiB white British group

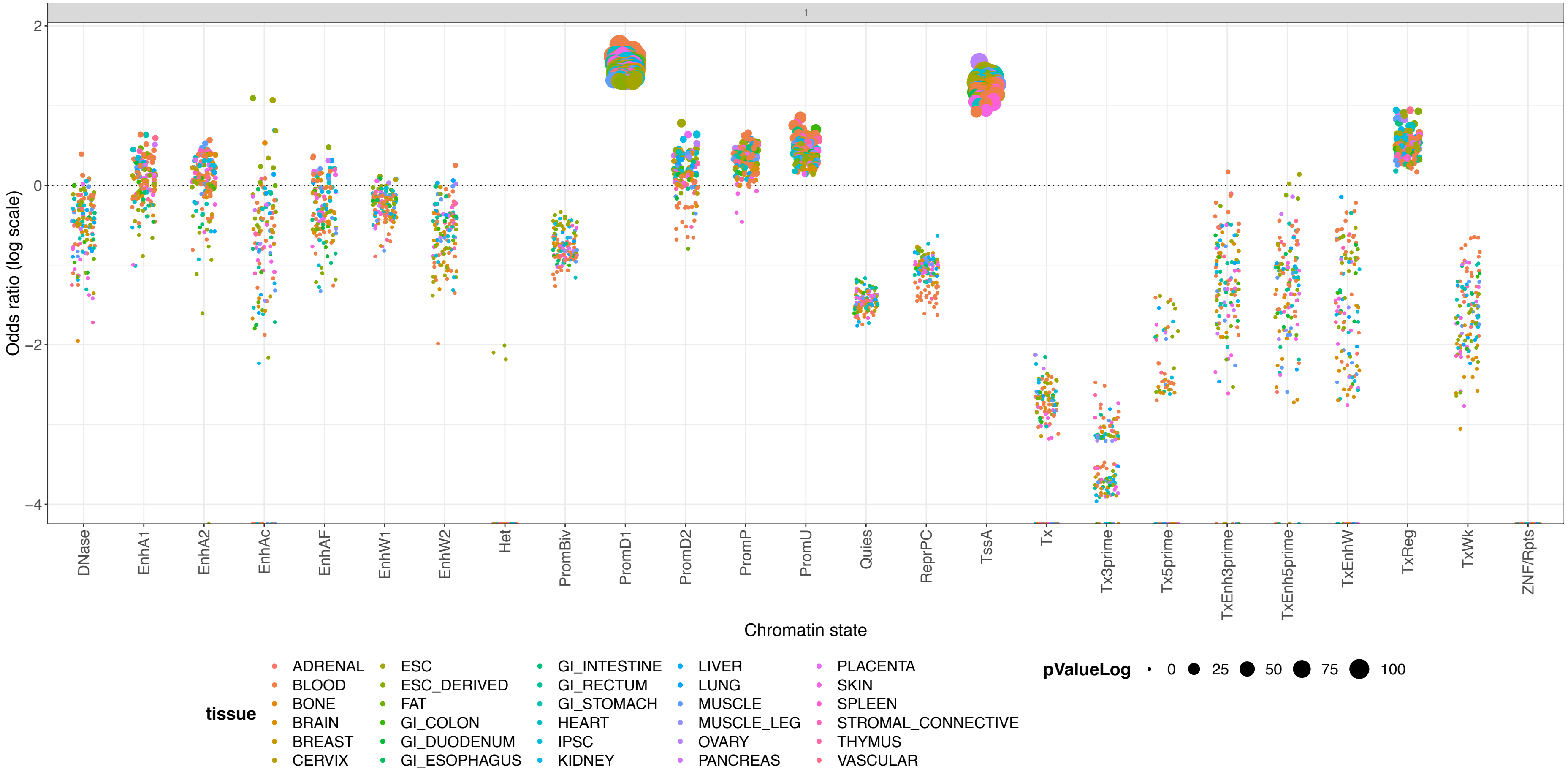


Figure 63: Bubble plot showing the enrichment for the Roadmap Epigenomics 25 chromatin states, for trans-correlating DNAm sites r>0.9, in the BiB Pakistani group

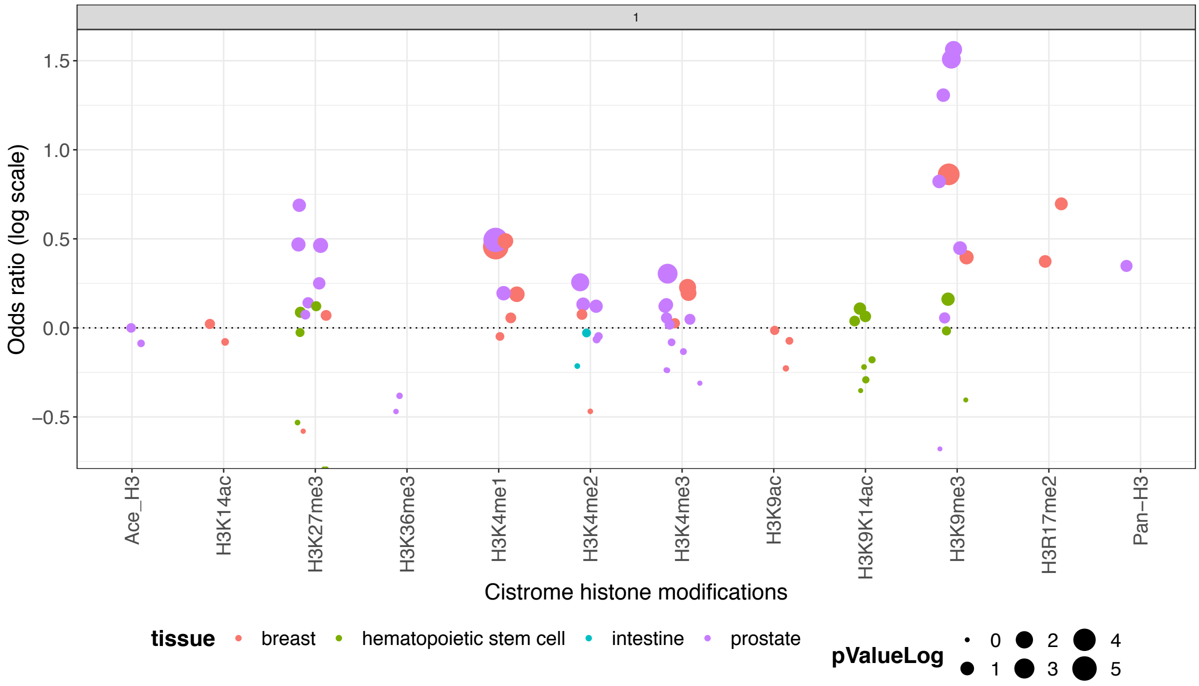


Figure 64: Bubble plot showing the enrichment for Cistrome histone modifications, for cis-correlating DNAm sites r>0.9, in the BiB white British group

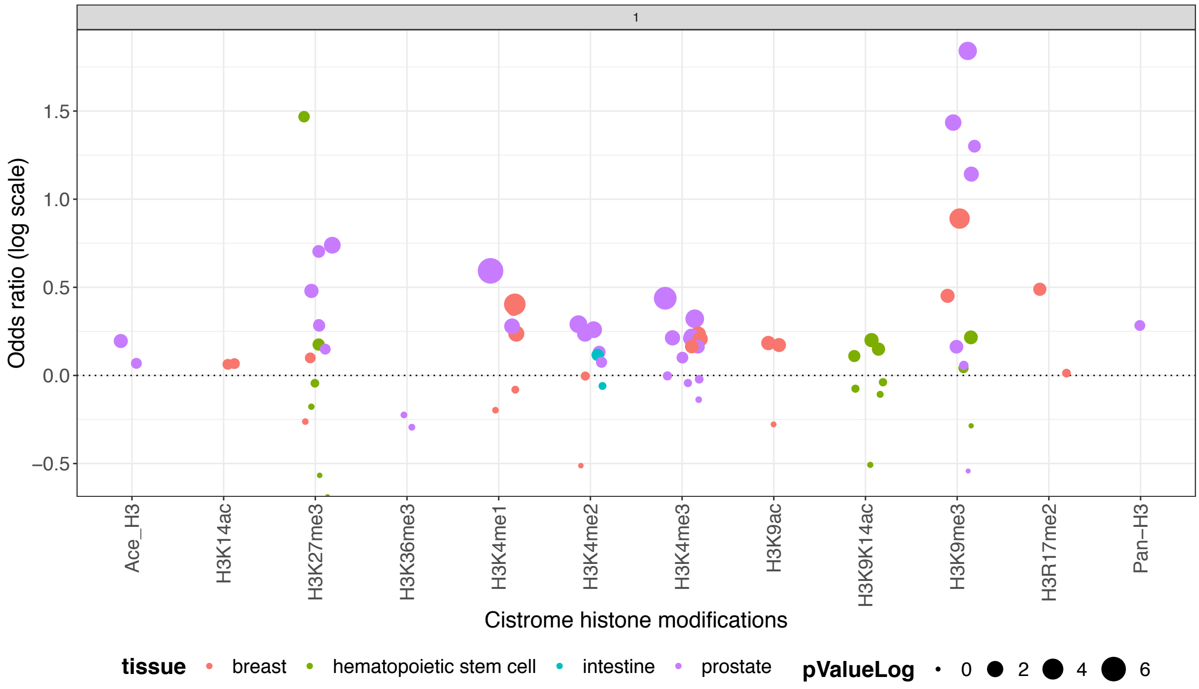


Figure 65: Bubble plot showing the enrichment for Cistrome histone modifications, for cis-correlating DNAm sites r>0.9, in the BiB Pakistani group

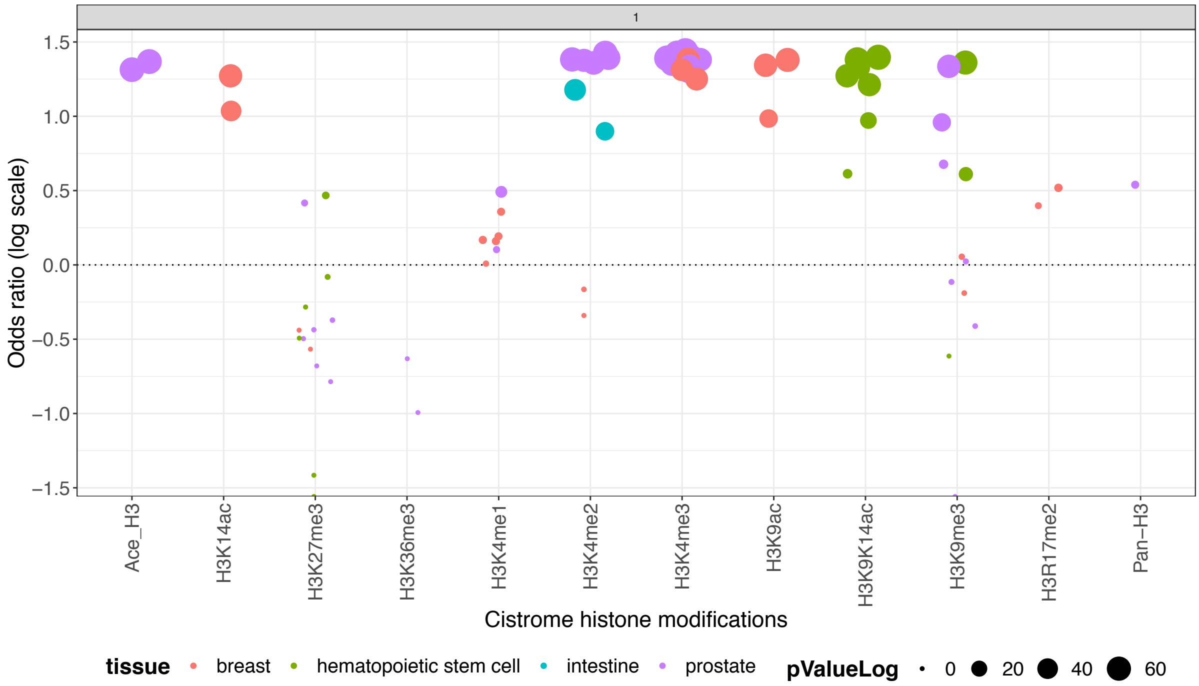


Figure 66: Bubble plot showing the enrichment for Cistrome histone modifications, for trans-correlating DNAm sites r>0.9, in the BiB white British group

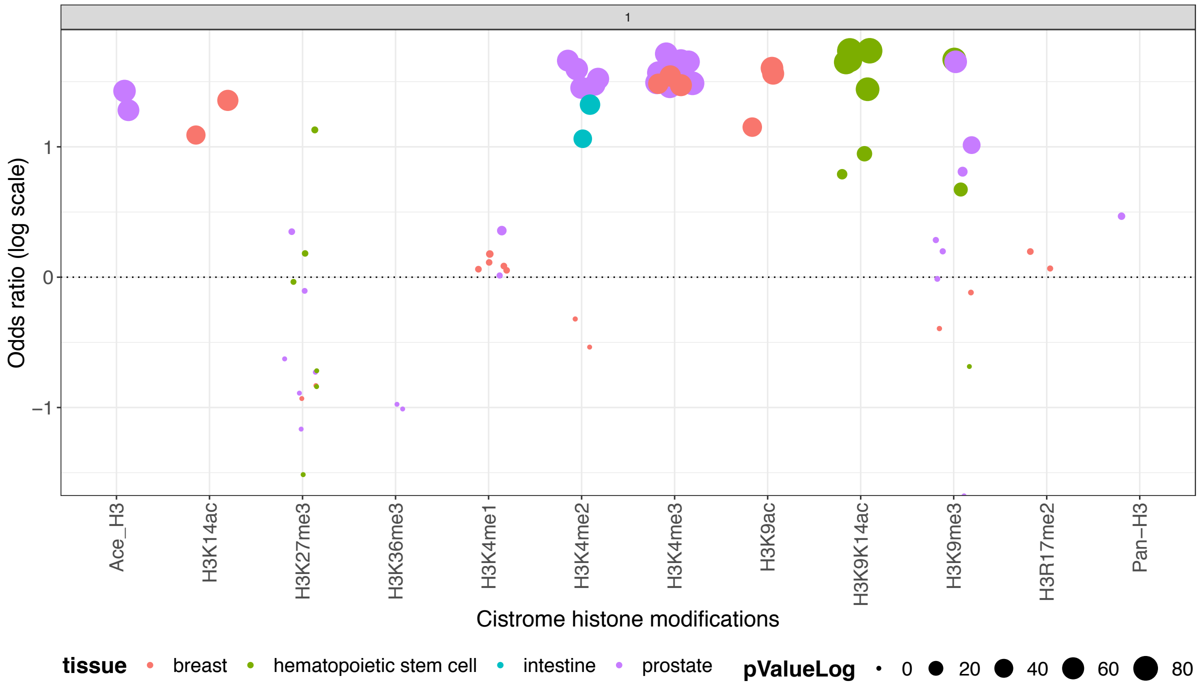


Figure 67: Bubble plot showing the enrichment for Cistrome histone modifications, for trans-correlating DNAm sites r>0.9, in the BiB Pakistani group

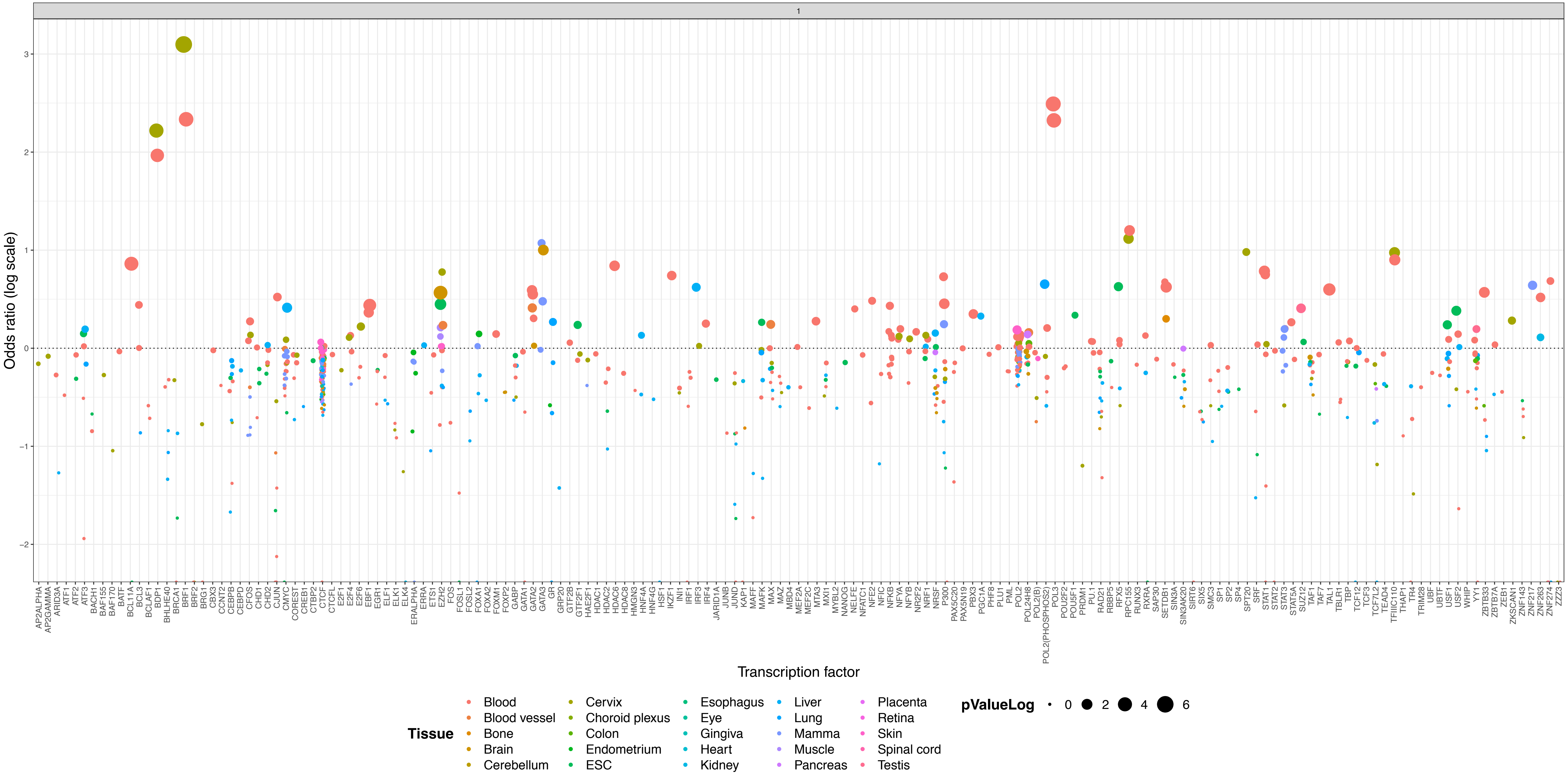


Figure 68: Bubble plot showing the enrichment for the ENCODE transcription factor binding sites, for cis-correlating DNAm sites r>0.9, in the BiB white British group

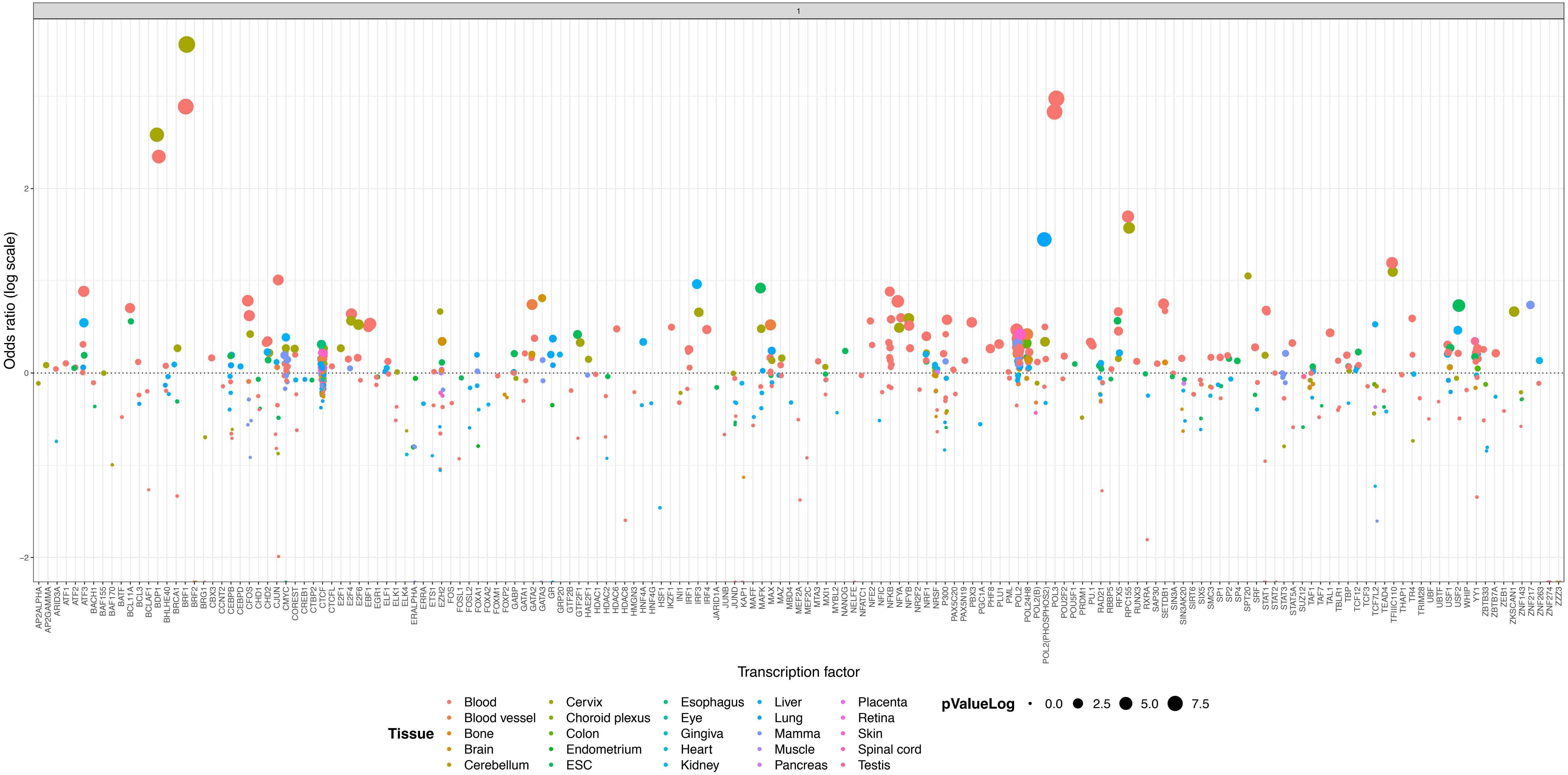


Figure 69: Bubble plot showing the enrichment for the ENCODE transcription factor binding sites, for cis-correlating DNAm sites r>0.9, in the BiB Pakistani group

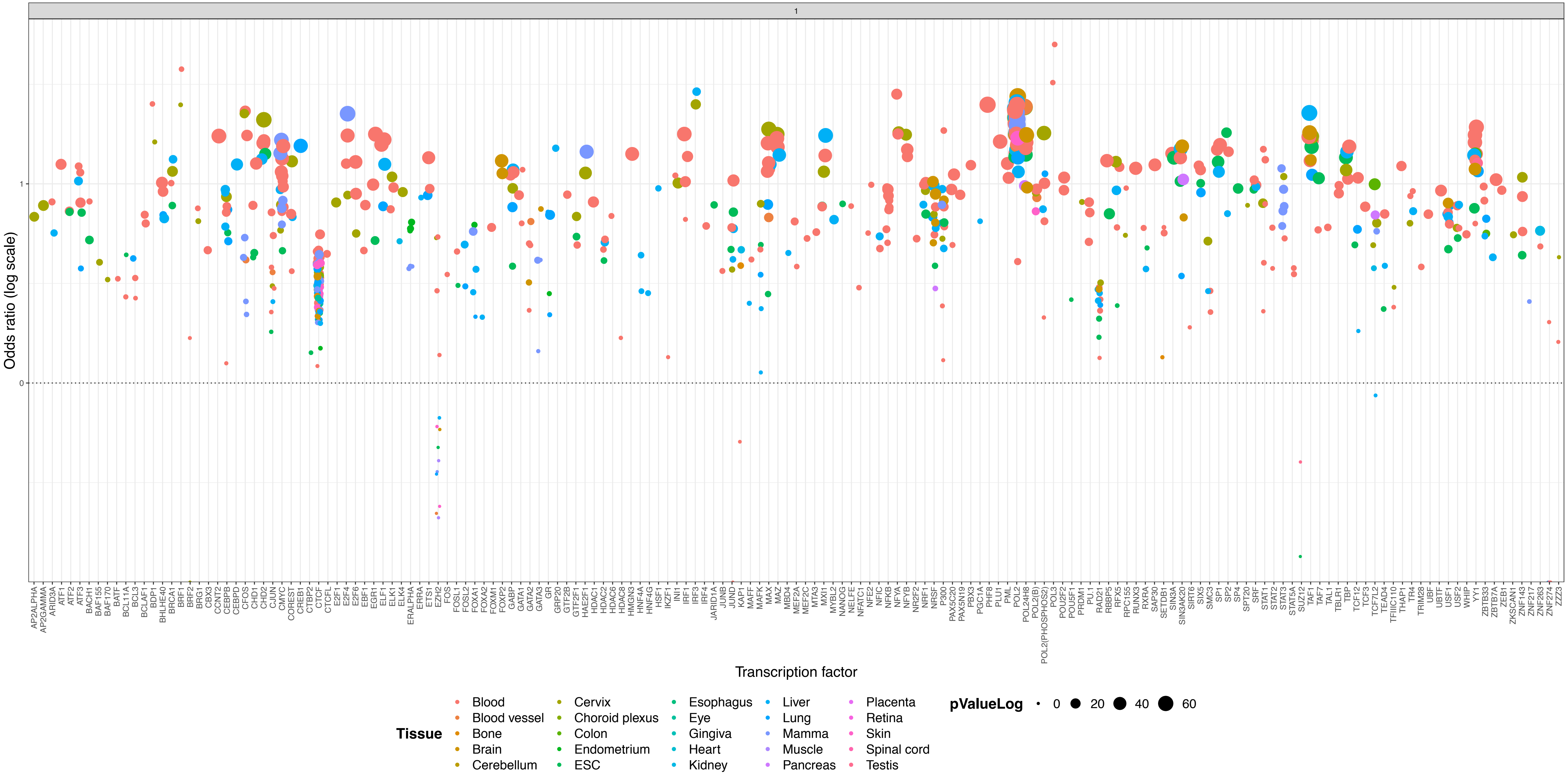


Figure 70: Bubble plot showing the enrichment for the ENCODE transcription factor binding sites, for trans-correlating DNAm sites r>0.9, in the BiB white British group

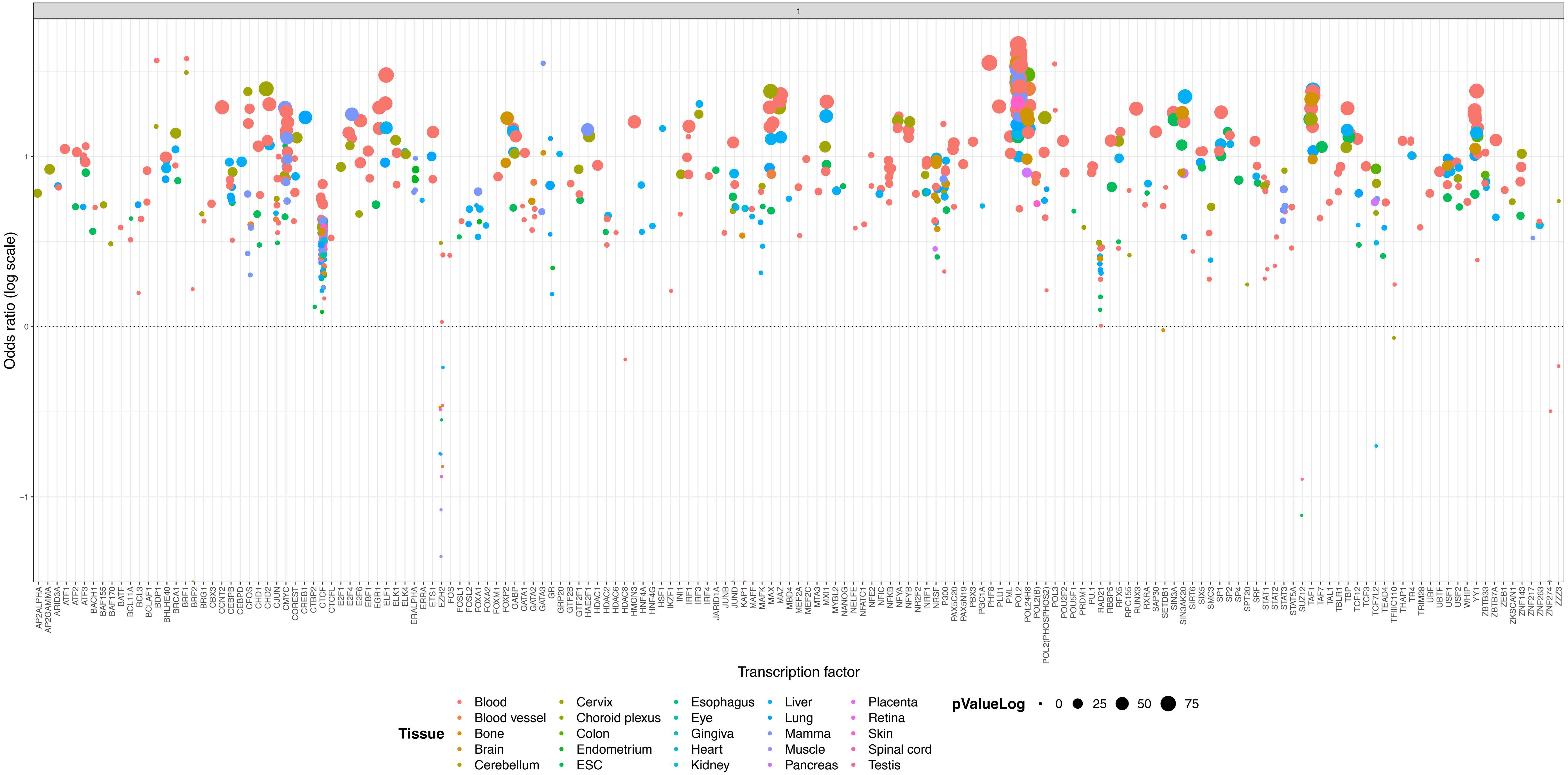


Figure 71: Bubble plot showing the enrichment for the ENCODE transcription factor binding sites, for trans-correlating DNAm sites r>0.9, in the BiB Pakistani group

5.3.5 Trans correlation structure

5.3.5.1 Visualisation of trans structure

5.3.5.1.1 Circos plots

The circos plots in Figure 72 illustrate similar patterns of inter-chromosomal trans correlations across the genome in the two BiB ethnic groups. The plots illustrate that the correlations are spread throughout the genome; that some sites have numerous correlations; and that the sites appear quite interconnected. The plot also demonstrates the larger number of high correlations in the Pakistani group (as shown in Table 22).

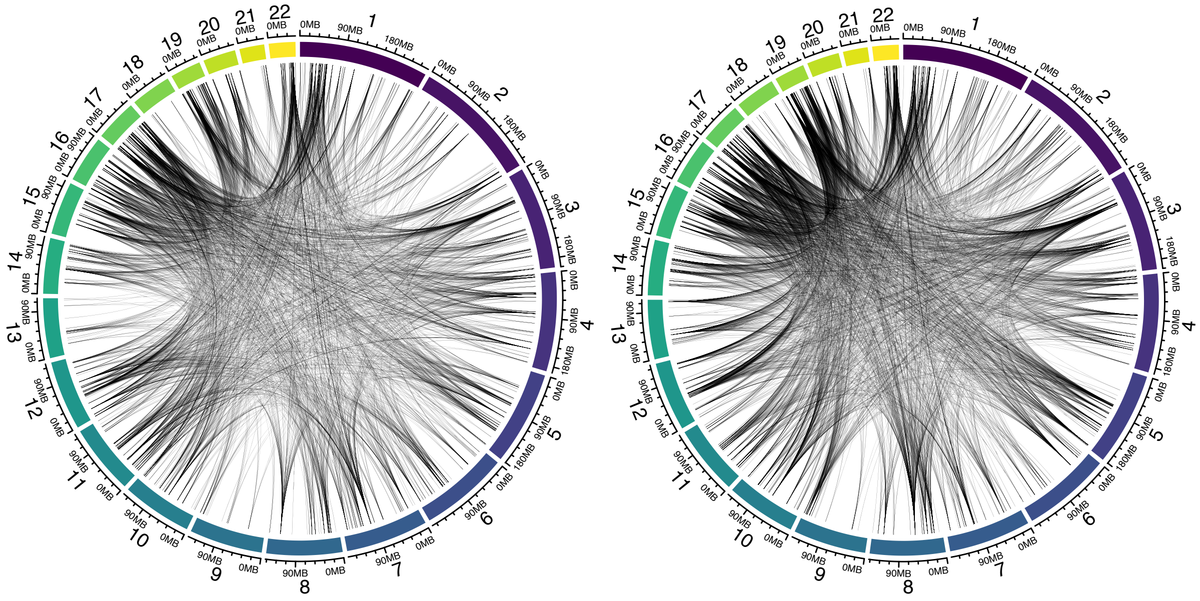


Figure 72: Circos plots illustrating high (>0.9) inter-chromosomal trans correlations in the white British group (left) and the Pakistani group (right).

5.3.5.1.2 Cytoscape plots

Cytoscape plots in Figure 73 illustrate how many more *trans* correlations there are in BiB as compared to ARIES. They show a different pattern to ARIES, in that there are multiple interconnected networks, and a number of groups with fewer than 5 nodes. This structure may be due to the greater number of correlations in BiB; if the threshold of 0.9 were lowered in ARIES it is possible the same pattern would occur.

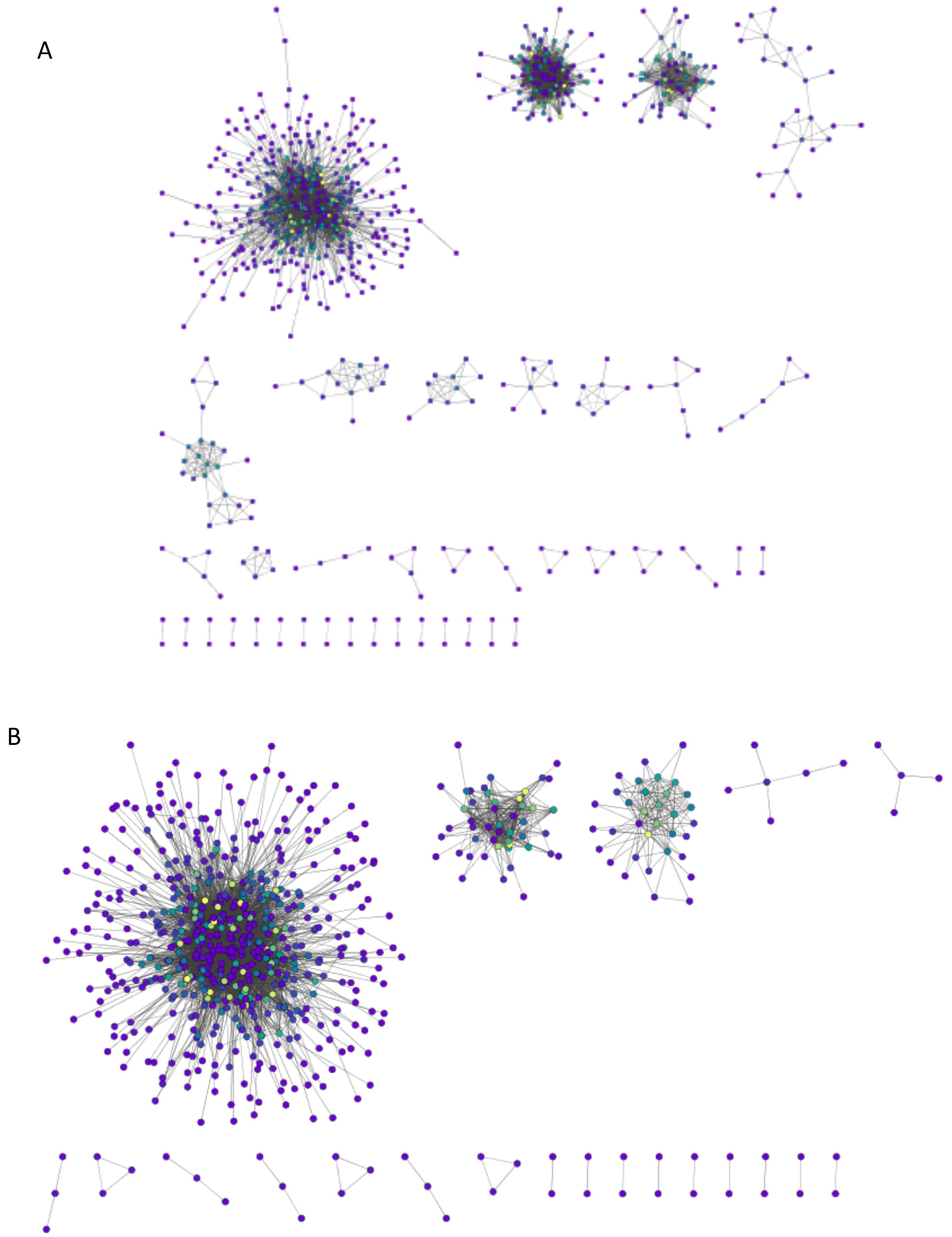


Figure 73: Cytoscape plots of trans correlation networks in BiB white British (A) and BiB Pakistani (B) groups

5.3.6 Consistency in correlations between ARIES and BiB

I tested whether strong correlations replicate between BiB and ARIES, to assess whether strong correlations were consistent, or cohort-specific. One would not expect batch effects, which could induce correlations between DNAm sites, to replicate. However it is important to bear in mind that technical effects such as probe design would replicate across the two studies, and we cannot exclude that these are the reasons for the high correlations.

To assess the preservation of high correlations between ARIES and BiB, and between the two ethnic groups in BiB, I created mean difference plots. Mean difference plots are described above in section 5.2.12.

5.3.6.1 Cis correlations

The *cis* mean difference plot between ARIES and the BiB white British participants (Figure 74) shows that correlations are higher in BiB than in ARIES by a mean of 0.085. As the 95% confidence intervals include a difference of 0, this provides weak evidence of a difference in correlations between the two cohorts.

When comparing *cis* correlations between the white British and Pakistani ethnic groups in BiB, the mean difference between the two groups is much smaller; the Pakistani group have correlations that are higher by 0.003. This difference is stable for correlations between 0.8 and 1. This shown in Figure 75, and illustrates that there is a much greater difference in correlations between DNAm sites between cohorts of the same ethnicity than there is between two ethnic groups in the same cohort. It also provides further evidence that cis correlation structure does not appear to differ between ethnic groups. Whether the differences between BiB and ARIES are due to cohort effects, or unaccounted for batch effects, is unclear. It is unlikely to be array effects because they were both measured using the same technology.

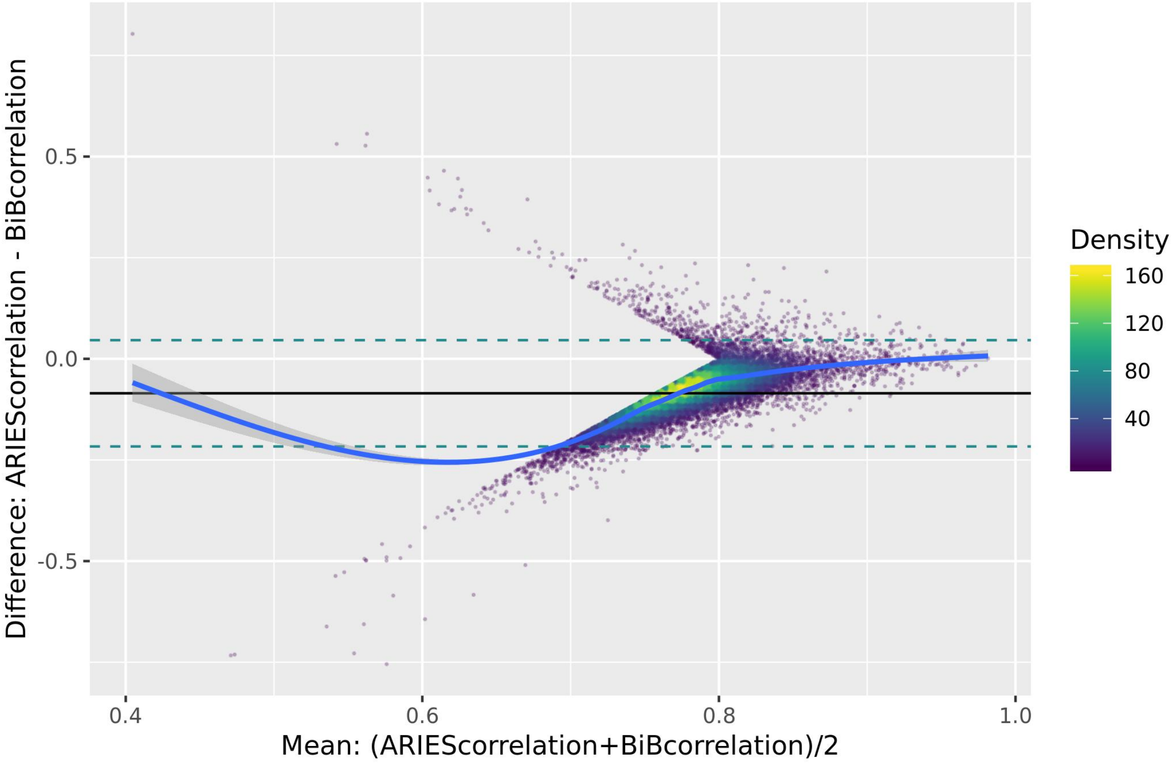


Figure 74: Cis correlation mean difference plot, which plots the mean against the difference in correlation, between ARIES and the BiB white British group, at birth, for cis correlations r>0.8. The solid black line represents the mean difference in correlation; the dashed green lines are the 95% confidence intervals around the mean; and the blue line is a smoothed regression line.

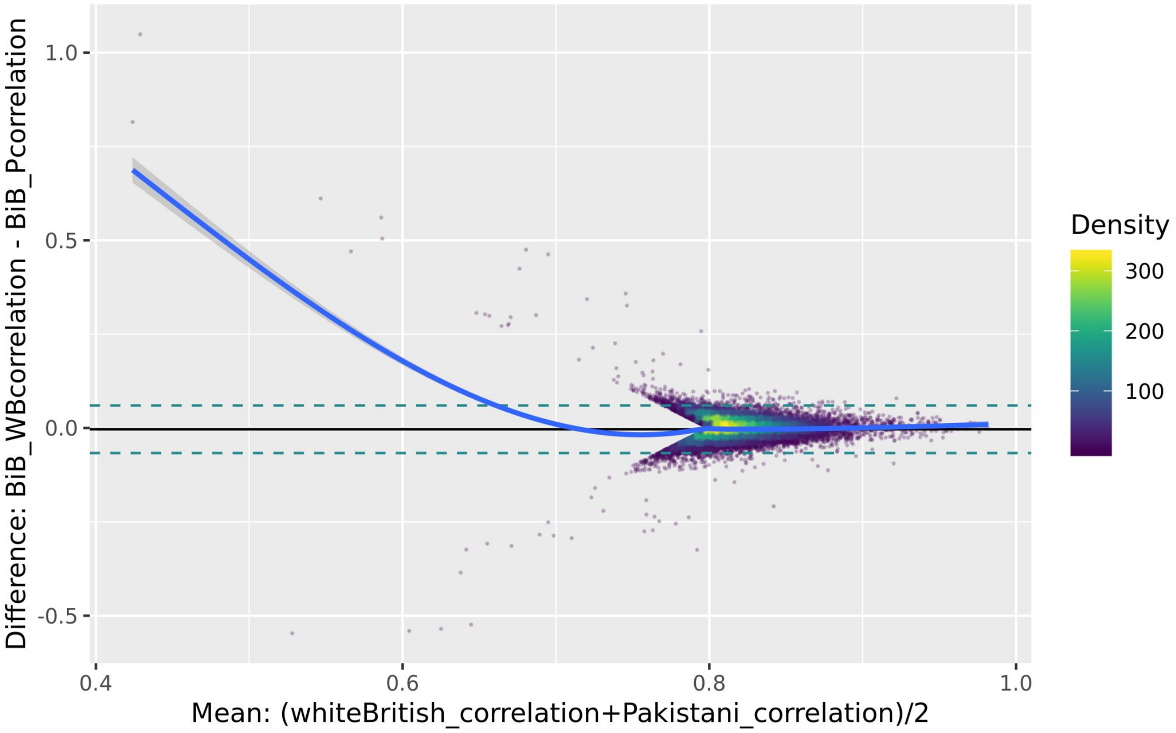


Figure 75: Cis correlation mean difference plot, which plots the mean against the difference in correlation, between the BiB white British and Pakistani ethnic groups, at birth, for cis correlations r>0.8. The solid black line represents the mean difference in correlation; the dashed green lines are the 95% confidence intervals around the mean; and the blue line is a smoothed regression line.

5.3.6.2 Trans correlations

When comparing *trans* correlations r>0.8 between ARIES and the BiB white British group the *trans* correlations are higher by a mean of 0.088 in BiB, replicating what we see for the *cis* correlations. Again we see that there is a smaller difference for correlations between around 0.9 and 1, illustrated by the regression line on the plot (Figure 76). Again the 95% confidence intervals show that this is weak evidence of a difference between the two cohorts. What is different to the *cis* correlations is the range of difference between the two cohorts; there are relatively large numbers (although a small proportion overall) of *trans* correlations that have a difference of up to 1; so in effect there are some strong (r>0.8) correlations, in both ARIES and BiB, that have a correlation close to 0 in the other cohort.

When comparing *trans* correlations between the two ethnic groups in BiB, correlations in the Pakistani group are higher by a mean of 0.005. The 95% confidence intervals do not cross 0, and so there is little evidence for a difference between the correlations in the two ethnic groups; and much weaker evidence than for the difference between ARIES and the

BiB white British group. This is illustrated in Figure 77. However there is a much greater range for a small number of the *trans* correlations between the two ethnic groups in BiB; some of the differences are over 1.5, which means that positive correlations in one ethnic group are negative in the other. These sites might indicate differential regulation between the two ethnicities, although it is notable that the figure illustrates how few correlations there are, compared to the absolute number of *trans* correlations.

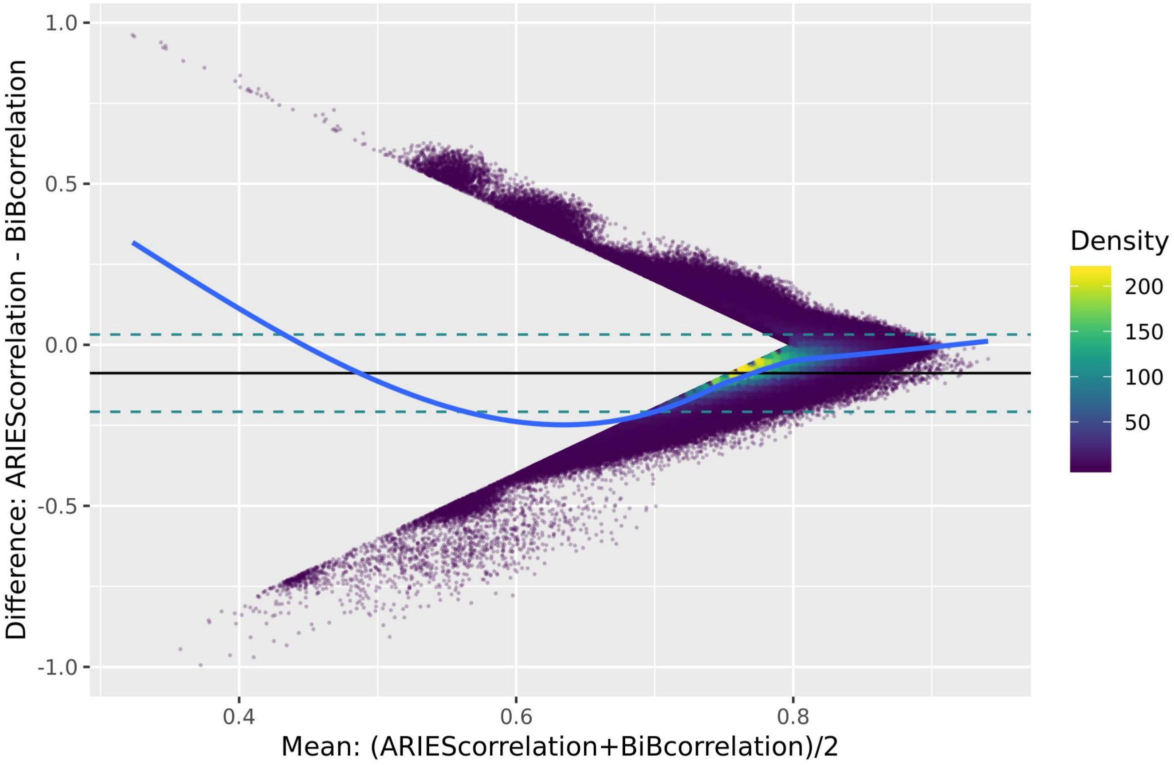


Figure 76: Trans correlation mean difference plot, which plots the mean against the difference in correlation, between ARIES and the BiB white British group, at birth, for trans correlations r>0.8

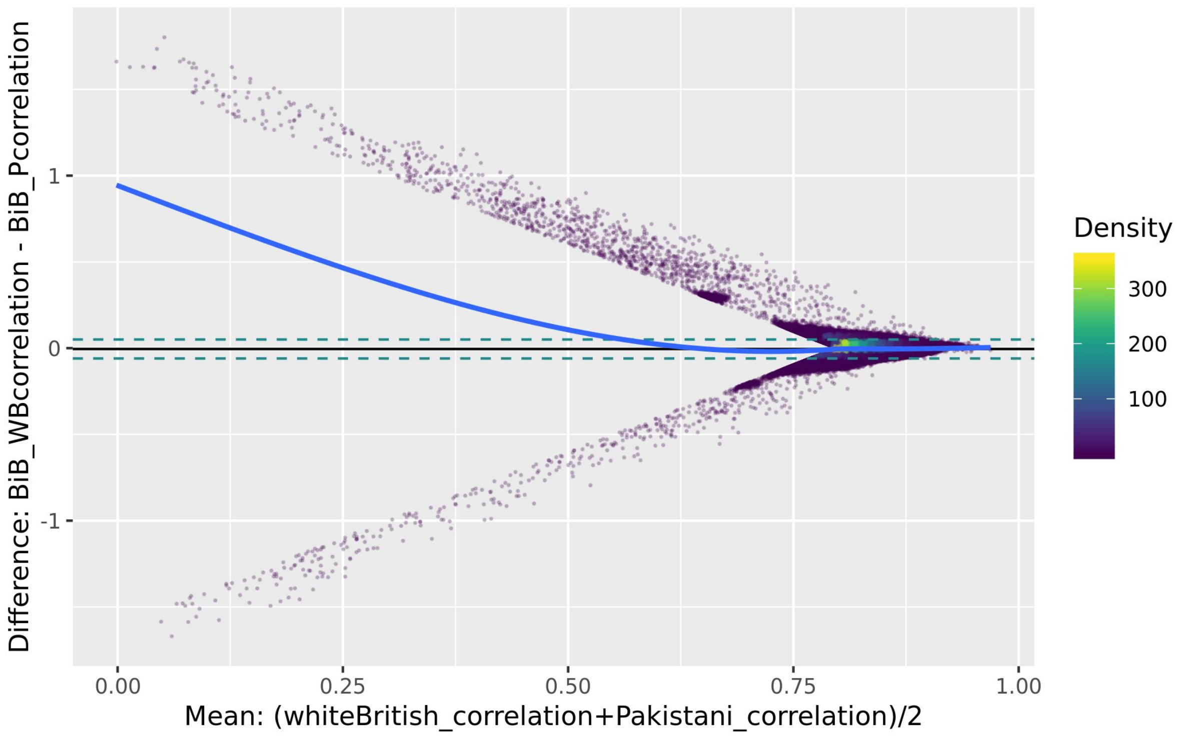


Figure 77: Trans correlation mean difference plot, which plots the mean against the difference in correlation, between the BiB white British and Pakistani ethnic groups, at birth, for trans correlations r>0.8.

Figure 78 shows examples of a correlation that replicates between the cohorts, and a correlation that did not replicate. These were constructed using the normalised but unadjusted beta values, and provide illustrations of the covariances of these DNAm sites.

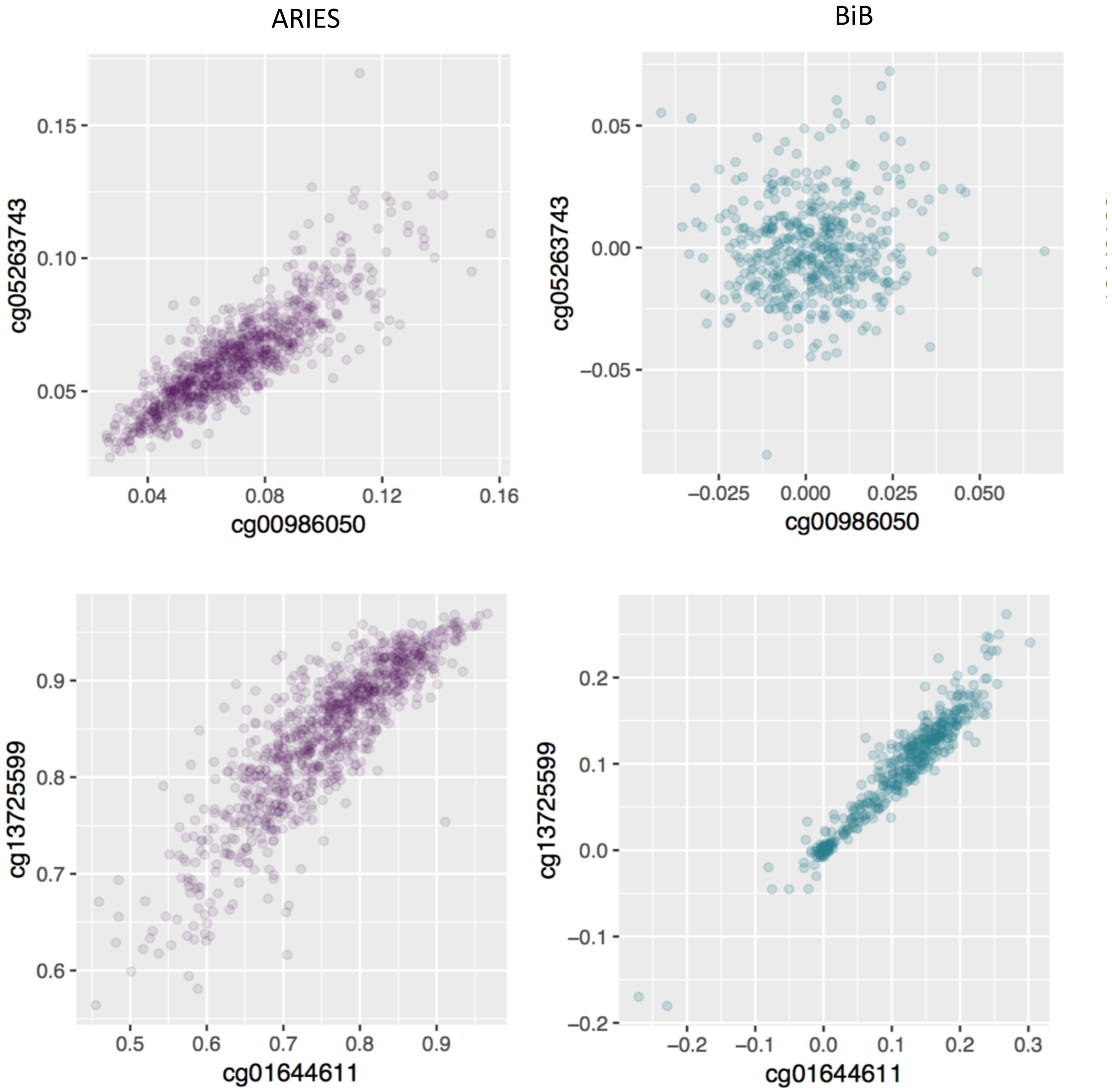


Figure 78: Scatter plots showing examples of correlations that do and do not replicate across the cohorts. ARIES is in purple (left) and BiB white British samples are in green (right). Top panels: A high trans correlation in ARIES that did not replicate in BiB. Bottom panel: a high cis correlation that replicated across both cohorts.

5.4 Discussion

5.4.1 Summary of findings

In this Chapter I have shown that the correlation structure I described in ARIES is generally very well preserved both across datasets and across ethnicities. I show that in fact correlations are broadly more similar between different ethnicities in the same cohort than they are between individuals of the same ethnicity in different cohorts. Of particular importance, the functional annotations of the strong correlations are similar, adding weight to the findings of the enrichments for chromatin states and transcription factor binding sites in Chapter 3. The fact that chromatin state enrichments are almost identical in both BiB ethnic groups and in ARIES for both *cis* and *trans* correlations suggests these sites are very likely to be of biological importance. The stability of correlation structure is important as this means correlations between DNAm sites are more likely to represent a fundamental part of genome biology; they are present in the face of genetic and environmental diversity.

5.4.2 Cis correlation structure

Cis correlation structure replicates solidly between ARIES and BiB. It replicates across the two ethnicities in BiB, so we can be fairly confident that the structure is in fact real; as long as it is not an artefact of the array technology, which was common to both studies. However this seems unlikely as previous work has described the same structure using other methods (Eckhardt et al., 2006). The separation of positive and negative correlations in this thesis was the first demonstration that negative correlations are not distance-based; the replication of this in BiB suggests that this is a stable feature.

Cis correlations were enriched for the same few chromatin states and transcription factor binding sites (TFBS) as in ARIES. This suggests a core function of these cis correlating sites, including enrichment for RNA polymerase III binding sites and chromatin states associated with poised promotors. This suggests that cis correlating DNAm sites may have a role in gene transcription. Some interesting further work could identify whether these DNAm sites are associated with nearby gene expression, and investigate local eQTLs, to further delineate this relationship.

One notable feature that was not replicated from the ARIES analysis was the adjustment for genetic influence on *cis* correlation decay. It is still unclear why this might be, and is the subject of continuing investigation.

5.4.3 Trans correlations

The *trans* correlation structure was also broadly replicated, although it is notable how many more strong trans correlations there were in BiB than in ARIES. It is not clear why this might be; whether it is to do with the differing arrays (although they use the same technology; whether it might be due to the cohort; or whether it might be due to confounding in ARIES induced by sample type at birth being confounded by slide. It is interesting that, even though there are many more trans correlations >0.9 in BiB, the same chromatin enrichments are found, suggesting that highly correlated sites are specifically located. The enrichment for a greater variety of transcription factor binding sites in BiB is likely a reflection of the greater number of high *trans* correlations. It seems that almost all TFBS are enriched; this may reflect a diverse role of *trans* correlating DNAm sites in genome function; and it may illustrate networks of transcription factors (Zhu, Wang, & Qian, 2016). The converse of this is that it is possible that there may be more enrichments in BiB due to measurement error; ARIES has almost twice as many participants and so the measurement there may simply be more accurate.

5.4.4 Comparison of correlations between datasets and ethnicities

The comparison of stability of correlations >0.8 across datasets and ethnicities demonstrated the greater similarity within cohort than within the same ethnic group. Correlation structure between DNAm sites is clearly very similar between ethnicities, which suggests DNAm correlation structure may be a more fundamental aspect of genome biology. The greater difference between cohorts of the same ethnicity may represent cohort effects, or it may represent batch effects that were not accounted for; however the difference between the cohorts was generally low and the functional enrichments were the same, suggesting both cohorts are showing the same function of DNAm. Of course we cannot discount the possibility that these DNAm sites are correlating across datasets for technical reasons rather than biological ones. Although the enrichments point to biological function, we cannot discount that these technical issues map to these enrichments due to some unknown genomic or probe design features.

A limitation of this analysis is that comparing correlations may not be the most appropriate method to assess similarity in fine correlation structure. Mean difference plots were developed to assess individual observations rather than composites (Bland & Altman, 1999; Giavarina, 2015), and so these analyses may not be entirely appropriate. Methods for comparing the structure of matrices are becoming available, which may be a much more informative and appropriate line of work.

5.4.5 Further work

The further work that naturally follows on from this chapter is, in the first instance, to identify the issues with the adjustment for genetic influence on cis correlation structure. This is important because one would expect the results, at least for the Europeans, to be similar as mQTLs would be shared. The similarity of the datasets in the other comparisons make it unlikely that it is to do with a large issue in the data; but all avenues are being considered in the ongoing analysis.

Another key piece of work would be replicating the Hi-C overlap analysis from Chapter 3. As the paper that inspired the analysis was quire recent (G. Li et al., 2019) there was not scope to replicate the analysis in this thesis, but it will be the next step in the analysis.

5.4.6 Summary

This chapter has demonstrated that correlation structure is broadly stable and replicable between two cohorts of the same age. It has also demonstrated that correlation structure differs very little between ethnicities. This provides an important validation that the data used in DNAm correlation networks is likely to be stable, and is likely to be illustrating the same biological features in different datasets. It replicates the biological meaning behind correlations described in Chapter 3, and shows that correlations between DNAm sites, especially *trans* sites, may have roles in functional regulation across the genome.

6 Chapter 6

Validation of systems biology networks

6.1 Introduction

6.1.1 Summary

In chapter 5 I demonstrated that DNAm correlation structure is largely replicable between two cohorts, ARIES and BiB, with highly correlating sites associating with similar biological properties. In chapter 4, I demonstrated that in ARIES a correlation network analysis identifies discrete modules of highly correlated DNAm sites, which reflect a number of stable and persistent biological pathways that DNAm may be involved in. Complex traits and biological pathways are the result of many biological interactions and environmental exposures, all of which can alter DNAm. As an example there is a clear cohort-specific effect between ARIES and BiB in that correlations >0.8 tend to be systematically higher in BiB than they are in ARIES. As a result, it is possible specific pathways of DNAm that I identified in ARIES are cohort-specific, which would make the interpretation of their functional relevance challenging. To identify whether the pathways really represent stable, persistent, and functional biological pathways that DNAm is involved in, in this chapter I repeat the WGCNA analysis in BiB.

I use two approaches in this chapter to assess the preservation of biologically meaningful networks that DNAm is involved in at birth. I firstly construct separate DNAm networks in both the ethnic groups of BiB, and seek to identify any trait- or GO-associated modules. This analysis will serve to identify whether there are any DNAm modules that are specific to individuals of European and South Asian descent. I then construct consensus networks, combining ARIES with both groups of BiB, to identify modules that are present in both datasets. This will identify whether there are any modules that are truly persistent, and therefore that may have fundamental roles in development. If they are present in both ethnicities, this could suggest a more core function in development than we might be able to infer from just the white British samples.

6.1.2 Consensus networks

Consensus network analysis can detect modules of DNAm sites that are shared by multiple datasets (Langfelder & Horvath, 2007), and was developed as part of the WGCNA package (Langfelder & Horvath, 2008). When seeking to identify core biological pathways, consensus analysis is much more powerful than comparing single networks.

6.1.2.1 Consensus network analysis in the literature

Consensus network analysis has been applied to broad questions and has generated many novel insights. In the gene expression literature, consensus networks have been used to identify; molecular networks affected by CAG repeat length in Huntington's (Langfelder et al., 2016); novel, druggable targets in frontotemporal dementia (Swarup et al., 2019); to predict stage and grade of ovarian cancers (Sun et al., 2017); to identify gene expression changes associated with Bisphenol-A (BPA) exposure (Maertens et al., 2018); the discovery of novel embryonic stem cell markers related to differentiation (J. J. Kim et al., 2014); identify potential predictors of progression in head and neck cancers (Sanati, Iancu, Wu, Jacobs, & McWeeney, 2018); and microglia expression profiles associated with neurodegenerative conditions (Holtman et al., 2015). In the DNA methylation literature, consensus network analysis has been used to identify multiple modules of DNAm associated with Huntington's disease (Horvath et al., 2016); an age-related module detectable across blood and brain over 16 datasets (Horvath et al., 2012); and differential cortisol stress reactivity in individuals who have experienced childhood trauma (Houtepen et al., 2016).

6.1.3 Hypotheses

H.6.1. Network modules that are related to core biological functions will be persistent across all three datasets.

H.6.2. Results in Chapter 5 illustrated that DNAm correlation structure is more similar between the two ethnic groups in BiB than between the BiB white British ethnic group and ARIES. I therefore expect there will be a greater preservation of network structure between the two groups of BiB.

H.6.3. Consensus modules preserved between the two BiB groups, but not with ARIES, have a higher likelihood of representing batch effects.

6.1.4 Aims

A.6.1. I will construct a single co-methylation network in each ancestry group of BiB. These networks will be assessed for association with traits, cell type proportions, and gene ontology terms, as were the ARIES networks in chapter 4.

A.6.2. I will create consensus co-methylation networks; one between the individuals of European ancestry, and one between ARIES and both ancestry groups in BiB. This will identify whether any DNAm network modules differ between individuals of European and South Asian descent at birth. The network including all three datasets will be most likely to contain modules that represent fundamental DNAm pathways.

6.2 Methods

6.2.1 Data

The datasets used in this chapter were the ARIES birth timepoint, and the two BiB ethnicities (white British and Pakistani). ARIES is a subsample of ALSPAC, where 1,000 mothers and their children were selected to have DNA methylation profiled. This was at birth (cord blood), 7 years and 15-17 years. ARIES data is described in detail in chapter 2 section 2.2.1.1. DNAm was measured in blood (at birth, either white cells or blood spots), using the Illumina 450k array. 849 individuals in ARIES with cord blood methylation data were included in this analysis.

Born in Bradford (BiB) is a birth cohort following over 12,000 mothers and their children. The cohort is described in detail in chapter 2 (section 2.2.2). Briefly, a subsample of 1,000 mothers and their children were selected to have DNAm data generated from blood samples provided by the mothers during pregnancy, and from cord blood from the children. The subsample was specifically selected to have 500 participants of white British origin, and 500 participants of Pakistani origin. DNAm was measured using the Illumina Infinium MethylationEPIC BeadChip array. Because I used BiB as a validation for ARIES (which was measured by the Illumina hm450 array), and also because of computing considerations, I subset the EPIC data to DNAm sites which are also present on the hm450 array.

6.2.1.1 Adjusting for population stratification and relatedness (BiB)

In BiB, after normalisation I used genetic data to remove individuals related above 0.125, and I regressed out population stratification (using 20 genetic PCs) in the white British and Pakistani groups separately. These steps are described in detail in chapter 2 section 2.2.2.5, and the number of participants remaining after removing related individuals can be found in Figure 79.

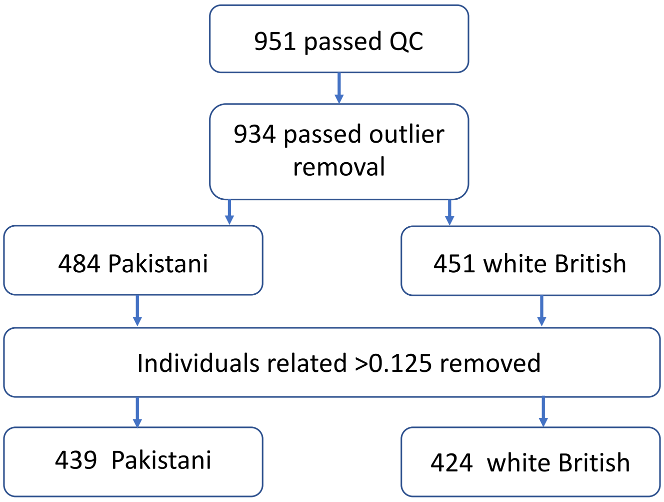


Figure 79: Flow diagram of BiB child participant numbers

6.2.1.2 Adjustments

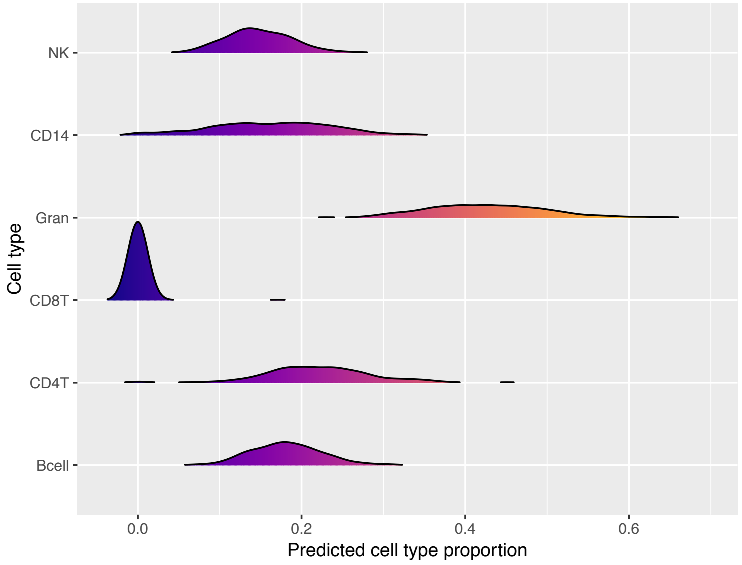
For ARIES, I filtered sites on the 450k array, as detailed in chapter 2, section 2.2.1.7.1). Briefly, all non-autosomal probes, probes in the HLA region, poorly functioning probes and probes containing SNPs were removed from the dataset. This left 394,842 DNAm sites. For BiB, I subset the EPIC data to sites also present on the 450k array, and filtered DNAm sites as detailed in chapter 2, section 2.2.2.7.1). Briefly, all non-autosomal probes, probes in the HLA region, poorly functioning probes and probes containing SNPs were removed from the dataset. This left 369,796 DNAm sites. As consensus analysis requires the same sites to be present in all datasets, for the consensus network analysis I subset both ARIES and BiB to sites common to both datasets. This left 368,093 DNAm sites for the consensus analysis. For both cohorts separately, all outlying values (> 10 standard deviations from the mean) were replaced with the mean for that probe, over three iterations.

As discussed in chapter 2, all datasets had sex and estimated cell counts regressed out using a linear model. Slide was also regressed out for all datasets, to account for batch effects, using a linear model. A random effects model was not used because batch effects were removed as a random effect during normalisation; slide was regressed out after normalisation because the normalisation did not completely remove associations with slide in either ARIES or BiB. ARIES also had sample type regressed out, as ARIES used more than one sample type; however it should be noted that sample type was confounded with slide at birth in ARIES, and so residual effects of slide and sample type are likely to remain in the data. Individuals in ARIES were excluded from the analysis if they were the only individual on a slide (as detailed in chapter 2, section 2.2.1.3).

6.2.1.3 Estimating cell counts

Blood cell type proportions were estimated and adjusted for as detailed in chapter 2 section 2.2.1.6 for ARIES, and chapter 2 section 2.2.2.6 for BiB. Cell counts were estimated using `meffil` (Min et al., 2018); the Houseman algorithm (Houseman et al., 2012) is used to estimate cell counts with this function. Josine Min estimated the ARIES cord blood cell type proportions using the Gervin reference panel (Gervin et al., 2016). I estimated cell counts for BiB using the Andrews and Bakulski cord blood reference panel (Bakulski et al., 2016), again using `meffil`.

In comparing there two cohorts, there are some important factors to bear in mind. Firstly, the sample types were different between the two cohorts – all BiB samples were whole blood from umbilical cord, whereas ARIES at birth was composed of 18.1% blood spots and 81.9% white cells from umbilical cord blood. Greater cell type variation in BiB may have led to greater variance in the data, and so greater potential for correlation. To assess this I created density plots of predicted cell type in ARIES and BiB at birth; in BiB there is likely more variation due to the presence of nucleated red blood cells (these were not adjusted for in ARIES because of the use of white cells). There are also quite different distributions of NK and CD8T cells in the two cohorts, with a proportion around 0 for CD8T in ARIES, and a proportion around 0 for NK in BiB. These differing proportions could lead to differences in correlation profile. In addition, in ARIES sample type was confounded by slide, so there would undoubtedly be sample type and slide effects in the ARIES data that were not accounted for. These plots are found in Figure 80.



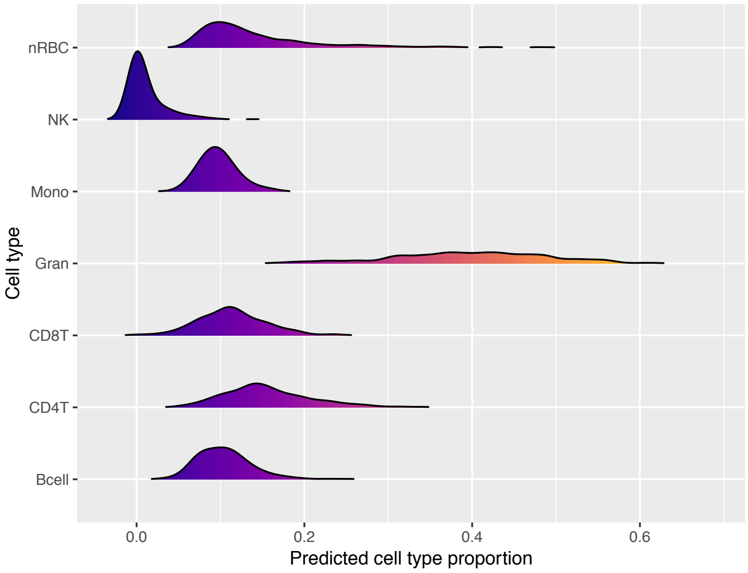


Figure 80: Density plots to compare the predicted proportions of blood cell types in ARIES at birth (top) and BiB white British participants (bottom)

6.2.1.4 Phenotype data

Phenotype data that were assessed for relationships with WGCNA modules are described in chapter 2, section 2.2.1.9 for ARIES, and chapter 2 section 2.2.2.9 for BiB.

6.2.2 WGCNA

Single dataset WGCNA analyses were run in BiB using the protocol described in chapter 2, section 2.6. The networks were created using the same specification as for ARIES; the only parameter that varied was the soft threshold power, which was calculated for each dataset separately. The powers I used can be found in results section 6.3.2. Consensus WGCNA analyses were run using the protocol described in section 6.2.2.5 below.

6.2.2.1 Removing outlying samples

Outlying samples may skew the clustering of DNAm sites, and therefore affect the modules it can detect. As a result it is recommended to cluster samples based on DNAm values, using a hierarchical cluster dendrogram. This method is described in chapter 2, section 2.6.1, and was done for each of the BiB ethnic groups and ARIES separately. The number of participants remaining in each group is displayed in Table 23.

Dataset	Number before clustering	Number after clustering
ARIES (birth)	849	829
BiB white British	424	416
BiB Pakistani	439	429
Total	**1712**	**1674**

Table 23: Number of participants in ARIES and both groups of BiB before and after hierarchical clustering to remove outlying samples before WGCNA analysis.

6.2.2.2 Calculating soft threshold power

As WGCNA raises the correlations between DNAm sites to a power to de-emphasise the weak correlations (known as soft thresholding because it doesn't impose an exclusionary threshold), a power must be selected. The selection is based on the power for which the network would reach scale-free topology, a network theory that assumes non-random connections and highly connected hub nodes (Barabasi, 2009; Barabasi & Albert, 1999). The details and method for calculating this power is in chapter 2, section 2.6.2.

6.2.2.3 Blockwise network construction: single networks

The single WGCNA networks were constructed using the `blockwiseModules` method because I used over 360,000 DNAm sites in the network, and it is not computationally possible to construct a single-block WGCNA network of that size. This was done separately for both BiB ethnic groups, and the detail of the method and settings that I used is described in chapter 2, section 2.6.3, as I used the same network settings (bar the soft threshold power) in both ARIES and BiB. The consensus WGCNA network was constructed using the `blockwiseConsensusModules` method for the same reason, and is described in section 6.2.2.5 below.

6.2.2.4 Preservation of single network modules

To assess how well the single network modules were preserved between ARIES and BiB, I ran the network preservation analysis described in chapter 2, section 2.6.4.1 between ARIES and both ethnicities of BiB separately.

6.2.2.5 Consensus network analysis method

Consensus networks were constructed using R scripts that I adapted from the tutorials available on the UCLA WGCNA website (Langfelder & Horvath, 2016). I constructed the consensus network between all three cord blood datasets – ARIES, the BiB white British ethnic group and The BiB Pakistani ethnic group. Consensus networks were created using the `blockwiseConsensusModules` function of the WGCNA R package (Langfelder & Horvath, 2008). As with the single network creation, there are many options to refine the modules that are created. I used similar options in the consensus module creation to the single networks (code shown can be found in the GitHub link below). The soft threshold power was not the same for any of the datasets, so I used a power of 7 for both the consensus networks because it was midway between the powers for the individual datasets.

Github file path: shwatkins/PhD/WGCNA_analysis/consensus_1.R

6.2.2.6 Association of WGCNA modules with phenotypes

To identify whether WGCNA modules are associated with traits of interest, they can be associated with the module eigengenes (which are essentially the first principal components of each module). The details of the general analysis can be found in chapter 2 section 2.6.4; the details of the phenotypes along with the covariates used for ARIES can be found in chapter 2, section 2.2.1.9, and chapter 2, section 2.2.2.9 for BiB. Code for this analysis was based on the WGCNA single network tutorials available on the WGCNA website (Langfelder & Horvath, 2016), although I altered the code to use a regression model to test each trait, rather than just a correlation, as a regression model is a more powerful tool and can include covariates. The command used to associate phenotypes or cell counts with DNAm modules was:

```
lm(as.matrix(ModuleEigengenes) ~ trait + covariates)
```

Github file path: shwatkins/PhD/WGCNA_analysis/consensus_2.R

6.2.2.7 Preservation of consensus eigengene networks

The preservation of consensus networks cannot be assessed in the same way as single networks. Because consensus networks are created from multiple datasets, which cannot just be combined as a matrix, alternate methods have to be used to those for single network analysis. The method to assess consensus network preservation is to study the relationships between module eigengenes. This is informative because changes in the relationships between module eigengenes may reflect differences in the biological pathways that the modules detect (Langfelder & Horvath, 2007). Methods to do this were developed by the WGCNA team, and I adapted the short command (shown below) available from the consensus network tutorial on the UCLA website (Langfelder & Horvath, 2016). The `WGCNA` function `plotEigengeneNetworks` assesses the preservation of the eigengene network between the constituent datasets of the consensus network, and is a simple command run with settings for the resulting plot:

```
plotEigengeneNetworks(MET, setLabels, marDendro = c(0,2,2,1),
marHeatmap = c(3,3,2,1), zlimPreservation = c(0.5, 1),
xLabelsAngle = 90)
```

Where `MET` is a list of the eigengenes, and `setlabels` are the labels of the datasets. The function plots a dendrogram and heatmap of the relationship between the consensus module eigengenes in each individual dataset, and a preservation bar plot and heatmap to illustrate how the module eigengene relationships are preserved between the datasets.

6.2.3 Gene ontology and KEGG pathway analysis

To assess whether the WGCNA modules were associated with identifiable biological functions, I ran the gene ontology and KEGG pathway enrichment analyses, using the function `gometh` from the R package `missMethyl` (Phipson et al., 2016), as detailed in chapter 2 section 2.7. This was run for each of the BiB single networks, and for ARIES and both groups of BiB for the consensus network.

6.3 Results

6.3.1 Removal of outlying samples: BiB

The number of samples removed after being identified as outliers using hierarchical clustering are detailed in Table 24. The cluster trees with the cut heights are shown in Figure 81 and Figure 82.

Dataset	Number before clustering	Number after clustering
BiB white British	424	416
BiB Pakistani	439	429

Table 24: Number of participants removed during hierarchical clustering of DNAm samples

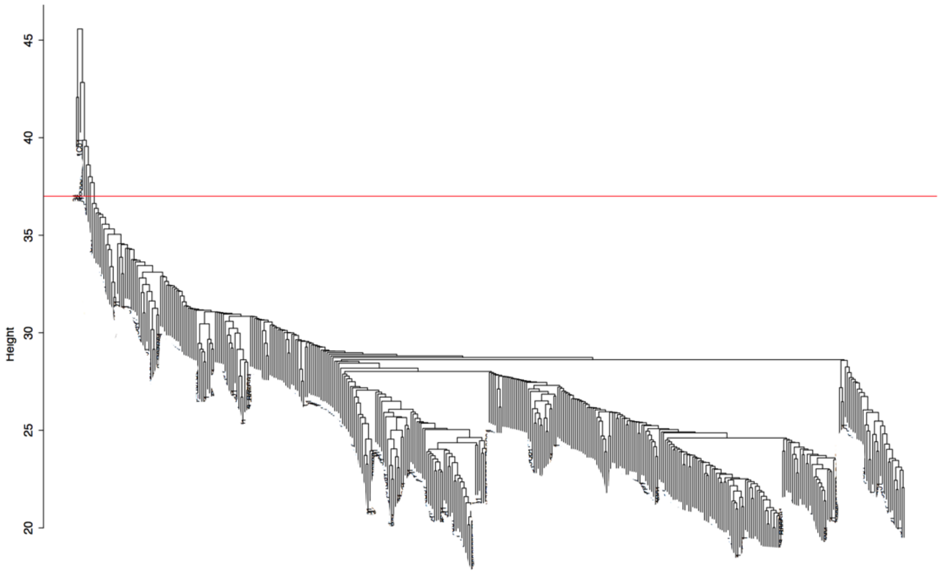


Figure 81: Hierarchical clustering of samples, in the BiB white British group at birth. Red line indicates height at which samples were cut.

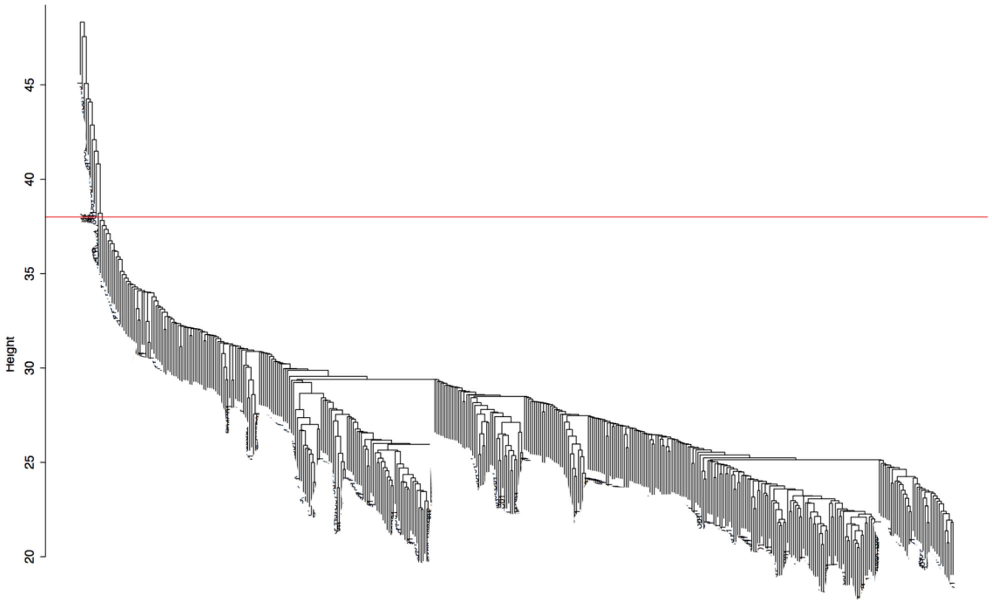


Figure 82: Hierarchical clustering of samples, in the BiB Pakistani group at birth. Red line indicates height at which samples were cut.

6.3.2 Soft power threshold

To construct DNAm networks in BiB and ARIES, pre-processed DNAm data was used to calculate the power at which the data would reach scale-free topology, which is discussed above in section 6.2.2.2. Scale-free topology is reached at a model R2 of 0.9. For the ARIES sample at birth, the soft threshold power was calculated in chapter 4 section 4.5.2 (where it reached scale free topology at a power of 8). The DNAm data for the BiB white British ethnic group reached scale free topology at a power of 6. For the BiB Pakistani group, scale free topology was reached at a power of 5 (see Figure 83 below). Reaching an acceptable soft threshold power should indicate that there is no major driver in the data; the power graphs indicate that this is the case for both the BiB groups.

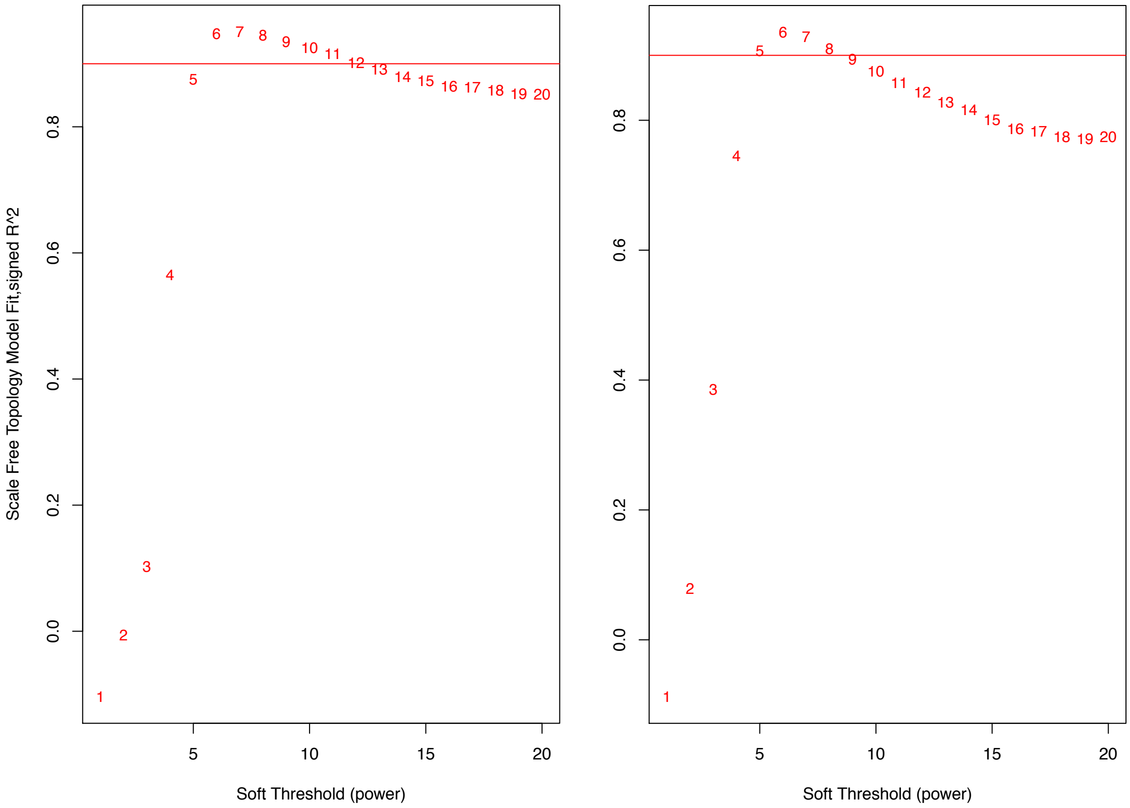


Figure 83: Plots to determine the power at which the network will reach scale-free topology. For the white British group (left)and the Pakistani group (right).

6.3.3 Single network construction

6.3.3.1 White British group

In the white British group the single network analysis resulted in 13 modules, with 127,895 DNAm sites being unassigned to a module. The modules ranged in size from 91,587 to 34 DNAm sites. The modules and their sizes are summarised in Table 25 below.

Module	Number of DNAm sites	Module	Number of DNAm sites	Module	Number of DNAm sites
black	1397	grey (unassigned)	127895	salmon	34
blue	57655	magenta	116	tan	69
brown	29971	pink	674	turquoise	91587
green	21727	purple	77	yellow	29965
greenyellow	71	red	8558		

Table 25: Module sizes and colours in the BiB white British group

6.3.3.2 Pakistani group

In the Pakistani group the network resulted in 12 modules, with 145,255 probes being unassigned to a module. The modules ranged in size from 77,506 to 34 members. The modules and their sizes are summarised in Table 26 below.

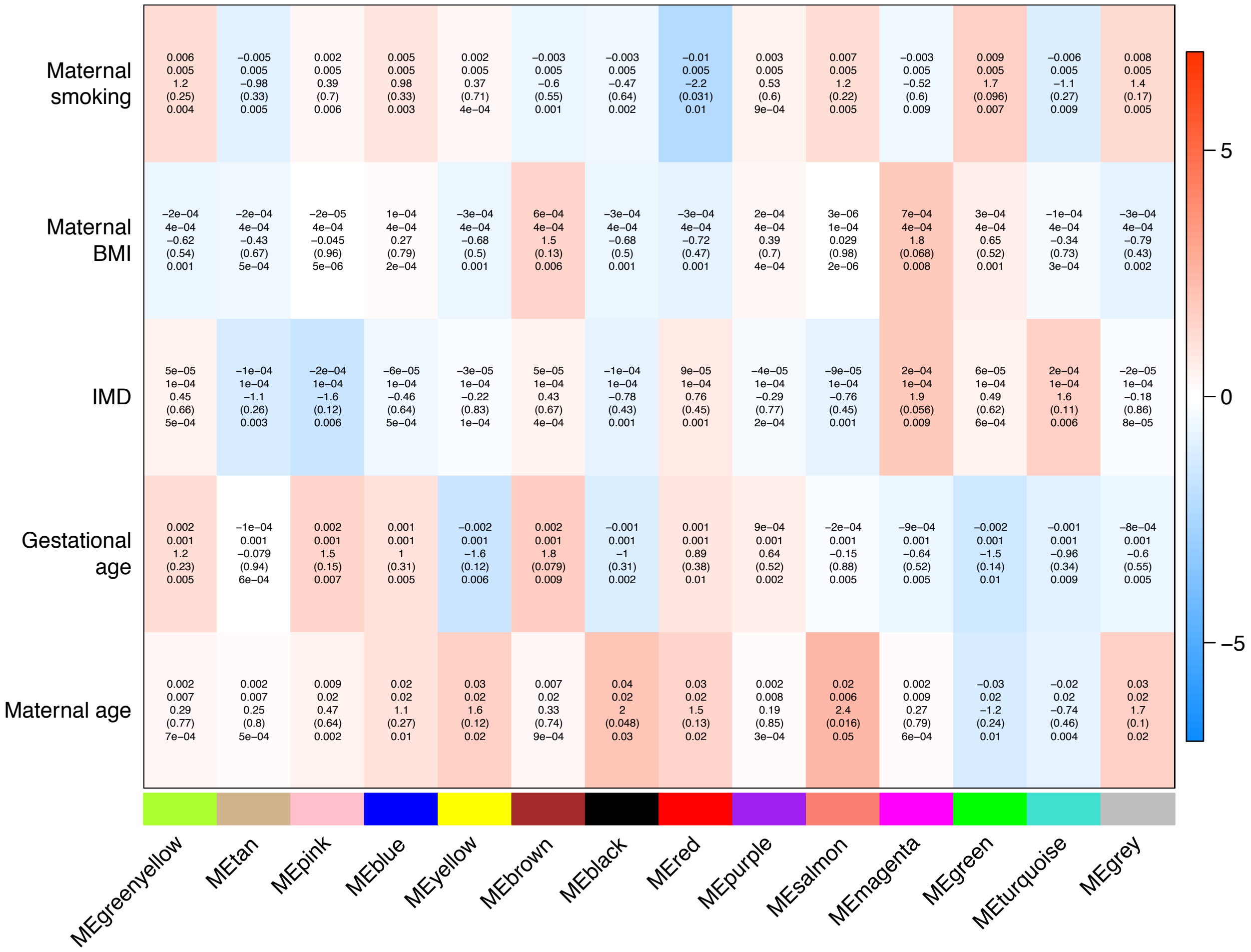
Module	Number of DNAm sites	Module	Number of DNAm sites	Module	Number of DNAm sites
black	9685	grey (unassigned)	145255	tan	50
blue	59431	magenta	1132	turquoise	77506
brown	25200	pink	9125	yellow	21061
green	11440	purple	94		
greenyellow	58	red	9759		

Table 26: Module sizes and colours in the BiB Pakistani group

6.3.4 Association of single network modules with traits

6.3.4.1 White British group

In the white British group, none of the WGCNA modules are associated with the traits tested in this analysis, as illustrated in Figure 84. The lack of association between maternal smoking may be more informative here than in ARIES, because smoking rates in the white British group in BiB were much higher – 34% of the mothers smoked during pregnancy, compared to 10.1% in ARIES (although with twice as many participants, ARIES has more power to detect an effect). This strengthens the idea from chapter 4 that smoking may not alter DNAm in a concerted way at a group of sites (that are present on the 450k array).



	MEgreenyellow	MEtan	MEpink	MEblue	MEyellow	MEbrown	MEblack	MEred	MEpurple	MEsalmon	MEmagenta	MEgreen	MEturquoise	MEgrey
Maternal smoking	0.006 0.005 1.2 (0.25) 0.004	−0.005 0.005 −0.98 (0.33) 0.005	0.002 0.005 0.39 (0.7) 0.006	0.005 0.005 0.98 (0.33) 0.003	0.002 0.005 0.37 (0.71) 4e−04	−0.003 0.005 −0.6 (0.55) 0.001	−0.003 0.005 −0.47 (0.64) 0.002	−0.01 0.005 −2.2 (0.031) 0.01	0.003 0.005 0.53 (0.6) 9e−04	0.007 0.005 1.2 (0.22) 0.005	−0.003 0.005 −0.52 (0.6) 0.009	0.009 0.005 1.7 (0.096) 0.007	−0.006 0.005 −1.1 (0.27) 0.009	0.008 0.005 1.4 (0.17) 0.005
Maternal BMI	−2e−04 4e−04 −0.62 (0.54) 0.001	−2e−04 4e−04 −0.43 (0.67) 5e−04	−2e−05 4e−04 −0.045 (0.96) 5e−06	1e−04 4e−04 0.27 (0.79) 2e−04	−3e−04 4e−04 −0.68 (0.5) 0.001	6e−04 4e−04 1.5 (0.13) 0.006	−3e−04 4e−04 −0.68 (0.5) 0.001	−3e−04 4e−04 −0.72 (0.47) 0.001	2e−04 4e−04 0.39 (0.7) 4e−04	3e−06 1e−04 0.029 (0.98) 2e−06	7e−04 4e−04 1.8 (0.068) 0.008	3e−04 4e−04 0.65 (0.52) 0.001	−1e−04 4e−04 −0.34 (0.73) 3e−04	−3e−04 4e−04 −0.79 (0.43) 0.002
IMD	5e−05 1e−04 0.45 (0.66) 5e−04	−1e−04 1e−04 −1.1 (0.26) 0.003	−2e−04 1e−04 −1.6 (0.12) 0.006	−6e−05 1e−04 −0.46 (0.64) 5e−04	−3e−05 1e−04 −0.22 (0.83) 1e−04	5e−05 1e−04 0.43 (0.67) 4e−04	−1e−04 1e−04 −0.78 (0.43) 0.001	9e−05 1e−04 0.76 (0.45) 0.001	−4e−05 1e−04 −0.29 (0.77) 2e−04	−9e−05 1e−04 −0.76 (0.45) 0.001	2e−04 1e−04 1.9 (0.056) 0.009	6e−05 1e−04 0.49 (0.62) 6e−04	2e−04 1e−04 1.6 (0.11) 0.006	−2e−05 1e−04 −0.18 (0.86) 8e−05
Gestational age	0.002 0.001 1.2 (0.23) 0.005	−1e−04 0.001 −0.079 (0.94) 6e−04	0.002 0.001 1.5 (0.15) 0.007	0.001 0.001 1 (0.31) 0.005	−0.002 0.001 −1.6 (0.12) 0.006	0.002 0.001 1.8 (0.079) 0.009	−0.001 0.001 −1 (0.31) 0.002	0.001 0.001 0.89 (0.38) 0.01	9e−04 0.001 0.64 (0.52) 0.002	−2e−04 0.001 −0.15 (0.88) 0.005	−9e−04 0.001 −0.64 (0.52) 0.005	−0.002 0.001 −1.5 (0.14) 0.01	−0.001 0.001 −0.96 (0.34) 0.009	−8e−04 0.001 −0.6 (0.55) 0.005
Maternal age	0.002 0.007 0.29 (0.77) 7e−04	0.002 0.007 0.25 (0.8) 5e−04	0.009 0.02 0.47 (0.64) 0.002	0.02 0.02 1.1 (0.27) 0.01	0.03 0.02 1.6 (0.12) 0.02	0.007 0.02 0.33 (0.74) 9e−04	0.04 0.02 2 (0.048) 0.03	0.03 0.02 1.5 (0.13) 0.02	0.002 0.008 0.19 (0.85) 3e−04	0.02 0.006 2.4 (0.016) 0.05	0.002 0.009 0.27 (0.79) 6e−04	−0.03 0.02 −1.2 (0.24) 0.01	−0.02 0.02 −0.74 (0.46) 0.004	0.03 0.02 1.7 (0.1) 0.02

*Figure 84: Heatmap of regression association between module eigengenes and traits in the BiB white British group. Data displayed are **beta; standard error; t score; p-value; r-squared.** Abbreviations: BMI (body mass index); IMD (index of multiple deprivation).*

6.3.4.2 Pakistani group

Smoking rates were very low in the Pakistani group, and as no smokers were left after hierarchical clustering to identify and remove outlying samples (see section 6.2.2.1 above), it was not possible to assess the association of maternal smoking with the modules in the children of Pakistani descent. None of the modules had significant associations with any of the traits I selected to test, as illustrated in Figure 85.

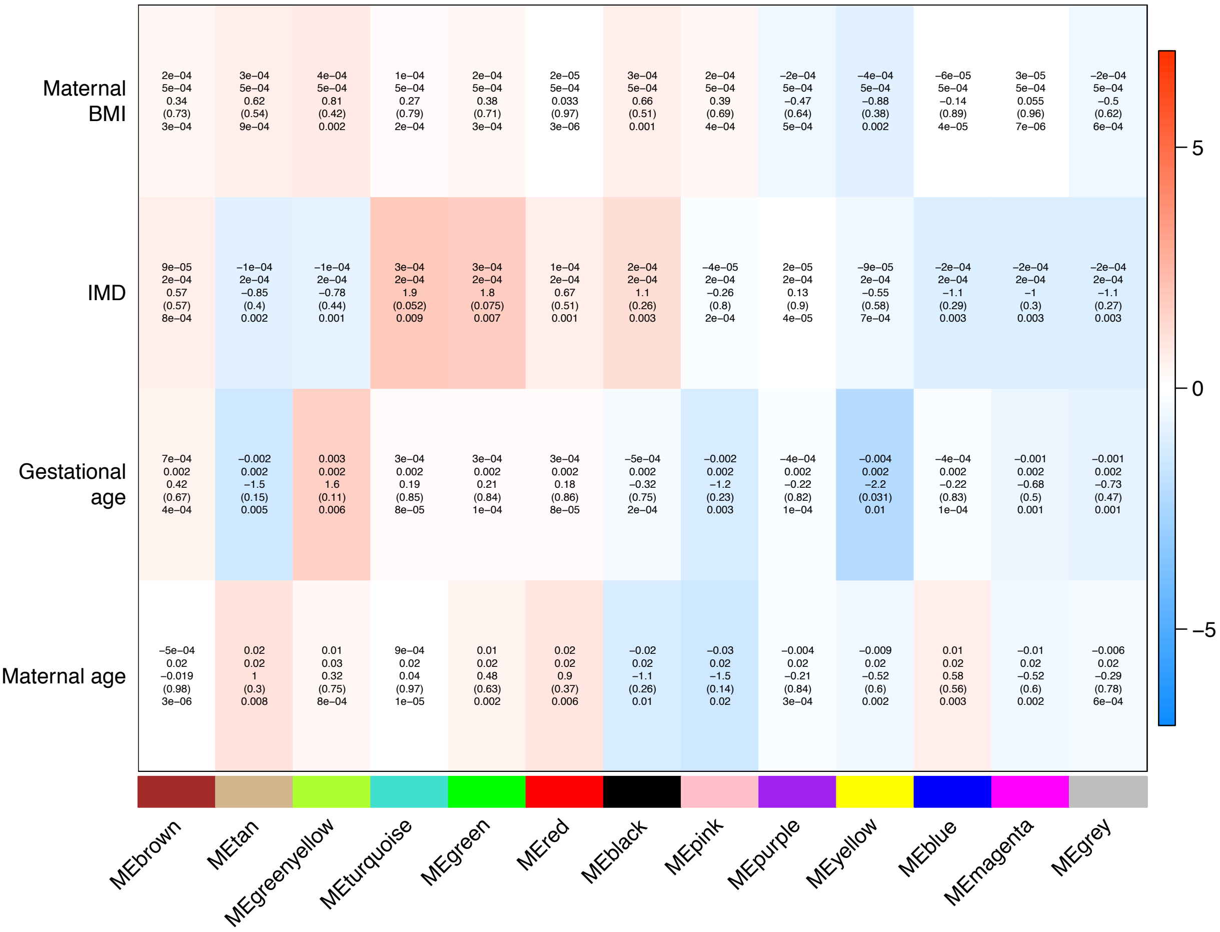


*Figure 85: Heatmap of regression association between module eigengenes and traits in the BiB Pakistani group. Data displayed are **beta; standard error; t score; p-value; r-squared***

6.3.4.3 Preservation of single network modules

6.3.4.3.1 Preservation between ARIES and BiB (Europeans)

Preservation summary statistics have been developed to assess the preservation of modules from one dataset in a separate dataset. Here I assessed the preservation of modules identified in ARIES at birth in the BiB white British ethnic group. The influence of module size on Zsummary is quite evident from Figure 86(B). All modules bar the MEdarkturquoise have strong preservation (>10) according to the Zsummary. However using the median rank, some of the smaller modules appear to be better preserved (which includes the MEdarkturquoise module). Looking at the median rank, the top preserved modules in the dataset tend to be modules that are not enriched for functional annotations in ARIES.

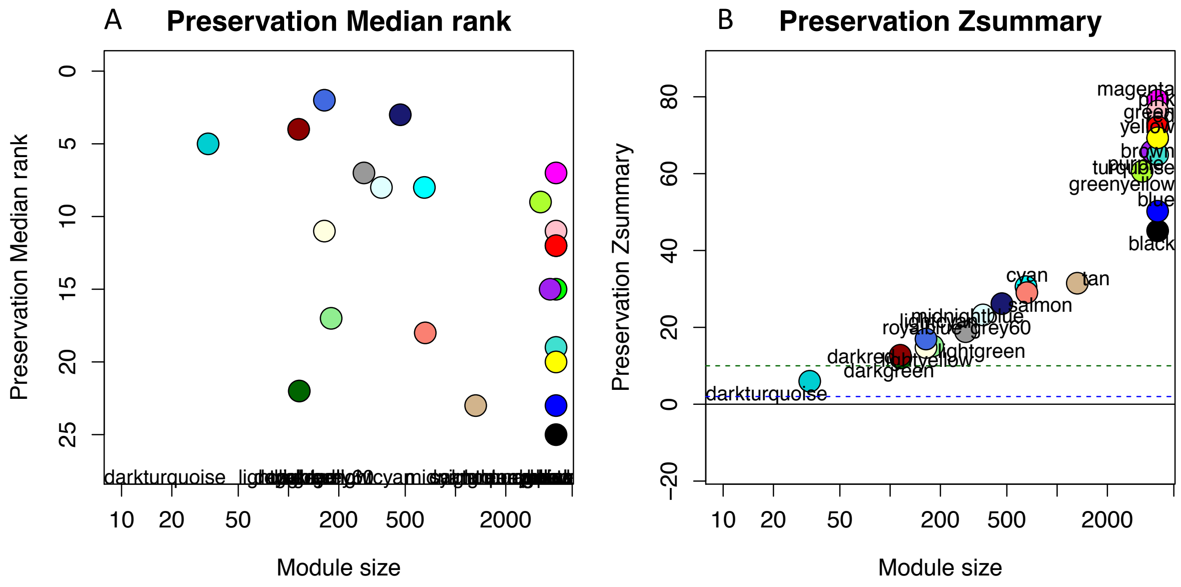


Figure 86: Plots of preservation summary statistics, preservation median rank and Zsummary, for preservation of ARIES modules in the BiB white British group

6.3.4.3.2 Preservation between ARIES and BiB (trans-ethnicity)

The first thing about the preservation plots in Figure 87 is that they are quite strikingly similar to those between ARIES and the white British BiB participants. Again most modules are very well preserved using the module preservation statistics, and those that have functional annotations in ARIES tend to be toward the bottom of the median rank (shown in Figure 87).

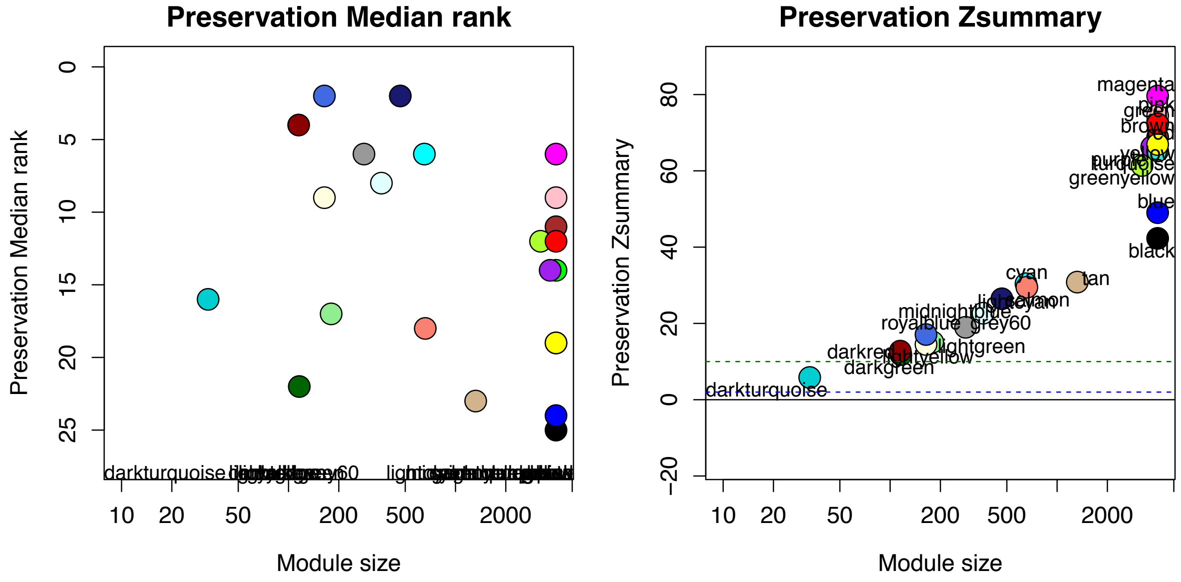


Figure 87: Plots of preservation summary statistics, preservation median rank and Zsummary, for preservation of ARIES modules in the BiB Pakistani group

6.3.5 Consensus network

6.3.5.1 Consensus network modules

A consensus network analysis was conducted using common DNAm sites between ARIES and both BiB ethnic groups, to ascertain whether there are DNAm modules which persist in all three datasets. 23 modules of co-methylated DNAm sites, plus one group of 'unassigned' DNAm sites, were detected in the consensus analysis. The first thing to note is that the unassigned MEgrey module contains only 15,997 DNAm sites. This is in stark contrast to the single networks for both BiB and ARIES, which all have over 100,000 DNAm sites which are unassigned to a module. It is likely the grey module is so much smaller in this consensus network because of the much greater sample size; the consensus network is double the size of ARIES, with 1,674 participants. Greater sample size will create more variation in the data, which will create more potential for correlation, and greater sample size will create more power to find relationships between DNAm sites. The consensus module sizes are listed below in Table 27.

Consensus module	Number of DNAm sites	Consensus module	Number of DNAm sites	Consensus module	Number of DNAm sites
black	20450	greenyellow	2483	pink	4762
blue	88140	grey (unassigned)	15997	purple	2508
brown	42007	grey60	349	red	25677
cyan	977	lightcyan	378	royalblue	192
darkgreen	46	lightgreen	305	salmon	1050
darkred	115	lightyellow	281	tan	1701
darkturquoise	30	magenta	4274	turquoise	88744
green	28858	midnightblue	488	yellow	38281

Table 27: Consensus module sizes and colours in the multi-ethnic consensus network between ARIES and BiB.

6.3.5.2 Association of the consensus network module with phenotypes

With consensus WGCNA networks, individuals in each dataset receive a score for each consensus module eigengene (much like in the single networks). These module eigengene scores can then be associated with traits, as I have done for the single network analyses. The advantage of testing these same traits with the consensus network is that the consensus network may have more power to detect DNAm co-methylation modules, which may enable detection of a module-trait relationship that is not possible to detect in a smaller sample.

6.3.5.2.1 ARIES

In ARIES, there is an association between asthma at 7 years and the MEdarkturquoise consensus module that almost survives correction for multiple testing (p=0.00033, effect size = -0.02, se=0.004, variance explained =0.02; Bonferroni adjusted threshold = 0.00031). None of the other modules have associations with the traits tested. The module-trait relationship heatmap can be found in Figure 88.

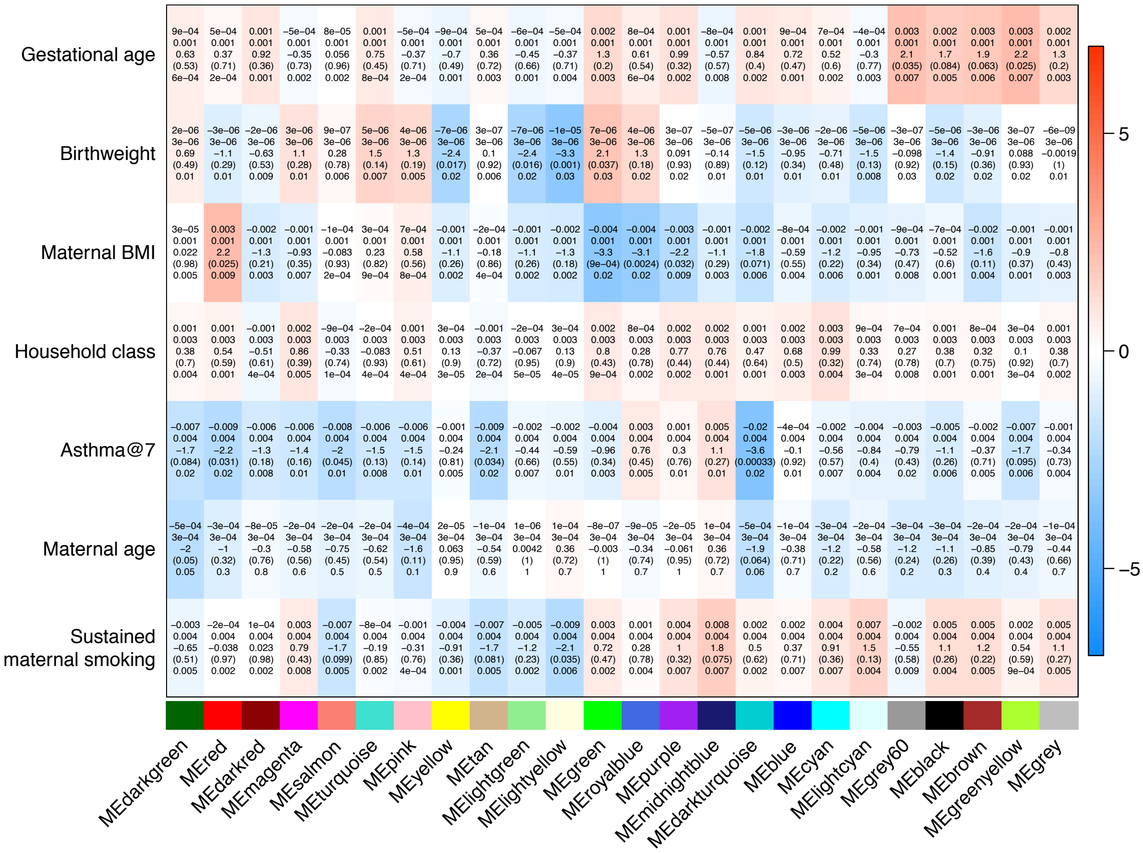
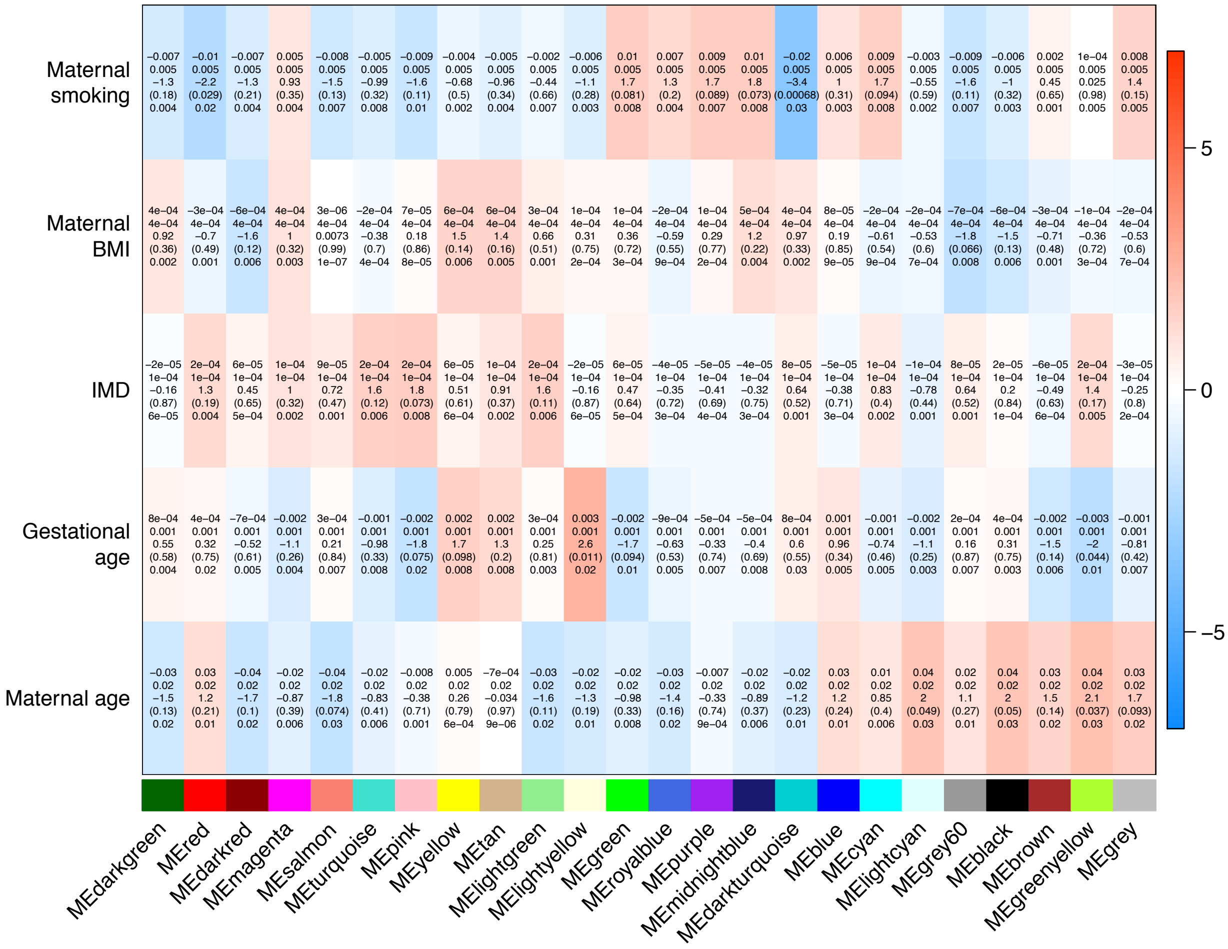


*Figure 88: Heatmap of regression association between consensus module eigengenes and traits in ARIES. Data displayed are **beta; standard error; t score; p-value; r-squared.** Abbreviations: Asthma@7 (asthma measured at 7 years old); BMI (body mass index).*

6.3.5.2.2 BiB white British

In the BiB white British ethnic group, the MEdarkturquoise consensus module is close to surviving multiple testing for association with maternal smoking (p=0.00068, effect size=-0.02, se=0.005, variance explained=0.03; Bonferroni corrected p value = 0.00043). This provides the possibility that maternal smoking during pregnancy could affect pathways of

DNAm, but it may be that as the effect is small it requires a larger sample size to detect it. This model explains a very small amount of variance in the phenotype. The module-trait relationship heatmap can be found in Figure 89.



	MEdarkgreen	MEred	MEdarkred	MEmagenta	MEsalmon	MEturquoise	MEpink	MEyellow	MEtan	MElightgreen	MElightyellow	MEgreen	MEroyalblue	MEpurple	MEmidnightblue	MEdarkturquoise	MEblue	MEcyan	MElightcyan	MEgrey60	MEblack	MEbrown	MEgreenyellow	MEgrey
Maternal smoking	−0.007 0.005 −1.3 (0.18) 0.004	−0.01 0.005 −2.2 (0.029) 0.02	−0.007 0.005 −1.3 (0.21) 0.004	0.005 0.005 0.93 (0.35) 0.004	−0.008 0.005 −1.5 (0.13) 0.007	−0.005 0.005 −0.99 (0.32) 0.008	−0.009 0.005 −1.6 (0.11) 0.01	−0.004 0.005 −0.68 (0.5) 0.002	−0.005 0.005 −0.96 (0.34) 0.004	−0.002 0.005 −0.44 (0.66) 0.007	−0.006 0.005 −1.1 (0.28) 0.003	0.01 0.005 1.7 (0.081) 0.008	0.007 0.005 1.3 (0.2) 0.004	0.009 0.005 1.7 (0.089) 0.007	0.01 0.005 1.8 (0.073) 0.008	−0.02 0.005 −3.4 (0.00068) 0.03	0.006 0.005 1 (0.31) 0.003	0.009 0.005 1.7 (0.094) 0.008	−0.003 0.005 −0.55 (0.59) 0.002	−0.009 0.005 −1.6 (0.11) 0.007	−0.006 0.005 −1 (0.32) 0.003	0.002 0.005 0.45 (0.65) 0.001	1e−04 0.005 0.025 (0.98) 0.005	0.008 0.005 1.4 (0.15) 0.005
Maternal BMI	4e−04 4e−04 0.92 (0.36) 0.002	−3e−04 4e−04 −0.7 (0.49) 0.001	−6e−04 4e−04 −1.6 (0.12) 0.006	4e−04 4e−04 1 (0.32) 0.003	3e−06 4e−04 0.0073 (0.99) 1e−07	−2e−04 4e−04 −0.38 (0.7) 4e−04	7e−05 4e−04 0.18 (0.86) 8e−05	6e−04 4e−04 1.5 (0.14) 0.006	6e−04 4e−04 1.4 (0.16) 0.005	3e−04 4e−04 0.66 (0.51) 0.001	1e−04 4e−04 0.31 (0.75) 2e−04	1e−04 4e−04 0.36 (0.72) 3e−04	−2e−04 4e−04 −0.59 (0.55) 9e−04	1e−04 4e−04 0.29 (0.77) 2e−04	5e−04 4e−04 1.2 (0.22) 0.004	4e−04 4e−04 0.97 (0.33) 0.002	8e−05 4e−04 0.19 (0.85) 9e−05	−2e−04 4e−04 −0.61 (0.54) 9e−04	−2e−04 4e−04 −0.53 (0.6) 7e−04	−7e−04 4e−04 −1.8 (0.066) 0.008	−6e−04 4e−04 −1.5 (0.13) 0.006	−3e−04 4e−04 −0.71 (0.48) 0.001	−1e−04 4e−04 −0.36 (0.72) 3e−04	−2e−04 4e−04 −0.53 (0.6) 7e−04
IMD	−2e−05 1e−04 −0.16 (0.87) 6e−05	2e−04 1e−04 1.3 (0.19) 0.004	6e−05 1e−04 0.45 (0.65) 5e−04	1e−04 1e−04 1 (0.32) 0.002	9e−05 1e−04 0.72 (0.47) 0.001	2e−04 1e−04 1.6 (0.12) 0.006	2e−04 1e−04 1.8 (0.073) 0.008	6e−05 1e−04 0.51 (0.61) 6e−04	1e−04 1e−04 0.91 (0.37) 0.002	2e−04 1e−04 1.6 (0.11) 0.006	−2e−05 1e−04 −0.16 (0.87) 6e−05	6e−05 1e−04 0.47 (0.64) 5e−04	−4e−05 1e−04 −0.35 (0.72) 3e−04	−5e−05 1e−04 −0.41 (0.69) 4e−04	−4e−05 1e−04 −0.32 (0.75) 3e−04	8e−05 1e−04 0.64 (0.52) 0.001	−5e−05 1e−04 −0.38 (0.71) 3e−04	1e−04 1e−04 0.83 (0.4) 0.002	−1e−04 1e−04 −0.78 (0.44) 0.001	8e−05 1e−04 0.64 (0.52) 0.001	2e−05 1e−04 0.2 (0.84) 1e−04	−6e−05 1e−04 −0.49 (0.63) 6e−04	2e−04 1e−04 1.4 (0.17) 0.005	−3e−05 1e−04 −0.25 (0.8) 2e−04
Gestational age	8e−04 0.001 0.55 (0.58) 0.004	4e−04 0.001 0.32 (0.75) 0.02	−7e−04 0.001 −0.52 (0.61) 0.005	−0.002 0.001 −1.1 (0.26) 0.004	3e−04 0.001 0.21 (0.84) 0.007	−0.001 0.001 −0.98 (0.33) 0.008	−0.002 0.001 −1.8 (0.075) 0.02	0.002 0.001 1.7 (0.098) 0.008	0.002 0.001 1.3 (0.2) 0.008	3e−04 0.001 0.25 (0.81) 0.003	0.003 0.001 2.6 (0.011) 0.02	−0.002 0.001 −1.7 (0.094) 0.01	−9e−04 0.001 −0.63 (0.53) 0.005	−5e−04 0.001 −0.33 (0.74) 0.007	−5e−04 0.001 −0.4 (0.69) 0.008	8e−04 0.001 0.6 (0.55) 0.03	0.001 0.001 0.96 (0.34) 0.005	−0.001 0.001 −0.74 (0.46) 0.005	−0.002 0.001 −1.1 (0.25) 0.003	2e−04 0.001 0.16 (0.87) 0.007	4e−04 0.001 0.31 (0.75) 0.003	−0.002 0.001 −1.5 (0.14) 0.006	−0.003 0.001 −2 (0.044) 0.01	−0.001 0.001 −0.81 (0.42) 0.007
Maternal age	−0.03 0.02 −1.5 (0.13) 0.02	0.03 0.02 1.2 (0.21) 0.01	−0.04 0.02 −1.7 (0.1) 0.02	−0.02 0.02 −0.87 (0.39) 0.006	−0.04 0.02 −1.8 (0.074) 0.03	−0.02 0.02 −0.83 (0.41) 0.006	−0.008 0.02 −0.38 (0.71) 0.001	0.005 0.02 0.26 (0.79) 6e−04	−7e−04 0.02 −0.034 (0.97) 9e−06	−0.03 0.02 −1.6 (0.11) 0.02	−0.02 0.02 −1.3 (0.19) 0.01	−0.02 0.02 −0.98 (0.33) 0.008	−0.03 0.02 −1.4 (0.16) 0.02	−0.007 0.02 −0.33 (0.74) 9e−04	−0.02 0.02 −0.89 (0.37) 0.006	−0.02 0.02 −1.2 (0.23) 0.01	0.03 0.02 1.2 (0.24) 0.01	0.01 0.02 0.85 (0.4) 0.006	0.04 0.02 2 (0.049) 0.03	0.02 0.02 1.1 (0.27) 0.01	0.04 0.02 2 (0.05) 0.03	0.03 0.02 1.5 (0.14) 0.02	0.04 0.02 2.1 (0.037) 0.03	0.03 0.02 1.7 (0.093) 0.02

*Figure 89: Heatmap of regression association between consensus module eigengenes and traits in the BiB white British ethnic group. Data displayed are **beta; standard error; t score; p-value; r-squared.** Abbreviations: BMI (body mass index); IMD (index of multiple deprivation).*

6.3.5.2.3 BiB Pakistani

None of the consensus network modules are associated with the traits tested in the Pakistani group of BiB. The module-trait relationship heatmap can be found in Figure 90.

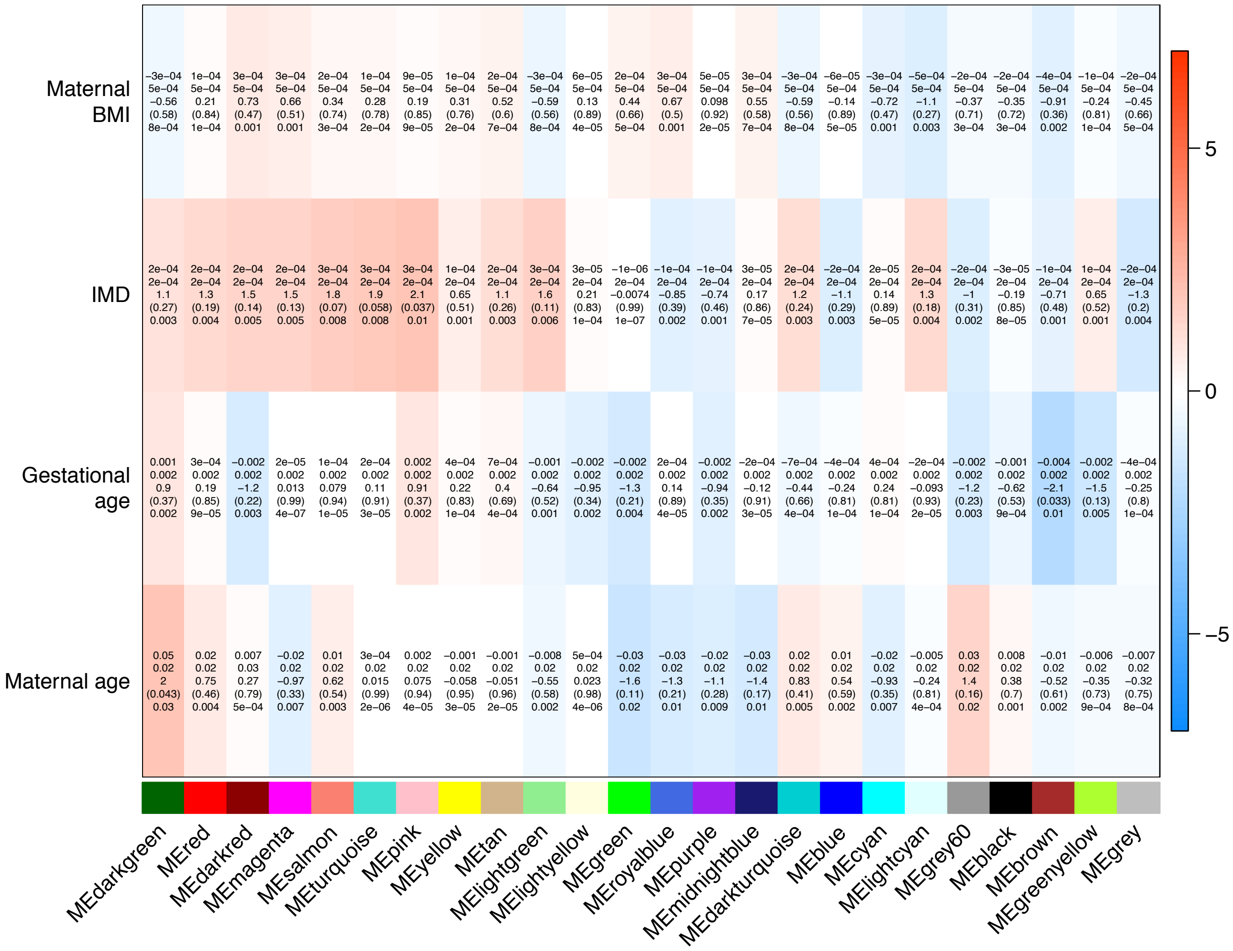


*Figure 90: Heatmap of regression association between consensus module eigengenes and traits in the BiB white British ethnic group. Data displayed are **beta; standard error; t score; p-value; r-squared.** Abbreviations: BMI (body mass index); IMD (index of multiple deprivation).*

6.3.5.2.4 Eigengene network preservation

In WGCNA, eigengenes are a mathematical representation of network modules, essentially equal to the first principal component of the module. From the consensus network analysis, 23 modules of DNAm sites that are highly co-methylated within the three datasets were identified. As with single WGCNA network analyses, these modules are not necessarily co-methylated in all the participants, so some modules may be more specific to particular datasets. The eigengene preservation analysis seeks to identify whether relationships between module eigengenes is preserved between datasets. This illustrates whether there is higher order organisation in methylation modules, and so whether there might be coordinated regulation between them. A lack of preservation may indicate biological differences or perturbations, and provides the illustration of how similar the modules are between the datasets.

The eigengene network preservation analysis shows that there is broadly similar clustering of module eigengenes in all three datasets (Figure 91 D, H and L). The clustering is more similar between the two ethnic groups of BiB than between the white British participants in BiB and ARIES. This is also reflected in the higher preservation of eigengene relationships between the two ethnic groups of BiB (Figure 91 I and K) than between the white British participants in BiB and ARIES (Figure 91 E and G). The overall network preservation summary statistic D (which is an aggregate measure showing the preservation of adjacency between the datasets (Langfelder & Horvath, 2007)) shows that the preservation between the two ethnic groups of BiB is very high (D=0.97). The preservation between ARIES and the white British group of BiB (D=0.93), and between ARIES and the Pakistani group of BiB (D=0.92), is still quite high which means that DNAm co-methylation structure is generally highly preserved, with a few exceptions (nicely illustrated by the white areas on the preservation heatmaps between ARIES and the white British (Figure 91 G) and Pakistani (Figure 91 J) participants of BiB).

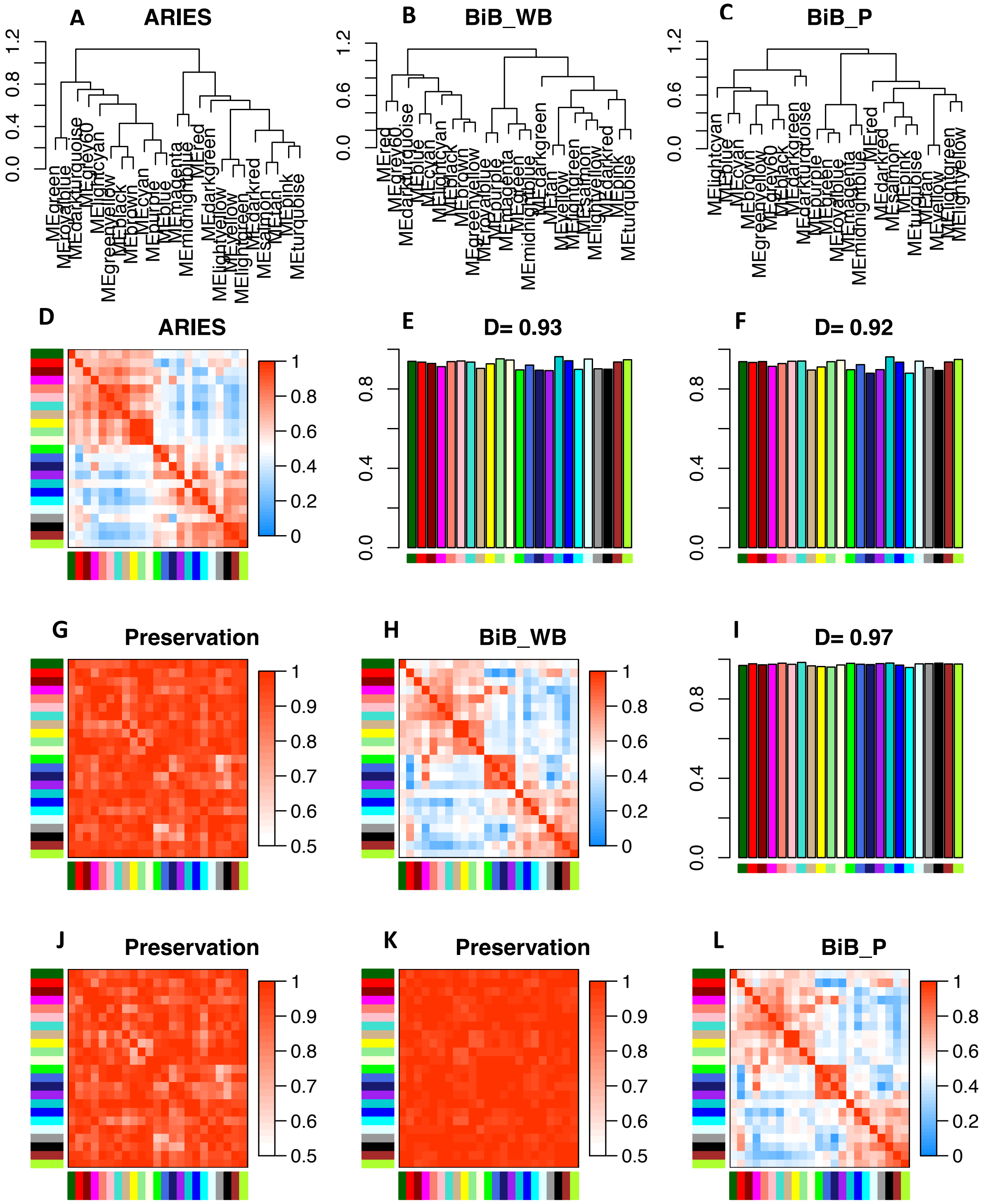


Figure 91: Eigengene network preservation in ARIES and both the BiB ethnic groups. ***A-C****: eigengene dendrograms, showing the clustering of the module eigengenes each individual dataset. The heatmaps (****D, H*** *and* ***L****) also illustrate the relationships between the module eigengenes in the individual datasets.* ***E*** *and* ***G*** *illustrate the preservation of the relationships between eigengenes between ARIES and the BiB white British group;* ***F*** *and* ***J*** *illustrate this between ARIES and the BiB Pakistani group;* ***I*** *and* ***K*** *illustrate this between the two BiB ethnic groups. D, above the barplots, is an overall measure of preservation of the module relationships.*

6.3.5.2.5 Functional annotation of consensus network

I conducted a gene ontology analysis to identify whether any of the consensus modules had clear biological functions. For each of the three groups separately, module membership was calculated for all DNAm sites for all modules. This means module membership is not exclusive, which is much more representative of biology. For each group, the top module members (kME>0.7) were assessed for gene ontology as compared to a background of all probes used in the analysis. Modules which had ontologies associated close to or below FDR p<0.05 are presented in Table 28; the top 20 gene ontology and KEGG enrichments are in Appendix 6 if there was an FDR significant association for any of the three datasets.

There are two modules which are strongly enriched for gene ontology terms relating to *intracellular organelles* in all three datasets. The MEturquoise module is strongly enriched in all three datasets for gene ontology terms relating to *intracellular organelles*. In both BiB groups, it is enriched for KEGG pathways relating to *RNA transport and degradation*, although this GO term does not near FDR significance in ARIES. As such it is hard to say whether this is a co-methylation module relating to genes with functions relating to RNA activity. The MEpink module is also (slightly less) strongly enriched in all three datasets for gene ontology terms relating to *intracellular organelles*. However the MEpinkmodule is not enriched for any KEGG terms, so it is not clear what the organelles might be.

In ARIES, the MEblue module is enriched for GO terms relating to *detecting chemical stimuli*, and fairly strongly enriched for the KEGG *olfactory transduction* pathway. However neither of the BiB groups feature these enrichments. It is possible this module is ARIES-specific.

The MEgrey60 module is associated with immune-related GO terms in the BiB Pakistani group, and with immune-related KEGG pathway terms in both the BiB ethnicities, suggesting this could represent a module to do with either immune function, or be related to cellular proportions. The MEmagenta is also associated with immune-related GO terms in the BiB white British ethnic group, but not in the other groups.

The MElightgreen module is associated with GO terms FDR p<0.05 in all datasets; however the GO terms are relatively different, so it is hard to know if these enrichments are meaningful.

The MEred module is moderately associated with the GO term *RNA Polymerase II activity* in both of the BiB ethnic groups. This term just misses FDR significance in ARIES (FDR p-value 0.45). It is possible that this represents a module which reflects the highly *trans*-correlating DNAm sites which are enriched for Pol2 binding sites, that were identified in both ARIES and BiB chapters 3 and 5. It may be that this GO term did not reach significance because of the greater similarities in module eigengene structure between the two groups of BiB.

The MEbrown module from the consensus analysis is quite weakly associated with gene ontology terms to do with *regulation of anatomical structure morphogenesis* in ARIES. As it does not pass the FDR significance threshold of 0.05 for ARIES, and the module is not associated with this GO term in BiB, it may be that the module with this weak association in all the ARIES timepoints is not preserved across datasets. This module is not in Table 28 as it did not pass the FDR threshold, but it can be found in Appendix 6.

Consensus module	**Cohort (n DNAm sites kME>0.7)**	**Gene ontology**	**FDR P-value**
Blue	ARIES (5,034)	Detection of chemical stimulus involved in sensory perception of smell Olfactory receptor activity Detection of chemical stimulus involved in sensory perception	0.006 0.006 0.006
Dark red	BiB Pakistani (14)	Sin3 complex Sin3-type complex	0.02 0.02
Grey60	BiB Pakistani (14)	T cell costimulation lymphocyte costimulation alpha-beta T cell receptor complex	0.03 0.03 0.03
Light green	ARIES (189)	homophilic cell adhesion via plasma membrane adhesion molecules cell-cell adhesion via plasma-membrane adhesion molecules	8.78E-05 0.01
Light green	BiB white British (4)	negative regulation of wound healing, spreading of epidermal cells regulation of wound healing, spreading of epidermal cells	0.001 0.001

		dystroglycan binding	0.002
Light green	BiB Pakistani (44)	organic cyclic compound biosynthetic process	0.04
Magenta	BiB white British (543)	neutrophil degranulation	0.0002
		neutrophil activation involved in immune response	0.0002
		neutrophil activation	0.0002
Pink	ARIES (1657)	intracellular membrane-bounded organelle	3.12E-10
		intracellular part	1.05E-09
		intracellular organelle	9.79E-09
Pink	BiB white British (3688)	intracellular membrane-bounded organelle	5.49E-21
		intracellular organelle	4.50E-19
		intracellular part	6.42E-18
Pink	BiB Pakistani (2308)	intracellular membrane-bounded organelle	5.88E-14
		nucleus	1.44E-11
		intracellular organelle	8.88E-11
Purple	BiB white British (186)	Rho protein signal transduction	0.08
		small GTPase mediated signal transduction	0.08
		Cytosol	0.1
Purple	BiB Pakistani (106)	small GTPase binding	0.06
		small GTPase mediated signal transduction	0.06
Red	BiB white British (618)	RNA polymerase II transcription factor activity, sequence-specific DNA binding	8.49E-11
		DNA binding transcription factor activity	2.06E-08
		transcription regulator activity	4.72E-06
Red	BiB Pakistani (494)	RNA polymerase II transcription factor activity, sequence-specific DNA binding	6.95E-06
		multicellular organism development	9.55E-05
		multicellular organismal process	9.55E-05
Tan	BiB white British (213)	intracellular membrane-bounded organelle	0.04
		membrane-bounded organelle	0.05
		cellular protein metabolic process	0.09
Turquoise	ARIES (6054)	intracellular membrane-bounded organelle	1.72E-31
		intracellular part	7.05E-31

		intracellular organelle	1.27E-30
Turquoise	BiB white British (11897)	intracellular part intracellular organelle intracellular membrane-bounded organelle	5.45E-55 1.17E-53 2.91E-53
Turquoise	BiB Pakistani (11500)	intracellular membrane-bounded organelle intracellular part intracellular organelle	1.66E-58 1.68E-58 3.08E-57
Yellow	ARIES (598)	intracellular membrane-bounded organelle intracellular part transferase complex, transferring phosphorus-containing groups	0.05 0.05 0.05
Yellow	BiB white British (465)	cellular metabolic process metabolic process cellular catabolic process	0.002 0.002 0.004

Table 28: Summary of up to the top 3 gene ontology terms for modules that were close to, or below, FDR p<0.05, in the trans-ethnicity consensus network.

Consensus module	Cohort (n DNAm sites kME>0.7)	KEGG pathway	FDR P-value
Blue	ARIES (5,034)	Olfactory transduction	3.48E-05
Grey60	BiB white British (16)	Th1 and Th2 cell differentiation	0.09
		Natural killer cell mediated cytotoxicity	0.09
		Th17 cell differentiation	0.09
Grey60	BiB Pakistani (14)	Th1 and Th2 cell differentiation	0.0009
		Th17 cell differentiation	0.0009
		PD-L1 expression and PD-1 checkpoint pathway in cancer	0.0009
Magenta	BiB white British (543)	Amoebiasis	0.06
		Hematopoietic cell lineage	0.06
		Metabolic pathways	0.06
Tan	ARIES (715)	Pyruvate metabolism	0.1
Turquoise	BiB white British (11897)	Spliceosome	0.02
		RNA degradation	0.02
		Viral carcinogenesis	0.04

Turquoise	BiB Pakistani (11500)	RNA transport	0.03
		Apoptosis	0.03
		RNA degradation	0.03

Table 29: Summary of up to the top 3 KEGG pathway terms for modules that were close to, or below, FDR p<0.05, in the trans-ethnicity consensus network.

6.4 Discussion

6.4.1 Summary of findings

This Chapter shows that correlation network analysis can identify stable, biologically relevant groups of highly co-methylated DNAm sites across two population cohorts of newborns. It also shows that co-methylation is very strongly preserved across ethnicities; more so than across datasets.

6.4.2 DNAm networks in BiB

I created single DNAm networks in BiB to investigate whether there might be any ethnicity-specific associations between DNAm modules and the phenotypes I selected to test. The networks are about half the size of those in ARIES, in that they have fewer modules. As there are over 20 modules in the consensus network, it is likely the smaller module numbers in the BiB standalone networks is due to a lack of power. What is also interesting in terms of network structure is the small number of DNAm sites (just over 15,000) that are not assigned to a module in the consensus analysis. This is a large departure from the >100,000 that are unassigned in the single networks. This could illustrate the need for larger sample sizes in DNAm network analysis when a population approach is taken, as opposed to case-control. It may also illustrate the power of consensus analysis in combining datasets to find biological signals.

6.4.3 Association of consensus network modules with phenotypes

The consensus network analysis identified a module, MEdarkturquoise, that was close to being associated with future asthma in ARIES, and maternal smoking in the white British individuals in BiB. As these findings did not pass the multiple testing threshold we cannot say for certain whether they represent true associations; however the associations of asthma and smoking in the same module is interesting because these phenotypes are relatively interconnected (DiFranza, Aligne, & Weitzman, 2004), to the extent that maternal smoking is controlled for in DNAm studies relating to asthma (Reese et al., 2019). It is possible this module could represent DNAm relating to confounding factors associated with both maternal smoking and asthma; it could be that it represents confounding by cell type proportions; or it could even be that the module represents some biological mediation between the two phenotypes (although I do not find good evidence for that). This consensus analysis had a relatively large sample size for DNAm, especially compared to many published WGCNA network analyses. It is possible that because of the population nature of this study, rather than case-control, phenotypes that are of relatively low frequency in a population may need larger sample sizes to be detected. This might explain why associations with these phenotypes did not show up in the single network analyses, which have lower power; so further analysis could look to include other cohorts with relevant DNAm data. It might also explain why a smoking association would show up in BiB rather than ARIES, as BiB had a much higher rate of smoking.

6.4.4 Functional enrichments of consensus DNAm modules

The MEgrey60 module is associated with immune-related GO terms in the BiB Pakistani group, and with immune-related KEGG pathway terms in both the BiB ethnicities. This could represent the different cell type panels used in the two cohorts; or it could represent an immune activation pathway that we do not see in ARIES for some reason.

The enrichment of the MEblue module in ARIES for GO terms relating to *detecting chemical stimuli*, and the KEGG *olfactory transduction* pathway, is interesting as these terms overlap. Epigenetic marks, including methylation, have been identified as having a key role in the control of olfactory receptor expression (X. J. Tian, Zhang, Sannerud, & Xing, 2016). This seems like quite a clear association, although it is unclear why no module at birth in the ARIES single network (chapter 4) was associated with these functions (either at birth or the other timepoints), and it is presumably not an issue of power that detected a module with this function in the consensus network, as the functions are not enriched in the MEblue module in either of the BiB groups.

The MEred module is enriched for GO terms relating to RNA polymerase II in both the BiB groups. The enrichment did not pass FDR significance in ARIES, but the term is further down the table. As a module enriched for these terms is found in the single DNAm networks in ARIES at birth, it adds weight to the idea that this module might have common functions in all three datasets; however it is possible that the between-cohort differences induce changes in the module structure between ARIES and BiB. As Pol2 is one of the key transcription factor binding site enrichments for the strongly correlated trans DNAm sites in Chapters 3 and 5, overall this seems like a clear demonstration of these highly correlated sites functioning in a interconnected network.

The MEturquoise module in the consensus network is very strongly associated with GO terms to do with *intracellular organelles* for all three datasets. It is notable how much smaller the enrichments are in ARIES, which is likely to be reflective of the different network structure relative to the two groups of BiB. In BiB this module is enriched for KEGG pathways relating to *RNA transport* and *degradation,* and *splicosome.* This may point to a function of this module in the regulation of alternative splicing, in which DNAm has an established role (Oberdoerffer, 2012; Shukla et al., 2011; Yearim et al., 2015).

6.4.5 Preservation of consensus eigengene network structure

The strong trans-ethnicity preservation of module eigengene relationships between the two groups of BiB is quite striking. This reinforces the findings of Chapter 5 that trans-ethnicity correlations are more similar than different cohorts of the same ethnicity. The network analysis adds a new dimension to this, because whilst the correlation analysis identifies similarities in pairwise associations, the network looks at the big picture connectivity between all sites. This connectivity across the genome (as measured by the 450k) is extremely well preserved, and points towards stable and persistent relationships between DNAm that are likely to have consistent biological functions. We see this in the functional annotations of the modules, where BiB are often enriched for the same terms, and less so with ARIES. The eigengene preservation between ARIES and both groups of BiB is still high; and is virtually the same with both of the ethnicities of BiB, suggesting the differences are likely to be cohort, platform, or batch based.

6.4.6 Summary

This chapter has illustrated the high degree of preservation between DNAm networks in trans-cohort and trans-ethnicity analysis. This preservation points to a highly conserved network structure in DNAm data that might be termed a normative network. Modules in this network have identified and relevant biological functions, and point to the strong utility of network analysis to infer biological pathways in which DNAm is involved. The importance of establishing a normative network structure lies in having this as a reference for DNAm activity; with this reference we can then identify if differences we find between diseases or exposures are perturbing a pathway of normal function, which would provide powerful mechanistic insights beyond what can be achieved with single site analyses.

7 Discussion

7.1 Summary of findings

This thesis has demonstrated that DNAm data has a stable correlation structure, both in cis and in trans, that persists over time, across datasets, and across ethnicity. I have shown that highly correlating DNAm sites are enriched for functional annotations, which demonstrate that co-methylation is likely to be an important biological process. I have shown how correlations in trans are fundamentally different to those in cis, with differing functional annotations. I have also demonstrated for the first time in humans that trans-correlating DNAm sites are likely to represent inter-chromosomal contacts, and thus they are likely to represent shared regulation. Correlation between DNAm sites in inter- and intra-chromosomal chromatin contact regions was demonstrated in mice in a recent paper (G. Li et al., 2019).

The stable and functional correlation structure I have illustrated shows that co-methylation networks of DNAm are likely to illuminate functional biological pathways that DNAm is involved in. To identify whether this can indeed be shown using DNAm array data, I created DNAm co-methylation networks which I show to be broadly well preserved across development, cohorts, and ethnicity; I show that co-methylation of DNAm is in fact more preserved across ethnicity than across cohorts; I show that a number of phenotypes that are relevant to human development are not necessarily related to coordinated pathways of DNAm regulation; and I show that coordinated activity or regulation of DNAm is likely to play a role in the relationship between eosinophils and asthma.

7.2 Wider relevance of correlations between DNAm sites

7.2.1 EWAS and regional analysis

The chromatin contact work demonstrates a mechanism by which EWAS hits might be linked. The stable trans-chromosomal correlations between DNAm sites also demonstrate a mechanism by which EWAS hits could be linked. Correlation analysis provides the ability to identify how DNAm sites may be involved in a pathway, as opposed to single sites where the functional effect is uncertain. However this also highlights an issue within EWAS, namely that correlation between DNAm sites ought to be considered when taking the p-value threshold and sample size into account (Saffari et al., 2018); this is analogous to the need in GWAS to account for LD. Although this has long been suspected for DNAm sites correlating in *cis*, this thesis has demonstrated that there are noteworthy *trans* correlations that may also want to be taken into account in EWAS studies. This has not previously been accounted for in the EWAS literature because there has not been a consensus about the correlations to adjust for; the changeable nature of DNAm makes this much harder to account for then LD, which does not change. However this thesis has established that stable and persistent correlations between DNAm sites can be identified. Confirmation of whether these correlations also persist into adulthood could then provide indication of which *trans*-correlating DNAm sites may require adjustment in EWAS studies, and the data for researchers to have the option to do so.

7.2.2 Biological meaning underlying DNAm correlation networks

This thesis has also contributed a greater understanding of the biological meaning underlying correlations between DNAm sites. I have shown that trans-correlations sites, which make up the majority of DNAm correlations in such analyses, are biologically meaningful and likely to represent both trans-acting networks such as those involving transcription factors, and inter-chromosomal chromatin contacts. Future work with DNAm networks might benefit from assessing whether module members associate with particular transcription factor binding sites, or whether they contain overlapping chromatin sites, for increased interpretation of the consequences of the DNAm network modules.

7.2.3 Application of network modules

In this work I have shown that networks of DNAm identify modules of highly co-methylated sites in a normal population sample. In Chapter 4 I show that groups of highly related DNAm sites that may form part of a biological pathway are associated with asthma, a common disease. Even if this analysis does not highlight a mechanistic pathway, and DNAm is merely a consequence of disease pathophysiology, this group of co-methylated sites could represent a biomarker. Biomarkers have been identified from single studies, but a strength of using a network-based biomarker is that networks do not depend on a single node. This means that slight disruptions in the network (that could represent stochastic variation or confounding) would be less likely to alter the effectiveness of the biomarker; the use of a network as a biomarker would also increase confidence in its effectiveness as disruption of a whole pathway would represent a much more substantial change in biology. Finally, I have shown the network modules I have identified to be stable over time (including the asthma-related module), which would be an essential feature of a biomarker.

7.2.4 DNAm in inter-chromosomal chromatin contacts

The demonstration in a population sample of humans that DNAm is coordinated between sites that are involved in *trans*-chromosomal chromatin contacts has much wider relevance in the DNAm literature. This finding means that the function of trans-chromosomal contacts may either require or induce co-methylation. Delineating the purpose of this will lead to a greater understanding of genome function, and could lead to a greater understanding of how DNAm may play a role in health and disease. For example, it is possible that differential DNAm in disease somehow either impacts these trans-chromosomal contacts, or is changed by them, thus potentially altering functional regulation of the genome. Greater understanding of the role of DNAm in these contacts could thereby lead to new mechanistic insights into disease.

7.2.5 Relevance to effects of genotype of DNAm correlation

The analysis in Chapter 3 demonstrates that cis correlations between DNAm sites are likely to be quite strongly affected by genotype. To say this with certainty requires the resolution of this analysis in Chapter 5. Strongly correlating DNAm sites often have mQTLs, and these are the sites I have shown to have functional enrichments that suggest they are involved in genome regulation. The demonstration of this in chapter 3 should underscore the arguments others have made that it is important to at least consider, and maybe account for the effects of genetic variation on DNAm (Hannon et al., 2018; Lappalainen & Greally, 2017).

7.3 Strengths and limitations

This thesis has a number of strengths and limitations that bear consideration.

7.3.1 Tissue specificity

This study only considered DNA methylation in blood. DNAm is cell-type- and tissue-specific, and as such the applicability of these results to DNAm in other tissues is likely to be fairly limited (Hannon et al., 2015). It would be very interesting to see whether the correlation structure does replicate across tissues, particularly the degree to which *cis* and *trans* chromosomal correlations are tissue or cell type specific. The Gene Expression Omnibus (GEO) repository holds a large number of publicly available datasets that could be used to investigate this. Datasets available include 450k profiling of brain, skin, saliva, breast, heart, and artery.

The effects of disparate cell types is likely to impede the biological insights we are able to gain from this analysis, as different cell type have different methylation profiles, and so almost certainly have different correlation structures (although whether the big picture would still be the same is an interesting question). As such this work is likely to reflect an amalgamation of correlation structures that are present in the various blood cell types. The advantage of doing this work with blood is that it is the tissue that tends to be available for large cohorts (and so better sample size and phenotypic information), and it is the tissue that would be most likely to be used as a biomarker. As clinical biomarker analysis might be less likely to sort individual cell types for testing, blood-based analysis has relevance despite its drawbacks.

One avenue of enquiry to address the issues that this study encountered with two cohorts with different blood sample types might be to explore the correlations in ARIES at birth using homogeneous sample types (for example, all the participants with white cells). This would reduce the sample size by around 20%, but could provide an important insight as to whether the sample type issue has induced greater differences between the two cohorts. Alternatively, a dataset that would address some of the concerns about cell heterogeneity would be the Blueprint consortium's profiling of distinct blood cell types, to see to what degree cell type in blood does influence methylation.

7.3.2 Longitudinal analysis

The longitudinal nature of the ARIES dataset makes this the first study (to my knowledge) to demonstrate the consistency of DNAm correlations over time. This is a key finding because

DNAm is changeable, and the extent to which a correlation structure would be of use to other datasets with participants of different ages has been questionable. I have shown that it is broadly the same, and plan to release the correlation matrix for wider use so that other studies are able to take persistently highly correlated sites into account if they choose. The longitudinal analysis has also shown that DNAm correlation is persistent, and so may represent fundamental biological processes.

7.3.3 Trans-ethnicity replication

The trans-ethnicity replication is a strength for a number of reasons. Firstly, studies of DNAm have tended to be European-centric, which means that mechanistic insights, biomarkers and potential drug targets risk being more relevant for Europeans. This perpetuates health inequalities, and so it is important to identify whether DNAm represents the same functions in other ethnicities. The trans-ethnicity replication is also a strength because the correlation structure replicating in individuals from different genetic and environmental backgrounds points to a much more stable action of DNAm. This makes it more likely that correlations between DNAm sites are meaningful; it also makes the use of correlations between DNAm sites much more relevant for drug targets or biomarkers, because such features would need to be stable with regard to environmental influences and natural genetic variation. This finding should be interpreted with caution, in the sense that it should not deter further cross-ancestry research looking at correlation between DNAm sites; conversely, it should be seen as an encouragement to conduct more analyses using multiple ethnic groups to enable delineation of more stable biological signals (Tang, 2006).

7.3.4 Replication of the adjustment for cis genetic influences on cis correlation structure

The lack of replication of adjusting for *cis* genetic influence on *cis* correlation structure is a key weakness of this work. Although the effect seems strong in ARIES and is in line with what one might expect, without replication it is not clear whether this finding is accurate. As such, investigations into the cause of the lack of replication are ongoing, including the possibility of an error in my pre-processing of the BiB genetic data (which was highlighted by plotting the genetic principal components against the 1000 Genomes ethnic groups. Further work might also consider the effect sizes of the mQTLs in ALSPAC and BiB.

7.3.5 Assessing correlation analysis against a null distribution

An aspect of analysis that has not been addressed in Chapters 3 and 5 is the assessment of the correlation structure against a null distribution, to test whether the correlation distribution is significantly different to that expected by chance. This is one of the greatest weaknesses of this study, and one that would be good to address. The logistics of generating a permuted dataset with a reasonable number of permutations for a ~400,000 x ~400,000 matrix might be rather challenging; the Rdata files containing the matrices take up around 600Gb, and so even generating 100 permutations would require 60Tb of storage. Whilst there might be ways around this, it was not within the scope of this thesis to conduct this analysis.

7.4 Future directions

There are many directions further work could take from here. Specific recommendations are detailed in the Chapter discussions, and here I summarise what I consider to be the most promising avenues.

7.4.1 Preservation of DNAm correlation structure into adulthood

I have shown that DNAm correlations persist between birth and adolescence, however it is not clear whether this might change into adulthood. As both ARIES and BiB feature DNAm data for the mothers, the next logical step would be to assess DNAm correlation structure in the mothers, how well it preserves compared to their children, and whether it remains enriched for the same functional annotations (such as *cis* and *trans* enrichments for chromatin states and transcription factor binding sites; and inter-chromosomal chromatin contacts). ARIES provides additional opportunities here as the mothers have DNAm data generated during pregnancy and 12-18 years later, which would also allow for assessment of preservation of correlation structure over time within a group of adults.

7.4.2 Functional relevance of negative correlations

The assessment of negative correlations between DNAm sites is another important analysis that would immediately follow on from this work. I have shown that these correlations appear to have greater genetic influence, and that *cis* correlations are not distance-based in the way that positive correlations are. Negative correlations between DNAm sites could represent inhibitory pathways, may be likely to be enriched for different genomic features to the positive correlations. Negative correlations are often removed from WGCNA networks because of the problems inherent in interpreting multiple connected positive and negative paths (Langfelder, 2013); and so their function warrants further investigation as they may provide biological insights that cannot be gained from studying only positive correlations.

7.4.3 Further investigation of trans-chromosomal contacts

Co-methylation related to *trans*-chromosomal chromatin contacts is a recent development in the literature (G. Li et al., 2019). Li et al developed a method named Methyl-HiC, which isolates chromatin contact regions and then bisulfite sequences these regions to ascertain the methylation status of the CpG sites in these regions. They found that regions with the same features (eg CTCF binding sites, sites within the same TAD, and sites with concordant chromatin states) had more highly correlated methylation states, suggesting that DNAm states are related to chromatin contacts. I have demonstrated that co-methylation can be identified at sites known to have these contacts (Rao et al., 2014). The next step, after replication of these findings in another cohort, would be to investigate whether DNAm at these trans-chromosomal contact sites is perturbed in relation to diseases or exposures, as it is possible that differences in DNAm at these sites could represent a change in the chromatin contact. It is unclear if DNAm has functional involvement in chromatin contacts, or if it is simply a consequence, but that would make no difference to it being used as a biomarker in this way.

The use of chromatin contacts to identify disease biomarkers or pathophysiology is an idea that has been around for almost a decade (Crutchley, Wang, Ferraiuolo, & Dostie, 2010). It has been demonstrated that *cis*-chromatin contacts are linked to genetic risk variants associated with numerous neuropsychiatric illnesses (Song et al., 2019), the gene that is dysregulated in cystic fibrosis (Moisan et al., 2016), and they can potentially be used as biomarkers for early melanoma detection (Jakub et al., 2015). Trans-chromosomal chromatin contacts have been used to identify biomarkers for methotrexate response in rheumatoid arthritis (Carini et al., 2018), biomarkers for quick diagnosis of amyotrophic lateral sclerosis/motor neurone disease (ALS/MND) (Salter et al., 2018). This demonstrates a clear functional utility of identifying whether the chromatin contacts are in some way dysregulated. The studies that test for biomarkers often have small sample sizes, and so if chromatin contacts could be investigated using DNAm from large cohort studies, this could give rise to greatly increased sample sizes using data already in existence.

The methodology developed by GoDMC and implemented in this thesis for DNAm data to detect chromatin contacts could be used in a larger number of cohorts to identify chromatin contact sites where co-methylation is apparent and stable as measured by a DNAm array, to establish a baseline of contacts that can be detected using DNAm data. These could be validated against the data generated by (G. Li et al., 2019), which has been made publicly available. The differential methylation of these sites could then be tested between groups with either diseases or exposures of interest. It is probable that these contacts would not be picked up by co-methylation WGCNA networks because of the module size requirements (it seems unlikely that co-methylation at such sites would reach the conventional minimum module size of 30). Of course the ideal analysis here would instead be to confirm these contacts using methods such as that developed in (G. Li et al., 2019), that capture both chromatin contacts and DNAm; but the feasibility of this in large cohorts that have rich phenotypic data is likely to be low.

7.5 Summary

This thesis has shown that correlations between DNAm sites are stable, persistent, and are associated with relevant functional annotations. They can illustrate complex genomic functions such as inter-chromosomal chromatin contacts, and could be used to investigate networks of transcription factors. They can be used to create biologically meaningful networks, which likely illustrate stable biological functions of DNAm, and these networks can be used to identify biological pathways involved in disease in a population cohort. There are many future avenues of research resulting from the novel findings in this thesis, that could lead toward both a deeper understanding of genome biology, and increased application to clinical insights and tools.

Thank you for reading.

8 References

Abascal-Palacios, G., Ramsay, E. P., Beuron, F., Morris, E., & Vannini, A. (2018). Structural basis of RNA polymerase III transcription initiation. *Nature, 553*(7688), 301-+. doi:10.1038/nature25441

Adkins, R. M., Krushkal, J., Tylavsky, F. A., & Thomas, F. (2011). Racial differences in gene-specific DNA methylation levels are present at birth. *Birth Defects Res A Clin Mol Teratol, 91*(8), 728-736. doi:10.1002/bdra.20770

Albert, R. (2005). Scale-free networks in cell biology. *J Cell Sci, 118*(Pt 21), 4947-4957. doi:10.1242/jcs.02714

Ali-Khan, S. E., Krakowski, T., Tahir, R., & Daar, A. S. (2011). The use of race, ethnicity and ancestry in human genetic research. *Hugo J, 5*(1-4), 47-63. doi:10.1007/s11568-011-9154-5

Alisch, R. S., Barwick, B. G., Chopra, P., Myrick, L. K., Satten, G. A., Conneely, K. N., & Warren, S. T. (2012). Age-associated DNA methylation in pediatric populations. *Genome Res, 22*(4), 623-632. doi:10.1101/gr.125187.111

Allum, F., Hedman, A. K., Shao, X., Cheung, W. A., Vijay, J., Guenard, F., . . . Grundberg, E. (2019). Dissecting features of epigenetic variants underlying cardiometabolic risk using full-resolution epigenome profiling in regulatory elements. *Nat Commun, 10*(1), 1209. doi:10.1038/s41467-019-09184-z

Aluru, N. (2017). Epigenetic effects of environmental chemicals: insights from zebrafish. *Curr Opin Toxicol, 6*, 26-33. doi:10.1016/j.cotox.2017.07.004

Ambatipudi, S., Cuenin, C., Hernandez-Vargas, H., Ghantous, A., Le Calvez-Kelm, F., Kaaks, R., . . . Herceg, Z. (2016). Tobacco smoking-associated genome-wide DNA methylation changes in the EPIC study. *Epigenomics, 8*(5), 599-618. doi:10.2217/epi-2016-0001

Anand, S. S., Yusuf, S., Vuksan, V., Devanesen, S., Teo, K. K., Montague, P. A., . . . McQueen, M. (2000). Differences in risk factors, atherosclerosis, and cardiovascular disease between ethnic groups in Canada: the Study of Health Assessment and Risk in Ethnic groups (SHARE). *Lancet, 356*(9226), 279-284. doi:10.1016/s0140-6736(00)02502-2

Andrews, N. P., Husain, M., Dakak, N., & Quyyumi, A. A. (2001). Platelet inhibitory effect of nitric oxide in the human coronary circulation: impact of endothelial dysfunction. *J Am Coll Cardiol, 37*(2), 510-516. doi:10.1016/s0735-1097(00)01114-1

Arathimos, R., Suderman, M., Sharp, G. C., Burrows, K., Granell, R., Tilling, K., . . . Relton, C. L. (2017). Epigenome-wide association study of asthma and wheeze in childhood and adolescence. *Clin Epigenetics, 9*, 112. doi:10.1186/s13148-017-0414-7

Aulchenko, Y. S., Ripke, S., Isaacs, A., & van Duijn, C. M. (2007). GenABEL: an R library for genome-wide association analysis. *Bioinformatics, 23*(10), 1294-1296. doi:10.1093/bioinformatics/btm108

Baccarelli, A., Wright, R. O., Bollati, V., Tarantini, L., Litonjua, A. A., Suh, H. H., . . . Schwartz, J. (2009). Rapid DNA methylation changes after exposure to traffic particles. *Am J Respir Crit Care Med, 179*(7), 572-578. doi:10.1164/rccm.200807-1097OC

Bakulski, K. M., Feinberg, J. I., Andrews, S. V., Yang, J., Brown, S., S, L. M., . . . Fallin, M. D. (2016). DNA methylation of cord blood cell types: Applications for mixed cell birth studies. *Epigenetics, 11*(5), 354-362. doi:10.1080/15592294.2016.1161875

Barabasi, A. L. (2009). Scale-free networks: a decade and beyond. *Science, 325*(5939), 412-413. doi:10.1126/science.1173299

Barabasi, A. L., & Albert, R. (1999). Emergence of scaling in random networks. *Science, 286*(5439), 509-512. doi:10.1126/science.286.5439.509

Barabasi, A. L., & Oltvai, Z. N. (2004). Network biology: understanding the cell's functional organization. *Nat Rev Genet, 5*(2), 101-113. doi:10.1038/nrg1272

Barajas-Olmos, F., Centeno-Cruz, F., Zerrweck, C., Imaz-Rosshandler, I., Martinez-Hernandez, A., Cordova, E. J., . . . Orozco, L. (2018). Altered DNA methylation in liver and adipose tissues derived from individuals with obesity and type 2 diabetes. *BMC Med Genet, 19*(1), 28. doi:10.1186/s12881-018-0542-8

Bell, J. T., Pai, A. A., Pickrell, J. K., Gaffney, D. J., Pique-Regi, R., Degner, J. F., . . . Pritchard, J. K. (2011). DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines (vol 12, pg R10, 2011). *Genome Biology, 12*(6). doi:ARTN 405

10.1186/gb-2011-12-6-405

Bernhart, S. H., Kretzmer, H., Holdt, L. M., Juhling, F., Ammerpohl, O., Bergmann, A. K., . . . Hoffmann, S. (2016). Changes of bivalent chromatin coincide with increased expression of developmental genes in cancer. *Sci Rep, 6*, 37393. doi:10.1038/srep37393

Bestor, T. H., & Ingram, V. M. (1983). Two DNA methyltransferases from murine erythroleukemia cells: purification, sequence specificity, and mode of interaction with DNA. *Proc Natl Acad Sci U S A, 80*(18), 5559-5563. doi:10.1073/pnas.80.18.5559

Bibikova, M., Barnes, B., Tsan, C., Ho, V., Klotzle, B., Le, J. M., . . . Shen, R. (2011). High density DNA methylation array with single CpG site resolution. *Genomics, 98*(4), 288-295. doi:10.1016/j.ygeno.2011.07.007

Bird, A. (2007). Perceptions of epigenetics. *Nature, 447*(7143), 396-398. doi:10.1038/nature05913

Bird, A. P. (1978). Use of restriction enzymes to study eukaryotic DNA methylation: II. The symmetry of methylated sites supports semi-conservative copying of the methylation pattern. *J Mol Biol, 118*(1), 49-60. doi:10.1016/0022-2836(78)90243-7

Bland, J. M., & Altman, D. G. (1999). Measuring agreement in method comparison studies. *Stat Methods Med Res, 8*(2), 135-160. doi:10.1177/096228029900800204

Bocklandt, S., Lin, W., Sehl, M. E., Sanchez, F. J., Sinsheimer, J. S., Horvath, S., & Vilain, E. (2011). Epigenetic Predictor of Age. *PLoS One, 6*(6). doi:ARTN e14821

10.1371/journal.pone.0014821

Bonder, M. J., Luijk, R., Zhernakova, D. V., Moed, M., Deelen, P., Vermaat, M., . . . Heijmans, B. T. (2017). Disease variants alter transcription factor levels and methylation of their binding sites. *Nat Genet, 49*(1), 131-138. doi:10.1038/ng.3721

Boyd, A., Golding, J., Macleod, J., Lawlor, D. A., Fraser, A., Henderson, J., . . . Davey Smith, G. (2013). Cohort Profile: the 'children of the 90s'--the index offspring of the Avon Longitudinal Study of Parents and Children. *Int J Epidemiol, 42*(1), 111-127. doi:10.1093/ije/dys064

Breitling, L. P., Yang, R., Korn, B., Burwinkel, B., & Brenner, H. (2011). Tobacco-smoking-related differential DNA methylation: 27K discovery and replication. *Am J Hum Genet, 88*(4), 450-457. doi:10.1016/j.ajhg.2011.03.003

Bulik-Sullivan, B. K., Loh, P. R., Finucane, H. K., Ripke, S., Yang, J., Schizophrenia Working Group of the Psychiatric Genomics, C., . . . Neale, B. M. (2015). LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat Genet, 47*(3), 291-295. doi:10.1038/ng.3211

Busch, R., Qiu, W., Lasky-Su, J., Morrow, J., Criner, G., & DeMeo, D. (2016). Differential DNA methylation marks and gene comethylation of COPD in African-Americans with COPD exacerbations. *Respir Res, 17*(1), 143. doi:10.1186/s12931-016-0459-8

Butcher, L. M., & Beck, S. (2015). Probe Lasso: a novel method to rope in differentially methylated regions with 450K DNA methylation data. *Methods, 72*, 21-28. doi:10.1016/j.ymeth.2014.10.036

Butte, A. J., & Kohane, I. S. (2000). Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. *Pac Symp Biocomput*, 418-429. doi:10.1142/9789814447331_0040

Caramaschi, D., Sharp, G. C., Nohr, E. A., Berryman, K., Lewis, S. J., Davey Smith, G., & Relton, C. L. (2017). Exploring a causal role of DNA methylation in the relationship between maternal vitamin B12 during pregnancy and child's IQ at age 8, cognitive performance and educational attainment: a two-step Mendelian randomization study. *Hum Mol Genet, 26*(15), 3001-3013. doi:10.1093/hmg/ddx164

Cardenas, A., Faleschini, S., Cortes Hidalgo, A., Rifas-Shiman, S. L., Baccarelli, A. A., DeMeo, D. L., . . . Burris, H. H. (2019). Prenatal maternal antidepressants, anxiety, and depression and offspring DNA methylation: epigenome-wide associations at birth and persistence into early childhood. *Clin Epigenetics, 11*(1), 56. doi:10.1186/s13148-019-0653-x

Carini, C., Hunter, E., Scottish Early Rheumatoid Arthritis Inception cohort, I., Ramadass, A. S., Green, J., Akoulitchev, A., . . . Goodyear, C. S. (2018). Chromosome conformation signatures define predictive markers of inadequate response to methotrexate in early rheumatoid arthritis. *J Transl Med, 16*(1), 18. doi:10.1186/s12967-018-1387-9

Carlson, M. (2019). org.Hs.eg.db: Genome wide annotation for Human.

Carter, S. L., Brechbuhler, C. M., Griffin, M., & Bond, A. T. (2004). Gene co-expression network topology provides a framework for molecular characterization of cellular state. *Bioinformatics, 20*(14), 2242-2250. doi:10.1093/bioinformatics/bth234

Cedar, H., & Bergman, Y. (2009). Linking DNA methylation and histone modification: patterns and paradigms. *Nat Rev Genet, 10*(5), 295-304. doi:10.1038/nrg2540

Cedar, H., Solage, A., Glaser, G., & Razin, A. (1979). Direct detection of methylated cytosine in DNA by use of the restriction enzyme MspI. *Nucleic Acids Res, 6*(6), 2125-2132. doi:10.1093/nar/6.6.2125

Centers for Disease, C., & Prevention. (2004). Asthma prevalence and control characteristics by race/ethnicity--United States, 2002. *MMWR Morb Mortal Wkly Rep, 53*(7), 145-148.

Chambers, J. C., Loh, M., Lehne, B., Drong, A., Kriebel, J., Motta, V., . . . Kooner, J. S. (2015). Epigenome-wide association of DNA methylation markers in peripheral blood from Indian Asians and Europeans with incident type 2 diabetes: a nested case-control study. *Lancet Diabetes Endocrinol, 3*(7), 526-534. doi:10.1016/S2213-8587(15)00127-8

Choate, L. A., & Danko, C. G. (2016). Poised for development. *Nat Genet, 48*(8), 822-823. doi:10.1038/ng.3628

Christensen, B. C., Houseman, E. A., Marsit, C. J., Zheng, S., Wrensch, M. R., Wiemels, J. L., . . . Kelsey, K. T. (2009). Aging and environmental exposures alter tissue-specific DNA methylation dependent upon CpG island context. *PLoS Genet, 5*(8), e1000602. doi:10.1371/journal.pgen.1000602

Christiansen, L., Lenart, A., Tan, Q., Vaupel, J. W., Aviv, A., McGue, M., & Christensen, K. (2016). DNA methylation age is associated with mortality in a longitudinal Danish twin study. *Aging Cell, 15*(1), 149-154. doi:10.1111/acel.12421

Chu, S. K., & Yang, H. C. (2017). Interethnic DNA methylation difference and its implications in pharmacoepigenetics. *Epigenomics, 9*(11), 1437-1454. doi:10.2217/epi-2017-0046

Chundru, V. K., Marioni, R. E., Pendergast, J. G. D., Lin, T., Beveridge, A. J., Martin, N. G., . . . McRae, A. F. (2020). Rare Genetic Variants Underlie Outlying levels of DNA Methylation and Gene-Expression. *bioRxiv*.

Chung, K. F. (2015). Targeting the interleukin pathway in the treatment of asthma. *Lancet, 386*(9998), 1086-1096. doi:10.1016/S0140-6736(15)00157-9

Ciccone, D. N., Su, H., Hevi, S., Gay, F., Lei, H., Bajko, J., . . . Chen, T. (2009). KDM1B is a histone H3K4 demethylase required to establish maternal genomic imprints. *Nature, 461*(7262), 415-418. doi:10.1038/nature08315

Coit, P., Ognenovski, M., Gensterblum, E., Maksimowicz-McKinnon, K., Wren, J. D., & Sawalha, A. H. (2015). Ethnicity-specific epigenetic variation in naive CD4+ T cells and the susceptibility to autoimmunity. *Epigenetics Chromatin, 8*, 49. doi:10.1186/s13072-015-0037-1

Cooper, R. S. (2001). Social inequality, ethnicity and cardiovascular disease. *Int J Epidemiol, 30 Suppl 1*, S48-52. doi:10.1093/ije/30.suppl_1.s48

Crutchley, J. L., Wang, X. Q., Ferraiuolo, M. A., & Dostie, J. (2010). Chromatin conformation signatures: ideal human disease biomarkers? *Biomark Med, 4*(4), 611-629. doi:10.2217/bmm.10.68

de Jong, S., Boks, M. P., Fuller, T. F., Strengman, E., Janson, E., de Kovel, C. G., . . . Ophoff, R. A. (2012). A gene co-expression network in whole blood of schizophrenia patients is independent of antipsychotic-use and enriched for brain-expressed genes. *PLoS One, 7*(6), e39498. doi:10.1371/journal.pone.0039498

de Vocht, F., Suderman, M., Ruano-Ravina, A., Thomas, R., Wakeford, R., Relton, C., . . . Boyd, A. (2019). Residential exposure to radon and DNA methylation across the lifecourse: an exploratory study in the ALSPAC birth cohort. *Wellcome Open Res, 4*, 3. doi:10.12688/wellcomeopenres.14991.2

Deaton, A. M., & Bird, A. (2011). CpG islands and the regulation of transcription. *Genes Dev, 25*(10), 1010-1022. doi:10.1101/gad.2037511

Dedeurwaerder, S., Defrance, M., Bizet, M., Calonne, E., Bontempi, G., & Fuks, F. (2014). A comprehensive overview of Infinium HumanMethylation450 data processing. *Brief Bioinform, 15*(6), 929-941. doi:10.1093/bib/bbt054

Delaneau, O., Marchini, J., Genomes Project, C., & Genomes Project, C. (2014). Integrating sequence and array data to create an improved 1000 Genomes Project haplotype reference panel. *Nat Commun, 5*, 3934. doi:10.1038/ncomms4934

Demerath, E. W., Guan, W., Grove, M. L., Aslibekyan, S., Mendelson, M., Zhou, Y. H., . . . Boerwinkle, E. (2015). Epigenome-wide association study (EWAS) of BMI, BMI change and waist circumference in African American adults identifies multiple replicated loci. *Hum Mol Genet, 24*(15), 4464-4479. doi:10.1093/hmg/ddv161

Department for Communities and Local Government. (2015). The English Indices of Deprivation 2015: Statistical Release. Retrieved from https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/465791/English_Indices_of_Deprivation_2015_-_Statistical_Release.pdf

Dick, K. J., Nelson, C. P., Tsaprouni, L., Sandling, J. K., Aissi, D., Wahl, S., . . . Samani, N. J. (2014). DNA methylation and body-mass index: a genome-wide analysis. *Lancet, 383*(9933), 1990-1998. doi:10.1016/S0140-6736(13)62674-4

DiFranza, J. R., Aligne, C. A., & Weitzman, M. (2004). Prenatal and postnatal environmental tobacco smoke exposure and children's health. *Pediatrics, 113*(4 Suppl), 1007-1015.

Domcke, S., Bardet, A. F., Adrian Ginno, P., Hartl, D., Burger, L., & Schubeler, D. (2015). Competition between DNA methylation and transcription factors determines binding of NRF1. *Nature, 528*(7583), 575-579. doi:10.1038/nature16462

Dunn, E. C., Soare, T. W., Zhu, Y., Simpkin, A. J., Suderman, M. J., Klengel, T., . . . Relton, C. L. (2019). Sensitive Periods for the Effect of Childhood Adversity on DNA Methylation: Results From a Prospective, Longitudinal Study. *Biol Psychiatry, 85*(10), 838-849. doi:10.1016/j.biopsych.2018.12.023

Eckhardt, F., Lewin, J., Cortese, R., Rakyan, V. K., Attwood, J., Burger, M., . . . Beck, S. (2006). DNA methylation profiling of human chromosomes 6, 20 and 22. *Nat Genet, 38*(12), 1378-1385. doi:10.1038/ng1909

Elliott, H. R., Tillin, T., McArdle, W. L., Ho, K., Duggirala, A., Frayling, T. M., . . . Relton, C. L. (2014). Differences in smoking associated DNA methylation patterns in South Asians and Europeans. *Clin Epigenetics, 6*(1), 4. doi:10.1186/1868-7083-6-4

Elliott, H. R., Walia, G. K., Duggirala, A., Groom, A., Reddy, S. U., Chandak, G. R., . . . Relton, C. L. (2013). Migration and DNA methylation: a comparison of methylation patterns in type 2 diabetes susceptibility genes between indians and europeans. *J Diabetes Res Clin Metab, 2*, 6. doi:10.7243/2050-0866-2-6

Epp, A., Sullivan, K. C., Herr, A. B., & Strait, R. T. (2016). Immunoglobulin Glycosylation Effects in Allergy and Immunity. *Curr Allergy Asthma Rep, 16*(11), 79. doi:10.1007/s11882-016-0658-x

Ernst, J., & Kellis, M. (2010). Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nat Biotechnol, 28*(8), 817-825. doi:10.1038/nbt.1662

Ernst, J., & Kellis, M. (2015). Large-scale imputation of epigenomic datasets for systematic annotation of diverse human tissues. *Nat Biotechnol, 33*(4), 364-376. doi:10.1038/nbt.3157

Ernst, J., Kheradpour, P., Mikkelsen, T. S., Shoresh, N., Ward, L. D., Epstein, C. B., . . . Bernstein, B. E. (2011). Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature, 473*(7345), 43-49. doi:10.1038/nature09906

Feinberg, A. P., Koldobskiy, M. A., & Gondor, A. (2016). Epigenetic modulators, modifiers and mediators in cancer aetiology and progression. *Nat Rev Genet, 17*(5), 284-299. doi:10.1038/nrg.2016.13

Felix, J. F., Joubert, B. R., Baccarelli, A. A., Sharp, G. C., Almqvist, C., Annesi-Maesano, I., . . . London, S. J. (2018). Cohort Profile: Pregnancy And Childhood Epigenetics (PACE) Consortium. *Int J Epidemiol, 47*(1), 22-23u. doi:10.1093/ije/dyx190

Felsenfeld, G. (2014). A brief history of epigenetics. *Cold Spring Harb Perspect Biol, 6*(1). doi:10.1101/cshperspect.a018200

Feng, S., Jacobsen, S. E., & Reik, W. (2010). Epigenetic reprogramming in plant and animal development. *Science, 330*(6004), 622-627. doi:10.1126/science.1190614

Fortin, J. P., Labbe, A., Lemire, M., Zanke, B. W., Hudson, T. J., Fertig, E. J., . . . Hansen, K. D. (2014). Functional normalization of 450k methylation array data improves replication in large cancer studies. *Genome Biol, 15*(12), 503. doi:10.1186/s13059-014-0503-2

Fraga, M. F., Ballestar, E., Paz, M. F., Ropero, S., Setien, F., Ballestar, M. L., . . . Esteller, M. (2005). Epigenetic differences arise during the lifetime of monozygotic twins. *Proc Natl Acad Sci U S A, 102*(30), 10604-10609. doi:10.1073/pnas.0500398102

Fraser, H. B., Lam, L. L., Neumann, S. M., & Kobor, M. S. (2012). Population-specificity of human DNA methylation. *Genome Biol, 13*(2), R8. doi:10.1186/gb-2012-13-2-r8

Fuks, F. (2005). DNA methylation and histone modifications: teaming up to silence genes. *Curr Opin Genet Dev, 15*(5), 490-495. doi:10.1016/j.gde.2005.08.002

Galanter, J. M., Gignoux, C. R., Oh, S. S., Torgerson, D., Pino-Yanes, M., Thakuri, N., . . . Zaitlen, N. (2017). Differential methylation between ethnic sub-groups reflects the effect of genetic ancestry and environmental exposures. *Elife, 6*. doi:ARTN e20532

10.7554/eLife.20532

Garg, P., Joshi, R. S., Watson, C., & Sharp, A. J. (2018). A survey of inter-individual variation in DNA methylation identifies environmentally responsive co-regulated networks of epigenetic variation in the human genome. *PLoS Genet, 14*(10), e1007707. doi:10.1371/journal.pgen.1007707

Gaunt, T. R., Shihab, H. A., Hemani, G., Min, J. L., Woodward, G., Lyttleton, O., . . . Relton, C. L. (2016). Systematic identification of genetic influences on methylation across the human life course. *Genome Biol, 17*, 61. doi:10.1186/s13059-016-0926-z

Genomes Project, C., Auton, A., Brooks, L. D., Durbin, R. M., Garrison, E. P., Kang, H. M., . . . Abecasis, G. R. (2015). A global reference for human genetic variation. *Nature, 526*(7571), 68-74. doi:10.1038/nature15393

Gervin, K., Page, C. M., Aass, H. C., Jansen, M. A., Fjeldstad, H. E., Andreassen, B. K., . . . Lyle, R. (2016). Cell type specific DNA methylation in cord blood: A 450K-reference data set and cell count-based validation of estimated cell type composition. *Epigenetics, 11*(9), 690-698. doi:10.1080/15592294.2016.1214782

Giavarina, D. (2015). Understanding Bland Altman analysis. *Biochem Med (Zagreb), 25*(2), 141-151. doi:10.11613/BM.2015.015

Gomez, L., Odom, G. J., Young, J. I., Martin, E. R., Liu, L., Chen, X., . . . Wang, L. (2019). coMethDMR: accurate identification of co-methylated and differentially methylated regions in epigenome-wide association studies with continuous phenotypes. *Nucleic Acids Res, 47*(17), e98. doi:10.1093/nar/gkz590

Granell, R., Henderson, A. J., & Sterne, J. A. (2016). Associations of wheezing phenotypes with late asthma outcomes in the Avon Longitudinal Study of Parents and Children: A population-based birth cohort. *J Allergy Clin Immunol, 138*(4), 1060-1070 e1011. doi:10.1016/j.jaci.2016.01.046

Groom, A., Elliott, H. R., Embleton, N. D., & Relton, C. L. (2011). Epigenetics and child health: basic principles. *Archives of Disease in Childhood, 96*(9), 863-869. doi:10.1136/adc.2009.165712

Gruzieva, O., Xu, C. J., Yousefi, P., Relton, C., Merid, S. K., Breton, C. V., . . . Melen, E. (2019). Prenatal Particulate Air Pollution and DNA Methylation in Newborns: An Epigenome-Wide Meta-Analysis. *Environ Health Perspect, 127*(5), 57012. doi:10.1289/EHP4522

Gu, Z., Gu, L., Eils, R., Schlesner, M., & Brors, B. (2014). circlize Implements and enhances circular visualization in R. *Bioinformatics, 30*(19), 2811-2812. doi:10.1093/bioinformatics/btu393

Gutierrez-Arcelus, M., Lappalainen, T., Montgomery, S. B., Buil, A., Ongen, H., Yurovsky, A., . . . Dermitzakis, E. T. (2013). Passive and active DNA methylation and the interplay with genetic variation in gene regulation. *Elife, 2*, e00523. doi:10.7554/eLife.00523

Hamilton, O. K. L., Zhang, Q., McRae, A. F., Walker, R. M., Morris, S. W., Redmond, P., . . . Marioni, R. E. (2019). An epigenetic score for BMI based on DNA methylation correlates with poor physical health and major disease in the Lothian Birth Cohort. *Int J Obes (Lond), 43*(9), 1795-1802. doi:10.1038/s41366-018-0262-3

Hannon, E., Knox, O., Sugden, K., Burrage, J., Wong, C. C. Y., Belsky, D. W., . . . Mill, J. (2018). Characterizing genetic and environmental influences on variable DNA methylation using monozygotic and dizygotic twins. *PLoS Genet, 14*(8), e1007544. doi:10.1371/journal.pgen.1007544

Hannon, E., Lunnon, K., Schalkwyk, L., & Mill, J. (2015). Interindividual methylomic variation across blood, cortex, and cerebellum: implications for epigenetic studies of neurological and neuropsychiatric phenotypes. *Epigenetics, 10*(11), 1024-1032. doi:10.1080/15592294.2015.1100786

Hannon, E., Schendel, D., Ladd-Acosta, C., Grove, J., Hansen, C. S., Hougaard, D. M., . . . Mill, J. (2019). Variable DNA methylation in neonates mediates the association between prenatal smoking and birth weight. *Philos Trans R Soc Lond B Biol Sci, 374*(1770), 20180120. doi:10.1098/rstb.2018.0120

Hannum, G., Guinney, J., Zhao, L., Zhang, L., Hughes, G., Sadda, S., . . . Zhang, K. (2013). Genome-wide methylation profiles reveal quantitative views of human aging rates. *Mol Cell, 49*(2), 359-367. doi:10.1016/j.molcel.2012.10.016

Hansen, K. D. (2016). IlluminaHumanMethylation450kanno.ilmn12.hg19: Annotation for Illumina's 450k methylation arrays (Version 0.6.0).

Harrell, F. E. J., & others., w. c. f. C. D. a. m. (2019). Hmisc: Harrell Miscellaneous. Retrieved from https://CRAN.R-project.org/package=Hmisc

He, Y. F., Li, B. Z., Li, Z., Liu, P., Wang, Y., Tang, Q., . . . Xu, G. L. (2011). Tet-mediated formation of 5-carboxylcytosine and its excision by TDG in mammalian DNA. *Science, 333*(6047), 1303-1307. doi:10.1126/science.1210944

Heintzman, N. D., Stuart, R. K., Hon, G., Fu, Y., Ching, C. W., Hawkins, R. D., . . . Ren, B. (2007). Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat Genet, 39*(3), 311-318. doi:10.1038/ng1966

Heiss, J. A., & Just, A. C. (2019). Improved filtering of DNA methylation microarray data by detection p values and its impact on downstream analyses. *Clin Epigenetics, 11*(1), 15. doi:10.1186/s13148-019-0615-3

Hendrich, B., & Bird, A. (1998). Identification and characterization of a family of mammalian methyl-CpG binding proteins. *Mol Cell Biol, 18*(11), 6538-6547. doi:10.1128/mcb.18.11.6538

Hibler, E., Huang, L., Andrade, J., & Spring, B. (2019). Impact of a diet and activity health promotion intervention on regional patterns of DNA methylation. *Clin Epigenetics, 11*(1), 133. doi:10.1186/s13148-019-0707-0

Holtman, I. R., Raj, D. D., Miller, J. A., Schaafsma, W., Yin, Z., Brouwer, N., . . . Eggen, B. J. (2015). Induction of a common microglia gene expression signature by aging and neurodegenerative conditions: a co-expression meta-analysis. *Acta Neuropathol Commun, 3*, 31. doi:10.1186/s40478-015-0203-5

Hon, G., Wang, W., & Ren, B. (2009). Discovery and annotation of functional chromatin signatures in the human genome. *Plos Computational Biology, 5*(11), e1000566. doi:10.1371/journal.pcbi.1000566

Horvath, S. (2013). DNA methylation age of human tissues and cell types. *Genome Biology, 14*(10). doi:ARTN R115

10.1186/gb-2013-14-10-r115

Horvath, S., Langfelder, P., Kwak, S., Aaronson, J., Rosinski, J., Vogt, T. F., . . . Yang, X. W. (2016). Huntington's disease accelerates epigenetic aging of human brain and disrupts DNA methylation levels. *Aging (Albany NY), 8*(7), 1485-1512. doi:10.18632/aging.101005

Horvath, S., Zhang, Y., Langfelder, P., Kahn, R. S., Boks, M. P., van Eijk, K., . . . Ophoff, R. A. (2012). Aging effects on DNA methylation modules in human brain and blood tissue. *Genome Biol, 13*(10), R97. doi:10.1186/gb-2012-13-10-r97

Houseman, E. A., Accomando, W. P., Koestler, D. C., Christensen, B. C., Marsit, C. J., Nelson, H. H., . . . Kelsey, K. T. (2012). DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics, 13*, 86. doi:10.1186/1471-2105-13-86

Houtepen, L. C., Hardy, R., Maddock, J., Kuh, D., Anderson, E. L., Relton, C. L., . . . Howe, L. D. (2018). Childhood adversity and DNA methylation in two population-based cohorts. *Transl Psychiatry, 8*(1), 266. doi:10.1038/s41398-018-0307-3

Houtepen, L. C., Vinkers, C. H., Carrillo-Roa, T., Hiemstra, M., van Lier, P. A., Meeus, W., . . . Boks, M. P. (2016). Genome-wide DNA methylation levels and altered cortisol stress reactivity following childhood trauma in humans. *Nat Commun, 7*, 10967. doi:10.1038/ncomms10967

Howie, B., Marchini, J., & Stephens, M. (2011). Genotype imputation with thousands of genomes. *G3 (Bethesda), 1*(6), 457-470. doi:10.1534/g3.111.001198

Howie, B. N., Donnelly, P., & Marchini, J. (2009). A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet, 5*(6), e1000529. doi:10.1371/journal.pgen.1000529

Hughes, A., Smart, M., Gorrie-Stone, T., Hannon, E., Mill, J., Bao, Y., . . . Kumari, M. (2018). Socioeconomic Position and DNA Methylation Age Acceleration Across the Life Course. *Am J Epidemiol, 187*(11), 2346-2354. doi:10.1093/aje/kwy155

Husquin, L. T., Rotival, M., Fagny, M., Quach, H., Zidane, N., McEwen, L. M., . . . Quintana-Murci, L. (2018). Exploring the genetic basis of human population differences in DNA methylation and their causal impact on immune gene regulation. *Genome Biol, 19*(1), 222. doi:10.1186/s13059-018-1601-3

Ideker, T., Galitski, T., & Hood, L. (2001). A new approach to decoding life: systems biology. *Annu Rev Genomics Hum Genet, 2*, 343-372. doi:10.1146/annurev.genom.2.1.343

International HapMap, C. (2005). A haplotype map of the human genome. *Nature, 437*(7063), 1299-1320. doi:10.1038/nature04226

International HapMap, C., Altshuler, D. M., Gibbs, R. A., Peltonen, L., Altshuler, D. M., Gibbs, R. A., . . . McEwen, J. E. (2010). Integrating common and rare genetic variation in diverse human populations. *Nature, 467*(7311), 52-58. doi:10.1038/nature09298

Ito, S., Shen, L., Dai, Q., Wu, S. C., Collins, L. B., Swenberg, J. A., . . . Zhang, Y. (2011). Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine. *Science, 333*(6047), 1300-1303. doi:10.1126/science.1210597

Jaffe, A. E., Gao, Y., Deep-Soboslay, A., Tao, R., Hyde, T. M., Weinberger, D. R., & Kleinman, J. E. (2016). Mapping DNA methylation across development, genotype and schizophrenia in the human frontal cortex. *Nat Neurosci, 19*(1), 40-47. doi:10.1038/nn.4181

Jaffe, A. E., Murakami, P., Lee, H., Leek, J. T., Fallin, M. D., Feinberg, A. P., & Irizarry, R. A. (2012). Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies. *Int J Epidemiol, 41*(1), 200-209. doi:10.1093/ije/dyr238

Jakub, J. W., Grotz, T. E., Jordan, P., Hunter, E., Pittelkow, M., Ramadass, A., . . . Markovic, S. (2015). A pilot study of chromosomal aberrations and epigenetic changes in peripheral blood samples to identify patients with melanoma. *Melanoma Res, 25*(5), 406-411. doi:10.1097/CMR.0000000000000182

Javierre, B. M., Burren, O. S., Wilder, S. P., Kreuzhuber, R., Hill, S. M., Sewitz, S., . . . Fraser, P. (2016). Lineage-Specific Genome Architecture Links Enhancers and Non-coding Disease Variants to Target Gene Promoters. *Cell, 167*(5), 1369-1384 e1319. doi:10.1016/j.cell.2016.09.037

Jeziorska, D. M., Murray, R. J. S., De Gobbi, M., Gaentzsch, R., Garrick, D., Ayyub, H., . . . Tufarelli, C. (2017). DNA methylation of intragenic CpG islands depends on their transcriptional activity during differentiation and disease. *Proc Natl Acad Sci U S A, 114*(36), E7526-E7535. doi:10.1073/pnas.1703087114

Joubert, B. R., den Dekker, H. T., Felix, J. F., Bohlin, J., Ligthart, S., Beckett, E., . . . London, S. J. (2016). Maternal plasma folate impacts differential DNA methylation in an epigenome-wide meta-analysis of newborns. *Nat Commun, 7*, 10577. doi:10.1038/ncomms10577

Joubert, B. R., Felix, J. F., Yousefi, P., Bakulski, K. M., Just, A. C., Breton, C., . . . London, S. J. (2016). DNA Methylation in Newborns and Maternal Smoking in Pregnancy: Genome-wide Consortium Meta-analysis. *Am J Hum Genet, 98*(4), 680-696. doi:10.1016/j.ajhg.2016.02.019

Kanehisa, M., & Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res, 28*(1), 27-30. doi:10.1093/nar/28.1.27

Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., & Tanabe, M. (2016). KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Research, 44*(D1), D457-D462. doi:10.1093/nar/gkv1070

Kato, N., Loh, M., Takeuchi, F., Verweij, N., Wang, X., Zhang, W. H., . . . Consortium, I. (2015). Trans-ancestry genome-wide association study identifies 12 genetic loci influencing blood pressure and implicates a role for DNA methylation. *Nat Genet, 47*(11), 1282-+. doi:10.1038/ng.3405

Keet, C. A., McCormack, M. C., Pollack, C. E., Peng, R. D., McGowan, E., & Matsui, E. C. (2015). Neighborhood poverty, urban residence, race/ethnicity, and asthma: Rethinking the inner-city asthma epidemic. *J Allergy Clin Immunol, 135*(3), 655-662. doi:10.1016/j.jaci.2014.11.022

Kim, J. J., Khalid, O., Namazi, A., Tu, T. G., Elie, O., Lee, C., & Kim, Y. (2014). Discovery of consensus gene signature and intermodular connectivity defining self-renewal of human embryonic stem cells. *Stem Cells, 32*(6), 1468-1479. doi:10.1002/stem.1675

Kim, S., Yu, N. K., & Kaang, B. K. (2015). CTCF as a multifunctional protein in genome regulation and gene expression. *Experimental and Molecular Medicine, 47*. doi:ARTN e166

10.1038/emm.2015.33

Kitano, H. (2002). Systems biology: A brief overview. *Science, 295*(5560), 1662-1664. doi:DOI 10.1126/science.1069492

Klasic, M., Kristic, J., Korac, P., Horvat, T., Markulin, D., Vojta, A., . . . Zoldos, V. (2016). DNA hypomethylation upregulates expression of the MGAT3 gene in HepG2 cells and leads to changes in N-glycosylation of secreted glycoproteins. *Sci Rep, 6*, 24363. doi:10.1038/srep24363

Kohli, R. M., & Zhang, Y. (2013). TET enzymes, TDG and the dynamics of DNA demethylation. *Nature, 502*(7472), 472-479. doi:10.1038/nature12750

Kuan, P. F., & Chiang, D. Y. (2012). Integrating prior knowledge in multiple testing under dependence with applications to detecting differential DNA methylation. *Biometrics, 68*(3), 774-783. doi:10.1111/j.1541-0420.2011.01730.x

Kupers, L. K., Monnereau, C., Sharp, G. C., Yousefi, P., Salas, L. A., Ghantous, A., . . . Felix, J. F. (2019). Meta-analysis of epigenome-wide association studies in neonates reveals widespread differential DNA methylation associated with birthweight. *Nat Commun, 10*(1), 1893. doi:10.1038/s41467-019-09671-3

Kwon, N. H., Kim, J. S., Lee, J. Y., Oh, M. J., & Choi, D. C. (2008). DNA methylation and the expression of IL-4 and IFN-gamma promoter genes in patients with bronchial asthma. *J Clin Immunol, 28*(2), 139-146. doi:10.1007/s10875-007-9148-1

Langevin, S. M., Houseman, E. A., Christensen, B. C., Wiencke, J. K., Nelson, H. H., Karagas, M. R., . . . Kelsey, K. T. (2011). The influence of aging, environmental exposures and local sequence features on the variation of DNA methylation in blood. *Epigenetics, 6*(7), 908-919. doi:10.4161/epi.6.7.16431

Langfelder, P. (2013). *Signed vs. Unsigned Topological Overlap Matrix*

Technical report. Retrieved from

Langfelder, P., Cantle, J. P., Chatzopoulou, D., Wang, N., Gao, F., Al-Ramahi, I., . . . Yang, X. W. (2016). Integrated genomics and proteomics define huntingtin CAG length-dependent networks in mice. *Nat Neurosci, 19*(4), 623-633. doi:10.1038/nn.4256

Langfelder, P., & Horvath, S. (2007). Eigengene networks for studying the relationships between co-expression modules. *BMC Syst Biol, 1*, 54. doi:10.1186/1752-0509-1-54

Langfelder, P., & Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics, 9*, 559. doi:10.1186/1471-2105-9-559

Langfelder, P., & Horvath, S. (2016). Tutorials for the WGCNA package. Retrieved from https://horvath.genetics.ucla.edu/html/CoexpressionNetwork/Rpackages/WGCNA/Tutorials/

Langfelder, P., & Horvath, S. (2017). WGCNA package FAQ. Retrieved from https://horvath.genetics.ucla.edu/html/CoexpressionNetwork/Rpackages/WGCNA/faq.html

Langfelder, P., Luo, R., Oldham, M. C., & Horvath, S. (2011). Is My Network Module Preserved and Reproducible? *Plos Computational Biology, 7*(1). doi:ARTN e1001057

10.1371/journal.pcbi.1001057

Lappalainen, T., & Greally, J. M. (2017). Associating cellular epigenetic models with human phenotypes. *Nat Rev Genet, 18*(7), 441-451. doi:10.1038/nrg.2017.32

Lawrence, M., Huber, W., Pages, H., Aboyoun, P., Carlson, M., Gentleman, R., . . . Carey, V. J. (2013). Software for computing and annotating genomic ranges. *Plos Computational Biology, 9*(8), e1003118. doi:10.1371/journal.pcbi.1003118

Lee, D. U., Agarwal, S., & Rao, A. (2002). Th2 lineage commitment and efficient IL-4 production involves extended demethylation of the IL-4 gene. *Immunity, 16*(5), 649-660. doi:10.1016/s1074-7613(02)00314-x

Leek J, J. W., Jaffe A, Parker H, Storey J. (2011). The SVA package for removing batch effects and other unwanted variation in high-throughput experiments. Retrieved from https://bioconductor.org/packages/release/bioc/vignettes/sva/inst/doc/sva.pdf

Leek, J. T., Scharpf, R. B., Bravo, H. C., Simcha, D., Langmead, B., Johnson, W. E., . . . Irizarry, R. A. (2010). Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat Rev Genet, 11*(10), 733-739. doi:10.1038/nrg2825

Leek, J. T., & Storey, J. D. (2007). Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet, 3*(9), 1724-1735. doi:10.1371/journal.pgen.0030161

Lesch, B. J., Silber, S. J., McCarrey, J. R., & Page, D. C. (2016). Parallel evolution of male germline epigenetic poising and somatic development in animals. *Nat Genet, 48*(8), 888-894. doi:10.1038/ng.3591

Levine, M. E., Lu, A. T., Quach, A., Chen, B. H., Assimes, T. L., Bandinelli, S., . . . Horvath, S. (2018). An epigenetic biomarker of aging for lifespan and healthspan. *Aging (Albany NY), 10*(4), 573-591. doi:10.18632/aging.101414

Li, B., Carey, M., & Workman, J. L. (2007). The role of chromatin during transcription. *Cell, 128*(4), 707-719. doi:10.1016/j.cell.2007.01.015

Li, G., Liu, Y., Zhang, Y., Kubo, N., Yu, M., Fang, R., . . . Ren, B. (2019). Joint profiling of DNA methylation and chromatin architecture in single cells. *Nat Methods, 16*(10), 991-993. doi:10.1038/s41592-019-0502-z

Liao, J., Karnik, R., Gu, H., Ziller, M. J., Clement, K., Tsankov, A. M., . . . Meissner, A. (2015). Targeted disruption of DNMT1, DNMT3A and DNMT3B in human embryonic stem cells. *Nat Genet, 47*(5), 469-478. doi:10.1038/ng.3258

Lienert, F., Wirbelauer, C., Som, I., Dean, A., Mohn, F., & Schubeler, D. (2011). Identification of genetic elements that autonomously determine DNA methylation states. *Nat Genet, 43*(11), 1091-1097. doi:10.1038/ng.946

Lio, C. J., & Rao, A. (2019). TET Enzymes and 5hmC in Adaptive and Innate Immune Systems. *Front Immunol, 10,* 210. doi:10.3389/fimmu.2019.00210

Lister, R., Pelizzola, M., Dowen, R. H., Hawkins, R. D., Hon, G., Tonti-Filippini, J., . . . Ecker, J. R. (2009). Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature, 462*(7271), 315-322. doi:10.1038/nature08514

Liu, J., Carnero-Montoro, E., van Dongen, J., Lent, S., Nedeljkovic, I., Ligthart, S., . . . van Duijn, C. M. (2019). An integrative cross-omics analysis of DNA methylation sites of glucose and insulin homeostasis. *Nature Communications, 10*. doi:ARTN 2581

10.1038/s41467-019-10487-4

Liu, Y., Li, X., Aryee, M. J., Ekstrom, T. J., Padyukov, L., Klareskog, L., . . . Feinberg, A. P. (2014). GeMes, clusters of DNA methylation under genetic control, can inform genetic and epigenetic analysis of disease. *Am J Hum Genet, 94*(4), 485-495. doi:10.1016/j.ajhg.2014.02.011

Lyko, F. (2018). The DNA methyltransferase family: a versatile toolkit for epigenetic regulation. *Nat Rev Genet, 19*(2), 81-92. doi:10.1038/nrg.2017.80

Maeder, M. L., Angstman, J. F., Richardson, M. E., Linder, S. J., Cascio, V. M., Tsai, S. Q., . . . Joung, J. K. (2013). Targeted DNA demethylation and activation of endogenous genes using programmable TALE-TET1 fusion proteins. *Nat Biotechnol, 31*(12), 1137-1142. doi:10.1038/nbt.2726

Maertens, A., Tran, V., Kleensang, A., & Hartung, T. (2018). Weighted Gene Correlation Network Analysis (WGCNA) Reveals Novel Transcription Factors Associated With Bisphenol A Dose-Response. *Front Genet, 9*, 508. doi:10.3389/fgene.2018.00508

Maiti, A., & Drohat, A. C. (2011). Thymine DNA glycosylase can rapidly excise 5-formylcytosine and 5-carboxylcytosine: potential implications for active demethylation of CpG sites. *J Biol Chem, 286*(41), 35334-35338. doi:10.1074/jbc.C111.284620

Mandaviya, P. R., Joehanes, R., Brody, J., Castillo-Fernandez, J. E., Dekkers, K. F., Do, A. N., . . . Heil, S. G. (2019). Association of dietary folate and vitamin B-12 intake with genome-wide DNA methylation in blood: a large-scale epigenome-wide association analysis in 5841 individuals. *Am J Clin Nutr, 110*(2), 437-450. doi:10.1093/ajcn/nqz031

Mansell, G., Gorrie-Stone, T. J., Bao, Y. C., Kumari, M., Schalkwyk, L. S., Mill, J., & Hannon, E. (2019). Guidance for DNA methylation studies: statistical insights from the Illumina EPIC array. *BMC Genomics, 20*. doi:ARTN 366

10.1186/s12864-019-5761-7

Marees, A. T., de Kluiver, H., Stringer, S., Vorspan, F., Curis, E., Marie-Claire, C., & Derks, E. M. (2018). A tutorial on conducting genome-wide association studies: Quality control and statistical analysis. *Int J Methods Psychiatr Res, 27*(2), e1608. doi:10.1002/mpr.1608

Marioni, R. E., Shah, S., McRae, A. F., Chen, B. H., Colicino, E., Harris, S. E., . . . Deary, I. J. (2015). DNA methylation age of blood predicts all-cause mortality in later life. *Genome Biol, 16*, 25. doi:10.1186/s13059-015-0584-6

Markunas, C. A., Wilcox, A. J., Xu, Z., Joubert, B. R., Harlid, S., Panduri, V., . . . Taylor, J. A. (2016). Maternal Age at Delivery Is Associated with an Epigenetic Signature in Both Newborns and Adults. *PLoS One, 11*(7), e0156361. doi:10.1371/journal.pone.0156361

Martin, E. M., & Fry, R. C. (2016). A cross-study analysis of prenatal exposures to environmental contaminants and the epigenome: support for stress-responsive transcription factor occupancy as a mediator of gene-specific CpG methylation patterning. *Environ Epigenet, 2*(1). doi:10.1093/eep/dvv011

Martino, D., Loke, Y. J., Gordon, L., Ollikainen, M., Cruickshank, M. N., Saffery, R., & Craig, J. M. (2013). Longitudinal, genome-scale analysis of DNA methylation in twins from birth to 18 months of age reveals rapid epigenetic change in early life and pair-specific effects of discordance. *Genome Biol, 14*(5), R42. doi:10.1186/gb-2013-14-5-r42

Martino, D. J., Tulic, M. K., Gordon, L., Hodder, M., Richman, T. R., Metcalfe, J., . . . Saffery, R. (2011). Evidence for age-related and individual-specific changes in DNA methylation profile of mononuclear cells during early immune development in humans. *Epigenetics, 6*(9), 1085-1094. doi:10.4161/epi.6.9.16401

Matsumura, Y., Nakaki, R., Inagaki, T., Yoshida, A., Kano, Y., Kimura, H., . . . Sakai, J. (2015). H3K4/H3K9me3 Bivalent Chromatin Domains Targeted by Lineage-Specific DNA Methylation Pauses Adipocyte Differentiation. *Mol Cell, 60*(4), 584-596. doi:10.1016/j.molcel.2015.10.025

McCaw, Z. (2019). Rank Normal Transformation Omnibus Test. Retrieved from https://cran.rstudio.com/web/packages/RNOmni/index.html

McGuinness, D., McGlynn, L. M., Johnson, P. C., MacIntyre, A., Batty, G. D., Burns, H., . . . Shiels, P. G. (2012). Socio-economic status is associated with epigenetic differences in the pSoBid cohort. *Int J Epidemiol, 41*(1), 151-160. doi:10.1093/ije/dyr215

McRae, A. F., Powell, J. E., Henders, A. K., Bowdler, L., Hemani, G., Shah, S., . . . Montgomery, G. W. (2014). Contribution of genetic variation to transgenerational inheritance of DNA methylation. *Genome Biol, 15*(5), R73. doi:10.1186/gb-2014-15-5-r73

McWilliams, J. M., Meara, E., Zaslavsky, A. M., & Ayanian, J. Z. (2009). Differences in control of cardiovascular disease and diabetes by race, ethnicity, and education: U.S. trends from 1999 to 2006 and effects of medicare coverage. *Ann Intern Med, 150*(8), 505-515. doi:10.7326/0003-4819-150-8-200904210-00005

Mendelson, M. M., Marioni, R. E., Joehanes, R., Liu, C., Hedman, A. K., Aslibekyan, S., . . . Deary, I. J. (2017). Association of Body Mass Index with DNA Methylation and Gene Expression in Blood Cells and Relations to Cardiometabolic Disease: A Mendelian Randomization Approach. *PLoS Med, 14*(1), e1002215. doi:10.1371/journal.pmed.1002215

Messerschmidt, D. M., Knowles, B. B., & Solter, D. (2014). DNA methylation dynamics during epigenetic reprogramming in the germline and preimplantation embryos. *Genes & Development, 28*(8), 812-828. doi:10.1101/gad.234294.113

Michels, K. B., Binder, A. M., Dedeurwaerder, S., Epstein, C. B., Greally, J. M., Gut, I., . . . Irizarry, R. A. (2013). Recommendations for the design and analysis of epigenome-wide association studies. *Nature Methods, 10*(10), 949-955. doi:10.1038/nmeth.2632

Mikkelsen, T. S., Ku, M., Jaffe, D. B., Issac, B., Lieberman, E., Giannoukos, G., . . . Bernstein, B. E. (2007). Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature, 448*(7153), 553-560. doi:10.1038/nature06008

Min, J. L., Hemani, G., Davey Smith, G., Relton, C., & Suderman, M. (2018). Meffil: efficient normalization and analysis of very large DNA methylation datasets. *Bioinformatics, 34*(23), 3983-3989. doi:10.1093/bioinformatics/bty476

Moen, E. L., Zhang, X., Mu, W., Delaney, S. M., Wing, C., McQuade, J., . . . Zhang, W. (2013). Genome-wide variation of cytosine modifications between European and African populations and the implications for complex traits. *Genetics, 194*(4), 987-996. doi:10.1534/genetics.113.151381

Moisan, S., Berlivet, S., Ka, C., Le Gac, G., Dostie, J., & Ferec, C. (2016). Analysis of long-range interactions in primary human cells identifies cooperative CFTR regulatory elements. *Nucleic Acids Res, 44*(6), 2564-2576. doi:10.1093/nar/gkv1300

Mullner, D. (2013). fastcluster: Fast Hierarchical, Agglomerative Clustering Routines for R and Python. *Journal of Statistical Software, 53*(9), 1-18.

Naeem, H., Wong, N. C., Chatterton, Z., Hong, M. K., Pedersen, J. S., Corcoran, N. M., . . . Macintyre, G. (2014). Reducing the risk of false discovery enabling identification of biologically significant genome-wide methylation status using the HumanMethylation450 array. *BMC Genomics, 15*, 51. doi:10.1186/1471-2164-15-51

Nan, X., Ng, H. H., Johnson, C. A., Laherty, C. D., Turner, B. M., Eisenman, R. N., & Bird, A. (1998). Transcriptional repression by the methyl-CpG-binding protein MeCP2 involves a histone deacetylase complex. *Nature, 393*(6683), 386-389. doi:10.1038/30764

Nguyen, A. B., Moser, R., & Chou, W. Y. (2014). Race and health profiles in the United States: an examination of the social gradient through the 2009 CHIS adult survey. *Public Health, 128*(12), 1076-1086. doi:10.1016/j.puhe.2014.10.003

Nicodemus-Johnson, J., Naughton, K. A., Sudi, J., Hogarth, K., Naurekas, E. T., Nicolae, D. L., . . . Ober, C. (2016). Genome-Wide Methylation Study Identifies an IL-13-induced Epigenetic Signature in Asthmatic Airways. *Am J Respir Crit Care Med, 193*(4), 376-385. doi:10.1164/rccm.201506-1243OC

Northstone, K., Lewcock, M., Groom, A., Boyd, A., Macleod, J., Timpson, N., & Wells, N. (2019). The Avon Longitudinal Study of Parents and Children (ALSPAC): an update on the enrolled sample of index children in 2019. *Wellcome Open Res, 4*, 51. doi:10.12688/wellcomeopenres.15132.1

Oberdoerffer, S. (2012). A conserved role for intragenic DNA methylation in alternative pre-mRNA splicing. *Transcription, 3*(3), 106-109. doi:10.4161/trns.19816

Okano, M., Bell, D. W., Haber, D. A., & Li, E. (1999). DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. *Cell, 99*(3), 247-257. doi:10.1016/s0092-8674(00)81656-6

Okuda, S., Yamada, T., Hamajima, M., Itoh, M., Katayama, T., Bork, P., . . . Kanehisa, M. (2008). KEGG Atlas mapping for global analysis of metabolic pathways. *Nucleic Acids Res, 36*(Web Server issue), W423-426. doi:10.1093/nar/gkn282

Oldham, M. C., Konopka, G., Iwamoto, K., Langfelder, P., Kato, T., Horvath, S., & Geschwind, D. H. (2008). Functional organization of the transcriptome in human brain. *Nat Neurosci, 11*(11), 1271-1282. doi:10.1038/nn.2207

Ong, M. L., & Holbrook, J. D. (2014). Novel region discovery method for Infinium 450K DNA methylation data reveals changes associated with aging in muscle and neuronal pathways. *Aging Cell, 13*(1), 142-155. doi:10.1111/acel.12159

Ooi, S. K., Qiu, C., Bernstein, E., Li, K., Jia, D., Yang, Z., . . . Bestor, T. H. (2007). DNMT3L connects unmethylated lysine 4 of histone H3 to de novo methylation of DNA. *Nature, 448*(7154), 714-717. doi:10.1038/nature05987

Otani, J., Kimura, H., Sharif, J., Endo, T. A., Mishima, Y., Kawakami, T., . . . Tajima, S. (2013). Cell cycle-dependent turnover of 5-hydroxymethyl cytosine in mouse embryonic stem cells. *PLoS One, 8*(12), e82961. doi:10.1371/journal.pone.0082961

Panico, L., Stuart, B., Bartley, M., & Kelly, Y. (2014). Asthma trajectories in early childhood: identifying modifiable factors. *PLoS One, 9*(11), e111922. doi:10.1371/journal.pone.0111922

Park, S. L., Patel, Y. M., Loo, L. W. M., Mullen, D. J., Offringa, I. A., Maunakea, A., . . . Le Marchand, L. (2018). Association of internal smoking dose with blood DNA methylation in three racial/ethnic populations. *Clinical Epigenetics, 10*. doi:ARTN 110

10.1186/s13148-018-0543-7

Pedersen, B. S., Schwartz, D. A., Yang, I. V., & Kechris, K. J. (2012). Comb-p: software for combining, analyzing, grouping and correcting spatially correlated P-values. *Bioinformatics, 28*(22), 2986-2988. doi:10.1093/bioinformatics/bts545

Perfilyev, A., Dahlman, I., Gillberg, L., Rosqvist, F., Iggman, D., Volkov, P., . . . Ling, C. (2017). Impact of polyunsaturated and saturated fat overfeeding on the DNA-methylation pattern in human adipose tissue: a randomized controlled trial. *Am J Clin Nutr, 105*(4), 991-1000. doi:10.3945/ajcn.116.143164

Perrier, F., Viallon, V., Ambatipudi, S., Ghantous, A., Cuenin, C., Hernandez-Vargas, H., . . . Romieu, I. (2019). Association of leukocyte DNA methylation changes with dietary folate and alcohol intake in the EPIC study. *Clin Epigenetics, 11*(1), 57. doi:10.1186/s13148-019-0637-x

Peters, T. J., Buckley, M. J., Statham, A. L., Pidsley, R., Samaras, K., R, V. L., . . . Molloy, P. L. (2015). De novo identification of differentially methylated regions in the human genome. *Epigenetics Chromatin, 8*, 6. doi:10.1186/1756-8935-8-6

Phipson, B., Maksimovic, J., & Oshlack, A. (2016). missMethyl: an R package for analyzing data from Illumina's HumanMethylation450 platform. *Bioinformatics, 32*(2), 286-288. doi:10.1093/bioinformatics/btv560

Pidsley, R., Zotenko, E., Peters, T. J., Lawrence, M. G., Risbridger, G. P., Molloy, P., . . . Clark, S. J. (2016). Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling. *Genome Biol, 17*(1), 208. doi:10.1186/s13059-016-1066-1

Prendergast, G. C., & Ziff, E. B. (1991). Methylation-sensitive sequence-specific DNA binding by the c-Myc basic region. *Science, 251*(4990), 186-189. doi:10.1126/science.1987636

Price, A. J., Collado-Torres, L., Ivanov, N. A., Xia, W., Burke, E. E., Shin, J. H., . . . Jaffe, A. E. (2019). Divergent neuronal DNA methylation patterns across human cortical development reveal critical periods and a unique role of CpH methylation. *Genome Biol, 20*(1), 196. doi:10.1186/s13059-019-1805-1

Prince, C., Hammerton, G., Taylor, A. E., Anderson, E. L., Timpson, N. J., Davey Smith, G., . . . Richmond, R. C. (2019). Investigating the impact of cigarette smoking behaviours on DNA methylation patterns in adolescence. *Hum Mol Genet, 28*(1), 155-165. doi:10.1093/hmg/ddy316

Probst, A. V., Dunleavy, E., & Almouzni, G. (2009). Epigenetic inheritance during the cell cycle. *Nat Rev Mol Cell Biol, 10*(3), 192-206. doi:10.1038/nrm2640

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., . . . Sham, P. C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet, 81*(3), 559-575. doi:10.1086/519795

Rahmani, E., Shenhav, L., Schweiger, R., Yousefi, P., Huen, K., Eskenazi, B., . . . Halperin, E. (2017). Genome-wide methylation data mirror ancestry information. *Epigenetics Chromatin, 10*, 1. doi:10.1186/s13072-016-0108-y

Ramos-Lopez, O., Riezu-Boj, J. I., Milagro, F. I., Martinez, J. A., & Project, M. (2018). DNA methylation signatures at endoplasmic reticulum stress genes are associated with adiposity and insulin resistance. *Mol Genet Metab, 123*(1), 50-58. doi:10.1016/j.ymgme.2017.11.011

Rao, S. S., Huntley, M. H., Durand, N. C., Stamenova, E. K., Bochkov, I. D., Robinson, J. T., . . . Aiden, E. L. (2014). A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell, 159*(7), 1665-1680. doi:10.1016/j.cell.2014.11.021

Razin, A., & Riggs, A. D. (1980). DNA methylation and gene function. *Science, 210*(4470), 604-610. doi:10.1126/science.6254144

Reese, S. E., Xu, C. J., den Dekker, H. T., Lee, M. K., Sikdar, S., Ruiz-Arenas, C., . . . London, S. J. (2019). Epigenome-wide meta-analysis of DNA methylation and childhood asthma. *J Allergy Clin Immunol, 143*(6), 2062-2074. doi:10.1016/j.jaci.2018.11.043

Reinius, L. E., Acevedo, N., Joerink, M., Pershagen, G., Dahlen, S. E., Greco, D., . . . Kere, J. (2012). Differential DNA methylation in purified human blood cells: implications for cell lineage and studies on disease susceptibility. *PLoS One, 7*(7), e41361. doi:10.1371/journal.pone.0041361

Relton, C. L., Gaunt, T., McArdle, W., Ho, K., Duggirala, A., Shihab, H., . . . Davey Smith, G. (2015). Data Resource Profile: Accessible Resource for Integrated Epigenomic Studies (ARIES). *Int J Epidemiol, 44*(4), 1181-1190. doi:10.1093/ije/dyv072

Richmond, R. C., Sharp, G. C., Ward, M. E., Fraser, A., Lyttleton, O., McArdle, W. L., . . . Relton, C. L. (2016). DNA Methylation and BMI: Investigating Identified Methylation Sites at HIF3A in a Causal Framework. *Diabetes, 65*(5), 1231-1244. doi:10.2337/db15-0996

Richmond, R. C., Simpkin, A. J., Woodward, G., Gaunt, T. R., Lyttleton, O., McArdle, W. L., . . . Relton, C. L. (2015). Prenatal exposure to maternal smoking and offspring DNA methylation across the lifecourse: findings from the Avon Longitudinal Study of Parents and Children (ALSPAC). *Hum Mol Genet, 24*(8), 2201-2217. doi:10.1093/hmg/ddu739

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., & Smyth, G. K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res, 43*(7), e47. doi:10.1093/nar/gkv007

Roadmap Epigenomics, C., Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., . . . Kellis, M. (2015). Integrative analysis of 111 reference human epigenomes. *Nature, 518*(7539), 317-330. doi:10.1038/nature14248

Roberts, S., Suderman, M., Zammit, S., Watkins, S. H., Hannon, E., Mill, J., . . . Fisher, H. L. (2019). Longitudinal investigation of DNA methylation changes preceding adolescent psychotic experiences. *Transl Psychiatry, 9*(1), 69. doi:10.1038/s41398-019-0407-8

Saffari, A., Silver, M. J., Zavattari, P., Moi, L., Columbano, A., Meaburn, E. L., & Dudbridge, F. (2018). Estimation of a significance threshold for epigenome-wide association studies. *Genet Epidemiol, 42*(1), 20-33. doi:10.1002/gepi.22086

Sainsbury, S., Bernecky, C., & Cramer, P. (2015). Structural basis of transcription initiation by RNA polymerase II. *Nat Rev Mol Cell Biol, 16*(3), 129-143. doi:10.1038/nrm3952

Salter, M., Corfield, E., Ramadass, A., Grand, F., Green, J., Westra, J., . . . Akoulitchev, A. (2018). Initial Identification of a Blood-Based Chromosome Conformation Signature for Aiding in the Diagnosis of Amyotrophic Lateral Sclerosis. *EBioMedicine, 33*, 169-184. doi:10.1016/j.ebiom.2018.06.015

Sanati, N., Iancu, O. D., Wu, G., Jacobs, J. E., & McWeeney, S. K. (2018). Network-Based Predictors of Progression in Head and Neck Squamous Cell Carcinoma. *Front Genet, 9*, 183. doi:10.3389/fgene.2018.00183

Sandoval, J., Heyn, H., Moran, S., Serra-Musach, J., Pujana, M. A., Bibikova, M., & Esteller, M. (2011). Validation of a DNA methylation microarray for 450,000 CpG sites in the human genome. *Epigenetics, 6*(6), 692-702. doi:10.4161/epi.6.6.16196

Saunderson, E. A., Stepper, P., Gomm, J. J., Hoa, L., Morgan, A., Allen, M. D., . . . Ficz, G. (2017). Hit-and-run epigenetic editing prevents senescence entry in primary breast cells from healthy donors. *Nat Commun, 8*(1), 1450. doi:10.1038/s41467-017-01078-2

Saxonov, S., Berg, P., & Brutlag, D. L. (2006). A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proc Natl Acad Sci U S A, 103*(5), 1412-1417. doi:10.1073/pnas.0510310103

Sbihi, H., Koehoorn, M., Tamburic, L., & Brauer, M. (2017). Asthma Trajectories in a Population-based Birth Cohort. Impacts of Air Pollution and Greenness. *Am J Respir Crit Care Med, 195*(5), 607-613. doi:10.1164/rccm.201601-0164OC

Schroeder, J. W., Conneely, K. N., Cubells, J. C., Kilaru, V., Newport, D. J., Knight, B. T., . . . Smith, A. K. (2011). Neonatal DNA methylation patterns associate with gestational age. *Epigenetics, 6*(12), 1498-1504. doi:10.4161/epi.6.12.18296

Schultz, M. D., He, Y. P., Whitaker, J. W., Hariharan, M., Mukamel, E. A., Leung, D., . . . Ecker, J. R. (2015). Human body epigenome maps reveal noncanonical DNA methylation variation. *Nature, 523*(7559), 212-U189. doi:10.1038/nature14465

Shah, S., Bonder, M. J., Marioni, R. E., Zhu, Z., McRae, A. F., Zhernakova, A., . . . Visscher, P. M. (2015). Improving Phenotypic Prediction by Combining Genetic and Epigenetic Associations. *Am J Hum Genet, 97*(1), 75-85. doi:10.1016/j.ajhg.2015.05.014

Shah, S., Mcrae, A. F., Marioni, R. E., Harris, S. E., Gibson, J., Henders, A. K., . . . Visscher, P. M. (2014). Genetic and environmental exposures constrain epigenetic drift over the human life course. *Genome Research, 24*(11), 1725-1733. doi:10.1101/gr.176933.114

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., . . . Ideker, T. (2003). Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Research, 13*(11), 2498-2504. doi:10.1101/gr.1239303

Sharp, G. C., Lawlor, D. A., Richmond, R. C., Fraser, A., Simpkin, A., Suderman, M., . . . Relton, C. L. (2015). Maternal pre-pregnancy BMI and gestational weight gain, offspring DNA methylation and later offspring adiposity: findings from the Avon Longitudinal Study of Parents and Children. *Int J Epidemiol, 44*(4), 1288-1304. doi:10.1093/ije/dyv042

Sheffield, N. C., & Bock, C. (2016). LOLA: enrichment analysis for genomic region sets and regulatory elements in R and Bioconductor. *Bioinformatics, 32*(4), 587-589. doi:10.1093/bioinformatics/btv612

Shenker, N. S., Ueland, P. M., Polidoro, S., van Veldhoven, K., Ricceri, F., Brown, R., . . . Vineis, P. (2013). DNA methylation as a long-term biomarker of exposure to tobacco smoke. *Epidemiology, 24*(5), 712-716. doi:10.1097/EDE.0b013e31829d5cb3

Shoemaker, R., Deng, J., Wang, W., & Zhang, K. (2010). Allele-specific methylation is prevalent and is contributed by CpG-SNPs in the human genome. *Genome Res, 20*(7), 883-889. doi:10.1101/gr.104695.109

Shukla, S., Kavak, E., Gregory, M., Imashimizu, M., Shutinoski, B., Kashlev, M., . . . Oberdoerffer, S. (2011). CTCF-promoted RNA polymerase II pausing links DNA methylation to splicing. *Nature, 479*(7371), 74-79. doi:10.1038/nature10442

Siegfried, Z., Eden, S., Mendelsohn, M., Feng, X., Tsuberi, B. Z., & Cedar, H. (1999). DNA methylation represses transcription in vivo. *Nat Genet, 22*(2), 203-206. doi:10.1038/9727

Simons, R. L., Lei, M. K., Beach, S. R., Philibert, R. A., Cutrona, C. E., Gibbons, F. X., & Barr, A. (2016). Economic hardship and biological weathering: The epigenetics of aging in a U.S. sample of black women. *Soc Sci Med, 150*, 192-200. doi:10.1016/j.socscimed.2015.12.001

Simpkin, A. J., Suderman, M., Gaunt, T. R., Lyttleton, O., McArdle, W. L., Ring, S. M., . . . Relton, C. L. (2015). Longitudinal analysis of DNA methylation associated with birth weight and gestational age. *Hum Mol Genet, 24*(13), 3752-3763. doi:10.1093/hmg/ddv119

Sinkkonen, L., Hugenschmidt, T., Berninger, P., Gaidatzis, D., Mohn, F., Artus-Revel, C. G., . . . Filipowicz, W. (2008). MicroRNAs control de novo DNA methylation through regulation of transcriptional repressors in mouse embryonic stem cells. *Nat Struct Mol Biol, 15*(3), 259-267. doi:10.1038/nsmb.1391

Sinsheimer, R. L. (1955). The action of pancreatic deoxyribonuclease. II. Isomeric dinucleotides. *J Biol Chem, 215*(2), 579-583.

Slatkin, M. (2008). Linkage disequilibrium--understanding the evolutionary past and mapping the medical future. *Nat Rev Genet, 9*(6), 477-485. doi:10.1038/nrg2361

Slieker, R. C., Roost, M. S., van Iperen, L., Suchiman, H. E. D., Tobi, E. W., Carlotti, F., . . . Lopes, S. M. C. D. (2015). DNA Methylation Landscapes of Human Fetal Development. *Plos Genetics, 11*(10). doi:ARTN e1005583

10.1371/journal.pgen.1005583

Smallwood, S. A., & Kelsey, G. (2012). De novo DNA methylation: a germ cell perspective. *Trends Genet, 28*(1), 33-42. doi:10.1016/j.tig.2011.09.004

Smith, J. A., Zhao, W., Wang, X., Ratliff, S. M., Mukherjee, B., Kardia, S. L. R., . . . Needham, B. L. (2017). Neighborhood characteristics influence DNA methylation of genes involved in stress response and inflammation: The Multi-Ethnic Study of Atherosclerosis. *Epigenetics, 12*(8), 662-673. doi:10.1080/15592294.2017.1341026

Smith, T., Noble, M., Noble, S., Wright, G., McLennan, D., & Plunkett, E. (2015). The English indices of deprivation 2015. *London: Department for Communities and Local Government*.

Smith, Z. D., & Meissner, A. (2013). DNA methylation: roles in mammalian development. *Nat Rev Genet, 14*(3), 204-220. doi:10.1038/nrg3354

Somineni, H. K., Zhang, X., Biagini Myers, J. M., Kovacic, M. B., Ulm, A., Jurcak, N., . . . Ji, H. (2016). Ten-eleven translocation 1 (TET1) methylation is associated with childhood asthma and traffic-related air pollution. *J Allergy Clin Immunol, 137*(3), 797-805 e795. doi:10.1016/j.jaci.2015.10.021

Song, M., Yang, X., Ren, X., Maliskova, L., Li, B., Jones, I. R., . . . Shen, Y. (2019). Mapping cis-regulatory chromatin contacts in neural cells links neuropsychiatric disorder risk variants to target genes. *Nat Genet, 51*(8), 1252-1262. doi:10.1038/s41588-019-0472-1

Spiers, H., Hannon, E., Schalkwyk, L. C., Smith, R., Wong, C. C., O'Donovan, M. C., . . . Mill, J. (2015). Methylomic trajectories across human fetal brain development. *Genome Res, 25*(3), 338-352. doi:10.1101/gr.180273.114

Stadler, M. B., Murr, R., Burger, L., Ivanek, R., Lienert, F., Scholer, A., . . . Schubeler, D. (2011). DNA-binding factors shape the mouse methylome at distal regulatory regions. *Nature, 480*(7378), 490-495. doi:10.1038/nature10716

Stein, R., Gruenbaum, Y., Pollack, Y., Razin, A., & Cedar, H. (1982). Clonal inheritance of the pattern of DNA methylation in mouse cells. *Proc Natl Acad Sci U S A, 79*(1), 61-65. doi:10.1073/pnas.79.1.61

Sterne, J. A., & Davey Smith, G. (2001). Sifting the evidence-what's wrong with significance tests? *BMJ, 322*(7280), 226-231. doi:10.1136/bmj.322.7280.226

Stuart, J. M., Segal, E., Koller, D., & Kim, S. K. (2003). A gene-coexpression network for global discovery of conserved genetic modules. *Science, 302*(5643), 249-255. doi:10.1126/science.1087447

Suderman, M., Staley, J. R., French, R., Arathimos, R., Simpkin, A., & Tilling, K. (2018). dmrff: identifying differentially methylated regions efficiently with power and control. *bioRxiv*.

Sun, Q., Zhao, H., Zhang, C., Hu, T., Wu, J., Lin, X., . . . Zhu, T. (2017). Gene co-expression network reveals shared modules predictive of stage and grade in serous ovarian cancers. *Oncotarget, 8*(26), 42983-42996. doi:10.18632/oncotarget.17785

Sutendra, G., Kinnaird, A., Dromparis, P., Paulin, R., Stenson, T. H., Haromy, A., . . . Michelakis, E. D. (2014). A nuclear pyruvate dehydrogenase complex is important for the generation of acetyl-CoA and histone acetylation. *Cell, 158*(1), 84-97. doi:10.1016/j.cell.2014.04.046

Swarup, V., Hinz, F. I., Rexach, J. E., Noguchi, K., Toyoshiba, H., Oda, A., . . . Gen, I. F. D. (2019). Identification of evolutionarily conserved gene networks mediating neurodegenerative dementia. *Nature Medicine, 25*(1), 152-+. doi:10.1038/s41591-018-0223-3

Tahiliani, M., Koh, K. P., Shen, Y., Pastor, W. A., Bandukwala, H., Brudno, Y., . . . Rao, A. (2009). Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. *Science, 324*(5929), 930-935. doi:10.1126/science.1170116

Tang, H. (2006). Confronting ethnicity-specific disease risk. *Nat Genet, 38*(1), 13-15. doi:10.1038/ng0106-13

Teschendorff, A. E., & Relton, C. L. (2018). Statistical and integrative system-level analysis of DNA methylation data. *Nat Rev Genet, 19*(3), 129-147. doi:10.1038/nrg.2017.86

Teschendorff, A. E., West, J., & Beck, S. (2013). Age-associated epigenetic drift: implications, and a case of epigenetic thrift? *Hum Mol Genet, 22*(R1), R7-R15. doi:10.1093/hmg/ddt375

Teschendorff, A. E., Zhuang, J., & Widschwendter, M. (2011). Independent surrogate variable analysis to deconvolve confounding factors in large-scale microarray profiling studies. *Bioinformatics, 27*(11), 1496-1505. doi:10.1093/bioinformatics/btr171

Tian, C., Gregersen, P. K., & Seldin, M. F. (2008). Accounting for ancestry: population substructure and genome-wide association studies. *Hum Mol Genet, 17*(R2), R143-150. doi:10.1093/hmg/ddn268

Tian, X. J., Zhang, H., Sannerud, J., & Xing, J. H. (2016). Achieving diverse and monoallelic olfactory receptor selection through dual-objective optimization design. *Proceedings of the National Academy of Sciences of the United States of America, 113*(21), E2889-E2898. doi:10.1073/pnas.1601722113

Tiedemann, R. L., Putiri, E. L., Lee, J. H., Hlady, R. A., Kashiwagi, K., Ordog, T., . . . Robertson, K. D. (2014). Acute depletion redefines the division of labor among DNA methyltransferases in methylating the human genome. *Cell Rep, 9*(4), 1554-1566. doi:10.1016/j.celrep.2014.10.013

Tillin, T., Hughes, A. D., Godsland, I. F., Whincup, P., Forouhi, N. G., Welsh, P., . . . Chaturvedi, N. (2013). Insulin resistance and truncal obesity as important determinants of the greater incidence of diabetes in Indian Asians and African Caribbeans compared with Europeans: the Southall And Brent REvisited (SABRE) cohort. *Diabetes Care, 36*(2), 383-393. doi:10.2337/dc12-0544

Tillin, T., Hughes, A. D., Mayet, J., Whincup, P., Sattar, N., Forouhi, N. G., . . . Chaturvedi, N. (2013). The relationship between metabolic risk factors and incident cardiovascular disease in Europeans, South Asians, and African Caribbeans: SABRE (Southall and Brent Revisited) -- a prospective population-based study. *J Am Coll Cardiol, 61*(17), 1777-1786. doi:10.1016/j.jacc.2012.12.046

Trinchera, M., Zulueta, A., Caretti, A., & Dall'Olio, F. (2014). Control of Glycosylation-Related Genes by DNA Methylation: the Intriguing Case of the B3GALT5 Gene and Its Distinct Promoters. *Biology (Basel), 3*(3), 484-497. doi:10.3390/biology3030484

Turowski, T. W., & Tollervey, D. (2016). Transcription by RNA polymerase III: insights into mechanism and regulation. *Biochem Soc Trans, 44*(5), 1367-1375. doi:10.1042/BST20160062

Tykocinski, L. O., Hajkova, P., Chang, H. D., Stamm, T., Sozeri, O., Lohning, M., . . . Radbruch, A. (2005). A critical control element for interleukin-4 memory expression in T helper lymphocytes. *J Biol Chem, 280*(31), 28177-28185. doi:10.1074/jbc.M502038200

van Dongen, J., Nivard, M. G., Willemsen, G., Hottenga, J. J., Helmer, Q., Dolan, C. V., . . . Boomsma, D. I. (2016). Genetic and environmental influences interact with age and sex in shaping the human methylome. *Nat Commun, 7*, 11115. doi:10.1038/ncomms11115

van Eijk, K. R., de Jong, S., Boks, M. P., Langeveld, T., Colas, F., Veldink, J. H., . . . Ophoff, R. A. (2012). Genetic analysis of DNA methylation and gene expression levels in whole blood of healthy human subjects. *BMC Genomics, 13*, 636. doi:10.1186/1471-2164-13-636

Vinkhuyzen, A. A., Wray, N. R., Yang, J., Goddard, M. E., & Visscher, P. M. (2013). Estimation and partition of heritability in human populations using whole-genome analysis methods. *Annu Rev Genet, 47*, 75-95. doi:10.1146/annurev-genet-111212-133258

Wahl, A., Kasela, S., Carnero-Montoro, E., van Iterson, M., Stambuk, J., Sharma, S., . . . Gieger, C. (2018). IgG glycosylation and DNA methylation are interconnected with smoking. *Biochim Biophys Acta Gen Subj, 1862*(3), 637-648. doi:10.1016/j.bbagen.2017.10.012

Wahl, S., Drong, A., Lehne, B., Loh, M., Scott, W. R., Kunze, S., . . . Chambers, J. C. (2017). Epigenome-wide association study of body mass index, and the adverse outcomes of adiposity. *Nature, 541*(7635), 81-86. doi:10.1038/nature20784

Wang, J., Zhuang, J., Iyer, S., Lin, X., Whitfield, T. W., Greven, M. C., . . . Weng, Z. (2012). Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. *Genome Res, 22*(9), 1798-1812. doi:10.1101/gr.139105.112

Wang, Q., Huang, J., Sun, H., Liu, J., Wang, J., Wang, Q., . . . Zhang, Y. (2014). CR Cistrome: a ChIP-Seq database for chromatin regulators and histone modification linkages in human and mouse. *Nucleic Acids Res, 42*(Database issue), D450-458. doi:10.1093/nar/gkt1151

Wickham, H. (2007). Reshaping data with the reshape package. *Journal of Statistical Software, 21*(12), 1-20.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*: Springer-Verlag New York.

Wigler, M., Levy, D., & Perucho, M. (1981). The somatic replication of DNA methylation. *Cell, 24*(1), 33-40. doi:10.1016/0092-8674(81)90498-0

Wilke, C. O. (2018). ggridges: Ridgeline Plots in 'ggplot2' (Version Version 0.5.1). Retrieved from https://CRAN.R-project.org/package=ggridges

Willer, C. J., Li, Y., & Abecasis, G. R. (2010). METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics, 26*(17), 2190-2191. doi:10.1093/bioinformatics/btq340

Wilson, L. E., Xu, Z., Harlid, S., White, A. J., Troester, M. A., Sandler, D. P., & Taylor, J. A. (2019). Alcohol and DNA Methylation: An Epigenome-Wide Association Study in

Blood and Normal Breast Tissue. *Am J Epidemiol, 188*(6), 1055-1065. doi:10.1093/aje/kwz032

Wong, C. C. Y., Smith, R. G., Hannon, E., Ramaswami, G., Parikshak, N. N., Assary, E., . . . Mill, J. (2019). Genome-wide DNA methylation profiling identifies convergent molecular signatures associated with idiopathic and syndromic autism in post-mortem human brain tissue. *Hum Mol Genet, 28*(13), 2201-2211. doi:10.1093/hmg/ddz052

Wright, J., Small, N., Raynor, P., Tuffnell, D., Bhopal, R., Cameron, N., . . . Born in Bradford Scientific Collaborators, G. (2013). Cohort Profile: the Born in Bradford multi-ethnic family cohort study. *Int J Epidemiol, 42*(4), 978-991. doi:10.1093/ije/dys112

Wu, X., & Zhang, Y. (2017). TET-mediated active DNA demethylation: mechanism, function and beyond. *Nat Rev Genet, 18*(9), 517-534. doi:10.1038/nrg.2017.33

Xiao, C. L., Zhu, S., He, M., Chen, Zhang, Q., Chen, Y., . . . Yan, G. R. (2018). N(6)-Methyladenine DNA Modification in the Human Genome. *Mol Cell, 71*(2), 306-318 e307. doi:10.1016/j.molcel.2018.06.015

Xu, C. J., Bonder, M. J., Soderhall, C., Bustamante, M., Baiz, N., Gehring, U., . . . Koppelman, G. H. (2017). The emerging landscape of dynamic DNA methylation in early childhood. *BMC Genomics, 18*(1), 25. doi:10.1186/s12864-016-3452-1

Xu, C. J., Soderhall, C., Bustamante, M., Baiz, N., Gruzieva, O., Gehring, U., . . . Koppelman, G. H. (2018). DNA methylation in childhood asthma: an epigenome-wide meta-analysis. *Lancet Respir Med, 6*(5), 379-388. doi:10.1016/S2213-2600(18)30052-3

Xue, Z., Huang, K., Cai, C., Cai, L., Jiang, C. Y., Feng, Y., . . . Fan, G. (2013). Genetic programs in human and mouse early embryos revealed by single-cell RNA sequencing. *Nature, 500*(7464), 593-597. doi:10.1038/nature12364

Yearim, A., Gelfman, S., Shayevitch, R., Melcer, S., Glaich, O., Mallm, J. P., . . . Ast, G. (2015). HP1 is involved in regulating the global impact of DNA methylation on alternative splicing. *Cell Rep, 10*(7), 1122-1134. doi:10.1016/j.celrep.2015.01.038

Yudell, M., Roberts, D., DeSalle, R., & Tishkoff, S. (2016). SCIENCE AND SOCIETY. Taking race out of human genetics. *Science, 351*(6273), 564-565. doi:10.1126/science.aac4951

Zaimi, I., Pei, D., Koestler, D. C., Marsit, C. J., De Vivo, I., Tworoger, S. S., . . . Michaud, D. S. (2018). Variation in DNA methylation of human blood over a 1-year period using the Illumina MethylationEPIC array. *Epigenetics, 13*(10-11), 1056-1071. doi:10.1080/15592294.2018.1530008

Zhang, B., & Horvath, S. (2005). A general framework for weighted gene co-expression network analysis. *Stat Appl Genet Mol Biol, 4*, Article17. doi:10.2202/1544-6115.1128

Zhang, Q., Vallerga, C. L., Walker, R. M., Lin, T., Henders, A. K., Montgomery, G. W., . . . Visscher, P. M. (2019). Improved precision of epigenetic clock estimates across tissues and its implication for biological ageing. *Genome Med, 11*(1), 54. doi:10.1186/s13073-019-0667-1

Zhang, W., Spector, T. D., Deloukas, P., Bell, J. T., & Engelhardt, B. E. (2015). Predicting genome-wide DNA methylation using methylation marks, genomic position, and DNA regulatory elements. *Genome Biol, 16*, 14. doi:10.1186/s13059-015-0581-9

Zhang, X., Biagini Myers, J. M., Yadagiri, V. K., Ulm, A., Chen, X., Weirauch, M. T., . . . Ji, H. (2017). Nasal DNA methylation differentiates corticosteroid treatment response in pediatric asthma: A pilot study. *PLoS One, 12*(10), e0186150. doi:10.1371/journal.pone.0186150

Zhang, Y. Y., Rohde, C., Tierling, S., Jurkowski, T. P., Bock, C., Santacruz, D., . . . Jeltsch, A. (2009). DNA Methylation Analysis of Chromosome 21 Gene Promoters at Single Base Pair and Single Allele Resolution. *Plos Genetics, 5*(3). doi:ARTN e1000438

10.1371/journal.pgen.1000438

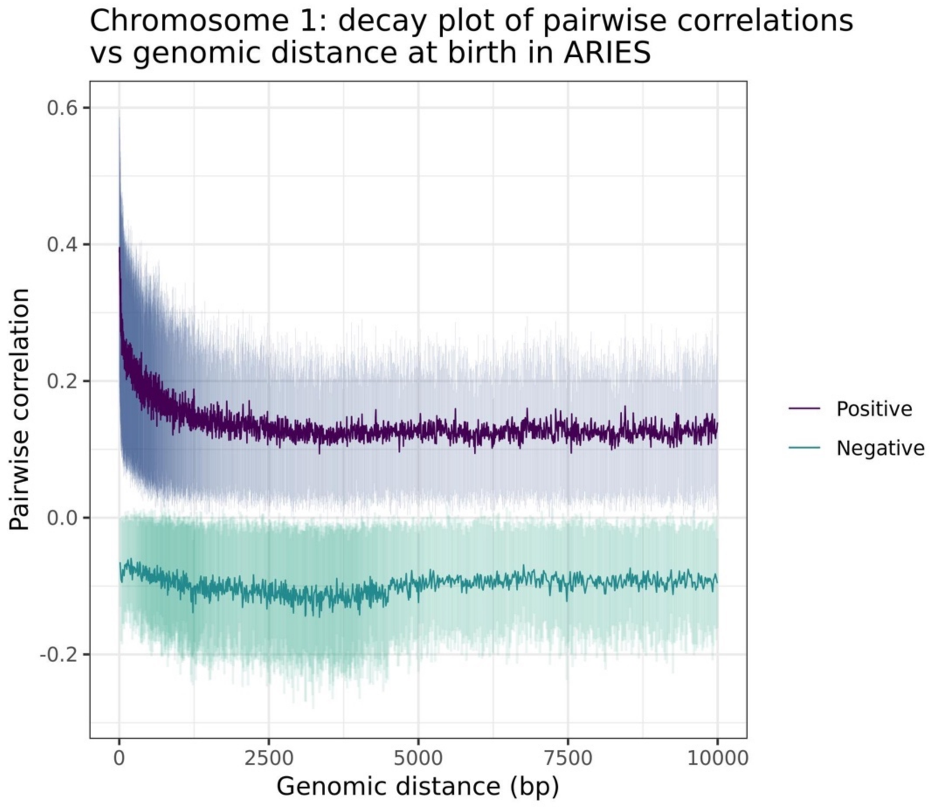
Zhou, W., Laird, P. W., & Shen, H. (2017). Comprehensive characterization, annotation and innovative use of Infinium DNA methylation BeadChip probes. *Nucleic Acids Res, 45*(4), e22. doi:10.1093/nar/gkw967

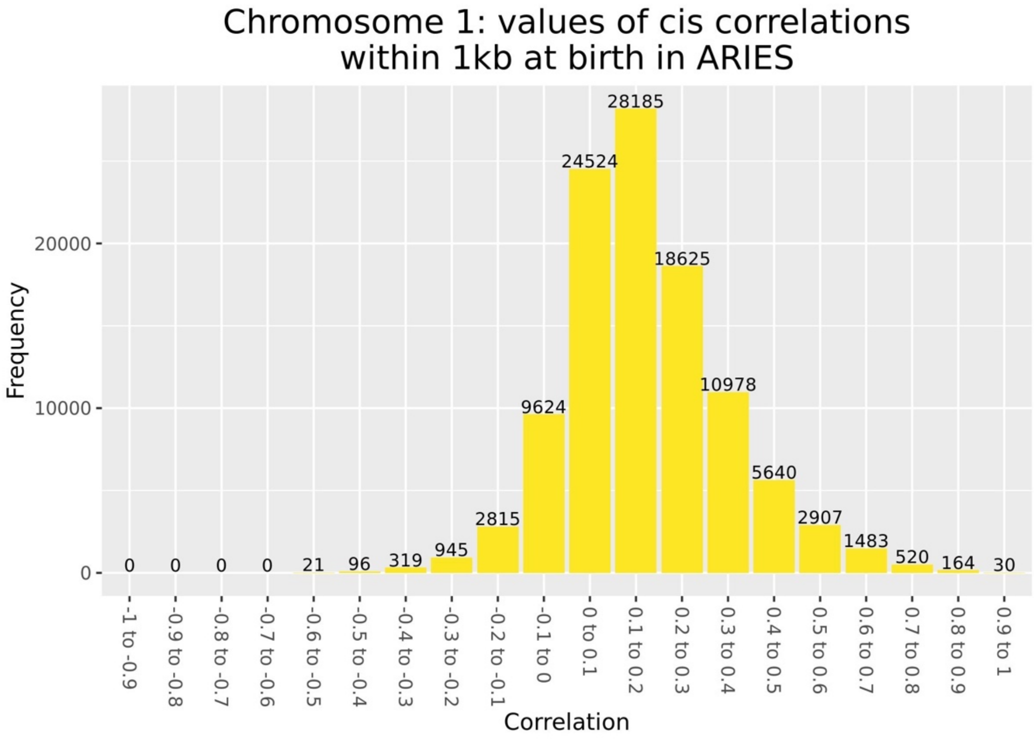
Zhu, H., Wang, G., & Qian, J. (2016). Transcription factors as readers and effectors of DNA methylation. *Nat Rev Genet, 17*(9), 551-565. doi:10.1038/nrg.2016.83

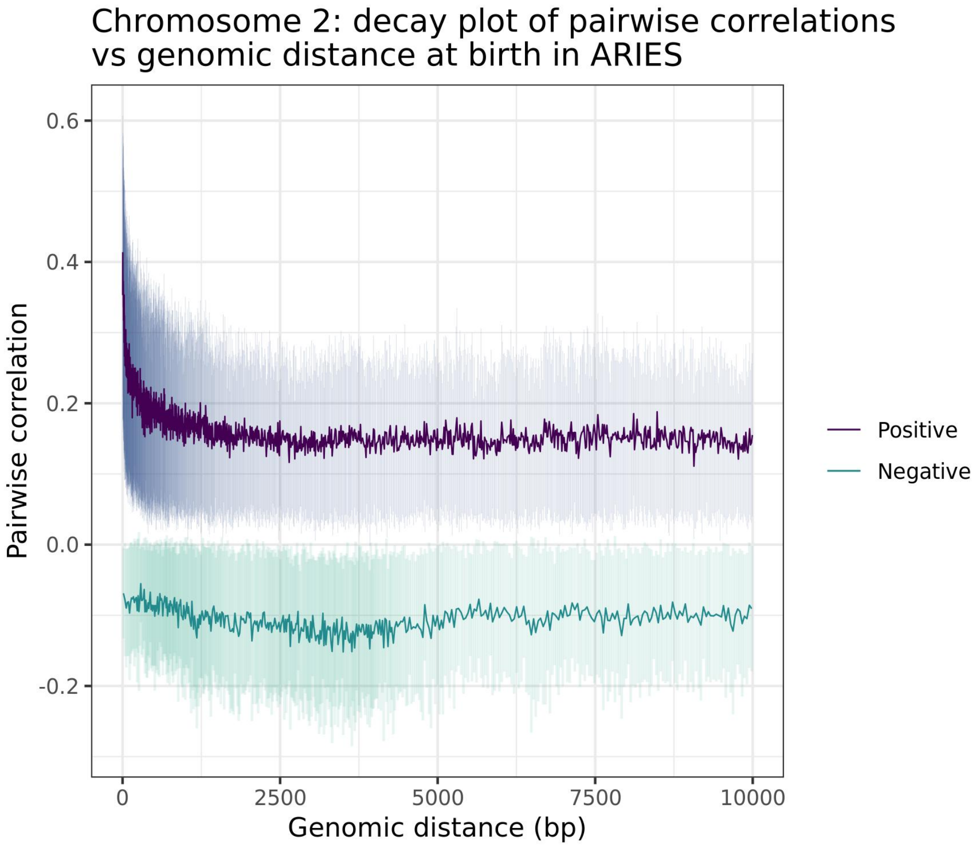
Appendix 1

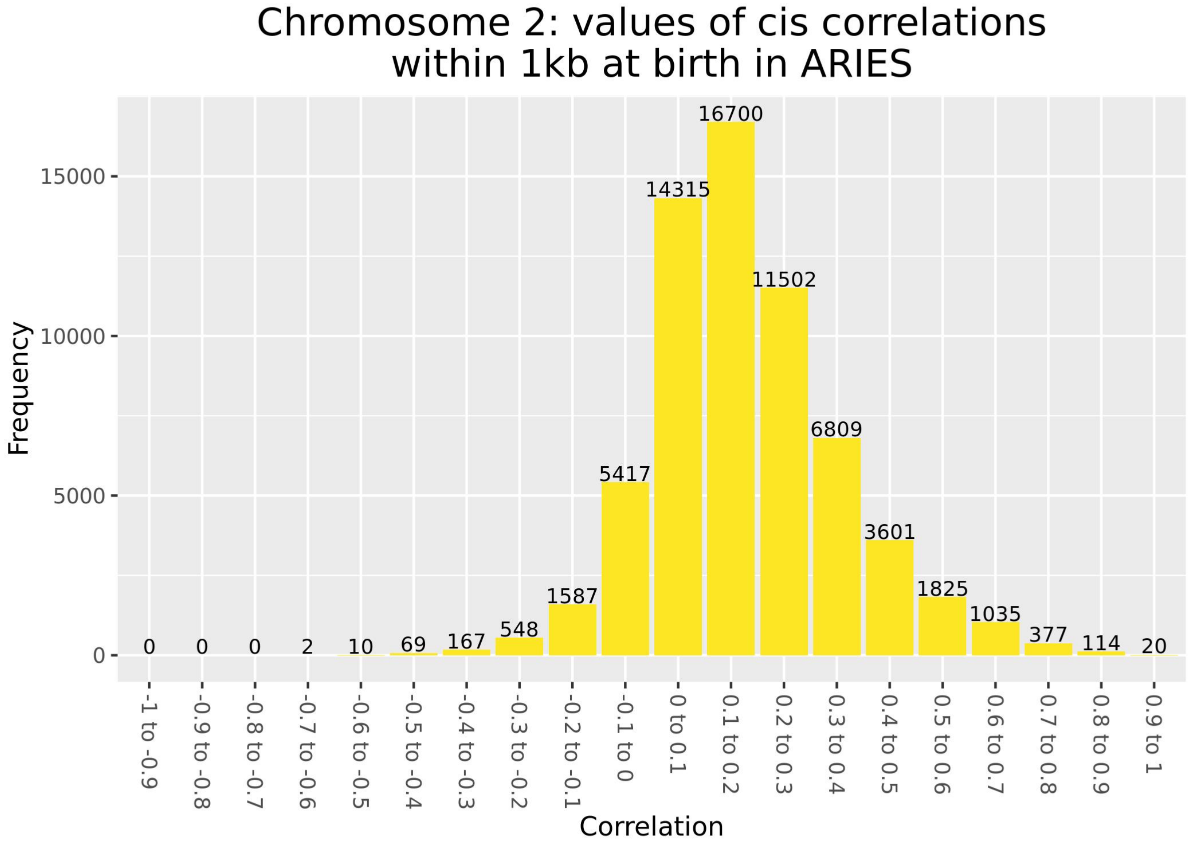
Cis correlation plots for chromosomes 1:5 and 15:19, for each ARIES timepoint

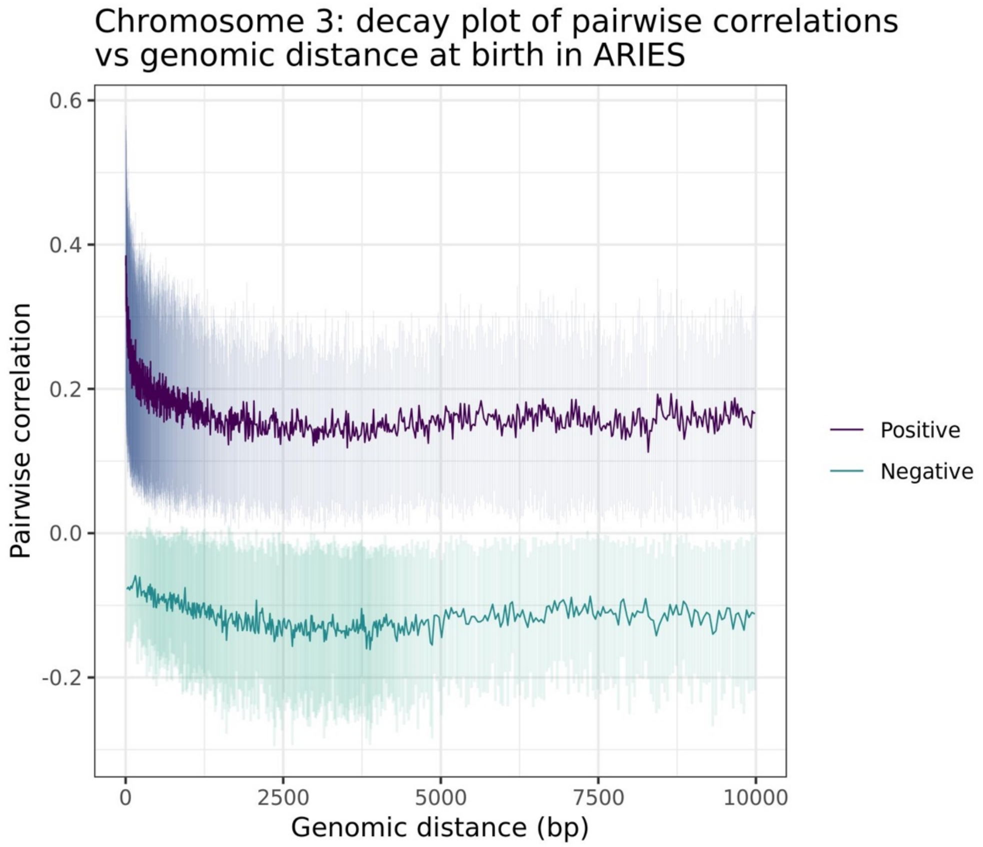
Birth

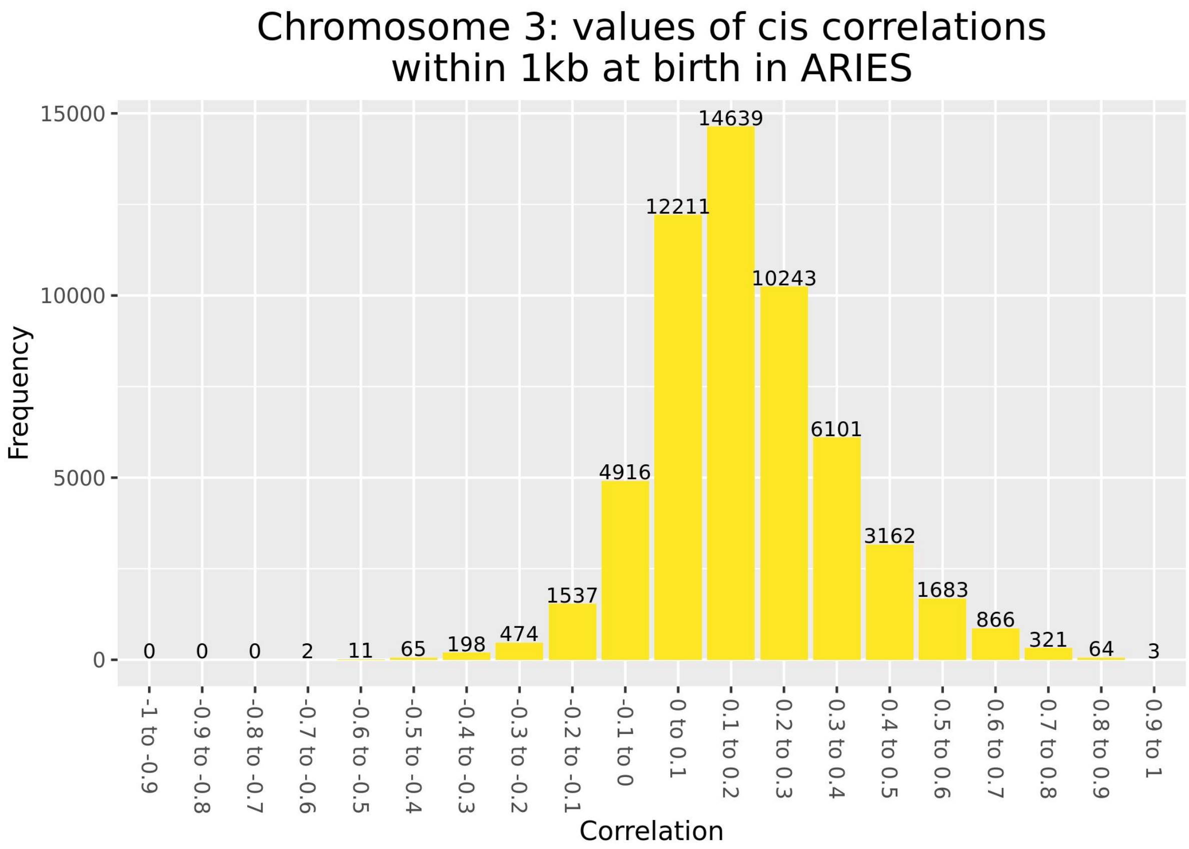


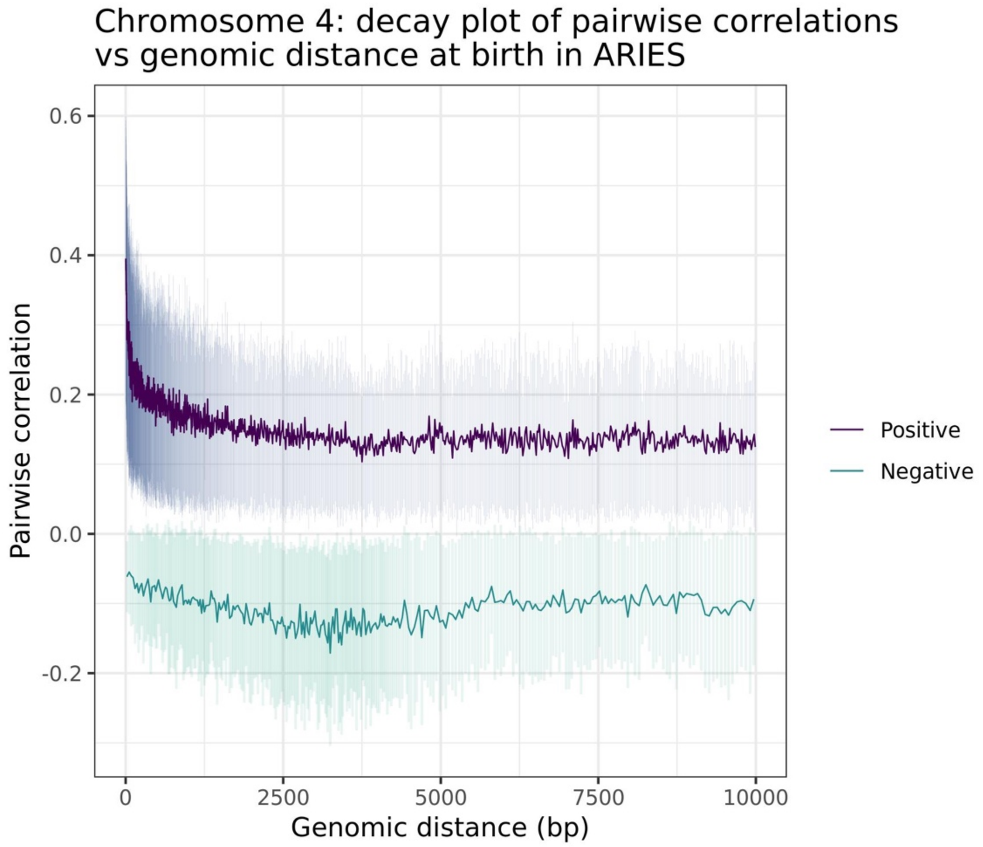



Chromosome 2: decay plot of pairwise correlations
vs genomic distance at birth in ARIES
Pairwise correlation
0.6
0.4
0.2
0.0
-0.2
0
2500
5000
7500
10000
Genomic distance (bp)
Positive
Negative

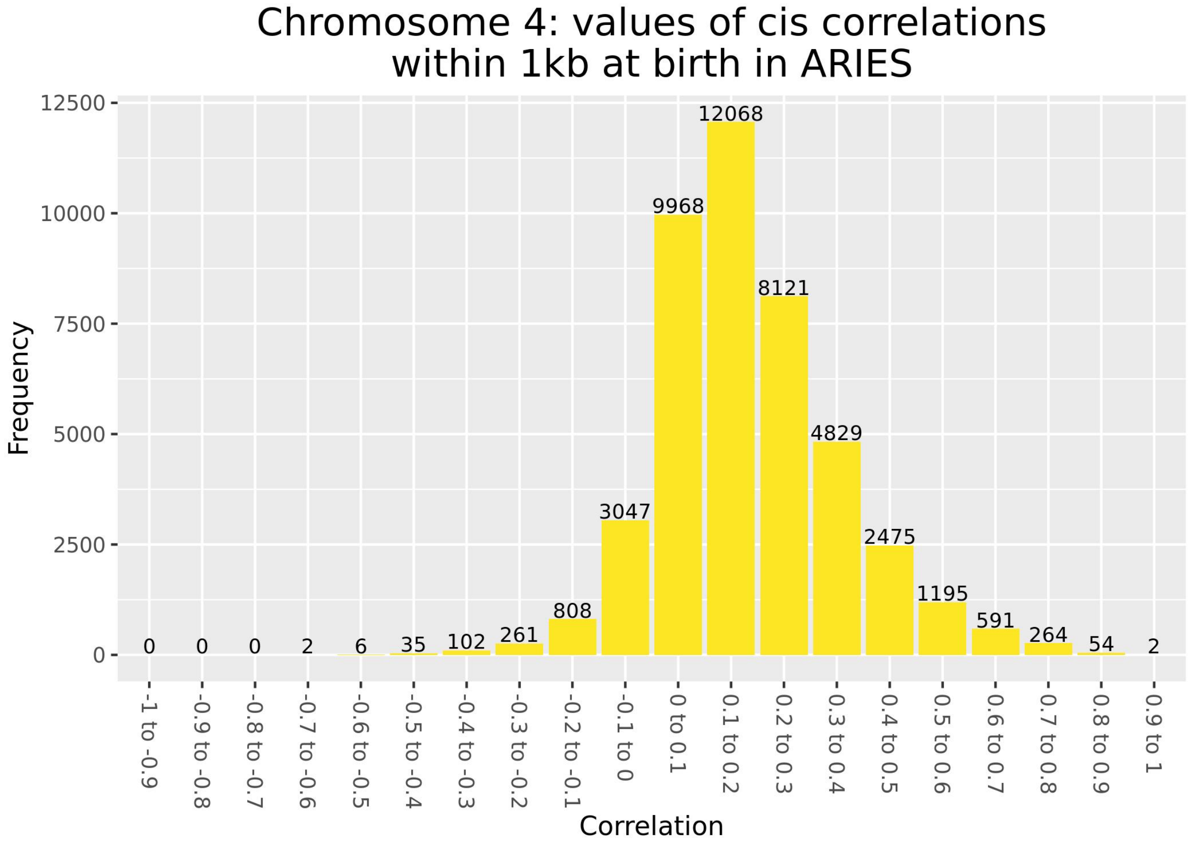

Chromosome 2: values of cis correlations
within 1kb at birth in ARIES
Frequency
15000
10000
5000
0
0
0
0
2
10
69
167
548
1587
5417
14315
16700
11502
6809
3601
1825
1035
377
114
20
-1 to -0.9
-0.9 to -0.8
-0.8 to -0.7
-0.7 to -0.6
-0.6 to -0.5
-0.5 to -0.4
-0.4 to -0.3
-0.3 to -0.2
-0.2 to -0.1
-0.1 to 0
0 to 0.1
0.1 to 0.2
0.2 to 0.3
0.3 to 0.4
0.4 to 0.5
0.5 to 0.6
0.6 to 0.7
0.7 to 0.8
0.8 to 0.9
0.9 to 1
Correlation


Chromosome 3: decay plot of pairwise correlations vs genomic distance at birth in ARIES
Pairwise correlation
0.6
0.4
0.2
0.0
-0.2
0
2500
5000
7500
10000
Genomic distance (bp)
Positive
Negative

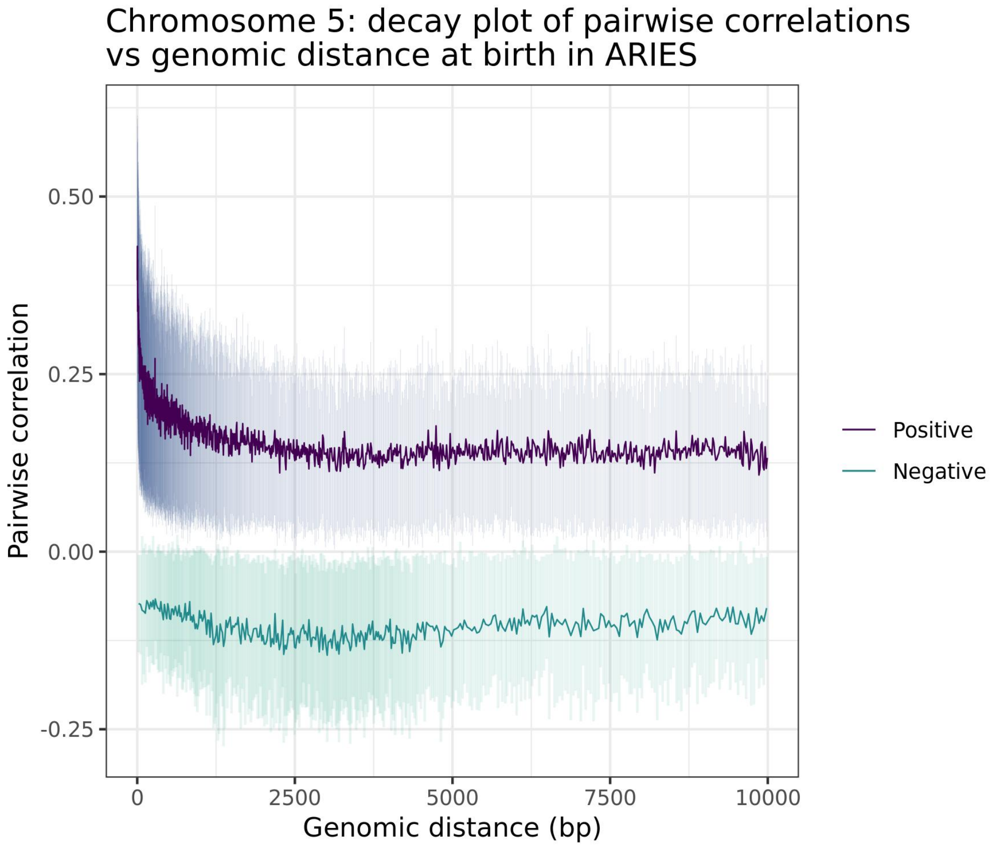

Chromosome 3: values of cis correlations within 1kb at birth in ARIES
Frequency
15000
10000
5000
0
0
0
0
2
11
65
198
474
1537
4916
12211
14639
10243
6101
3162
1683
866
321
64
3
-1 to -0.9
-0.9 to -0.8
-0.8 to -0.7
-0.7 to -0.6
-0.6 to -0.5
-0.5 to -0.4
-0.4 to -0.3
-0.3 to -0.2
-0.2 to -0.1
-0.1 to 0
0 to 0.1
0.1 to 0.2
0.2 to 0.3
0.3 to 0.4
0.4 to 0.5
0.5 to 0.6
0.6 to 0.7
0.7 to 0.8
0.8 to 0.9
0.9 to 1
Correlation



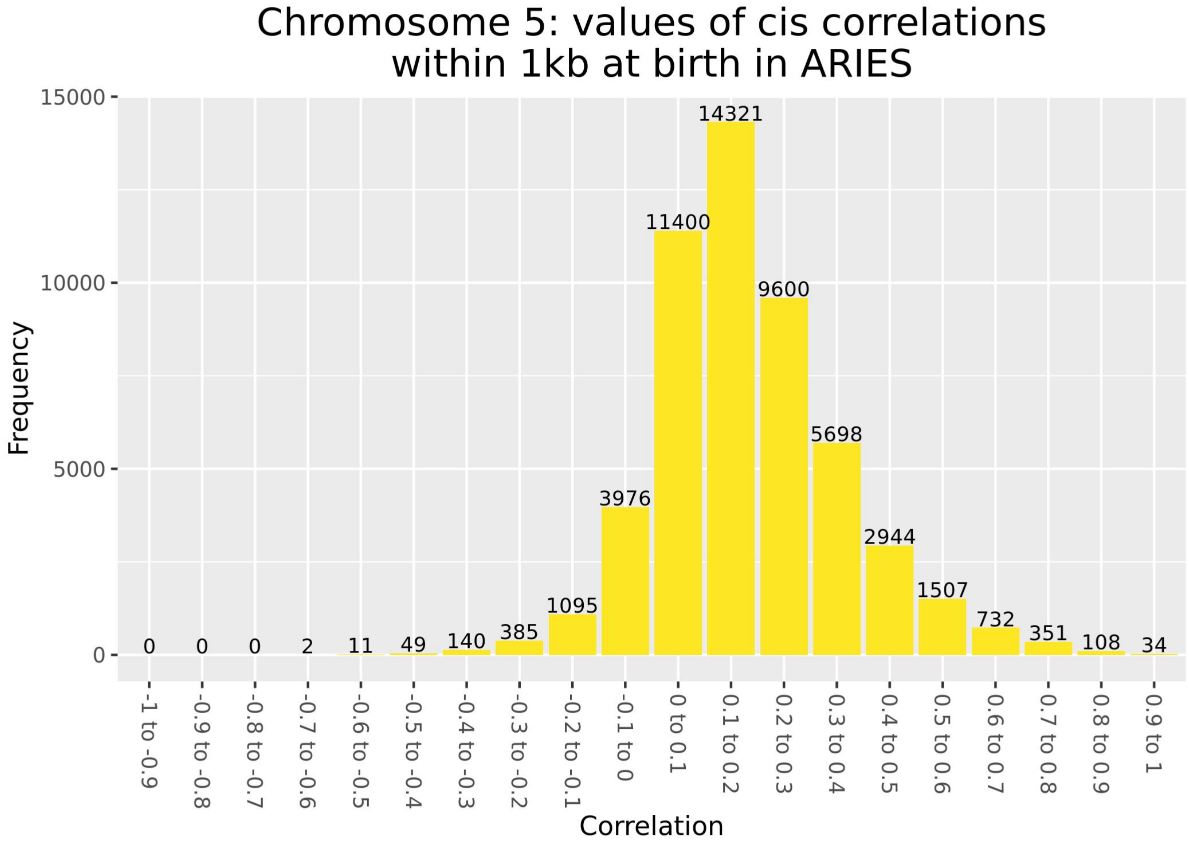
Chromosome 4: decay plot of pairwise correlations vs genomic distance at birth in ARIES



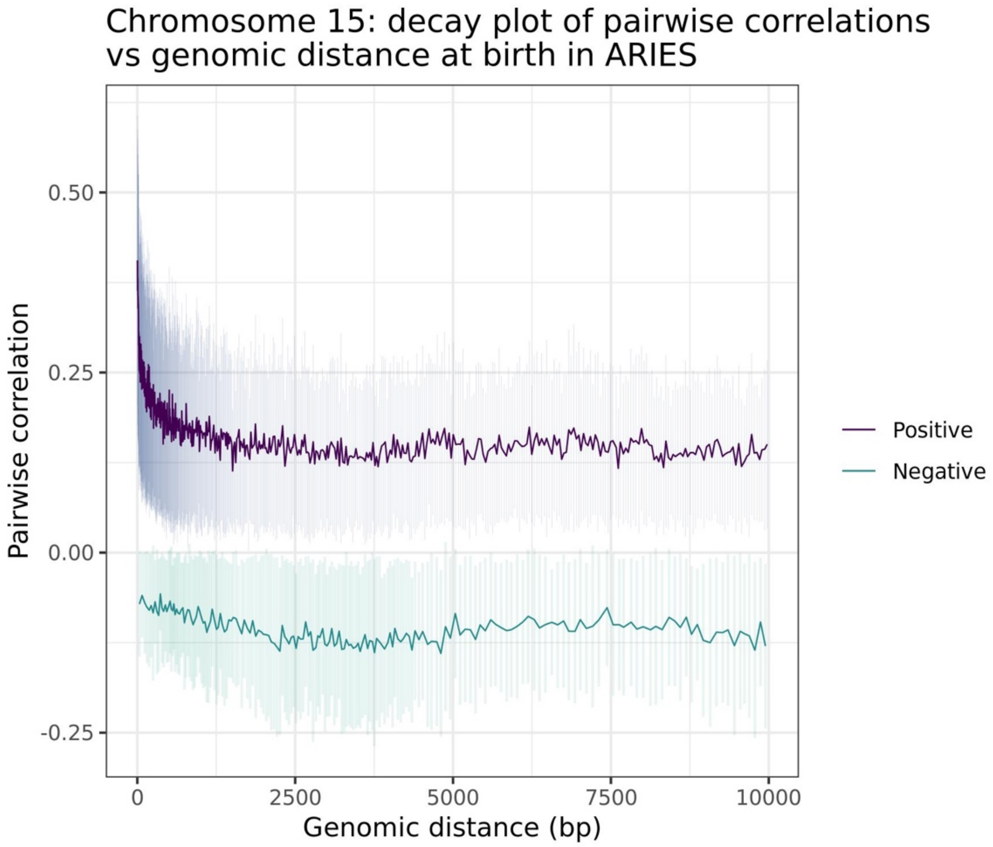
Chromosome 4: values of cis correlations within 1kb at birth in ARIES

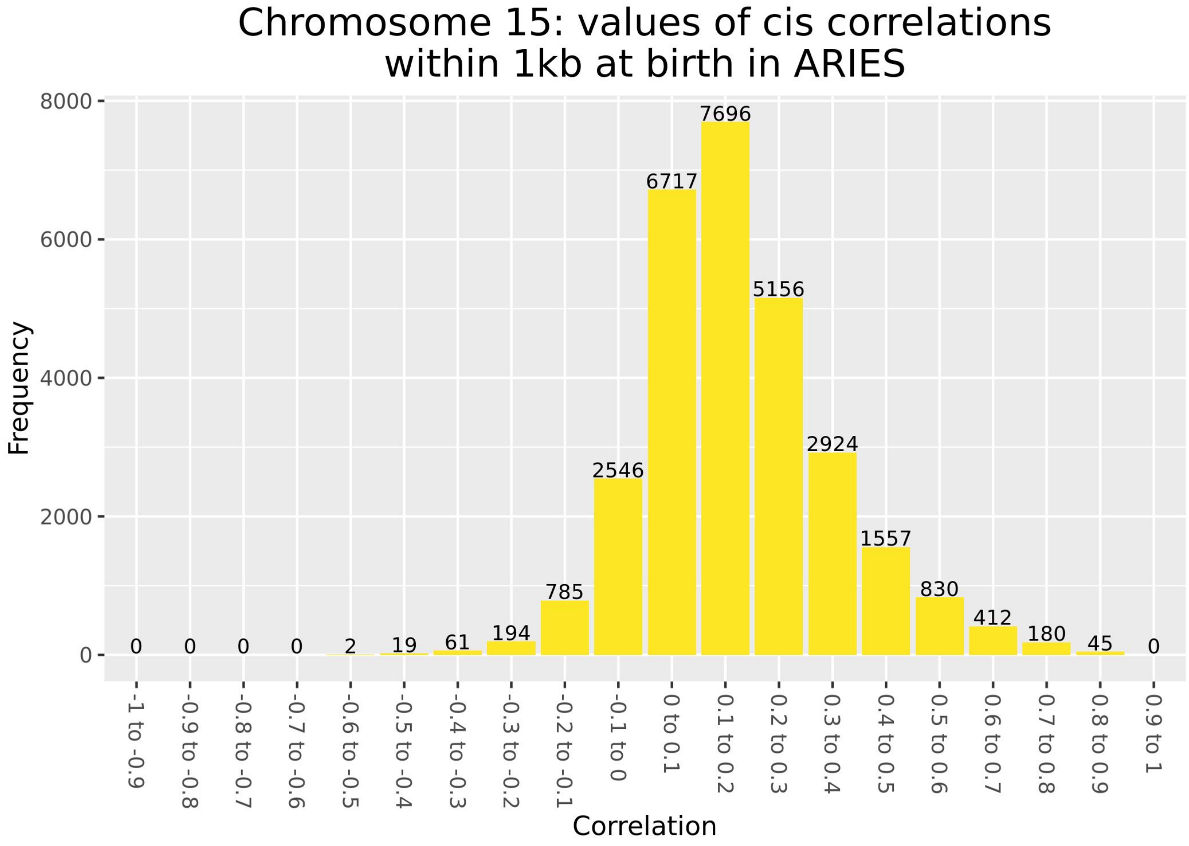


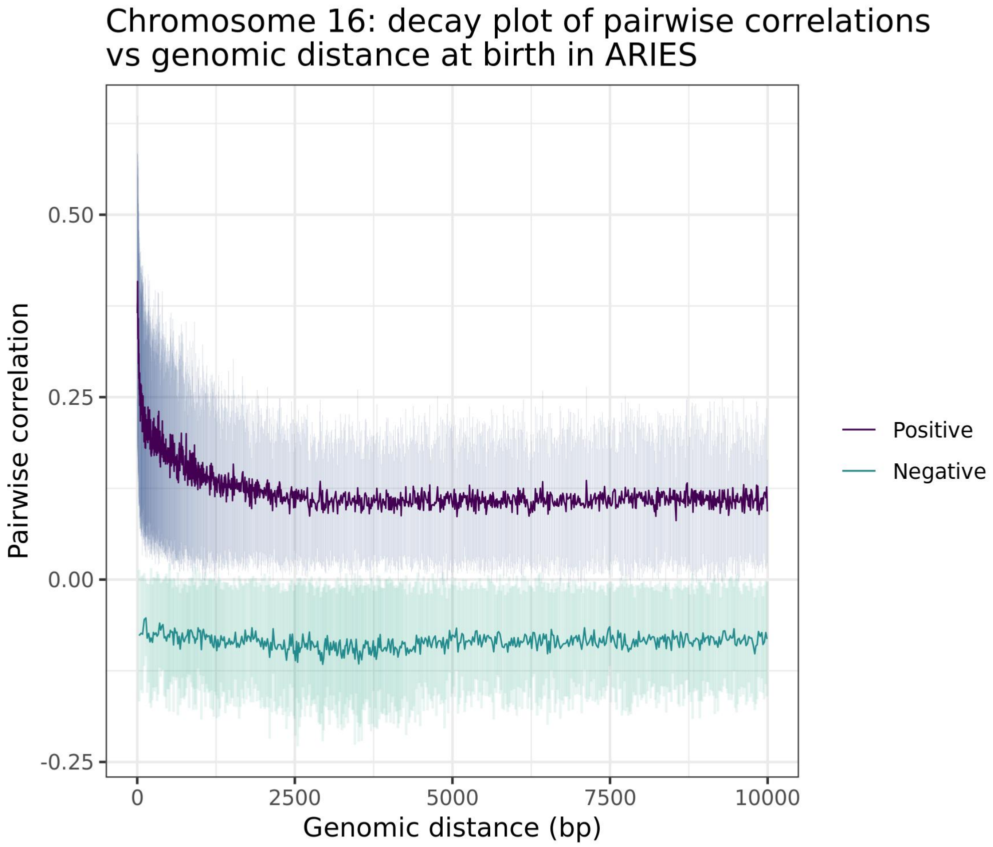
Chromosome 5: decay plot of pairwise correlations vs genomic distance at birth in ARIES



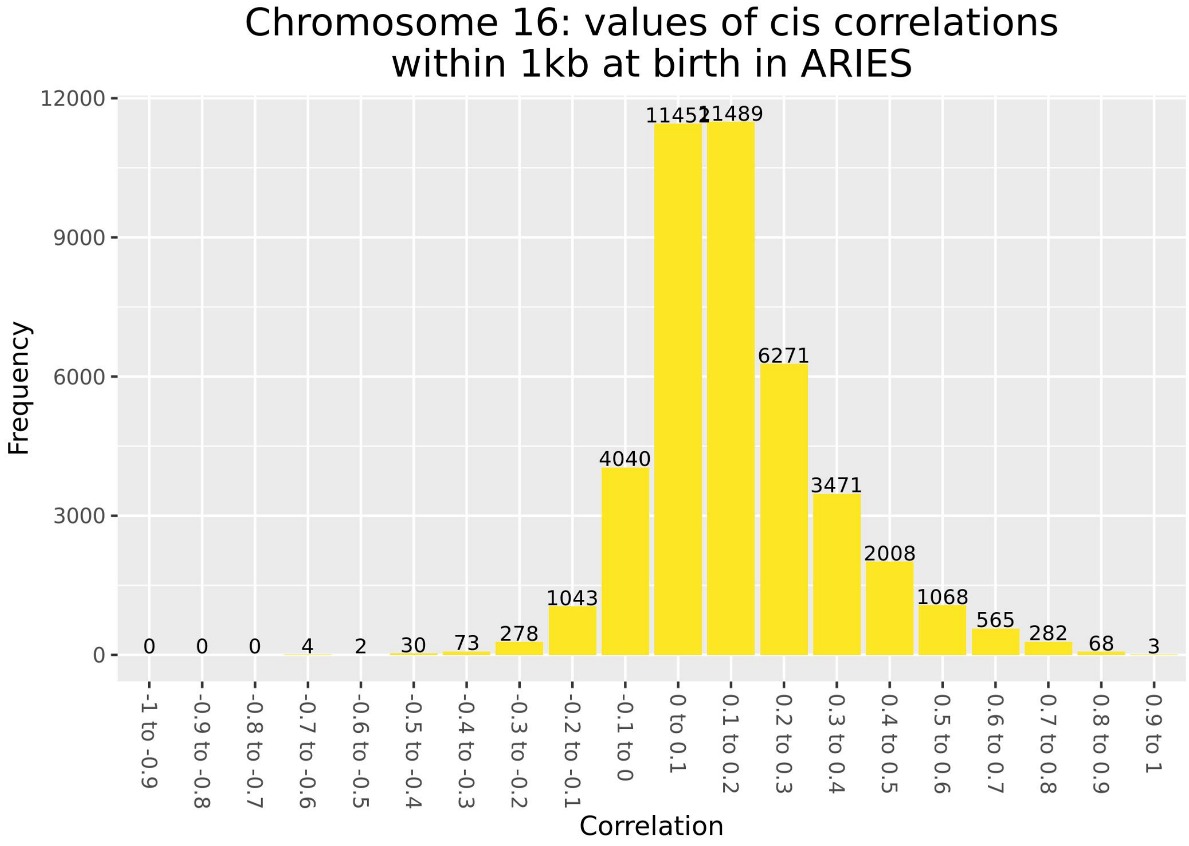
Chromosome 5: values of cis correlations within 1kb at birth in ARIES


Chromosome 15: decay plot of pairwise correlations
vs genomic distance at birth in ARIES
Pairwise correlation
0.50
0.25
0.00
-0.25
0
2500
5000
7500
10000
Genomic distance (bp)
Positive
Negative

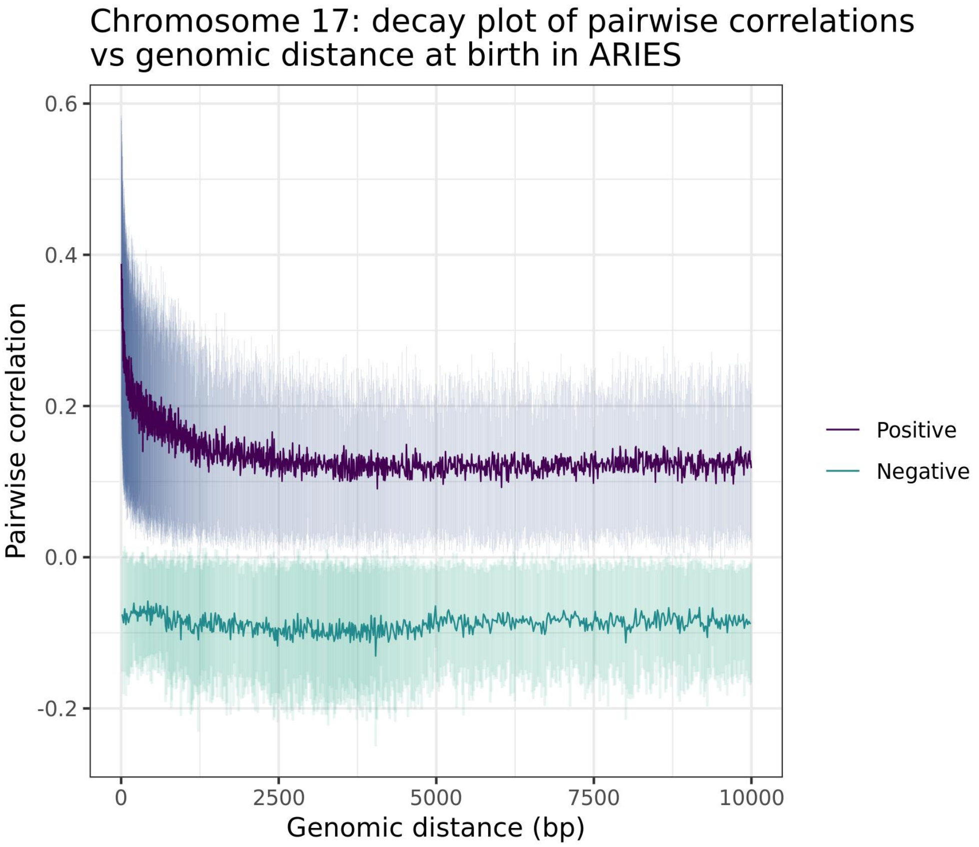

Chromosome 15: values of cis correlations
within 1kb at birth in ARIES
Frequency
8000
6000
4000
2000
0
0
0
0
0
2
19
61
194
785
2546
6717
7696
5156
2924
1557
830
412
180
45
0
-1 to -0.9
-0.9 to -0.8
-0.8 to -0.7
-0.7 to -0.6
-0.6 to -0.5
-0.5 to -0.4
-0.4 to -0.3
-0.3 to -0.2
-0.2 to -0.1
-0.1 to 0
0 to 0.1
0.1 to 0.2
0.2 to 0.3
0.3 to 0.4
0.4 to 0.5
0.5 to 0.6
0.6 to 0.7
0.7 to 0.8
0.8 to 0.9
0.9 to 1
Correlation

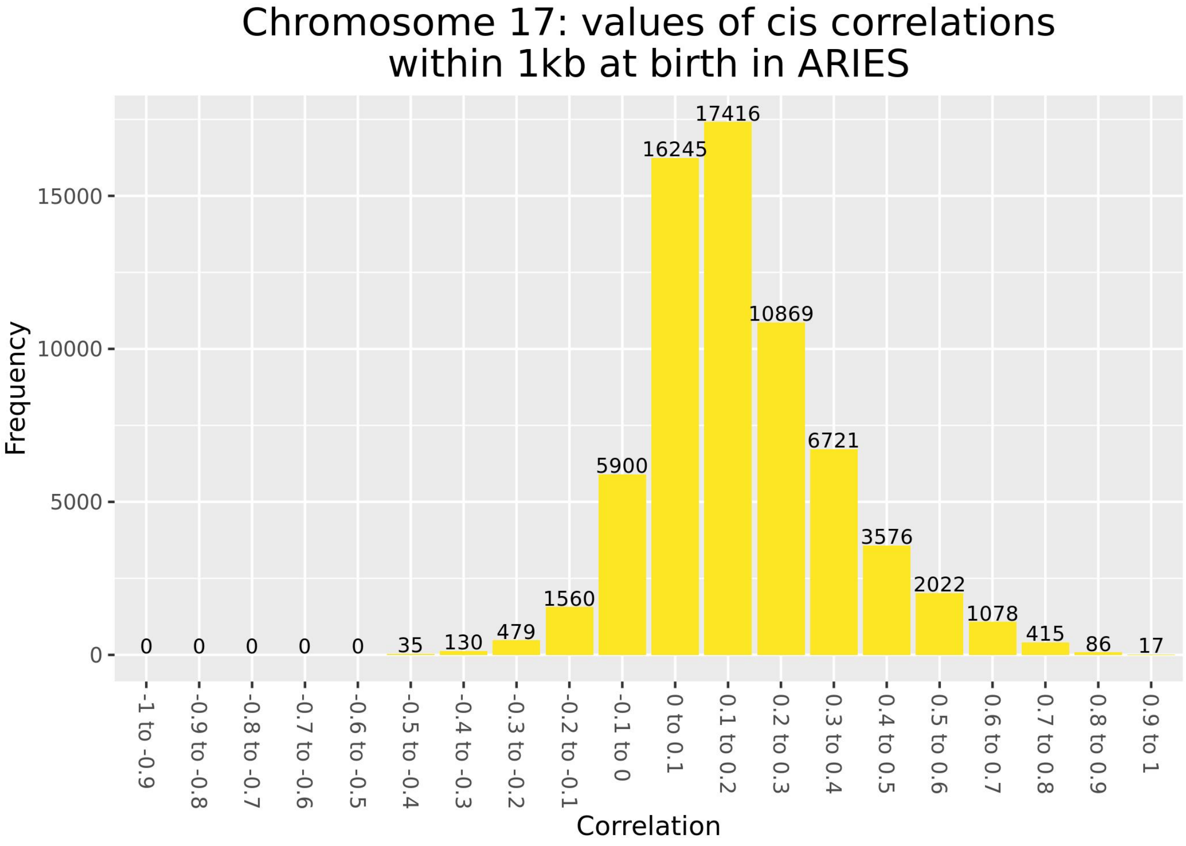


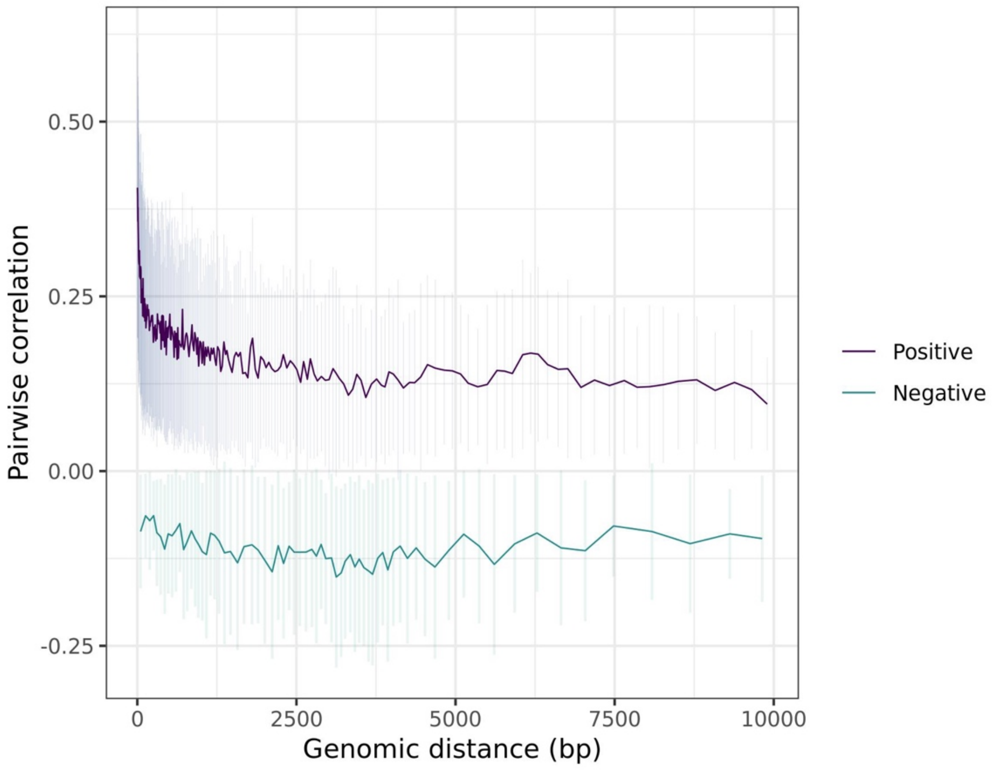
Chromosome 16: decay plot of pairwise correlations vs genomic distance at birth in ARIES



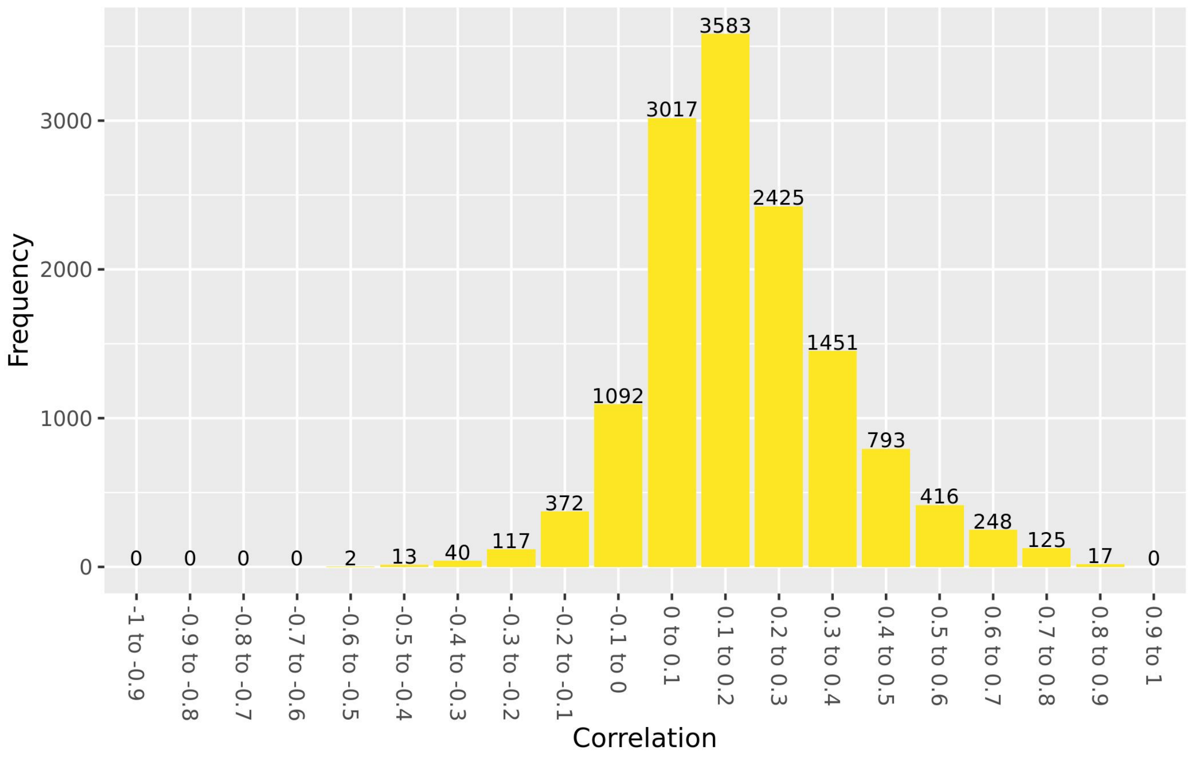
Chromosome 16: values of cis correlations within 1kb at birth in ARIES


Chromosome 17: decay plot of pairwise correlations
vs genomic distance at birth in ARIES
Pairwise correlation
0.6
0.4
0.2
0.0
-0.2
0
2500
5000
7500
10000
Genomic distance (bp)
Positive
Negative


Chromosome 17: values of cis correlations
within 1kb at birth in ARIES
Frequency
15000
10000
5000
0
0
0
0
0
0
35
130
479
1560
5900
16245
17416
10869
6721
3576
2022
1078
415
86
17
-1 to -0.9
-0.9 to -0.8
-0.8 to -0.7
-0.7 to -0.6
-0.6 to -0.5
-0.5 to -0.4
-0.4 to -0.3
-0.3 to -0.2
-0.2 to -0.1
-0.1 to 0
0 to 0.1
0.1 to 0.2
0.2 to 0.3
0.3 to 0.4
0.4 to 0.5
0.5 to 0.6
0.6 to 0.7
0.7 to 0.8
0.8 to 0.9
0.9 to 1
Correlation

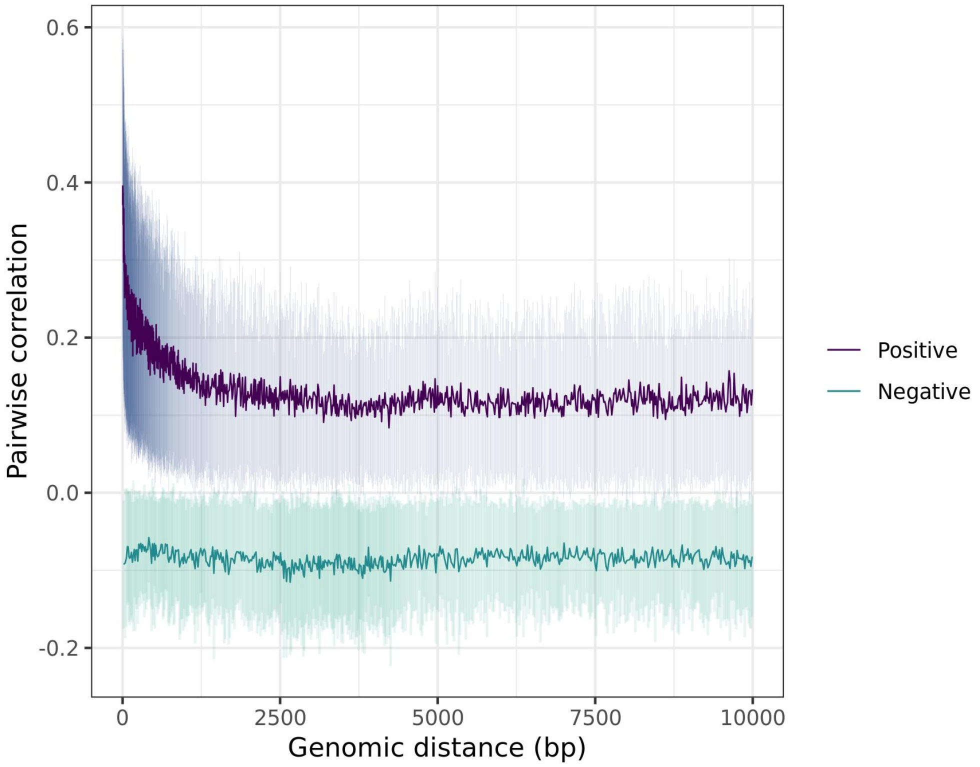


Chromosome 18: decay plot of pairwise correlations vs genomic distance at birth in ARIES

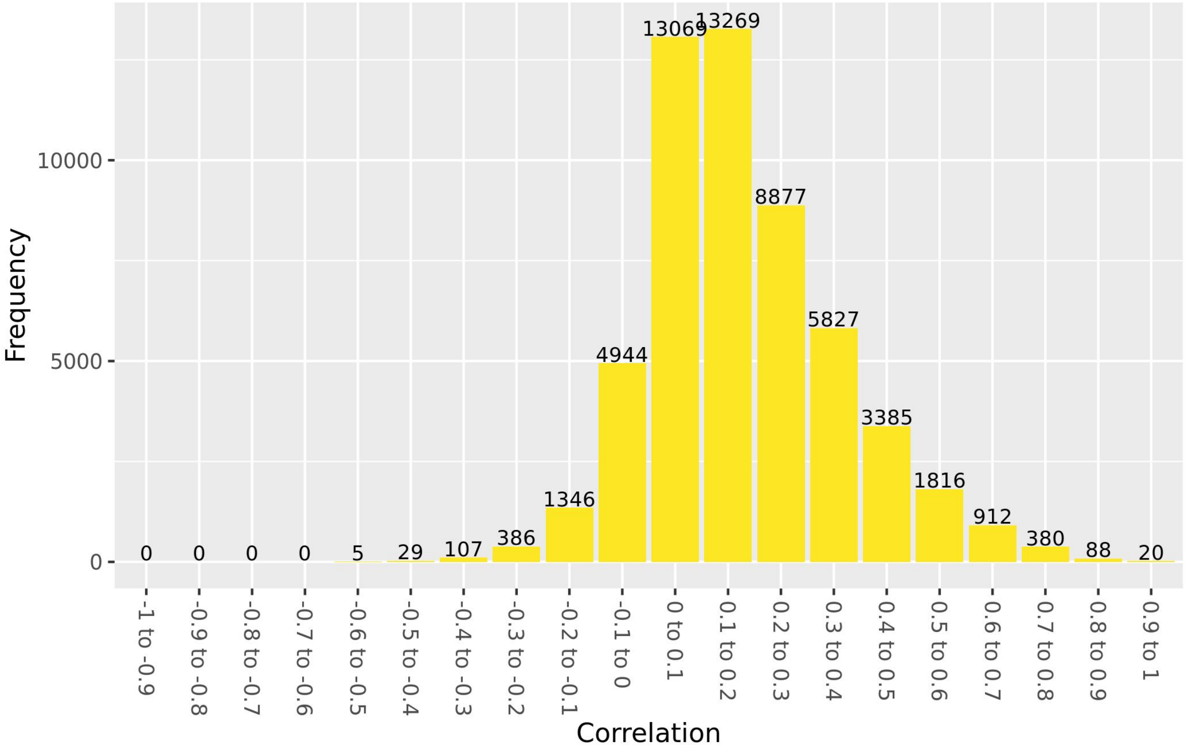


Chromosome 18: values of cis correlations within 1kb at birth in ARIES

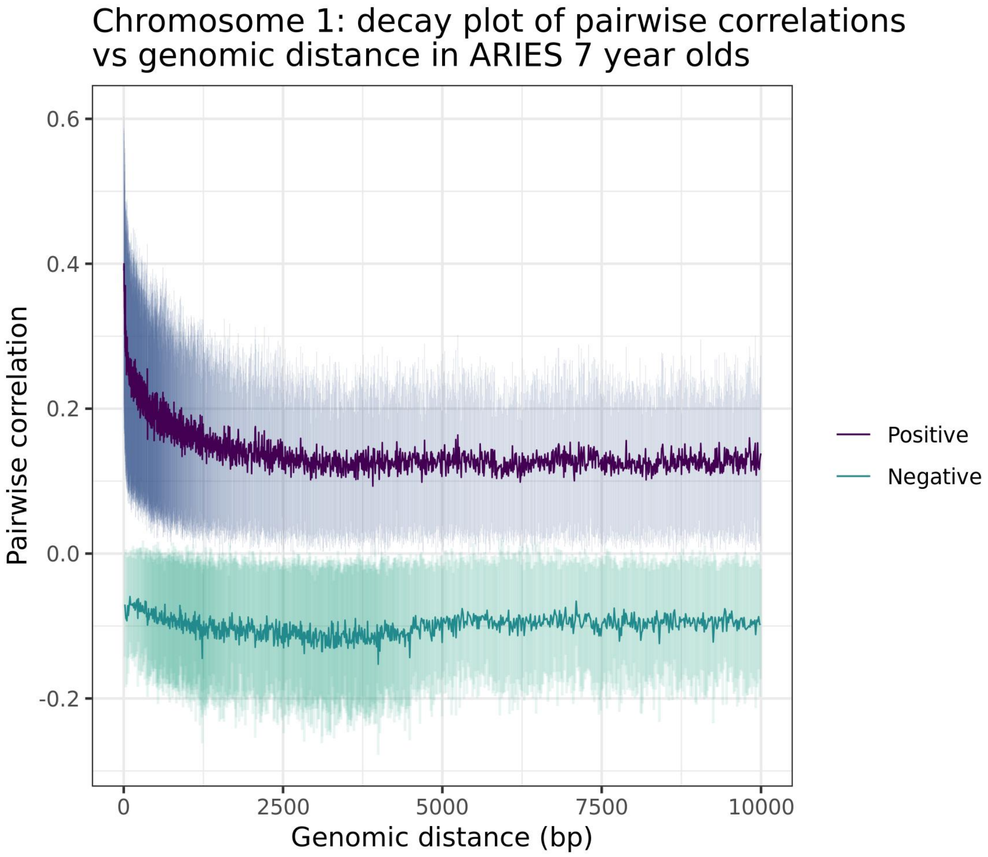
Chromosome 19: decay plot of pairwise correlations vs genomic distance at birth in ARIES

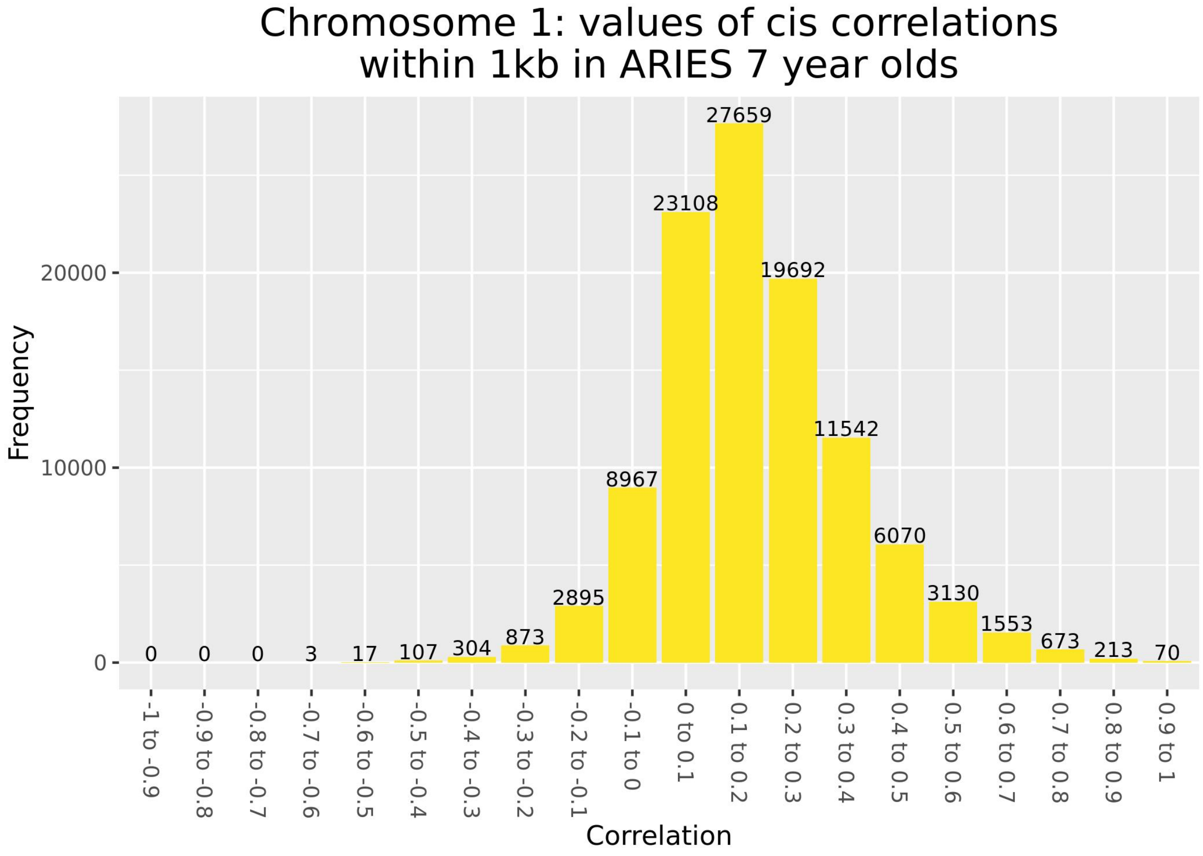


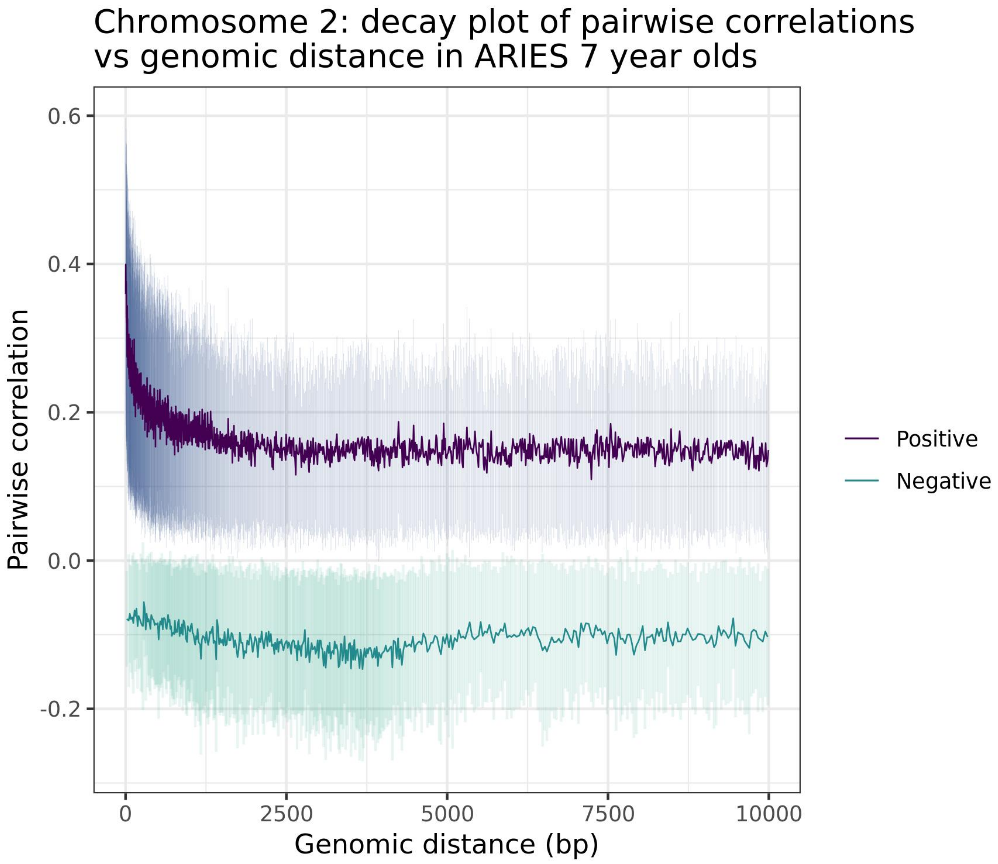
Chromosome 19: values of cis correlations within 1kb at birth in ARIES

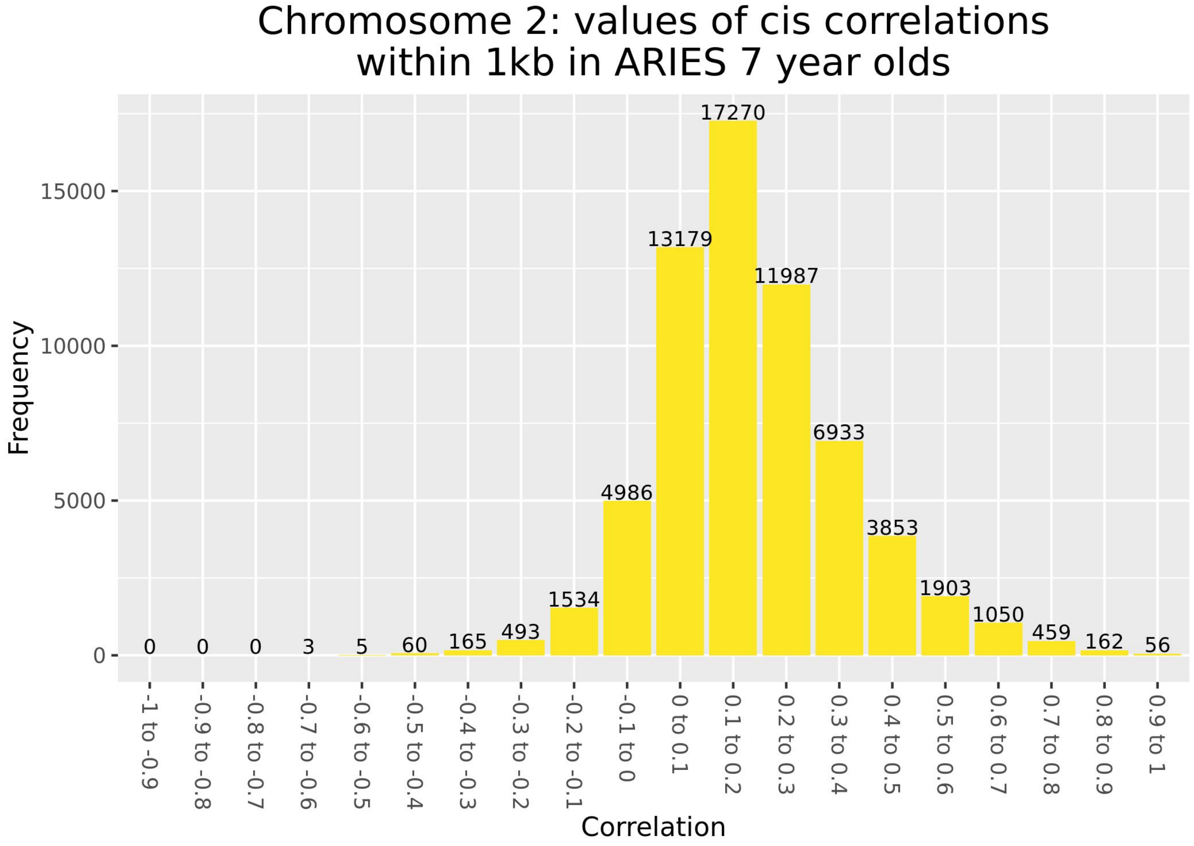


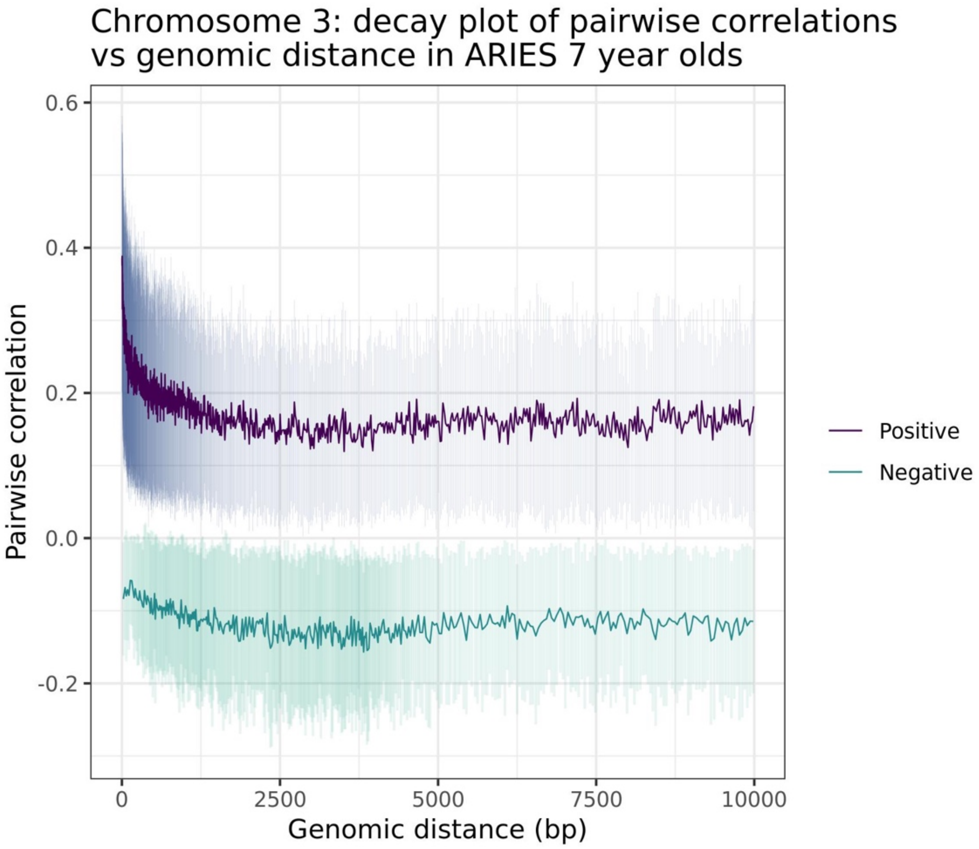
7 years

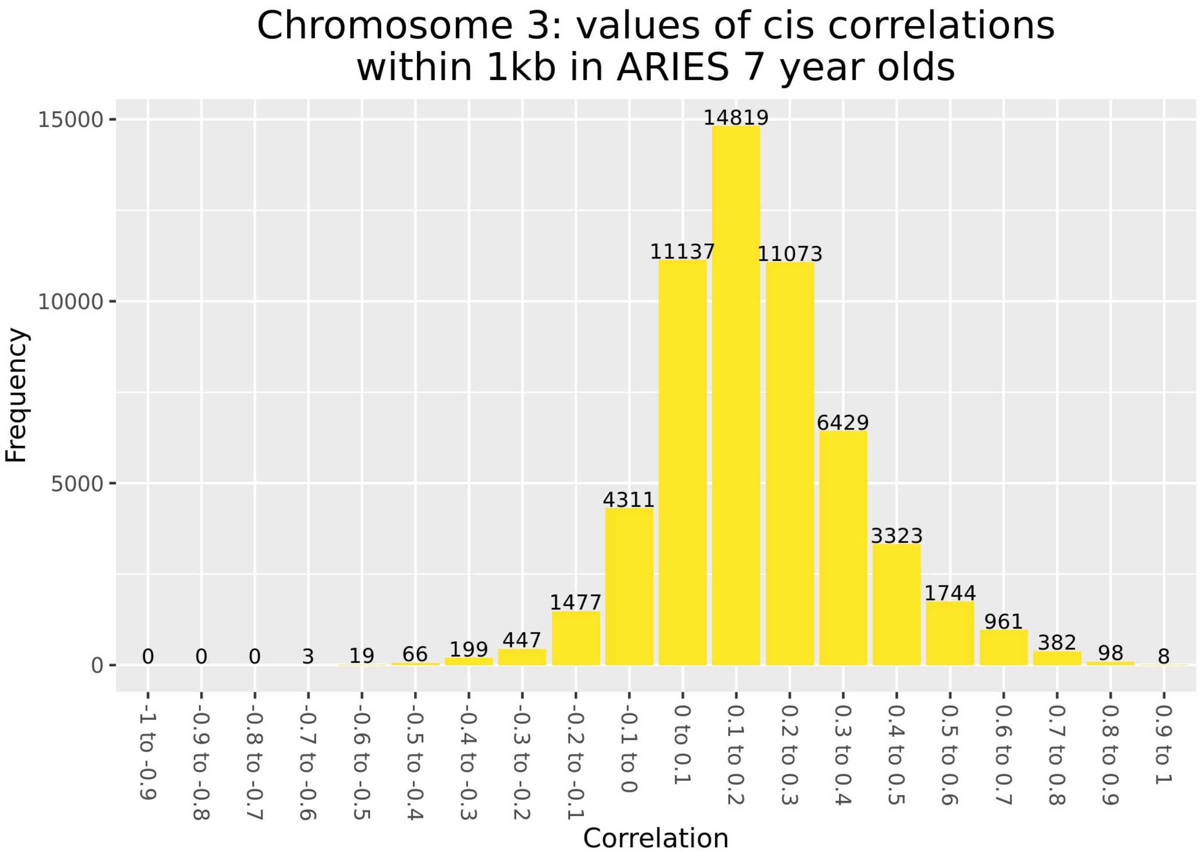

Chromosome 1: decay plot of pairwise correlations
vs genomic distance in ARIES 7 year olds
Pairwise correlation
0.6
0.4
0.2
0.0
-0.2
0
2500
5000
7500
10000
Genomic distance (bp)
Positive
Negative

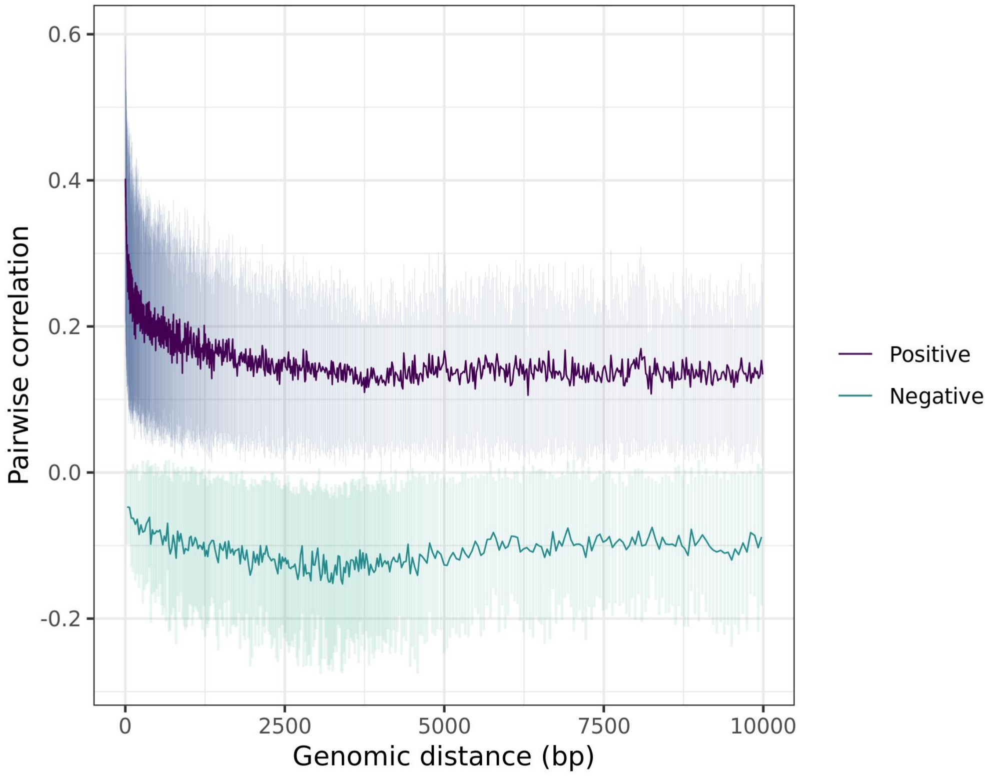

Chromosome 1: values of cis correlations
within 1kb in ARIES 7 year olds
Frequency
20000
10000
0
0
0
0
3
17
107
304
873
2895
8967
23108
27659
19692
11542
6070
3130
1553
673
213
70
-1 to -0.9
-0.9 to -0.8
-0.8 to -0.7
-0.7 to -0.6
-0.6 to -0.5
-0.5 to -0.4
-0.4 to -0.3
-0.3 to -0.2
-0.2 to -0.1
-0.1 to 0
0 to 0.1
0.1 to 0.2
0.2 to 0.3
0.3 to 0.4
0.4 to 0.5
0.5 to 0.6
0.6 to 0.7
0.7 to 0.8
0.8 to 0.9
0.9 to 1
Correlation


Chromosome 2: decay plot of pairwise correlations
vs genomic distance in ARIES 7 year olds
Pairwise correlation
0.6
0.4
0.2
0.0
-0.2
0
2500
5000
7500
10000
Genomic distance (bp)
Positive
Negative

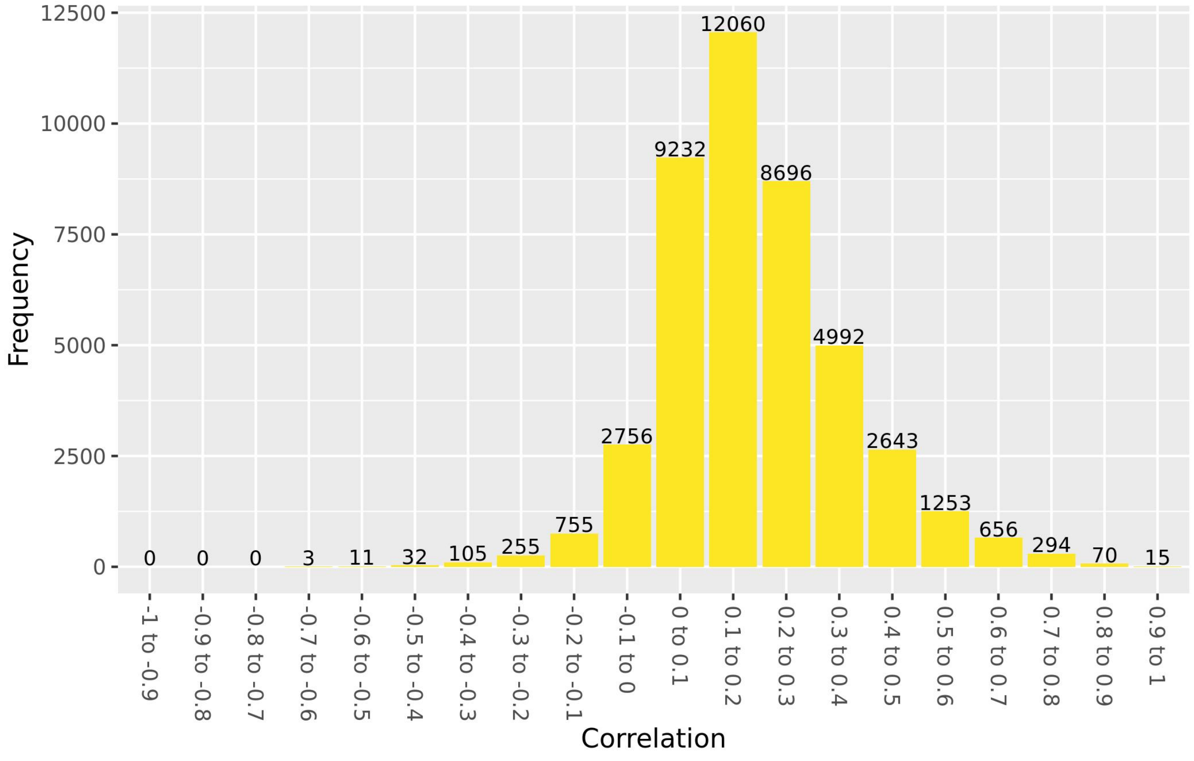

Chromosome 2: values of cis correlations
within 1kb in ARIES 7 year olds
Frequency
15000
10000
5000
0
0
0
0
3
5
60
165
493
1534
4986
13179
17270
11987
6933
3853
1903
1050
459
162
56
-1 to -0.9
-0.9 to -0.8
-0.8 to -0.7
-0.7 to -0.6
-0.6 to -0.5
-0.5 to -0.4
-0.4 to -0.3
-0.3 to -0.2
-0.2 to -0.1
-0.1 to 0
0 to 0.1
0.1 to 0.2
0.2 to 0.3
0.3 to 0.4
0.4 to 0.5
0.5 to 0.6
0.6 to 0.7
0.7 to 0.8
0.8 to 0.9
0.9 to 1
Correlation


Chromosome 3: decay plot of pairwise correlations
vs genomic distance in ARIES 7 year olds
Pairwise correlation
0.6
0.4
0.2
0.0
-0.2
0
2500
5000
7500
10000
Genomic distance (bp)
Positive
Negative

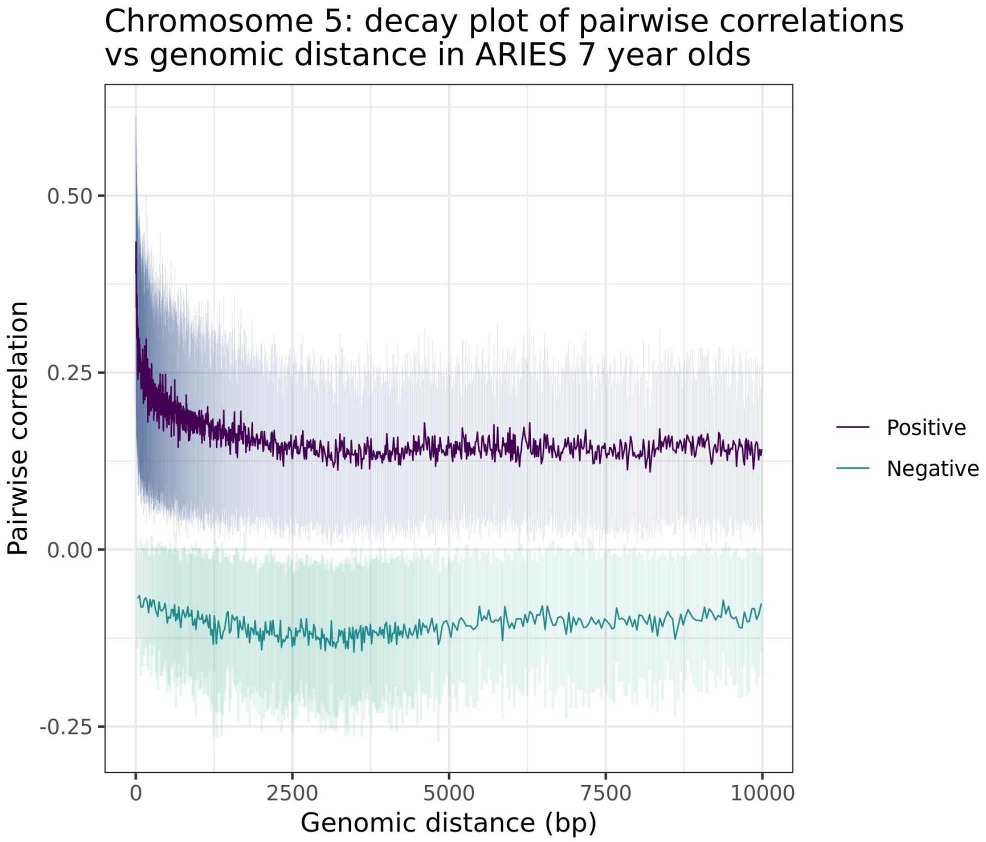

Chromosome 3: values of cis correlations
within 1kb in ARIES 7 year olds
Frequency
15000
10000
5000
0
0
0
0
3
19
66
199
447
1477
4311
11137
14819
11073
6429
3323
1744
961
382
98
8
-1 to -0.9
-0.9 to -0.8
-0.8 to -0.7
-0.7 to -0.6
-0.6 to -0.5
-0.5 to -0.4
-0.4 to -0.3
-0.3 to -0.2
-0.2 to -0.1
-0.1 to 0
0 to 0.1
0.1 to 0.2
0.2 to 0.3
0.3 to 0.4
0.4 to 0.5
0.5 to 0.6
0.6 to 0.7
0.7 to 0.8
0.8 to 0.9
0.9 to 1
Correlation

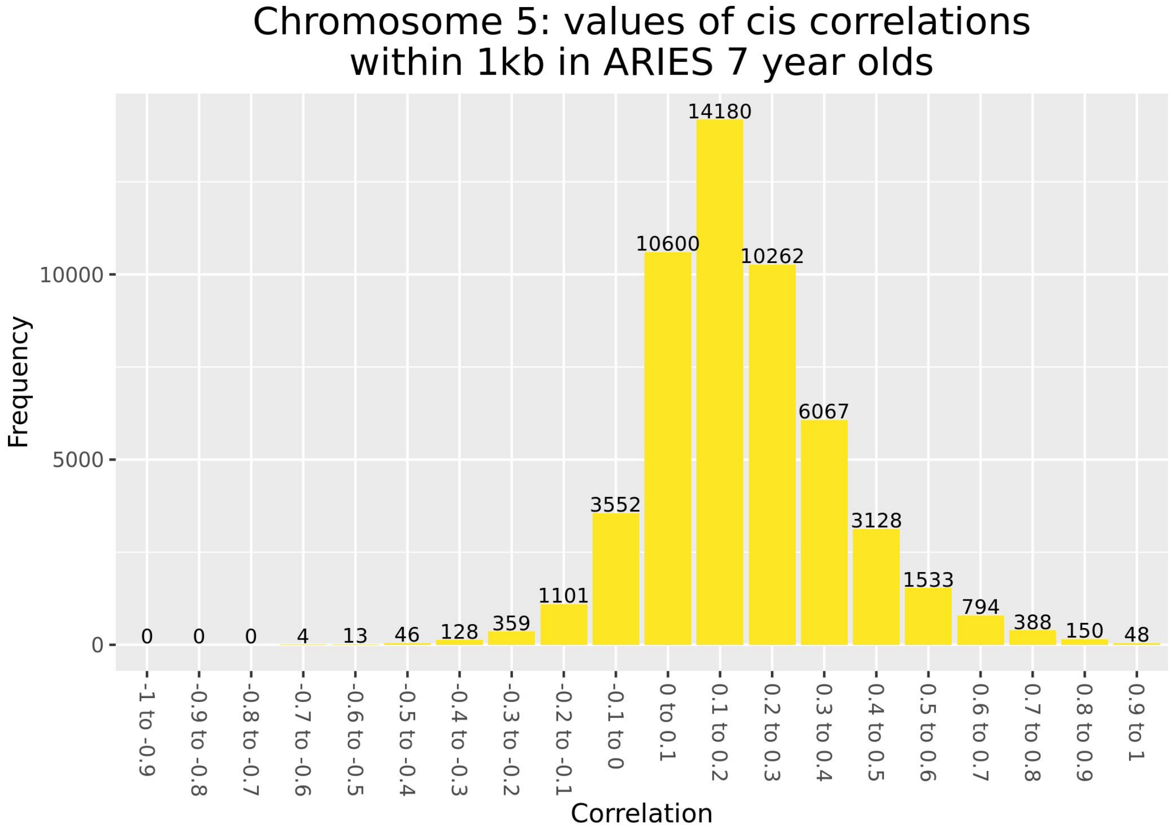


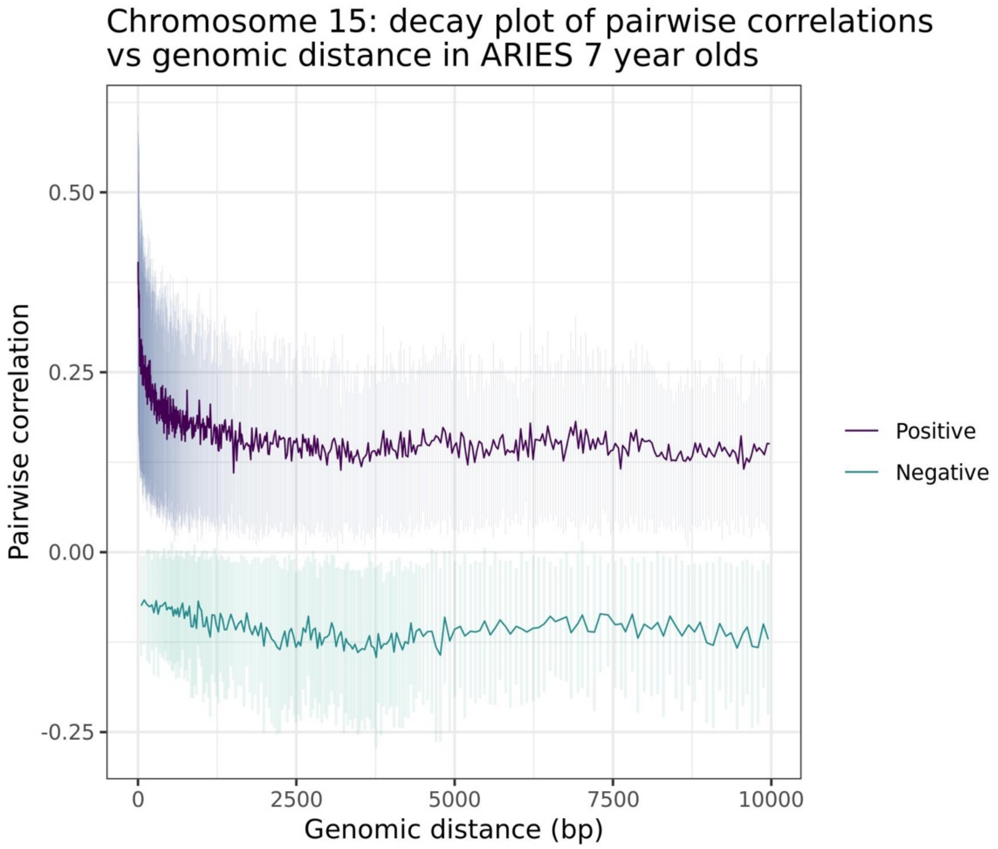
Chromosome 4: decay plot of pairwise correlations vs genomic distance in ARIES 7 year olds



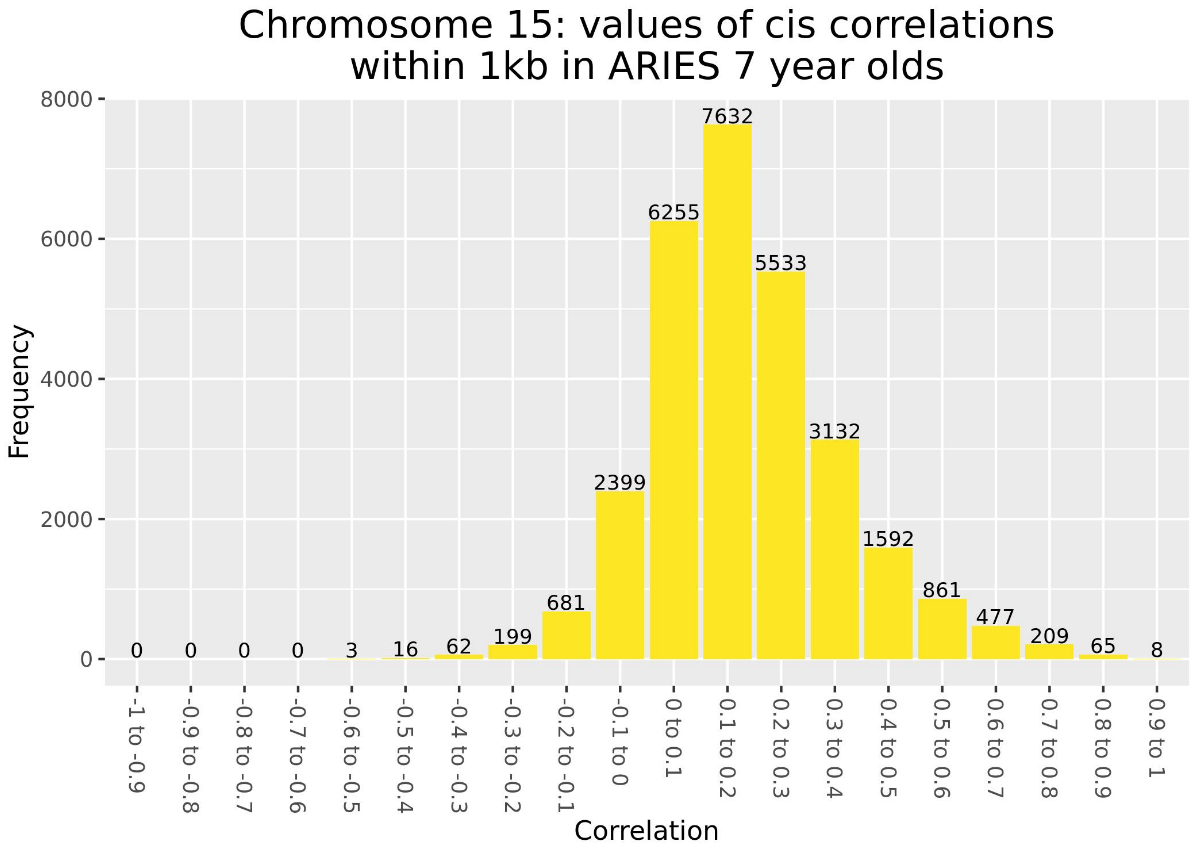
Chromosome 4: values of cis correlations within 1kb in ARIES 7 year olds


Chromosome 5: decay plot of pairwise correlations
vs genomic distance in ARIES 7 year olds
Pairwise correlation
0.50
0.25
0.00
-0.25
0
2500
5000
7500
10000
Genomic distance (bp)
Positive
Negative

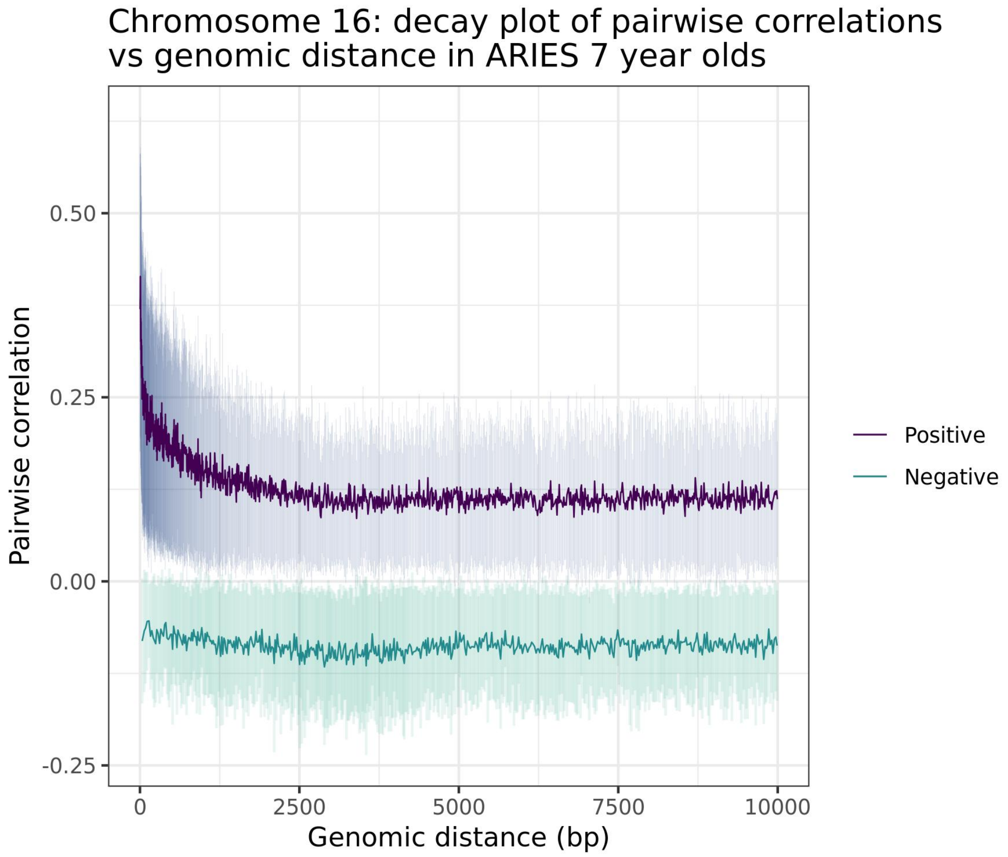

Chromosome 5: values of cis correlations
within 1kb in ARIES 7 year olds
Frequency
10000
5000
0
0
0
0
4
13
46
128
359
1101
3552
10600
14180
10262
6067
3128
1533
794
388
150
48
-1 to -0.9
-0.9 to -0.8
-0.8 to -0.7
-0.7 to -0.6
-0.6 to -0.5
-0.5 to -0.4
-0.4 to -0.3
-0.3 to -0.2
-0.2 to -0.1
-0.1 to 0
0 to 0.1
0.1 to 0.2
0.2 to 0.3
0.3 to 0.4
0.4 to 0.5
0.5 to 0.6
0.6 to 0.7
0.7 to 0.8
0.8 to 0.9
0.9 to 1
Correlation

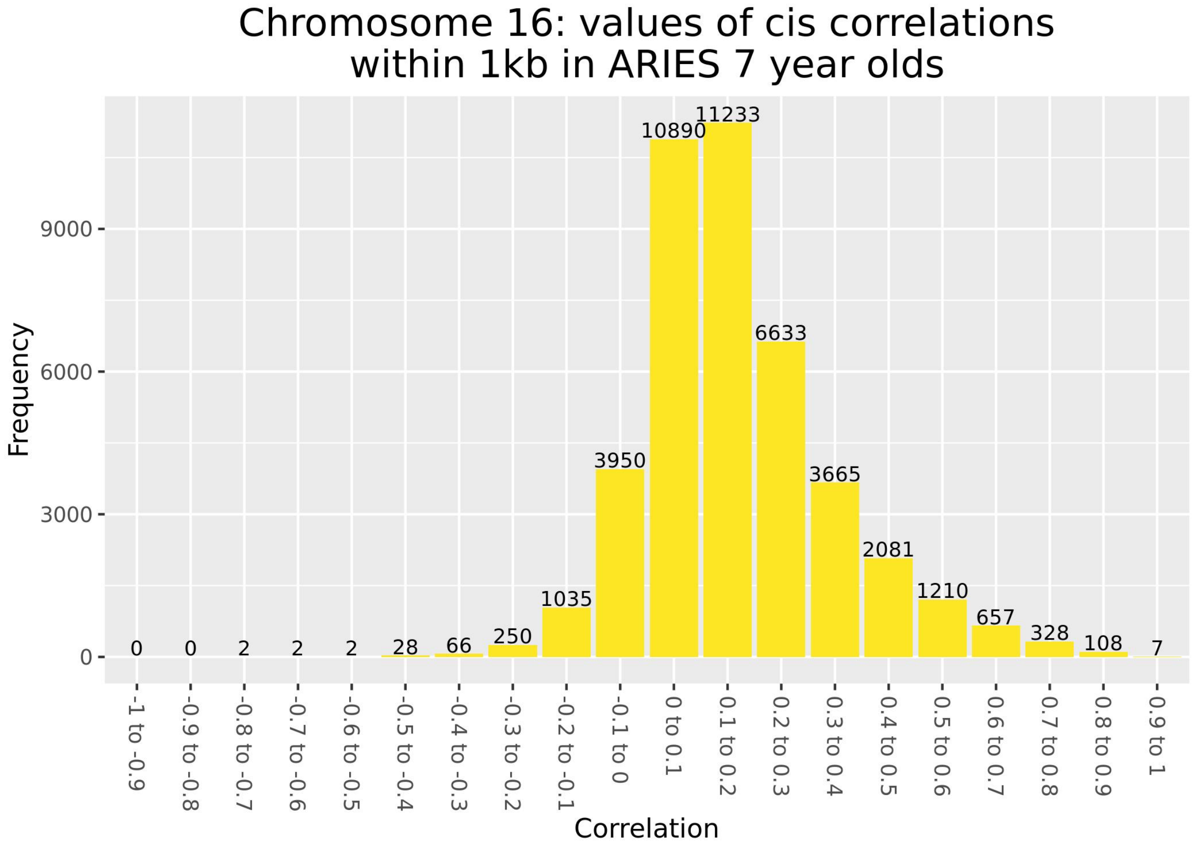


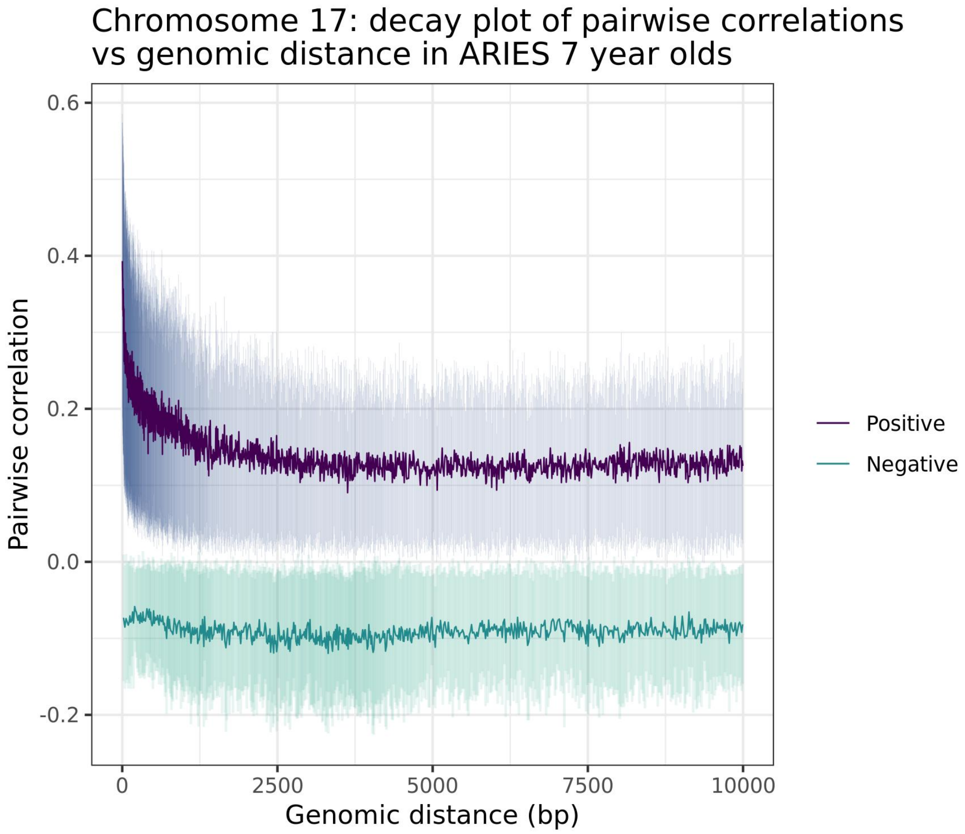
Chromosome 15: decay plot of pairwise correlations vs genomic distance in ARIES 7 year olds

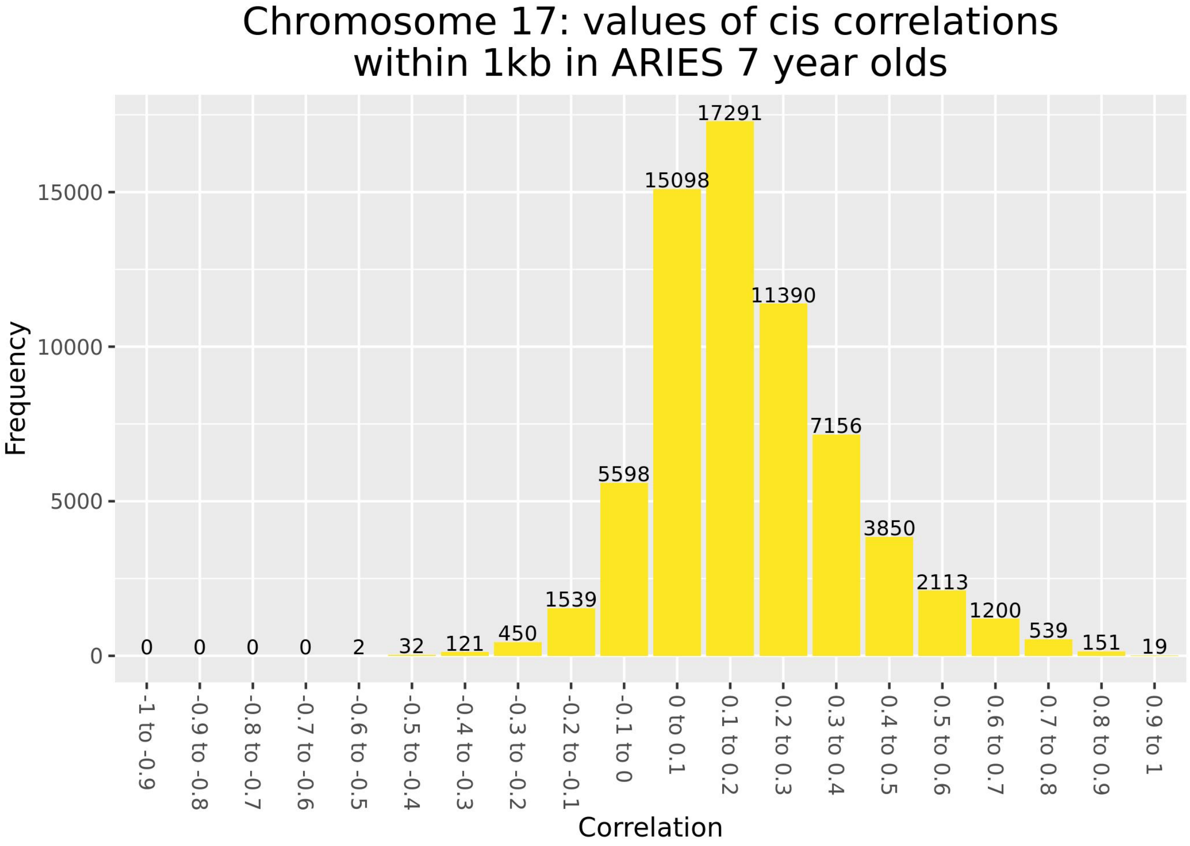


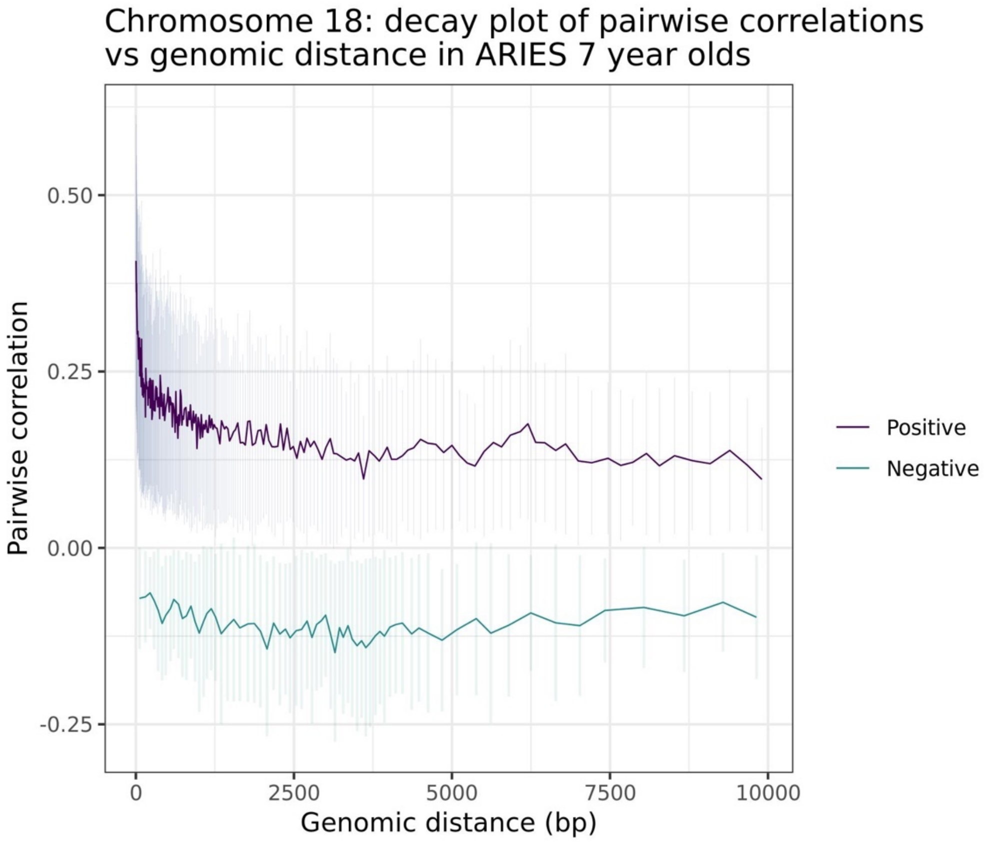
Chromosome 15: values of cis correlations within 1kb in ARIES 7 year olds

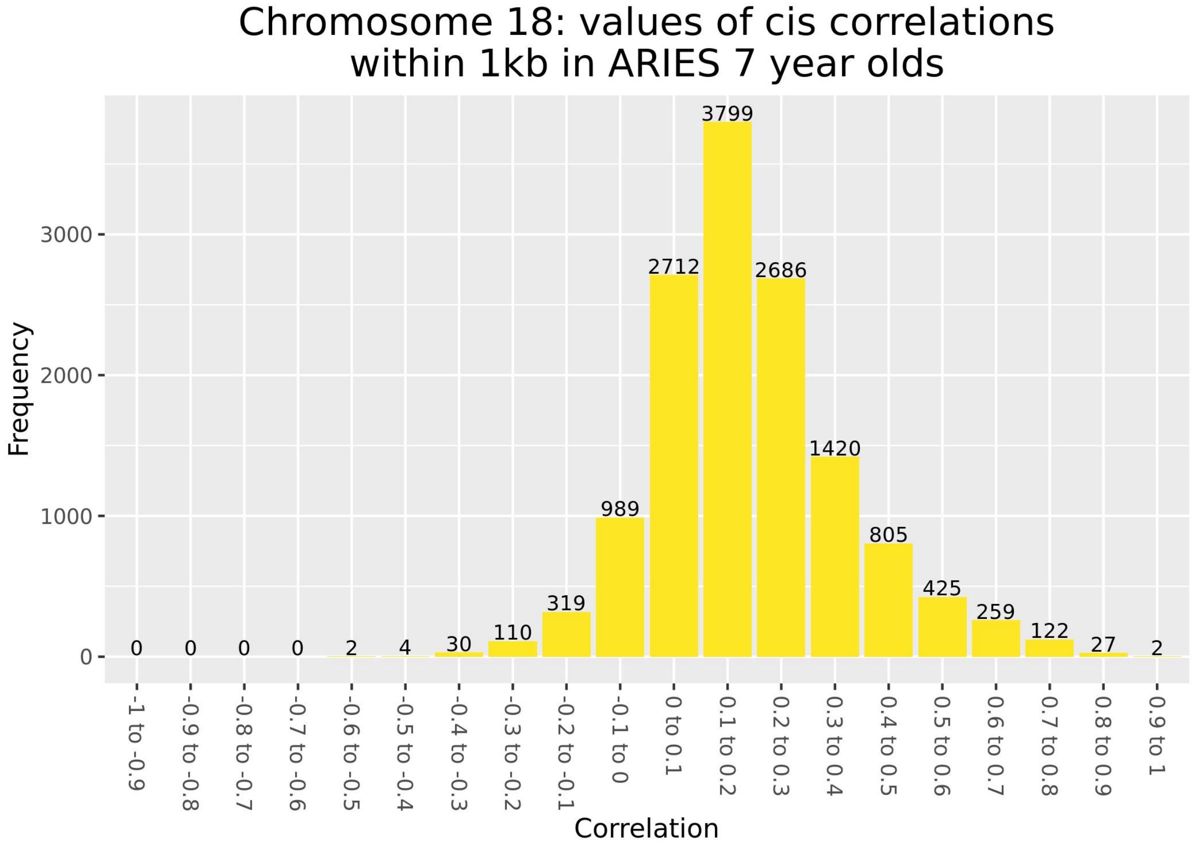

Chromosome 16: decay plot of pairwise correlations
vs genomic distance in ARIES 7 year olds
Pairwise correlation
0.50
0.25
0.00
-0.25
0
2500
5000
7500
10000
Genomic distance (bp)
Positive
Negative

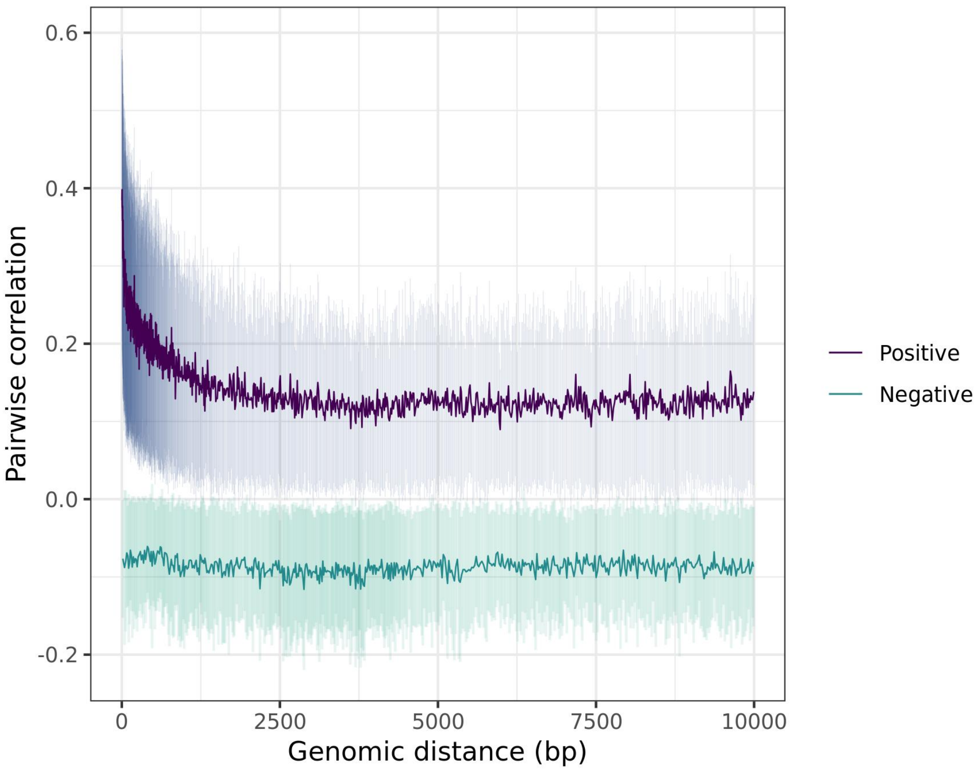

Chromosome 16: values of cis correlations
within 1kb in ARIES 7 year olds
Frequency
9000
6000
3000
0
0
0
2
2
2
28
66
250
1035
3950
10890
11233
6633
3665
2081
1210
657
328
108
7
-1 to -0.9
-0.9 to -0.8
-0.8 to -0.7
-0.7 to -0.6
-0.6 to -0.5
-0.5 to -0.4
-0.4 to -0.3
-0.3 to -0.2
-0.2 to -0.1
-0.1 to 0
0 to 0.1
0.1 to 0.2
0.2 to 0.3
0.3 to 0.4
0.4 to 0.5
0.5 to 0.6
0.6 to 0.7
0.7 to 0.8
0.8 to 0.9
0.9 to 1
Correlation


Chromosome 17: decay plot of pairwise correlations
vs genomic distance in ARIES 7 year olds
Pairwise correlation
0.6
0.4
0.2
0.0
-0.2
0
2500
5000
7500
10000
Genomic distance (bp)
Positive
Negative

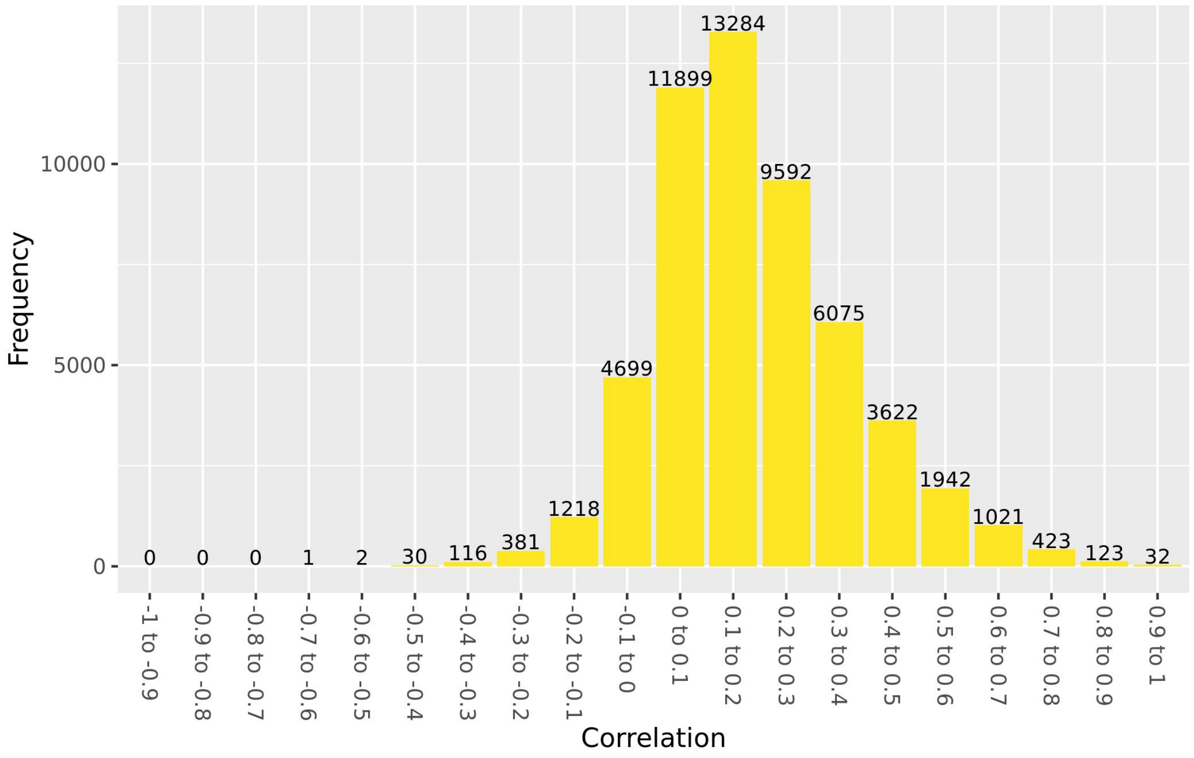

Chromosome 17: values of cis correlations
within 1kb in ARIES 7 year olds
Frequency
15000
10000
5000
0
0
0
0
0
2
32
121
450
1539
5598
15098
17291
11390
7156
3850
2113
1200
539
151
19
-1 to -0.9
-0.9 to -0.8
-0.8 to -0.7
-0.7 to -0.6
-0.6 to -0.5
-0.5 to -0.4
-0.4 to -0.3
-0.3 to -0.2
-0.2 to -0.1
-0.1 to 0
0 to 0.1
0.1 to 0.2
0.2 to 0.3
0.3 to 0.4
0.4 to 0.5
0.5 to 0.6
0.6 to 0.7
0.7 to 0.8
0.8 to 0.9
0.9 to 1
Correlation


Chromosome 18: decay plot of pairwise correlations
vs genomic distance in ARIES 7 year olds
Pairwise correlation
0.50
0.25
0.00
-0.25
0
2500
5000
7500
10000
Genomic distance (bp)
Positive
Negative


Chromosome 18: values of cis correlations
within 1kb in ARIES 7 year olds
Frequency
3000
2000
1000
0
0
0
0
0
2
4
30
110
319
989
2712
3799
2686
1420
805
425
259
122
27
2
-1 to -0.9
-0.9 to -0.8
-0.8 to -0.7
-0.7 to -0.6
-0.6 to -0.5
-0.5 to -0.4
-0.4 to -0.3
-0.3 to -0.2
-0.2 to -0.1
-0.1 to 0
0 to 0.1
0.1 to 0.2
0.2 to 0.3
0.3 to 0.4
0.4 to 0.5
0.5 to 0.6
0.6 to 0.7
0.7 to 0.8
0.8 to 0.9
0.9 to 1
Correlation

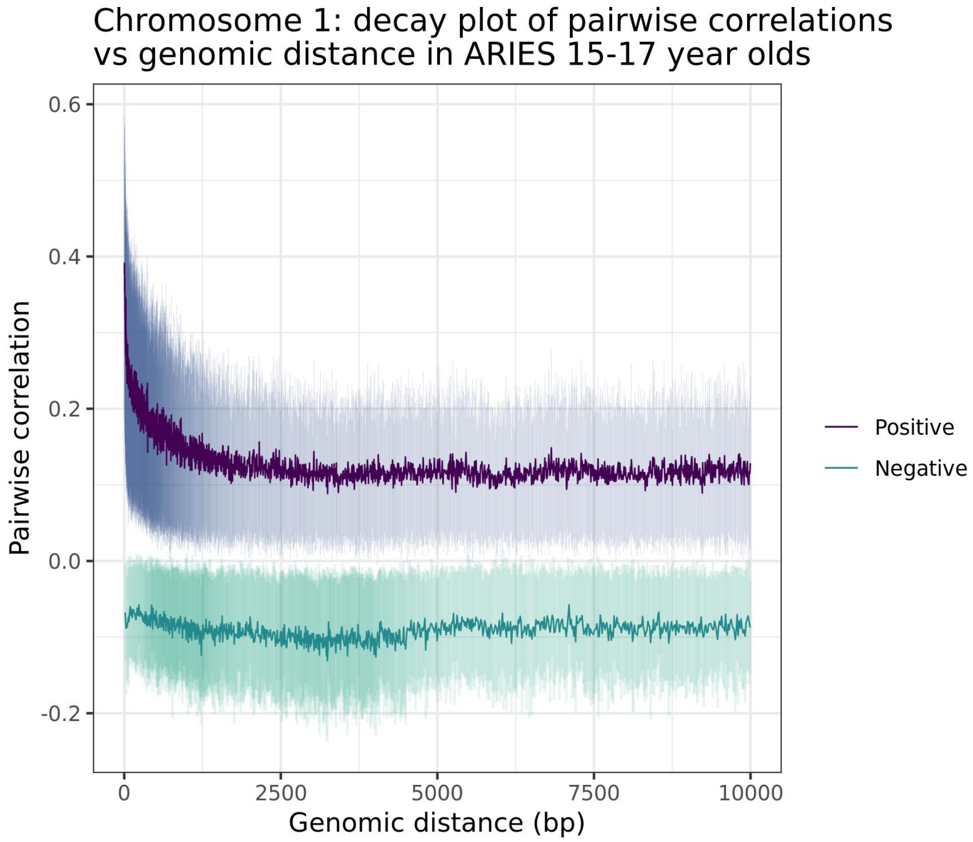


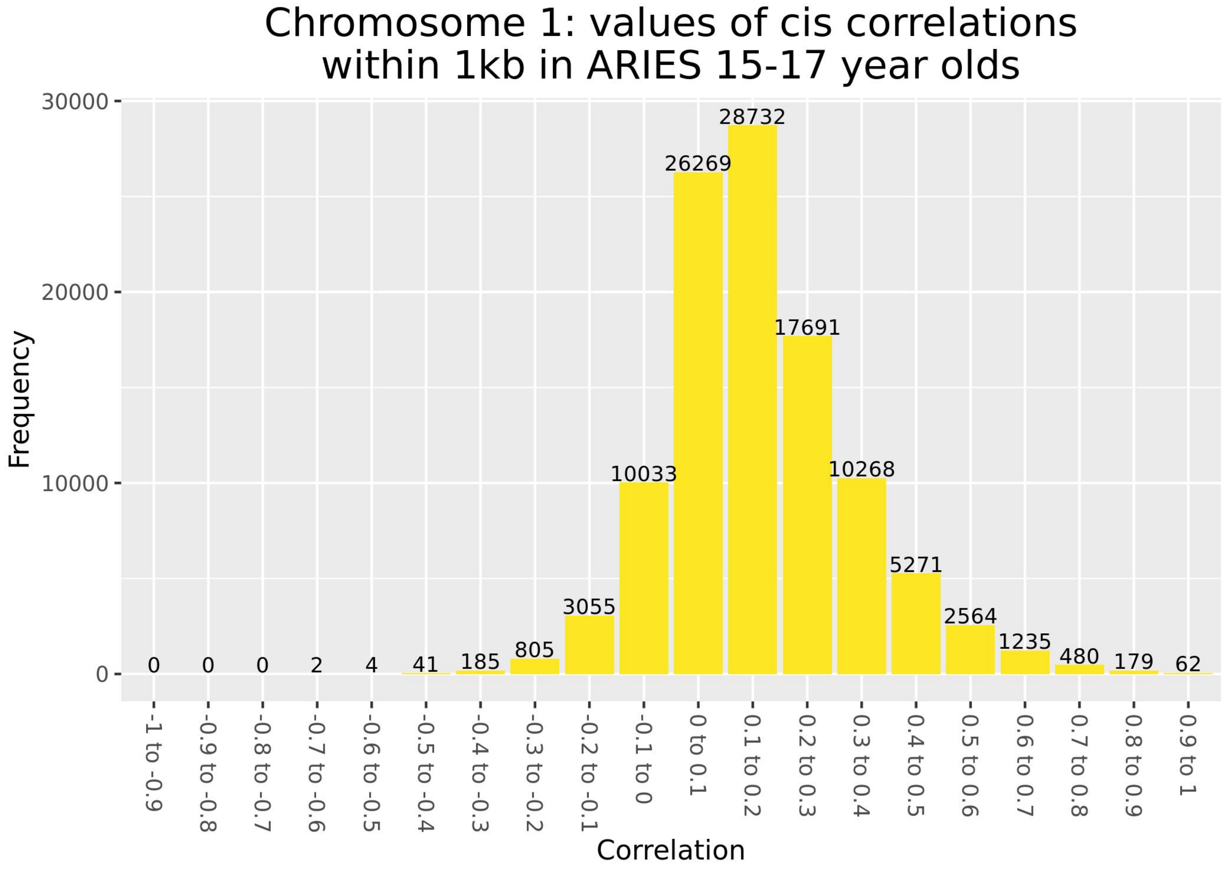
Chromosome 19: decay plot of pairwise correlations vs genomic distance in ARIES 7 year olds

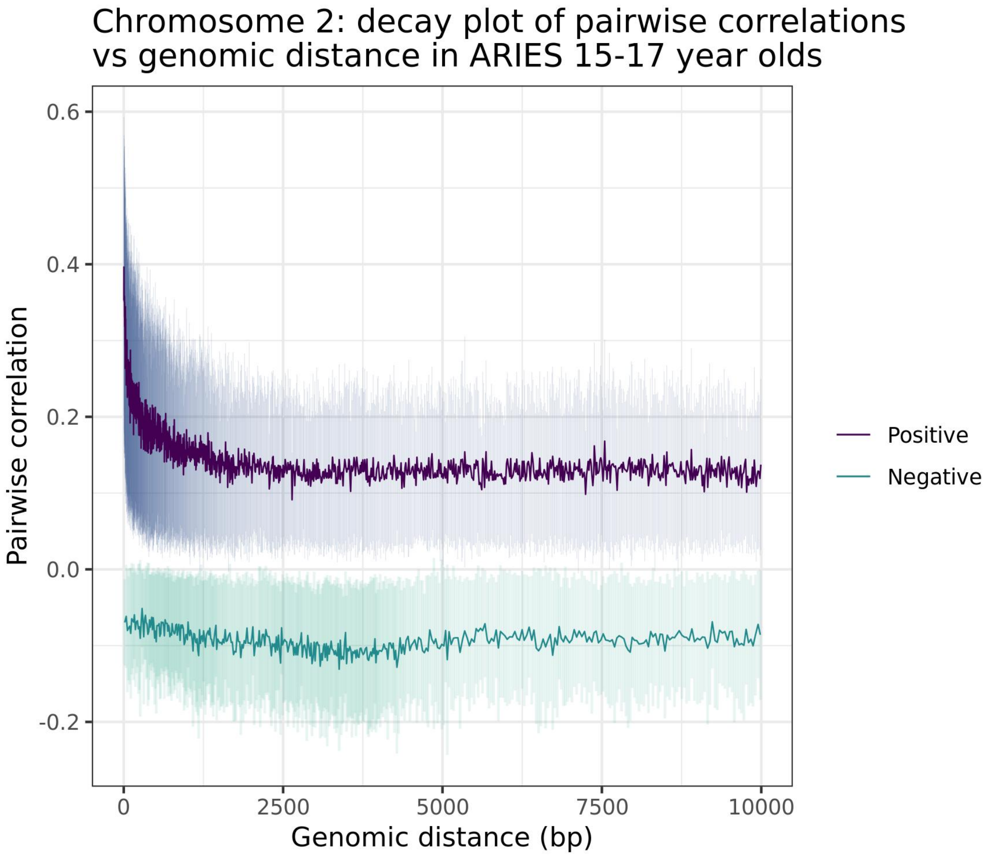


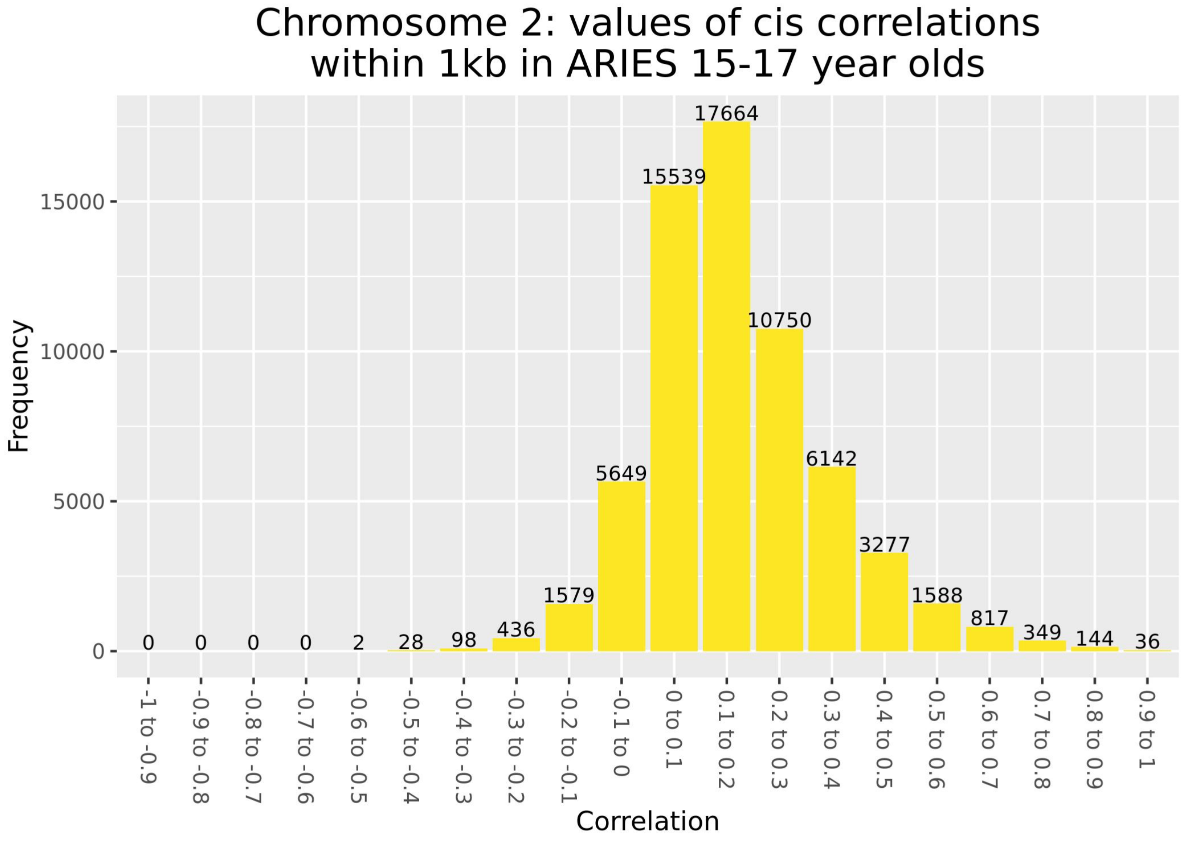
Chromosome 19: values of cis correlations within 1kb in ARIES 7 year olds

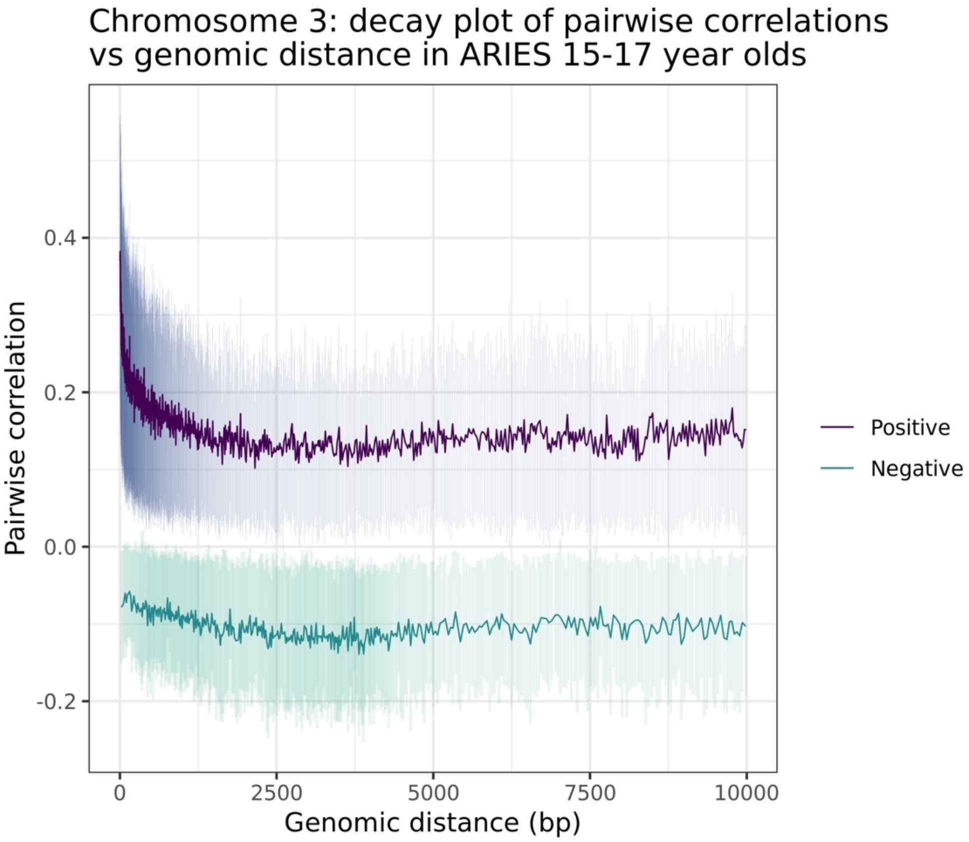
15-17 years

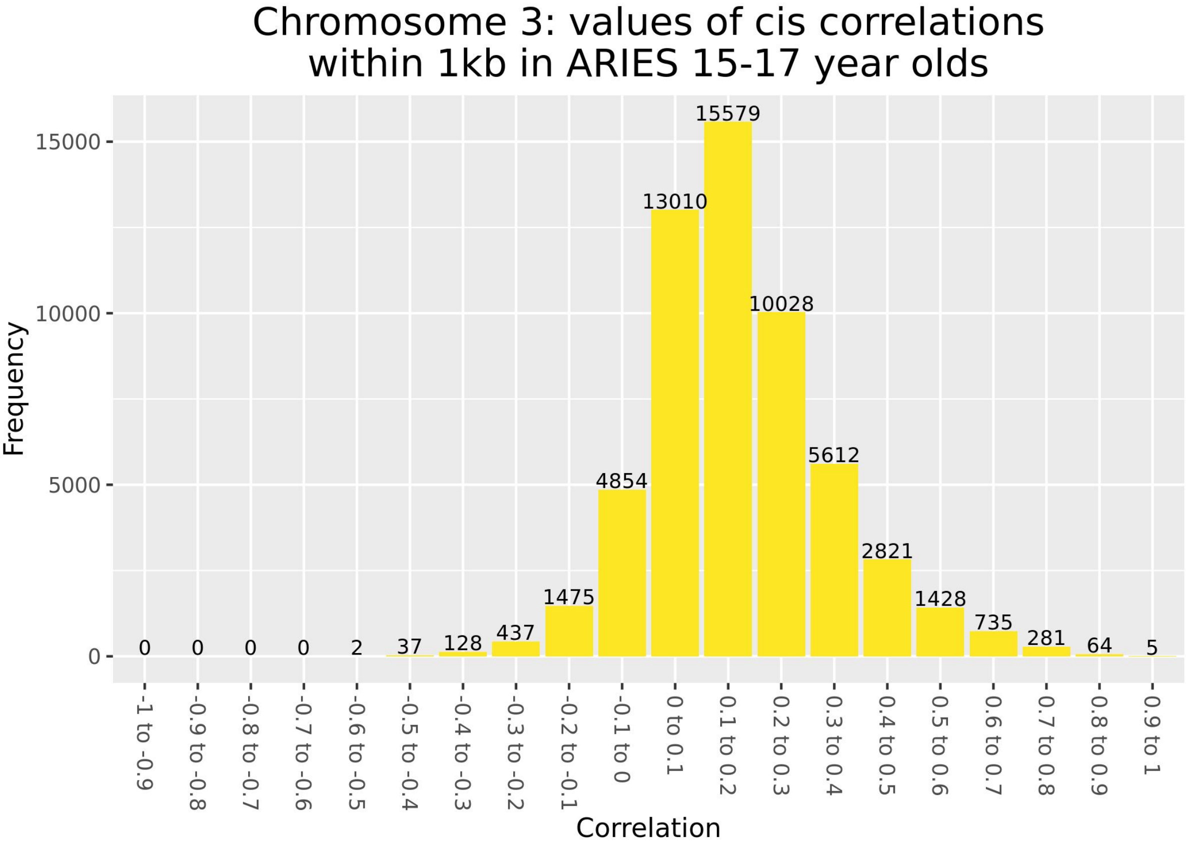

Chromosome 1: decay plot of pairwise correlations
vs genomic distance in ARIES 15-17 year olds
Pairwise correlation
0.6
0.4
0.2
0.0
-0.2
0
2500
5000
7500
10000
Genomic distance (bp)
Positive
Negative

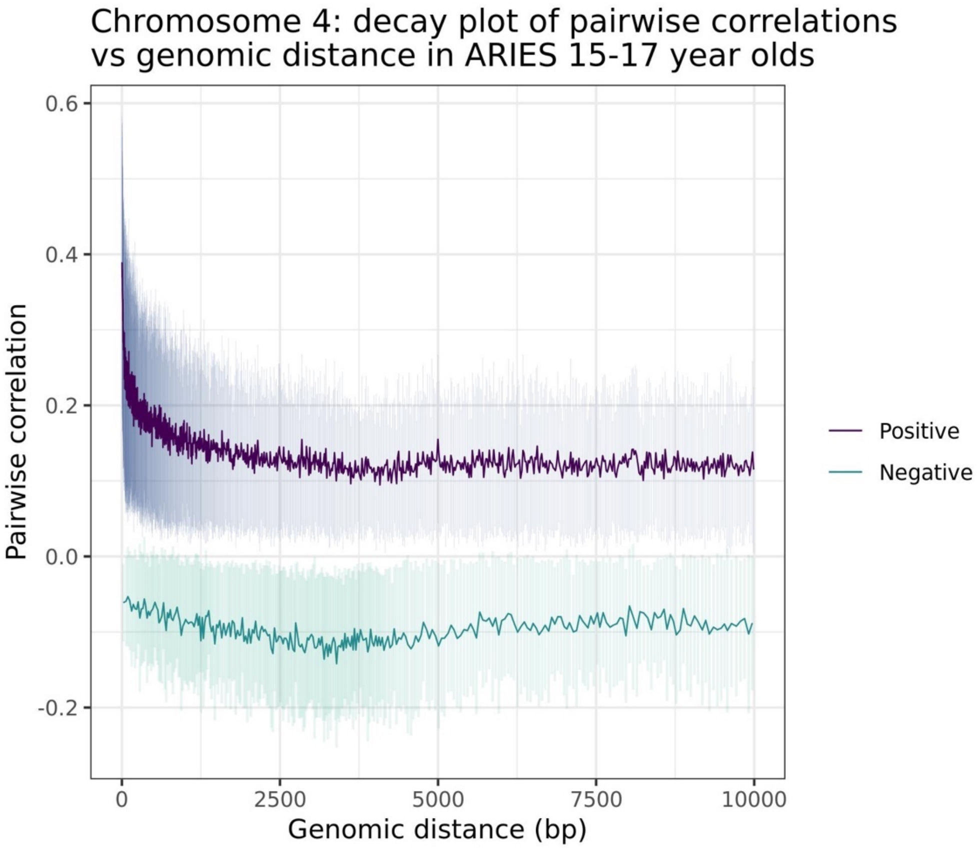

Chromosome 1: values of cis correlations
within 1kb in ARIES 15-17 year olds
Frequency
30000
20000
10000
0
0
0
0
2
4
41
185
805
3055
10033
26269
28732
17691
10268
5271
2564
1235
480
179
62
-1 to -0.9
-0.9 to -0.8
-0.8 to -0.7
-0.7 to -0.6
-0.6 to -0.5
-0.5 to -0.4
-0.4 to -0.3
-0.3 to -0.2
-0.2 to -0.1
-0.1 to 0
0 to 0.1
0.1 to 0.2
0.2 to 0.3
0.3 to 0.4
0.4 to 0.5
0.5 to 0.6
0.6 to 0.7
0.7 to 0.8
0.8 to 0.9
0.9 to 1
Correlation

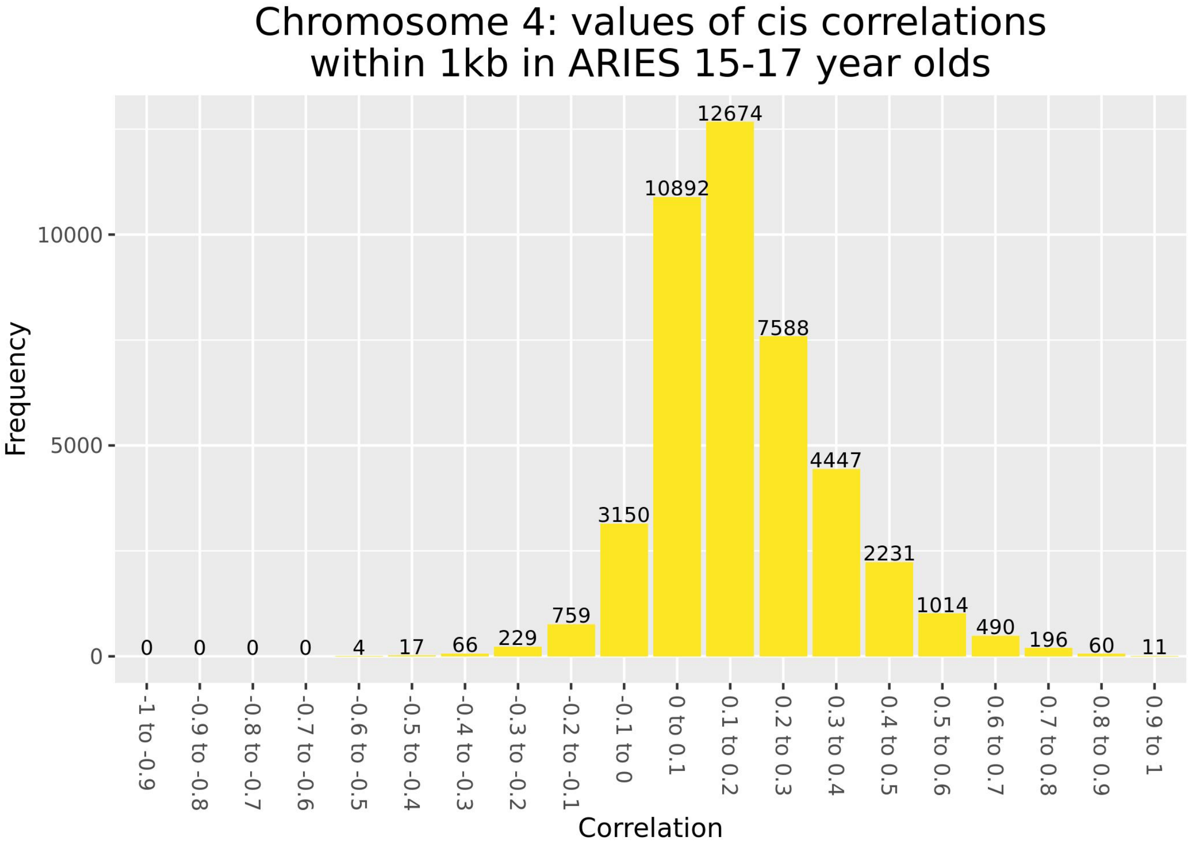

Chromosome 2: decay plot of pairwise correlations
vs genomic distance in ARIES 15-17 year olds
Pairwise correlation
0.6
0.4
0.2
0.0
-0.2
0
2500
5000
7500
10000
Genomic distance (bp)
Positive
Negative

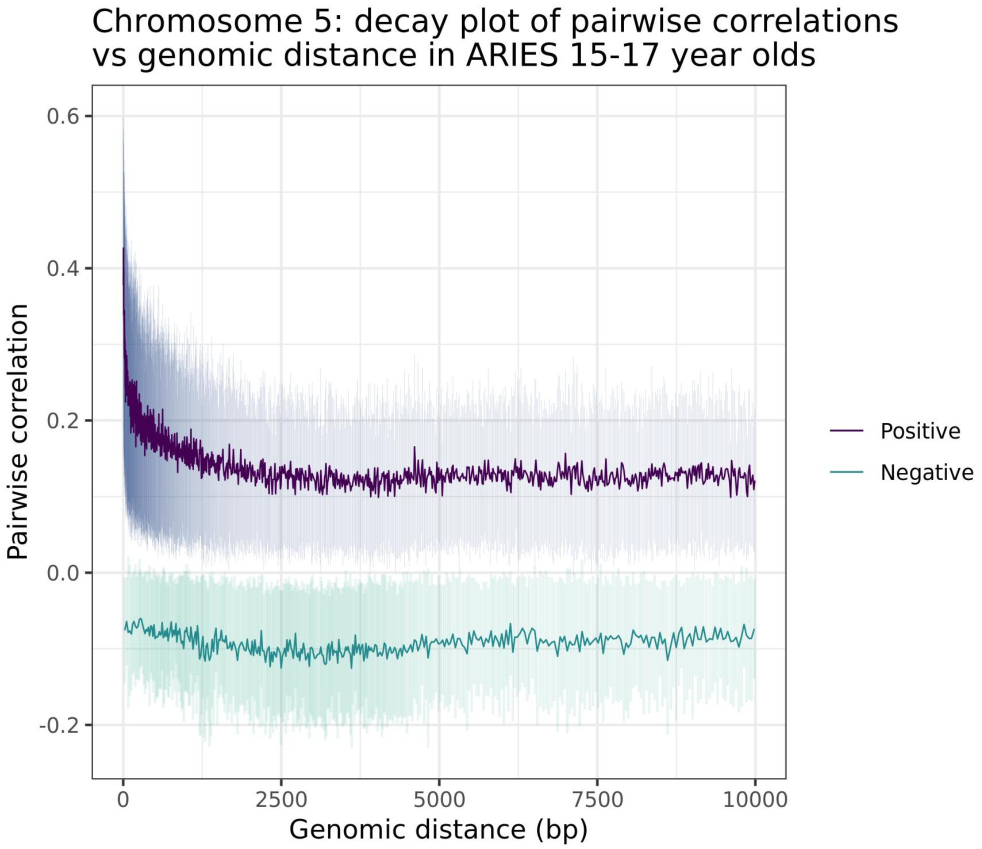

Chromosome 2: values of cis correlations
within 1kb in ARIES 15-17 year olds
Frequency
15000
10000
5000
0
0
0
0
0
2
28
98
436
1579
5649
15539
17664
10750
6142
3277
1588
817
349
144
36
-1 to -0.9
-0.9 to -0.8
-0.8 to -0.7
-0.7 to -0.6
-0.6 to -0.5
-0.5 to -0.4
-0.4 to -0.3
-0.3 to -0.2
-0.2 to -0.1
-0.1 to 0
0 to 0.1
0.1 to 0.2
0.2 to 0.3
0.3 to 0.4
0.4 to 0.5
0.5 to 0.6
0.6 to 0.7
0.7 to 0.8
0.8 to 0.9
0.9 to 1
Correlation

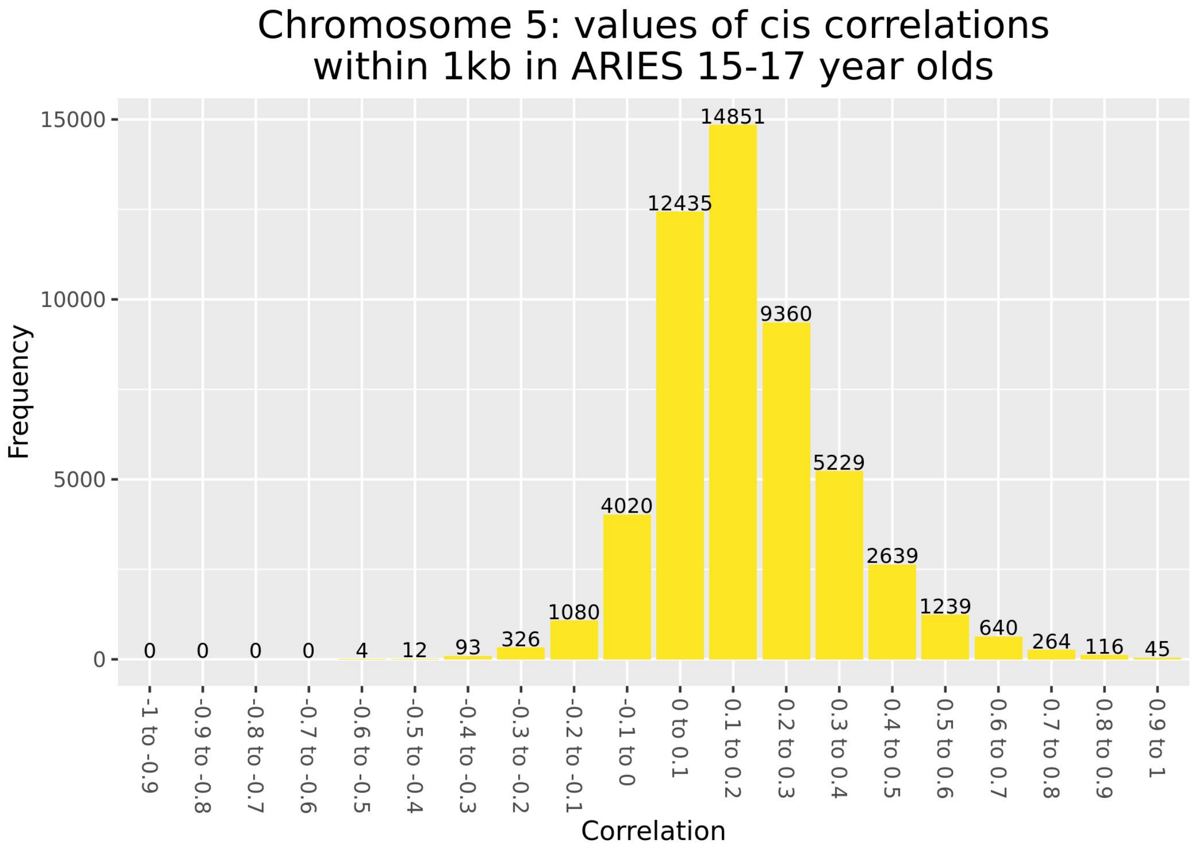

Chromosome 3: decay plot of pairwise correlations
vs genomic distance in ARIES 15-17 year olds
Pairwise correlation
0.4
0.2
0.0
-0.2
0
2500
5000
7500
10000
Genomic distance (bp)
Positive
Negative

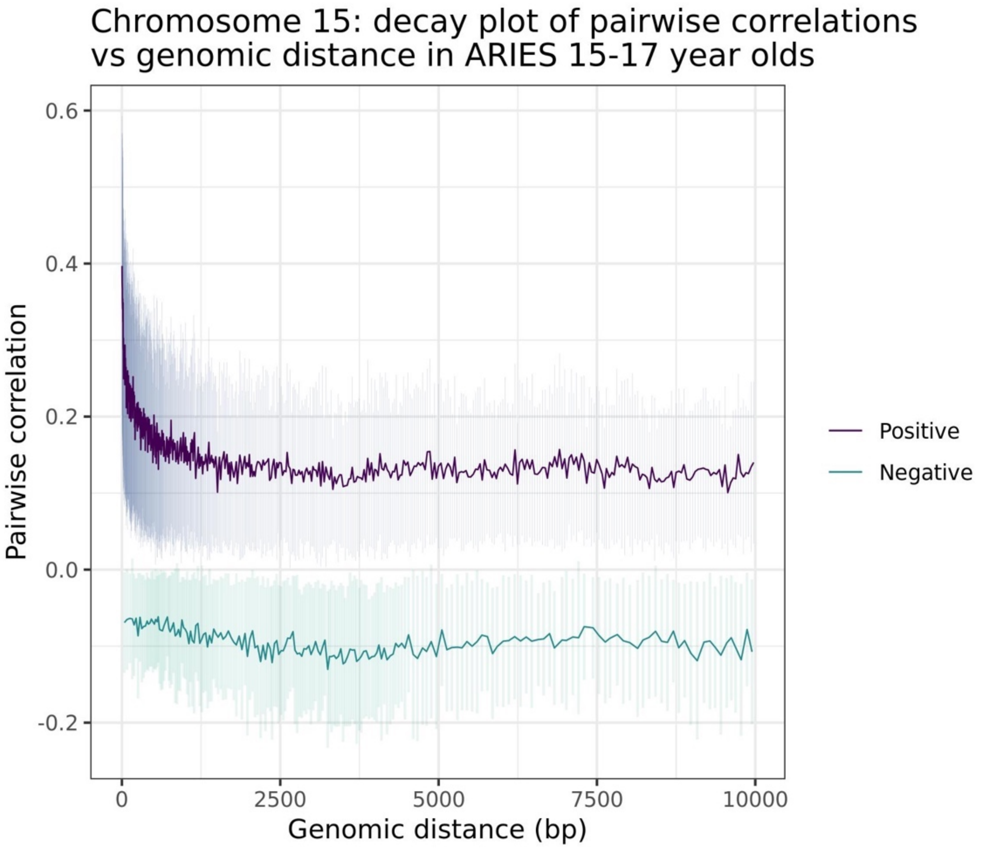

Chromosome 3: values of cis correlations
within 1kb in ARIES 15-17 year olds
Frequency
15000
10000
5000
0
0
0
0
0
2
37
128
437
1475
4854
13010
15579
10028
5612
2821
1428
735
281
64
5
-1 to -0.9
-0.9 to -0.8
-0.8 to -0.7
-0.7 to -0.6
-0.6 to -0.5
-0.5 to -0.4
-0.4 to -0.3
-0.3 to -0.2
-0.2 to -0.1
-0.1 to 0
0 to 0.1
0.1 to 0.2
0.2 to 0.3
0.3 to 0.4
0.4 to 0.5
0.5 to 0.6
0.6 to 0.7
0.7 to 0.8
0.8 to 0.9
0.9 to 1
Correlation


Chromosome 4: decay plot of pairwise correlations vs genomic distance in ARIES 15-17 year olds

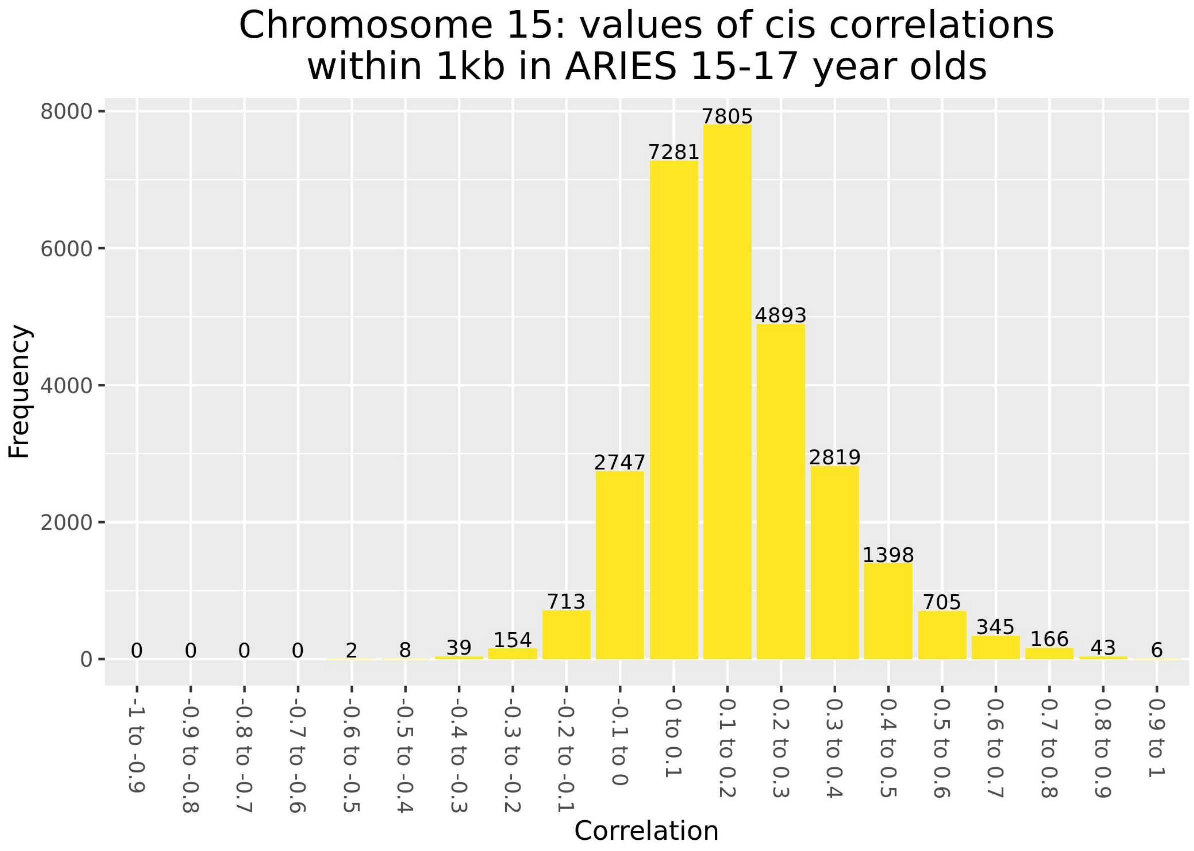

Chromosome 4: values of cis correlations within 1kb in ARIES 15-17 year olds


Chromosome 5: decay plot of pairwise correlations vs genomic distance in ARIES 15-17 year olds

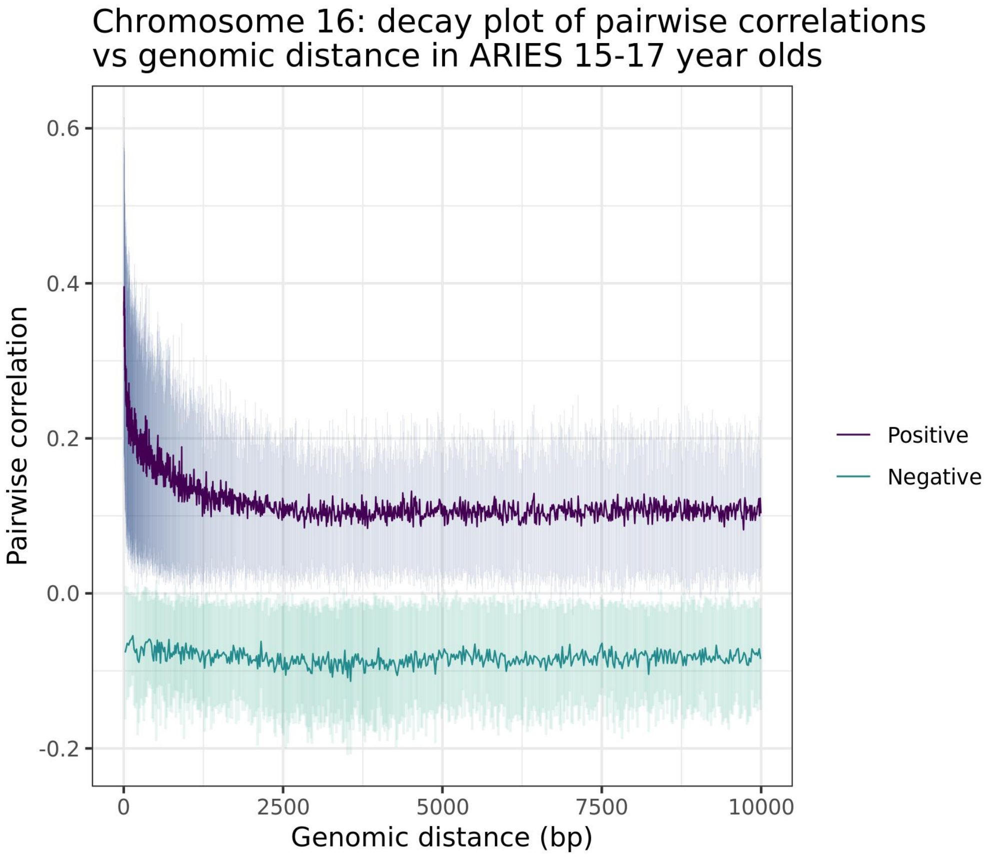

Chromosome 5: values of cis correlations within 1kb in ARIES 15-17 year olds

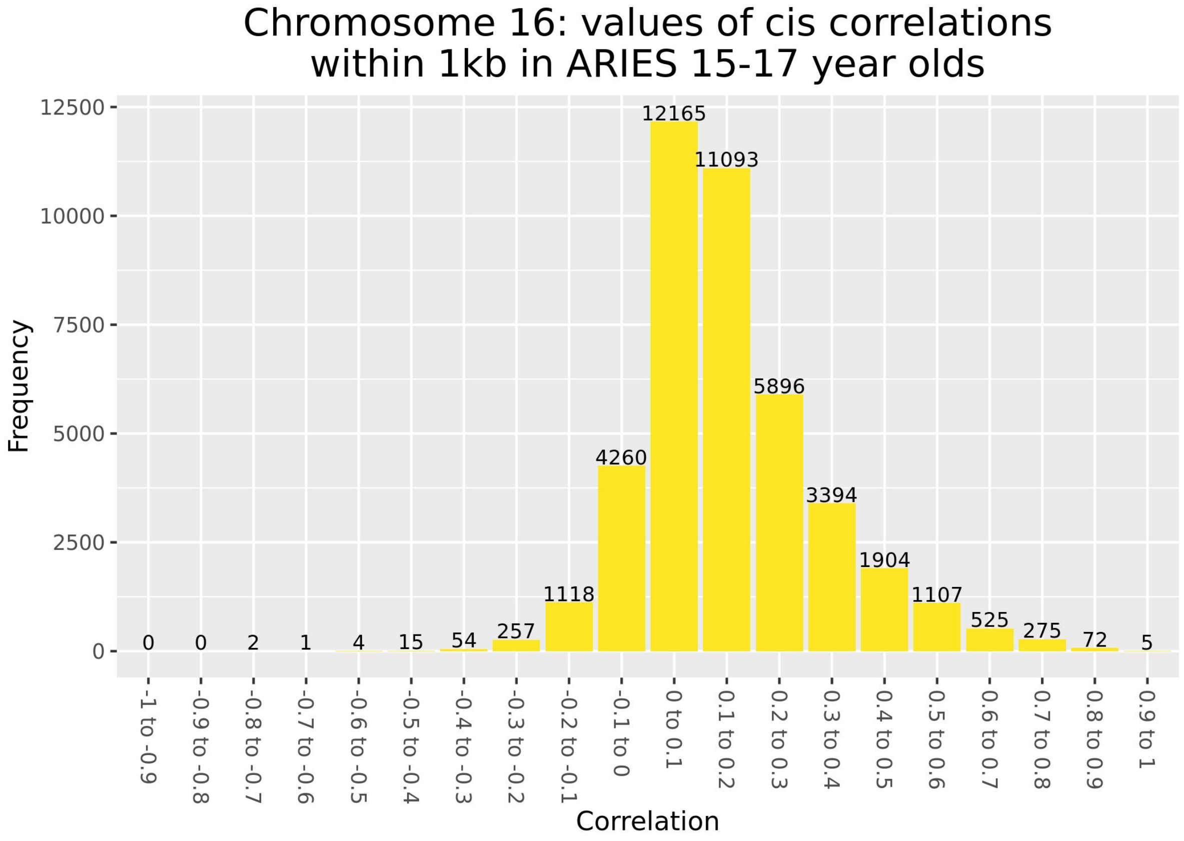


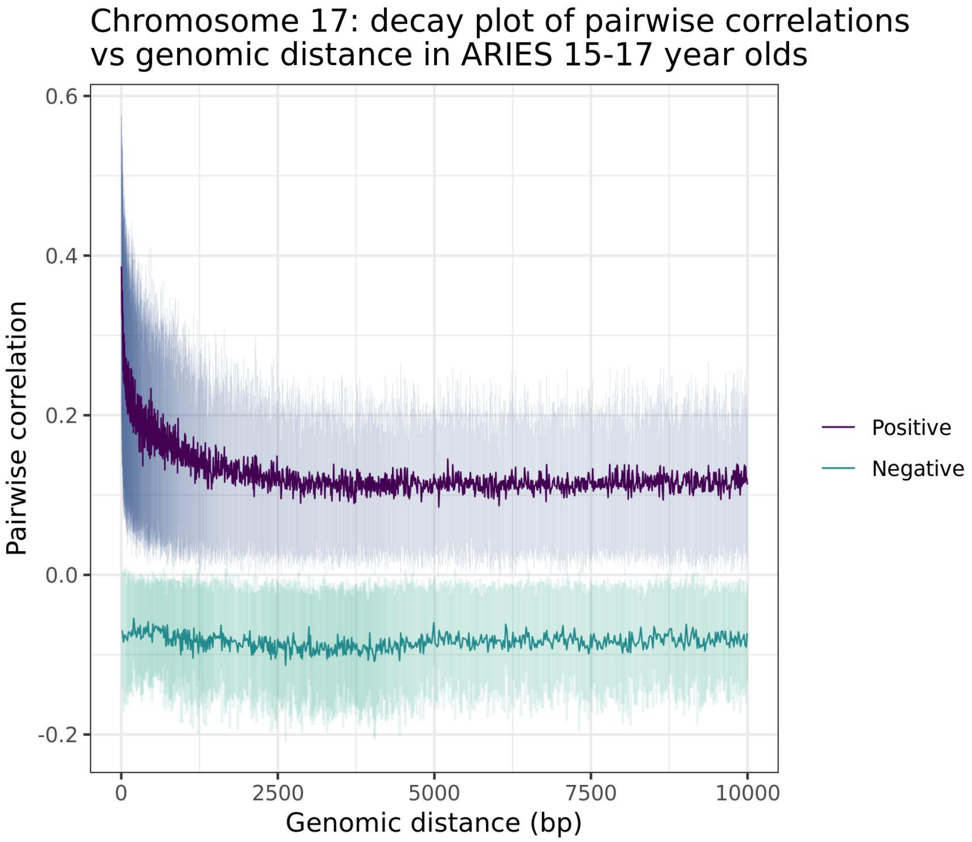
Chromosome 15: decay plot of pairwise correlations vs genomic distance in ARIES 15-17 year olds

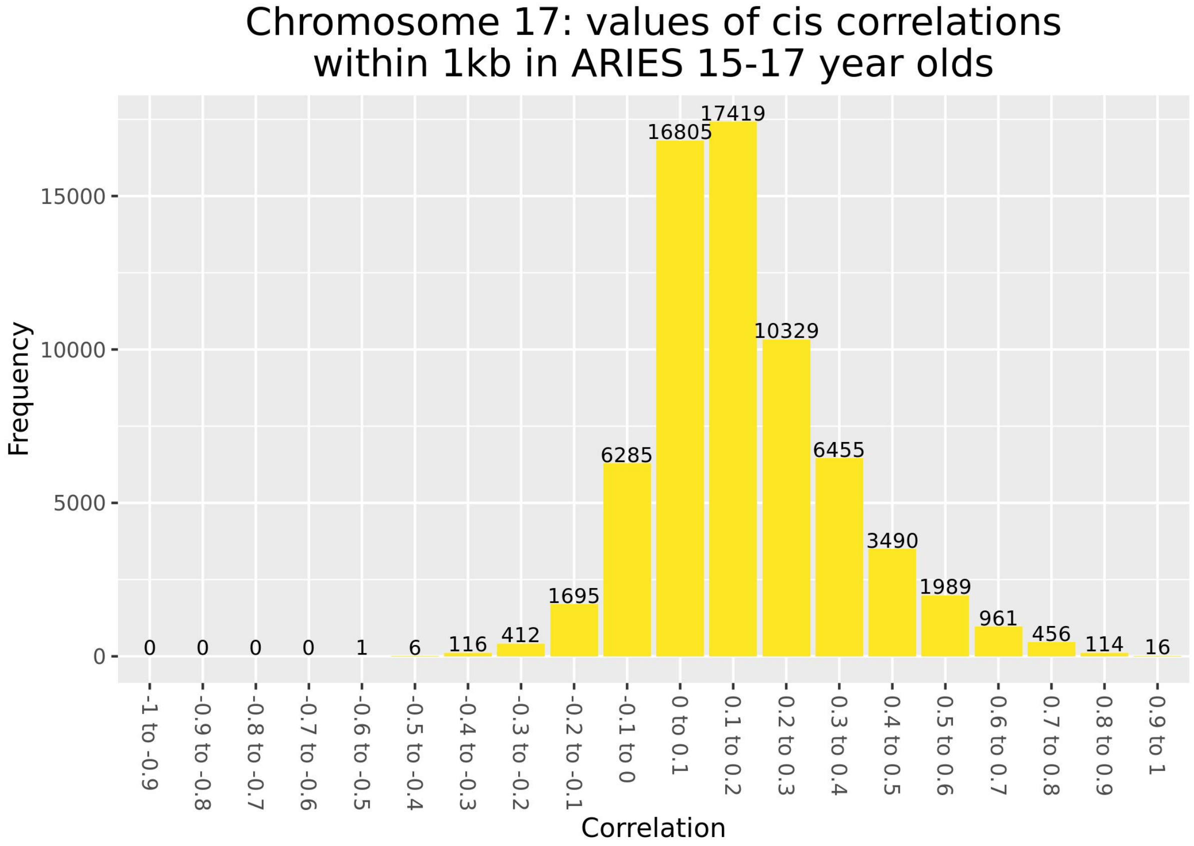


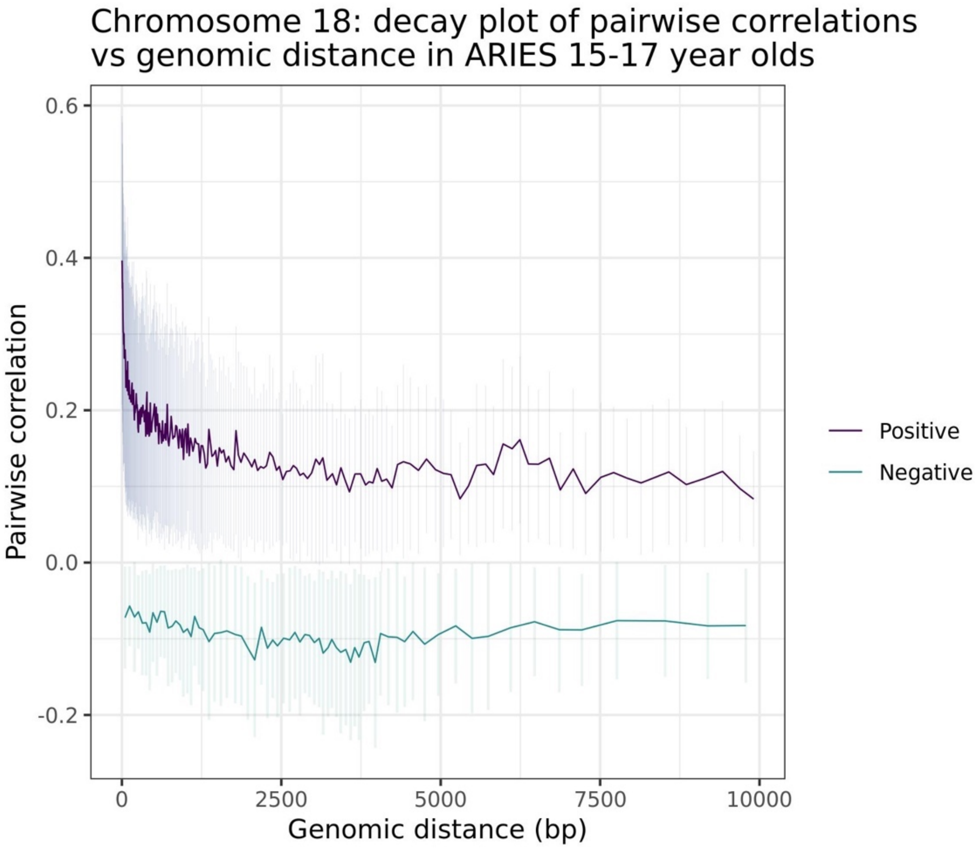
Chromosome 15: values of cis correlations within 1kb in ARIES 15-17 year olds

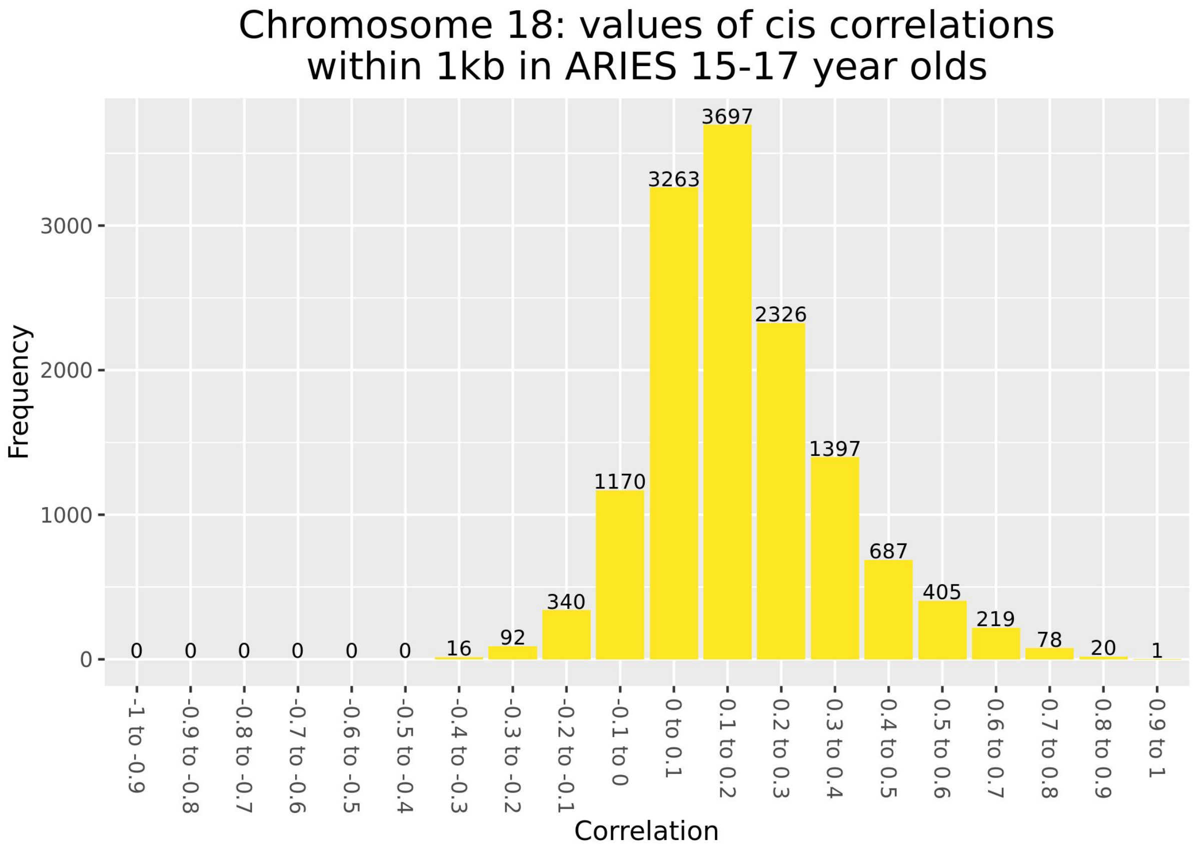

Chromosome 16: decay plot of pairwise correlations
vs genomic distance in ARIES 15-17 year olds
Pairwise correlation
0.6
0.4
0.2
0.0
-0.2
0
2500
5000
7500
10000
Genomic distance (bp)
Positive
Negative

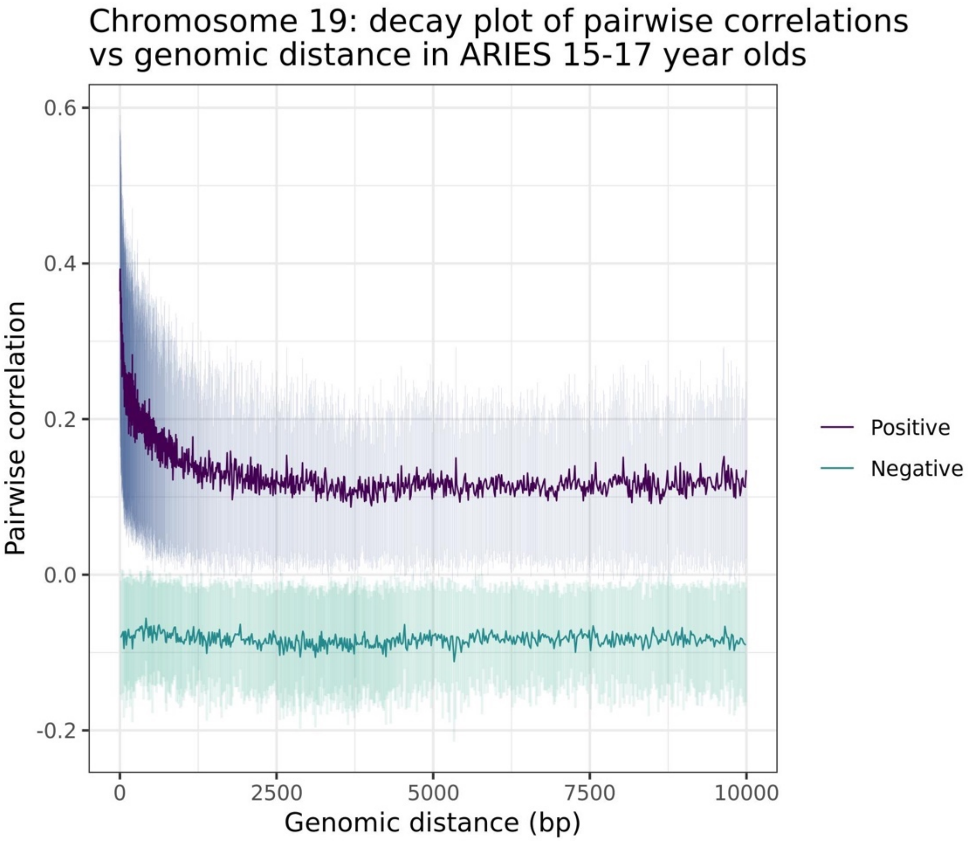

Chromosome 16: values of cis correlations
within 1kb in ARIES 15-17 year olds
Frequency
12500
10000
7500
5000
2500
0
0
0
2
1
4
15
54
257
1118
4260
12165
11093
5896
3394
1904
1107
525
275
72
5
-1 to -0.9
-0.9 to -0.8
-0.8 to -0.7
-0.7 to -0.6
-0.6 to -0.5
-0.5 to -0.4
-0.4 to -0.3
-0.3 to -0.2
-0.2 to -0.1
-0.1 to 0
0 to 0.1
0.1 to 0.2
0.2 to 0.3
0.3 to 0.4
0.4 to 0.5
0.5 to 0.6
0.6 to 0.7
0.7 to 0.8
0.8 to 0.9
0.9 to 1
Correlation

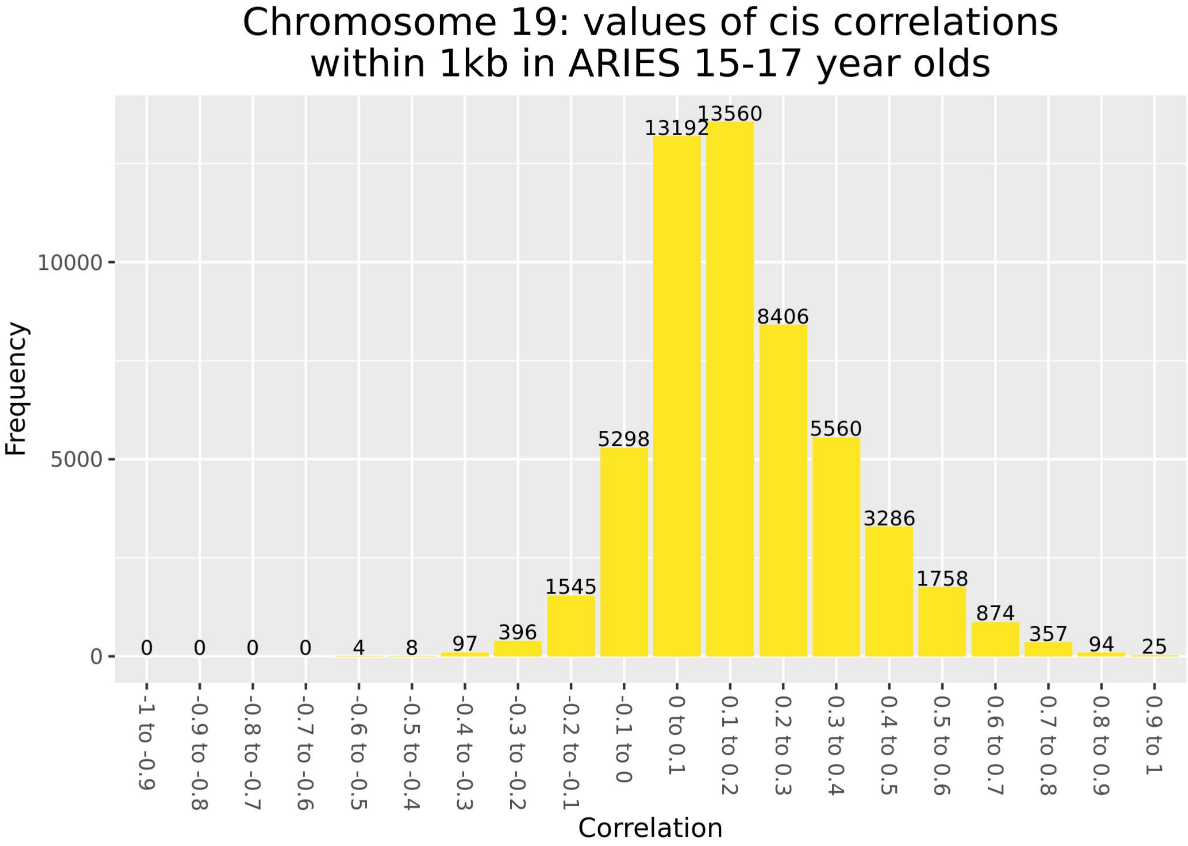

Chromosome 17: decay plot of pairwise correlations vs genomic distance in ARIES 15-17 year olds


Chromosome 17: values of cis correlations within 1kb in ARIES 15-17 year olds


Chromosome 18: decay plot of pairwise correlations
vs genomic distance in ARIES 15-17 year olds
Pairwise correlation
0.6
0.4
0.2
0.0
-0.2
0
2500
5000
7500
10000
Genomic distance (bp)
Positive
Negative

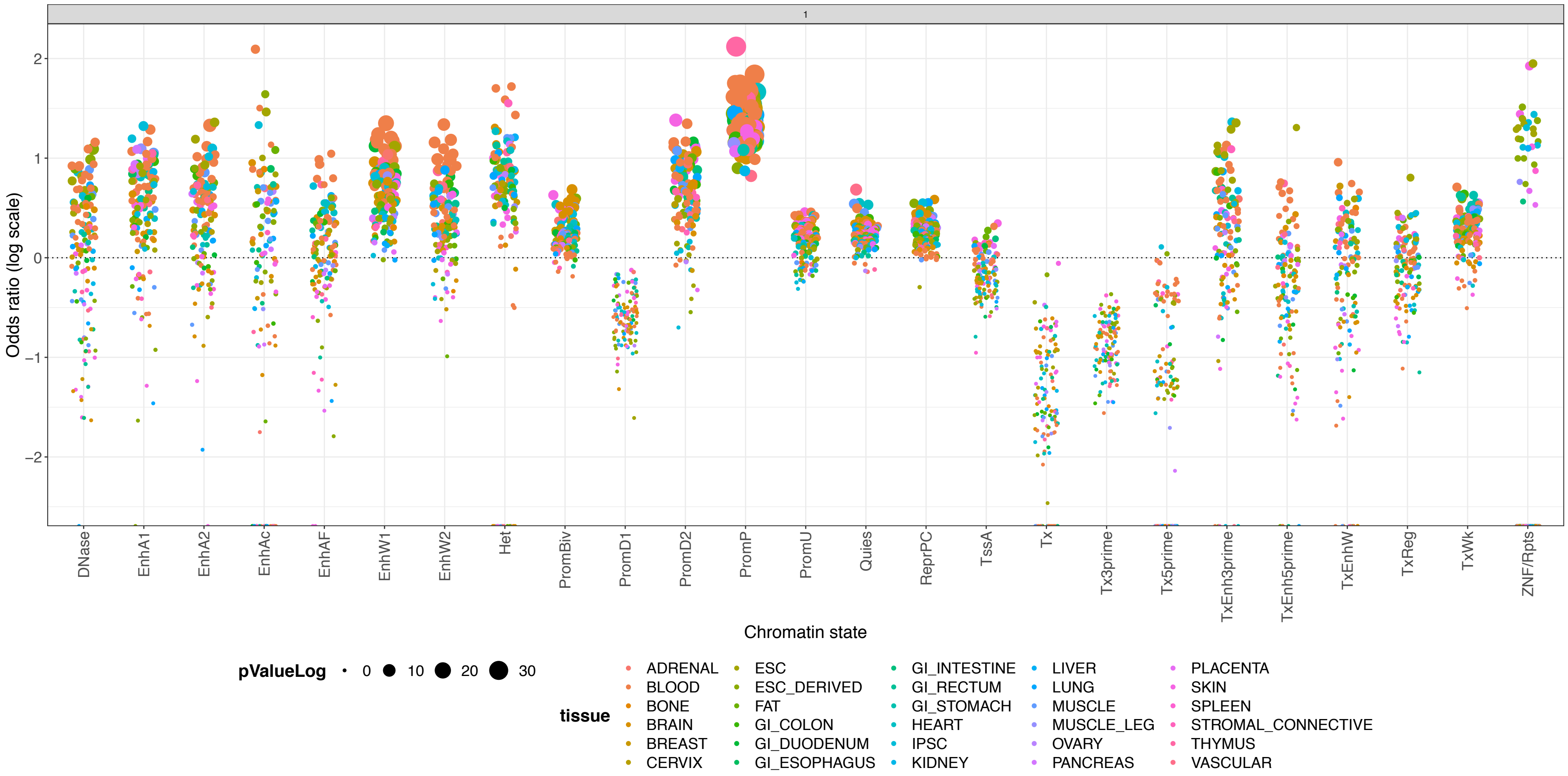

Chromosome 18: values of cis correlations
within 1kb in ARIES 15-17 year olds
Frequency
3000
2000
1000
0
0
0
0
0
0
0
16
92
340
1170
3263
3697
2326
1397
687
405
219
78
20
1
-1 to -0.9
-0.9 to -0.8
-0.8 to -0.7
-0.7 to -0.6
-0.6 to -0.5
-0.5 to -0.4
-0.4 to -0.3
-0.3 to -0.2
-0.2 to -0.1
-0.1 to 0
0 to 0.1
0.1 to 0.2
0.2 to 0.3
0.3 to 0.4
0.4 to 0.5
0.5 to 0.6
0.6 to 0.7
0.7 to 0.8
0.8 to 0.9
0.9 to 1
Correlation


Chromosome 19: decay plot of pairwise correlations
vs genomic distance in ARIES 15-17 year olds
Pairwise correlation
0.6
0.4
0.2
0.0
-0.2
0
2500
5000
7500
10000
Genomic distance (bp)
Positive
Negative

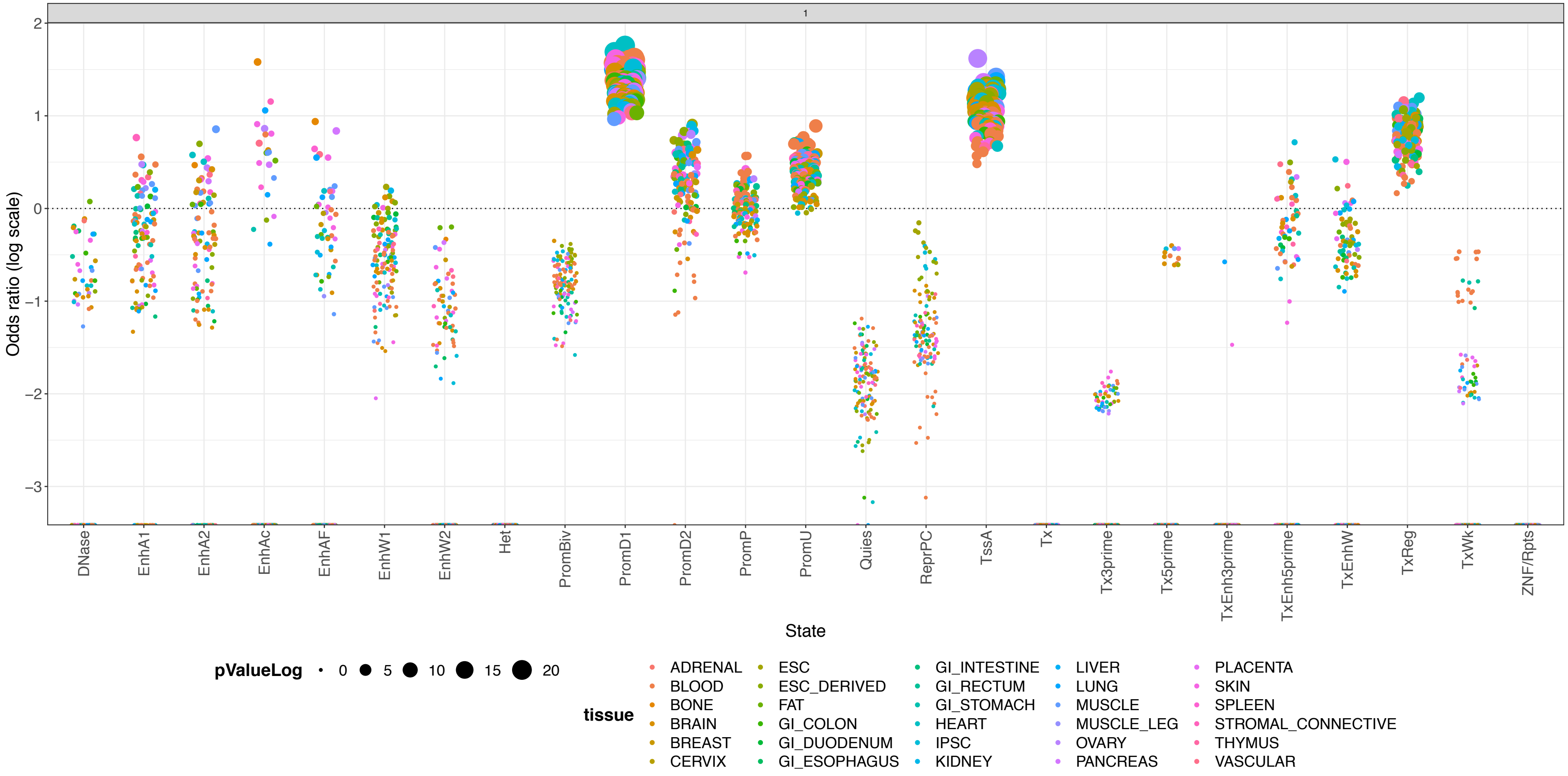

Chromosome 19: values of cis correlations
within 1kb in ARIES 15-17 year olds
Frequency
10000
5000
0
0
0
0
0
4
8
97
396
1545
5298
13192
13560
8406
5560
3286
1758
874
357
94
25
-1 to -0.9
-0.9 to -0.8
-0.8 to -0.7
-0.7 to -0.6
-0.6 to -0.5
-0.5 to -0.4
-0.4 to -0.3
-0.3 to -0.2
-0.2 to -0.1
-0.1 to 0
0 to 0.1
0.1 to 0.2
0.2 to 0.3
0.3 to 0.4
0.4 to 0.5
0.5 to 0.6
0.6 to 0.7
0.7 to 0.8
0.8 to 0.9
0.9 to 1
Correlation

Appendix 2: Genomic region enrichments in ARIES

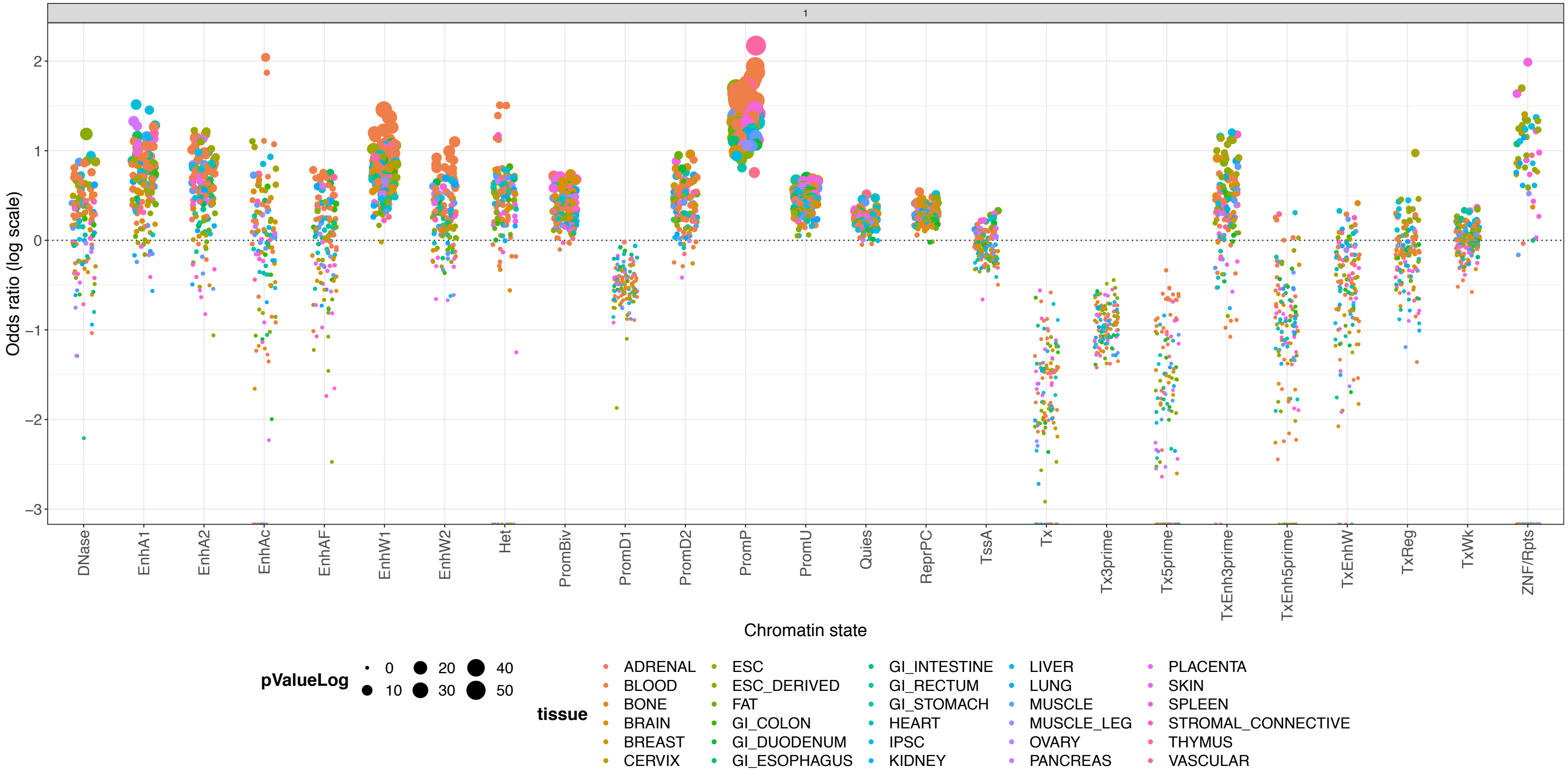
Chromatin states associated with cis correlating DNAm sites >0.9 at birth in ARIES



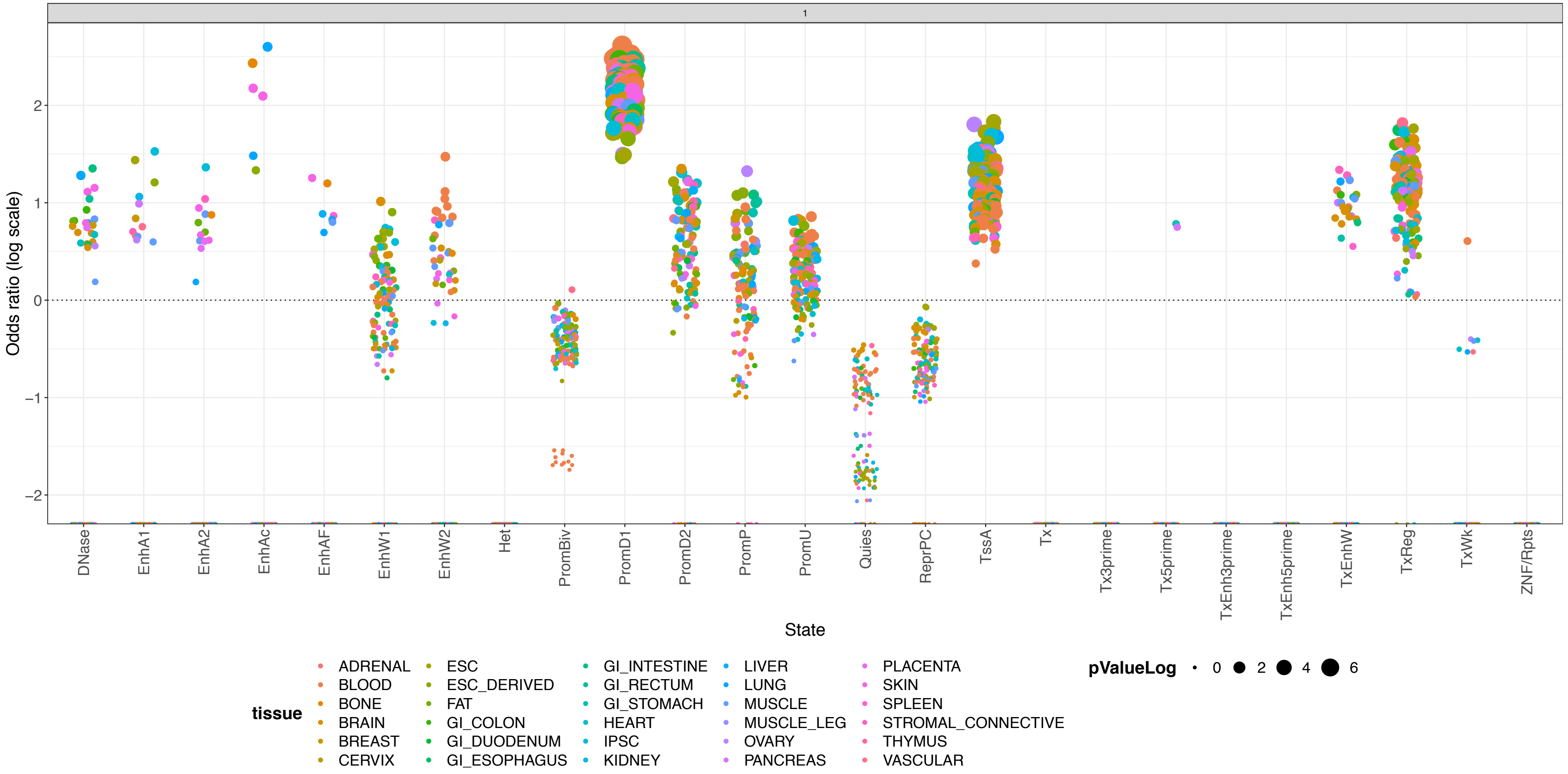
Chromatin states associated with trans correlating DNAm sites >0.9 at birth in ARIES



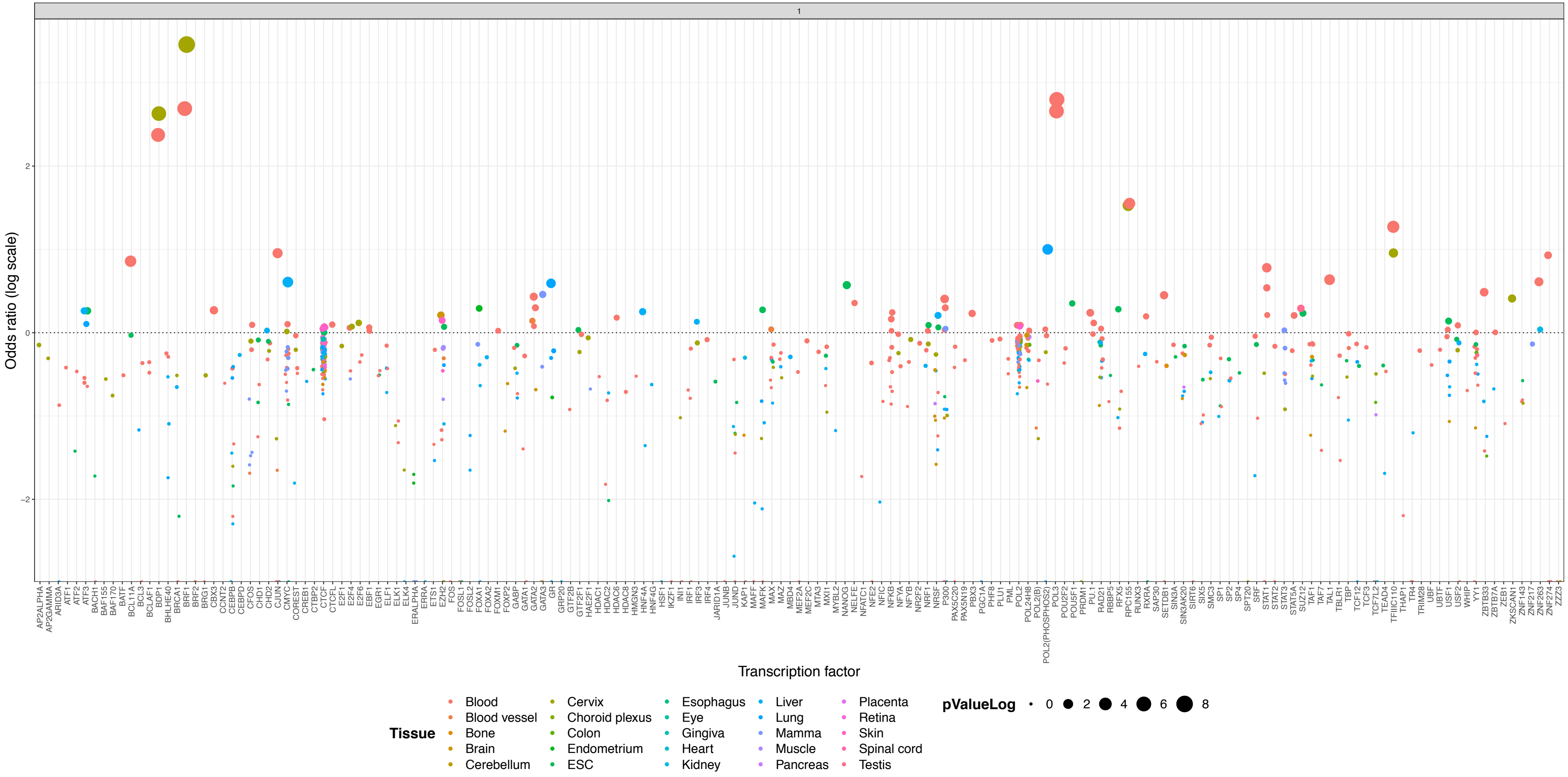
Chromatin states associated with cis correlating DNAm sites >0.9 at 15-17 years in ARIES



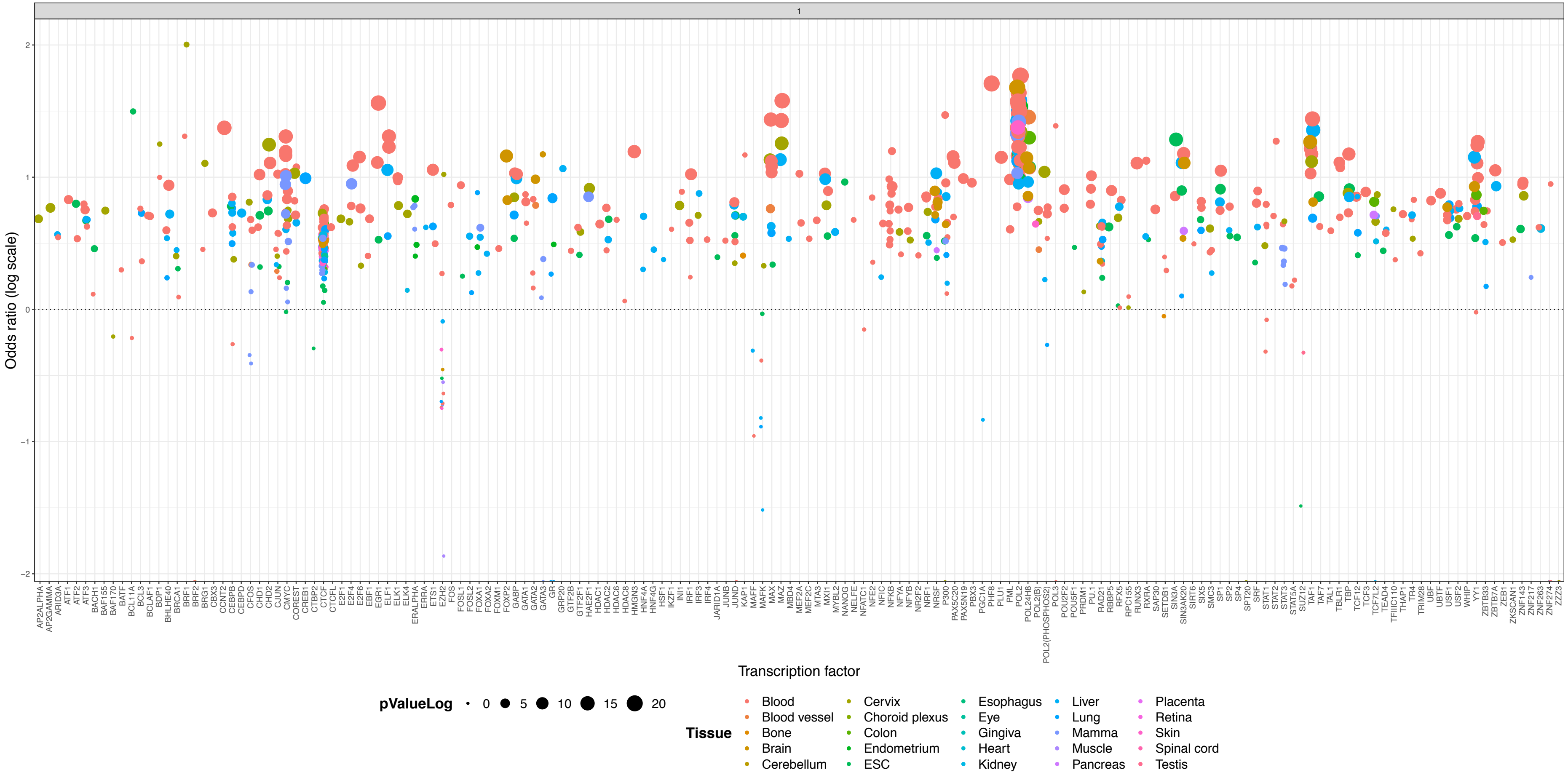
Chromatin states associated with trans correlating DNAm sites >0.9 at 15-17 years in ARIES



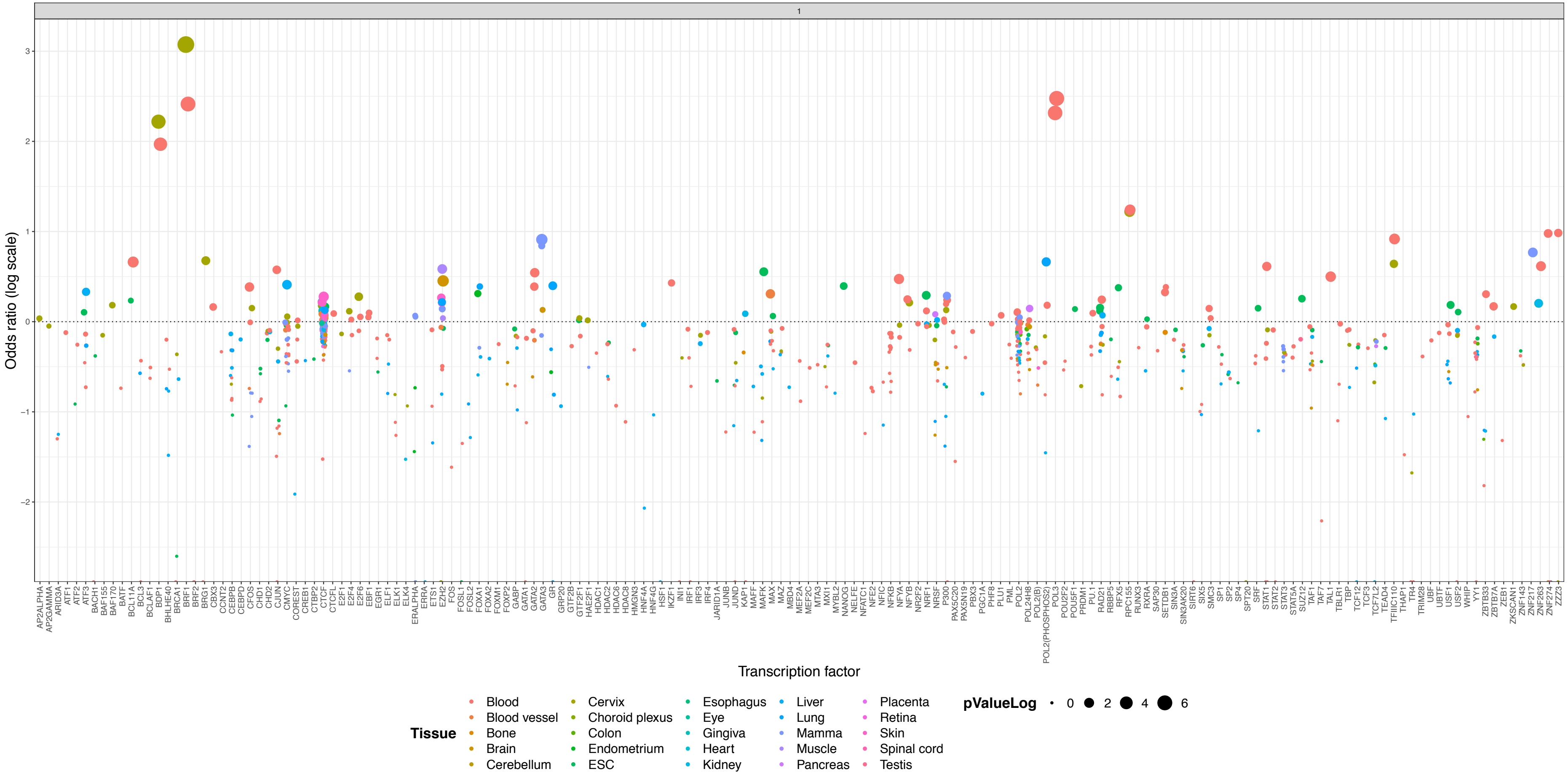
Transcription factor binding site enrichment for cis correlating DNAm sites >0.9 at birth in ARIES



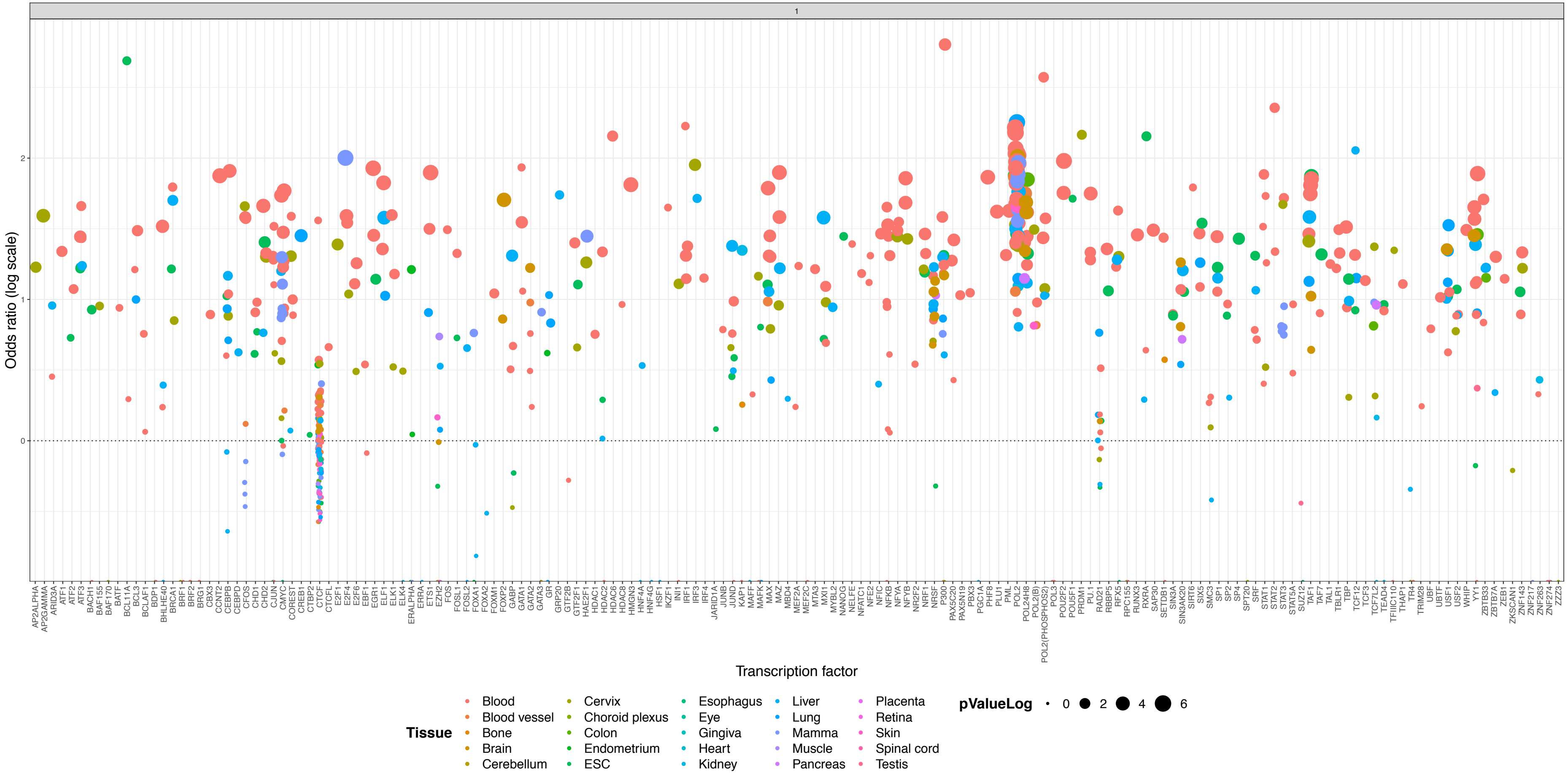
Transcription factor binding site enrichment for trans correlating DNAm sites >0.9 at birth in ARIES



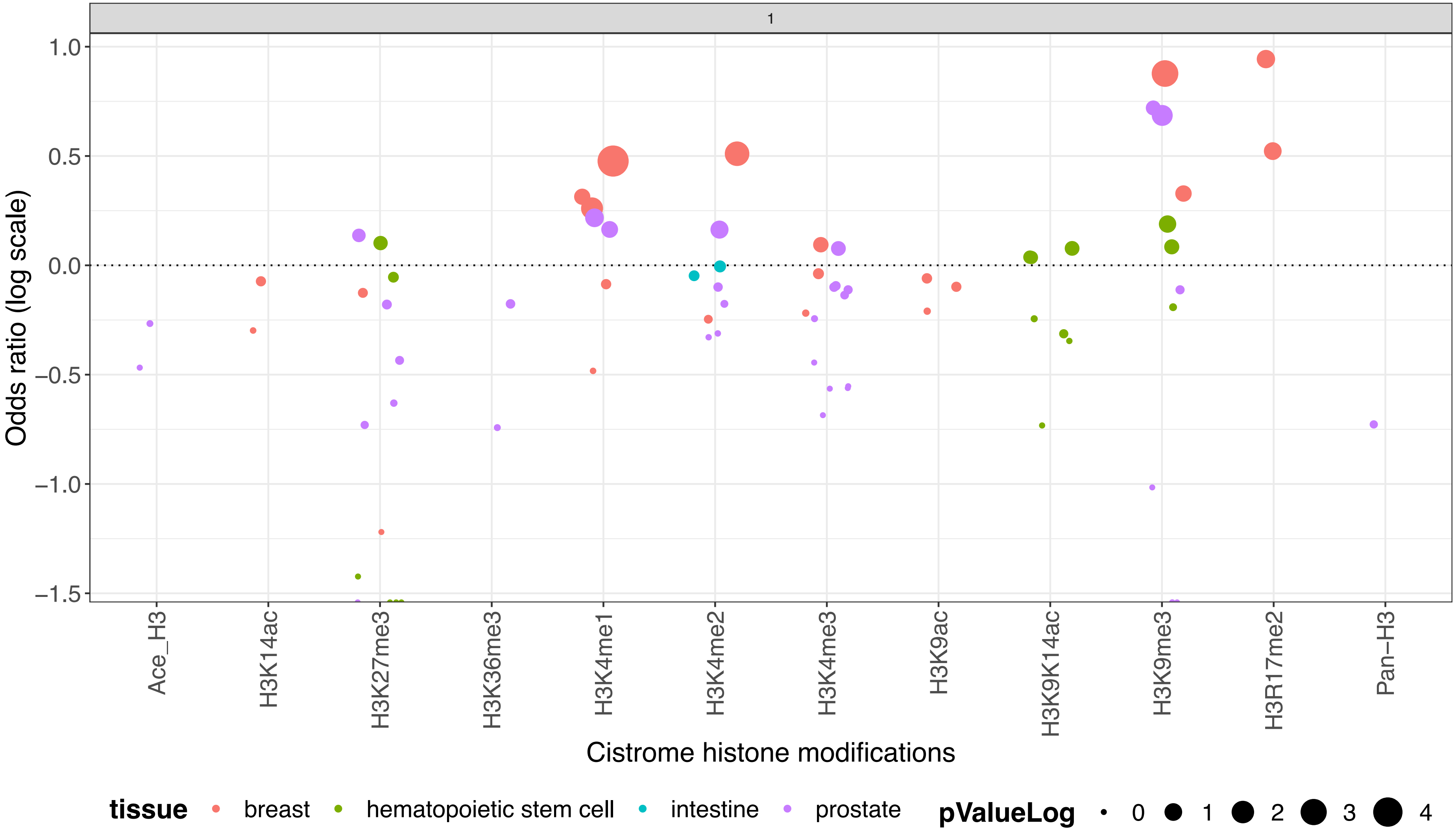
Transcription factor binding site enrichment for cis correlating DNAm sites >0.9 at 15-17 years in ARIES



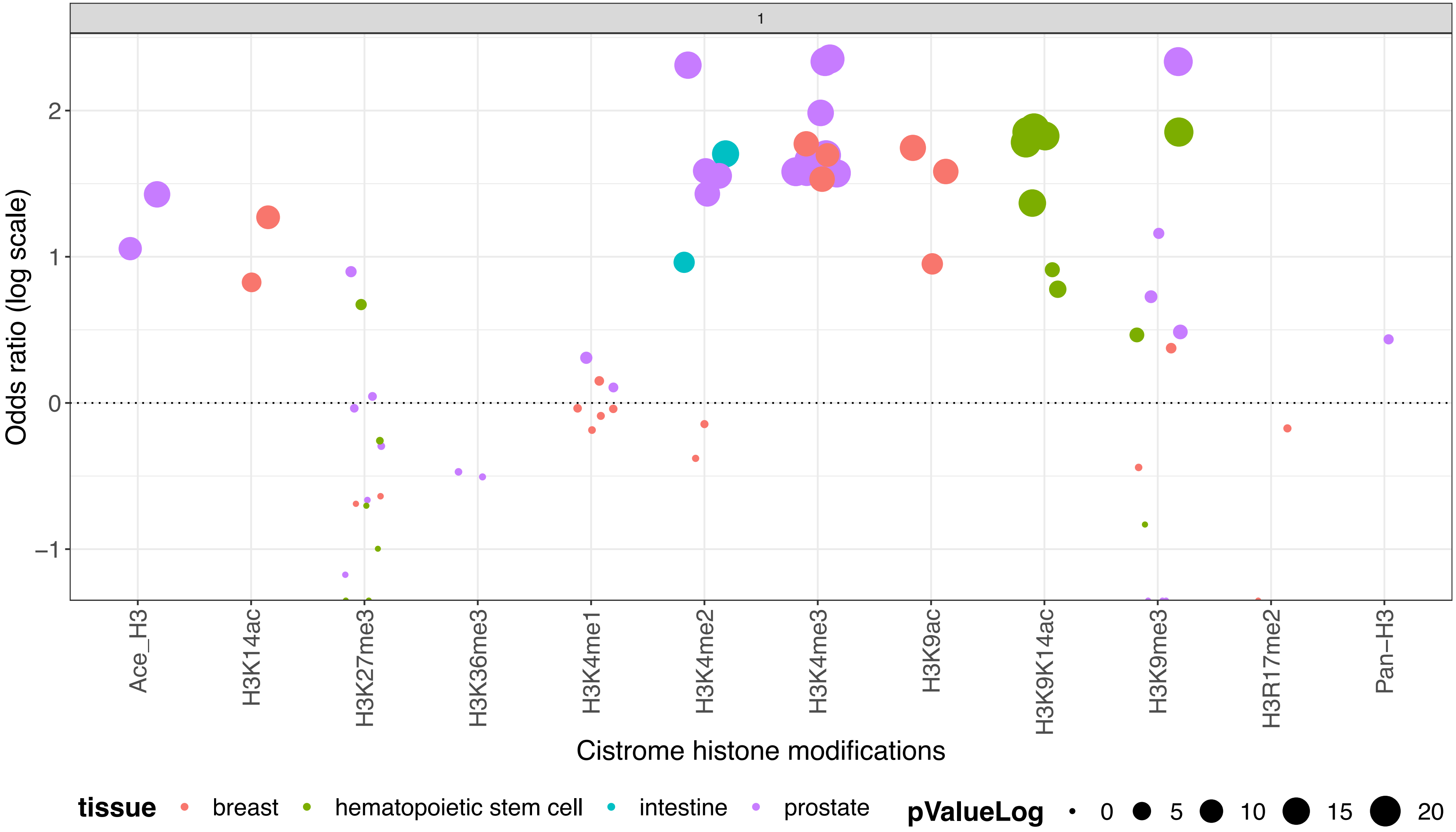
Transcription factor binding site enrichment for trans correlating DNAm sites >0.9 at 15-17 years in ARIES



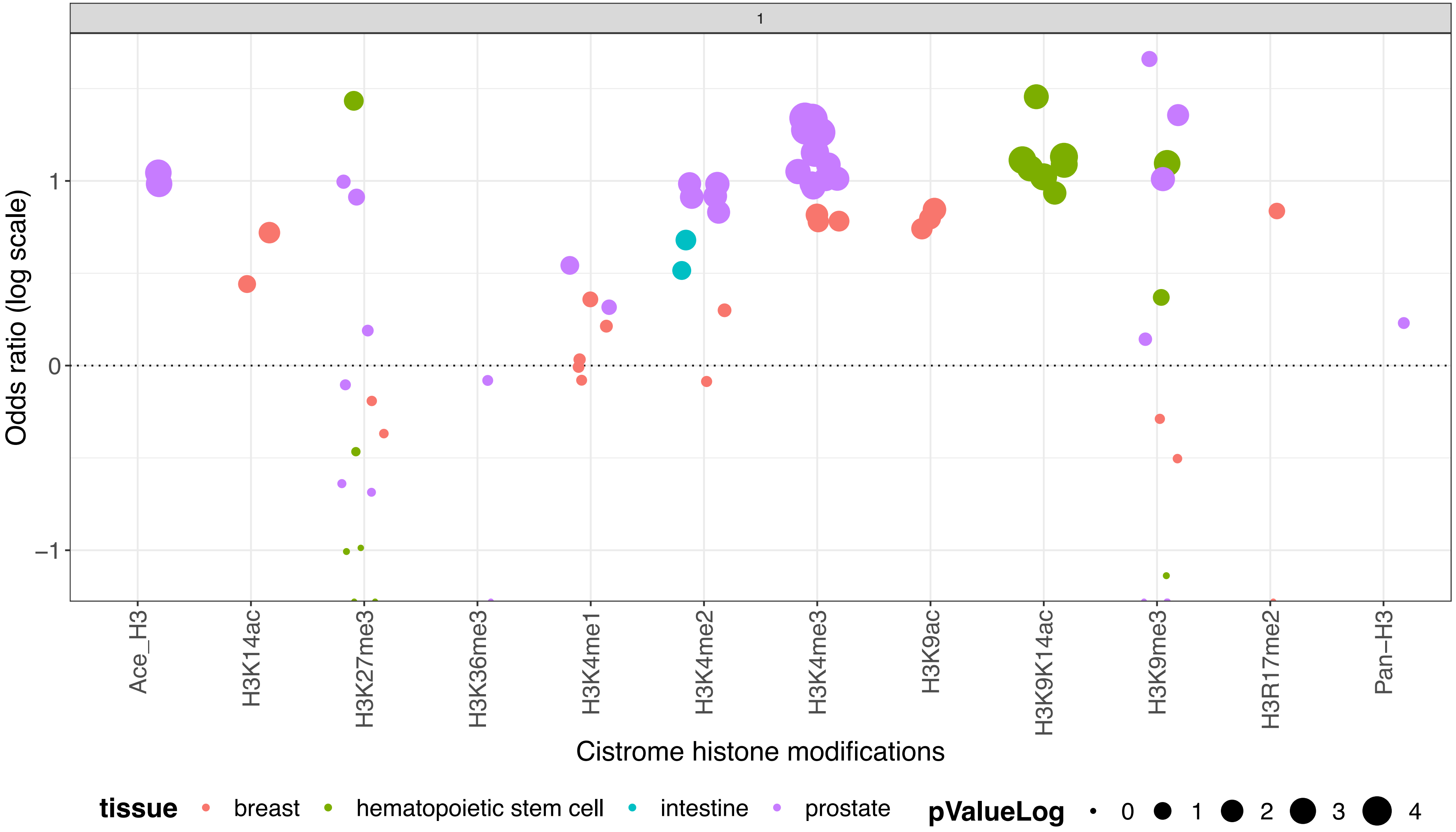
Histone modification enrichment in cis correlating sites >0.9 at birth in ARIES



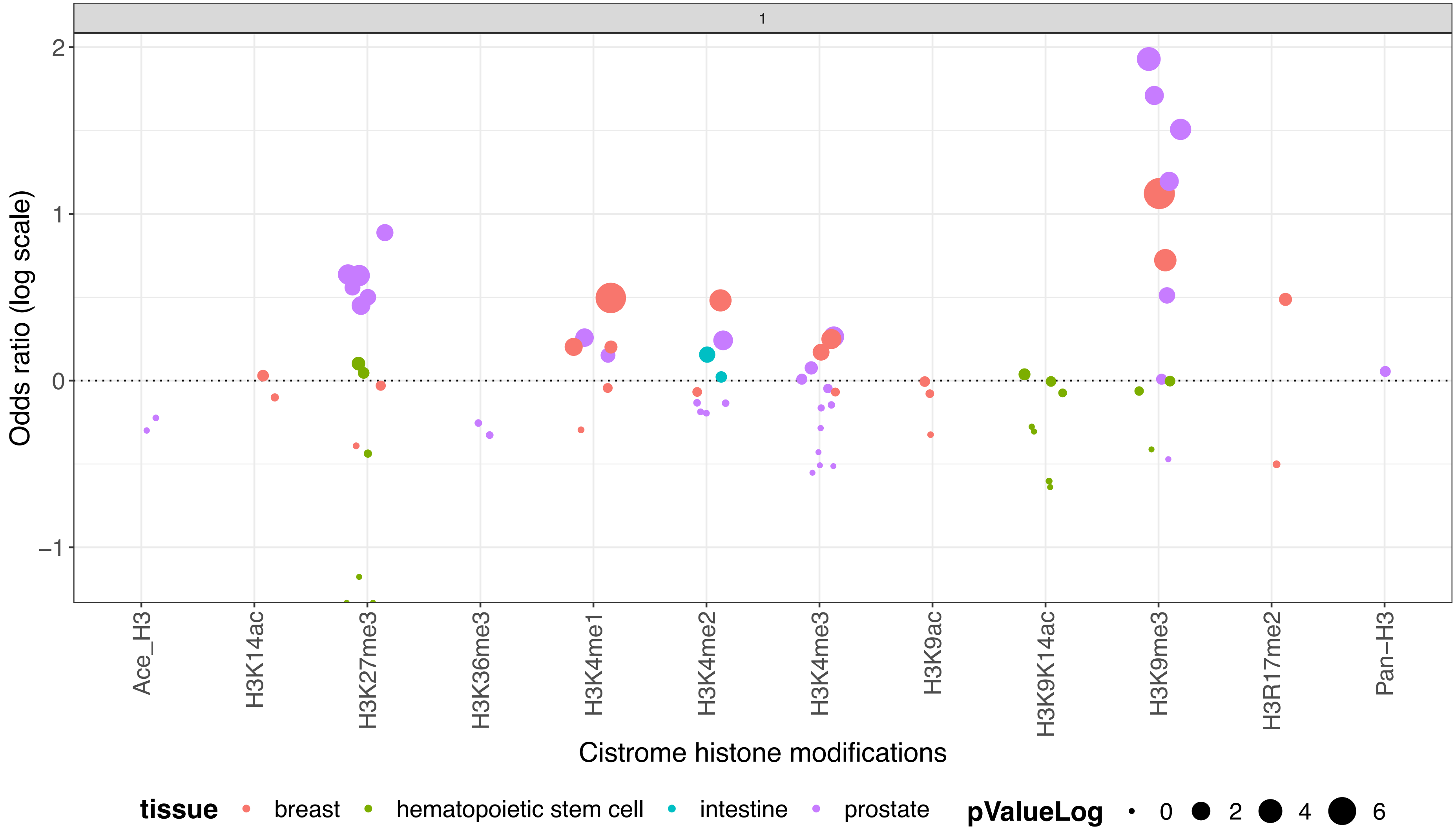
Histone modification enrichment in cis correlating sites >0.9 at birth in ARIES



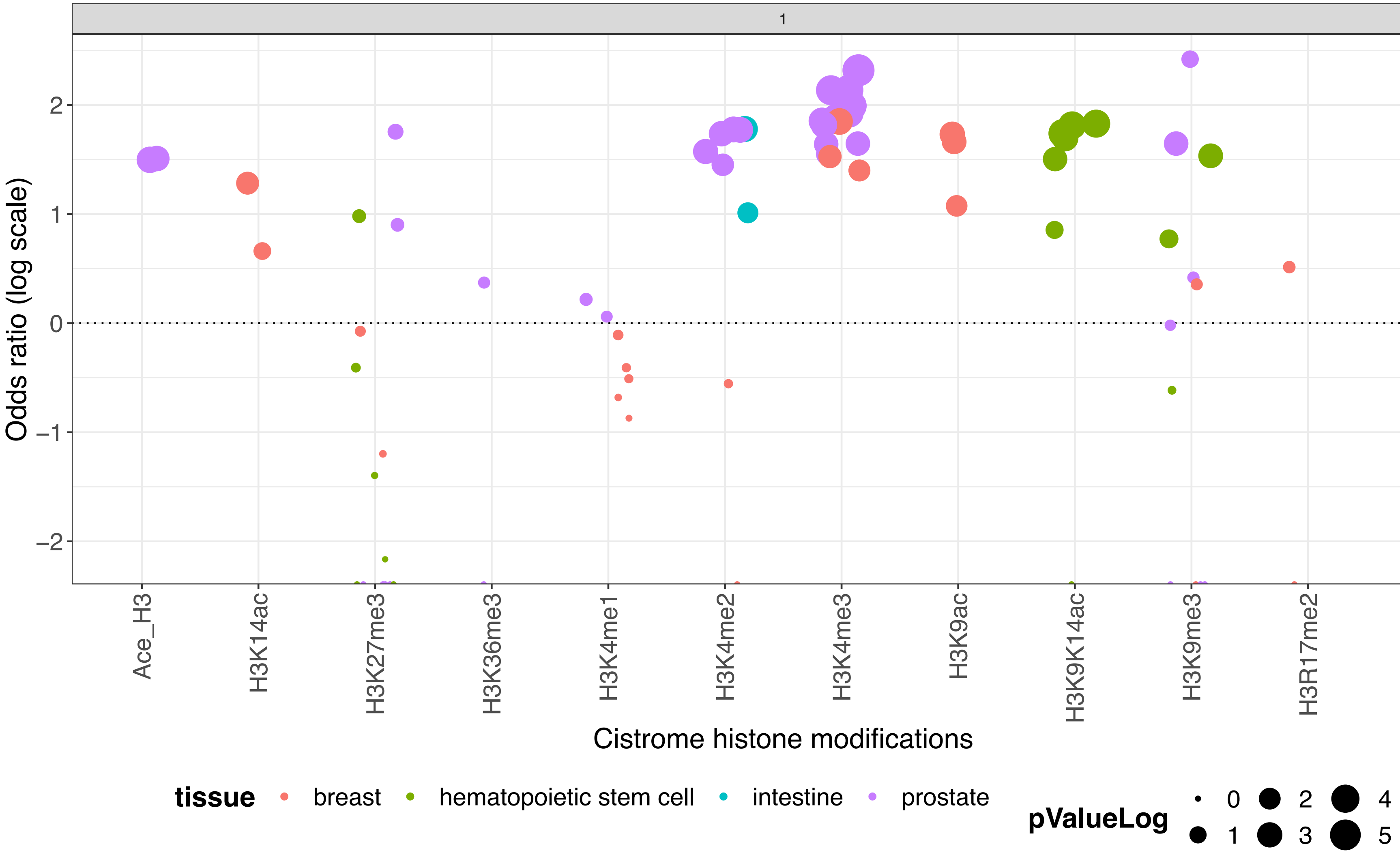
Histone modification enrichment in trans correlating sites >0.9 at birth in ARIES



Histone modification enrichment in cis correlating sites >0.9 at 15-17 years in ARIES



Histone modification enrichment in cis correlating sites >0.9 at 15-17 years in ARIES



Appendix 3: the top 20 asthma-associated DNAm sites in ARIES 7 year olds, detailing correlation with asthma, and strength of relationship to the MEsalmon eigengene

TargetID	chromosome	moduleColor	GS.asthma	p.GS.asthma	MM.salmon	p.MM.salmon
cg19434937	12	salmon	-0.2233026	3.62E-11	0.78197859	3.98E-178
cg11988722	20	salmon	-0.2225069	4.26E-11	0.76025089	1.02E-162
cg11699125	1	salmon	-0.2171643	1.26E-10	0.83004913	1.50E-219
cg12425700	4	salmon	-0.2162484	1.51E-10	0.66513318	8.14E-111
cg06558622	7	salmon	-0.2123459	3.25E-10	0.75885556	8.74E-162
cg14612966	9	salmon	-0.2092316	5.94E-10	0.74309664	1.14E-151
cg10644885	19	salmon	-0.2086235	6.67E-10	0.75171548	4.16E-157
cg05541460	22	salmon	-0.2051219	1.30E-09	0.77941693	3.22E-176
cg09278187	1	salmon	-0.2045053	1.46E-09	0.64270364	2.93E-101
cg04983687	16	salmon	-0.2007131	2.95E-09	0.77674202	2.97E-174
cg24376793	1	salmon	-0.1994883	3.69E-09	0.73305592	1.32E-145
cg09247486	1	salmon	-0.1993805	3.76E-09	0.71074845	4.57E-133
cg13458609	9	salmon	-0.1985409	4.38E-09	0.74265317	2.14E-151
cg04549076	11	salmon	-0.198536	4.39E-09	0.70587397	1.75E-130
cg06824199	1	salmon	-0.1968816	5.92E-09	0.79392548	2.24E-187
cg27469152	17	salmon	-0.1959088	7.05E-09	0.76768853	8.40E-168
cg05786348	2	salmon	-0.1937259	1.04E-08	0.72496298	6.51E-141
cg04497992	16	salmon	-0.1932618	1.13E-08	0.81820395	3.24E-208
cg26396815	4	salmon	-0.1929121	1.20E-08	0.79247817	3.18E-186
cg17041511	17	salmon	-0.1929101	1.20E-08	0.72312122	7.21E-140

Appendix 4: Gene ontology enrichments for ARIES WGCNA modules

Modules are included if there is an FDR significant enrichment.

Black module – ARIES – Birth – gene ontology

	Term	Ont	N	DE	P.DE	FDR
GO:0044428	nuclear part	CC	4030	479	1.29E-07	0.00286595
GO:0005634	nucleus	CC	6501	700	3.16E-06	0.02983137
GO:0031981	nuclear lumen	CC	3701	439	4.03E-06	0.02983137
GO:0043231	intracellular membrane-bounded organelle	CC	9793	997	6.19E-06	0.03430957
GO:0007049	cell cycle	BP	1759	219	2.85E-05	0.12627576
GO:0044422	organelle part	CC	8524	874	3.83E-05	0.14164622
GO:0051172	negative regulation of nitrogen compound metabolic process	BP	2308	266	6.84E-05	0.1806049
GO:0010605	negative regulation of macromolecule metabolic process	BP	2540	289	7.17E-05	0.1806049
GO:0070262	peptidyl-serine dephosphorylation	BP	16	8	7.87E-05	0.1806049
GO:0070013	intracellular organelle lumen	CC	4745	523	9.99E-05	0.1806049
GO:0031974	membrane-enclosed lumen	CC	4745	523	9.99E-05	0.1806049
GO:0043233	organelle lumen	CC	4745	523	9.99E-05	0.1806049
GO:0044446	intracellular organelle part	CC	8305	850	0.00010584	0.1806049
GO:0009892	negative regulation of metabolic process	BP	2759	309	0.00013231	0.19733985
GO:0006091	generation of precursor metabolites and energy	BP	457	65	0.00013344	0.19733985
GO:0022402	cell cycle process	BP	1308	164	0.00015185	0.20372766
GO:0006401	RNA catabolic process	BP	356	54	0.00015613	0.20372766
GO:0006998	nuclear envelope organization	BP	84	19	0.00017537	0.21612053
GO:0031324	negative regulation of cellular metabolic process	BP	2454	277	0.00025633	0.29927121
GO:0043229	intracellular organelle	CC	11413	1118	0.00029377	0.32583332

Black module – ARIES – Birth – KEGG pathway

	Pathway	N	DE	P.DE	FDR
path:hsa00620	Pyruvate metabolism	38	11	0.00021513	0.07228427
path:hsa05010	Alzheimer disease	157	24	0.01196631	0.89850749
path:hsa04722	Neurotrophin signaling pathway	112	20	0.01421989	0.89850749
path:hsa04922	Glucagon signaling pathway	97	16	0.01777748	0.89850749
path:hsa04910	Insulin signaling pathway	130	21	0.02202609	0.89850749
path:hsa04150	mTOR signaling pathway	146	24	0.02333567	0.89850749
path:hsa03013	RNA transport	140	20	0.0236936	0.89850749
path:hsa05133	Pertussis	69	11	0.02422186	0.89850749
path:hsa01200	Carbon metabolism	105	16	0.02454995	0.89850749
path:hsa03018	RNA degradation	72	12	0.02853545	0.89850749
path:hsa05416	Viral myocarditis	37	8	0.02941542	0.89850749
path:hsa03060	Protein export	23	5	0.04769356	1
path:hsa05170	Human immunodeficiency virus 1 infection	179	24	0.04963949	1
path:hsa03015	mRNA surveillance pathway	77	13	0.04965554	1
path:hsa00190	Oxidative phosphorylation	114	15	0.05894632	1
path:hsa04612	Antigen processing and presentation	37	6	0.06104081	1
path:hsa05203	Viral carcinogenesis	151	22	0.06426776	1
path:hsa04310	Wnt signaling pathway	152	23	0.06772709	1
path:hsa05031	Amphetamine addiction	64	11	0.06791655	1
path:hsa03050	Proteasome	43	7	0.06994362	1

Blue module – ARIES – Birth – Gene ontology

	Term	Ont	N	DE	P.DE	FDR
GO:0005575	cellular_component	CC	16397	4262	4.65E-10	1.03E-05
GO:0071944	cell periphery	CC	4868	1401	6.03E-09	4.85E-05
GO:0007155	cell adhesion	BP	1256	420	9.49E-09	4.85E-05
GO:0022610	biological adhesion	BP	1263	422	9.88E-09	4.85E-05
GO:0044425	membrane part	CC	5997	1671	1.09E-08	4.85E-05
GO:0005886	plasma membrane	CC	4765	1364	3.35E-08	0.0001238
GO:0005623	cell	CC	14968	3926	5.11E-08	0.00016186
GO:0044464	cell part	CC	14941	3919	6.34E-08	0.00017567
GO:0031224	intrinsic component of membrane	CC	4837	1345	2.93E-07	0.0007227
GO:0016020	membrane	CC	8277	2254	3.97E-07	0.00082851
GO:0016021	integral component of membrane	CC	4713	1312	4.11E-07	0.00082851
GO:0008150	biological_process	BP	15599	4052	1.70E-06	0.00313656
GO:0043167	ion binding	MF	5651	1579	7.77E-06	0.01304436
GO:0031344	regulation of cell projection organization	BP	573	206	8.23E-06	0.01304436
GO:0045202	synapse	CC	823	281	1.32E-05	0.01957738
GO:0140096	catalytic activity, acting on a protein	MF	2123	631	2.01E-05	0.02785706
GO:0120035	regulation of plasma membrane bounded cell projection organization	BP	563	201	2.25E-05	0.02941441
GO:0032989	cellular component morphogenesis	BP	1034	344	2.51E-05	0.03094841
GO:0044420	extracellular matrix component	CC	111	51	2.80E-05	0.03172845
GO:0007156	homophilic cell adhesion via plasma membrane adhesion molecules	BP	150	66	2.86E-05	0.03172845

Blue module – ARIES – Birth – KEGG pathway

	Pathway	N	DE	P.DE	FDR
path:hsa04151	PI3K-Akt signaling pathway	326	114	0.00087464	0.1628594
path:hsa04520	Adherens junction	69	32	0.0009694	0.1628594
path:hsa04512	ECM-receptor interaction	85	36	0.00223574	0.25040272
path:hsa04510	Focal adhesion	190	70	0.00379181	0.30451377
path:hsa04072	Phospholipase D signaling pathway	142	54	0.00453145	0.30451377
path:hsa04141	Protein processing in endoplasmic reticulum	153	52	0.00682759	0.38234491
path:hsa04120	Ubiquitin mediated proteolysis	128	45	0.01241735	0.52602958
path:hsa04621	NOD-like receptor signaling pathway	153	50	0.01252451	0.52602958
path:hsa05165	Human papillomavirus infection	300	100	0.01413079	0.52672852
path:hsa04730	Long-term depression	57	23	0.02065854	0.52672852
path:hsa01521	EGFR tyrosine kinase inhibitor resistance	77	31	0.02088501	0.52672852
path:hsa04923	Regulation of lipolysis in adipocytes	54	22	0.02456411	0.52672852
path:hsa05222	Small cell lung cancer	85	32	0.02501294	0.52672852
path:hsa05215	Prostate cancer	93	35	0.02511433	0.52672852
path:hsa04630	JAK-STAT signaling pathway	135	44	0.02776902	0.52672852
path:hsa05412	Arrhythmogenic right ventricular cardiomyopathy (ARVC)	74	29	0.02947216	0.52672852
path:hsa04726	Serotonergic synapse	107	37	0.03102444	0.52672852
path:hsa04724	Glutamatergic synapse	111	40	0.03367144	0.52672852
path:hsa00062	Fatty acid elongation	25	11	0.03442342	0.52672852
path:hsa04152	AMPK signaling pathway	117	42	0.03442833	0.52672852

Brown module – ARIES – Birth – Gene ontology

	Term	Ont	N	DE	P.DE	FDR
GO:0022603	regulation of anatomical structure morphogenesis	BP	1041	18	2.82E-06	0.06254947
GO:0050793	regulation of developmental process	BP	2358	27	7.49E-06	0.08304349
GO:0048640	negative regulation of developmental growth	BP	91	6	1.54E-05	0.11354804
GO:0051239	regulation of multicellular organismal process	BP	2744	28	2.52E-05	0.13983725
GO:0030517	negative regulation of axon extension	BP	33	4	5.05E-05	0.22389768
GO:0010771	negative regulation of cell morphogenesis involved in differentiation	BP	81	5	0.00012189	0.31596665
GO:0019932	second-messenger-mediated signaling	BP	328	8	0.00012937	0.31596665
GO:0051960	regulation of nervous system development	BP	778	14	0.0001316	0.31596665
GO:0009653	anatomical structure morphogenesis	BP	2494	26	0.00015591	0.31596665
GO:0097512	cardiac myofibril	CC	7	2	0.00015764	0.31596665
GO:2000026	regulation of multicellular organismal development	BP	1852	21	0.00016323	0.31596665
GO:0022604	regulation of cell morphogenesis	BP	435	10	0.00017092	0.31596665
GO:0060560	developmental growth involved in morphogenesis	BP	201	7	0.00026483	0.4275581
GO:0021514	ventral spinal cord interneuron differentiation	BP	17	3	0.00029397	0.4275581
GO:0021776	smoothened signaling pathway involved in spinal cord motor neuron cell fate specification	BP	3	2	0.00030839	0.4275581
GO:0021775	smoothened signaling pathway involved in ventral spinal cord interneuron specification	BP	3	2	0.00030839	0.4275581
GO:0043086	negative regulation of catalytic activity	BP	769	11	0.00037116	0.4488993
GO:0007154	cell communication	BP	6025	41	0.00038063	0.4488993
GO:0050771	negative regulation of axonogenesis	BP	56	4	0.00038449	0.4488993

Brown module – ARIES – Birth – KEGG pathway

	Pathway	N	DE	P.DE	FDR
path:hsa05410	Hypertrophic cardiomyopathy (HCM)	86	3	0.01026352	1
path:hsa05414	Dilated cardiomyopathy (DCM)	91	3	0.01422944	1
path:hsa05152	Tuberculosis	142	3	0.01981452	1
path:hsa04810	Regulation of actin cytoskeleton	202	4	0.02084418	1
path:hsa04340	Hedgehog signaling pathway	49	2	0.03540922	1
path:hsa04260	Cardiac muscle contraction	78	2	0.0405711	1
path:hsa05217	Basal cell carcinoma	61	2	0.05185859	1
path:hsa04390	Hippo signaling pathway	150	3	0.06262019	1
path:hsa00730	Thiamine metabolism	15	1	0.06765476	1
path:hsa04145	Phagosome	116	2	0.06788438	1
path:hsa04625	C-type lectin receptor signaling pathway	98	2	0.07024962	1
path:hsa00604	Glycosphingolipid biosynthesis - ganglio series	14	1	0.07289388	1
path:hsa05412	Arrhythmogenic right ventricular cardiomyopathy (ARVC)	74	2	0.07591522	1
path:hsa04151	PI3K-Akt signaling pathway	326	4	0.08652407	1
path:hsa04940	Type I diabetes mellitus	21	1	0.08727262	1
path:hsa04966	Collecting duct acid secretion	26	1	0.09586269	1
path:hsa04014	Ras signaling pathway	221	3	0.10681767	1
path:hsa05200	Pathways in cancer	496	5	0.10710773	1
path:hsa00514	Other types of O-glycan biosynthesis	22	1	0.11305837	1
path:hsa04950	Maturity onset diabetes of the young	26	1	0.12614913	1

Green module – ARIES – Birth – Gene ontology

	Term	Ont	N	DE	P.DE	FDR
GO:0000981	RNA polymerase II transcription factor activity, sequence-specific DNA binding	MF	1027	54	1.93E-14	4.28E-10
GO:0043565	sequence-specific DNA binding	MF	1052	51	1.58E-12	1.75E-08
GO:0003700	DNA binding transcription factor activity	MF	1404	56	6.36E-12	4.70E-08
GO:0140110	transcription regulator activity	MF	1733	58	2.49E-09	1.38E-05
GO:0003677	DNA binding	MF	2233	63	3.15E-08	0.00013605
GO:1990837	sequence-specific double-stranded DNA binding	MF	733	35	3.68E-08	0.00013605
GO:0000976	transcription regulatory region sequence-specific DNA binding	MF	698	34	4.94E-08	0.00015641
GO:0003002	regionalization	BP	319	22	1.24E-07	0.00034302
GO:0045165	cell fate commitment	BP	248	19	2.14E-07	0.00052685
GO:0044212	transcription regulatory region DNA binding	MF	825	36	2.48E-07	0.00052685
GO:0001067	regulatory region nucleic acid binding	MF	826	36	2.61E-07	0.00052685
GO:0003690	double-stranded DNA binding	MF	809	35	3.01E-07	0.00055638
GO:0007389	pattern specification process	BP	404	24	3.29E-07	0.00056185
GO:0048856	anatomical structure development	BP	5393	115	5.17E-07	0.00081947
GO:0021879	forebrain neuron differentiation	BP	49	9	8.98E-07	0.00132512
GO:0032501	multicellular organismal process	BP	6780	127	9.56E-07	0.00132512
GO:0000977	RNA polymerase II regulatory region sequence-specific DNA binding	MF	624	29	1.67E-06	0.00215622
GO:0001012	RNA polymerase II regulatory region DNA binding	MF	627	29	1.83E-06	0.00215622
GO:0001228	transcriptional activator activity, RNA polymerase II transcription regulatory region sequence-specific DNA binding	MF	396	22	1.87E-06	0.00215622
GO:0007275	multicellular organism development	BP	4943	107	2.00E-06	0.00215622

Green module – ARIES – Birth – KEGG pathway

	Pathway	N	DE	P.DE	FDR
path:hsa04672	Intestinal immune network for IgA production	31	2	0.0264386	1
path:hsa00515	Mannose type O-glycan biosynthesis	22	2	0.04405629	1
path:hsa05226	Gastric cancer	144	6	0.05085027	1
path:hsa00040	Pentose and glucuronate interconversions	30	2	0.05429066	1
path:hsa00532	Glycosaminoglycan biosynthesis - chondroitin sulfate / dermatan sulfate	19	2	0.05444488	1
path:hsa04950	Maturity onset diabetes of the young	26	2	0.06945812	1
path:hsa05202	Transcriptional misregulation in cancer	158	6	0.07032233	1
path:hsa00230	Purine metabolism	123	4	0.11643244	1
path:hsa05216	Thyroid cancer	34	2	0.12567571	1
path:hsa04015	Rap1 signaling pathway	205	6	0.1573056	1
path:hsa00603	Glycosphingolipid biosynthesis - globo and isoglobo series	13	1	0.15734762	1
path:hsa04080	Neuroactive ligand-receptor interaction	311	6	0.16427503	1
path:hsa04614	Renin-angiotensin system	20	1	0.17874483	1
path:hsa01210	2-Oxocarboxylic acid metabolism	16	1	0.19400464	1
path:hsa04550	Signaling pathways regulating pluripotency of stem cells	133	4	0.21912908	1
path:hsa00220	Arginine biosynthesis	18	1	0.224629	1
path:hsa04924	Renin secretion	67	2	0.233958	1
path:hsa04514	Cell adhesion molecules (CAMs)	116	3	0.24395071	1
path:hsa04916	Melanogenesis	101	3	0.25516541	1
path:hsa00531	Glycosaminoglycan degradation	18	1	0.26327635	1

Light green module – ARIES – Birth – Gene ontology

	Term	Ont	N	DE	P.DE	FDR
GO:0005675	holo TFIIH complex	CC	9	2	3.22E-05	0.55830299
GO:0008353	RNA polymerase II carboxy-terminal domain kinase activity	MF	14	2	7.05E-05	0.55830299
GO:0006368	transcription elongation from RNA polymerase II promoter	BP	99	3	0.00012009	0.55830299
GO:0032806	carboxy-terminal domain protein kinase complex	CC	21	2	0.00018804	0.55830299
GO:0006354	DNA-templated transcription, elongation	BP	121	3	0.00022432	0.55830299
GO:0090534	calcium ion-transporting ATPase complex	CC	1	1	0.00025462	0.55830299
GO:1902081	negative regulation of calcium ion import into sarcoplasmic reticulum	BP	1	1	0.00025462	0.55830299
GO:1902080	regulation of calcium ion import into sarcoplasmic reticulum	BP	1	1	0.00025462	0.55830299
GO:0006294	nucleotide-excision repair, preincision complex assembly	BP	25	2	0.00025849	0.55830299
GO:0006362	transcription elongation from RNA polymerase I promoter	BP	27	2	0.00029509	0.55830299
GO:0006363	termination of RNA polymerase I transcription	BP	28	2	0.000324	0.55830299
GO:0006370	7-methylguanosine mRNA capping	BP	30	2	0.00035558	0.55830299
GO:0016538	cyclin-dependent protein serine/threonine kinase regulator activity	MF	29	2	0.00035651	0.55830299
GO:0009452	7-methylguanosine RNA capping	BP	31	2	0.00037752	0.55830299
GO:0036260	RNA capping	BP	31	2	0.00037752	0.55830299
GO:0006361	transcription initiation from RNA polymerase I promoter	BP	33	2	0.00046222	0.64083929
GO:0045737	positive regulation of cyclin-dependent protein serine/threonine kinase activity	BP	32	2	0.00058106	0.75820848
GO:0006357	regulation of transcription by RNA polymerase II	BP	2044	8	0.00068478	0.79185998
GO:1904031	positive regulation of cyclin-dependent protein kinase activity	BP	35	2	0.00068508	0.79185998

Light green module – ARIES – Birth – KEGG pathway

	Pathway	N	DE	P.DE	FDR
path:hsa03420	Nucleotide excision repair	39	2	0.0005828	0.10415963
path:hsa03022	Basal transcription factors	37	2	0.00062	0.10415963
path:hsa04919	Thyroid hormone signaling pathway	113	2	0.00764461	0.85619677
path:hsa00600	Sphingolipid metabolism	45	1	0.04089905	1
path:hsa04330	Notch signaling pathway	52	1	0.06019142	1
path:hsa05211	Renal cell carcinoma	64	1	0.0702	1
path:hsa04658	Th1 and Th2 cell differentiation	74	1	0.07473722	1
path:hsa05414	Dilated cardiomyopathy (DCM)	91	1	0.10278116	1
path:hsa04066	HIF-1 signaling pathway	103	1	0.10831159	1
path:hsa04071	Sphingolipid signaling pathway	115	1	0.11712936	1
path:hsa04110	Cell cycle	120	1	0.11856228	1
path:hsa04611	Platelet activation	120	1	0.12125348	1
path:hsa04120	Ubiquitin mediated proteolysis	128	1	0.1236633	1
path:hsa04261	Adrenergic signaling in cardiomyocytes	141	1	0.15031382	1
path:hsa04022	cGMP-PKG signaling pathway	159	1	0.16400508	1
path:hsa05170	Human immunodeficiency virus 1 infection	179	1	0.16583195	1
path:hsa04020	Calcium signaling pathway	182	1	0.19053541	1
path:hsa04024	cAMP signaling pathway	208	1	0.20260459	1
path:hsa05165	Human papillomavirus infection	300	1	0.29376856	1
path:hsa04151	PI3K-Akt signaling pathway	326	1	0.3061199	1

Purple module – ARIES – Birth – Gene ontology

	Term	Ont	N	DE	P.DE	FDR
GO:0007156	homophilic cell adhesion via plasma membrane adhesion molecules	BP	150	23	5.11E-13	1.13E-08
GO:0098742	cell-cell adhesion via plasma-membrane adhesion molecules	BP	231	25	1.28E-11	1.42E-07
GO:0098609	cell-cell adhesion	BP	717	38	1.99E-09	1.47E-05
GO:0005887	integral component of plasma membrane	CC	1531	54	1.10E-08	6.11E-05
GO:0016021	integral component of membrane	CC	4713	106	3.09E-08	0.00012525
GO:0007267	cell-cell signaling	BP	1554	59	3.39E-08	0.00012525
GO:0031226	intrinsic component of plasma membrane	CC	1589	54	4.16E-08	0.00013167
GO:0031224	intrinsic component of membrane	CC	4837	107	7.59E-08	0.00021055
GO:0032501	multicellular organismal process	BP	6780	143	1.20E-06	0.00294998
GO:0005509	calcium ion binding	MF	655	30	2.45E-06	0.00544051
GO:0007155	cell adhesion	BP	1256	46	3.45E-06	0.00695395
GO:0022610	biological adhesion	BP	1263	46	4.05E-06	0.00749451
GO:0044459	plasma membrane part	CC	2544	70	9.09E-06	0.01551589
GO:0022832	voltage-gated channel activity	MF	180	14	1.49E-05	0.02100216
GO:0005244	voltage-gated ion channel activity	MF	180	14	1.49E-05	0.02100216
GO:0048731	system development	BP	4404	108	1.61E-05	0.02100216
GO:0007399	nervous system development	BP	2159	70	1.61E-05	0.02100216
GO:0044425	membrane part	CC	5997	118	6.78E-05	0.08356331
GO:0010469	regulation of signaling receptor activity	BP	488	18	7.16E-05	0.08358948
GO:0022843	voltage-gated cation channel activity	MF	134	11	0.00010307	0.11431852

Purple module – ARIES – Birth – KEGG pathway

	Pathway	N	DE	P.DE	FDR
path:hsa05412	Arrhythmogenic right ventricular cardiomyopathy (ARVC)	74	7	0.00127391	0.4082946
path:hsa05414	Dilated cardiomyopathy (DCM)	91	7	0.00243033	0.4082946
path:hsa05410	Hypertrophic cardiomyopathy (HCM)	86	6	0.00609731	0.68289864
path:hsa04514	Cell adhesion molecules (CAMs)	116	6	0.01402955	1
path:hsa04080	Neuroactive ligand-receptor interaction	311	9	0.02627902	1
path:hsa04918	Thyroid hormone synthesis	68	4	0.02974584	1
path:hsa04260	Cardiac muscle contraction	78	4	0.03258762	1
path:hsa04261	Adrenergic signaling in cardiomyocytes	141	6	0.05398787	1
path:hsa04010	MAPK signaling pathway	276	9	0.09698496	1
path:hsa00524	Neomycin, kanamycin and gentamicin biosynthesis	5	1	0.11385042	1
path:hsa05217	Basal cell carcinoma	61	3	0.15113492	1
path:hsa04974	Protein digestion and absorption	86	3	0.16341557	1
path:hsa04921	Oxytocin signaling pathway	149	5	0.17012248	1
path:hsa04640	Hematopoietic cell lineage	76	2	0.18029802	1
path:hsa04934	Cushing syndrome	150	5	0.18131351	1
path:hsa05110	Vibrio cholerae infection	48	2	0.19258079	1
path:hsa04910	Insulin signaling pathway	130	4	0.20343131	1
path:hsa04914	Progesterone-mediated oocyte maturation	84	3	0.20624024	1
path:hsa05014	Amyotrophic lateral sclerosis (ALS)	47	2	0.20873572	1
path:hsa01210	2-Oxocarboxylic acid metabolism	16	1	0.20928237	1

Turquoise module – ARIES – Birth – Gene ontology

	Term	Ont	N	DE	P.DE	FDR
GO:0043231	intracellular membrane-bounded organelle	CC	9793	5095	7.70E-81	1.71E-76
GO:0044424	intracellular part	CC	12998	6422	3.12E-78	3.46E-74
GO:0005622	intracellular	CC	13234	6490	8.13E-72	6.01E-68
GO:0043229	intracellular organelle	CC	11413	5728	1.13E-69	6.27E-66
GO:0043227	membrane-bounded organelle	CC	11374	5669	1.53E-63	6.81E-60
GO:0044446	intracellular organelle part	CC	8305	4324	2.17E-58	8.03E-55
GO:0005634	nucleus	CC	6501	3482	2.80E-54	8.87E-51
GO:0043226	organelle	CC	12279	6004	6.40E-54	1.78E-50
GO:0044422	organelle part	CC	8524	4397	1.10E-53	2.70E-50
GO:0031981	nuclear lumen	CC	3701	2153	1.41E-50	3.13E-47
GO:0044428	nuclear part	CC	4030	2309	1.93E-49	3.90E-46
GO:0005654	nucleoplasm	CC	3170	1883	2.12E-49	3.92E-46
GO:0070013	intracellular organelle lumen	CC	4745	2638	5.15E-49	7.62E-46
GO:0031974	membrane-enclosed lumen	CC	4745	2638	5.15E-49	7.62E-46
GO:0043233	organelle lumen	CC	4745	2638	5.15E-49	7.62E-46
GO:0005737	cytoplasm	CC	10570	5231	4.23E-47	5.87E-44
GO:0044237	cellular metabolic process	BP	9902	4909	1.42E-45	1.85E-42
GO:0044444	cytoplasmic part	CC	8871	4454	7.84E-44	9.66E-41
GO:0006139	nucleobase-containing compound metabolic process	BP	5371	2875	4.65E-41	5.43E-38
GO:0034641	cellular nitrogen compound metabolic process	BP	6161	3213	5.03E-41	5.58E-38

Turquoise module – ARIES – Birth – KEGG pathway

	Pathway	N	DE	P.DE	FDR
path:hsa04141	Protein processing in endoplasmic reticulum	153	106	1.54E-09	5.16E-07
path:hsa05220	Chronic myeloid leukemia	72	54	6.72E-06	0.00080997
path:hsa04140	Autophagy - animal	129	87	7.23E-06	0.00080997
path:hsa05203	Viral carcinogenesis	151	98	2.10E-05	0.00166804
path:hsa05210	Colorectal cancer	83	59	2.48E-05	0.00166804
path:hsa04110	Cell cycle	120	79	3.50E-05	0.00195811
path:hsa04150	mTOR signaling pathway	146	96	4.10E-05	0.00196923
path:hsa05161	Hepatitis B	140	89	4.87E-05	0.00204416
path:hsa04070	Phosphatidylinositol signaling system	92	62	7.56E-05	0.0025881
path:hsa04071	Sphingolipid signaling pathway	115	75	8.11E-05	0.0025881
path:hsa01100	Metabolic pathways	1379	672	8.47E-05	0.0025881
path:hsa04211	Longevity regulating pathway	86	59	0.00017355	0.00485951
path:hsa05166	Human T-cell leukemia virus 1 infection	189	114	0.00022507	0.00581713
path:hsa04714	Thermogenesis	209	121	0.00028993	0.00695828
path:hsa04932	Non-alcoholic fatty liver disease (NAFLD)	140	83	0.00048681	0.010324
path:hsa05200	Pathways in cancer	496	273	0.00049162	0.010324
path:hsa05169	Epstein-Barr virus infection	158	95	0.00055835	0.01055544
path:hsa03015	mRNA surveillance pathway	77	52	0.00056547	0.01055544
path:hsa05212	Pancreatic cancer	71	49	0.0005978	0.01057165
path:hsa04115	p53 signaling pathway	71	48	0.00068589	0.01152288

Yellow module – ARIES – Birth – Gene ontology

	Term	Ont	N	DE	P.DE	FDR
GO:0043229	intracellular organelle	CC	11413	1118	2.85E-15	6.32E-11
GO:0044424	intracellular part	CC	12998	1230	6.55E-14	7.26E-10
GO:0043231	intracellular membrane-bounded organelle	CC	9793	987	1.32E-13	9.72E-10
GO:0005622	intracellular	CC	13234	1244	1.75E-13	9.72E-10
GO:0044422	organelle part	CC	8524	877	4.37E-13	1.94E-09
GO:0044446	intracellular organelle part	CC	8305	859	5.27E-13	1.95E-09
GO:0043226	organelle	CC	12279	1162	4.52E-12	1.43E-08
GO:0005634	nucleus	CC	6501	689	1.52E-10	4.22E-07
GO:0032991	protein-containing complex	CC	4588	514	3.41E-10	8.41E-07
GO:0044428	nuclear part	CC	4030	471	4.15E-10	9.21E-07
GO:0031981	nuclear lumen	CC	3701	440	4.78E-10	9.64E-07
GO:0005515	protein binding	MF	10522	1025	6.77E-10	1.25E-06
GO:0043227	membrane-bounded organelle	CC	11374	1080	7.92E-10	1.35E-06
GO:0005829	cytosol	CC	4663	511	1.11E-09	1.76E-06
GO:1901363	heterocyclic compound binding	MF	5520	583	1.88E-09	2.78E-06
GO:0070013	intracellular organelle lumen	CC	4745	527	2.38E-09	2.93E-06
GO:0031974	membrane-enclosed lumen	CC	4745	527	2.38E-09	2.93E-06
GO:0043233	organelle lumen	CC	4745	527	2.38E-09	2.93E-06
GO:0097159	organic cyclic compound binding	MF	5603	585	6.84E-09	7.98E-06
GO:0034641	cellular nitrogen compound metabolic process	BP	6161	627	5.98E-08	6.64E-05

Yellow module – ARIES – Birth – KEGG pathway

	Pathway	N	DE	P.DE	FDR
path:hsa04142	Lysosome	116	21	0.00128566	0.43198121
path:hsa05131	Shigellosis	62	13	0.00570118	0.83407523
path:hsa00310	Lysine degradation	57	12	0.00831829	0.83407523
path:hsa05110	Vibrio cholerae infection	48	10	0.01037717	0.83407523
path:hsa03460	Fanconi anemia pathway	47	10	0.01270614	0.83407523
path:hsa04714	Thermogenesis	209	28	0.01720719	0.83407523
path:hsa00062	Fatty acid elongation	25	6	0.01784013	0.83407523
path:hsa03050	Proteasome	43	8	0.02173095	0.83407523
path:hsa00520	Amino sugar and nucleotide sugar metabolism	47	9	0.0223413	0.83407523
path:hsa00052	Galactose metabolism	28	6	0.02519911	0.84669014
path:hsa03013	RNA transport	140	19	0.02857665	0.86291331
path:hsa05132	Salmonella infection	72	11	0.03134701	0.86291331
path:hsa04330	Notch signaling pathway	52	10	0.03689363	0.86291331
path:hsa04910	Insulin signaling pathway	130	19	0.03787922	0.86291331
path:hsa04141	Protein processing in endoplasmic reticulum	153	20	0.04304502	0.86291331
path:hsa04114	Oocyte meiosis	112	16	0.04462877	0.86291331
path:hsa05034	Alcoholism	118	17	0.04489413	0.86291331
path:hsa04660	T cell receptor signaling pathway	96	14	0.0462275	0.86291331
path:hsa05130	Pathogenic Escherichia coli infection	187	23	0.05646593	0.99855538
path:hsa00430	Taurine and hypotaurine metabolism	11	3	0.06096896	1

Blue module – ARIES – 7 years – Gene ontology

	Term	Ont	N	DE	P.DE	FDR
GO:0005575	cellular_component	CC	16397	5422	6.31E-12	1.40E-07
GO:0016740	transferase activity	MF	2251	863	1.25E-09	1.39E-05
GO:0005623	cell	CC	14968	4986	4.56E-09	3.37E-05
GO:0140096	catalytic activity, acting on a protein	MF	2123	812	6.59E-09	3.65E-05
GO:0044464	cell part	CC	14941	4975	9.18E-09	4.07E-05
GO:0008150	biological_process	BP	15599	5154	4.17E-08	0.0001542
GO:0043412	macromolecule modification	BP	4002	1442	9.60E-08	0.00030415
GO:0016020	membrane	CC	8277	2834	1.40E-06	0.00388489
GO:0071944	cell periphery	CC	4868	1723	2.35E-06	0.00579361
GO:0006464	cellular protein modification process	BP	3822	1372	3.21E-06	0.00647432
GO:0036211	protein modification process	BP	3822	1372	3.21E-06	0.00647432
GO:0044425	membrane part	CC	5997	2070	4.84E-06	0.00824236
GO:0043167	ion binding	MF	5651	1983	5.14E-06	0.00824236
GO:0003674	molecular_function	MF	15583	5137	5.20E-06	0.00824236
GO:0003824	catalytic activity	MF	5428	1888	8.42E-06	0.01245298
GO:0005886	plasma membrane	CC	4765	1680	9.55E-06	0.0131843
GO:0061659	ubiquitin-like protein ligase activity	MF	214	100	1.01E-05	0.0131843
GO:0007155	cell adhesion	BP	1256	492	1.31E-05	0.01543652
GO:0019787	ubiquitin-like protein transferase activity	MF	443	187	1.32E-05	0.01543652
GO:0004842	ubiquitin-protein transferase activity	MF	421	179	1.41E-05	0.01569096

Blue module – ARIES – 7 years – KEGG pathway

	Pathway	N	DE	P.DE	FDR
path:hsa04520	Adherens junction	69	38	0.0005686	0.19048231
path:hsa04740	Olfactory transduction	343	127	0.00132859	0.22253948
path:hsa04514	Cell adhesion molecules (CAMs)	116	53	0.00357007	0.39858034
path:hsa04141	Protein processing in endoplasmic reticulum	152	64	0.00475917	0.39858034
path:hsa04510	Focal adhesion	190	82	0.00964635	0.52502346
path:hsa00561	Glycerolipid metabolism	56	27	0.00973446	0.52502346
path:hsa04512	ECM-receptor interaction	85	40	0.01124425	0.52502346
path:hsa04120	Ubiquitin mediated proteolysis	129	55	0.01253787	0.52502346
path:hsa04072	Phospholipase D signaling pathway	142	62	0.01561478	0.58121674
path:hsa04530	Tight junction	162	67	0.01819763	0.60962067
path:hsa00534	Glycosaminoglycan biosynthesis - heparan sulfate / heparin	23	13	0.02068349	0.62990633
path:hsa04151	PI3K-Akt signaling pathway	326	128	0.02547139	0.7085999
path:hsa00310	Lysine degradation	55	26	0.02756478	0.7085999
path:hsa00512	Mucin type O-glycan biosynthesis	29	15	0.02961313	0.7085999
path:hsa05412	Arrhythmogenic right ventricular cardiomyopathy (ARVC)	74	34	0.0397113	0.83212875
path:hsa04621	NOD-like receptor signaling pathway	150	57	0.04563803	0.83212875
path:hsa03440	Homologous recombination	39	18	0.04883616	0.83212875
path:hsa03015	mRNA surveillance pathway	77	33	0.05137813	0.83212875
path:hsa04724	Glutamatergic synapse	111	47	0.05564404	0.83212875
path:hsa04961	Endocrine and other factor-regulated calcium reabsorption	49	22	0.06154503	0.83212875

Grey60 module – ARIES – 7 years – Gene ontology

	Term	Ont	N	DE	P.DE	FDR
GO:0044446	intracellular organelle part	CC	8305	297	4.11E-09	9.12E-05
GO:0043227	membrane-bounded organelle	CC	11374	371	1.09E-08	0.00011543
GO:0044422	organelle part	CC	8524	300	1.56E-08	0.00011543
GO:0043231	intracellular membrane-bounded organelle	CC	9793	333	3.05E-08	0.00013178
GO:0070013	intracellular organelle lumen	CC	4745	192	4.16E-08	0.00013178
GO:0031974	membrane-enclosed lumen	CC	4745	192	4.16E-08	0.00013178
GO:0043233	organelle lumen	CC	4745	192	4.16E-08	0.00013178
GO:0043229	intracellular organelle	CC	11413	370	7.50E-08	0.00020784
GO:0044428	nuclear part	CC	4030	170	8.67E-08	0.00021371
GO:0034641	cellular nitrogen compound metabolic process	BP	6161	226	1.67E-07	0.00037138
GO:0031981	nuclear lumen	CC	3701	158	2.11E-07	0.00042535
GO:0016071	mRNA metabolic process	BP	749	48	2.94E-07	0.00054388
GO:0003735	structural constituent of ribosome	MF	147	17	5.08E-07	0.00086754
GO:1990904	ribonucleoprotein complex	CC	707	45	6.50E-07	0.00103011
GO:0005840	ribosome	CC	219	21	6.97E-07	0.00103047
GO:0043226	organelle	CC	12279	384	1.10E-06	0.00152114
GO:0000228	nuclear chromosome	CC	505	36	1.57E-06	0.00205484
GO:0046483	heterocycle metabolic process	BP	5533	206	1.84E-06	0.00226625
GO:0000956	nuclear-transcribed mRNA catabolic process	BP	193	19	1.95E-06	0.00226654
GO:0006725	cellular aromatic compound metabolic process	BP	5573	207	2.13E-06	0.00226654

Grey60 module – ARIES – 7 years – KEGG pathway

	Pathway	N	DE	P.DE	FDR
path:hsa03010	Ribosome	121	15	1.20E-06	0.0004017
path:hsa03050	Proteasome	43	5	0.00602898	0.68581634
path:hsa05168	Herpes simplex virus 1 infection	415	18	0.00641084	0.68581634
path:hsa00630	Glyoxylate and dicarboxylate metabolism	30	4	0.00818885	0.68581634
path:hsa03460	Fanconi anemia pathway	47	5	0.01575295	0.92473988
path:hsa03018	RNA degradation	72	6	0.01656251	0.92473988
path:hsa00670	One carbon pool by folate	19	3	0.01984821	0.94987882
path:hsa03022	Basal transcription factors	37	4	0.02552748	1
path:hsa00190	Oxidative phosphorylation	114	7	0.03123473	1
path:hsa00020	Citrate cycle (TCA cycle)	28	3	0.03603544	1
path:hsa00260	Glycine, serine and threonine metabolism	34	3	0.0416848	1
path:hsa04932	Non-alcoholic fatty liver disease (NAFLD)	140	8	0.04264341	1
path:hsa05134	Legionellosis	52	4	0.04770485	1
path:hsa04068	FoxO signaling pathway	121	7	0.06071614	1
path:hsa03040	Spliceosome	115	7	0.06150593	1
path:hsa00780	Biotin metabolism	3	1	0.08273057	1
path:hsa03013	RNA transport	140	7	0.09860108	1
path:hsa05169	Epstein-Barr virus infection	158	8	0.10112058	1
path:hsa04137	Mitophagy - animal	60	4	0.10810733	1
path:hsa04918	Thyroid hormone synthesis	68	4	0.11065056	1

Light cyan module – ARIES – 7 years – Gene ontology

	Term	Ont	N	DE	P.DE	FDR
GO:0042105	alpha-beta T cell receptor complex	CC	5	3	5.46E-08	0.00097616
GO:0006955	immune response	BP	1761	14	8.80E-08	0.00097616
GO:0002376	immune system process	BP	2550	16	4.18E-07	0.00308877
GO:0050851	antigen receptor-mediated signaling pathway	BP	201	6	9.26E-07	0.0051374
GO:0046649	lymphocyte activation	BP	564	8	3.39E-06	0.01187304
GO:1903037	regulation of leukocyte cell-cell adhesion	BP	261	6	3.70E-06	0.01187304
GO:0042110	T cell activation	BP	404	7	4.16E-06	0.01187304
GO:0002250	adaptive immune response	BP	311	6	5.46E-06	0.01187304
GO:0050778	positive regulation of immune response	BP	643	8	5.73E-06	0.01187304
GO:0002757	immune response-activating signal transduction	BP	439	7	6.13E-06	0.01187304
GO:0042101	T cell receptor complex	CC	19	3	6.19E-06	0.01187304
GO:0050852	T cell receptor signaling pathway	BP	163	5	6.72E-06	0.01187304
GO:0007159	leukocyte cell-cell adhesion	BP	293	6	6.96E-06	0.01187304
GO:0002684	positive regulation of immune system process	BP	901	9	8.62E-06	0.01365927
GO:0002764	immune response-regulating signaling pathway	BP	468	7	9.24E-06	0.01365927
GO:0002429	immune response-activating cell surface receptor signaling pathway	BP	304	6	1.03E-05	0.014301
GO:0050865	regulation of cell activation	BP	467	7	1.12E-05	0.01465598
GO:0002253	activation of immune response	BP	505	7	1.35E-05	0.01662967
GO:1903039	positive regulation of leukocyte cell-cell adhesion	BP	191	5	1.65E-05	0.01889148

Light cyan module – ARIES – 7 years – KEGG pathway

	Pathway	N	DE	P.DE	FDR
path:hsa04658	Th1 and Th2 cell differentiation	74	4	1.08E-05	0.0033987
path:hsa05142	Chagas disease (American trypanosomiasis)	94	4	2.03E-05	0.0033987
path:hsa05169	Epstein-Barr virus infection	158	4	0.00018121	0.02023463
path:hsa05170	Human immunodeficiency virus 1 infection	179	4	0.00026226	0.0219645
path:hsa05235	PD-L1 expression and PD-1 checkpoint pathway in cancer	85	3	0.00047481	0.02720001
path:hsa04659	Th17 cell differentiation	87	3	0.00048716	0.02720001
path:hsa04660	T cell receptor signaling pathway	96	3	0.00069868	0.03343668
path:hsa05162	Measles	115	3	0.00098479	0.0412379
path:hsa04640	Hematopoietic cell lineage	76	2	0.00479042	0.17831
path:hsa05166	Human T-cell leukemia virus 1 infection	189	3	0.00543221	0.18197901
path:hsa04650	Natural killer cell mediated cytotoxicity	93	2	0.01164225	0.35455931
path:hsa05332	Graft-versus-host disease	15	1	0.01636847	0.44581181
path:hsa05330	Allograft rejection	15	1	0.01730016	0.44581181
path:hsa05165	Human papillomavirus infection	300	3	0.02268587	0.54284038
path:hsa05320	Autoimmune thyroid disease	21	1	0.02687242	0.60015074
path:hsa04940	Type I diabetes mellitus	21	1	0.03453543	0.72308547
path:hsa05340	Primary immunodeficiency	30	1	0.04470481	0.88094771
path:hsa05143	African trypanosomiasis	35	1	0.05061211	0.94194755
path:hsa00513	Various types of N-glycan biosynthesis	37	1	0.06664182	1
path:hsa05216	Thyroid cancer	34	1	0.06812478	1

Magenta module – ARIES – 7 years – Gene ontology

	Term	Ont	N	DE	P.DE	FDR
GO:0034641	cellular nitrogen compound metabolic process	BP	6161	458	2.63E-06	0.05834429
GO:0043231	intracellular membrane-bounded organelle	CC	9793	687	1.27E-05	0.14051312
GO:1901991	negative regulation of mitotic cell cycle phase transition	BP	229	32	2.69E-05	0.19925438
GO:1901988	negative regulation of cell cycle phase transition	BP	246	33	4.12E-05	0.20696408
GO:0002520	immune system development	BP	840	82	5.59E-05	0.20696408
GO:1901363	heterocyclic compound binding	MF	5520	409	5.92E-05	0.20696408
GO:0102521	tRNA-4-demethylwyosine synthase activity	MF	2	2	7.86E-05	0.20696408
GO:0097159	organic cyclic compound binding	MF	5603	412	8.76E-05	0.20696408
GO:0003676	nucleic acid binding	MF	3766	289	9.28E-05	0.20696408
GO:0048534	hematopoietic or lymphoid organ development	BP	799	78	9.92E-05	0.20696408
GO:0031088	platelet dense granule membrane	CC	5	4	0.00010263	0.20696408
GO:0044271	cellular nitrogen compound biosynthetic process	BP	4785	357	0.00013145	0.24300479
GO:0071786	endoplasmic reticulum tubular network organization	BP	16	6	0.00019199	0.29523472
GO:0045930	negative regulation of mitotic cell cycle	BP	306	37	0.00019363	0.29523472
GO:0010467	gene expression	BP	5109	376	0.00021772	0.29523472
GO:0046483	heterocycle metabolic process	BP	5533	409	0.00022479	0.29523472
GO:1901360	organic cyclic compound metabolic process	BP	5779	423	0.00022625	0.29523472
GO:0006139	nucleobase-containing compound metabolic process	BP	5371	399	0.00024881	0.30662784
GO:1901987	regulation of cell cycle phase transition	BP	428	47	0.00026507	0.3094788
GO:0045910	negative regulation of DNA recombination	BP	21	7	0.00029733	0.32978892

Magenta module – ARIES – 7 years – KEGG pathway

	Pathway	N	DE	P.DE	FDR
path:hsa00230	Purine metabolism	123	18	0.00147871	0.32308405
path:hsa04612	Antigen processing and presentation	37	7	0.00262627	0.32308405
path:hsa04115	p53 signaling pathway	71	12	0.00336207	0.32308405
path:hsa05169	Epstein-Barr virus infection	158	20	0.00473993	0.32308405
path:hsa05203	Viral carcinogenesis	151	20	0.00482215	0.32308405
path:hsa05231	Choline metabolism in cancer	93	13	0.0146838	0.55547332
path:hsa05210	Colorectal cancer	83	12	0.01522395	0.55547332
path:hsa04110	Cell cycle	120	15	0.01591604	0.55547332
path:hsa05217	Basal cell carcinoma	61	10	0.01833677	0.55547332
path:hsa04550	Signaling pathways regulating pluripotency of stem cells	133	17	0.01915557	0.55547332
path:hsa05215	Prostate cancer	93	13	0.02045239	0.55547332
path:hsa05168	Herpes simplex virus 1 infection	415	30	0.02129147	0.55547332
path:hsa04390	Hippo signaling pathway	150	19	0.02179001	0.55547332
path:hsa03013	RNA transport	140	15	0.02321381	0.55547332
path:hsa00500	Starch and sucrose metabolism	30	5	0.02551915	0.56992772
path:hsa01100	Metabolic pathways	1340	96	0.03217339	0.62993851
path:hsa01212	Fatty acid metabolism	53	8	0.03281626	0.62993851
path:hsa05216	Thyroid cancer	34	6	0.03571528	0.62993851
path:hsa04142	Lysosome	116	13	0.03572786	0.62993851
path:hsa04218	Cellular senescence	143	16	0.04288356	0.71507493

Pink module – ARIES – 7 years – Gene ontology

	Term	Ont	N	DE	P.DE	FDR
GO:0043312	neutrophil degranulation	BP	452	45	9.01E-15	8.38E-11
GO:0002283	neutrophil activation involved in immune response	BP	455	45	1.05E-14	8.38E-11
GO:0043299	leukocyte degranulation	BP	496	47	1.43E-14	8.38E-11
GO:0002274	myeloid leukocyte activation	BP	585	52	1.70E-14	8.38E-11
GO:0042119	neutrophil activation	BP	465	45	2.08E-14	8.38E-11
GO:0002275	myeloid cell activation involved in immune response	BP	504	47	2.57E-14	8.38E-11
GO:0002446	neutrophil mediated immunity	BP	467	45	2.73E-14	8.38E-11
GO:0036230	granulocyte activation	BP	471	45	3.02E-14	8.38E-11
GO:0002444	myeloid leukocyte mediated immunity	BP	509	47	3.60E-14	8.87E-11
GO:0002366	leukocyte activation involved in immune response	BP	629	52	1.90E-13	4.22E-10
GO:0002263	cell activation involved in immune response	BP	633	52	2.63E-13	5.30E-10
GO:0002443	leukocyte mediated immunity	BP	676	51	4.69E-12	8.67E-09
GO:0006955	immune response	BP	1761	94	6.15E-12	1.05E-08
GO:0045055	regulated exocytosis	BP	700	53	1.43E-11	2.27E-08
GO:0002252	immune effector process	BP	1016	64	6.56E-11	9.70E-08
GO:0002376	immune system process	BP	2550	123	1.04E-10	1.45E-07
GO:0045321	leukocyte activation	BP	1063	67	1.61E-10	2.10E-07
GO:0030141	secretory granule	CC	785	52	4.55E-10	5.61E-07
GO:0044433	cytoplasmic vesicle part	CC	1349	76	2.57E-09	3.00E-06
GO:0001775	cell activation	BP	1204	70	3.22E-09	3.57E-06

Pink module – ARIES – 7 years – KEGG pathway

	Pathway	N	DE	P.DE	FDR
path:hsa05146	Amoebiasis	95	11	0.00010962	0.0367213
path:hsa00620	Pyruvate metabolism	38	6	0.00037846	0.06339264
path:hsa03320	PPAR signaling pathway	71	7	0.00202981	0.15720293
path:hsa01100	Metabolic pathways	1340	54	0.00229601	0.15720293
path:hsa04261	Adrenergic signaling in cardiomyocytes	141	12	0.00277599	0.15720293
path:hsa04742	Taste transduction	77	7	0.00281557	0.15720293
path:hsa04971	Gastric acid secretion	73	8	0.00362046	0.17326471
path:hsa00601	Glycosphingolipid biosynthesis - lacto and neolacto series	27	4	0.00453407	0.18986427
path:hsa04725	Cholinergic synapse	110	10	0.00684446	0.21325246
path:hsa04640	Hematopoietic cell lineage	76	6	0.00691761	0.21325246
path:hsa04911	Insulin secretion	81	8	0.00700232	0.21325246
path:hsa04922	Glucagon signaling pathway	97	8	0.00870838	0.24310896
path:hsa04020	Calcium signaling pathway	182	13	0.00950382	0.24490621
path:hsa04928	Parathyroid hormone synthesis, secretion and action	101	9	0.01036144	0.24793446
path:hsa04621	NOD-like receptor signaling pathway	150	9	0.01268581	0.27572946
path:hsa00240	Pyrimidine metabolism	54	5	0.0144145	0.27572946
path:hsa05214	Glioma	72	7	0.01462046	0.27572946
path:hsa04932	Non-alcoholic fatty liver disease (NAFLD)	140	9	0.01481531	0.27572946
path:hsa04710	Circadian rhythm	30	4	0.01604642	0.28292377
path:hsa04510	Focal adhesion	190	13	0.01940006	0.32495097

Salmon module – ARIES – 7 years – Gene ontology

	Term	Ont	N	DE	P.DE	FDR
GO:0048029	monosaccharide binding	MF	67	6	4.36E-06	0.05537625
GO:0032941	secretion by tissue	BP	38	5	5.09E-06	0.05537625
GO:0030246	carbohydrate binding	MF	241	9	7.49E-06	0.05537625
GO:0032673	regulation of interleukin-4 production	BP	23	4	1.50E-05	0.08315022
GO:0032633	interleukin-4 production	BP	29	4	4.33E-05	0.16082908
GO:0001817	regulation of cytokine production	BP	578	13	4.35E-05	0.16082908
GO:0002215	defense response to nematode	BP	3	2	6.61E-05	0.20957507
GO:0030157	pancreatic juice secretion	BP	12	3	8.30E-05	0.23016408
GO:0043560	insulin receptor substrate binding	MF	11	3	0.00011083	0.2715919
GO:0001816	cytokine production	BP	640	13	0.00012243	0.2715919
GO:0032753	positive regulation of interleukin-4 production	BP	18	3	0.00020274	0.40886053
GO:0007589	body fluid secretion	BP	86	5	0.00024186	0.43032583
GO:0050878	regulation of body fluid levels	BP	452	11	0.0002731	0.43032583
GO:0032674	regulation of interleukin-5 production	BP	18	3	0.00029662	0.43032583
GO:0047184	1-acylglycerophosphocholine O-acyltransferase activity	MF	4	2	0.00030965	0.43032583
GO:0071211	protein targeting to vacuole involved in autophagy	BP	3	2	0.00031038	0.43032583
GO:0032634	interleukin-5 production	BP	19	3	0.00033057	0.4313578
GO:0002702	positive regulation of production of molecular mediator of immune response	BP	69	4	0.00067338	0.82987217
GO:0051239	regulation of multicellular organismal process	BP	2744	33	0.00078152	0.912448
GO:0070314	G1 to G0 transition	BP	8	2	0.00091416	1

Salmon module – ARIES – 7 years – KEGG pathway

	Pathway	N	DE	P.DE	FDR
path:hsa05310	Asthma	14	3	2.20E-05	0.00735464
path:hsa00564	Glycerophospholipid metabolism	91	4	0.00363919	0.57178158
path:hsa00010	Glycolysis / Gluconeogenesis	64	3	0.00532915	0.57178158
path:hsa04930	Type II diabetes mellitus	44	3	0.00682724	0.57178158
path:hsa00030	Pentose phosphate pathway	24	2	0.00961237	0.58931847
path:hsa00500	Starch and sucrose metabolism	30	2	0.01268249	0.58931847
path:hsa04910	Insulin signaling pathway	130	4	0.01362785	0.58931847
path:hsa05218	Melanoma	68	3	0.01697818	0.58931847
path:hsa00051	Fructose and mannose metabolism	32	2	0.02010254	0.58931847
path:hsa05100	Bacterial invasion of epithelial cells	72	3	0.02061125	0.58931847
path:hsa04914	Progesterone-mediated oocyte maturation	84	3	0.02318635	0.58931847
path:hsa04144	Endocytosis	227	5	0.02450287	0.58931847
path:hsa01200	Carbon metabolism	105	3	0.02504323	0.58931847
path:hsa04530	Tight junction	162	4	0.02742961	0.58931847
path:hsa04973	Carbohydrate digestion and absorption	38	2	0.02812347	0.58931847
path:hsa00520	Amino sugar and nucleotide sugar metabolism	47	2	0.03169404	0.58931847
path:hsa04380	Osteoclast differentiation	113	3	0.0338545	0.58931847
path:hsa05206	MicroRNAs in cancer	281	5	0.0342016	0.58931847
path:hsa00440	Phosphonate and phosphinate metabolism	5	1	0.03431582	0.58931847
path:hsa04750	Inflammatory mediator regulation of TRP channels	95	3	0.03629361	0.58931847

Turquoise module – ARIES – 7 years – Gene ontology

	Term	Ont	N	DE	P.DE	FDR
GO:0044424	intracellular part	CC	12998	6630	9.90E-88	2.20E-83
GO:0043231	intracellular membrane-bounded organelle	CC	9793	5240	2.42E-84	2.68E-80
GO:0005622	intracellular	CC	13234	6704	6.41E-82	4.74E-78
GO:0043229	intracellular organelle	CC	11413	5895	3.78E-73	2.09E-69
GO:0043227	membrane-bounded organelle	CC	11374	5842	2.06E-68	9.14E-65
GO:0043226	organelle	CC	12279	6200	4.26E-61	1.57E-57
GO:0044446	intracellular organelle part	CC	8305	4443	3.26E-60	1.03E-56
GO:0005634	nucleus	CC	6501	3588	1.73E-58	4.79E-55
GO:0044422	organelle part	CC	8524	4521	7.36E-56	1.81E-52
GO:0044237	cellular metabolic process	BP	9902	5080	3.78E-53	8.40E-50
GO:0044428	nuclear part	CC	4030	2376	1.05E-52	2.13E-49
GO:0031981	nuclear lumen	CC	3701	2209	2.25E-52	4.15E-49
GO:0005654	nucleoplasm	CC	3170	1934	7.54E-52	1.29E-48
GO:0005737	cytoplasm	CC	10570	5386	6.68E-50	1.06E-46
GO:0070013	intracellular organelle lumen	CC	4745	2700	6.89E-49	8.99E-46
GO:0031974	membrane-enclosed lumen	CC	4745	2700	6.89E-49	8.99E-46
GO:0043233	organelle lumen	CC	4745	2700	6.89E-49	8.99E-46
GO:0008152	metabolic process	BP	10529	5316	2.81E-48	3.47E-45
GO:0044238	primary metabolic process	BP	9767	4966	7.15E-47	8.34E-44
GO:0071704	organic substance metabolic process	BP	10077	5101	2.05E-46	2.28E-43

Turquoise module – ARIES – 7 years – KEGG pathway

	Pathway	N	DE	P.DE	FDR
path:hsa04141	Protein processing in endoplasmic reticulum	152	102	2.52E-07	8.43E-05
path:hsa05220	Chronic myeloid leukemia	72	56	1.71E-06	0.00028613
path:hsa05203	Viral carcinogenesis	151	102	4.39E-06	0.00036786
path:hsa04110	Cell cycle	120	83	4.39E-06	0.00036786
path:hsa04140	Autophagy - animal	129	87	2.68E-05	0.0017961
path:hsa05161	Hepatitis B	140	89	0.00016734	0.00934321
path:hsa04115	p53 signaling pathway	71	50	0.00025882	0.01238635
path:hsa01100	Metabolic pathways	1340	663	0.0003071	0.01285984
path:hsa05210	Colorectal cancer	83	57	0.000388	0.01444205
path:hsa05169	Epstein-Barr virus infection	158	97	0.0005755	0.01927926
path:hsa05016	Huntington disease	179	106	0.00063334	0.01928803
path:hsa04144	Endocytosis	227	135	0.00070231	0.01960614
path:hsa05166	Human T-cell leukemia virus 1 infection	189	114	0.00082417	0.02089866
path:hsa05135	Yersinia infection	112	72	0.0008937	0.02089866
path:hsa04071	Sphingolipid signaling pathway	115	73	0.00093576	0.02089866
path:hsa04722	Neurotrophin signaling pathway	112	73	0.00109495	0.02158135
path:hsa04714	Thermogenesis	209	121	0.00111898	0.02158135
path:hsa05131	Shigellosis	62	43	0.0011596	0.02158135
path:hsa05223	Non-small cell lung cancer	62	44	0.00134595	0.0228452
path:hsa04120	Ubiquitin mediated proteolysis	129	80	0.00141876	0.0228452

Yellow module – ARIES – 7 years – Gene ontology

	Term	Ont	N	DE	P.DE	FDR
GO:0022603	regulation of anatomical structure morphogenesis	BP	1041	23	9.58E-08	0.00212431
GO:0032269	negative regulation of cellular protein metabolic process	BP	1172	18	2.87E-05	0.31783204
GO:0051248	negative regulation of protein metabolic process	BP	1226	18	5.54E-05	0.40939269
GO:0032268	regulation of cellular protein metabolic process	BP	2535	27	0.0002126	1
GO:0051246	regulation of protein metabolic process	BP	2704	28	0.00026204	1
GO:0050793	regulation of developmental process	BP	2358	28	0.00027565	1
GO:0045765	regulation of angiogenesis	BP	335	8	0.00047697	1
GO:0040029	regulation of gene expression, epigenetic	BP	423	8	0.00055688	1
GO:0036515	serotonergic neuron axon guidance	BP	4	2	0.00062738	1
GO:0035195	gene silencing by miRNA	BP	275	6	0.00073031	1
GO:0036514	dopaminergic neuron axon guidance	BP	5	2	0.00075275	1
GO:1904938	planar cell polarity pathway involved in axon guidance	BP	5	2	0.00075275	1
GO:0035194	posttranscriptional gene silencing by RNA	BP	282	6	0.00088894	1
GO:1901342	regulation of vasculature development	BP	365	8	0.00088907	1
GO:0016441	posttranscriptional gene silencing	BP	283	6	0.00091472	1
GO:1905330	regulation of morphogenesis of an epithelium	BP	168	6	0.0009473	1
GO:0001738	morphogenesis of a polarized epithelium	BP	128	5	0.00096449	1
GO:0016324	apical plasma membrane	CC	281	7	0.00097735	1
GO:0022604	regulation of cell morphogenesis	BP	435	10	0.0010297	1
GO:0017148	negative regulation of translation	BP	369	7	0.00111215	1

Yellow module – ARIES – 7 years – KEGG pathway

	Pathway	N	DE	P.DE	FDR
path:hsa05110	Vibrio cholerae infection	48	3	0.00258882	0.86725592
path:hsa04966	Collecting duct acid secretion	26	2	0.00754173	1
path:hsa04145	Phagosome	116	3	0.01752917	1
path:hsa05206	MicroRNAs in cancer	281	5	0.0218567	1
path:hsa05323	Rheumatoid arthritis	72	2	0.03835091	1
path:hsa05120	Epithelial cell signaling in Helicobacter pylori infection	65	2	0.04262029	1
path:hsa00190	Oxidative phosphorylation	114	2	0.0784104	1
path:hsa00730	Thiamine metabolism	15	1	0.08134563	1
path:hsa04721	Synaptic vesicle cycle	77	2	0.08395278	1
path:hsa04310	Wnt signaling pathway	152	3	0.0966696	1
path:hsa04390	Hippo signaling pathway	150	3	0.10211084	1
path:hsa05410	Hypertrophic cardiomyopathy (HCM)	86	2	0.10873999	1
path:hsa03020	RNA polymerase	28	1	0.11005115	1
path:hsa04940	Type I diabetes mellitus	21	1	0.11022159	1
path:hsa04012	ErbB signaling pathway	80	2	0.11728001	1
path:hsa01521	EGFR tyrosine kinase inhibitor resistance	77	2	0.12492209	1
path:hsa05414	Dilated cardiomyopathy (DCM)	91	2	0.1317623	1
path:hsa05165	Human papillomavirus infection	300	4	0.1404613	1
path:hsa04136	Autophagy - other	29	1	0.15745059	1
path:hsa05152	Tuberculosis	142	2	0.16331037	1

Black module – ARIES – 15-17 years – Gene ontology

	Term	Ont	N	DE	P.DE	FDR
GO:0006955	immune response	BP	1761	53	1.37E-10	3.04E-06
GO:0002443	leukocyte mediated immunity	BP	676	30	4.06E-10	4.06E-06
GO:0002274	myeloid leukocyte activation	BP	585	28	6.34E-10	4.06E-06
GO:0043312	neutrophil degranulation	BP	452	24	1.01E-09	4.06E-06
GO:0002283	neutrophil activation involved in immune response	BP	455	24	1.08E-09	4.06E-06
GO:0045055	regulated exocytosis	BP	700	31	1.13E-09	4.06E-06
GO:0043299	leukocyte degranulation	BP	496	25	1.28E-09	4.06E-06
GO:0042119	neutrophil activation	BP	465	24	1.54E-09	4.26E-06
GO:0002275	myeloid cell activation involved in immune response	BP	504	25	2.08E-09	4.53E-06
GO:0002446	neutrophil mediated immunity	BP	467	24	2.22E-09	4.53E-06
GO:0036230	granulocyte activation	BP	471	24	2.25E-09	4.53E-06
GO:0002444	myeloid leukocyte mediated immunity	BP	509	25	2.56E-09	4.73E-06
GO:0002366	leukocyte activation involved in immune response	BP	629	28	2.84E-09	4.85E-06
GO:0002263	cell activation involved in immune response	BP	633	28	3.42E-09	5.42E-06
GO:0002376	immune system process	BP	2550	66	5.22E-09	7.73E-06
GO:0002252	immune effector process	BP	1016	36	5.84E-09	8.10E-06
GO:0006887	exocytosis	BP	806	31	6.63E-08	8.66E-05
GO:0001775	cell activation	BP	1204	39	1.23E-07	0.00015214
GO:0030141	secretory granule	CC	785	28	2.66E-07	0.000311
GO:0045321	leukocyte activation	BP	1063	35	2.88E-07	0.00031956

Black module – ARIES – 15-17 years – KEGG pathway

	Pathway	N	DE	P.DE	FDR
path:hsa04261	Adrenergic signaling in cardiomyocytes	141	10	8.61E-05	0.02883699
path:hsa04611	Platelet activation	120	8	0.00043803	0.07337056
path:hsa04725	Cholinergic synapse	110	8	0.00065851	0.07353313
path:hsa04971	Gastric acid secretion	73	6	0.00126124	0.09107216
path:hsa03320	PPAR signaling pathway	71	5	0.00135929	0.09107216
path:hsa04071	Sphingolipid signaling pathway	115	7	0.00213287	0.1190851
path:hsa04921	Oxytocin signaling pathway	149	8	0.00313935	0.15024044
path:hsa04970	Salivary secretion	83	5	0.00562811	0.23567712
path:hsa04972	Pancreatic secretion	94	5	0.00742186	0.26224358
path:hsa04928	Parathyroid hormone synthesis, secretion and action	101	6	0.00782817	0.26224358
path:hsa04932	Non-alcoholic fatty liver disease (NAFLD)	140	6	0.00965175	0.26364014
path:hsa04742	Taste transduction	77	4	0.0102257	0.26364014
path:hsa04072	Phospholipase D signaling pathway	142	7	0.01087144	0.26364014
path:hsa04911	Insulin secretion	81	5	0.0110178	0.26364014
path:hsa04923	Regulation of lipolysis in adipocytes	54	4	0.01188213	0.26536749
path:hsa04371	Apelin signaling pathway	134	6	0.01649464	0.34535661
path:hsa05216	Thyroid cancer	34	3	0.01927633	0.37985708
path:hsa04916	Melanogenesis	101	5	0.02539379	0.46576063
path:hsa04713	Circadian entrainment	95	5	0.02641627	0.46576063
path:hsa04621	NOD-like receptor signaling pathway	150	5	0.02898044	0.4854224

Blue module – ARIES – 15-17 years – Gene ontology

	Term	Ont	N	DE	P.DE	FDR
GO:0043231	intracellular membrane-bounded organelle	CC	9793	5069	9.07E-98	2.01E-93
GO:0044424	intracellular part	CC	12998	6354	7.22E-94	8.01E-90
GO:0005622	intracellular	CC	13234	6421	2.09E-87	1.54E-83
GO:0043229	intracellular organelle	CC	11413	5685	1.19E-85	6.58E-82
GO:0044446	intracellular organelle part	CC	8305	4338	6.26E-79	2.78E-75
GO:0043227	membrane-bounded organelle	CC	11374	5611	8.61E-75	3.19E-71
GO:0044422	organelle part	CC	8524	4407	6.02E-73	1.91E-69
GO:0043226	organelle	CC	12279	5957	1.74E-69	4.82E-66
GO:0031981	nuclear lumen	CC	3701	2185	9.07E-68	2.11E-64
GO:0044428	nuclear part	CC	4030	2345	9.53E-68	2.11E-64
GO:0005634	nucleus	CC	6501	3480	4.32E-67	8.71E-64
GO:0070013	intracellular organelle lumen	CC	4745	2666	1.04E-65	1.65E-62
GO:0031974	membrane-enclosed lumen	CC	4745	2666	1.04E-65	1.65E-62
GO:0043233	organelle lumen	CC	4745	2666	1.04E-65	1.65E-62
GO:0005654	nucleoplasm	CC	3170	1911	6.02E-65	8.91E-62
GO:0005737	cytoplasm	CC	10570	5172	9.21E-55	1.28E-51
GO:0044237	cellular metabolic process	BP	9902	4833	2.82E-48	3.68E-45
GO:0044444	cytoplasmic part	CC	8871	4396	3.44E-48	4.23E-45
GO:0006139	nucleobase-containing compound metabolic process	BP	5371	2838	1.92E-43	2.25E-40
GO:0005829	cytosol	CC	4663	2491	3.57E-43	3.96E-40

Blue module – ARIES – 15-17 years – KEGG pathway

	Pathway	N	DE	P.DE	FDR
path:hsa04140	Autophagy - animal	129	89	4.05E-07	8.76E-05
path:hsa04141	Protein processing in endoplasmic reticulum	152	98	5.23E-07	8.76E-05
path:hsa03015	mRNA surveillance pathway	77	57	2.12E-06	0.00021803
path:hsa04110	Cell cycle	120	81	2.60E-06	0.00021803
path:hsa04142	Lysosome	116	75	1.71E-05	0.00114327
path:hsa05220	Chronic myeloid leukemia	72	52	2.69E-05	0.0014007
path:hsa05203	Viral carcinogenesis	151	96	2.93E-05	0.0014007
path:hsa05166	Human T-cell leukemia virus 1 infection	189	115	4.15E-05	0.00173807
path:hsa03013	RNA transport	140	85	4.84E-05	0.00180013
path:hsa04714	Thermogenesis	209	121	0.00010235	0.00342863
path:hsa04120	Ubiquitin mediated proteolysis	129	81	0.00012453	0.00364404
path:hsa03040	Spliceosome	115	73	0.00013053	0.00364404
path:hsa04115	p53 signaling pathway	71	49	0.00015251	0.00393017
path:hsa05210	Colorectal cancer	83	56	0.00017349	0.00415143
path:hsa05010	Alzheimer disease	157	93	0.00023221	0.00518598
path:hsa03018	RNA degradation	72	47	0.00028199	0.00581696
path:hsa04071	Sphingolipid signaling pathway	115	72	0.00030821	0.00581696
path:hsa04218	Cellular senescence	143	88	0.00031255	0.00581696
path:hsa04932	Non-alcoholic fatty liver disease (NAFLD)	140	82	0.00042219	0.00738646
path:hsa04210	Apoptosis	128	76	0.00044098	0.00738646

Brown module – ARIES – 15-17 years – Gene ontology

	Term	Ont	N	DE	P.DE	FDR
GO:0044446	intracellular organelle part	CC	8305	494	5.02E-12	1.11E-07
GO:0044422	organelle part	CC	8524	502	1.01E-11	1.12E-07
GO:0043231	intracellular membrane-bounded organelle	CC	9793	553	3.03E-10	2.24E-06
GO:0043229	intracellular organelle	CC	11413	617	8.17E-10	4.53E-06
GO:0031981	nuclear lumen	CC	3701	257	7.92E-09	3.52E-05
GO:0043226	organelle	CC	12279	644	1.09E-08	4.02E-05
GO:0043227	membrane-bounded organelle	CC	11374	606	1.79E-08	5.41E-05
GO:0044428	nuclear part	CC	4030	272	1.95E-08	5.41E-05
GO:0044424	intracellular part	CC	12998	672	8.36E-08	0.00016845
GO:0070013	intracellular organelle lumen	CC	4745	302	9.11E-08	0.00016845
GO:0031974	membrane-enclosed lumen	CC	4745	302	9.11E-08	0.00016845
GO:0043233	organelle lumen	CC	4745	302	9.11E-08	0.00016845
GO:0032991	protein-containing complex	CC	4588	290	1.61E-07	0.00027449
GO:0005622	intracellular	CC	13234	678	3.03E-07	0.0004616
GO:0005829	cytosol	CC	4663	288	3.12E-07	0.0004616
GO:0044427	chromosomal part	CC	770	71	9.07E-07	0.00120947
GO:1901363	heterocyclic compound binding	MF	5520	327	9.27E-07	0.00120947
GO:0097159	organic cyclic compound binding	MF	5603	329	1.88E-06	0.00232205
GO:0005694	chromosome	CC	889	78	2.32E-06	0.00270936
GO:0005654	nucleoplasm	CC	3170	216	4.15E-06	0.00460556

Brown module – ARIES – 15-17 years – KEGG pathway

	Pathway	N	DE	P.DE	FDR
path:hsa05131	Shigellosis	62	9	0.00542058	1
path:hsa04660	T cell receptor signaling pathway	96	11	0.00713664	1
path:hsa05130	Pathogenic Escherichia coli infection	52	7	0.01056008	1
path:hsa04714	Thermogenesis	209	18	0.0133913	1
path:hsa00062	Fatty acid elongation	25	4	0.02958161	1
path:hsa05110	Vibrio cholerae infection	48	6	0.032044	1
path:hsa03040	Spliceosome	115	11	0.0346044	1
path:hsa01040	Biosynthesis of unsaturated fatty acids	26	4	0.04409008	1
path:hsa04659	Th17 cell differentiation	87	8	0.04621293	1
path:hsa03018	RNA degradation	72	7	0.05741593	1
path:hsa03320	PPAR signaling pathway	71	6	0.06092393	1
path:hsa01212	Fatty acid metabolism	53	6	0.06592539	1
path:hsa04658	Th1 and Th2 cell differentiation	74	7	0.06990417	1
path:hsa05135	Yersinia infection	112	10	0.06992003	1
path:hsa00520	Amino sugar and nucleotide sugar metabolism	47	5	0.08218753	1
path:hsa05418	Fluid shear stress and atherosclerosis	130	10	0.08994579	1
path:hsa00514	Other types of O-glycan biosynthesis	21	3	0.09865071	1
path:hsa05100	Bacterial invasion of epithelial cells	72	7	0.09882295	1
path:hsa04064	NF-kappa B signaling pathway	87	7	0.10059613	1
path:hsa03460	Fanconi anemia pathway	47	5	0.10369358	1

Cyan module – ARIES – 15-17 years – Gene ontology

	Term	Ont	N	DE	P.DE	FDR
GO:0001730	2'-5'-oligoadenylate synthetase activity	MF	4	3	1.58E-07	0.00350786
GO:0060700	regulation of ribonuclease activity	BP	8	3	7.95E-06	0.08815097
GO:0043901	negative regulation of multi-organism process	BP	149	7	1.45E-05	0.10695609
GO:1903901	negative regulation of viral life cycle	BP	66	5	2.62E-05	0.14523258
GO:0010033	response to organic substance	BP	2960	36	5.16E-05	0.17651038
GO:0050896	response to stimulus	BP	8077	70	5.34E-05	0.17651038
GO:0042221	response to chemical	BP	4202	44	6.13E-05	0.17651038
GO:0006955	immune response	BP	1761	23	6.75E-05	0.17651038
GO:0048525	negative regulation of viral process	BP	81	5	7.16E-05	0.17651038
GO:0015698	inorganic anion transport	BP	164	7	0.00010409	0.23091169
GO:0045071	negative regulation of viral genome replication	BP	49	4	0.00014004	0.25946729
GO:0071702	organic substance transport	BP	2480	30	0.00014036	0.25946729
GO:0002252	immune effector process	BP	1016	16	0.00015987	0.27279547
GO:0006820	anion transport	BP	555	12	0.00017977	0.27748009
GO:0034341	response to interferon-gamma	BP	156	6	0.00018763	0.27748009
GO:0045055	regulated exocytosis	BP	700	13	0.00020323	0.28176278
GO:0051707	response to other organism	BP	760	13	0.00023603	0.29359498
GO:0043207	response to external biotic stimulus	BP	761	13	0.00023823	0.29359498
GO:0002700	regulation of production of molecular mediator of immune response	BP	102	5	0.00026504	0.3094411
GO:0006887	exocytosis	BP	806	14	0.00029388	0.3259577

Cyan module – ARIES – 15-17 years – KEGG pathway

	Pathway	N	DE	P.DE	FDR
path:hsa05164	Influenza A	129	6	9.26E-05	0.0310353
path:hsa02010	ABC transporters	41	3	0.00204934	0.23551035
path:hsa05160	Hepatitis C	137	5	0.00210905	0.23551035
path:hsa00790	Folate biosynthesis	26	2	0.00642444	0.53804681
path:hsa04923	Regulation of lipolysis in adipocytes	54	3	0.00863474	0.53959883
path:hsa01523	Antifolate resistance	29	2	0.01050895	0.53959883
path:hsa05168	Herpes simplex virus 1 infection	415	6	0.0112752	0.53959883
path:hsa05146	Amoebiasis	95	3	0.02354376	0.89667289
path:hsa04973	Carbohydrate digestion and absorption	38	2	0.02709942	0.89667289
path:hsa04060	Cytokine-cytokine receptor interaction	255	4	0.02765552	0.89667289
path:hsa05321	Inflammatory bowel disease (IBD)	48	2	0.02944299	0.89667289
path:hsa05162	Measles	115	3	0.03447528	0.90380398
path:hsa04066	HIF-1 signaling pathway	103	3	0.03964538	0.90380398
path:hsa04714	Thermogenesis	209	4	0.03998134	0.90380398
path:hsa04621	NOD-like receptor signaling pathway	150	3	0.04046883	0.90380398
path:hsa04630	JAK-STAT signaling pathway	135	3	0.04460123	0.9338383
path:hsa04620	Toll-like receptor signaling pathway	82	2	0.07206383	1
path:hsa05161	Hepatitis B	140	3	0.07617209	1
path:hsa04976	Bile secretion	70	2	0.07657595	1
path:hsa05211	Renal cell carcinoma	64	2	0.07796726	1

Green module – ARIES – 15-17 years – Gene ontology

	Term	Ont	N	DE	P.DE	FDR
GO:0044446	intracellular organelle part	CC	8305	136	1.39E-06	0.01395633
GO:1990904	ribonucleoprotein complex	CC	707	25	2.36E-06	0.01395633
GO:0044428	nuclear part	CC	4030	81	2.59E-06	0.01395633
GO:0005654	nucleoplasm	CC	3170	68	4.19E-06	0.01395633
GO:0070013	intracellular organelle lumen	CC	4745	90	4.40E-06	0.01395633
GO:0031974	membrane-enclosed lumen	CC	4745	90	4.40E-06	0.01395633
GO:0043233	organelle lumen	CC	4745	90	4.40E-06	0.01395633
GO:0044422	organelle part	CC	8524	137	5.74E-06	0.01591993
GO:0031981	nuclear lumen	CC	3701	75	7.38E-06	0.01818195
GO:0032991	protein-containing complex	CC	4588	86	1.33E-05	0.02037025
GO:0043231	intracellular membrane-bounded organelle	CC	9793	151	1.37E-05	0.02037025
GO:0010468	regulation of gene expression	BP	4310	81	1.37E-05	0.02037025
GO:0006996	organelle organization	BP	3526	71	1.38E-05	0.02037025
GO:0010467	gene expression	BP	5109	92	1.39E-05	0.02037025
GO:0009892	negative regulation of metabolic process	BP	2759	58	1.42E-05	0.02037025
GO:0006139	nucleobase-containing compound metabolic process	BP	5371	96	1.62E-05	0.02037025
GO:0003676	nucleic acid binding	MF	3766	73	1.62E-05	0.02037025
GO:0051276	chromosome organization	BP	1079	31	1.73E-05	0.02037025
GO:0045934	negative regulation of nucleobase-containing compound metabolic process	BP	1277	35	1.74E-05	0.02037025
GO:0043229	intracellular organelle	CC	11413	168	1.84E-05	0.02041948

Green module – ARIES – 15-17 years – KEGG pathway

	Pathway	N	DE	P.DE	FDR
path:hsa05167	Kaposi sarcoma-associated herpesvirus infection	160	9	0.00016502	0.05528228
path:hsa05016	Huntington disease	179	9	0.0004276	0.07162285
path:hsa05010	Alzheimer disease	157	8	0.00082558	0.09218998
path:hsa04714	Thermogenesis	209	9	0.00130855	0.10959106
path:hsa03010	Ribosome	121	6	0.00356756	0.21892959
path:hsa05012	Parkinson disease	122	6	0.00392113	0.21892959
path:hsa00920	Sulfur metabolism	10	2	0.00609018	0.26391594
path:hsa04215	Apoptosis - multiple species	30	3	0.00630247	0.26391594
path:hsa04932	Non-alcoholic fatty liver disease (NAFLD)	140	6	0.00785449	0.27698247
path:hsa01524	Platinum drug resistance	66	4	0.00826813	0.27698247
path:hsa05416	Viral myocarditis	37	3	0.01185917	0.33672034
path:hsa00190	Oxidative phosphorylation	114	5	0.01206344	0.33672034
path:hsa05203	Viral carcinogenesis	151	6	0.01393318	0.33672034
path:hsa04115	p53 signaling pathway	71	4	0.01407189	0.33672034
path:hsa05168	Herpes simplex virus 1 infection	415	10	0.01707682	0.38138224
path:hsa04217	Necroptosis	128	5	0.0192404	0.40284583
path:hsa04210	Apoptosis	128	5	0.02088151	0.41148863
path:hsa05145	Toxoplasmosis	89	4	0.02616814	0.47319668
path:hsa04140	Autophagy - animal	129	5	0.02686936	0.47319668
path:hsa05152	Tuberculosis	142	5	0.02825055	0.47319668

Light cyan module – ARIES – 15-17 years - Gene ontology

	Term	Ont	N	DE	P.DE	FDR
GO:0003723	RNA binding	MF	1657	53	1.94E-09	4.31E-05
GO:0044422	organelle part	CC	8524	155	2.40E-07	0.00224176
GO:0044446	intracellular organelle part	CC	8305	152	3.03E-07	0.00224176
GO:0006725	cellular aromatic compound metabolic process	BP	5573	111	2.52E-06	0.01113377
GO:0003676	nucleic acid binding	MF	3766	82	2.88E-06	0.01113377
GO:0043229	intracellular organelle	CC	11413	186	3.29E-06	0.01113377
GO:0034641	cellular nitrogen compound metabolic process	BP	6161	118	3.56E-06	0.01113377
GO:0016071	mRNA metabolic process	BP	749	28	4.02E-06	0.01113377
GO:0046483	heterocycle metabolic process	BP	5533	109	7.81E-06	0.0179254
GO:0036002	pre-mRNA binding	MF	31	6	8.08E-06	0.0179254
GO:0043227	membrane-bounded organelle	CC	11374	184	9.52E-06	0.01919613
GO:0006139	nucleobase-containing compound metabolic process	BP	5371	106	1.42E-05	0.02629274
GO:0000228	nuclear chromosome	CC	505	21	1.65E-05	0.02702518
GO:1901360	organic cyclic compound metabolic process	BP	5779	111	1.77E-05	0.02702518
GO:0005694	chromosome	CC	889	30	1.83E-05	0.02702518
GO:0044454	nuclear chromosome part	CC	474	20	2.24E-05	0.03104585
GO:0090304	nucleic acid metabolic process	BP	4784	96	3.51E-05	0.04358619
GO:0033554	cellular response to stress	BP	1817	47	3.54E-05	0.04358619
GO:0006974	cellular response to DNA damage stimulus	BP	748	26	4.15E-05	0.04849272
GO:0043226	organelle	CC	12279	192	4.75E-05	0.05273845

Light cyan module – ARIES – 15-17 years – KEGG pathway

	Pathway	N	DE	P.DE	FDR
path:hsa04932	Non-alcoholic fatty liver disease (NAFLD)	140	8	0.00073058	0.2331857
path:hsa03010	Ribosome	121	7	0.00156308	0.2331857
path:hsa05169	Epstein-Barr virus infection	158	8	0.00208823	0.2331857
path:hsa03050	Proteasome	43	4	0.00285563	0.23915894
path:hsa00190	Oxidative phosphorylation	114	6	0.00413278	0.27689605
path:hsa04068	FoxO signaling pathway	121	6	0.00666242	0.30338517
path:hsa05012	Parkinson disease	122	6	0.00669286	0.30338517
path:hsa04110	Cell cycle	120	6	0.00827591	0.30338517
path:hsa04211	Longevity regulating pathway	86	5	0.00876972	0.30338517
path:hsa04136	Autophagy - other	29	3	0.00905627	0.30338517
path:hsa05215	Prostate cancer	93	5	0.01292785	0.36137966
path:hsa05016	Huntington disease	179	7	0.01294494	0.36137966
path:hsa05161	Hepatitis B	140	6	0.01481988	0.38189688
path:hsa05203	Viral carcinogenesis	151	6	0.02377866	0.53384235
path:hsa05010	Alzheimer disease	157	6	0.02390339	0.53384235
path:hsa04714	Thermogenesis	209	7	0.02699399	0.55946882
path:hsa05134	Legionellosis	52	3	0.02965269	0.55946882
path:hsa03040	Spliceosome	115	5	0.03006101	0.55946882
path:hsa00670	One carbon pool by folate	19	2	0.033184	0.57497573
path:hsa03460	Fanconi anemia pathway	47	3	0.03482869	0.57497573

Light green module – ARIES – 15-17 years – Gene ontology

	Term	Ont	N	DE	P.DE	FDR
GO:0032673	regulation of interleukin-4 production	BP	23	4	3.27E-09	7.26E-05
GO:0032633	interleukin-4 production	BP	29	4	9.83E-09	0.00010902
GO:0032753	positive regulation of interleukin-4 production	BP	18	3	4.14E-07	0.0030609
GO:0002215	defense response to nematode	BP	3	2	1.46E-06	0.00808215
GO:0032653	regulation of interleukin-10 production	BP	40	3	3.26E-06	0.01332138
GO:0032613	interleukin-10 production	BP	42	3	3.60E-06	0.01332138
GO:0032693	negative regulation of interleukin-10 production	BP	16	2	6.88E-05	0.21801931
GO:0001819	positive regulation of cytokine production	BP	375	4	0.00013754	0.3315192
GO:0032674	regulation of interleukin-5 production	BP	18	2	0.00013781	0.3315192
GO:0032634	interleukin-5 production	BP	19	2	0.00014945	0.3315192
GO:0001818	negative regulation of cytokine production	BP	223	3	0.00053093	0.8994218
GO:0051187	cofactor catabolic process	BP	53	2	0.00071094	0.8994218
GO:0001817	regulation of cytokine production	BP	578	4	0.00072682	0.8994218
GO:1905937	negative regulation of germ cell proliferation	BP	2	1	0.00080026	0.8994218
GO:2000255	negative regulation of male germ cell proliferation	BP	2	1	0.00080026	0.8994218
GO:0032468	Golgi calcium ion homeostasis	BP	1	1	0.00084025	0.8994218
GO:0004105	choline-phosphate cytidylyltransferase activity	MF	1	1	0.00086153	0.8994218
GO:0002366	leukocyte activation involved in immune response	BP	629	4	0.00087278	0.8994218
GO:0002263	cell activation involved in immune response	BP	633	4	0.00089901	0.8994218
GO:0042093	T-helper cell differentiation	BP	50	2	0.00091209	0.8994218

Light green module – ARIES – 15-17 years – KEGG pathway

	Pathway	N	DE	P.DE	FDR
path:hsa05310	Asthma	14	2	2.15E-05	0.00721414
path:hsa00440	Phosphonate and phosphinate metabolism	5	1	0.00423273	0.70898291
path:hsa04530	Tight junction	162	2	0.00928668	0.95766864
path:hsa04390	Hippo signaling pathway	150	2	0.01143485	0.95766864
path:hsa00062	Fatty acid elongation	25	1	0.01803988	0.98009773
path:hsa04144	Endocytosis	227	2	0.01878607	0.98009773
path:hsa01040	Biosynthesis of unsaturated fatty acids	26	1	0.02195191	0.98009773
path:hsa05165	Human papillomavirus infection	300	2	0.03462873	0.98009773
path:hsa04340	Hedgehog signaling pathway	46	1	0.04843048	0.98009773
path:hsa04930	Type II diabetes mellitus	44	1	0.05095294	0.98009773
path:hsa04137	Mitophagy - animal	60	1	0.05273974	0.98009773
path:hsa04924	Renin secretion	67	1	0.053927	0.98009773
path:hsa04929	GnRH secretion	60	1	0.06220615	0.98009773
path:hsa04927	Cortisol synthesis and secretion	60	1	0.06238587	0.98009773
path:hsa05218	Melanoma	68	1	0.06972502	0.98009773
path:hsa04260	Cardiac muscle contraction	78	1	0.0698896	0.98009773
path:hsa04350	TGF-beta signaling pathway	92	1	0.07845705	0.98009773
path:hsa00564	Glycerophospholipid metabolism	91	1	0.07897028	0.98009773
path:hsa04912	GnRH signaling pathway	88	1	0.07904236	0.98009773
path:hsa05412	Arrhythmogenic right ventricular cardiomyopathy (ARVC)	74	1	0.08114217	0.98009773

Magenta module – ARIES – 15-17 years – Gene ontology

	Term	Ont	N	DE	P.DE	FDR
GO:0044428	nuclear part	CC	4030	41	7.92E-07	0.01756811
GO:0005654	nucleoplasm	CC	3170	34	7.22E-06	0.0800448
GO:0044271	cellular nitrogen compound biosynthetic process	BP	4785	41	1.69E-05	0.12032395
GO:0031981	nuclear lumen	CC	3701	36	2.29E-05	0.12032395
GO:0034654	nucleobase-containing compound biosynthetic process	BP	4034	37	2.75E-05	0.12032395
GO:0019438	aromatic compound biosynthetic process	BP	4102	37	3.75E-05	0.12032395
GO:0018130	heterocycle biosynthetic process	BP	4094	37	3.80E-05	0.12032395
GO:0031326	regulation of cellular biosynthetic process	BP	4149	36	7.03E-05	0.18735354
GO:1901362	organic cyclic compound biosynthetic process	BP	4241	37	7.70E-05	0.18735354
GO:0006139	nucleobase-containing compound metabolic process	BP	5371	43	8.45E-05	0.18735354
GO:0009889	regulation of biosynthetic process	BP	4213	36	0.00010261	0.20693459
GO:0031328	positive regulation of cellular biosynthetic process	BP	1705	21	0.00013329	0.20854243
GO:2000112	regulation of cellular macromolecule biosynthetic process	BP	3908	34	0.00014109	0.20854243
GO:0009891	positive regulation of biosynthetic process	BP	1730	21	0.00015398	0.20854243
GO:0005634	nucleus	CC	6501	48	0.00016676	0.20854243
GO:0046483	heterocycle metabolic process	BP	5533	43	0.00016799	0.20854243
GO:0010557	positive regulation of macromolecule biosynthetic process	BP	1598	20	0.00017423	0.20854243
GO:0006725	cellular aromatic compound metabolic process	BP	5573	43	0.00018818	0.20854243
GO:0032991	protein-containing complex	CC	4588	38	0.00018908	0.20854243
GO:0010556	regulation of macromolecule biosynthetic process	BP	3988	34	0.00020145	0.20854243

Magenta module – ARIES – 15-17 years – KEGG pathway

	Pathway	N	DE	P.DE	FDR
path:hsa00310	Lysine degradation	55	2	0.03686129	1
path:hsa04137	Mitophagy - animal	60	2	0.04056206	1
path:hsa00730	Thiamine metabolism	15	1	0.07936462	1
path:hsa04068	FoxO signaling pathway	121	2	0.11939262	1
path:hsa04621	NOD-like receptor signaling pathway	150	2	0.12535055	1
path:hsa04110	Cell cycle	120	2	0.12696407	1
path:hsa04612	Antigen processing and presentation	37	1	0.13217645	1
path:hsa03013	RNA transport	140	2	0.13307182	1
path:hsa00230	Purine metabolism	123	2	0.13537954	1
path:hsa04710	Circadian rhythm	30	1	0.1455073	1
path:hsa04136	Autophagy - other	29	1	0.14626798	1
path:hsa04141	Protein processing in endoplasmic reticulum	152	2	0.16113497	1
path:hsa04218	Cellular senescence	143	2	0.17354994	1
path:hsa03440	Homologous recombination	39	1	0.17716804	1
path:hsa03420	Nucleotide excision repair	39	1	0.18810625	1
path:hsa05169	Epstein-Barr virus infection	158	2	0.20066176	1
path:hsa04150	mTOR signaling pathway	146	2	0.20725253	1
path:hsa05202	Transcriptional misregulation in cancer	157	2	0.20983122	1
path:hsa04310	Wnt signaling pathway	152	2	0.21305902	1
path:hsa05170	Human immunodeficiency virus 1 infection	179	2	0.22025312	1

Salmon module – ARIES – 15-17 years – Gene ontology

	Term	Ont	N	DE	P.DE	FDR
GO:0046649	lymphocyte activation	BP	564	9	1.78E-08	0.00018903
GO:0042110	T cell activation	BP	404	8	2.51E-08	0.00018903
GO:0050778	positive regulation of immune response	BP	643	9	3.58E-08	0.00018903
GO:0002757	immune response-activating signal transduction	BP	439	8	4.27E-08	0.00018903
GO:0050852	T cell receptor signaling pathway	BP	163	6	4.64E-08	0.00018903
GO:0007159	leukocyte cell-cell adhesion	BP	293	7	5.11E-08	0.00018903
GO:0002764	immune response-regulating signaling pathway	BP	468	8	6.78E-08	0.00021487
GO:0050865	regulation of cell activation	BP	467	8	8.34E-08	0.00023128
GO:0002253	activation of immune response	BP	505	8	1.10E-07	0.00027167
GO:0006955	immune response	BP	1761	12	1.62E-07	0.00034869
GO:0050851	antigen receptor-mediated signaling pathway	BP	201	6	1.73E-07	0.00034869
GO:0050776	regulation of immune response	BP	849	9	3.12E-07	0.00057619
GO:0030217	T cell differentiation	BP	216	6	3.82E-07	0.00065156
GO:0002684	positive regulation of immune system process	BP	901	9	6.79E-07	0.00101089
GO:1903037	regulation of leukocyte cell-cell adhesion	BP	261	6	6.84E-07	0.00101089
GO:0002694	regulation of leukocyte activation	BP	430	7	8.15E-07	0.00112969
GO:0002250	adaptive immune response	BP	311	6	1.01E-06	0.00124822
GO:0002696	positive regulation of leukocyte activation	BP	270	6	1.01E-06	0.00124822
GO:0050863	regulation of T cell activation	BP	281	6	1.09E-06	0.00127735
GO:0050867	positive regulation of cell activation	BP	281	6	1.26E-06	0.00139782

Salmon module – ARIES – 15-17 years – KEGG pathway

	Pathway	N	DE	P.DE	FDR
path:hsa04658	Th1 and Th2 cell differentiation	74	5	5.63E-08	1.89E-05
path:hsa05235	PD-L1 expression and PD-1 checkpoint pathway in cancer	85	4	5.10E-06	0.00043861
path:hsa04659	Th17 cell differentiation	87	4	5.49E-06	0.00043861
path:hsa05340	Primary immunodeficiency	30	3	5.79E-06	0.00043861
path:hsa04650	Natural killer cell mediated cytotoxicity	93	4	6.55E-06	0.00043861
path:hsa04660	T cell receptor signaling pathway	96	4	8.46E-06	0.00047218
path:hsa05166	Human T-cell leukemia virus 1 infection	189	4	0.0001418	0.006786
path:hsa05142	Chagas disease (American trypanosomiasis)	94	3	0.00025157	0.01053453
path:hsa05162	Measles	115	3	0.00041634	0.01549722
path:hsa05169	Epstein-Barr virus infection	158	3	0.00126172	0.04226775
path:hsa05170	Human immunodeficiency virus 1 infection	179	3	0.00166659	0.05075531
path:hsa05133	Pertussis	69	2	0.00265624	0.07415342
path:hsa04640	Hematopoietic cell lineage	76	2	0.00306178	0.07889982
path:hsa04064	NF-kappa B signaling pathway	87	2	0.00544847	0.13037412
path:hsa04670	Leukocyte transendothelial migration	104	2	0.00917002	0.20479704
path:hsa05165	Human papillomavirus infection	300	3	0.01040977	0.2177161
path:hsa05135	Yersinia infection	112	2	0.01104828	0.2177161
path:hsa05332	Graft-versus-host disease	15	1	0.01253235	0.23324096
path:hsa05330	Allograft rejection	15	1	0.01362663	0.24025902
path:hsa05320	Autoimmune thyroid disease	21	1	0.01972424	0.33038098

Turquoise module – ARIES – 15-17 years – Gene ontology

	Term	Ont	N	DE	P.DE	FDR
GO:0044425	membrane part	CC	5997	1069	9.74E-09	0.00019147
GO:0071944	cell periphery	CC	4868	903	2.19E-08	0.00019147
GO:0031224	intrinsic component of membrane	CC	4837	868	2.76E-08	0.00019147
GO:0005575	cellular_component	CC	16397	2639	3.45E-08	0.00019147
GO:0005886	plasma membrane	CC	4765	881	5.47E-08	0.00020778
GO:0016021	integral component of membrane	CC	4713	844	5.62E-08	0.00020778
GO:0007155	cell adhesion	BP	1256	277	2.71E-07	0.0008546
GO:0022610	biological adhesion	BP	1263	278	3.08E-07	0.0008546
GO:0016020	membrane	CC	8277	1423	1.57E-06	0.0038626
GO:0003008	system process	BP	1892	371	3.50E-06	0.00776851
GO:0000902	cell morphogenesis	BP	932	219	5.87E-06	0.01184569
GO:0030030	cell projection organization	BP	1388	299	8.24E-06	0.01523524
GO:0032989	cellular component morphogenesis	BP	1034	236	1.01E-05	0.0172588
GO:0048812	neuron projection morphogenesis	BP	586	149	1.39E-05	0.02199463
GO:0051865	protein autoubiquitination	BP	55	22	1.71E-05	0.02534026
GO:0120036	plasma membrane bounded cell projection organization	BP	1356	291	1.91E-05	0.02611964
GO:0120039	plasma membrane bounded cell projection morphogenesis	BP	600	151	2.00E-05	0.02611964
GO:0048858	cell projection morphogenesis	BP	602	151	2.28E-05	0.0280859
GO:0048666	neuron development	BP	992	226	2.79E-05	0.0326212
GO:0032990	cell part morphogenesis	BP	620	153	4.05E-05	0.0448945

Turquoise module – ARIES – 15-17 years – KEGG pathway

	Pathway	N	DE	P.DE	FDR
path:hsa04512	ECM-receptor interaction	85	29	0.00017564	0.05884008
path:hsa04740	Olfactory transduction	343	68	0.00150101	0.17742434
path:hsa04151	PI3K-Akt signaling pathway	326	77	0.00158887	0.17742434
path:hsa05100	Bacterial invasion of epithelial cells	72	22	0.00647455	0.48904933
path:hsa04510	Focal adhesion	190	47	0.00964849	0.48904933
path:hsa03008	Ribosome biogenesis in eukaryotes	65	17	0.01102706	0.48904933
path:hsa04152	AMPK signaling pathway	117	31	0.01228034	0.48904933
path:hsa04514	Cell adhesion molecules (CAMs)	116	29	0.0127658	0.48904933
path:hsa03320	PPAR signaling pathway	71	18	0.01313864	0.48904933
path:hsa04520	Adherens junction	69	20	0.01906889	0.60108088
path:hsa04080	Neuroactive ligand-receptor interaction	310	61	0.01973698	0.60108088
path:hsa04923	Regulation of lipolysis in adipocytes	54	16	0.02227973	0.62197592
path:hsa05165	Human papillomavirus infection	300	66	0.02560945	0.65993586
path:hsa00534	Glycosaminoglycan biosynthesis - heparan sulfate / heparin	23	8	0.03177239	0.76026797
path:hsa04742	Taste transduction	77	18	0.03727277	0.83242511
path:hsa05146	Amoebiasis	95	22	0.04603436	0.96384432
path:hsa04120	Ubiquitin mediated proteolysis	129	28	0.04943548	0.97115294
path:hsa05222	Small cell lung cancer	86	21	0.0572639	0.97115294
path:hsa04072	Phospholipase D signaling pathway	142	33	0.06069357	0.97115294
path:hsa00230	Purine metabolism	123	27	0.07035196	0.97115294

Yellow module – ARIES – 15-17 years – Gene ontology

	Term	Ont	N	DE	P.DE	FDR
GO:0022603	regulation of anatomical structure morphogenesis	BP	1041	35	9.45E-09	0.00020968
GO:0045765	regulation of angiogenesis	BP	335	13	3.09E-05	0.34284422
GO:1901342	regulation of vasculature development	BP	365	13	8.17E-05	0.41008043
GO:0050793	regulation of developmental process	BP	2358	46	8.42E-05	0.41008043
GO:0048646	anatomical structure formation involved in morphogenesis	BP	1069	27	9.28E-05	0.41008043
GO:0072359	circulatory system development	BP	1037	26	0.00011092	0.41008043
GO:2000026	regulation of multicellular organismal development	BP	1852	38	0.00019613	0.62153964
GO:0051960	regulation of nervous system development	BP	778	23	0.00023988	0.6320212
GO:0009653	anatomical structure morphogenesis	BP	2494	48	0.00025642	0.6320212
GO:0050770	regulation of axonogenesis	BP	154	9	0.00036863	0.78922598
GO:0051962	positive regulation of nervous system development	BP	462	16	0.00039136	0.78922598
GO:0051094	positive regulation of developmental process	BP	1236	28	0.00043666	0.80719426
GO:0022604	regulation of cell morphogenesis	BP	435	15	0.00056077	0.84974139
GO:0090335	regulation of brown fat cell differentiation	BP	12	3	0.00058467	0.84974139
GO:0045766	positive regulation of angiogenesis	BP	180	8	0.0006196	0.84974139
GO:0001525	angiogenesis	BP	528	15	0.00066979	0.84974139
GO:0045664	regulation of neuron differentiation	BP	566	18	0.00068919	0.84974139
GO:0060284	regulation of cell development	BP	812	22	0.00068951	0.84974139
GO:0051239	regulation of multicellular organismal process	BP	2744	47	0.00082215	0.9019198
GO:0010769	regulation of cell morphogenesis involved in differentiation	BP	252	11	0.00085038	0.9019198

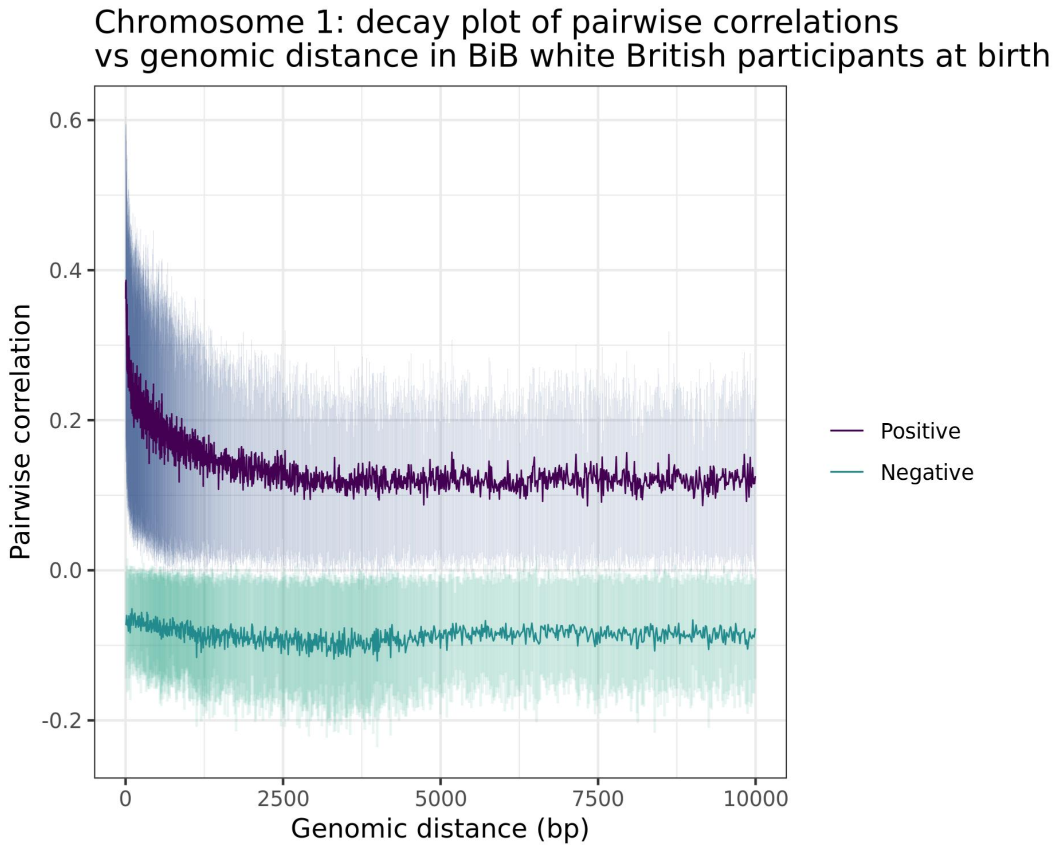
Yellow module – ARIES – 15-17 years – KEGG pathway

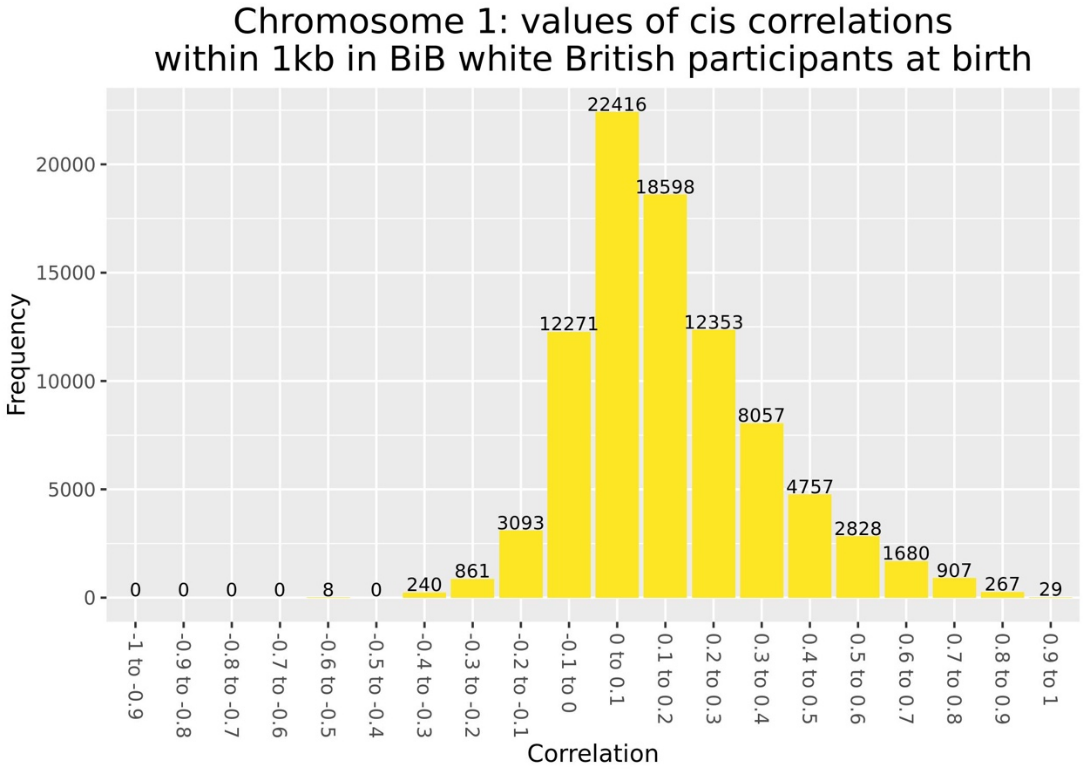
	Pathway	N	DE	P.DE	FDR
path:hsa05110	Vibrio cholerae infection	48	3	0.01412622	1
path:hsa04940	Type I diabetes mellitus	21	2	0.01862911	1
path:hsa04145	Phagosome	116	4	0.01890409	1
path:hsa05410	Hypertrophic cardiomyopathy (HCM)	86	4	0.01995691	1
path:hsa04966	Collecting duct acid secretion	26	2	0.02410875	1
path:hsa05414	Dilated cardiomyopathy (DCM)	91	4	0.02870722	1
path:hsa04931	Insulin resistance	104	4	0.03138465	1
path:hsa04260	Cardiac muscle contraction	78	3	0.04429927	1
path:hsa00330	Arginine and proline metabolism	44	2	0.04522319	1
path:hsa04010	MAPK signaling pathway	276	7	0.05428021	1
path:hsa04390	Hippo signaling pathway	150	5	0.05490034	1
path:hsa04371	Apelin signaling pathway	134	4	0.06571973	1
path:hsa05140	Leishmaniasis	50	2	0.06619841	1
path:hsa04512	ECM-receptor interaction	85	3	0.08891697	1
path:hsa04261	Adrenergic signaling in cardiomyocytes	141	4	0.09194572	1
path:hsa05203	Viral carcinogenesis	151	4	0.10004831	1
path:hsa00430	Taurine and hypotaurine metabolism	11	1	0.10459284	1
path:hsa00120	Primary bile acid biosynthesis	17	1	0.1046405	1
path:hsa05323	Rheumatoid arthritis	72	2	0.11055533	1
path:hsa04022	cGMP-PKG signaling pathway	159	4	0.11461195	1

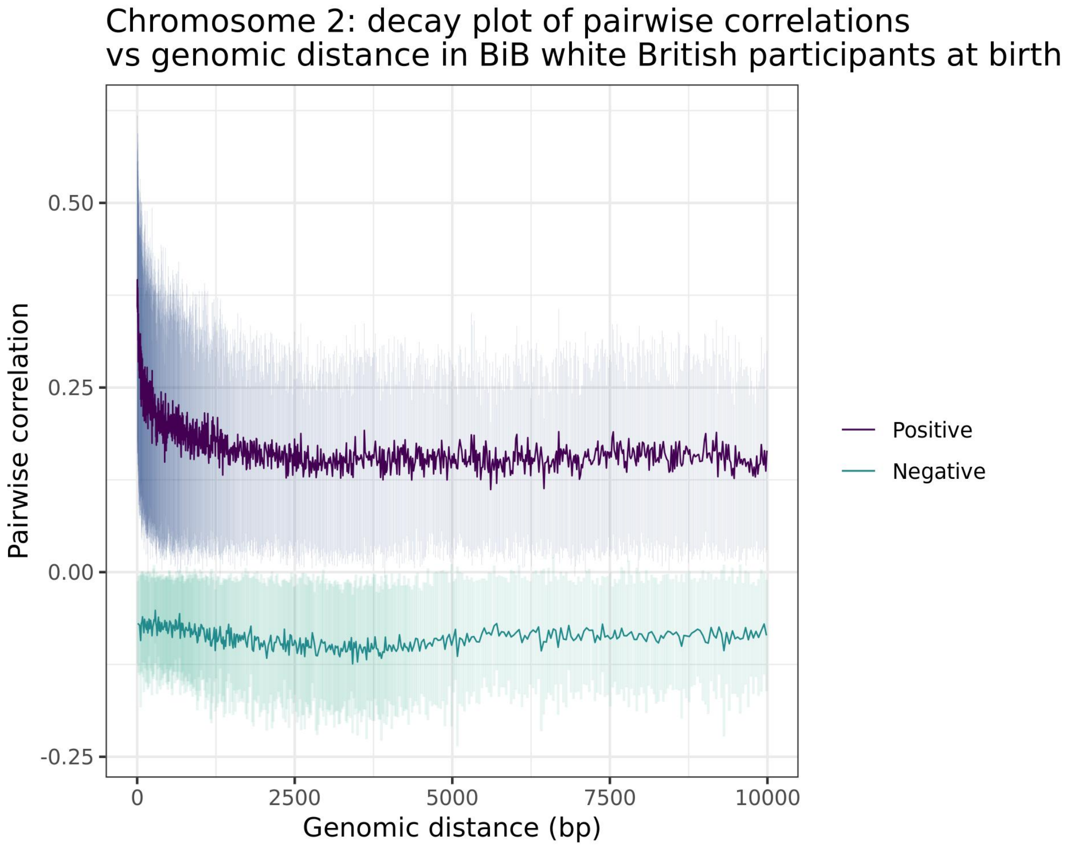
Appendix 5

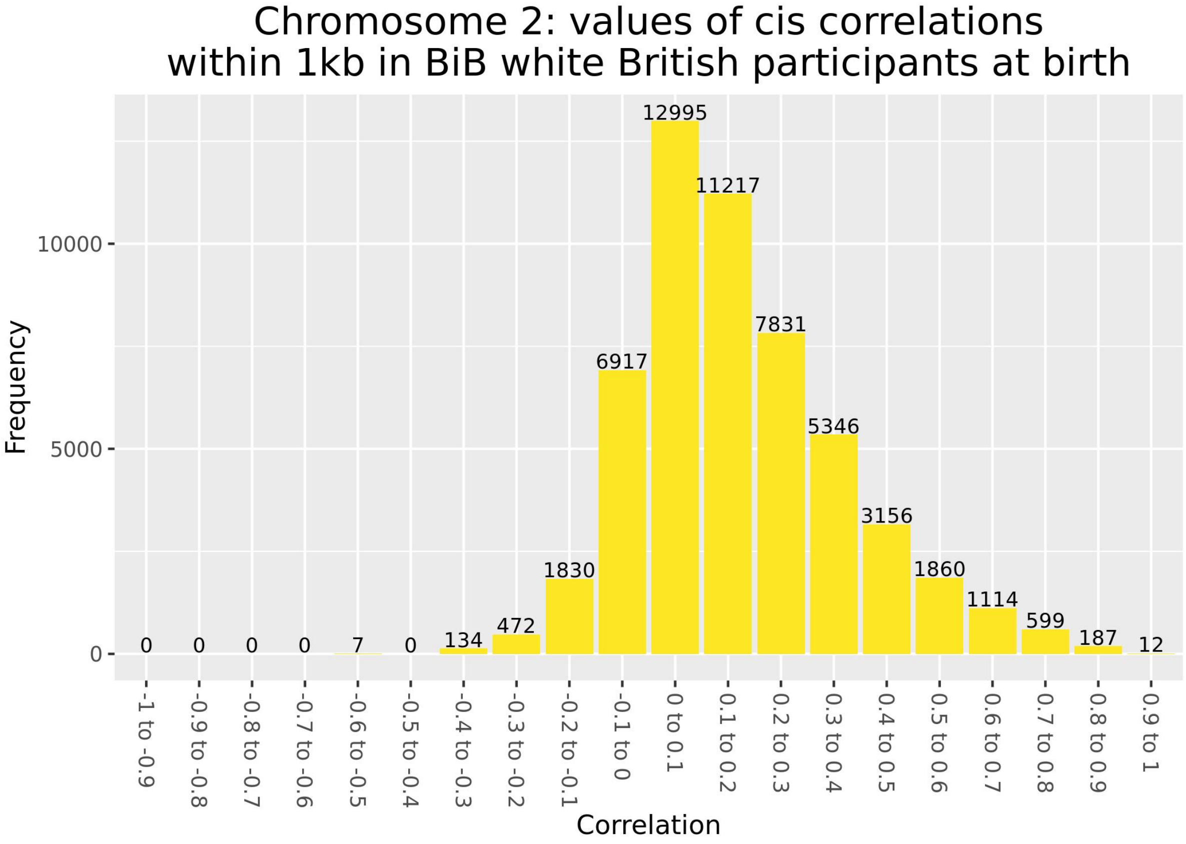
Cis correlation plots for chromosomes 1:5 and 15:19, for Born in Bradford.

White British ethnic group

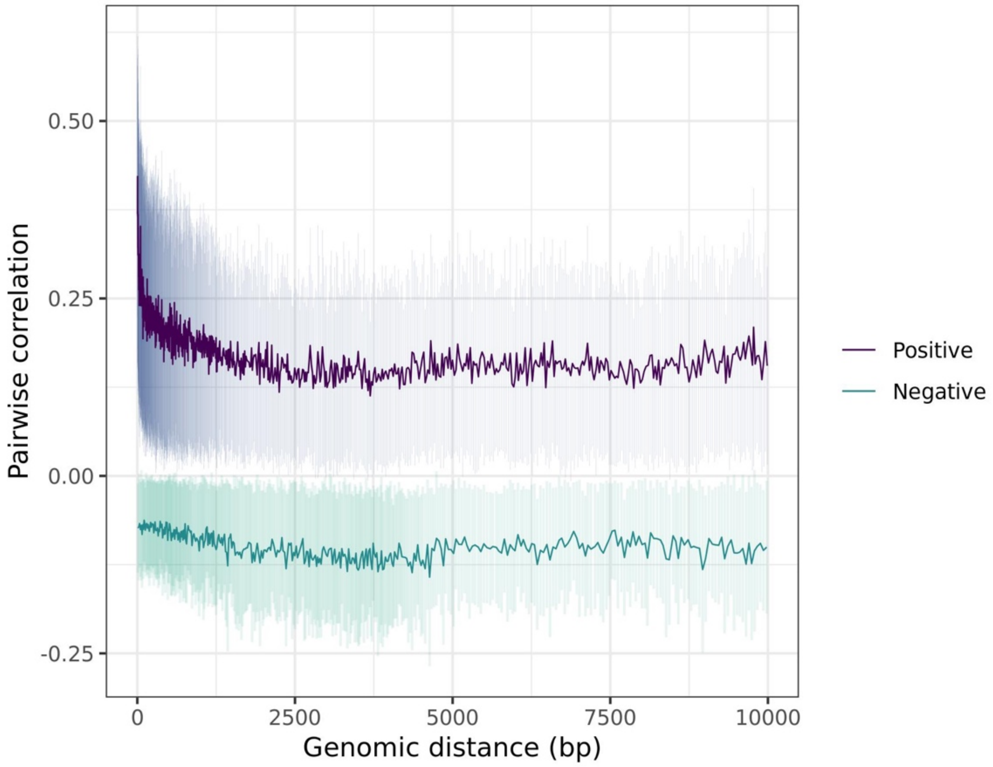




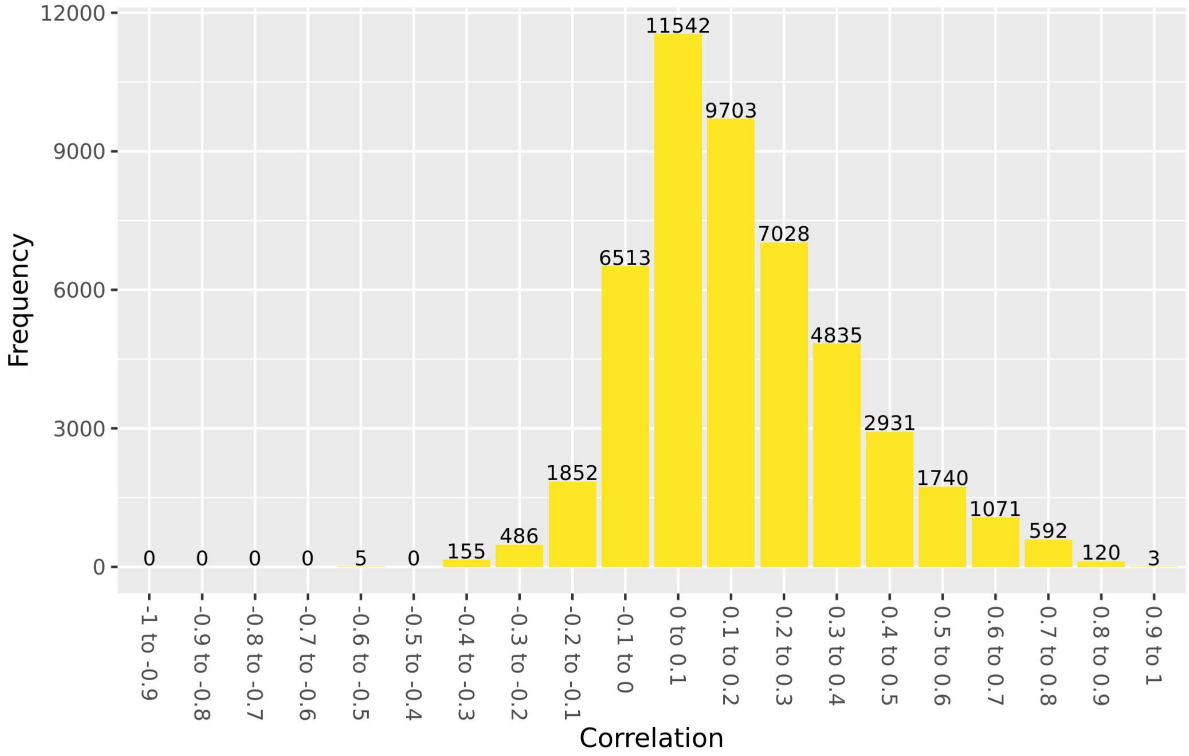

Chromosome 2: decay plot of pairwise correlations
vs genomic distance in BiB white British participants at birth
Pairwise correlation
0.50
0.25
0.00
-0.25
0
2500
5000
7500
10000
Genomic distance (bp)
Positive
Negative

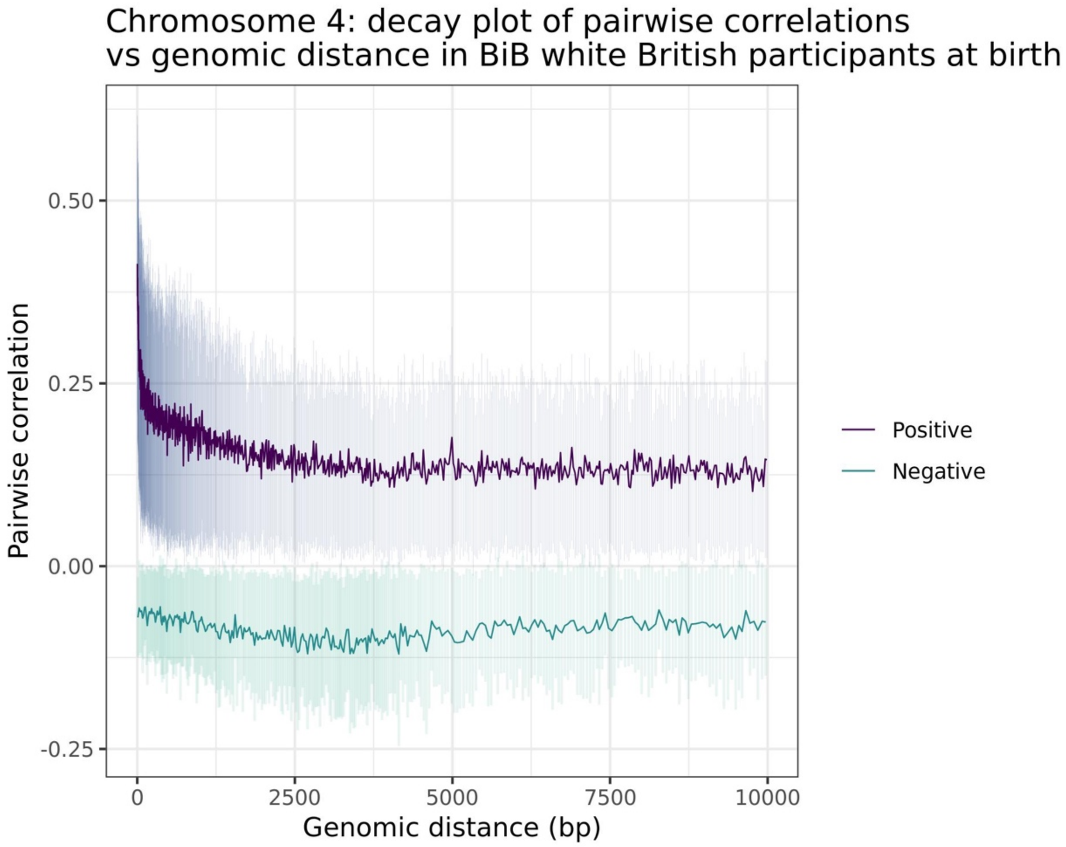

Chromosome 2: values of cis correlations
within 1kb in BiB white British participants at birth
Frequency
10000
5000
0
12995
11217
7831
6917
5346
3156
1860
1830
1114
599
472
187
134
12
7
0
0
0
0
0
-1 to -0.9
-0.9 to -0.8
-0.8 to -0.7
-0.7 to -0.6
-0.6 to -0.5
-0.5 to -0.4
-0.4 to -0.3
-0.3 to -0.2
-0.2 to -0.1
-0.1 to 0
0 to 0.1
0.1 to 0.2
0.2 to 0.3
0.3 to 0.4
0.4 to 0.5
0.5 to 0.6
0.6 to 0.7
0.7 to 0.8
0.8 to 0.9
0.9 to 1
Correlation

Chromosome 3: decay plot of pairwise correlations
vs genomic distance in BiB white British participants at birth

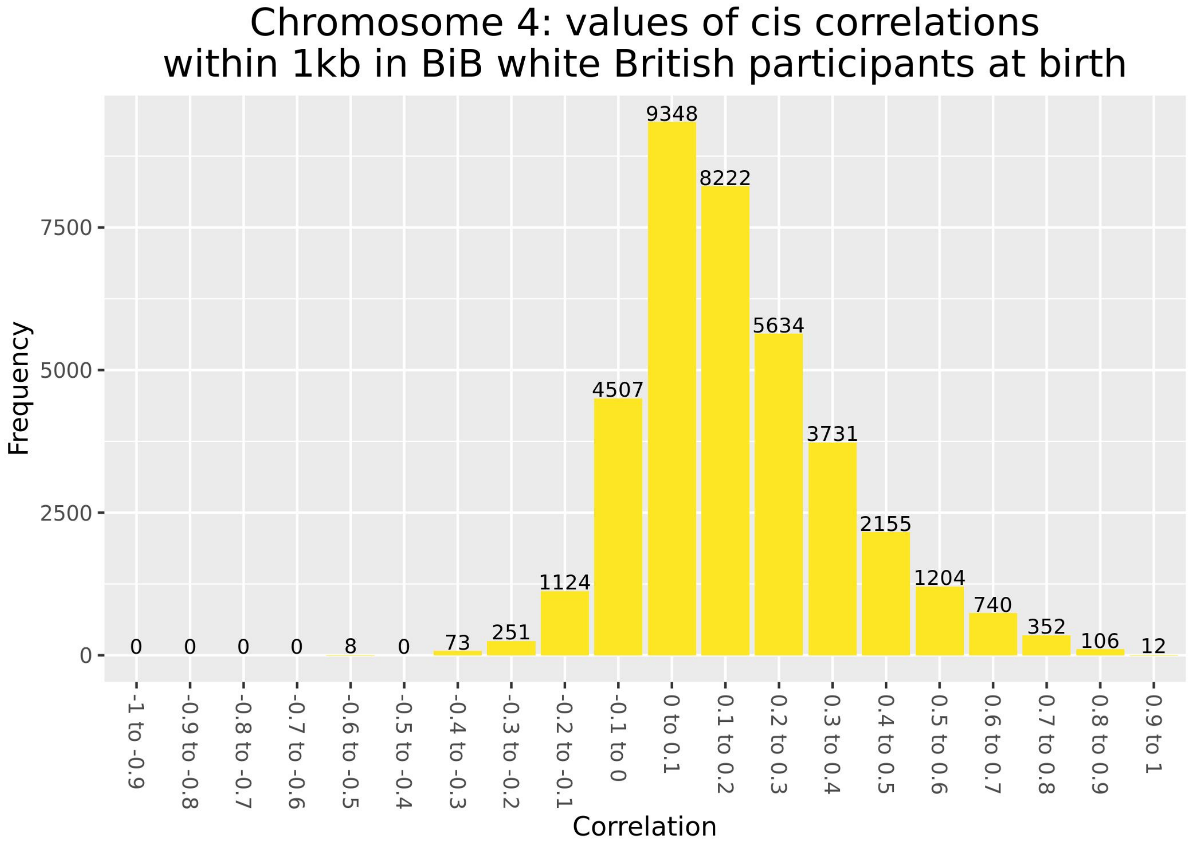


Chromosome 3: values of cis correlations
within 1kb in BiB white British participants at birth

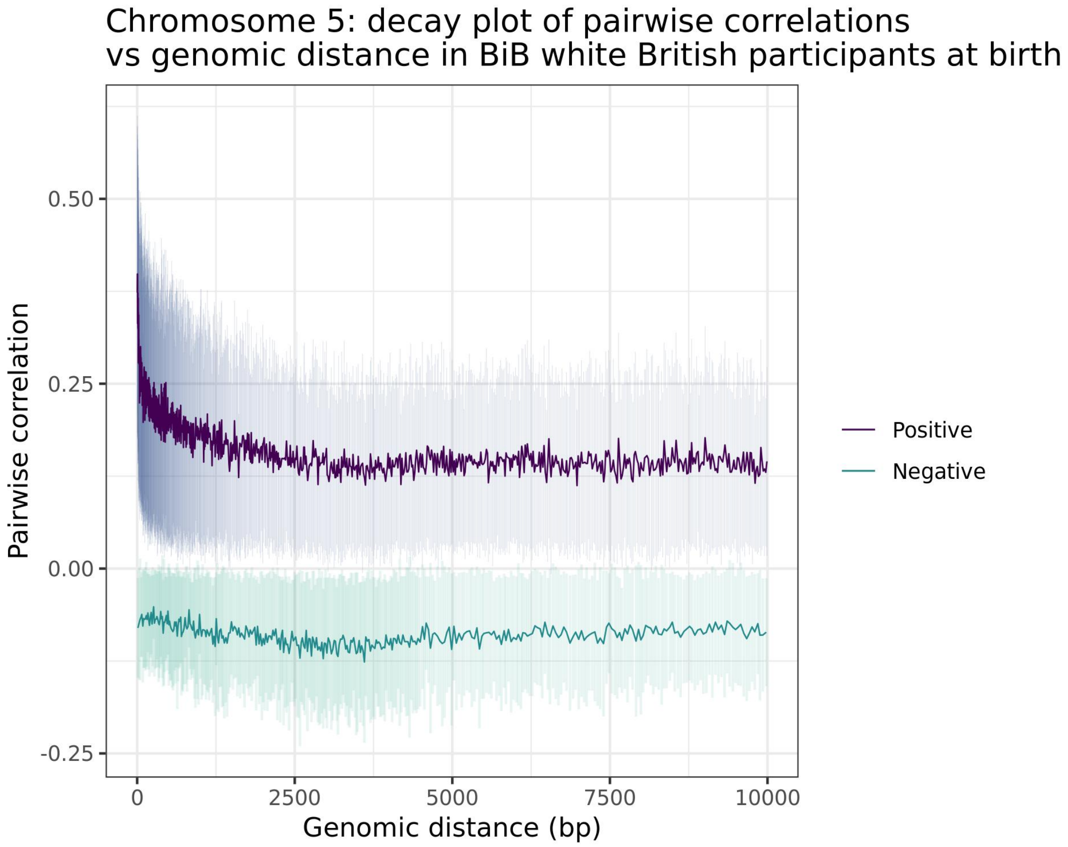


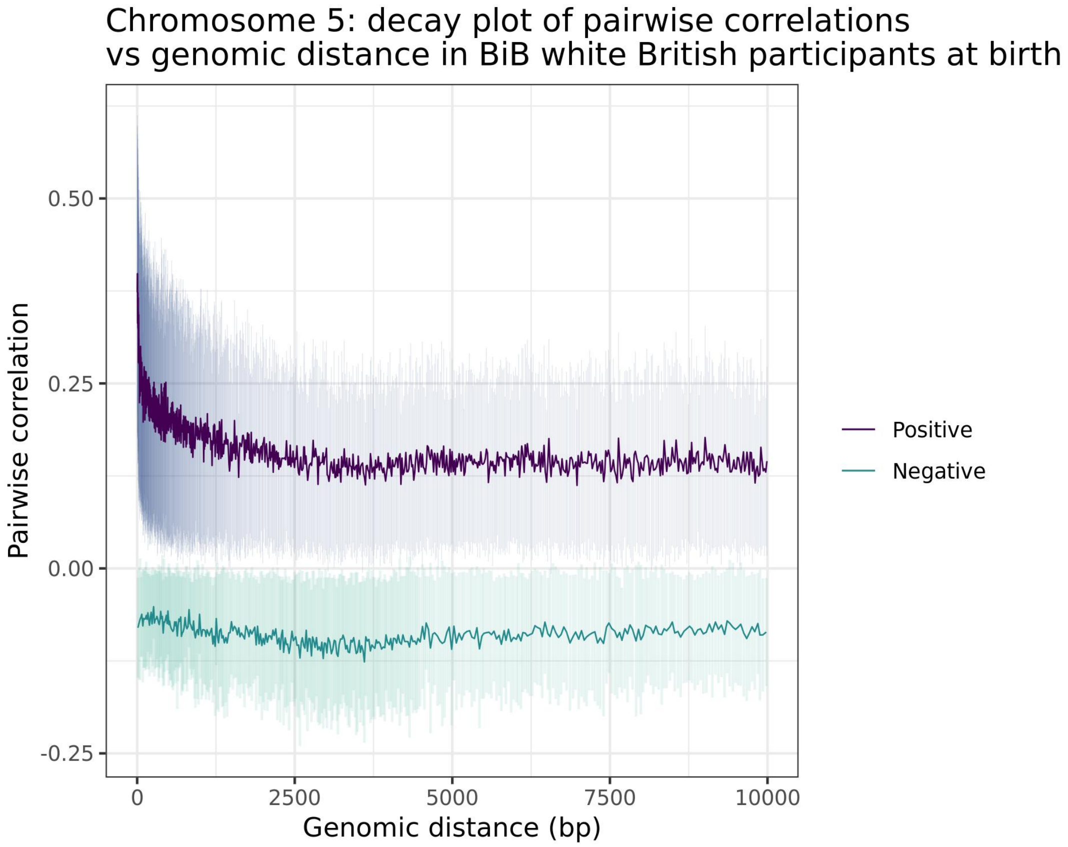


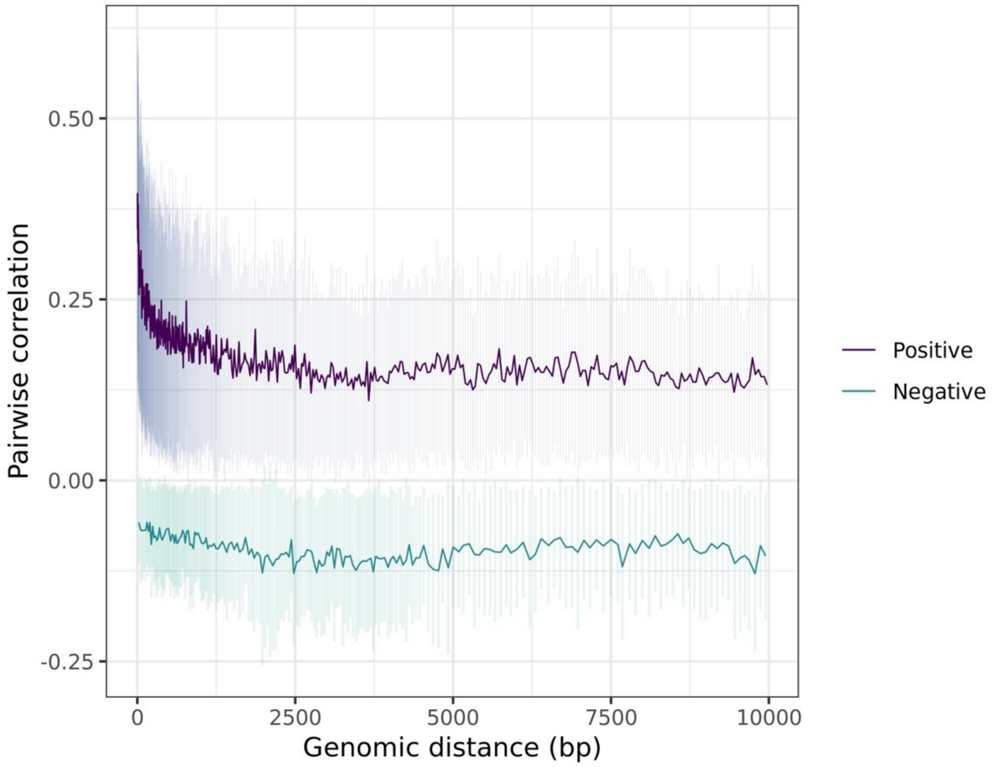
Chromosome 4: decay plot of pairwise correlations
vs genomic distance in BiB white British participants at birth



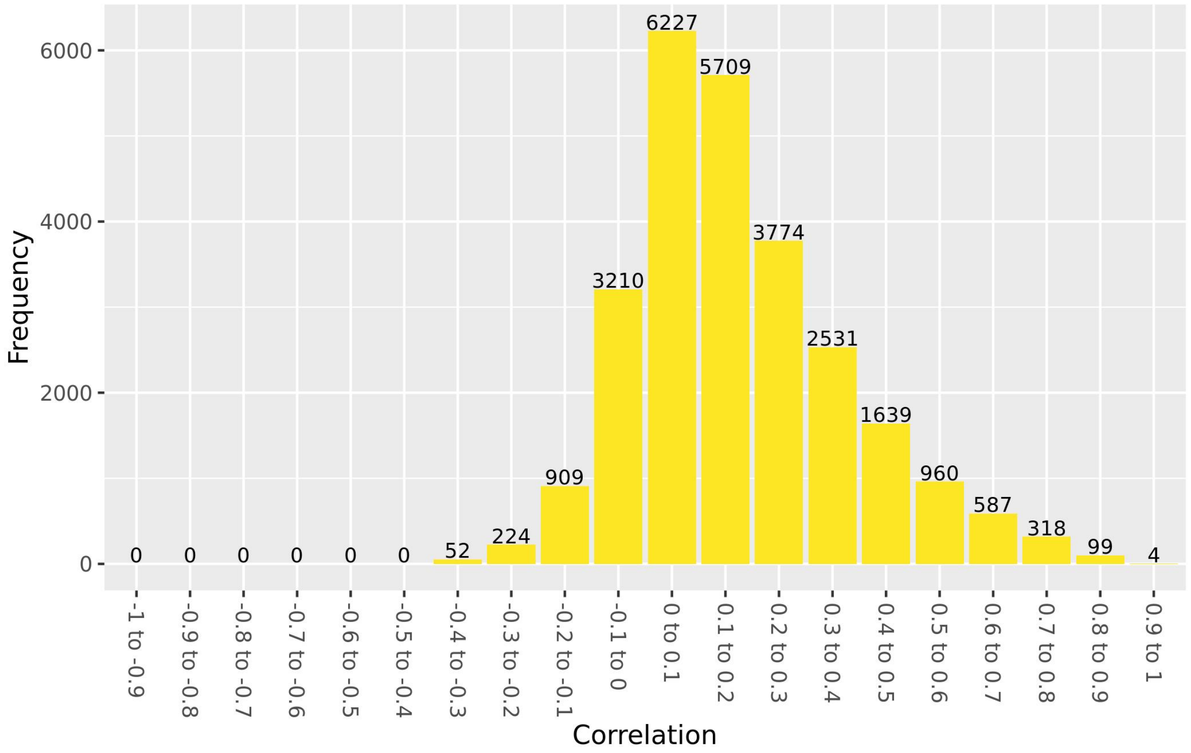
Chromosome 4: values of cis correlations
within 1kb in BiB white British participants at birth


Chromosome 5: decay plot of pairwise correlations
vs genomic distance in BiB white British participants at birth
Pairwise correlation
0.50
0.25
0.00
-0.25
0
2500
5000
7500
10000
Genomic distance (bp)
Positive
Negative

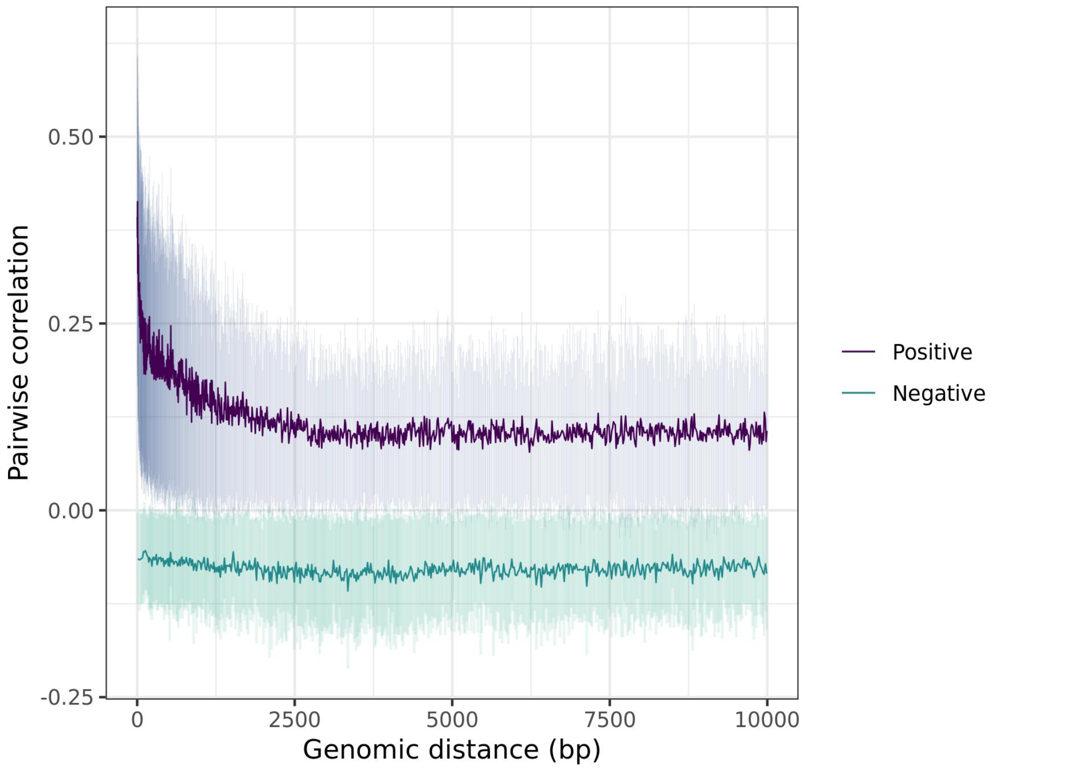

Chromosome 5: decay plot of pairwise correlations
vs genomic distance in BiB white British participants at birth
Pairwise correlation
0.50
0.25
0.00
-0.25
0
2500
5000
7500
10000
Genomic distance (bp)
Positive
Negative



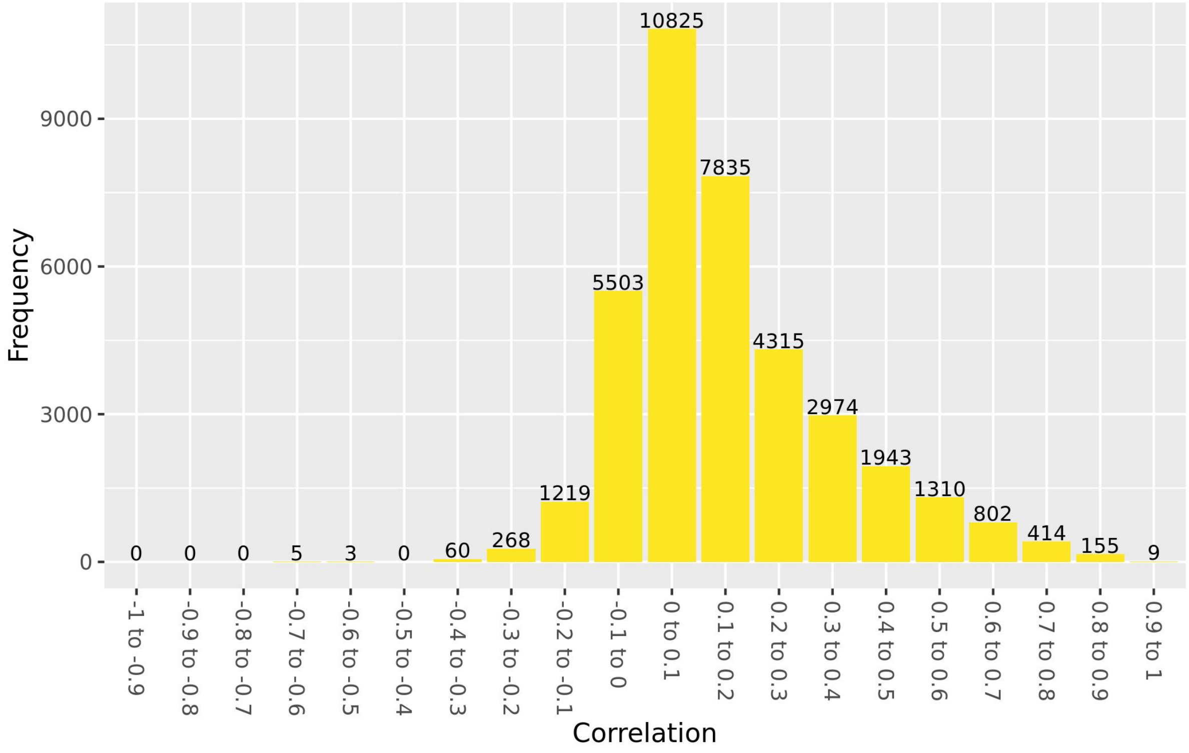
Chromosome 15: decay plot of pairwise correlations
vs genomic distance in BiB white British participants at birth



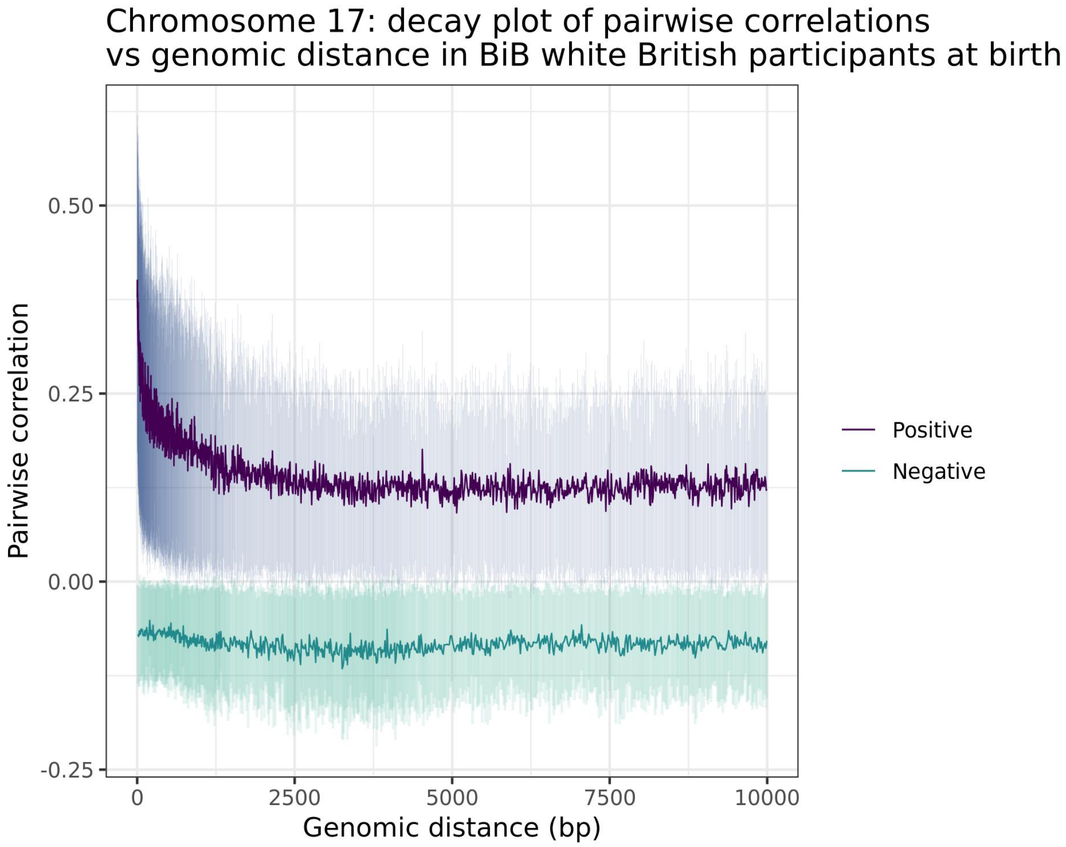
Chromosome 15: values of cis correlations
within 1kb in BiB white British participants at birth

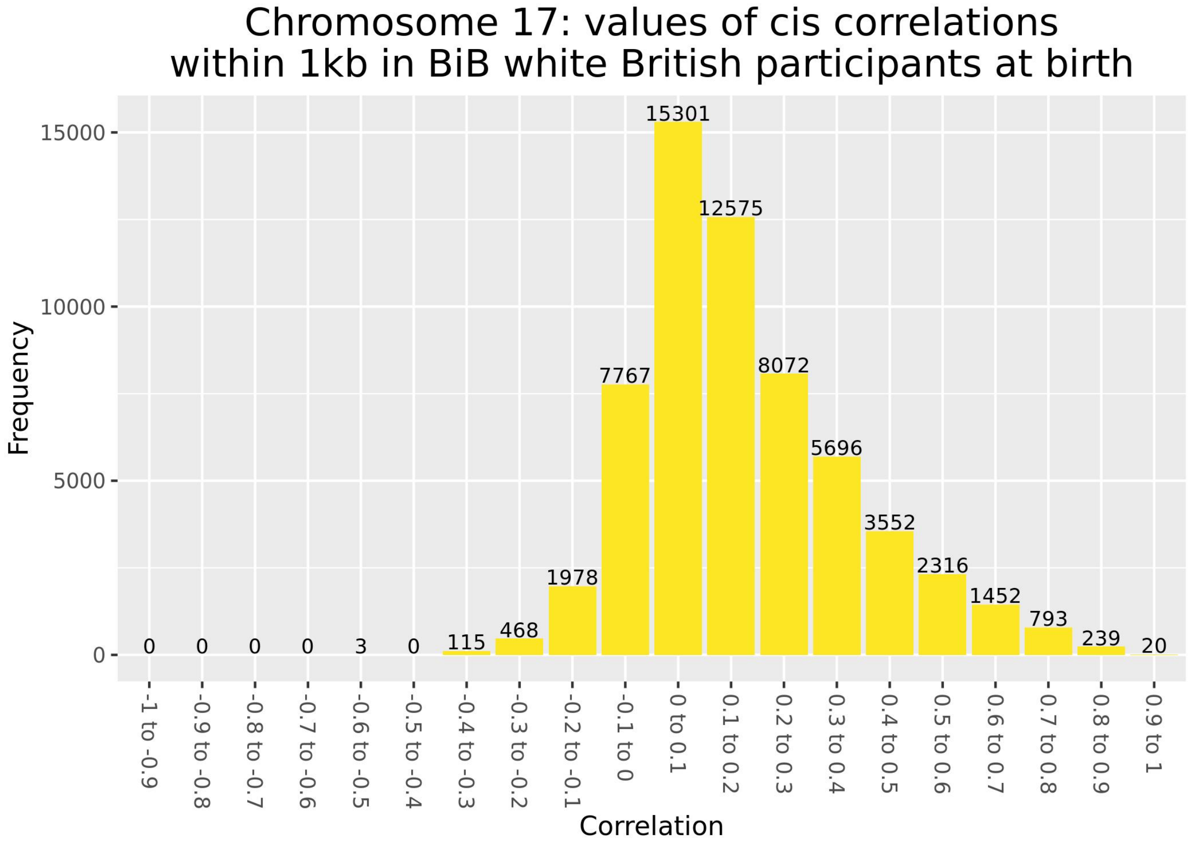


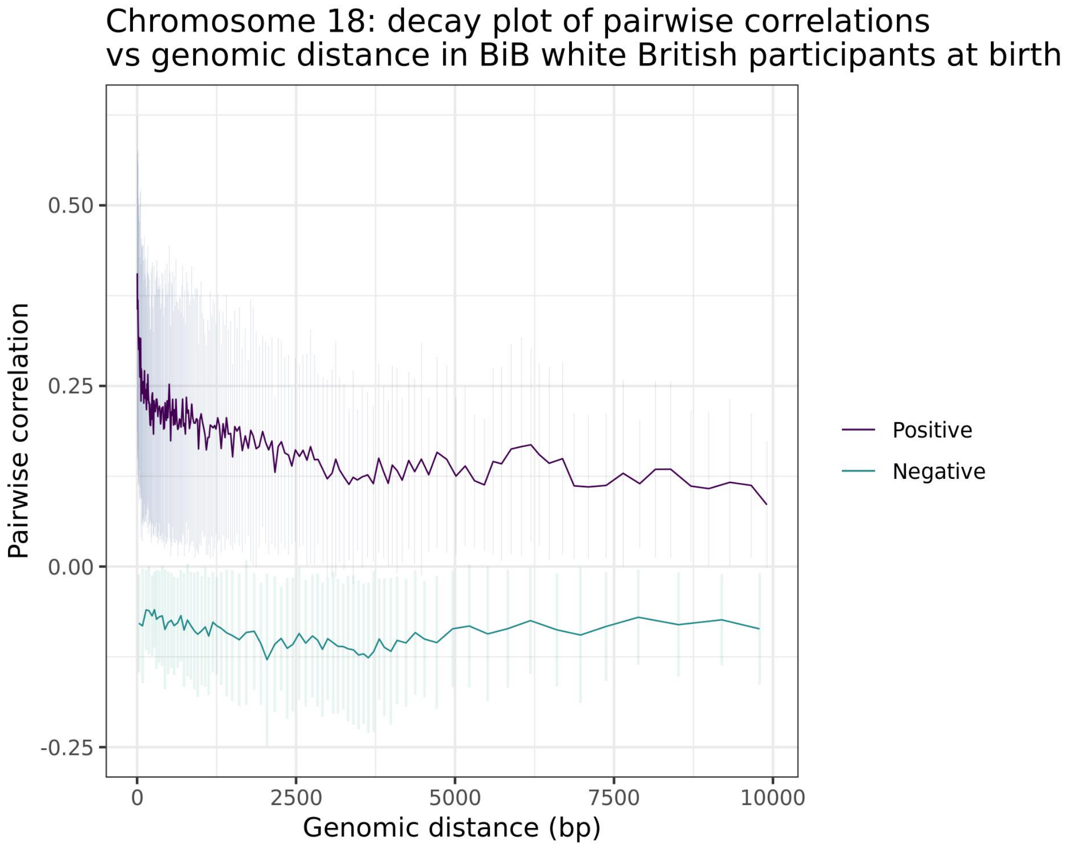
Chromosome 16: decay plot of pairwise correlations
vs genomic distance in BiB white British participants at birth



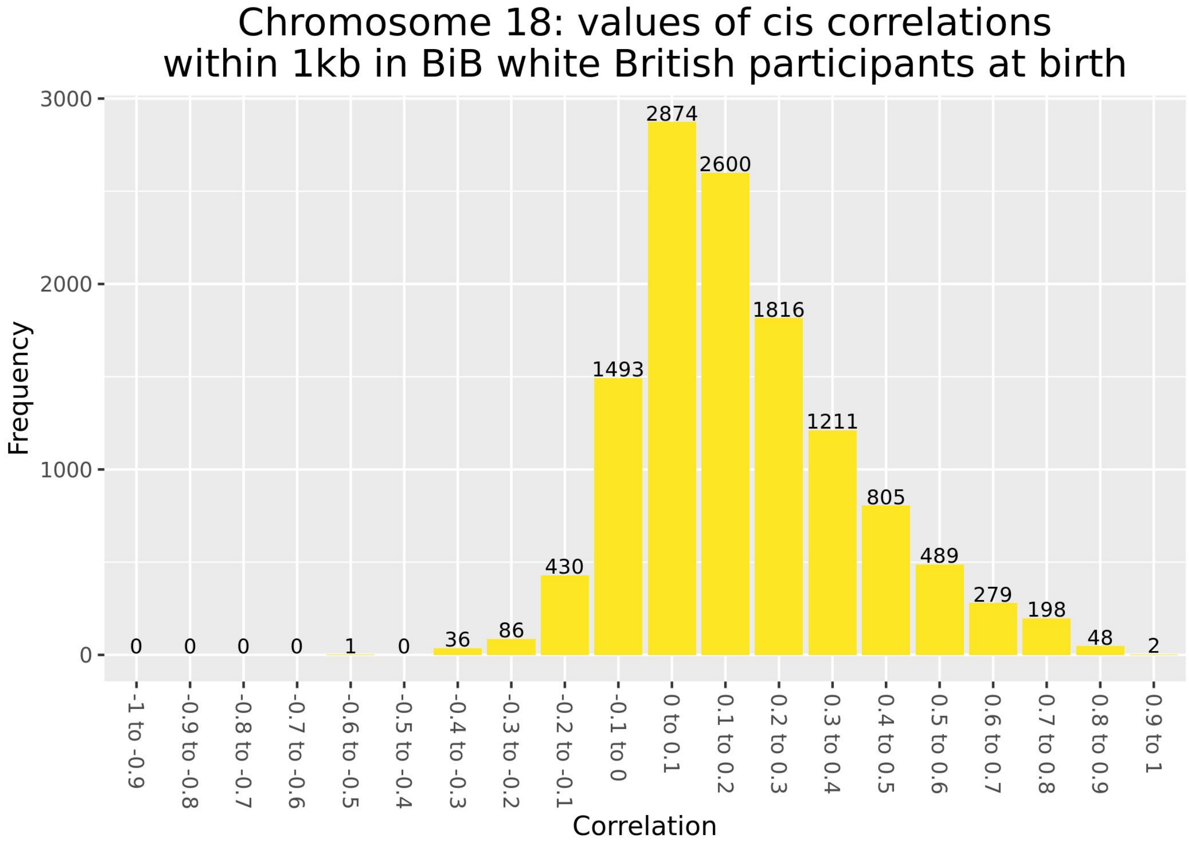
Chromosome 16: values of cis correlations
within 1kb in BiB white British participants at birth


Chromosome 17: decay plot of pairwise correlations
vs genomic distance in BiB white British participants at birth
Pairwise correlation
0.50
0.25
0.00
-0.25
0
2500
5000
7500
10000
Genomic distance (bp)
Positive
Negative


Chromosome 17: values of cis correlations
within 1kb in BiB white British participants at birth
Frequency
15000
10000
5000
0
0
0
0
0
3
0
115
468
1978
7767
15301
12575
8072
5696
3552
2316
1452
793
239
20
-1 to -0.9
-0.9 to -0.8
-0.8 to -0.7
-0.7 to -0.6
-0.6 to -0.5
-0.5 to -0.4
-0.4 to -0.3
-0.3 to -0.2
-0.2 to -0.1
-0.1 to 0
0 to 0.1
0.1 to 0.2
0.2 to 0.3
0.3 to 0.4
0.4 to 0.5
0.5 to 0.6
0.6 to 0.7
0.7 to 0.8
0.8 to 0.9
0.9 to 1
Correlation

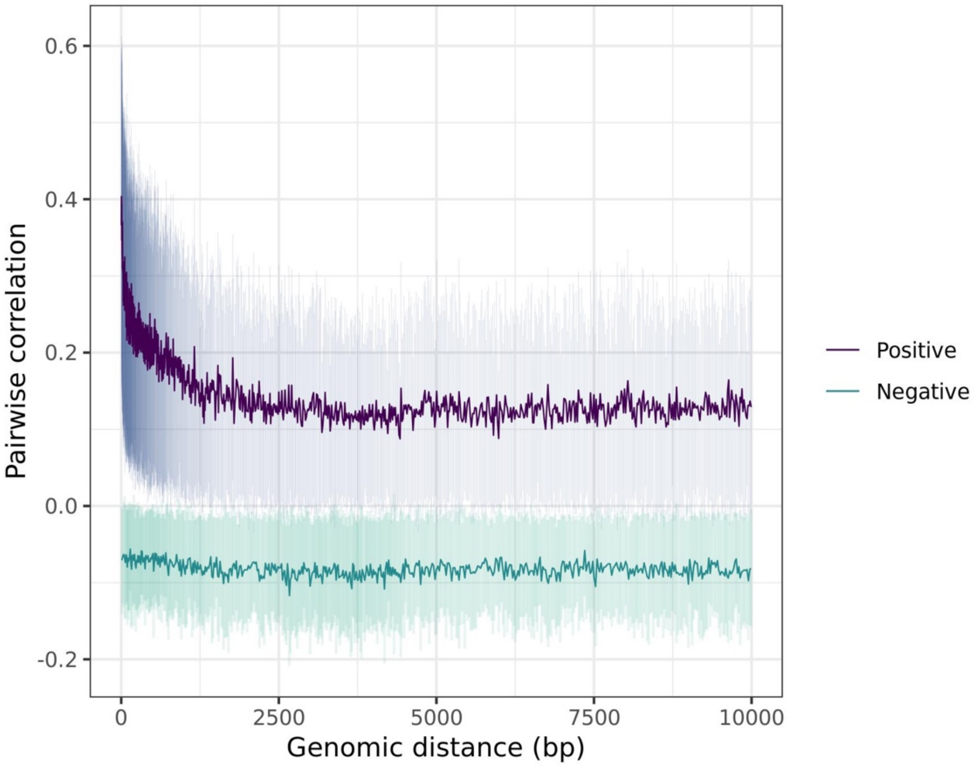


Chromosome 18: decay plot of pairwise correlations
vs genomic distance in BiB white British participants at birth
Pairwise correlation
0.50
0.25
0.00
-0.25
0
2500
5000
7500
10000
Genomic distance (bp)
Positive
Negative

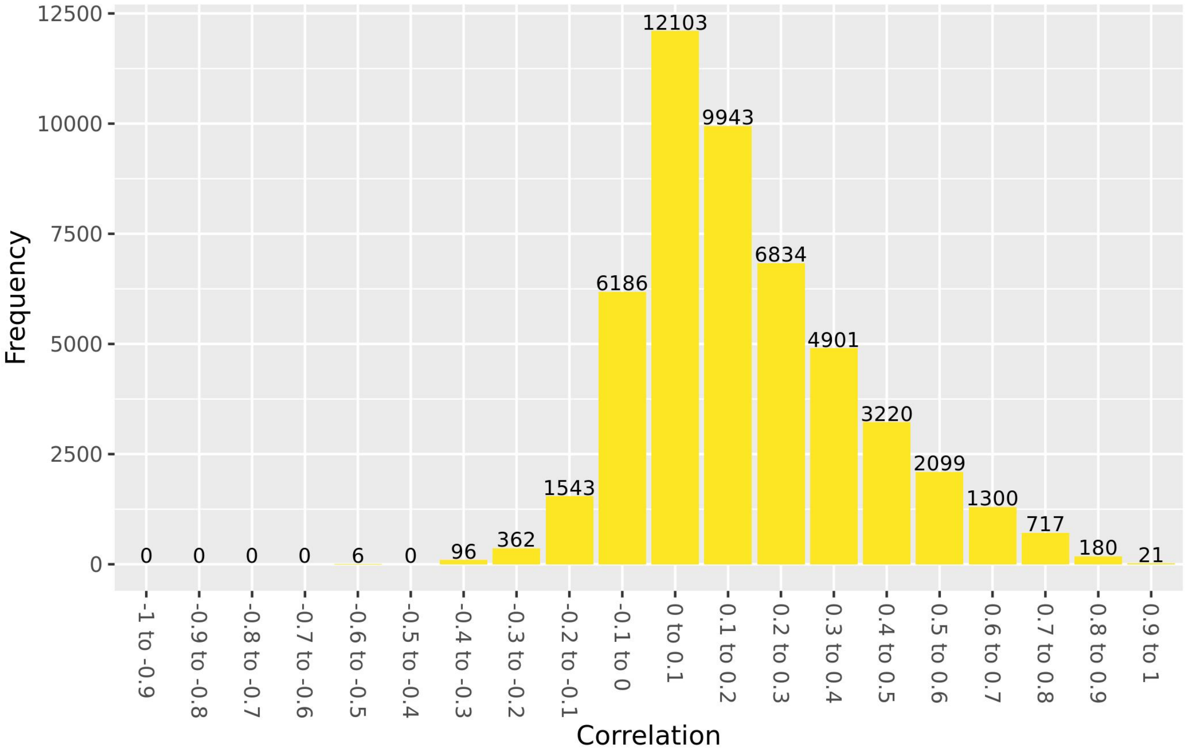


Chromosome 18: values of cis correlations
within 1kb in BiB white British participants at birth
Frequency
3000
2000
1000
0
0
0
0
0
1
0
36
86
430
1493
2874
2600
1816
1211
805
489
279
198
48
2
-1 to -0.9
-0.9 to -0.8
-0.8 to -0.7
-0.7 to -0.6
-0.6 to -0.5
-0.5 to -0.4
-0.4 to -0.3
-0.3 to -0.2
-0.2 to -0.1
-0.1 to 0
0 to 0.1
0.1 to 0.2
0.2 to 0.3
0.3 to 0.4
0.4 to 0.5
0.5 to 0.6
0.6 to 0.7
0.7 to 0.8
0.8 to 0.9
0.9 to 1
Correlation

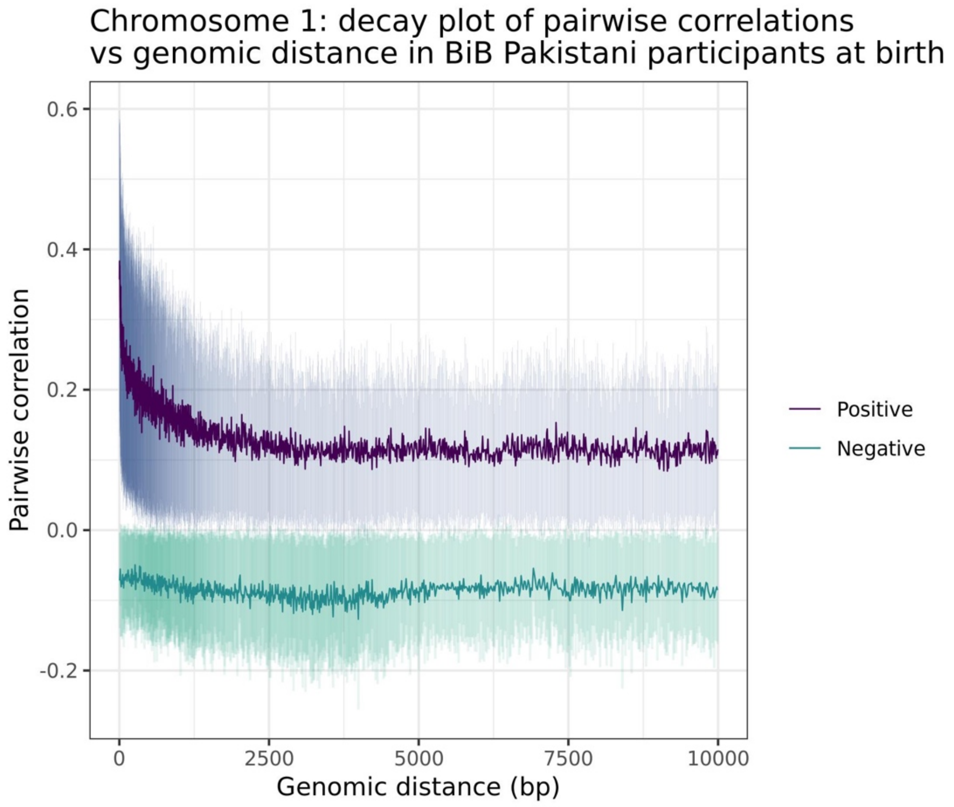
Chromosome 19: decay plot of pairwise correlations
vs genomic distance in BiB white British participants at birth

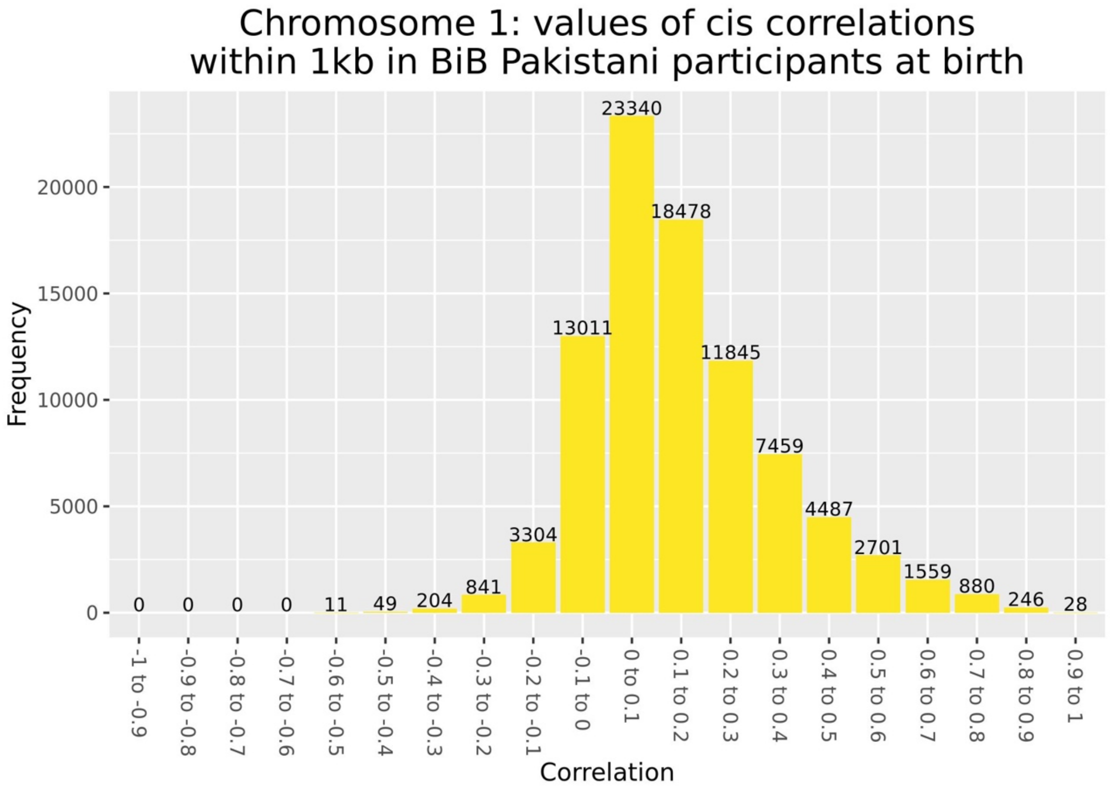


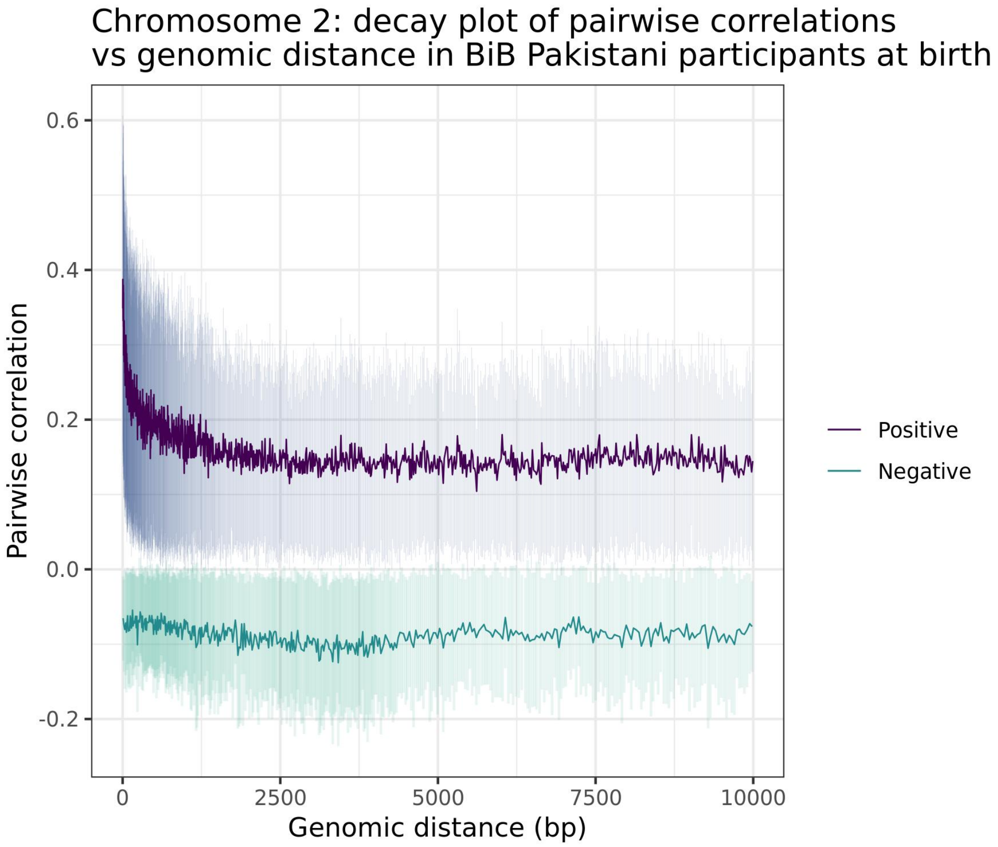
Chromosome 19: values of cis correlations
within 1kb in BiB white British participants at birth

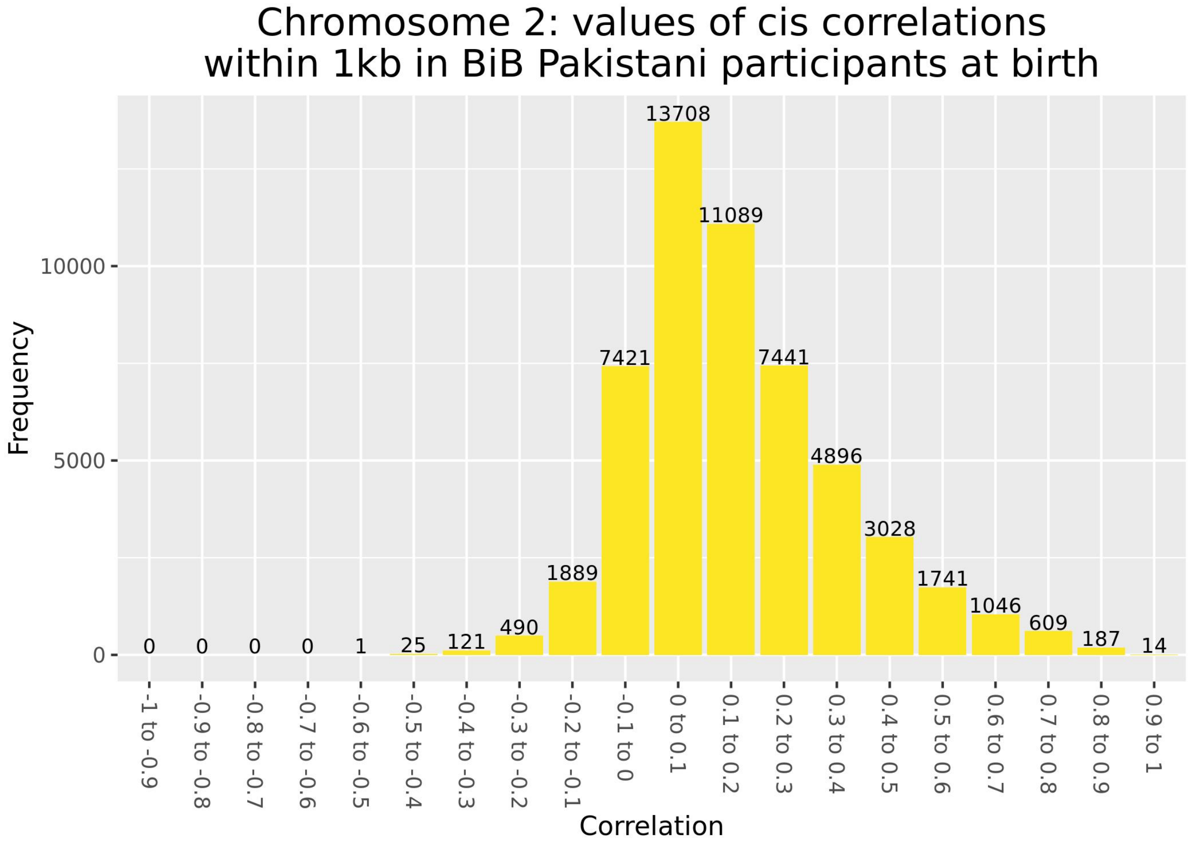


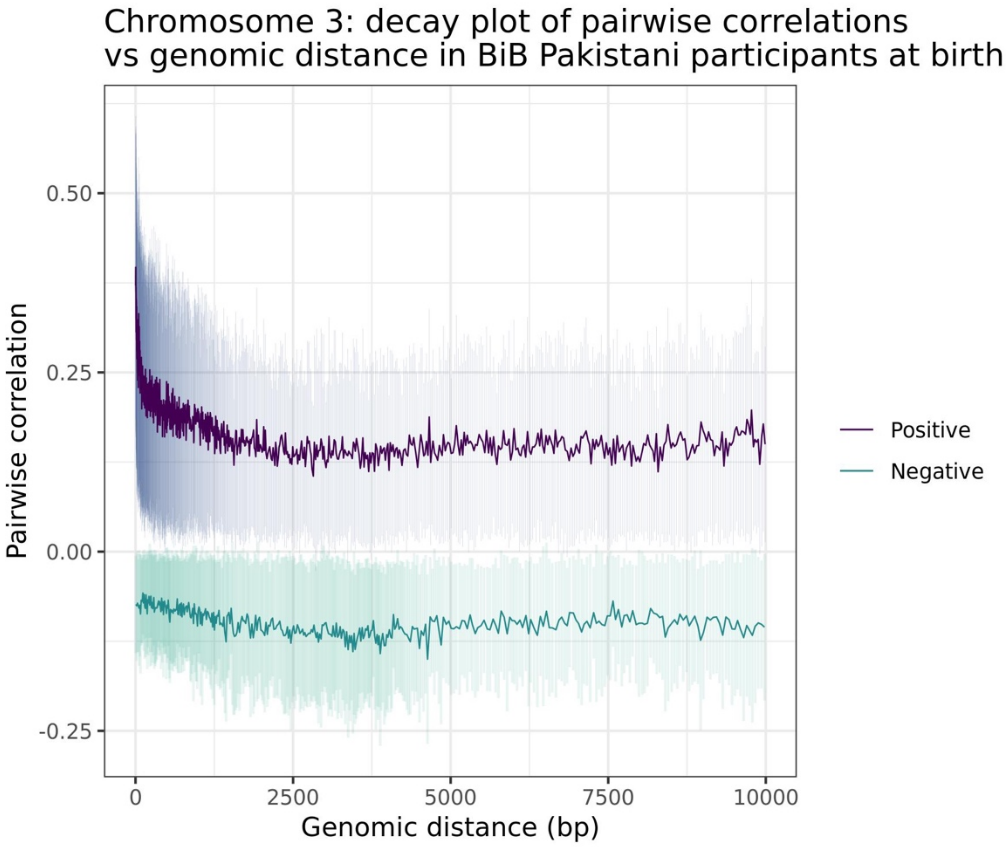
Pakistani ethnic group

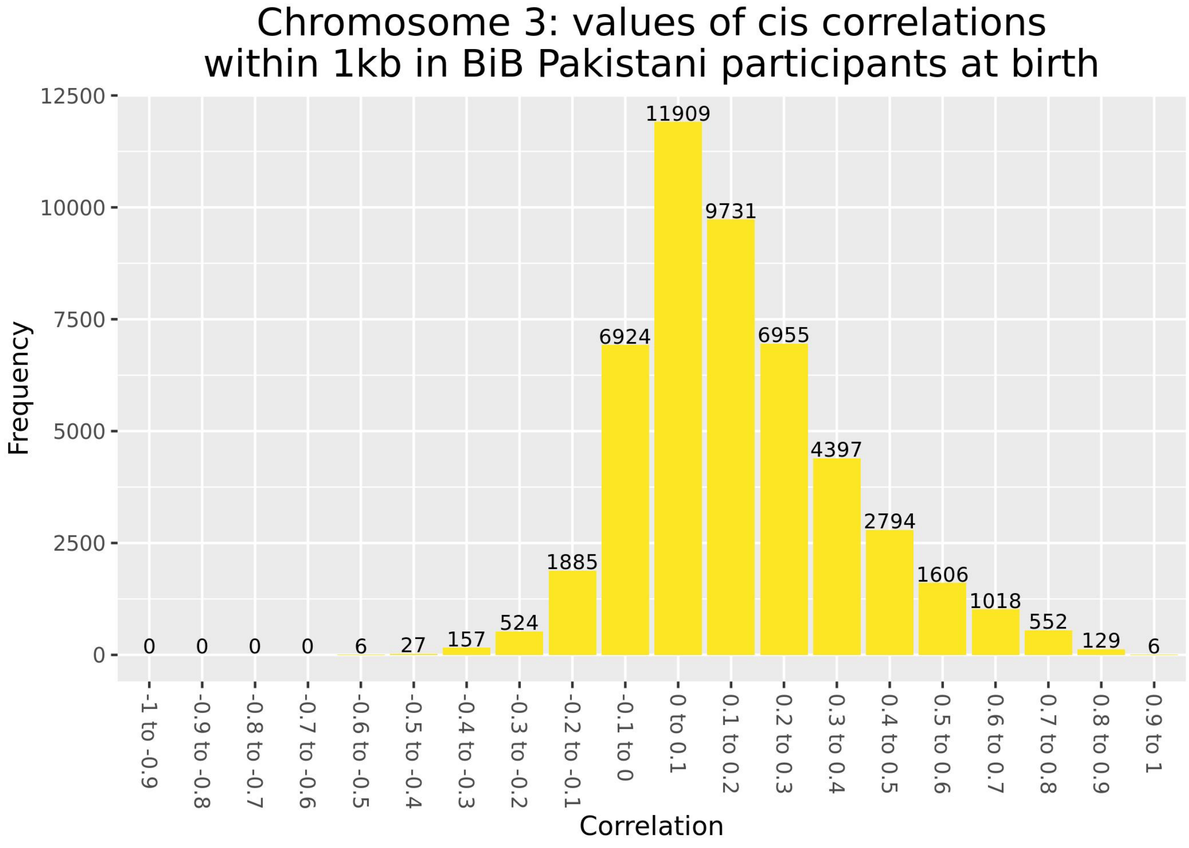


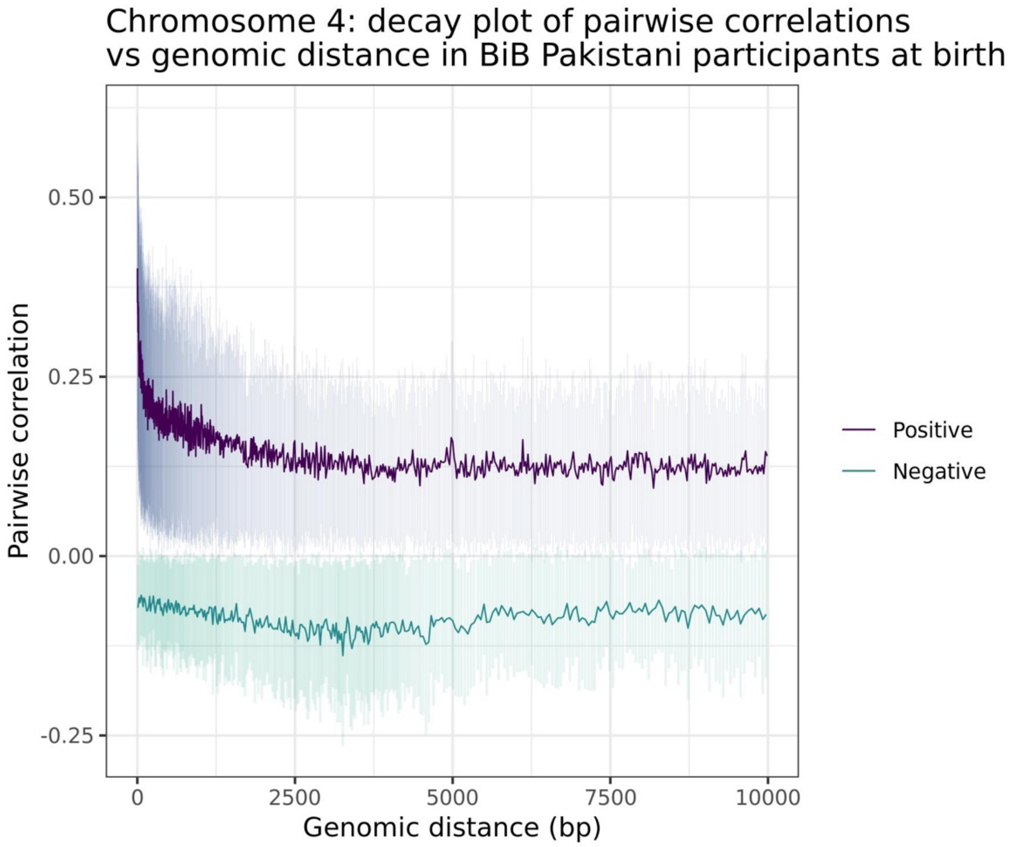


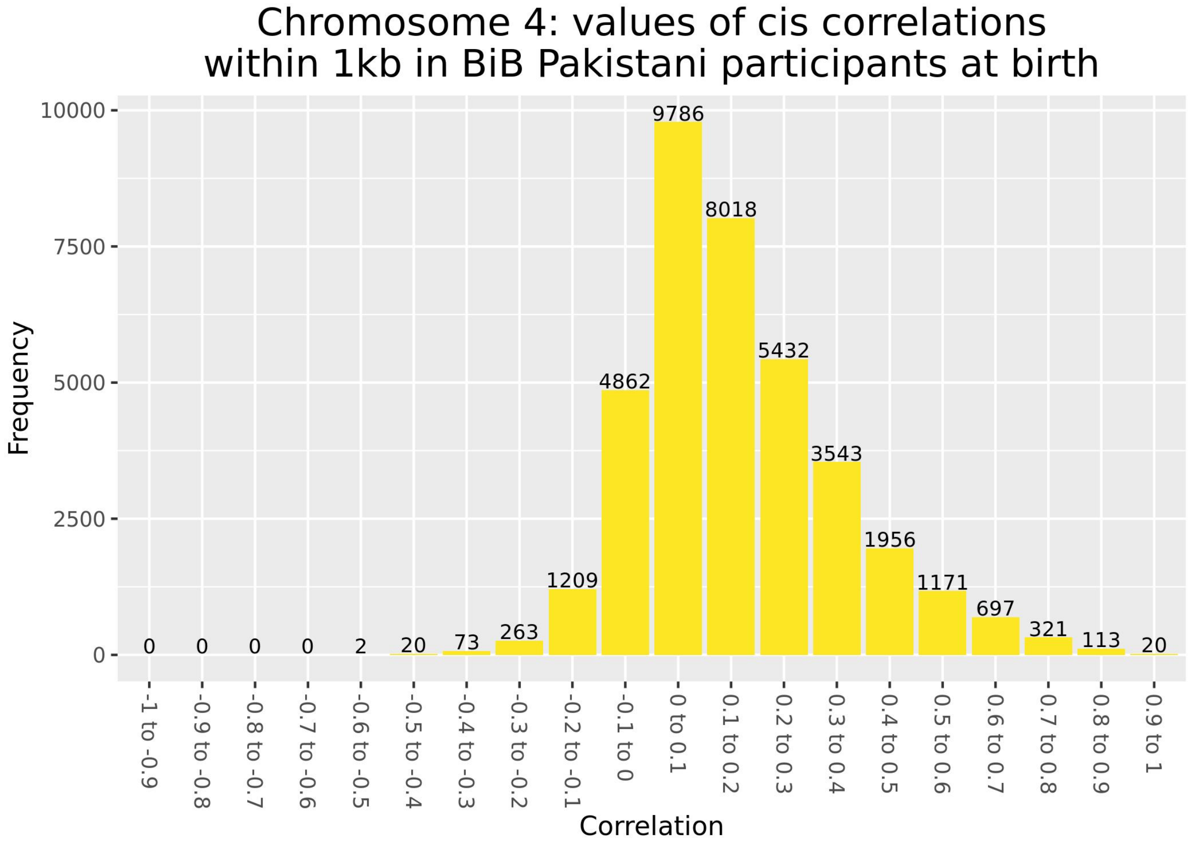

Chromosome 2: decay plot of pairwise correlations
vs genomic distance in BiB Pakistani participants at birth

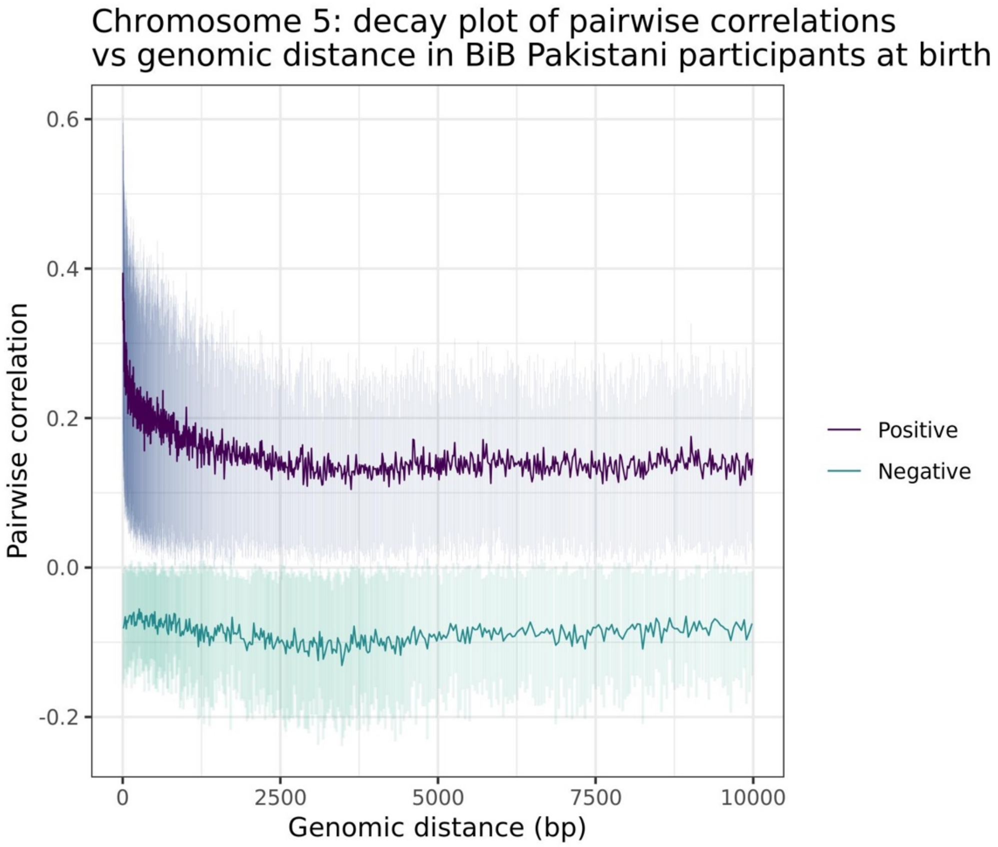

Chromosome 2: values of cis correlations
within 1kb in BiB Pakistani participants at birth

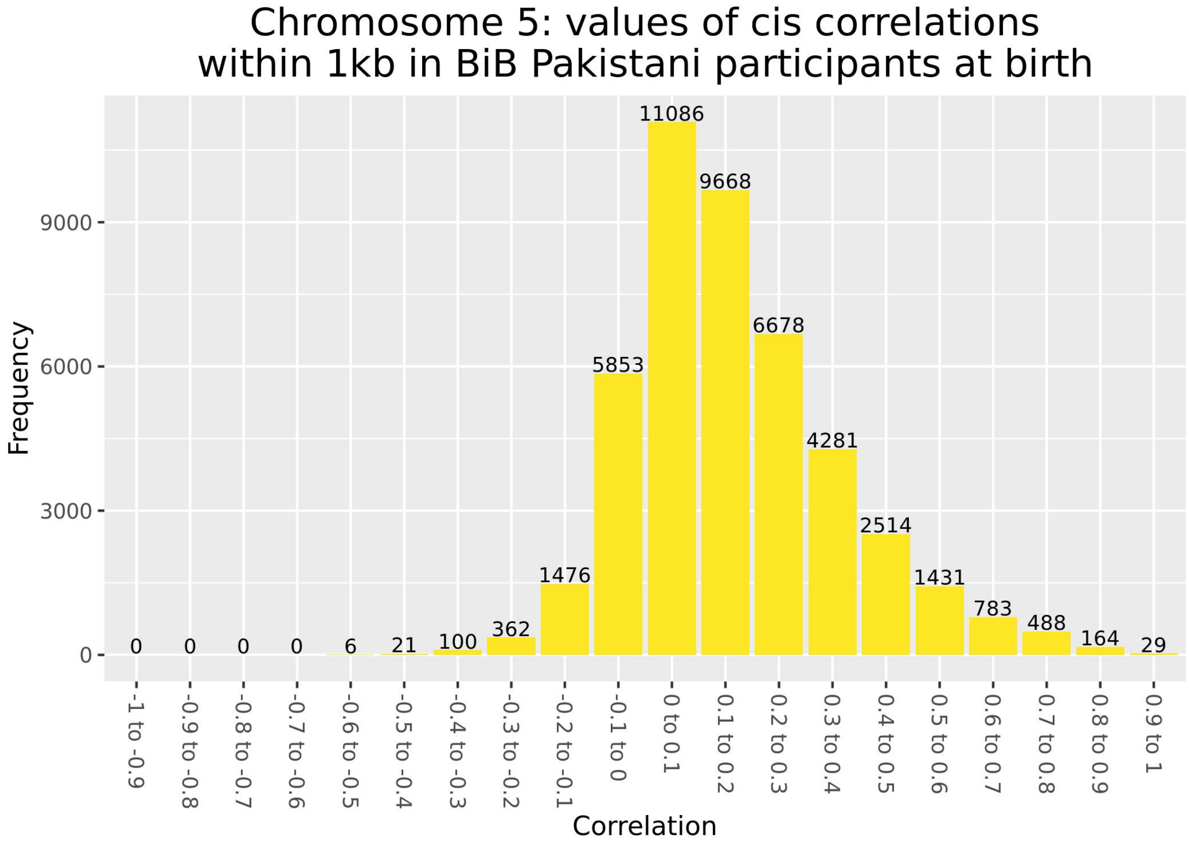

Chromosome 3: decay plot of pairwise correlations vs genomic distance in BiB Pakistani participants at birth

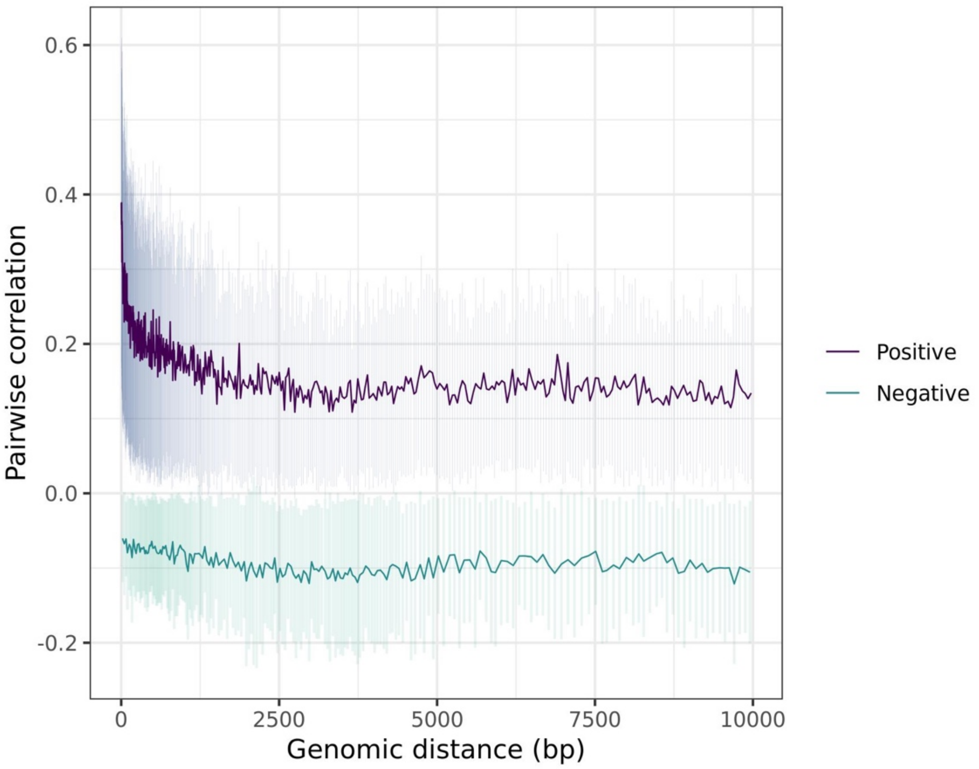

Chromosome 3: values of cis correlations within 1kb in BiB Pakistani participants at birth


Chromosome 4: decay plot of pairwise correlations
vs genomic distance in BiB Pakistani participants at birth
Pairwise correlation
0.50
0.25
0.00
-0.25
0
2500
5000
7500
10000
Genomic distance (bp)
Positive
Negative

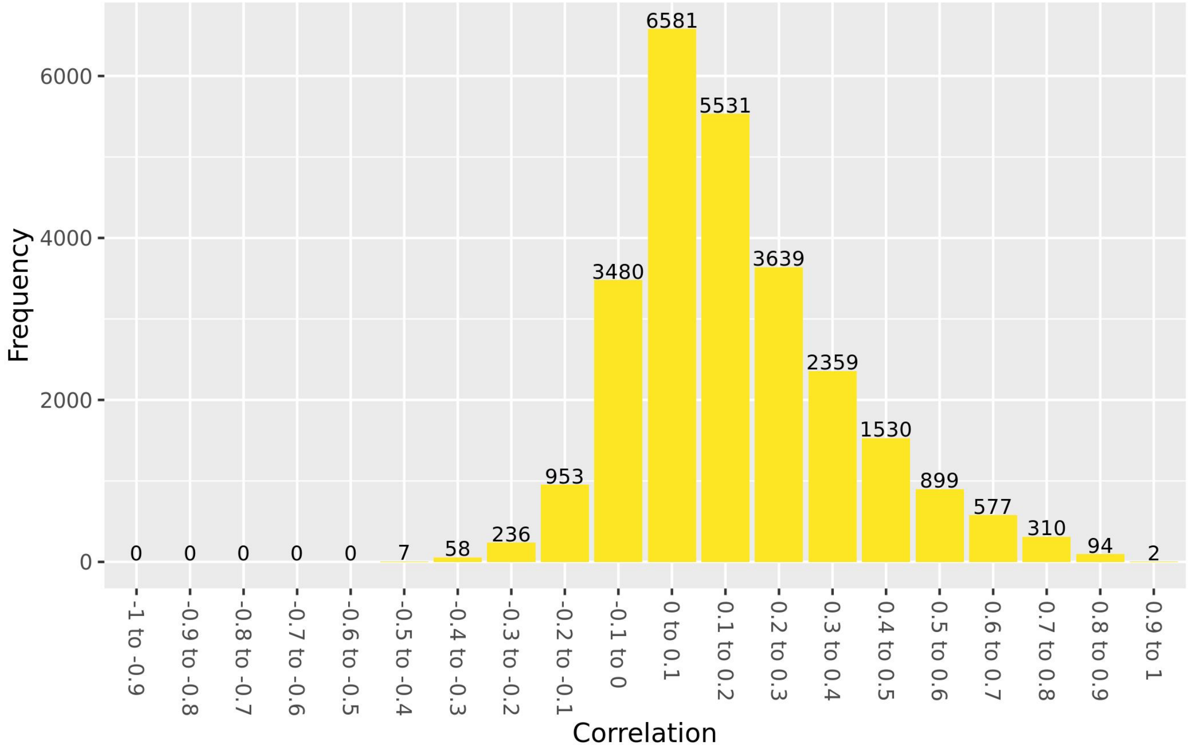

Chromosome 4: values of cis correlations
within 1kb in BiB Pakistani participants at birth
Frequency
10000
7500
5000
2500
0
0
0
0
0
2
20
73
263
1209
4862
9786
8018
5432
3543
1956
1171
697
321
113
20
-1 to -0.9
-0.9 to -0.8
-0.8 to -0.7
-0.7 to -0.6
-0.6 to -0.5
-0.5 to -0.4
-0.4 to -0.3
-0.3 to -0.2
-0.2 to -0.1
-0.1 to 0
0 to 0.1
0.1 to 0.2
0.2 to 0.3
0.3 to 0.4
0.4 to 0.5
0.5 to 0.6
0.6 to 0.7
0.7 to 0.8
0.8 to 0.9
0.9 to 1
Correlation


Chromosome 5: decay plot of pairwise correlations vs genomic distance in BiB Pakistani participants at birth

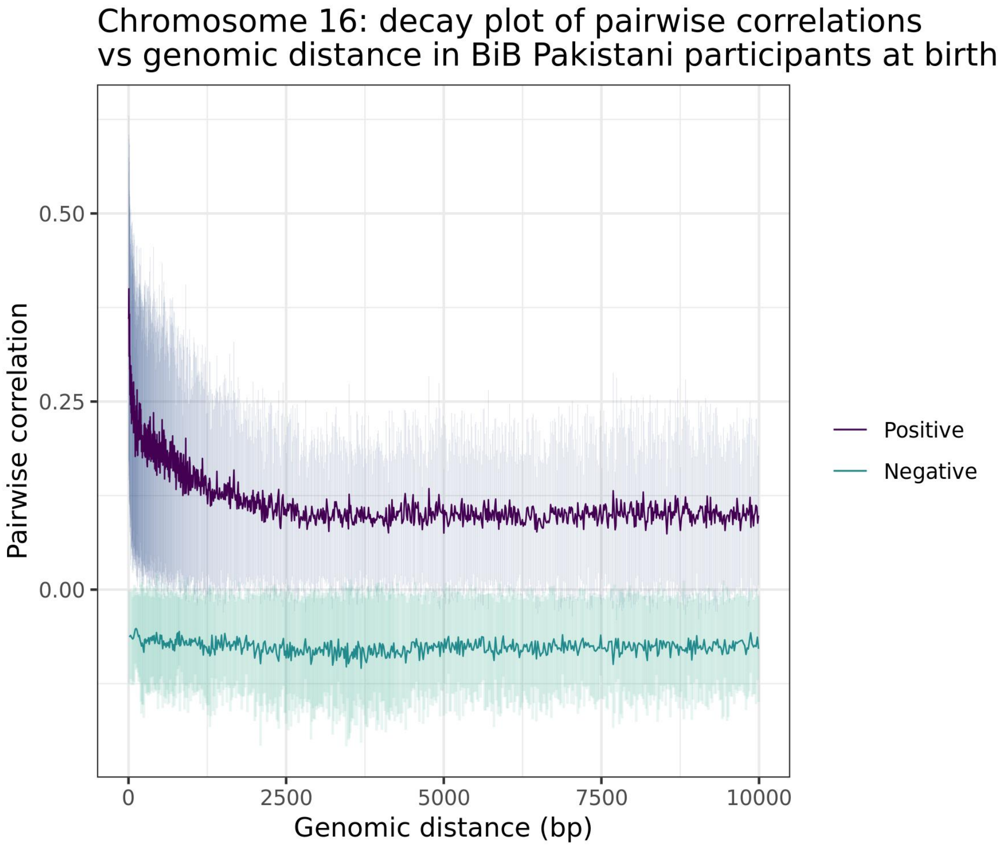

Chromosome 5: values of cis correlations within 1kb in BiB Pakistani participants at birth

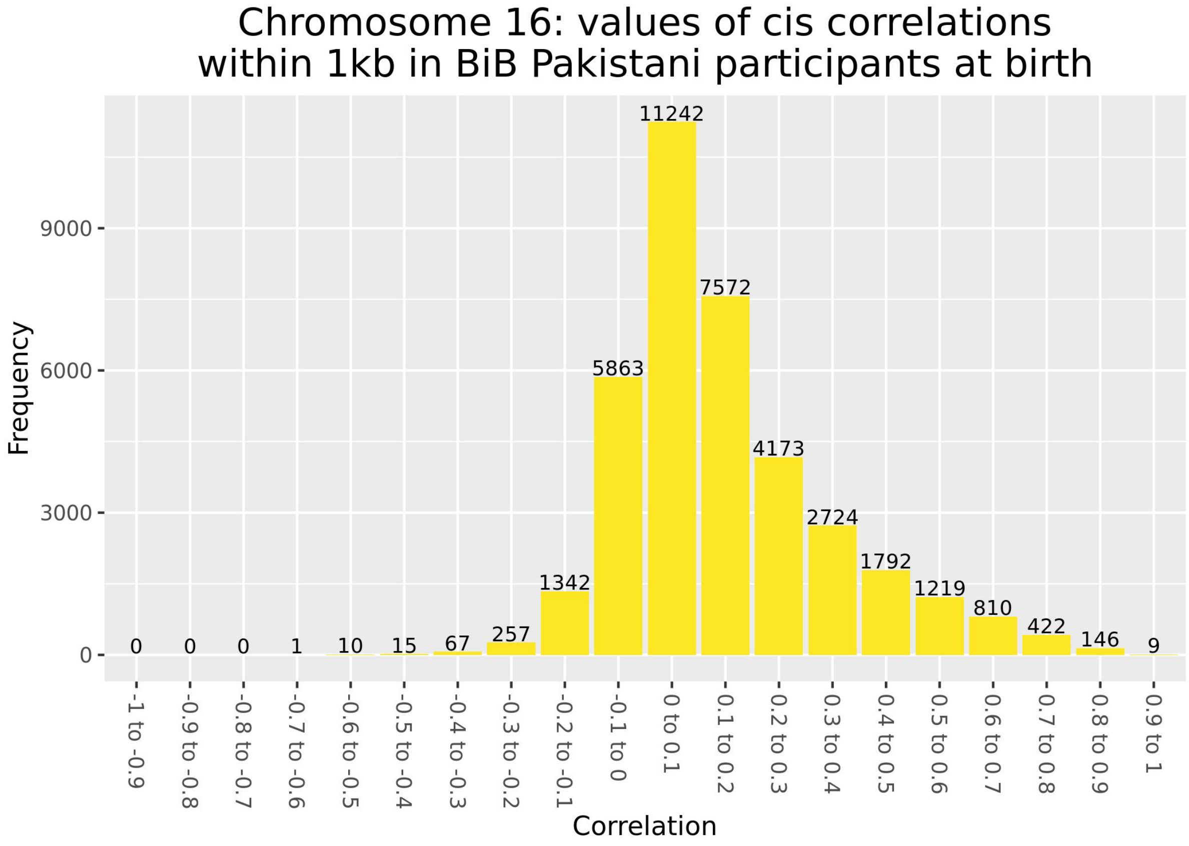


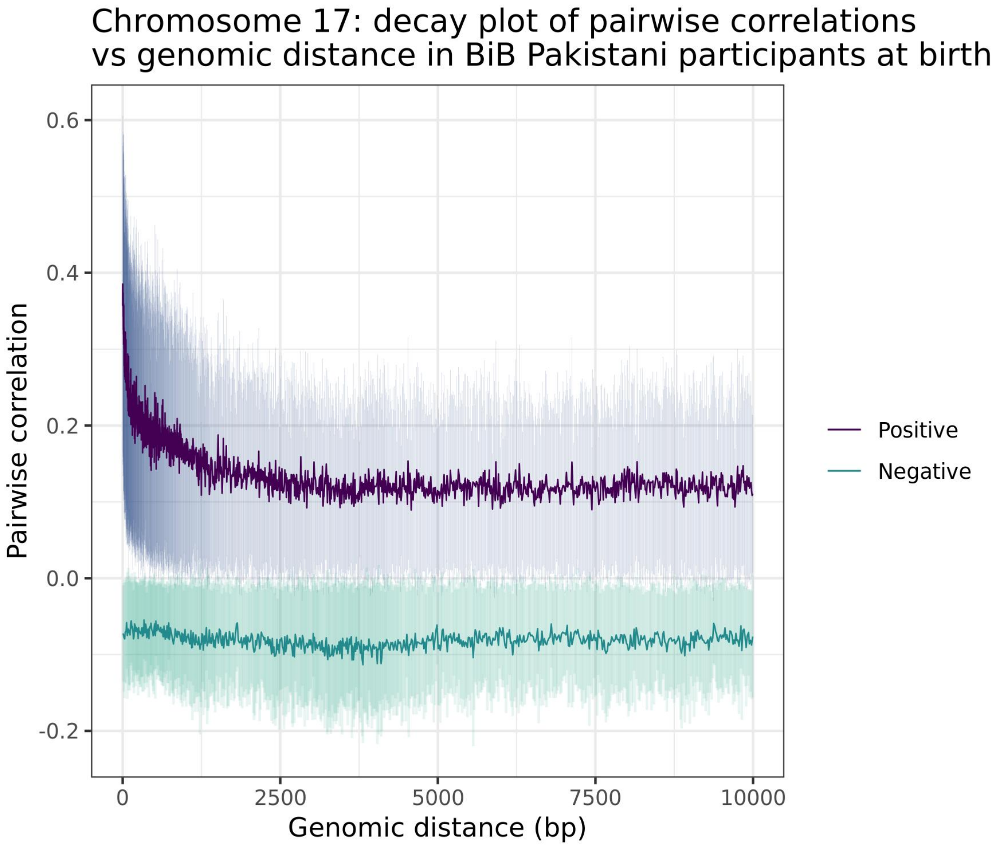
Chromosome 15: decay plot of pairwise correlations vs genomic distance in BiB Pakistani participants at birth



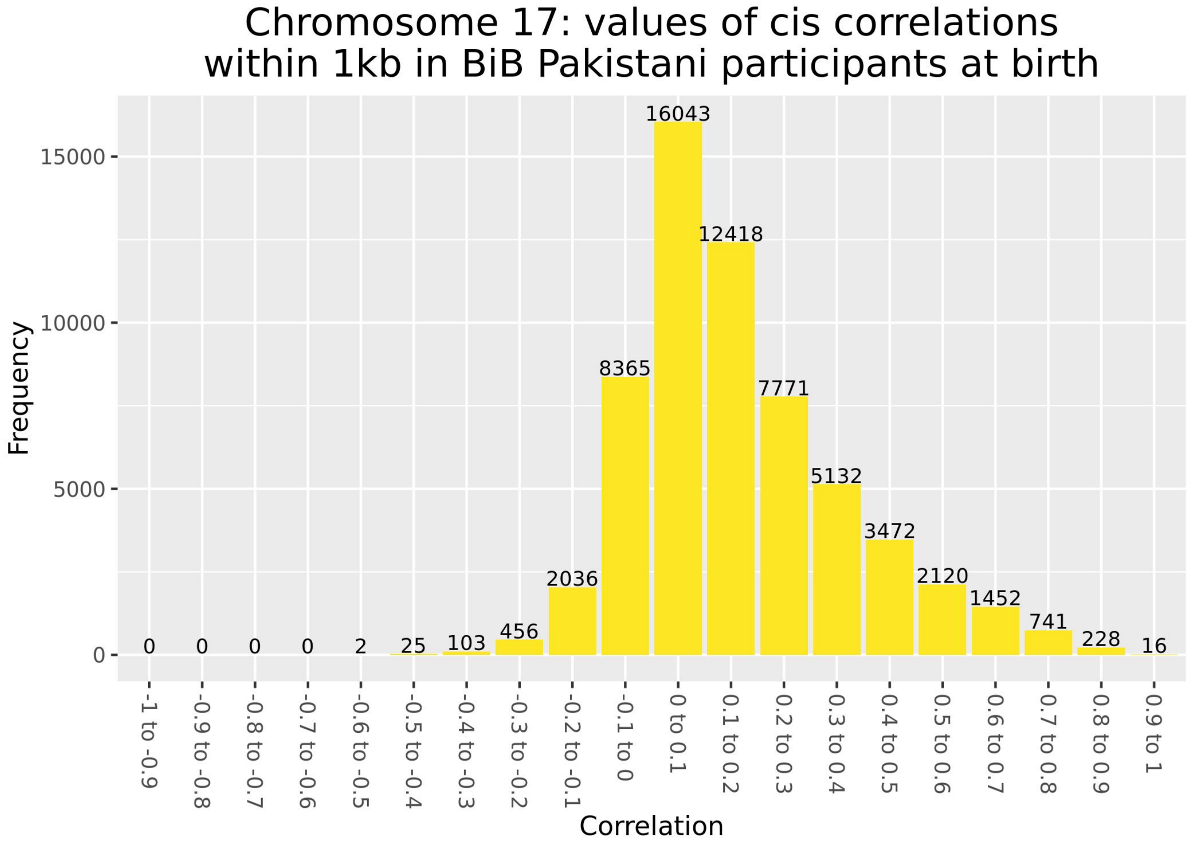
Chromosome 15: values of cis correlations within 1kb in BiB Pakistani participants at birth


Chromosome 16: decay plot of pairwise correlations
vs genomic distance in BiB Pakistani participants at birth
Pairwise correlation
0.50
0.25
0.00
0
2500
5000
7500
10000
Genomic distance (bp)
Positive
Negative

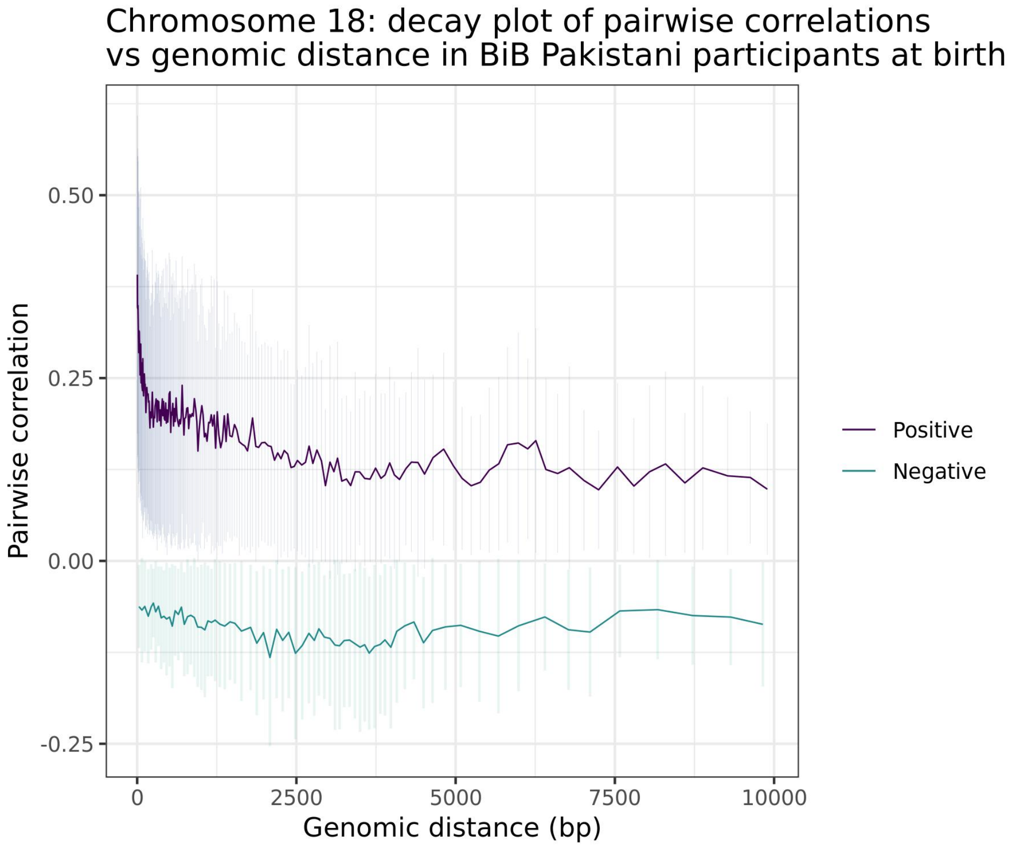

Chromosome 16: values of cis correlations
within 1kb in BiB Pakistani participants at birth
Frequency
9000
6000
3000
0
0
0
0
1
10
15
67
257
1342
5863
11242
7572
4173
2724
1792
1219
810
422
146
9
-1 to -0.9
-0.9 to -0.8
-0.8 to -0.7
-0.7 to -0.6
-0.6 to -0.5
-0.5 to -0.4
-0.4 to -0.3
-0.3 to -0.2
-0.2 to -0.1
-0.1 to 0
0 to 0.1
0.1 to 0.2
0.2 to 0.3
0.3 to 0.4
0.4 to 0.5
0.5 to 0.6
0.6 to 0.7
0.7 to 0.8
0.8 to 0.9
0.9 to 1
Correlation



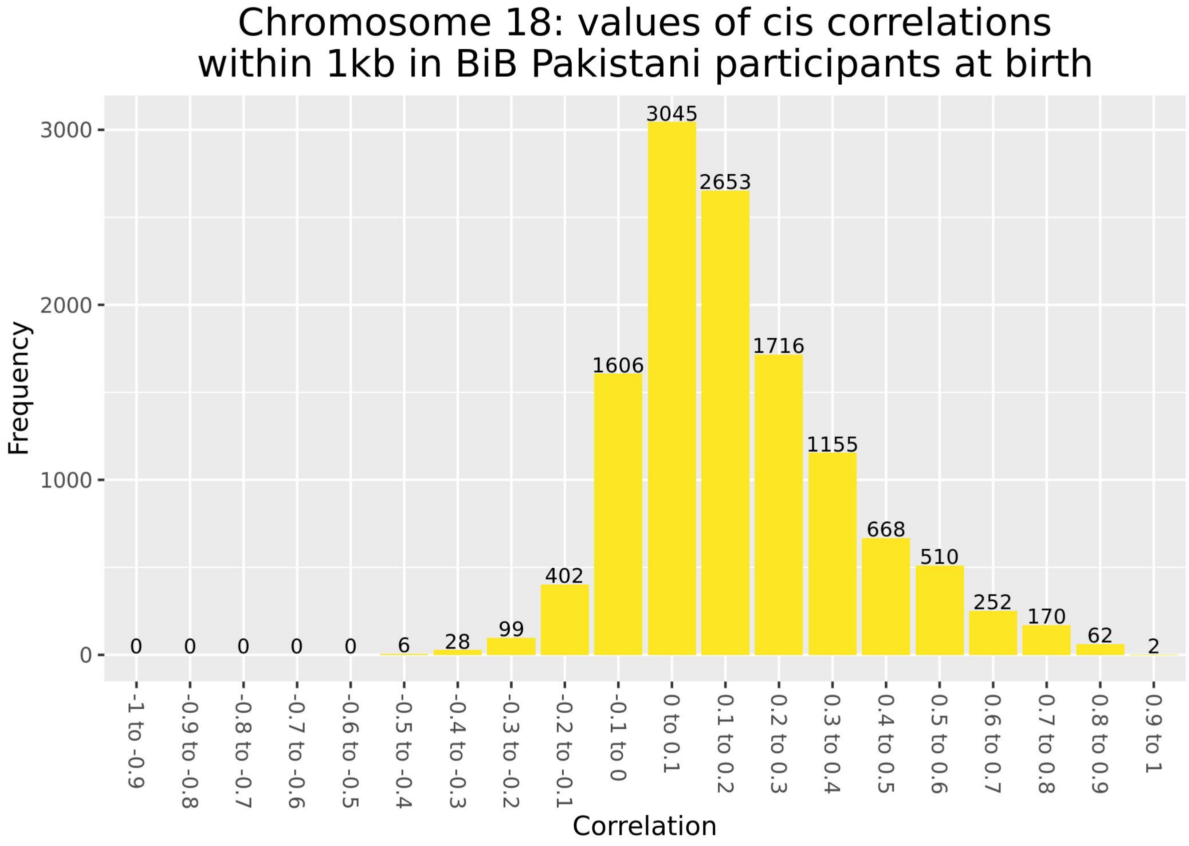
Chromosome 17: decay plot of pairwise correlations vs genomic distance in BiB Pakistani participants at birth



Chromosome 17: values of cis correlations within 1kb in BiB Pakistani participants at birth

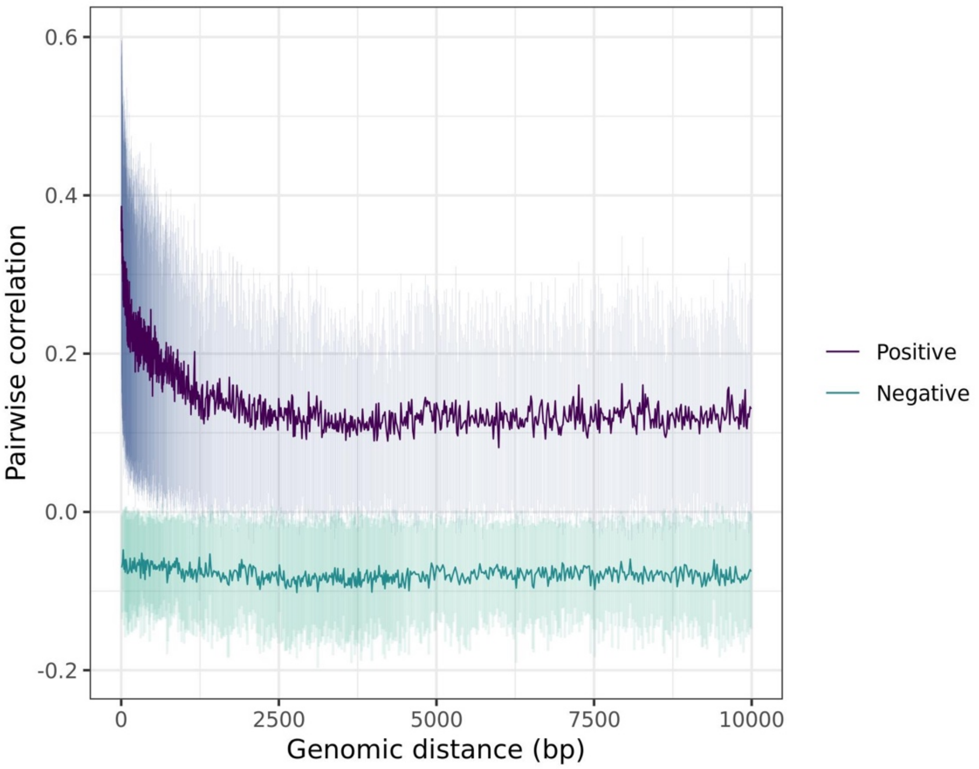


Chromosome 18: decay plot of pairwise correlations vs genomic distance in BiB Pakistani participants at birth

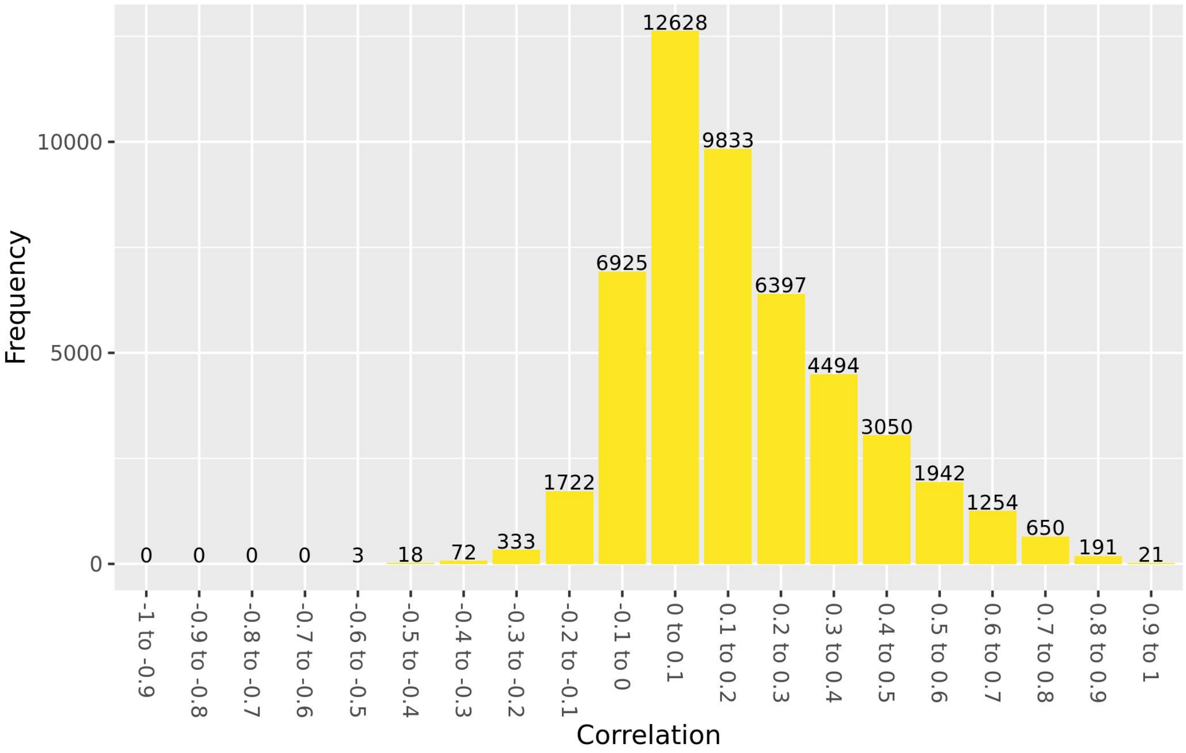


Chromosome 18: values of cis correlations within 1kb in BiB Pakistani participants at birth

Chromosome 19: decay plot of pairwise correlations
vs genomic distance in BiB Pakistani participants at birth



Chromosome 19: values of cis correlations
within 1kb in BiB Pakistani participants at birth



Appendix 6: Gene ontology and KEGG pathway enrichments for ARIES and BiB consensus WGCNA network. Modules are included in this appendix if there were FDR significant enrichments in one of the datasets.

Blue module – ARIES – gene ontology

	Term	Ont	N	DE	P.DE	FDR
GO:0050911	detection of chemical stimulus involved in sensory perception of smell	BP	168	42	5.15E-07	0.00559865
GO:0004984	olfactory receptor activity	MF	168	42	5.15E-07	0.00559865
GO:0050907	detection of chemical stimulus involved in sensory perception	BP	203	48	8.44E-07	0.00611412
GO:0009593	detection of chemical stimulus	BP	235	54	2.78E-06	0.01392593
GO:0007608	sensory perception of smell	BP	190	45	3.20E-06	0.01392593
GO:0045202	synapse	CC	780	182	1.91E-05	0.06820315
GO:0044420	extracellular matrix component	CC	110	37	2.20E-05	0.06820315
GO:0007606	sensory perception of chemical stimulus	BP	249	54	2.51E-05	0.06825819
GO:0001508	action potential	BP	119	39	2.85E-05	0.06890073
GO:0050906	detection of stimulus involved in sensory perception	BP	243	53	4.01E-05	0.08723358
GO:0035249	synaptic transmission, glutamatergic	BP	85	30	5.19E-05	0.10256115
GO:0051606	detection of stimulus	BP	387	81	7.93E-05	0.13214613
GO:0086001	cardiac muscle cell action potential	BP	59	23	8.19E-05	0.13214613
GO:0050877	nervous system process	BP	965	197	8.64E-05	0.13214613
GO:0071944	cell periphery	CC	4357	799	9.12E-05	0.13214613
GO:0097458	neuron part	CC	1349	284	0.00011346	0.14517015
GO:0004871	signal transducer activity	MF	1303	253	0.00013034	0.14517015
GO:0055024	regulation of cardiac muscle tissue development	BP	68	25	0.00013109	0.14517015
GO:0007155	cell adhesion	BP	1157	246	0.0001369	0.14517015
GO:0098772	molecular function regulator	MF	1538	306	0.0001435	0.14517015

Blue module – ARIES – KEGG pathways

	Pathway	N	DE	P.DE	FDR
path:hsa04740	Olfactory transduction	204	52	1.03E-07	3.48E-05
path:hsa04512	ECM-receptor interaction	82	26	0.00128038	0.14408521
path:hsa00512	Mucin type O-glycan biosynthesis	28	12	0.00135868	0.14408521
path:hsa04151	PI3K-Akt signaling pathway	308	74	0.0017153	0.14408521
path:hsa05146	Amoebiasis	83	24	0.00501764	0.33718555
path:hsa04724	Glutamatergic synapse	104	29	0.00849928	0.40162718
path:hsa04924	Renin secretion	58	17	0.00937504	0.40162718
path:hsa05412	Arrhythmogenic right ventricular cardiomyopathy (ARVC)	71	21	0.00956255	0.40162718
path:hsa05165	Human papillomavirus infection	288	65	0.0136197	0.42778083
path:hsa05235	PD-L1 expression and PD-1 checkpoint pathway in cancer	81	22	0.01481881	0.42778083
path:hsa04514	Cell adhesion molecules (CAMs)	110	28	0.01524132	0.42778083
path:hsa04923	Regulation of lipolysis in adipocytes	51	15	0.01527789	0.42778083
path:hsa04726	Serotonergic synapse	95	24	0.02084406	0.48141545
path:hsa04921	Oxytocin signaling pathway	141	34	0.02268885	0.48141545
path:hsa05231	Choline metabolism in cancer	89	23	0.02580576	0.48141545
path:hsa04015	Rap1 signaling pathway	194	45	0.02777379	0.48141545
path:hsa00770	Pantothenate and CoA biosynthesis	19	7	0.02886227	0.48141545
path:hsa04912	GnRH signaling pathway	83	21	0.03090492	0.48141545
path:hsa04114	Oocyte meiosis	104	25	0.03202537	0.48141545
path:hsa04914	Progesterone-mediated oocyte maturation	82	21	0.03285636	0.48141545

Blue module – BiB white British – gene ontology

	Term	Ont	N	DE	P.DE	FDR
GO:0030551	cyclic nucleotide binding	MF	35	9	0.0001215	1
GO:0071726	cellular response to diacyl bacterial lipopeptide	BP	4	3	0.00040412	1
GO:0071724	response to diacyl bacterial lipopeptide	BP	4	3	0.00040412	1
GO:0038124	toll-like receptor TLR6:TLR2 signaling pathway	BP	4	3	0.00040412	1
GO:0061092	positive regulation of phospholipid translocation	BP	3	3	0.00041572	1
GO:0061091	regulation of phospholipid translocation	BP	3	3	0.00041572	1
GO:0042328	heparan sulfate N-acetylglucosaminyltransferase activity	MF	4	3	0.00100102	1
GO:0007405	neuroblast proliferation	BP	52	10	0.00112608	1
GO:0072523	purine-containing compound catabolic process	BP	46	9	0.00117593	1
GO:0030552	cAMP binding	MF	22	6	0.00121515	1
GO:1900227	positive regulation of NLRP3 inflammasome complex assembly	BP	5	3	0.0013157	1
GO:0071221	cellular response to bacterial lipopeptide	BP	6	3	0.00158442	1
GO:0071220	cellular response to bacterial lipoprotein	BP	6	3	0.00158442	1
GO:0070339	response to bacterial lipopeptide	BP	6	3	0.00158442	1
GO:0006195	purine nucleotide catabolic process	BP	39	8	0.00174474	1
GO:0033391	chromatoid body	CC	12	4	0.00210242	1
GO:0030175	filopodium	CC	87	13	0.00213359	1
GO:0042496	detection of diacyl bacterial lipopeptide	BP	2	2	0.00220531	1
GO:0071638	negative regulation of monocyte chemotactic protein-1 production	BP	6	3	0.00224755	1
GO:0003779	actin binding	MF	365	36	0.00240628	1

Blue module – BiB white British – KEGG pathways

	Pathway	N	DE	P.DE	FDR
path:hsa04530	Tight junction	147	16	0.00845165	1
path:hsa04924	Renin secretion	58	8	0.01238119	1
path:hsa04514	Cell adhesion molecules (CAMs)	110	12	0.02155856	1
path:hsa05152	Tuberculosis	132	12	0.0494497	1
path:hsa03440	Homologous recombination	38	5	0.05685433	1
path:hsa04145	Phagosome	107	10	0.05798064	1
path:hsa04710	Circadian rhythm	26	4	0.05898466	1
path:hsa02010	ABC transporters	41	5	0.06411138	1
path:hsa04022	cGMP-PKG signaling pathway	151	13	0.07901521	1
path:hsa00770	Pantothenate and CoA biosynthesis	19	3	0.088394	1
path:hsa04010	MAPK signaling pathway	266	21	0.09588912	1
path:hsa00510	N-Glycan biosynthesis	46	5	0.10692218	1
path:hsa04512	ECM-receptor interaction	82	8	0.10903752	1
path:hsa04216	Ferroptosis	32	4	0.10945049	1
path:hsa04015	Rap1 signaling pathway	194	16	0.10995179	1
path:hsa00230	Purine metabolism	113	10	0.11083658	1
path:hsa04110	Cell cycle	115	10	0.11278643	1
path:hsa05202	Transcriptional misregulation in cancer	148	13	0.11402134	1
path:hsa04970	Salivary secretion	75	7	0.11449105	1
path:hsa04520	Adherens junction	69	7	0.11881017	1

Blue module – BiB Pakistani – gene ontology

	Term	Ont	N	DE	P.DE	FDR
GO:1904646	cellular response to amyloid-beta	BP	24	6	9.82E-05	1
GO:1904645	response to amyloid-beta	BP	26	6	0.0001904	1
GO:0030576	Cajal body organization	BP	2	2	0.0008194	1
GO:0030175	filopodium	CC	87	10	0.00088559	1
GO:0030551	cyclic nucleotide binding	MF	35	6	0.00095124	1
GO:0031527	filopodium membrane	CC	16	4	0.00156805	1
GO:0010838	positive regulation of keratinocyte proliferation	BP	8	3	0.0017869	1
GO:0090286	cytoskeletal anchoring at nuclear membrane	BP	7	3	0.00185716	1
GO:2000620	positive regulation of histone H4-K16 acetylation	BP	2	2	0.00256097	1
GO:0034993	meiotic nuclear membrane microtubule tethering complex	CC	9	3	0.00308798	1
GO:0034992	microtubule organizing center attachment site	CC	9	3	0.00308798	1
GO:0106094	nuclear membrane microtubule tethering complex	CC	9	3	0.00308798	1
GO:0106083	nuclear membrane protein complex	CC	9	3	0.00308798	1
GO:2000670	positive regulation of dendritic cell apoptotic process	BP	3	2	0.00318197	1
GO:0030426	growth cone	CC	137	12	0.00341821	1
GO:0031871	proteinase activated receptor binding	MF	3	2	0.00402775	1
GO:0034136	negative regulation of toll-like receptor 2 signaling pathway	BP	3	2	0.00411014	1
GO:0071726	cellular response to diacyl bacterial lipopeptide	BP	4	2	0.00441829	1
GO:0071724	response to diacyl bacterial lipopeptide	BP	4	2	0.00441829	1
GO:0038124	toll-like receptor TLR6:TLR2 signaling pathway	BP	4	2	0.00441829	1

Blue module – BiB Pakistani – KEGG pathways

	Pathway	N	DE	P.DE	FDR
path:hsa00531	Glycosaminoglycan degradation	16	3	0.01448496	1
path:hsa04514	Cell adhesion molecules (CAMs)	110	8	0.02926196	1
path:hsa04512	ECM-receptor interaction	82	6	0.06131366	1
path:hsa05218	Melanoma	66	5	0.07271789	1
path:hsa04216	Ferroptosis	32	3	0.09029298	1
path:hsa04520	Adherens junction	69	5	0.0914664	1
path:hsa04530	Tight junction	147	8	0.10954649	1
path:hsa04730	Long-term depression	55	4	0.11353357	1
path:hsa04740	Olfactory transduction	204	8	0.12221525	1
path:hsa03440	Homologous recombination	38	3	0.12329545	1
path:hsa05213	Endometrial cancer	54	4	0.12792495	1
path:hsa00290	Valine, leucine and isoleucine biosynthesis	4	1	0.13202389	1
path:hsa05210	Colorectal cancer	82	5	0.14121597	1
path:hsa03018	RNA degradation	66	4	0.15941578	1
path:hsa00534	Glycosaminoglycan biosynthesis - heparan sulfate / heparin	21	2	0.16434161	1
path:hsa00630	Glyoxylate and dicarboxylate metabolism	26	2	0.18711681	1
path:hsa05224	Breast cancer	139	7	0.18718131	1
path:hsa04961	Endocrine and other factor-regulated calcium reabsorption	48	3	0.21102211	1
path:hsa00350	Tyrosine metabolism	28	2	0.21510615	1
path:hsa05412	Arrhythmogenic right ventricular cardiomyopathy (ARVC)	71	4	0.22354092	1

Brown module – ARIES – gene ontology

	Term	Ont	N	DE	P.DE	FDR
GO:0050793	regulation of developmental process	BP	2170	17	1.03E-05	0.22469758
GO:0022603	regulation of anatomical structure morphogenesis	BP	940	11	2.76E-05	0.30002778
GO:0008360	regulation of cell shape	BP	146	5	4.24E-05	0.30714259
GO:0022604	regulation of cell morphogenesis	BP	416	7	0.00020095	1
GO:0060033	anatomical structure regression	BP	11	2	0.00056428	1
GO:0090527	actin filament reorganization	BP	9	2	0.00057395	1
GO:0007166	cell surface receptor signaling pathway	BP	2469	15	0.00059319	1
GO:0009653	anatomical structure morphogenesis	BP	2315	15	0.0006552	1
GO:0051094	positive regulation of developmental process	BP	1140	10	0.0007428	1
GO:0072359	circulatory system development	BP	941	9	0.0007579	1
GO:0046628	positive regulation of insulin receptor signaling pathway	BP	15	2	0.00110191	1
GO:2000177	regulation of neural precursor cell proliferation	BP	75	3	0.0014421	1
GO:0051270	regulation of cellular component movement	BP	827	8	0.00151791	1
GO:0030546	receptor activator activity	MF	14	2	0.00154602	1
GO:0009967	positive regulation of signal transduction	BP	1370	10	0.00158186	1
GO:0005159	insulin-like growth factor receptor binding	MF	15	2	0.00165551	1
GO:0009966	regulation of signal transduction	BP	2746	15	0.00165782	1
GO:0051128	regulation of cellular component organization	BP	2094	13	0.00171114	1
GO:0010771	negative regulation of cell morphogenesis involved in differentiation	BP	80	3	0.00192921	1
GO:0007165	signal transduction	BP	4963	21	0.00197611	1

Brown module – ARIES - KEGG pathway

	Pathway	N	DE	P.DE	FDR
path:hsa04151	PI3K-Akt signaling pathway	308	4	0.01024624	1
path:hsa00730	Thiamine metabolism	12	1	0.02059298	1
path:hsa04512	ECM-receptor interaction	82	2	0.02135699	1
path:hsa04940	Type I diabetes mellitus	18	1	0.04624878	1
path:hsa04950	Maturity onset diabetes of the young	23	1	0.05555702	1
path:hsa05322	Systemic lupus erythematosus	40	1	0.07292125	1
path:hsa04390	Hippo signaling pathway	147	2	0.07443405	1
path:hsa05150	Staphylococcus aureus infection	54	1	0.0757346	1
path:hsa04960	Aldosterone-regulated sodium reabsorption	35	1	0.08239336	1
path:hsa04810	Regulation of actin cytoskeleton	194	2	0.09283691	1
path:hsa05140	Leishmaniasis	45	1	0.09405476	1
path:hsa04510	Focal adhesion	184	2	0.09798931	1
path:hsa04610	Complement and coagulation cascades	62	1	0.10147683	1
path:hsa04640	Hematopoietic cell lineage	66	1	0.10785369	1
path:hsa05134	Legionellosis	47	1	0.11205088	1
path:hsa05206	MicroRNAs in cancer	220	2	0.11436036	1
path:hsa04144	Endocytosis	215	2	0.11680858	1
path:hsa04930	Type II diabetes mellitus	42	1	0.11707907	1
path:hsa05133	Pertussis	65	1	0.11754075	1
path:hsa04913	Ovarian steroidogenesis	44	1	0.11885383	1

Brown module – BiB white British – gene ontology

	Term	Ont	N	DE	P.DE	FDR
GO:0031714	C5a anaphylatoxin chemotactic receptor binding	MF	1	1	0.00058664	0.91067083
GO:0031715	C5L2 anaphylatoxin chemotactic receptor binding	MF	1	1	0.00058664	0.91067083
GO:0001970	positive regulation of activation of membrane attack complex	BP	2	1	0.00058664	0.91067083
GO:0001798	positive regulation of type IIa hypersensitivity	BP	2	1	0.00058664	0.91067083
GO:0002894	positive regulation of type II hypersensitivity	BP	2	1	0.00058664	0.91067083
GO:0001796	regulation of type IIa hypersensitivity	BP	2	1	0.00058664	0.91067083
GO:0002892	regulation of type II hypersensitivity	BP	2	1	0.00058664	0.91067083
GO:0001794	type IIa hypersensitivity	BP	2	1	0.00058664	0.91067083
GO:0002445	type II hypersensitivity	BP	2	1	0.00058664	0.91067083
GO:0001905	activation of membrane attack complex	BP	3	1	0.00058664	0.91067083
GO:0045917	positive regulation of complement activation	BP	3	1	0.00058664	0.91067083
GO:2000259	positive regulation of protein activation cascade	BP	3	1	0.00058664	0.91067083
GO:0001969	regulation of activation of membrane attack complex	BP	3	1	0.00058664	0.91067083
GO:0006957	complement activation, alternative pathway	BP	7	1	0.00058664	0.91067083
GO:0021541	ammon gyrus development	BP	1	1	0.00071942	0.97719612
GO:0038025	reelin receptor activity	MF	2	1	0.00071942	0.97719612
GO:0002885	positive regulation of hypersensitivity	BP	4	1	0.00089574	1
GO:0002883	regulation of hypersensitivity	BP	5	1	0.00089574	1
GO:0002922	positive regulation of humoral immune response	BP	9	1	0.0009082	1
GO:0030229	very-low-density lipoprotein particle receptor activity	MF	3	1	0.00097619	1

Brown module – BiB white British – KEGG pathway

path:hsa05150	Staphylococcus aureus infection	54	1	0.00204284	0.5790524
path:hsa05322	Systemic lupus erythematosus	40	1	0.00547908	0.5790524
path:hsa04610	Complement and coagulation cascades	62	1	0.00602376	0.5790524
path:hsa05140	Leishmaniasis	45	1	0.00775063	0.5790524
path:hsa05133	Pertussis	65	1	0.00872829	0.5790524
path:hsa05134	Legionellosis	47	1	0.01034022	0.5790524
path:hsa05142	Chagas disease (American trypanosomiasis)	89	1	0.01683318	0.76107721
path:hsa04145	Phagosome	107	1	0.01812089	0.76107721
path:hsa05152	Tuberculosis	132	1	0.02500128	0.93338125
path:hsa05167	Kaposi sarcoma-associated herpesvirus infection	148	1	0.03203891	1
path:hsa05203	Viral carcinogenesis	147	1	0.03539493	1
path:hsa05168	Herpes simplex virus 1 infection	372	1	0.03782349	1
path:hsa04080	Neuroactive ligand-receptor interaction	277	1	0.04672435	1
path:hsa04921	Oxytocin signaling pathway	141	0	1	1
path:hsa00472	D-Arginine and D-ornithine metabolism	1	0	1	1
path:hsa00780	Biotin metabolism	3	0	1	1
path:hsa00232	Caffeine metabolism	3	0	1	1
path:hsa00785	Lipoic acid metabolism	3	0	1	1
path:hsa00400	Phenylalanine, tyrosine and tryptophan biosynthesis	3	0	1	1
path:hsa00471	D-Glutamine and D-glutamate metabolism	4	0	1	1

Brown module – BiB Pakistani – gene ontology

	Term	Ont	N	DE	P.DE	FDR
GO:0031714	C5a anaphylatoxin chemotactic receptor binding	MF	1	1	0.00023965	0.47349031
GO:0031715	C5L2 anaphylatoxin chemotactic receptor binding	MF	1	1	0.00023965	0.47349031
GO:0001970	positive regulation of activation of membrane attack complex	BP	2	1	0.00023965	0.47349031
GO:0001798	positive regulation of type IIa hypersensitivity	BP	2	1	0.00023965	0.47349031
GO:0002894	positive regulation of type II hypersensitivity	BP	2	1	0.00023965	0.47349031
GO:0001796	regulation of type IIa hypersensitivity	BP	2	1	0.00023965	0.47349031
GO:0002892	regulation of type II hypersensitivity	BP	2	1	0.00023965	0.47349031
GO:0001794	type IIa hypersensitivity	BP	2	1	0.00023965	0.47349031
GO:0002445	type II hypersensitivity	BP	2	1	0.00023965	0.47349031
GO:0045917	positive regulation of complement activation	BP	3	1	0.00023965	0.47349031
GO:2000259	positive regulation of protein activation cascade	BP	3	1	0.00023965	0.47349031
GO:0002885	positive regulation of hypersensitivity	BP	4	1	0.00059965	0.87218883
GO:0002883	regulation of hypersensitivity	BP	5	1	0.00060818	0.87218883
GO:0001905	activation of membrane attack complex	BP	3	1	0.00062682	0.87218883
GO:0001969	regulation of activation of membrane attack complex	BP	3	1	0.00062682	0.87218883
GO:0045017	glycerolipid biosynthetic process	BP	239	2	0.00064211	0.87218883
GO:0009967	positive regulation of signal transduction	BP	1370	3	0.00070237	0.88781956
GO:0002524	hypersensitivity	BP	6	1	0.0007676	0.88781956
GO:0006957	complement activation, alternative pathway	BP	7	1	0.00077617	0.88781956
GO:2000427	positive regulation of apoptotic cell clearance	BP	3	1	0.00082729	0.89897407

Brown module – BiB Pakistani – KEGG pathway

	Pathway	N	DE	P.DE	FDR
path:hsa05322	Systemic lupus erythematosus	40	1	0.00539111	0.55795458
path:hsa05150	Staphylococcus aureus infection	54	1	0.00564183	0.55795458
path:hsa05140	Leishmaniasis	46	1	0.00834202	0.55795458
path:hsa05134	Legionellosis	48	1	0.00958301	0.55795458
path:hsa05133	Pertussis	65	1	0.00970975	0.55795458
path:hsa04610	Complement and coagulation cascades	68	1	0.00996347	0.55795458
path:hsa05142	Chagas disease (American trypanosomiasis)	89	1	0.01453586	0.69772134
path:hsa04145	Phagosome	107	1	0.02238238	0.83323844
path:hsa05152	Tuberculosis	133	1	0.02425682	0.83323844
path:hsa05167	Kaposi sarcoma-associated herpesvirus infection	148	1	0.02636649	0.83323844
path:hsa05203	Viral carcinogenesis	147	1	0.02727864	0.83323844
path:hsa04080	Neuroactive ligand-receptor interaction	277	1	0.04665969	1
path:hsa05168	Herpes simplex virus 1 infection	372	1	0.05633802	1
path:hsa00472	D-Arginine and D-ornithine metabolism	1	0	1	1
path:hsa00780	Biotin metabolism	3	0	1	1
path:hsa00232	Caffeine metabolism	3	0	1	1
path:hsa00785	Lipoic acid metabolism	3	0	1	1
path:hsa00400	Phenylalanine, tyrosine and tryptophan biosynthesis	3	0	1	1
path:hsa00471	D-Glutamine and D-glutamate metabolism	4	0	1	1
path:hsa00290	Valine, leucine and isoleucine biosynthesis	4	0	1	1

Dark red module – ARIES – gene ontology

	Term	Ont	N	DE	P.DE	FDR
GO:0031062	positive regulation of histone methylation	BP	26	3	8.60E-05	0.81856045
GO:0007099	centriole replication	BP	29	3	9.29E-05	0.81856045
GO:0098534	centriole assembly	BP	31	3	0.00011299	0.81856045
GO:0051574	positive regulation of histone H3-K9 methylation	BP	7	2	0.00023703	1
GO:0043569	negative regulation of insulin-like growth factor receptor signaling pathway	BP	8	2	0.00030339	1
GO:0031060	regulation of histone methylation	BP	48	3	0.00056125	1
GO:0006349	regulation of gene expression by genetic imprinting	BP	15	2	0.00098397	1
GO:0051298	centrosome duplication	BP	62	3	0.00120539	1
GO:0032039	integrator complex	CC	15	2	0.00126306	1
GO:0051571	positive regulation of histone H3-K4 methylation	BP	14	2	0.00126917	1
GO:0031058	positive regulation of histone modification	BP	71	3	0.00149965	1
GO:0016180	snRNA processing	BP	21	2	0.00162585	1
GO:0051570	regulation of histone H3-K9 methylation	BP	17	2	0.00173867	1
GO:2001022	positive regulation of response to DNA damage stimulus	BP	72	3	0.00181247	1
GO:0043567	regulation of insulin-like growth factor receptor signaling pathway	BP	18	2	0.00185804	1
GO:1905269	positive regulation of chromatin organization	BP	78	3	0.00187319	1
GO:0071514	genetic imprinting	BP	22	2	0.00190387	1
GO:0006396	RNA processing	BP	809	8	0.00219058	1
GO:0051567	histone H3-K9 methylation	BP	24	2	0.00279364	1
GO:0051569	regulation of histone H3-K4 methylation	BP	21	2	0.00282006	1

Dark red module – ARIES – KEGG

	Pathway	N	DE	P.DE	FDR
path:hsa03060	Protein export	21	1	0.05650899	1
path:hsa00350	Tyrosine metabolism	28	1	0.06672273	1
path:hsa00640	Propanoate metabolism	29	1	0.07150177	1
path:hsa03410	Base excision repair	29	1	0.07505706	1
path:hsa04151	PI3K-Akt signaling pathway	308	3	0.07526199	1
path:hsa00620	Pyruvate metabolism	33	1	0.08268697	1
path:hsa00250	Alanine, aspartate and glutamate metabolism	33	1	0.08589577	1
path:hsa00270	Cysteine and methionine metabolism	42	1	0.09670017	1
path:hsa04130	SNARE interactions in vesicular transport	32	1	0.09956187	1
path:hsa00010	Glycolysis / Gluconeogenesis	55	1	0.11599104	1
path:hsa04960	Aldosterone-regulated sodium reabsorption	35	1	0.11889029	1
path:hsa03440	Homologous recombination	38	1	0.1189837	1
path:hsa00600	Sphingolipid metabolism	38	1	0.12123856	1
path:hsa05134	Legionellosis	47	1	0.13324612	1
path:hsa05206	MicroRNAs in cancer	220	2	0.13546281	1
path:hsa03460	Fanconi anemia pathway	47	1	0.14457237	1
path:hsa01524	Platinum drug resistance	62	1	0.17683012	1
path:hsa05230	Central carbon metabolism in cancer	66	1	0.19740792	1
path:hsa03015	mRNA surveillance pathway	74	1	0.21662585	1
path:hsa04911	Insulin secretion	77	1	0.23829649	1

Dark red module – BiB white British – gene ontology

	Term	Ont	N	DE	P.DE	FDR
GO:1902512	positive regulation of apoptotic DNA fragmentation	BP	3	1	0.00115357	1
GO:1901300	positive regulation of hydrogen peroxide-mediated programmed cell death	BP	3	1	0.0014377	1
GO:1903626	positive regulation of DNA catabolic process	BP	4	1	0.00166063	1
GO:1905206	positive regulation of hydrogen peroxide-induced cell death	BP	4	1	0.00167615	1
GO:0016788	hydrolase activity, acting on ester bonds	MF	636	3	0.00175861	1
GO:1901033	positive regulation of response to reactive oxygen species	BP	5	1	0.00207483	1
GO:2001040	positive regulation of cellular response to drug	BP	5	1	0.00209445	1
GO:1902510	regulation of apoptotic DNA fragmentation	BP	5	1	0.00214037	1
GO:0010891	negative regulation of sequestering of triglyceride	BP	5	1	0.00217475	1
GO:0010898	positive regulation of triglyceride catabolic process	BP	6	1	0.0022699	1
GO:0031329	regulation of cellular catabolic process	BP	704	3	0.00247764	1
GO:1903624	regulation of DNA catabolic process	BP	6	1	0.002647	1
GO:0051006	positive regulation of lipoprotein lipase activity	BP	9	1	0.00328594	1
GO:0004726	non-membrane spanning protein tyrosine phosphatase activity	MF	7	1	0.0033955	1
GO:0009894	regulation of catabolic process	BP	792	3	0.00348821	1
GO:0061365	positive regulation of triglyceride lipase activity	BP	10	1	0.00352591	1
GO:0010896	regulation of triglyceride catabolic process	BP	10	1	0.00370178	1
GO:2000587	negative regulation of platelet-derived growth factor receptor-beta signaling pathway	BP	8	1	0.00375642	1
GO:2000586	regulation of platelet-derived growth factor receptor-beta signaling pathway	BP	10	1	0.00475228	1

Dark red module – BiB white British – KEGG

	Pathway	N	DE	P.DE	FDR
path:hsa04923	Regulation of lipolysis in adipocytes	51	1	0.02038793	1
path:hsa03040	Spliceosome	108	1	0.04609513	1
path:hsa04210	Apoptosis	118	1	0.049132	1
path:hsa00790	Folate biosynthesis	23	0	1	1
path:hsa00472	D-Arginine and D-ornithine metabolism	1	0	1	1
path:hsa00780	Biotin metabolism	3	0	1	1
path:hsa00232	Caffeine metabolism	3	0	1	1
path:hsa00785	Lipoic acid metabolism	3	0	1	1
path:hsa00400	Phenylalanine, tyrosine and tryptophan biosynthesis	3	0	1	1
path:hsa00471	D-Glutamine and D-glutamate metabolism	4	0	1	1
path:hsa00290	Valine, leucine and isoleucine biosynthesis	4	0	1	1
path:hsa00524	Neomycin, kanamycin and gentamicin biosynthesis	5	0	1	1
path:hsa00440	Phosphonate and phosphinate metabolism	5	0	1	1
path:hsa00750	Vitamin B6 metabolism	5	0	1	1
path:hsa04122	Sulfur relay system	6	0	1	1
path:hsa00740	Riboflavin metabolism	8	0	1	1
path:hsa00072	Synthesis and degradation of ketone bodies	8	0	1	1
path:hsa00130	Ubiquinone and other terpenoid-quinone biosynthesis	8	0	1	1
path:hsa00430	Taurine and hypotaurine metabolism	9	0	1	1
path:hsa05330	Allograft rejection	10	0	1	1

Dark red module – BiB Pakistani – gene ontology

	Term	Ont	N	DE	P.DE	FDR
GO:0016580	Sin3 complex	CC	13	3	7.43E-07	0.01614101
GO:0070822	Sin3-type complex	CC	16	3	1.49E-06	0.01620988
GO:0000118	histone deacetylase complex	CC	55	3	4.57E-05	0.32186165
GO:0045292	mRNA cis splicing, via spliceosome	BP	14	2	6.93E-05	0.32186165
GO:0017053	transcriptional repressor complex	CC	79	3	9.76E-05	0.32186165
GO:0005654	nucleoplasm	CC	2978	11	0.00011694	0.32186165
GO:0070013	intracellular organelle lumen	CC	4418	13	0.00013329	0.32186165
GO:0031974	membrane-enclosed lumen	CC	4418	13	0.00013329	0.32186165
GO:0043233	organelle lumen	CC	4418	13	0.00013329	0.32186165
GO:0046655	folic acid metabolic process	BP	20	2	0.00022146	0.48093677
GO:1902494	catalytic complex	CC	1195	7	0.00024342	0.48093677
GO:0034641	cellular nitrogen compound metabolic process	BP	5652	14	0.00027428	0.49674444
GO:0006760	folic acid-containing compound metabolic process	BP	28	2	0.00037706	0.57506406
GO:0001106	RNA polymerase II transcription corepressor activity	MF	24	2	0.00039247	0.57506406
GO:0044446	intracellular organelle part	CC	7678	16	0.00044946	0.57506406
GO:0031981	nuclear lumen	CC	3468	11	0.00046318	0.57506406
GO:0033737	1-pyrroline dehydrogenase activity	MF	1	1	0.00054279	0.57506406
GO:0047105	4-trimethylammoniobutyraldehyde dehydrogenase activity	MF	1	1	0.00054279	0.57506406
GO:0019145	aminobutyraldehyde dehydrogenase activity	MF	1	1	0.00054279	0.57506406
GO:0046483	heterocycle metabolic process	BP	5120	13	0.00055215	0.57506406

Dark red module – BiB Pakistani – KEGG

	Pathway	N	DE	P.DE	FDR
path:hsa03040	Spliceosome	108	2	0.00755959	0.97135713
path:hsa00053	Ascorbate and aldarate metabolism	22	1	0.01989958	0.97135713
path:hsa00340	Histidine metabolism	21	1	0.02099859	0.97135713
path:hsa00670	One carbon pool by folate	19	1	0.02216699	0.97135713
path:hsa00790	Folate biosynthesis	23	1	0.02394996	0.97135713
path:hsa01523	Antifolate resistance	28	1	0.02923125	0.97135713
path:hsa00071	Fatty acid degradation	34	1	0.03235679	0.97135713
path:hsa00410	beta-Alanine metabolism	30	1	0.03250751	0.97135713
path:hsa00380	Tryptophan metabolism	35	1	0.03261305	0.97135713
path:hsa03020	RNA polymerase	28	1	0.03270772	0.97135713
path:hsa00620	Pyruvate metabolism	33	1	0.03517927	0.97135713
path:hsa03050	Proteasome	34	1	0.03685491	0.97135713
path:hsa00280	Valine, leucine and isoleucine degradation	39	1	0.04120268	0.97135713
path:hsa00010	Glycolysis / Gluconeogenesis	55	1	0.04700711	0.97135713
path:hsa04623	Cytosolic DNA-sensing pathway	44	1	0.04831609	0.97135713
path:hsa00561	Glycerolipid metabolism	51	1	0.04862515	0.97135713
path:hsa00330	Arginine and proline metabolism	40	1	0.04914605	0.97135713
path:hsa00310	Lysine degradation	53	1	0.06182501	1
path:hsa03008	Ribosome biogenesis in eukaryotes	60	1	0.06398975	1
path:hsa04137	Mitophagy - animal	60	1	0.07235084	1

Grey60 module – BiB white British – gene ontology

	Term	Ont	N	DE	P.DE	FDR
GO:0042110	T cell activation	BP	369	4	5.15E-05	0.95348639
GO:0006955	immune response	BP	1554	6	9.64E-05	0.95348639
GO:0050690	regulation of defense response to virus by virus	BP	27	2	0.00013887	0.95348639
GO:0046649	lymphocyte activation	BP	512	4	0.00017549	0.95348639
GO:0031295	T cell costimulation	BP	49	2	0.0004296	1
GO:0031294	lymphocyte costimulation	BP	50	2	0.00044367	1
GO:0002376	immune system process	BP	2297	6	0.00097321	1
GO:0048619	embryonic hindgut morphogenesis	BP	1	1	0.00099516	1
GO:1903006	positive regulation of protein K63-linked deubiquitination	BP	2	1	0.00110295	1
GO:1903004	regulation of protein K63-linked deubiquitination	BP	2	1	0.00110295	1
GO:0002757	immune response-activating signal transduction	BP	397	3	0.00149387	1
GO:1901837	negative regulation of transcription of nucleolar large rRNA by RNA polymerase I	BP	3	1	0.00161752	1
GO:0002764	immune response-regulating signaling pathway	BP	423	3	0.00177428	1
GO:0045321	leukocyte activation	BP	961	4	0.0018887	1
GO:0042610	CD8 receptor binding	MF	2	1	0.00196976	1
GO:0016479	negative regulation of transcription by RNA polymerase I	BP	4	1	0.00213506	1
GO:0002253	activation of immune response	BP	456	3	0.0022083	1
GO:0030538	embryonic genitalia morphogenesis	BP	3	1	0.00233129	1
GO:0008454	alpha-1,3-mannosylglycoprotein 4-beta-N-acetylglucosaminyltransferase activity	MF	3	1	0.00242167	1
GO:0033553	rDNA heterochromatin	CC	4	1	0.00268111	1

Grey60 module – BiB white British – KEGG

	Pathway	N	DE	P.DE	FDR
path:hsa04658	Th1 and Th2 cell differentiation	67	2	0.0007961	0.08761379
path:hsa04650	Natural killer cell mediated cytotoxicity	78	2	0.00093583	0.08761379
path:hsa04659	Th17 cell differentiation	80	2	0.00102545	0.08761379
path:hsa05235	PD-L1 expression and PD-1 checkpoint pathway in cancer	81	2	0.00109314	0.08761379
path:hsa04660	T cell receptor signaling pathway	90	2	0.00130378	0.08761379
path:hsa05340	Primary immunodeficiency	29	1	0.01964211	0.95348797
path:hsa05216	Thyroid cancer	34	1	0.02226286	0.95348797
path:hsa00513	Various types of N-glycan biosynthesis	37	1	0.02562686	0.95348797
path:hsa00510	N-Glycan biosynthesis	46	1	0.02647881	0.95348797
path:hsa05213	Endometrial cancer	54	1	0.03675405	0.95348797
path:hsa05217	Basal cell carcinoma	61	1	0.03898589	0.95348797
path:hsa05221	Acute myeloid leukemia	59	1	0.04151685	0.95348797
path:hsa04640	Hematopoietic cell lineage	66	1	0.04187227	0.95348797
path:hsa04520	Adherens junction	69	1	0.04446866	0.95348797
path:hsa04064	NF-kappa B signaling pathway	79	1	0.0453614	0.95348797
path:hsa05412	Arrhythmogenic right ventricular cardiomyopathy (ARVC)	71	1	0.048007	0.95348797
path:hsa05142	Chagas disease (American trypanosomiasis)	89	1	0.0510469	0.95348797
path:hsa05210	Colorectal cancer	82	1	0.05324941	0.95348797
path:hsa04916	Melanogenesis	96	1	0.05837888	0.95348797
path:hsa04380	Osteoclast differentiation	105	1	0.05870955	0.95348797

Grey60 module – BiB Pakistani – gene ontology

	Term	Ont	N	DE	P.DE	FDR
GO:0031295	T cell costimulation	BP	49	3	2.87E-06	0.03063898
GO:0031294	lymphocyte costimulation	BP	50	3	2.92E-06	0.03063898
GO:0042105	alpha-beta T cell receptor complex	CC	5	2	4.49E-06	0.03063898
GO:0042608	T cell receptor binding	MF	6	2	5.64E-06	0.03063898
GO:0042110	T cell activation	BP	369	4	3.24E-05	0.14072785
GO:0042101	T cell receptor complex	CC	18	2	5.56E-05	0.19782691
GO:0050852	T cell receptor signaling pathway	BP	147	3	6.37E-05	0.19782691
GO:0050870	positive regulation of T cell activation	BP	161	3	8.02E-05	0.21781643
GO:1903039	positive regulation of leukocyte cell-cell adhesion	BP	169	3	9.27E-05	0.22391587
GO:0046649	lymphocyte activation	BP	512	4	0.00010909	0.23709412
GO:0050851	antigen receptor-mediated signaling pathway	BP	183	3	0.00012729	0.25149354
GO:0022409	positive regulation of cell-cell adhesion	BP	200	3	0.00015124	0.26824431
GO:0051251	positive regulation of lymphocyte activation	BP	210	3	0.00017089	0.26824431
GO:0030217	T cell differentiation	BP	200	3	0.0001728	0.26824431
GO:1903037	regulation of leukocyte cell-cell adhesion	BP	233	3	0.00024462	0.34316614
GO:0002696	positive regulation of leukocyte activation	BP	241	3	0.00025264	0.34316614
GO:0050867	positive regulation of cell activation	BP	251	3	0.00029248	0.37390613
GO:0007159	leukocyte cell-cell adhesion	BP	263	3	0.00033114	0.38457644
GO:0050863	regulation of T cell activation	BP	253	3	0.00033621	0.38457644
GO:0002429	immune response-activating cell surface receptor signaling pathway	BP	274	3	0.00038599	0.41943848

Grey60 module – BiB Pakistani – KEGG

	Pathway	N	DE	P.DE	FDR
path:hsa04658	Th1 and Th2 cell differentiation	67	3	5.98E-06	0.00094723
path:hsa04659	Th17 cell differentiation	80	3	8.98E-06	0.00094723
path:hsa05235	PD-L1 expression and PD-1 checkpoint pathway in cancer	81	3	9.87E-06	0.00094723
path:hsa05142	Chagas disease (American trypanosomiasis)	89	3	1.13E-05	0.00094723
path:hsa04660	T cell receptor signaling pathway	90	3	1.49E-05	0.00099914
path:hsa05162	Measles	104	3	2.34E-05	0.0013087
path:hsa05170	Human immunodeficiency virus 1 infection	165	3	6.80E-05	0.00326527
path:hsa05166	Human T-cell leukemia virus 1 infection	182	3	9.44E-05	0.00396576
path:hsa05340	Primary immunodeficiency	29	2	0.00014505	0.00541532
path:hsa04650	Natural killer cell mediated cytotoxicity	78	2	0.00073672	0.02269601
path:hsa04640	Hematopoietic cell lineage	67	2	0.00074302	0.02269601
path:hsa05169	Epstein-Barr virus infection	146	2	0.00269663	0.07550574
path:hsa05330	Allograft rejection	10	1	0.00530719	0.13717033
path:hsa05332	Graft-versus-host disease	12	1	0.00623269	0.14958451
path:hsa05320	Autoimmune thyroid disease	15	1	0.00757265	0.16962737
path:hsa04940	Type I diabetes mellitus	18	1	0.00938578	0.19710139
path:hsa05165	Human papillomavirus infection	288	2	0.0103292	0.2041536
path:hsa05143	African trypanosomiasis	31	1	0.01810915	0.32076684
path:hsa05216	Thyroid cancer	34	1	0.01853926	0.32076684
path:hsa03440	Homologous recombination	38	1	0.01909326	0.32076684

Light green module – ARIES – gene ontology

	Term	Ont	N	DE	P.DE	FDR
GO:0007156	homophilic cell adhesion via plasma membrane adhesion molecules	BP	135	15	4.04E-09	8.78E-05
GO:0098742	cell-cell adhesion via plasma-membrane adhesion molecules	BP	210	15	1.25E-06	0.01355944
GO:0005654	nucleoplasm	CC	2978	68	4.28E-05	0.31025456
GO:0031981	nuclear lumen	CC	3468	75	6.02E-05	0.32712246
GO:0044428	nuclear part	CC	3767	79	9.68E-05	0.42090921
GO:0005634	nucleus	CC	6047	110	0.00013073	0.4735133
GO:0046949	fatty-acyl-CoA biosynthetic process	BP	23	4	0.00042308	1
GO:0006633	fatty acid biosynthetic process	BP	107	7	0.00048055	1
GO:0043231	intracellular membrane-bounded organelle	CC	9069	147	0.0005533	1
GO:0046394	carboxylic acid biosynthetic process	BP	351	13	0.00059634	1
GO:0016053	organic acid biosynthetic process	BP	352	13	0.00060446	1
GO:0098609	cell-cell adhesion	BP	661	21	0.0007567	1
GO:0000462	maturation of SSU-rRNA from tricistronic rRNA transcript (SSU-rRNA, 5.8S rRNA, LSU-rRNA)	BP	32	4	0.00100063	1
GO:0003674	molecular_function	MF	14138	201	0.00108795	1
GO:0005488	binding	MF	12314	185	0.00112825	1
GO:0047499	calcium-independent phospholipase A2 activity	MF	5	2	0.00137783	1
GO:1901570	fatty acid derivative biosynthetic process	BP	67	5	0.0014671	1
GO:0072330	monocarboxylic acid biosynthetic process	BP	249	10	0.00147299	1
GO:0035337	fatty-acyl-CoA metabolic process	BP	31	4	0.00147881	1
GO:0035338	long-chain fatty-acyl-CoA biosynthetic process	BP	16	3	0.00172672	1

Light green module – ARIES – KEGG

	Pathway	N	DE	P.DE	FDR
path:hsa00062	Fatty acid elongation	24	3	0.00424188	0.44274553
path:hsa03020	RNA polymerase	28	3	0.00500714	0.44274553
path:hsa01040	Biosynthesis of unsaturated fatty acids	23	3	0.00508469	0.44274553
path:hsa03460	Fanconi anemia pathway	47	4	0.00527078	0.44274553
path:hsa05416	Viral myocarditis	34	3	0.0147506	0.99124065
path:hsa04962	Vasopressin-regulated water reabsorption	43	3	0.02228139	1
path:hsa00770	Pantothenate and CoA biosynthesis	19	2	0.02558854	1
path:hsa04659	Th17 cell differentiation	80	4	0.0264608	1
path:hsa01212	Fatty acid metabolism	48	3	0.03493174	1
path:hsa05132	Salmonella infection	67	3	0.05381271	1
path:hsa04218	Cellular senescence	135	5	0.05447431	1
path:hsa05135	Yersinia infection	106	4	0.06355406	1
path:hsa03022	Basal transcription factors	31	2	0.06693175	1
path:hsa03050	Proteasome	34	2	0.06711124	1
path:hsa04130	SNARE interactions in vesicular transport	32	2	0.07341975	1
path:hsa04612	Antigen processing and presentation	31	2	0.07426088	1
path:hsa05215	Prostate cancer	91	4	0.07563144	1
path:hsa05332	Graft-versus-host disease	12	1	0.08033007	1
path:hsa05330	Allograft rejection	10	1	0.08101456	1
path:hsa05100	Bacterial invasion of epithelial cells	69	3	0.10039667	1

Light green module – BiB white British – gene ontology

	Term	Ont	N	DE	P.DE	FDR
GO:1903690	negative regulation of wound healing, spreading of epidermal cells	BP	4	1	0.00110864	1
GO:1903689	regulation of wound healing, spreading of epidermal cells	BP	6	1	0.00135729	1
GO:0002162	dystroglycan binding	MF	8	1	0.00162098	1
GO:0045180	basal cortex	CC	5	1	0.00170878	1
GO:0005828	kinetochore microtubule	CC	7	1	0.00188599	1
GO:2001197	basement membrane assembly involved in embryonic body morphogenesis	BP	5	1	0.00191003	1
GO:1904261	positive regulation of basement membrane assembly involved in embryonic body morphogenesis	BP	5	1	0.00191003	1
GO:1904259	regulation of basement membrane assembly involved in embryonic body morphogenesis	BP	5	1	0.00191003	1
GO:0110011	regulation of basement membrane organization	BP	5	1	0.00191003	1
GO:0090091	positive regulation of extracellular matrix disassembly	BP	7	1	0.00211826	1
GO:0070831	basement membrane assembly	BP	8	1	0.00286633	1
GO:1901203	positive regulation of extracellular matrix assembly	BP	9	1	0.00328691	1
GO:0035313	wound healing, spreading of epidermal cells	BP	16	1	0.00368782	1
GO:0010715	regulation of extracellular matrix disassembly	BP	14	1	0.00376874	1
GO:1905331	negative regulation of morphogenesis of an epithelium	BP	15	1	0.00377475	1
GO:0010172	embryonic body morphogenesis	BP	11	1	0.00384577	1
GO:0051497	negative regulation of stress fiber assembly	BP	17	1	0.0038835	1
GO:1901201	regulation of extracellular matrix assembly	BP	13	1	0.00390118	1
GO:0051010	microtubule plus-end binding	MF	13	1	0.00431411	1

Light green module – BiB white British – KEGG

	Pathway	N	DE	P.DE	FDR
path:hsa05168	Herpes simplex virus 1 infection	372	1	0.07481026	1
path:hsa00472	D-Arginine and D-ornithine metabolism	1	0	1	1
path:hsa00780	Biotin metabolism	3	0	1	1
path:hsa00232	Caffeine metabolism	3	0	1	1
path:hsa00785	Lipoic acid metabolism	3	0	1	1
path:hsa00400	Phenylalanine, tyrosine and tryptophan biosynthesis	3	0	1	1
path:hsa00471	D-Glutamine and D-glutamate metabolism	4	0	1	1
path:hsa00290	Valine, leucine and isoleucine biosynthesis	4	0	1	1
path:hsa00524	Neomycin, kanamycin and gentamicin biosynthesis	5	0	1	1
path:hsa00440	Phosphonate and phosphinate metabolism	5	0	1	1
path:hsa00750	Vitamin B6 metabolism	5	0	1	1
path:hsa04122	Sulfur relay system	6	0	1	1
path:hsa00740	Riboflavin metabolism	8	0	1	1
path:hsa00072	Synthesis and degradation of ketone bodies	8	0	1	1
path:hsa00130	Ubiquinone and other terpenoid-quinone biosynthesis	8	0	1	1
path:hsa00430	Taurine and hypotaurine metabolism	9	0	1	1
path:hsa05330	Allograft rejection	10	0	1	1
path:hsa00920	Sulfur metabolism	10	0	1	1
path:hsa05310	Asthma	12	0	1	1
path:hsa05332	Graft-versus-host disease	12	0	1	1

Light green module – BiB Pakistani – gene ontology

	Term	Ont	N	DE	P.DE	FDR
GO:1901362	organic cyclic compound biosynthetic process	BP	3932	29	1.86E-06	0.04044006
GO:0018130	heterocycle biosynthetic process	BP	3805	27	1.40E-05	0.14567517
GO:1901360	organic cyclic compound metabolic process	BP	5332	32	2.14E-05	0.14567517
GO:0090185	negative regulation of kidney development	BP	15	3	2.68E-05	0.14567517
GO:0034654	nucleobase-containing compound biosynthetic process	BP	3749	26	3.85E-05	0.16741397
GO:0019438	aromatic compound biosynthetic process	BP	3808	26	4.86E-05	0.1760045
GO:0072111	cell proliferation involved in kidney development	BP	17	3	7.30E-05	0.22651979
GO:0046483	heterocycle metabolic process	BP	5120	30	0.0001007	0.27355206
GO:0006351	transcription, DNA-templated	BP	3256	23	0.00013281	0.27408532
GO:0097659	nucleic acid-templated transcription	BP	3267	23	0.0001427	0.27408532
GO:0032774	RNA biosynthetic process	BP	3281	23	0.00014913	0.27408532
GO:0031981	nuclear lumen	CC	3468	24	0.00017269	0.27408532
GO:0006139	nucleobase-containing compound metabolic process	BP	4977	29	0.00018226	0.27408532
GO:0005654	nucleoplasm	CC	2978	22	0.0001882	0.27408532
GO:0071440	regulation of histone H3-K14 acetylation	BP	5	2	0.00019504	0.27408532
GO:0090304	nucleic acid metabolic process	BP	4452	27	0.00022261	0.27408532
GO:0003677	DNA binding	MF	2090	17	0.00022897	0.27408532
GO:0070013	intracellular organelle lumen	CC	4418	27	0.0002713	0.27408532
GO:0031974	membrane-enclosed lumen	CC	4418	27	0.0002713	0.27408532
GO:0043233	organelle lumen	CC	4418	27	0.0002713	0.27408532

Light green module – BiB Pakistani – KEGG

	Pathway	N	DE	P.DE	FDR
path:hsa00310	Lysine degradation	53	3	0.00109712	0.36863077
path:hsa03013	RNA transport	127	3	0.00741419	1
path:hsa05202	Transcriptional misregulation in cancer	148	3	0.02320218	1
path:hsa03015	mRNA surveillance pathway	74	2	0.03013058	1
path:hsa00120	Primary bile acid biosynthesis	14	1	0.04950096	1
path:hsa05168	Herpes simplex virus 1 infection	372	3	0.05023645	1
path:hsa04928	Parathyroid hormone synthesis, secretion and action	100	2	0.05943172	1
path:hsa00030	Pentose phosphate pathway	24	1	0.06335636	1
path:hsa03030	DNA replication	32	1	0.06791382	1
path:hsa05321	Inflammatory bowel disease (IBD)	43	1	0.10900091	1
path:hsa00330	Arginine and proline metabolism	40	1	0.11556148	1
path:hsa04978	Mineral absorption	45	1	0.13318072	1
path:hsa01230	Biosynthesis of amino acids	57	1	0.13535594	1
path:hsa05134	Legionellosis	48	1	0.1539708	1
path:hsa03008	Ribosome biogenesis in eukaryotes	60	1	0.15739603	1
path:hsa04340	Hedgehog signaling pathway	47	1	0.16755081	1
path:hsa04658	Th1 and Th2 cell differentiation	67	1	0.20613037	1
path:hsa05217	Basal cell carcinoma	61	1	0.20774433	1
path:hsa04260	Cardiac muscle contraction	73	1	0.20808985	1
path:hsa04659	Th17 cell differentiation	80	1	0.23644931	1

Magenta module – ARIES – gene ontology

	Term	Ont	N	DE	P.DE	FDR
GO:1903515	calcium ion transport from cytosol to endoplasmic reticulum	BP	1	1	0.00132788	1
GO:0086039	calcium-transporting ATPase activity involved in regulation of cardiac muscle cell membrane potential	MF	1	1	0.00132788	1
GO:0031775	lutropin-choriogonadotropic hormone receptor binding	MF	1	1	0.00132788	1
GO:1903233	regulation of calcium ion-dependent exocytosis of neurotransmitter	BP	1	1	0.00132788	1
GO:1900753	doxorubicin transport	BP	1	1	0.00159748	1
GO:1900148	negative regulation of Schwann cell migration	BP	1	1	0.00174617	1
GO:1905045	negative regulation of Schwann cell proliferation involved in axon regeneration	BP	1	1	0.00174617	1
GO:1905044	regulation of Schwann cell proliferation involved in axon regeneration	BP	1	1	0.00174617	1
GO:0014011	Schwann cell proliferation involved in axon regeneration	BP	1	1	0.00174617	1
GO:1990036	calcium ion import into sarcoplasmic reticulum	BP	2	1	0.00176116	1
GO:0003863	3-methyl-2-oxobutanoate dehydrogenase (2-methylpropanoyl-transferring) activity	MF	2	1	0.00183545	1
GO:0003826	alpha-ketoacid dehydrogenase activity	MF	2	1	0.00183545	1
GO:0044329	canonical Wnt signaling pathway involved in positive regulation of cell-cell adhesion	BP	1	1	0.00191035	1
GO:0044328	canonical Wnt signaling pathway involved in positive regulation of endothelial cell migration	BP	1	1	0.00191035	1
GO:0044330	canonical Wnt signaling pathway involved in positive regulation of wound healing	BP	1	1	0.00191035	1
GO:0032470	positive regulation of endoplasmic reticulum calcium ion concentration	BP	2	1	0.00268311	1
GO:0071051	polyadenylation-dependent snoRNA 3'-end processing	BP	3	1	0.00297078	1

Magenta module – ARIES – KEGG

	Pathway	N	DE	P.DE	FDR
path:hsa00600	Sphingolipid metabolism	38	2	0.00093009	0.31250903
path:hsa04975	Fat digestion and absorption	31	1	0.03046451	1
path:hsa00640	Propanoate metabolism	29	1	0.03056133	1
path:hsa00280	Valine, leucine and isoleucine degradation	39	1	0.04089537	1
path:hsa00565	Ether lipid metabolism	42	1	0.04333755	1
path:hsa00561	Glycerolipid metabolism	51	1	0.0527499	1
path:hsa04923	Regulation of lipolysis in adipocytes	51	1	0.05781532	1
path:hsa03008	Ribosome biogenesis in eukaryotes	60	1	0.0608148	1
path:hsa03018	RNA degradation	66	1	0.07282576	1
path:hsa04260	Cardiac muscle contraction	73	1	0.0790747	1
path:hsa05212	Pancreatic cancer	69	1	0.0818489	1
path:hsa04520	Adherens junction	69	1	0.08516063	1
path:hsa03015	mRNA surveillance pathway	74	1	0.08631236	1
path:hsa04972	Pancreatic secretion	82	1	0.08778103	1
path:hsa00564	Glycerophospholipid metabolism	84	1	0.08990621	1
path:hsa05412	Arrhythmogenic right ventricular cardiomyopathy (ARVC)	71	1	0.0899189	1
path:hsa04666	Fc gamma R-mediated phagocytosis	82	1	0.09741603	1
path:hsa05410	Hypertrophic cardiomyopathy (HCM)	79	1	0.09758256	1
path:hsa05231	Choline metabolism in cancer	89	1	0.10318632	1
path:hsa04380	Osteoclast differentiation	105	1	0.10713878	1

Magenta module – BiB white British – gene ontology

GO:0043312	neutrophil degranulation	BP	406	30	2.32E-08	0.0001672
GO:0002283	neutrophil activation involved in immune response	BP	409	30	2.60E-08	0.0001672
GO:0042119	neutrophil activation	BP	417	30	3.74E-08	0.0001672
GO:0002446	neutrophil mediated immunity	BP	420	30	4.94E-08	0.0001672
GO:0036230	granulocyte activation	BP	423	30	5.01E-08	0.0001672
GO:0002443	leukocyte mediated immunity	BP	605	37	5.38E-08	0.0001672
GO:0043299	leukocyte degranulation	BP	447	31	5.59E-08	0.0001672
GO:0002263	cell activation involved in immune response	BP	566	36	6.15E-08	0.0001672
GO:0002275	myeloid cell activation involved in immune response	BP	454	31	8.05E-08	0.00019441
GO:0002444	myeloid leukocyte mediated immunity	BP	459	31	1.03E-07	0.00020971
GO:0006955	immune response	BP	1554	68	1.06E-07	0.00020971
GO:0002274	myeloid leukocyte activation	BP	528	34	1.29E-07	0.00023343
GO:0002366	leukocyte activation involved in immune response	BP	562	35	1.55E-07	0.00024797
GO:0045055	regulated exocytosis	BP	641	39	1.60E-07	0.00024797
GO:0045321	leukocyte activation	BP	961	48	2.06E-06	0.00298939
GO:0044433	cytoplasmic vesicle part	CC	1254	58	2.68E-06	0.00363452
GO:0030141	secretory granule	CC	709	36	8.41E-06	0.0107496
GO:0016192	vesicle-mediated transport	BP	1684	72	8.96E-06	0.01080509
GO:0001775	cell activation	BP	1094	51	9.52E-06	0.01080509
GO:0006887	exocytosis	BP	744	39	9.94E-06	0.01080509

Magenta module – BiB white British – KEGG

	Pathway	N	DE	P.DE	FDR
path:hsa05146	Amoebiasis	83	9	0.00044579	0.05804935
path:hsa04640	Hematopoietic cell lineage	66	7	0.00051549	0.05804935
path:hsa01100	Metabolic pathways	1249	49	0.0005183	0.05804935
path:hsa00620	Pyruvate metabolism	33	5	0.00078013	0.06553114
path:hsa04145	Phagosome	107	9	0.0012532	0.08421528
path:hsa03440	Homologous recombination	38	5	0.00316403	0.17718586
path:hsa04922	Glucagon signaling pathway	89	7	0.00932378	0.44754122
path:hsa00270	Cysteine and methionine metabolism	42	4	0.01190486	0.45455152
path:hsa00230	Purine metabolism	113	8	0.01238594	0.45455152
path:hsa00600	Sphingolipid metabolism	38	4	0.01537333	0.45455152
path:hsa05321	Inflammatory bowel disease (IBD)	43	4	0.01697218	0.45455152
path:hsa00020	Citrate cycle (TCA cycle)	25	3	0.02006222	0.45455152
path:hsa03410	Base excision repair	29	3	0.02090932	0.45455152
path:hsa01200	Carbon metabolism	94	6	0.02169328	0.45455152
path:hsa04931	Insulin resistance	97	7	0.02325236	0.45455152
path:hsa04720	Long-term potentiation	59	5	0.02380823	0.45455152
path:hsa00010	Glycolysis / Gluconeogenesis	55	4	0.0240921	0.45455152
path:hsa00240	Pyrimidine metabolism	48	4	0.02563157	0.45455152
path:hsa00350	Tyrosine metabolism	28	3	0.02734119	0.45455152
path:hsa04928	Parathyroid hormone synthesis, secretion and action	100	7	0.02821823	0.45455152

Magenta module – BiB Pakistani – gene ontology

	Term	Ont	N	DE	P.DE	FDR
GO:0098581	detection of external biotic stimulus	BP	17	4	0.00011187	0.79663602
GO:0045055	regulated exocytosis	BP	641	27	0.0001211	0.79663602
GO:0002263	cell activation involved in immune response	BP	566	24	0.00013018	0.79663602
GO:0009595	detection of biotic stimulus	BP	20	4	0.00016323	0.79663602
GO:0045321	leukocyte activation	BP	961	35	0.00021853	0.79663602
GO:0002443	leukocyte mediated immunity	BP	605	24	0.00022965	0.79663602
GO:0002366	leukocyte activation involved in immune response	BP	562	23	0.00029264	0.79663602
GO:0006955	immune response	BP	1554	47	0.00030728	0.79663602
GO:0002274	myeloid leukocyte activation	BP	528	22	0.00035555	0.79663602
GO:0043312	neutrophil degranulation	BP	406	18	0.0004418	0.79663602
GO:0002283	neutrophil activation involved in immune response	BP	409	18	0.00046703	0.79663602
GO:0043299	leukocyte degranulation	BP	447	19	0.00051742	0.79663602
GO:0042119	neutrophil activation	BP	417	18	0.00056225	0.79663602
GO:0002275	myeloid cell activation involved in immune response	BP	454	19	0.00062178	0.79663602
GO:0002446	neutrophil mediated immunity	BP	420	18	0.0006513	0.79663602
GO:0036230	granulocyte activation	BP	423	18	0.00065928	0.79663602
GO:0002444	myeloid leukocyte mediated immunity	BP	459	19	0.00070892	0.79663602
GO:0045085	negative regulation of interleukin-2 biosynthetic process	BP	3	2	0.00071483	0.79663602
GO:0001775	cell activation	BP	1094	37	0.00074076	0.79663602
GO:0070013	intracellular organelle lumen	CC	4418	116	0.00081379	0.79663602

Magenta module – BiB Pakistani – KEGG

	Pathway	N	DE	P.DE	FDR
path:hsa03440	Homologous recombination	38	4	0.00846502	0.99782561
path:hsa04145	Phagosome	107	6	0.01709728	0.99782561
path:hsa00620	Pyruvate metabolism	33	3	0.01977808	0.99782561
path:hsa04640	Hematopoietic cell lineage	67	4	0.02362374	0.99782561
path:hsa03320	PPAR signaling pathway	63	4	0.0267087	0.99782561
path:hsa04659	Th17 cell differentiation	80	5	0.02779886	0.99782561
path:hsa04261	Adrenergic signaling in cardiomyocytes	131	7	0.03101625	0.99782561
path:hsa05410	Hypertrophic cardiomyopathy (HCM)	79	5	0.03467631	0.99782561
path:hsa05146	Amoebiasis	83	5	0.03509611	0.99782561
path:hsa05144	Malaria	39	3	0.03857369	0.99782561
path:hsa00600	Sphingolipid metabolism	38	3	0.04070921	0.99782561
path:hsa05321	Inflammatory bowel disease (IBD)	43	3	0.04189662	0.99782561
path:hsa04720	Long-term potentiation	59	4	0.04196732	0.99782561
path:hsa04920	Adipocytokine signaling pathway	60	4	0.04941455	0.99782561
path:hsa04151	PI3K-Akt signaling pathway	308	12	0.05159628	0.99782561
path:hsa04260	Cardiac muscle contraction	73	4	0.05281619	0.99782561
path:hsa05414	Dilated cardiomyopathy (DCM)	87	5	0.05315035	0.99782561
path:hsa00240	Pyrimidine metabolism	48	3	0.05677867	0.99782561
path:hsa05010	Alzheimer disease	143	6	0.06216025	0.99782561
path:hsa05031	Amphetamine addiction	61	4	0.06332558	0.99782561

Pink module – ARIES – gene ontology

	Term	Ont	N	DE	P.DE	FDR
GO:0043231	intracellular membrane-bounded organelle	CC	9069	1028	1.44E-14	3.12E-10
GO:0044424	intracellular part	CC	11987	1285	9.64E-14	1.05E-09
GO:0043229	intracellular organelle	CC	10549	1154	1.56E-12	9.79E-09
GO:0005622	intracellular	CC	12200	1296	1.80E-12	9.79E-09
GO:0043227	membrane-bounded organelle	CC	10487	1139	9.02E-12	3.92E-08
GO:0044446	intracellular organelle part	CC	7678	880	3.23E-10	1.17E-06
GO:0043226	organelle	CC	11307	1203	8.71E-10	2.70E-06
GO:0044422	organelle part	CC	7875	895	1.28E-09	3.47E-06
GO:0070013	intracellular organelle lumen	CC	4418	544	2.32E-09	4.59E-06
GO:0031974	membrane-enclosed lumen	CC	4418	544	2.32E-09	4.59E-06
GO:0043233	organelle lumen	CC	4418	544	2.32E-09	4.59E-06
GO:0005488	binding	MF	12314	1291	3.57E-09	6.46E-06
GO:0005634	nucleus	CC	6047	702	4.44E-09	7.41E-06
GO:0034641	cellular nitrogen compound metabolic process	BP	5652	649	9.85E-09	1.53E-05
GO:0044428	nuclear part	CC	3767	477	1.13E-08	1.63E-05
GO:0003676	nucleic acid binding	MF	3455	423	1.72E-08	2.34E-05
GO:0044444	cytoplasmic part	CC	8175	900	4.99E-08	6.38E-05
GO:0031981	nuclear lumen	CC	3468	439	6.37E-08	7.70E-05
GO:0006139	nucleobase-containing compound metabolic process	BP	4977	581	8.44E-08	9.37E-05
GO:0044237	cellular metabolic process	BP	9068	973	8.63E-08	9.37E-05

Pink module – ARIES – KEGG

	Pathway	N	DE	P.DE	FDR
path:hsa04612	Antigen processing and presentation	31	10	0.00078447	0.18080132
path:hsa05203	Viral carcinogenesis	147	28	0.00121998	0.18080132
path:hsa05218	Melanoma	66	16	0.0018804	0.18080132
path:hsa04141	Protein processing in endoplasmic reticulum	135	24	0.0021524	0.18080132
path:hsa05220	Chronic myeloid leukemia	72	16	0.00453804	0.21979779
path:hsa03420	Nucleotide excision repair	35	8	0.00566319	0.21979779
path:hsa04115	p53 signaling pathway	68	15	0.00572003	0.21979779
path:hsa04390	Hippo signaling pathway	147	27	0.00626059	0.21979779
path:hsa05200	Pathways in cancer	468	68	0.0063541	0.21979779
path:hsa00620	Pyruvate metabolism	33	8	0.0065416	0.21979779
path:hsa03013	RNA transport	127	21	0.0103229	0.31531765
path:hsa05165	Human papillomavirus infection	288	44	0.01446858	0.40512023
path:hsa00380	Tryptophan metabolism	35	7	0.01639163	0.4236607
path:hsa04910	Insulin signaling pathway	120	20	0.02141892	0.49853257
path:hsa04922	Glucagon signaling pathway	89	16	0.02225592	0.49853257
path:hsa00240	Pyrimidine metabolism	48	9	0.02947646	0.57856229
path:hsa05217	Basal cell carcinoma	61	12	0.03103364	0.57856229
path:hsa05216	Thyroid cancer	34	8	0.0315249	0.57856229
path:hsa05214	Glioma	70	14	0.03337922	0.57856229
path:hsa00230	Purine metabolism	113	19	0.03443823	0.57856229

Pink module – BiB white British – gene ontology

	Term	Ont	N	DE	P.DE	FDR
GO:0043231	intracellular membrane-bounded organelle	CC	9069	1983	2.53E-25	5.49E-21
GO:0043229	intracellular organelle	CC	10549	2240	4.14E-23	4.50E-19
GO:0044424	intracellular part	CC	11987	2479	8.86E-22	6.42E-18
GO:0005634	nucleus	CC	6047	1390	6.20E-21	3.37E-17
GO:0043227	membrane-bounded organelle	CC	10487	2205	2.10E-20	9.11E-17
GO:0005622	intracellular	CC	12200	2505	3.81E-20	1.38E-16
GO:0005654	nucleoplasm	CC	2978	776	6.59E-19	1.80E-15
GO:0043226	organelle	CC	11307	2345	6.64E-19	1.80E-15
GO:0031981	nuclear lumen	CC	3468	876	8.19E-19	1.98E-15
GO:0044446	intracellular organelle part	CC	7678	1704	1.64E-18	3.25E-15
GO:0044428	nuclear part	CC	3767	938	1.64E-18	3.25E-15
GO:0044422	organelle part	CC	7875	1732	4.36E-17	7.89E-14
GO:0070013	intracellular organelle lumen	CC	4418	1050	8.29E-17	1.20E-13
GO:0031974	membrane-enclosed lumen	CC	4418	1050	8.29E-17	1.20E-13
GO:0043233	organelle lumen	CC	4418	1050	8.29E-17	1.20E-13
GO:0034641	cellular nitrogen compound metabolic process	BP	5652	1264	2.29E-16	3.11E-13
GO:0003676	nucleic acid binding	MF	3455	815	1.30E-14	1.66E-11
GO:0090304	nucleic acid metabolic process	BP	4452	1030	1.42E-14	1.72E-11
GO:0006139	nucleobase-containing compound metabolic process	BP	4977	1128	4.43E-14	5.07E-11
GO:0046483	heterocycle metabolic process	BP	5120	1151	1.28E-13	1.39E-10

Pink module – BiB white British – KEGG

	Pathway	N	DE	P.DE	FDR
path:hsa04110	Cell cycle	115	37	0.0010832	0.21132691
path:hsa04612	Antigen processing and presentation	31	14	0.0013662	0.21132691
path:hsa00130	Ubiquinone and other terpenoid-quinone biosynthesis	8	5	0.00188685	0.21132691
path:hsa04141	Protein processing in endoplasmic reticulum	135	38	0.00453841	0.35360068
path:hsa05203	Viral carcinogenesis	147	43	0.00526191	0.35360068
path:hsa04115	p53 signaling pathway	68	23	0.0092472	0.51784307
path:hsa03013	RNA transport	127	35	0.01088346	0.52240586
path:hsa05214	Glioma	70	24	0.01654781	0.69500822
path:hsa05216	Thyroid cancer	34	13	0.0196423	0.73331251
path:hsa00330	Arginine and proline metabolism	40	13	0.02480841	0.79668917
path:hsa04071	Sphingolipid signaling pathway	108	31	0.02608209	0.79668917
path:hsa05218	Melanoma	66	21	0.02960183	0.82456781
path:hsa00380	Tryptophan metabolism	35	10	0.03481821	0.82456781
path:hsa04614	Renin-angiotensin system	16	6	0.03812913	0.82456781
path:hsa04916	Melanogenesis	96	26	0.0447986	0.82456781
path:hsa04210	Apoptosis	118	30	0.04857024	0.82456781
path:hsa05200	Pathways in cancer	468	112	0.04870505	0.82456781
path:hsa05164	Influenza A	115	30	0.04924861	0.82456781
path:hsa03020	RNA polymerase	28	9	0.05260327	0.82456781
path:hsa03060	Protein export	21	7	0.05614032	0.82456781

Pink module – BiB Pakistani – gene ontology

	Term	Ont	N	DE	P.DE	FDR
GO:0043231	intracellular membrane-bounded organelle	CC	9069	1355	2.71E-18	5.88E-14
GO:0005634	nucleus	CC	6047	955	1.33E-15	1.44E-11
GO:0043229	intracellular organelle	CC	10549	1516	1.51E-14	8.88E-11
GO:0044428	nuclear part	CC	3767	652	1.63E-14	8.88E-11
GO:0044424	intracellular part	CC	11987	1679	6.99E-14	3.04E-10
GO:0031981	nuclear lumen	CC	3468	603	1.38E-13	4.98E-10
GO:0005654	nucleoplasm	CC	2978	534	1.91E-13	5.93E-10
GO:0005622	intracellular	CC	12200	1698	4.03E-13	1.02E-09
GO:0043227	membrane-bounded organelle	CC	10487	1493	4.23E-13	1.02E-09
GO:0070013	intracellular organelle lumen	CC	4418	722	1.86E-12	3.36E-09
GO:0031974	membrane-enclosed lumen	CC	4418	722	1.86E-12	3.36E-09
GO:0043233	organelle lumen	CC	4418	722	1.86E-12	3.36E-09
GO:0044446	intracellular organelle part	CC	7678	1155	5.58E-12	9.32E-09
GO:0043226	organelle	CC	11307	1585	1.23E-11	1.92E-08
GO:0044422	organelle part	CC	7875	1173	5.66E-11	8.20E-08
GO:0003676	nucleic acid binding	MF	3455	560	6.27E-11	8.52E-08
GO:0010467	gene expression	BP	4692	728	7.54E-10	9.64E-07
GO:0090304	nucleic acid metabolic process	BP	4452	697	1.49E-09	1.80E-06
GO:0044237	cellular metabolic process	BP	9068	1283	2.08E-09	2.38E-06
GO:0034641	cellular nitrogen compound metabolic process	BP	5652	844	2.35E-09	2.50E-06

Pink module – BiB Pakistani – KEGG

	Pathway	N	DE	P.DE	FDR
path:hsa04612	Antigen processing and presentation	31	11	0.00182257	0.27592302
path:hsa05203	Viral carcinogenesis	147	33	0.0028382	0.27592302
path:hsa05202	Transcriptional misregulation in cancer	148	36	0.00324926	0.27592302
path:hsa05200	Pathways in cancer	468	88	0.0032848	0.27592302
path:hsa04115	p53 signaling pathway	68	18	0.00654507	0.43982882
path:hsa05214	Glioma	70	19	0.00930841	0.48290948
path:hsa03013	RNA transport	127	26	0.01006061	0.48290948
path:hsa00970	Aminoacyl-tRNA biosynthesis	40	11	0.02214971	0.66300406
path:hsa05222	Small cell lung cancer	84	20	0.02268567	0.66300406
path:hsa00330	Arginine and proline metabolism	40	10	0.02353545	0.66300406
path:hsa05218	Melanoma	66	16	0.02589617	0.66300406
path:hsa04916	Melanogenesis	96	20	0.02750152	0.66300406
path:hsa00440	Phosphonate and phosphinate metabolism	5	3	0.02840088	0.66300406
path:hsa00100	Steroid biosynthesis	16	6	0.02919321	0.66300406
path:hsa00130	Ubiquinone and other terpenoid-quinone biosynthesis	8	3	0.02966446	0.66300406
path:hsa04141	Protein processing in endoplasmic reticulum	135	25	0.03157162	0.66300406
path:hsa04922	Glucagon signaling pathway	89	19	0.03515129	0.6711494
path:hsa04621	NOD-like receptor signaling pathway	135	22	0.03653007	0.6711494
path:hsa01230	Biosynthesis of amino acids	57	11	0.0420674	0.6711494
path:hsa04350	TGF-beta signaling pathway	90	19	0.04638538	0.6711494

Purple module – ARIES – gene ontology

	Term	Ont	N	DE	P.DE	FDR
GO:0045298	tubulin complex	CC	5	2	6.13E-05	1
GO:0003868	4-hydroxyphenylpyruvate dioxygenase activity	MF	2	1	0.00074285	1
GO:0033365	protein localization to organelle	BP	724	5	0.00159749	1
GO:0008281	sulfonylurea receptor activity	MF	2	1	0.00161217	1
GO:0072594	establishment of protein localization to organelle	BP	450	4	0.00179149	1
GO:0019948	SUMO activating enzyme activity	MF	2	1	0.00219964	1
GO:0031510	SUMO activating enzyme complex	CC	2	1	0.00219964	1
GO:0005220	inositol 1,4,5-trisphosphate-sensitive calcium-release channel activity	MF	2	1	0.00230483	1
GO:0008282	inward rectifying potassium channel	CC	4	1	0.00276561	1
GO:0007499	ectoderm and mesoderm interaction	BP	1	1	0.00279324	1
GO:0015016	[heparan sulfate]-glucosamine N-sulfotransferase activity	MF	4	1	0.00284046	1
GO:0039714	cytoplasmic viral factory	CC	1	1	0.00308282	1
GO:0072517	host cell viral assembly compartment	CC	1	1	0.00308282	1
GO:1900737	negative regulation of phospholipase C-activating G-protein coupled receptor signaling pathway	BP	1	1	0.00308282	1
GO:1900276	regulation of proteinase activated receptor activity	BP	1	1	0.00308282	1
GO:0039713	viral factory	CC	1	1	0.00308282	1
GO:1903779	regulation of cardiac conduction	BP	65	2	0.0033539	1
GO:0021691	cerebellar Purkinje cell layer maturation	BP	1	1	0.00368939	1
GO:0008267	poly-glutamine tract binding	MF	1	1	0.00368939	1
GO:1904674	positive regulation of somatic stem cell population maintenance	BP	2	1	0.00408207	1

Purple module – ARIES – KEGG

	Pathway	N	DE	P.DE	FDR
path:hsa00534	Glycosaminoglycan biosynthesis - heparan sulfate / heparin	21	1	0.03144455	1
path:hsa04710	Circadian rhythm	26	1	0.03546081	1
path:hsa02010	ABC transporters	41	1	0.04629223	1
path:hsa03460	Fanconi anemia pathway	47	1	0.05727497	1
path:hsa04924	Renin secretion	58	1	0.06968167	1
path:hsa04730	Long-term depression	55	1	0.07606422	1
path:hsa04929	GnRH secretion	57	1	0.07622766	1
path:hsa04918	Thyroid hormone synthesis	64	1	0.07664209	1
path:hsa04720	Long-term potentiation	59	1	0.07879663	1
path:hsa04927	Cortisol synthesis and secretion	53	1	0.08033948	1
path:hsa04970	Salivary secretion	75	1	0.09006723	1
path:hsa04971	Gastric acid secretion	70	1	0.09272717	1
path:hsa04972	Pancreatic secretion	82	1	0.09346871	1
path:hsa04540	Gap junction	79	1	0.09794268	1
path:hsa04625	C-type lectin receptor signaling pathway	90	1	0.09877105	1
path:hsa04912	GnRH signaling pathway	83	1	0.10431879	1
path:hsa04922	Glucagon signaling pathway	89	1	0.1085198	1
path:hsa04726	Serotonergic synapse	95	1	0.11499837	1
path:hsa04925	Aldosterone synthesis and secretion	84	1	0.11682656	1
path:hsa04070	Phosphatidylinositol signaling system	87	1	0.11806355	1

Purple module – BiB white British – gene ontology

	Term	Ont	N	DE	P.DE	FDR
GO:0007266	Rho protein signal transduction	BP	178	10	6.90E-06	0.08090549
GO:0007264	small GTPase mediated signal transduction	BP	508	17	7.45E-06	0.08090549
GO:0005829	cytosol	CC	4310	60	1.35E-05	0.0977215
GO:0045944	positive regulation of transcription by RNA polymerase II	BP	1011	24	2.27E-05	0.12352748
GO:0035767	endothelial cell chemotaxis	BP	25	4	8.61E-05	0.35708162
GO:0048522	positive regulation of cellular process	BP	4343	59	0.00010619	0.35708162
GO:0007265	Ras protein signal transduction	BP	404	13	0.00011501	0.35708162
GO:0035023	regulation of Rho protein signal transduction	BP	114	7	0.00013616	0.36989593
GO:0051056	regulation of small GTPase mediated signal transduction	BP	297	11	0.00020785	0.40090702
GO:0044444	cytoplasmic part	CC	8175	89	0.00021502	0.40090702
GO:0010628	positive regulation of gene expression	BP	1621	29	0.00024861	0.40090702
GO:0048260	positive regulation of receptor-mediated endocytosis	BP	42	4	0.00030409	0.40090702
GO:1903508	positive regulation of nucleic acid-templated transcription	BP	1289	25	0.00030882	0.40090702
GO:0045893	positive regulation of transcription, DNA-templated	BP	1289	25	0.00030882	0.40090702
GO:1902680	positive regulation of RNA biosynthetic process	BP	1290	25	0.00031036	0.40090702
GO:0048518	positive regulation of biological process	BP	4847	62	0.0003135	0.40090702
GO:0005856	cytoskeleton	CC	1861	31	0.0003136	0.40090702
GO:0031267	small GTPase binding	MF	499	14	0.00034342	0.41463636
GO:0030036	actin cytoskeleton organization	BP	572	15	0.00045647	0.5083966
GO:0051653	spindle localization	BP	38	4	0.00046786	0.5083966

Purple module – BiB white British – KEGG

	Pathway	N	DE	P.DE	FDR
path:hsa05135	Yersinia infection	106	5	0.00258765	0.52644662
path:hsa04144	Endocytosis	215	7	0.00422854	0.52644662
path:hsa05100	Bacterial invasion of epithelial cells	69	4	0.00470042	0.52644662
path:hsa04520	Adherens junction	69	3	0.03194946	1
path:hsa00471	D-Glutamine and D-glutamate metabolism	4	1	0.03243399	1
path:hsa05144	Malaria	38	2	0.03884151	1
path:hsa05321	Inflammatory bowel disease (IBD)	43	2	0.04456456	1
path:hsa04350	TGF-beta signaling pathway	90	3	0.05174795	1
path:hsa04714	Thermogenesis	190	4	0.07178089	1
path:hsa04722	Neurotrophin signaling pathway	109	3	0.07494778	1
path:hsa00730	Thiamine metabolism	12	1	0.07805796	1
path:hsa05132	Salmonella infection	67	2	0.1000704	1
path:hsa04720	Long-term potentiation	59	2	0.10330674	1
path:hsa05211	Renal cell carcinoma	62	2	0.1099062	1
path:hsa05205	Proteoglycans in cancer	182	4	0.10999469	1
path:hsa04010	MAPK signaling pathway	266	5	0.1122564	1
path:hsa00220	Arginine biosynthesis	17	1	0.11337038	1
path:hsa00120	Primary bile acid biosynthesis	14	1	0.11400239	1
path:hsa00340	Histidine metabolism	21	1	0.11801213	1
path:hsa00670	One carbon pool by folate	19	1	0.13286864	1

Purple module – BiB Pakistani – gene ontology

	Term	Ont	N	DE	P.DE	FDR
GO:0031267	small GTPase binding	MF	499	12	5.70E-06	0.0637
GO:0007264	small GTPase mediated signal transduction	BP	508	12	5.86E-06	0.0637
GO:0017016	Ras GTPase binding	MF	460	11	1.59E-05	0.10460468
GO:0007266	Rho protein signal transduction	BP	178	7	2.22E-05	0.10460468
GO:0051020	GTPase binding	MF	587	12	2.41E-05	0.10460468
GO:0005829	cytosol	CC	4310	35	2.92E-05	0.1056227
GO:0007265	Ras protein signal transduction	BP	404	9	0.00012979	0.40294634
GO:0005088	Ras guanyl-nucleotide exchange factor activity	MF	225	7	0.00018187	0.45180696
GO:0048522	positive regulation of cellular process	BP	4343	34	0.0001871	0.45180696
GO:0035767	endothelial cell chemotaxis	BP	25	3	0.00025952	0.52556975
GO:0035023	regulation of Rho protein signal transduction	BP	114	5	0.00026601	0.52556975
GO:0090087	regulation of peptide transport	BP	616	10	0.00037318	0.57871729
GO:0005737	cytoplasm	CC	9740	55	0.00038739	0.57871729
GO:0004711	ribosomal protein S6 kinase activity	MF	5	2	0.00039137	0.57871729
GO:0044444	cytoplasmic part	CC	8175	49	0.00039943	0.57871729
GO:0051879	Hsp90 protein binding	MF	29	3	0.00045569	0.61897469
GO:0030036	actin cytoskeleton organization	BP	572	10	0.0005004	0.63972106
GO:0045298	tubulin complex	CC	5	2	0.00057539	0.67622477
GO:0048260	positive regulation of receptor-mediated endocytosis	BP	42	3	0.0006327	0.67622477
GO:0045744	negative regulation of G-protein coupled receptor protein signaling pathway	BP	39	3	0.00067132	0.67622477

Purple module – BiB Pakistani – KEGG

	Pathway	N	DE	P.DE	FDR
path:hsa04010	MAPK signaling pathway	266	5	0.00954942	1
path:hsa04720	Long-term potentiation	59	2	0.03164166	1
path:hsa05218	Melanoma	66	2	0.04150696	1
path:hsa04540	Gap junction	79	2	0.04828311	1
path:hsa04914	Progesterone-mediated oocyte maturation	82	2	0.0583563	1
path:hsa01521	EGFR tyrosine kinase inhibitor resistance	74	2	0.05844513	1
path:hsa00340	Histidine metabolism	21	1	0.06117092	1
path:hsa04666	Fc gamma R-mediated phagocytosis	82	2	0.06351886	1
path:hsa04015	Rap1 signaling pathway	194	3	0.07228609	1
path:hsa03430	Mismatch repair	22	1	0.0782615	1
path:hsa04114	Oocyte meiosis	104	2	0.08035205	1
path:hsa04144	Endocytosis	215	3	0.08321644	1
path:hsa05135	Yersinia infection	106	2	0.08406855	1
path:hsa04931	Insulin resistance	97	2	0.0858822	1
path:hsa00515	Mannose type O-glycan biosynthesis	20	1	0.08654931	1
path:hsa04722	Neurotrophin signaling pathway	109	2	0.08985129	1
path:hsa04928	Parathyroid hormone synthesis, secretion and action	100	2	0.09172363	1
path:hsa03410	Base excision repair	29	1	0.09404488	1
path:hsa00532	Glycosaminoglycan biosynthesis - chondroitin sulfate / dermatan sulfate	18	1	0.09558562	1
path:hsa04080	Neuroactive ligand-receptor interaction	277	3	0.09604847	1

Red module – ARIES – gene ontology

	Term	Ont	N	DE	P.DE	FDR
GO:0048704	embryonic skeletal system morphogenesis	BP	92	5	3.88E-05	0.45207335
GO:0000981	RNA polymerase II transcription factor activity, sequence-specific DNA binding	MF	971	13	5.69E-05	0.45207335
GO:0043565	sequence-specific DNA binding	MF	994	13	6.24E-05	0.45207335
GO:0048562	embryonic organ morphogenesis	BP	273	7	0.00011071	0.4587149
GO:0003700	DNA binding transcription factor activity	MF	1323	14	0.00012461	0.4587149
GO:0007389	pattern specification process	BP	389	8	0.00012664	0.4587149
GO:0048706	embryonic skeletal system development	BP	120	5	0.00015573	0.48351072
GO:0005249	voltage-gated potassium channel activity	MF	85	4	0.00019784	0.537463
GO:0003002	regionalization	BP	308	7	0.0002281	0.55080278
GO:0048598	embryonic morphogenesis	BP	545	9	0.00030355	0.65969542
GO:0022832	voltage-gated channel activity	MF	169	5	0.00043567	0.78903081
GO:0005244	voltage-gated ion channel activity	MF	169	5	0.00043567	0.78903081
GO:0005267	potassium channel activity	MF	119	4	0.00076142	1
GO:0009952	anterior/posterior pattern specification	BP	186	5	0.00088095	1
GO:0043009	chordate embryonic development	BP	520	8	0.00091267	1
GO:0048568	embryonic organ development	BP	396	7	0.00102514	1
GO:0009792	embryo development ending in birth or egg hatching	BP	534	8	0.00108176	1
GO:0140110	transcription regulator activity	MF	1635	14	0.00111615	1
GO:0048705	skeletal system morphogenesis	BP	202	5	0.00125843	1
GO:0021700	developmental maturation	BP	233	5	0.00137671	1

Red module – ARIES – KEGG

	Pathway	N	DE	P.DE	FDR
path:hsa00120	Primary bile acid biosynthesis	14	1	0.04819698	1
path:hsa04744	Phototransduction	25	1	0.06841158	1
path:hsa04910	Insulin signaling pathway	120	2	0.06954224	1
path:hsa00515	Mannose type O-glycan biosynthesis	20	1	0.0703499	1
path:hsa04136	Autophagy - other	28	1	0.11025244	1
path:hsa02010	ABC transporters	41	1	0.12007347	1
path:hsa04979	Cholesterol metabolism	44	1	0.13379359	1
path:hsa00510	N-Glycan biosynthesis	46	1	0.14077449	1
path:hsa03320	PPAR signaling pathway	63	1	0.16355494	1
path:hsa04330	Notch signaling pathway	50	1	0.17930725	1
path:hsa04740	Olfactory transduction	204	1	0.18165312	1
path:hsa04730	Long-term depression	55	1	0.18658921	1
path:hsa04213	Longevity regulating pathway - multiple species	54	1	0.20506929	1
path:hsa05217	Basal cell carcinoma	61	1	0.21502456	1
path:hsa00562	Inositol phosphate metabolism	66	1	0.22262715	1
path:hsa05200	Pathways in cancer	468	3	0.2612683	1
path:hsa04211	Longevity regulating pathway	84	1	0.28138617	1
path:hsa04933	AGE-RAGE signaling pathway in diabetic complications	83	1	0.28374259	1
path:hsa04010	MAPK signaling pathway	266	2	0.28376667	1
path:hsa04350	TGF-beta signaling pathway	90	1	0.28604075	1

Red module – BiB white British – gene ontology

	Term	Ont	N	DE	P.DE	FDR
GO:0000981	RNA polymerase II transcription factor activity, sequence-specific DNA binding	MF	971	67	3.91E-15	8.49E-11
GO:0003700	DNA binding transcription factor activity	MF	1323	71	1.89E-12	2.06E-08
GO:0140110	transcription regulator activity	MF	1635	75	6.52E-10	4.72E-06
GO:0043565	sequence-specific DNA binding	MF	994	56	1.27E-09	6.92E-06
GO:0032501	multicellular organismal process	BP	6078	181	6.09E-09	2.65E-05
GO:0048856	anatomical structure development	BP	4953	161	7.99E-09	2.77E-05
GO:0001077	transcriptional activator activity, RNA polymerase II proximal promoter sequence-specific DNA binding	MF	255	25	9.19E-09	2.77E-05
GO:0001228	transcriptional activator activity, RNA polymerase II transcription regulatory region sequence-specific DNA binding	MF	381	31	1.02E-08	2.77E-05
GO:0032502	developmental process	BP	5296	167	1.66E-08	4.01E-05
GO:0007275	multicellular organism development	BP	4542	149	4.56E-08	9.91E-05
GO:0000982	transcription factor activity, RNA polymerase II proximal promoter sequence-specific DNA binding	MF	382	30	5.70E-08	0.0001127
GO:0006357	regulation of transcription by RNA polymerase II	BP	1920	80	7.02E-08	0.00012716
GO:0048731	system development	BP	4031	136	1.24E-07	0.00020292
GO:0045944	positive regulation of transcription by RNA polymerase II	BP	1011	52	1.31E-07	0.00020292
GO:0003008	system process	BP	1621	63	3.84E-07	0.00052174
GO:0003002	regionalization	BP	308	25	3.84E-07	0.00052174
GO:0007423	sensory organ development	BP	481	32	4.11E-07	0.00052546
GO:0007389	pattern specification process	BP	389	28	5.04E-07	0.00060817
GO:0009887	animal organ morphogenesis	BP	898	47	6.42E-07	0.00068034

Red module – BiB white British – KEGG

	Pathway	N	DE	P.DE	FDR
path:hsa04910	Insulin signaling pathway	120	8	0.00696465	1
path:hsa04022	cGMP-PKG signaling pathway	151	9	0.00808147	1
path:hsa05206	MicroRNAs in cancer	220	11	0.01307755	1
path:hsa04722	Neurotrophin signaling pathway	109	6	0.03968034	1
path:hsa04020	Calcium signaling pathway	171	8	0.05525996	1
path:hsa05224	Breast cancer	139	7	0.05555144	1
path:hsa04978	Mineral absorption	45	3	0.06023145	1
path:hsa04925	Aldosterone synthesis and secretion	84	5	0.06298033	1
path:hsa04350	TGF-beta signaling pathway	90	5	0.06539941	1
path:hsa05414	Dilated cardiomyopathy (DCM)	87	5	0.07110684	1
path:hsa04390	Hippo signaling pathway	147	7	0.09368466	1
path:hsa04550	Signaling pathways regulating pluripotency of stem cells	130	6	0.10386701	1
path:hsa04923	Regulation of lipolysis in adipocytes	51	3	0.1054492	1
path:hsa04514	Cell adhesion molecules (CAMs)	110	5	0.10725845	1
path:hsa04261	Adrenergic signaling in cardiomyocytes	131	6	0.1084279	1
path:hsa05412	Arrhythmogenic right ventricular cardiomyopathy (ARVC)	71	4	0.11240201	1
path:hsa05226	Gastric cancer	142	6	0.1403487	1
path:hsa00512	Mucin type O-glycan biosynthesis	28	2	0.14568636	1
path:hsa04080	Neuroactive ligand-receptor interaction	277	8	0.15333297	1
path:hsa04211	Longevity regulating pathway	84	4	0.15764826	1

Red module – BiB Pakistani – gene ontology

	Term	Ont	N	DE	P.DE	FDR
GO:0000981	RNA polymerase II transcription factor activity, sequence-specific DNA binding	MF	971	50	3.20E-10	6.95E-06
GO:0007275	multicellular organism development	BP	4542	127	1.55E-08	9.55E-05
GO:0032501	multicellular organismal process	BP	6078	150	1.55E-08	9.55E-05
GO:0048856	anatomical structure development	BP	4953	134	2.06E-08	9.55E-05
GO:0003700	DNA binding transcription factor activity	MF	1323	53	2.20E-08	9.55E-05
GO:0032502	developmental process	BP	5296	139	3.49E-08	0.00012645
GO:0043565	sequence-specific DNA binding	MF	994	44	2.75E-07	0.00085375
GO:0048731	system development	BP	4031	113	3.69E-07	0.00100329
GO:0030154	cell differentiation	BP	3499	99	6.57E-07	0.00158682
GO:0140110	transcription regulator activity	MF	1635	57	8.83E-07	0.00191945
GO:0001077	transcriptional activator activity, RNA polymerase II proximal promoter sequence-specific DNA binding	MF	255	19	1.81E-06	0.00356835
GO:0048869	cellular developmental process	BP	3671	101	2.34E-06	0.00394211
GO:0048513	animal organ development	BP	2920	86	2.36E-06	0.00394211
GO:0006357	regulation of transcription by RNA polymerase II	BP	1920	64	3.84E-06	0.00596516
GO:0009887	animal organ morphogenesis	BP	898	39	4.44E-06	0.00628637
GO:0001228	transcriptional activator activity, RNA polymerase II transcription regulatory region sequence-specific DNA binding	MF	381	23	4.70E-06	0.00628637
GO:0007389	pattern specification process	BP	389	23	4.92E-06	0.00628637
GO:0003002	regionalization	BP	308	20	8.17E-06	0.00946359
GO:0007399	nervous system development	BP	2029	70	8.27E-06	0.00946359

Red module – BiB Pakistani – KEGG

	Pathway	N	DE	P.DE	FDR
path:hsa04910	Insulin signaling pathway	120	6	0.02663726	1
path:hsa05414	Dilated cardiomyopathy (DCM)	87	5	0.03495451	1
path:hsa04722	Neurotrophin signaling pathway	109	5	0.05346646	1
path:hsa04514	Cell adhesion molecules (CAMs)	110	5	0.05413978	1
path:hsa05206	MicroRNAs in cancer	220	8	0.05560998	1
path:hsa05224	Breast cancer	139	6	0.06109757	1
path:hsa05412	Arrhythmogenic right ventricular cardiomyopathy (ARVC)	71	4	0.06348758	1
path:hsa04923	Regulation of lipolysis in adipocytes	51	3	0.06565579	1
path:hsa00920	Sulfur metabolism	10	1	0.08343293	1
path:hsa00430	Taurine and hypotaurine metabolism	9	1	0.11350962	1
path:hsa04213	Longevity regulating pathway - multiple species	54	3	0.11546142	1
path:hsa04020	Calcium signaling pathway	171	6	0.12060717	1
path:hsa05032	Morphine addiction	84	4	0.12127381	1
path:hsa04550	Signaling pathways regulating pluripotency of stem cells	130	5	0.12349781	1
path:hsa04261	Adrenergic signaling in cardiomyocytes	131	5	0.12821363	1
path:hsa00565	Ether lipid metabolism	42	2	0.13537184	1
path:hsa00562	Inositol phosphate metabolism	66	3	0.14664058	1
path:hsa04022	cGMP-PKG signaling pathway	151	5	0.15204585	1
path:hsa05218	Melanoma	66	3	0.15212681	1
path:hsa05216	Thyroid cancer	34	2	0.15751422	1

Tan module – ARIES – gene ontology

	Term	Ont	N	DE	P.DE	FDR
GO:0044446	intracellular organelle part	CC	7678	411	2.85E-05	0.42444514
GO:0044422	organelle part	CC	7875	419	3.91E-05	0.42444514
GO:0044428	nuclear part	CC	3767	226	7.94E-05	0.46252933
GO:0070013	intracellular organelle lumen	CC	4418	253	0.00012971	0.46252933
GO:0031974	membrane-enclosed lumen	CC	4418	253	0.00012971	0.46252933
GO:0043233	organelle lumen	CC	4418	253	0.00012971	0.46252933
GO:0043229	intracellular organelle	CC	10549	526	0.00014898	0.46252933
GO:0005634	nucleus	CC	6047	325	0.0002485	0.61419304
GO:0031981	nuclear lumen	CC	3468	207	0.00027204	0.61419304
GO:0044424	intracellular part	CC	11987	583	0.00029258	0.61419304
GO:0005829	cytosol	CC	4310	244	0.00031087	0.61419304
GO:0043231	intracellular membrane-bounded organelle	CC	9069	458	0.00039186	0.7096972
GO:0005737	cytoplasm	CC	9740	487	0.00059384	0.99276589
GO:1902235	regulation of endoplasmic reticulum stress-induced intrinsic apoptotic signaling pathway	BP	25	6	0.00086577	1
GO:0005739	mitochondrion	CC	1483	90	0.00094189	1
GO:0006399	tRNA metabolic process	BP	163	17	0.00123773	1
GO:0004470	malic enzyme activity	MF	4	3	0.00137712	1
GO:0044444	cytoplasmic part	CC	8175	414	0.0014367	1
GO:0043227	membrane-bounded organelle	CC	10487	513	0.00147059	1
GO:0043226	organelle	CC	11307	549	0.00155759	1

Tan module – ARIES – KEGG

	Pathway	N	DE	P.DE	FDR
path:hsa00620	Pyruvate metabolism	33	7	0.00028853	0.09694516
path:hsa04910	Insulin signaling pathway	120	13	0.00537481	0.83056244
path:hsa03013	RNA transport	127	12	0.01202772	0.83056244
path:hsa04922	Glucagon signaling pathway	89	10	0.01250881	0.83056244
path:hsa03050	Proteasome	34	5	0.01322655	0.83056244
path:hsa04612	Antigen processing and presentation	31	5	0.01629694	0.83056244
path:hsa05031	Amphetamine addiction	60	8	0.01730338	0.83056244
path:hsa05134	Legionellosis	47	6	0.02465	1
path:hsa00380	Tryptophan metabolism	35	4	0.03254628	1
path:hsa03420	Nucleotide excision repair	35	4	0.03681546	1
path:hsa05217	Basal cell carcinoma	61	7	0.03771972	1
path:hsa03022	Basal transcription factors	31	4	0.05113822	1
path:hsa05010	Alzheimer disease	143	11	0.05660057	1
path:hsa04970	Salivary secretion	75	7	0.06079176	1
path:hsa04931	Insulin resistance	97	9	0.07586284	1
path:hsa00630	Glyoxylate and dicarboxylate metabolism	26	3	0.0771631	1
path:hsa04925	Aldosterone synthesis and secretion	84	8	0.0794865	1
path:hsa03015	mRNA surveillance pathway	74	7	0.08231267	1
path:hsa04728	Dopaminergic synapse	115	10	0.08328651	1
path:hsa04261	Adrenergic signaling in cardiomyocytes	131	11	0.08382949	1

Tan module – BiB white British – gene ontology

	Term	Ont	N	DE	P.DE	FDR
GO:0043231	intracellular membrane-bounded organelle	CC	9069	169	1.90E-06	0.04131115
GO:0043227	membrane-bounded organelle	CC	10487	186	4.36E-06	0.04733641
GO:0044267	cellular protein metabolic process	BP	4452	98	1.20E-05	0.08589618
GO:0044248	cellular catabolic process	BP	1907	51	1.58E-05	0.08589618
GO:0009056	catabolic process	BP	2158	55	2.87E-05	0.12491887
GO:1901575	organic substance catabolic process	BP	1757	47	3.66E-05	0.13270163
GO:1901564	organonitrogen compound metabolic process	BP	5782	115	4.33E-05	0.13429702
GO:0044237	cellular metabolic process	BP	9068	163	5.67E-05	0.14025674
GO:0044446	intracellular organelle part	CC	7678	147	5.81E-05	0.14025674
GO:0044444	cytoplasmic part	CC	8175	151	0.00011461	0.22969234
GO:0019538	protein metabolic process	BP	4912	101	0.00011626	0.22969234
GO:0044424	intracellular part	CC	11987	201	0.00016069	0.27024747
GO:0044422	organelle part	CC	7875	148	0.00016188	0.27024747
GO:0010523	negative regulation of calcium ion transport into cytosol	BP	15	4	0.00017409	0.27024747
GO:0071704	organic substance metabolic process	BP	9213	162	0.00020635	0.28548225
GO:0043226	organelle	CC	11307	192	0.00021017	0.28548225
GO:0031975	envelope	CC	995	29	0.00024826	0.29974239
GO:0031967	organelle envelope	CC	995	29	0.00024826	0.29974239
GO:0035751	regulation of lysosomal lumen pH	BP	12	3	0.00026849	0.30207129
GO:0005739	mitochondrion	CC	1483	38	0.00030445	0.30207129

Tan module – BiB white British – KEGG

	Pathway	N	DE	P.DE	FDR
path:hsa03050	Proteasome	34	3	0.01098939	1
path:hsa05321	Inflammatory bowel disease (IBD)	43	3	0.01673542	1
path:hsa01100	Metabolic pathways	1249	27	0.01868831	1
path:hsa04218	Cellular senescence	135	6	0.02596908	1
path:hsa04977	Vitamin digestion and absorption	21	2	0.02722674	1
path:hsa04979	Cholesterol metabolism	44	3	0.02747038	1
path:hsa00531	Glycosaminoglycan degradation	16	2	0.0289514	1
path:hsa04115	p53 signaling pathway	68	4	0.02977008	1
path:hsa00790	Folate biosynthesis	23	2	0.03399544	1
path:hsa04659	Th17 cell differentiation	80	4	0.03537493	1
path:hsa03015	mRNA surveillance pathway	74	4	0.0358845	1
path:hsa03013	RNA transport	127	5	0.038965	1
path:hsa04744	Phototransduction	25	2	0.04120024	1
path:hsa00062	Fatty acid elongation	24	2	0.05105373	1
path:hsa00071	Fatty acid degradation	34	2	0.05735926	1
path:hsa00480	Glutathione metabolism	38	2	0.05948215	1
path:hsa04380	Osteoclast differentiation	105	4	0.06188123	1
path:hsa00310	Lysine degradation	53	3	0.06629678	1
path:hsa00410	beta-Alanine metabolism	30	2	0.07204196	1
path:hsa04114	Oocyte meiosis	104	4	0.08440322	1

Tan module – BiB Pakistani – gene ontology

	Term	Ont	N	DE	P.DE	FDR
GO:0008565	protein transporter activity	MF	77	3	0.00051785	1
GO:0034250	positive regulation of cellular amide metabolic process	BP	123	3	0.00209142	1
GO:0031523	Myb complex	CC	1	1	0.00228192	1
GO:0000412	histone peptidyl-prolyl isomerization	BP	1	1	0.00234058	1
GO:1901675	negative regulation of histone H3-K27 acetylation	BP	1	1	0.0028141	1
GO:0045819	positive regulation of glycogen catabolic process	BP	1	1	0.00284047	1
GO:0071344	diphosphate metabolic process	BP	1	1	0.00292717	1
GO:1903259	exon-exon junction complex disassembly	BP	1	1	0.00348055	1
GO:0050104	L-gulonate 3-dehydrogenase activity	MF	1	1	0.00353167	1
GO:0061084	negative regulation of protein refolding	BP	4	1	0.00361606	1
GO:0061083	regulation of protein refolding	BP	4	1	0.00361606	1
GO:0006399	tRNA metabolic process	BP	163	3	0.00424687	1
GO:0006064	glucuronate catabolic process	BP	4	1	0.00510701	1
GO:0019640	glucuronate catabolic process to xylulose 5-phosphate	BP	4	1	0.00510701	1
GO:1901159	xylulose 5-phosphate biosynthetic process	BP	4	1	0.00510701	1
GO:0051167	xylulose 5-phosphate metabolic process	BP	4	1	0.00510701	1
GO:0032558	adenyl deoxyribonucleotide binding	MF	2	1	0.00536648	1
GO:0032564	dATP binding	MF	2	1	0.00536648	1
GO:0046778	modification by virus of host mRNA processing	BP	2	1	0.00576885	1
GO:0030943	mitochondrion targeting sequence binding	MF	3	1	0.00593736	1

Tan module – BiB Pakistani – KEGG

	Pathway	N	DE	P.DE	FDR
path:hsa03015	mRNA surveillance pathway	74	2	0.01180015	1
path:hsa04922	Glucagon signaling pathway	89	2	0.01540806	1
path:hsa00591	Linoleic acid metabolism	20	1	0.03092045	1
path:hsa00592	alpha-Linolenic acid metabolism	20	1	0.03134392	1
path:hsa05202	Transcriptional misregulation in cancer	148	2	0.04923209	1
path:hsa00040	Pentose and glucuronate interconversions	28	1	0.05365353	1
path:hsa00590	Arachidonic acid metabolism	44	1	0.06290053	1
path:hsa03050	Proteasome	34	1	0.06315405	1
path:hsa00565	Ether lipid metabolism	42	1	0.07120127	1
path:hsa05131	Shigellosis	58	1	0.11674279	1
path:hsa05132	Salmonella infection	67	1	0.11867476	1
path:hsa04115	p53 signaling pathway	68	1	0.13973865	1
path:hsa00564	Glycerophospholipid metabolism	84	1	0.14904701	1
path:hsa00190	Oxidative phosphorylation	104	1	0.1577515	1
path:hsa04666	Fc gamma R-mediated phagocytosis	82	1	0.1701589	1
path:hsa04750	Inflammatory mediator regulation of TRP channels	90	1	0.17844505	1
path:hsa04217	Necroptosis	111	1	0.18120197	1
path:hsa04928	Parathyroid hormone synthesis, secretion and action	100	1	0.20817768	1
path:hsa05164	Influenza A	115	1	0.2083555	1
path:hsa05215	Prostate cancer	91	1	0.20968457	1

Turquoise module – ARIES – gene ontology

	Term	Ont	N	DE	P.DE	FDR
GO:0043231	intracellular membrane-bounded organelle	CC	9069	2847	7.90E-36	1.72E-31
GO:0044424	intracellular part	CC	11987	3589	6.49E-35	7.05E-31
GO:0043229	intracellular organelle	CC	10549	3230	1.76E-34	1.27E-30
GO:0005622	intracellular	CC	12200	3617	3.68E-30	2.00E-26
GO:0043227	membrane-bounded organelle	CC	10487	3177	1.38E-29	5.99E-26
GO:0043226	organelle	CC	11307	3377	8.27E-27	3.00E-23
GO:0044446	intracellular organelle part	CC	7678	2438	3.15E-25	9.77E-22
GO:0005634	nucleus	CC	6047	1958	2.24E-24	6.08E-21
GO:0044422	organelle part	CC	7875	2480	2.60E-23	6.27E-20
GO:0044237	cellular metabolic process	BP	9068	2724	1.72E-19	3.73E-16
GO:0005654	nucleoplasm	CC	2978	1061	2.74E-19	5.41E-16
GO:0044428	nuclear part	CC	3767	1291	1.08E-18	1.96E-15
GO:0006139	nucleobase-containing compound metabolic process	BP	4977	1607	2.14E-18	3.58E-15
GO:0031981	nuclear lumen	CC	3468	1198	2.65E-18	4.11E-15
GO:0005737	cytoplasm	CC	9740	2921	3.61E-18	5.23E-15
GO:0008152	metabolic process	BP	9634	2856	3.97E-18	5.39E-15
GO:0034641	cellular nitrogen compound metabolic process	BP	5652	1779	4.54E-18	5.80E-15
GO:0044238	primary metabolic process	BP	8929	2665	1.56E-17	1.89E-14
GO:0046483	heterocycle metabolic process	BP	5120	1638	2.13E-17	2.44E-14
GO:1901360	organic cyclic compound metabolic process	BP	5332	1694	3.15E-17	3.27E-14

Turquoise module – ARIES – KEGG

	Pathway	N	DE	P.DE	FDR
path:hsa05203	Viral carcinogenesis	147	62	0.00035253	0.11845084
path:hsa04070	Phosphatidylinositol signaling system	87	41	0.00070888	0.11909151
path:hsa05210	Colorectal cancer	82	37	0.00231705	0.25950949
path:hsa05169	Epstein-Barr virus infection	146	58	0.00440388	0.28687883
path:hsa04110	Cell cycle	115	46	0.00446822	0.28687883
path:hsa05130	Pathogenic Escherichia coli infection	177	64	0.00529436	0.28687883
path:hsa00970	Aminoacyl-tRNA biosynthesis	40	20	0.0067595	0.28687883
path:hsa04071	Sphingolipid signaling pathway	108	44	0.00713292	0.28687883
path:hsa03040	Spliceosome	108	43	0.00908945	0.28687883
path:hsa05010	Alzheimer disease	143	52	0.00936262	0.28687883
path:hsa05166	Human T-cell leukemia virus 1 infection	182	68	0.01026369	0.28687883
path:hsa04932	Non-alcoholic fatty liver disease (NAFLD)	128	46	0.01053957	0.28687883
path:hsa04141	Protein processing in endoplasmic reticulum	135	49	0.01109948	0.28687883
path:hsa05012	Parkinson disease	110	38	0.01328564	0.31885536
path:hsa05161	Hepatitis B	135	50	0.01447977	0.32434694
path:hsa04115	p53 signaling pathway	68	29	0.016057	0.33719705
path:hsa03013	RNA transport	127	46	0.01814347	0.34540622
path:hsa05034	Alcoholism	108	41	0.01953068	0.34540622
path:hsa04928	Parathyroid hormone synthesis, secretion and action	100	41	0.01963193	0.34540622
path:hsa01524	Platinum drug resistance	62	26	0.02127465	0.34540622

Turquoise module – BiB white British – gene ontology

	Term	Ont	N	DE	P.DE	FDR
GO:0044424	intracellular part	CC	11987	5390	2.51E-59	5.45E-55
GO:0043229	intracellular organelle	CC	10549	4840	1.07E-57	1.17E-53
GO:0043231	intracellular membrane-bounded organelle	CC	9069	4248	4.02E-57	2.91E-53
GO:0005622	intracellular	CC	12200	5447	1.78E-54	9.65E-51
GO:0043226	organelle	CC	11307	5071	1.29E-45	5.61E-42
GO:0043227	membrane-bounded organelle	CC	10487	4742	1.05E-44	3.79E-41
GO:0005634	nucleus	CC	6047	2932	1.33E-42	4.13E-39
GO:0044446	intracellular organelle part	CC	7678	3636	2.38E-41	6.46E-38
GO:0044422	organelle part	CC	7875	3700	4.58E-38	1.11E-34
GO:0044237	cellular metabolic process	BP	9068	4087	2.18E-32	4.75E-29
GO:0044428	nuclear part	CC	3767	1920	5.40E-32	1.07E-28
GO:0006139	nucleobase-containing compound metabolic process	BP	4977	2407	8.91E-32	1.61E-28
GO:0031981	nuclear lumen	CC	3468	1780	4.00E-31	6.70E-28
GO:1901360	organic cyclic compound metabolic process	BP	5332	2548	4.51E-31	7.00E-28
GO:0005737	cytoplasm	CC	9740	4383	7.80E-31	1.13E-27
GO:0034641	cellular nitrogen compound metabolic process	BP	5652	2669	8.60E-31	1.17E-27
GO:0046483	heterocycle metabolic process	BP	5120	2459	1.05E-30	1.34E-27
GO:0070013	intracellular organelle lumen	CC	4418	2182	1.51E-30	1.64E-27
GO:0031974	membrane-enclosed lumen	CC	4418	2182	1.51E-30	1.64E-27
GO:0043233	organelle lumen	CC	4418	2182	1.51E-30	1.64E-27

Turquoise module – BiB white British – KEGG

	Pathway	N	DE	P.DE	FDR
path:hsa03040	Spliceosome	108	66	0.00012324	0.02264686
path:hsa03018	RNA degradation	66	43	0.0001348	0.02264686
path:hsa05203	Viral carcinogenesis	147	83	0.00054627	0.04154183
path:hsa03013	RNA transport	127	71	0.00058677	0.04154183
path:hsa04110	Cell cycle	115	66	0.00064568	0.04154183
path:hsa04210	Apoptosis	118	66	0.00074182	0.04154183
path:hsa03015	mRNA surveillance pathway	74	46	0.00134304	0.06446609
path:hsa05166	Human T-cell leukemia virus 1 infection	182	98	0.00227235	0.09543887
path:hsa03008	Ribosome biogenesis in eukaryotes	60	34	0.00410052	0.15308625
path:hsa05161	Hepatitis B	135	72	0.0049239	0.16544297
path:hsa04668	TNF signaling pathway	95	51	0.00610652	0.18652629
path:hsa04137	Mitophagy - animal	60	37	0.00796339	0.20212521
path:hsa00970	Aminoacyl-tRNA biosynthesis	40	26	0.00824924	0.20212521
path:hsa05168	Herpes simplex virus 1 infection	372	151	0.0090435	0.20212521
path:hsa05169	Epstein-Barr virus infection	146	78	0.00977301	0.20212521
path:hsa04070	Phosphatidylinositol signaling system	87	50	0.0110183	0.20212521
path:hsa03460	Fanconi anemia pathway	47	29	0.01166643	0.20212521
path:hsa05167	Kaposi sarcoma-associated herpesvirus infection	148	75	0.01184385	0.20212521
path:hsa05010	Alzheimer disease	143	72	0.01231375	0.20212521
path:hsa04217	Necroptosis	111	56	0.01251679	0.20212521

Turquoise module – BiB Pakistani – gene ontology

	Term	Ont	N	DE	P.DE	FDR
GO:0043231	intracellular membrane-bounded organelle	CC	9069	4217	7.62E-63	1.66E-58
GO:0044424	intracellular part	CC	11987	5328	1.54E-62	1.68E-58
GO:0043229	intracellular organelle	CC	10549	4789	4.25E-61	3.08E-57
GO:0005622	intracellular	CC	12200	5384	1.23E-57	6.66E-54
GO:0043226	organelle	CC	11307	5012	6.16E-48	2.33E-44
GO:0043227	membrane-bounded organelle	CC	10487	4693	6.44E-48	2.33E-44
GO:0044446	intracellular organelle part	CC	7678	3618	4.33E-47	1.35E-43
GO:0005634	nucleus	CC	6047	2901	6.91E-44	1.88E-40
GO:0044422	organelle part	CC	7875	3680	2.06E-43	4.98E-40
GO:0044428	nuclear part	CC	3767	1923	2.67E-37	5.81E-34
GO:0031981	nuclear lumen	CC	3468	1784	2.47E-36	4.88E-33
GO:0070013	intracellular organelle lumen	CC	4418	2181	2.01E-35	3.12E-32
GO:0031974	membrane-enclosed lumen	CC	4418	2181	2.01E-35	3.12E-32
GO:0043233	organelle lumen	CC	4418	2181	2.01E-35	3.12E-32
GO:0005654	nucleoplasm	CC	2978	1560	2.60E-34	3.77E-31
GO:0044237	cellular metabolic process	BP	9068	4039	1.22E-33	1.66E-30
GO:0005737	cytoplasm	CC	9740	4339	1.60E-33	2.04E-30
GO:0008152	metabolic process	BP	9634	4241	9.05E-32	1.06E-28
GO:0006139	nucleobase-containing compound metabolic process	BP	4977	2376	9.31E-32	1.06E-28
GO:0034641	cellular nitrogen compound metabolic process	BP	5652	2640	9.78E-32	1.06E-28

Turquoise module – BiB Pakistani – KEGG

	Pathway	N	DE	P.DE	FDR
path:hsa03013	RNA transport	127	72	0.00018861	0.02667855
path:hsa04210	Apoptosis	118	67	0.00023768	0.02667855
path:hsa03018	RNA degradation	66	42	0.0002382	0.02667855
path:hsa04110	Cell cycle	115	66	0.0003931	0.03302059
path:hsa05012	Parkinson disease	110	58	0.00062844	0.04223113
path:hsa05203	Viral carcinogenesis	147	81	0.00097468	0.05458212
path:hsa05010	Alzheimer disease	143	75	0.0018775	0.09011993
path:hsa03040	Spliceosome	108	61	0.00238603	0.10021319
path:hsa03460	Fanconi anemia pathway	47	30	0.00403469	0.13753145
path:hsa03015	mRNA surveillance pathway	74	44	0.0040932	0.13753145
path:hsa05166	Human T-cell leukemia virus 1 infection	182	95	0.00509819	0.14933477
path:hsa04932	Non-alcoholic fatty liver disease (NAFLD)	128	65	0.00542451	0.14933477
path:hsa03410	Base excision repair	29	18	0.0065276	0.14933477
path:hsa00970	Aminoacyl-tRNA biosynthesis	40	26	0.00666545	0.14933477
path:hsa03008	Ribosome biogenesis in eukaryotes	60	33	0.00666673	0.14933477
path:hsa04070	Phosphatidylinositol signaling system	87	50	0.00784985	0.15531532
path:hsa04668	TNF signaling pathway	95	50	0.00785822	0.15531532
path:hsa05161	Hepatitis B	135	70	0.00841628	0.15710387
path:hsa04217	Necroptosis	111	56	0.00912327	0.16133778
path:hsa04141	Protein processing in endoplasmic reticulum	135	68	0.00975599	0.1639006

Yellow module – ARIES – gene ontology

	Term	Ont	N	DE	P.DE	FDR
GO:0043231	intracellular membrane-bounded organelle	CC	9069	400	4.12E-06	0.05471652
GO:0044424	intracellular part	CC	11987	500	7.13E-06	0.05471652
GO:0061695	transferase complex, transferring phosphorus-containing groups	CC	227	24	7.55E-06	0.05471652
GO:0043229	intracellular organelle	CC	10549	451	1.42E-05	0.07737092
GO:1902494	catalytic complex	CC	1195	76	1.80E-05	0.07750816
GO:1990234	transferase complex	CC	684	50	2.14E-05	0.07750816
GO:0044446	intracellular organelle part	CC	7678	349	3.30E-05	0.10233347
GO:0044422	organelle part	CC	7875	356	4.00E-05	0.10864634
GO:0055029	nuclear DNA-directed RNA polymerase complex	CC	115	14	8.41E-05	0.1919946
GO:0000428	DNA-directed RNA polymerase complex	CC	116	14	8.83E-05	0.1919946
GO:0030880	RNA polymerase complex	CC	119	14	0.00010994	0.21721425
GO:0005622	intracellular	CC	12200	501	0.00013948	0.2526181
GO:0016591	DNA-directed RNA polymerase II, holoenzyme	CC	93	12	0.00019699	0.32932604
GO:0007156	homophilic cell adhesion via plasma membrane adhesion molecules	BP	135	16	0.00035328	0.54841732
GO:0099518	vesicle cytoskeletal trafficking	BP	48	9	0.00039717	0.57544306
GO:0047496	vesicle transport along microtubule	BP	40	8	0.00049982	0.62246164
GO:0043226	organelle	CC	11307	467	0.00053173	0.62246164
GO:0038093	Fc receptor signaling pathway	BP	159	16	0.00056007	0.62246164
GO:0031082	BLOC complex	CC	20	5	0.00056774	0.62246164
GO:0031981	nuclear lumen	CC	3468	175	0.00057283	0.62246164

Yellow module – ARIES – KEGG

	Pathway	N	DE	P.DE	FDR
path:hsa03022	Basal transcription factors	31	6	0.00101153	0.33987305
path:hsa00480	Glutathione metabolism	38	5	0.00372914	0.62649576
path:hsa03020	RNA polymerase	28	4	0.01639936	1
path:hsa04142	Lysosome	108	9	0.02548917	1
path:hsa03050	Proteasome	34	4	0.03127469	1
path:hsa04152	AMPK signaling pathway	111	10	0.03849086	1
path:hsa03460	Fanconi anemia pathway	47	5	0.0484491	1
path:hsa01212	Fatty acid metabolism	48	5	0.05035365	1
path:hsa05416	Viral myocarditis	34	4	0.05611459	1
path:hsa04218	Cellular senescence	135	10	0.06138864	1
path:hsa00062	Fatty acid elongation	24	3	0.06493587	1
path:hsa01040	Biosynthesis of unsaturated fatty acids	23	3	0.07506239	1
path:hsa00983	Drug metabolism - other enzymes	61	5	0.08566777	1
path:hsa03420	Nucleotide excision repair	35	3	0.08948776	1
path:hsa04659	Th17 cell differentiation	80	6	0.09662179	1
path:hsa05418	Fluid shear stress and atherosclerosis	115	8	0.10162364	1
path:hsa01230	Biosynthesis of amino acids	57	4	0.10432521	1
path:hsa05110	Vibrio cholerae infection	47	4	0.12100191	1
path:hsa05135	Yersinia infection	106	7	0.12811749	1
path:hsa04940	Type I diabetes mellitus	18	2	0.12887925	1

Yellow module – BiB white British – gene ontology

	Term	Ont	N	DE	P.DE	FDR
GO:0044237	cellular metabolic process	BP	9068	330	7.60E-08	0.00165106
GO:0008152	metabolic process	BP	9634	343	1.83E-07	0.00198532
GO:0044248	cellular catabolic process	BP	1907	94	5.19E-07	0.00376148
GO:0071704	organic substance metabolic process	BP	9213	327	1.11E-06	0.00603208
GO:0044446	intracellular organelle part	CC	7678	291	1.72E-06	0.00746034
GO:0009056	catabolic process	BP	2158	102	2.29E-06	0.00830181
GO:0044238	primary metabolic process	BP	8929	317	6.61E-06	0.02051513
GO:0044424	intracellular part	CC	11987	405	1.07E-05	0.02902221
GO:0044422	organelle part	CC	7875	293	1.47E-05	0.0321391
GO:1901575	organic substance catabolic process	BP	1757	84	1.50E-05	0.0321391
GO:0006807	nitrogen compound metabolic process	BP	8506	303	1.63E-05	0.0321391
GO:0005829	cytosol	CC	4310	178	2.46E-05	0.04455492
GO:0044444	cytoplasmic part	CC	8175	296	3.32E-05	0.04559542
GO:0043603	cellular amide metabolic process	BP	975	50	3.76E-05	0.04559542
GO:0018211	peptidyl-tryptophan modification	BP	4	3	3.78E-05	0.04559542
GO:0018103	protein C-linked glycosylation	BP	4	3	3.78E-05	0.04559542
GO:0018406	protein C-linked glycosylation via 2'-alpha-mannosyl-L-tryptophan	BP	4	3	3.78E-05	0.04559542
GO:0018317	protein C-linked glycosylation via tryptophan	BP	4	3	3.78E-05	0.04559542
GO:0043231	intracellular membrane-bounded organelle	CC	9069	321	4.47E-05	0.05107458
GO:0005759	mitochondrial matrix	CC	394	26	6.59E-05	0.07165793

Yellow module – BiB white British – KEGG

	Pathway	N	DE	P.DE	FDR
path:hsa03050	Proteasome	34	5	0.00227141	0.76319237
path:hsa04130	SNARE interactions in vesicular transport	32	4	0.01719189	1
path:hsa00330	Arginine and proline metabolism	40	4	0.03028512	1
path:hsa03013	RNA transport	127	8	0.03794292	1
path:hsa00640	Propanoate metabolism	29	3	0.04464289	1
path:hsa04979	Cholesterol metabolism	44	4	0.04493208	1
path:hsa00910	Nitrogen metabolism	15	2	0.04560955	1
path:hsa04142	Lysosome	108	7	0.0521624	1
path:hsa00220	Arginine biosynthesis	17	2	0.06094264	1
path:hsa01100	Metabolic pathways	1249	46	0.06113493	1
path:hsa00051	Fructose and mannose metabolism	30	3	0.06501029	1
path:hsa03010	Ribosome	104	6	0.06557564	1
path:hsa04136	Autophagy - other	28	3	0.08018951	1
path:hsa00120	Primary bile acid biosynthesis	14	2	0.08173865	1
path:hsa04340	Hedgehog signaling pathway	47	4	0.09100253	1
path:hsa00531	Glycosaminoglycan degradation	16	2	0.10022247	1
path:hsa04977	Vitamin digestion and absorption	21	2	0.10107586	1
path:hsa00280	Valine, leucine and isoleucine degradation	39	3	0.10479699	1
path:hsa03015	mRNA surveillance pathway	74	5	0.1122989	1
path:hsa00900	Terpenoid backbone biosynthesis	20	2	0.11632236	1

Yellow module – BiB Pakistani – gene ontology

	Term	Ont	N	DE	P.DE	FDR
GO:0005829	cytosol	CC	4310	61	3.07E-05	0.56251164
GO:0044424	intracellular part	CC	11987	121	9.78E-05	0.56251164
GO:0044444	cytoplasmic part	CC	8175	93	0.0001003	0.56251164
GO:0009056	catabolic process	BP	2158	35	0.00011695	0.56251164
GO:0005737	cytoplasm	CC	9740	105	0.00013785	0.56251164
GO:1901575	organic substance catabolic process	BP	1757	30	0.00015936	0.56251164
GO:0003873	6-phosphofructo-2-kinase activity	MF	2	2	0.00018118	0.56251164
GO:0044248	cellular catabolic process	BP	1907	31	0.0002949	0.66201668
GO:0004331	fructose-2,6-bisphosphate 2-phosphatase activity	MF	3	2	0.00031153	0.66201668
GO:0006003	fructose 2,6-bisphosphate metabolic process	BP	3	2	0.00031153	0.66201668
GO:0005622	intracellular	CC	12200	121	0.00033507	0.66201668
GO:0035173	histone kinase activity	MF	17	3	0.00046709	0.84593687
GO:0019203	carbohydrate phosphatase activity	MF	7	2	0.0006363	0.98776704
GO:0050308	sugar-phosphatase activity	MF	7	2	0.0006363	0.98776704
GO:0032984	protein-containing complex disassembly	BP	279	9	0.00077264	1
GO:0043402	glucocorticoid mediated signaling pathway	BP	3	2	0.00077801	1
GO:0035184	histone threonine kinase activity	MF	7	2	0.00098203	1
GO:0035405	histone-threonine phosphorylation	BP	7	2	0.00098203	1
GO:0009057	macromolecule catabolic process	BP	1121	21	0.00129398	1
GO:0042989	sequestering of actin monomers	BP	6	2	0.00130854	1

Yellow module – BiB Pakistani – KEGG

	Pathway	N	DE	P.DE	FDR
path:hsa00531	Glycosaminoglycan degradation	16	2	0.01022558	1
path:hsa00230	Purine metabolism	113	4	0.02450958	1
path:hsa00760	Nicotinate and nicotinamide metabolism	35	2	0.0280122	1
path:hsa00051	Fructose and mannose metabolism	30	2	0.02803418	1
path:hsa03015	mRNA surveillance pathway	74	3	0.03595483	1
path:hsa00565	Ether lipid metabolism	42	2	0.04133455	1
path:hsa01100	Metabolic pathways	1249	16	0.05088448	1
path:hsa05130	Pathogenic Escherichia coli infection	177	4	0.06373636	1
path:hsa04310	Wnt signaling pathway	146	4	0.06453036	1
path:hsa05110	Vibrio cholerae infection	47	2	0.06953707	1
path:hsa05135	Yersinia infection	106	3	0.07169305	1
path:hsa04340	Hedgehog signaling pathway	47	2	0.08428938	1
path:hsa03013	RNA transport	127	3	0.09342284	1
path:hsa05120	Epithelial cell signaling in Helicobacter pylori infection	62	2	0.09637355	1
path:hsa05132	Salmonella infection	67	2	0.10030911	1
path:hsa04918	Thyroid hormone synthesis	64	2	0.10449929	1
path:hsa00340	Histidine metabolism	21	1	0.11699807	1
path:hsa00591	Linoleic acid metabolism	20	1	0.11922641	1
path:hsa00592	alpha-Linolenic acid metabolism	20	1	0.12005617	1
path:hsa00120	Primary bile acid biosynthesis	14	1	0.1283331	1