Author:
Berryman, Katie

Title:
An Epigenome-Wide Association Study of Eczema

An Epigenome-Wide Association Study of Eczema

Katie May Berryman

BA (Hons), Cambridge University

MA (Hons), Cambridge University

A dissertation submitted to the University of Bristol in accordance with the requirements for award of the degree of Masters in Research in the Faculty of Population Health Sciences

July 2019

Word count: 27,616

ABSTRACT

BACKGROUND

Eczema is an allergic disease which affects the skin of thousands of children worldwide. EWAS analyses on related atopic factors such as asthma and Immunoglobulin E have already been carried out, as has a GWAS on eczema.

METHODS

Epigenetic analyses were conducted to look at the relationship between DNA methylation and eczema. Avon Longitudinal Study of Parents and Children (ALSPAC) was used, which consisted of questionnaires about eczema status and methylation data at time points birth, age seven and age 15/17. In order to see if there was a relationship between methylation and whether a child had ever had eczema, I conducted a longitudinal study, and a cross-sectional study looking at the relationship between DNA methylation and whether a child had had eczema in the last 12 months. I adjusted for sex, surrogate variables, two socioeconomic status variables, maternal history and cell counts. I looked at environmental and genetic risk factors, including smoking, animal exposure and breastfeeding and mQTL's to investigate DNA methylation as a potential mediator.

RESULTS

25 CpG sites were found to be suggestively associated at $P<0.05$ with the presence of eczema. There were five CpG sites which showed continued association between the different timepoints, strengthening the findings. There were 24 CpG sites which had a similarly small p-value when looking at the association between risk factors and eczema, and eczema and risk factors. These methylation sites were identified in smoking, but none in animal exposure and breastfeeding.

CONCLUSION

Overall, I have found evidence for 25 weak associations when looking at DNA methylation and eczema, and 24 CpG sites which could potentially show a link between risk factors and eczema. Replication and/or validation would strengthen results, as would a meta-analysis. Future work may potentially help design a treatment.

ACKNOWLEDGEMENTS

Thanks to…

My supervisors Dr Lavinia Paternoster, Dr Hannah Elliott and Professor Caroline Relton for their dedication in helping me achieve my degree. They have continually pushed to get the most out of me and encouraged me when I felt stuck. Thank you for access to your paper, 'Multi-ancestry genome-wide association study of 21,000 cases and 95,000 controls identifies new risk loci for atopic dermatitis', your comments on my draft chapters and supporting me in what I am going to do next in my career

Gemma Sharp, for producing the skeleton EWAS script which I then edited and adapted to carry out my study

Raquel Granell, for providing a list of 15 confounders in her paper 'Associations of different phenotypes of wheezing illness in early childhood with environmental variables implicated in the aetiology of asthma' which I then used as a starting point for the confounders addressed in my study

Authors of the journal papers I read and cited throughout this study, including epigenetics and eczema

Epigenetic Epidemiology group and Epigenetics Journal club (weekly) and Atopy group (monthly) which have allowed me to push the boundaries of what I understand and am researching about atopic illnesses including eczema, through listening and expanding what I know, to presenting myself

Fellow students in the office, from all different topics and backgrounds, who have helped me with EWAS's, statistics and decoding difficult papers

Barbara, Lynn, Simon and Rob, for all their laughter and early morning chatter

Friends, who have provided a welcome distraction and space to debrief and chat. They have provided moral support and companionship throughout my degree

My family, my mum Sally, dad Martyn and brother Tom, for being a sounding board for all my research, including presentations, and their continued support. Thank you for guiding me and helping me complete the degree with strength and resilience

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

CHAPTER 1. INTRODUCTION

1.1 ATOPIC DISEASES

1.1.1 Atopic Dermatitis/Eczema

Disease description

Eczema, also known as atopic dermatitis (AD), is a chronic skin condition that affects many aspects of a person's health and wellbeing (1). It is an allergic condition, like asthma and hay fever, meaning the body is showing an exaggerated immune response to an environmental exposure (2). When the skin meets an allergen, a reaction occurs, and a rash can develop. Creams and ointments can be applied to the skin to sooth the rash, tablets can be taken, and dietary supplements can sometimes be used with effect (3). Eczema is more common in children, with the disease fading as a person approaches adulthood (4). However, sometimes it develops during adulthood or persists throughout life (4).

Terminology

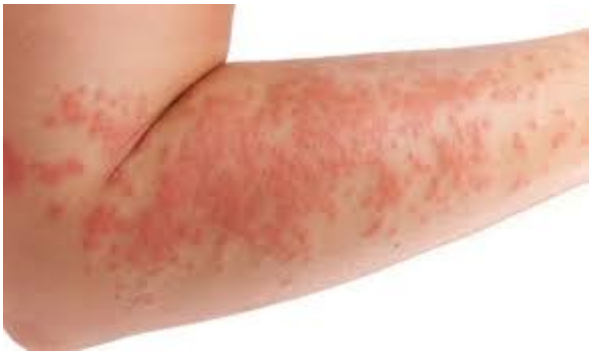
The terms 'Atopic dermatitis' and 'Eczema' are often used interchangeably. However, the World Allergy Organisation has described dermatitis as an umbrella term which includes eczema (5). This definition of 'eczema' covers both 'atopic' and 'non-atopic' illness. Other types of dermatitis include 'Irritant contact dermatitis' and 'Seborrheic dermatitis'. I shall refer to the illness as eczema, rather than atopic dermatitis, unless I am specifically referring not to eczema but to another form of dermatitis.

Biology of Eczema

Eczema is an allergic disease. The skin can become red, itchy and raw and flare up when exposed to particular allergens (1). These environmental allergens cause the immune system to react and produce many white blood cells, which include lymphocytes, monocytes, neutrophils, basophils and,

lastly, eosinophils (6). These white blood cells, which unlike red blood cells, contain a nucleus, travel to the site of invasion and produce antibodies to fight the infection. Eosinophils produce an antibody known as Immunoglobulin E (7) or the allergen antibody. This can affect symptoms in the skin, eyes, nose and throat. In eczema, symptoms almost entirely focus on the skin, as can be seen in figure 1.1 (2). It is a complex disease with both genetic and environmental risk factors.

Figure 1.1: A photograph of eczematous skin



Diagnosis

Diagnosis of eczema itself takes place based on an examination of the skin that is causing the distress and questions about the rash and a person's lifestyle. The doctor may investigate allergy testing, a skin biopsy or levels of Immunoglobulin E, antibodies produced by the immune system which occur with an allergic disease. Infections of Impetigo (8) or Furuncles (9) coinciding with the atopic disease may indicate the presence of eczema and positive or negative misdiagnosis may take place. It may be a very severe unidentified rash that is misdiagnosed as eczema, or a relatively mild limited rash wrongly diagnosed as something other than eczema. Questions asked by the doctor are often qualitative, such as "how bad is the itch?" or "whereabouts does it hurt?". Therefore, the bias could be under or over estimating the presence or severity of eczema. There are however two questions which were selected from Avon Longitudinal Study of Parents and Children (ALSPAC) which require information about whether a doctor has made a diagnosis of eczema. There is no laboratory test that a doctor can do to diagnose eczema, although patch tests can be carried out to diagnose other illnesses which cooccur with the disease such as skin infections. Examples of these are *Staphylococcus aureus* infection (10), which is the most likely infection to be present with eczema. A quarter of people have this bacteria residing in their nose, but this is found to be considerably higher in people with eczema. Examples of common *S. aureus* infections are Impetigo

(8) and Furuncles (9).

Related conditions and comorbidities

Trials and studies have been carried out on other atopic diseases, such as asthma (11) and hay fever (12). Asthma is a long-term condition affecting the airways of the lungs; allergens can trigger an inflammatory response and cause problems breathing. Hay fever is a reaction to pollen or other allergens in the air (13). It most commonly occurs during the summer months and can be triggered by grass, flower or plant pollen. Allergies such as eczema, asthma and hay fever belong to a group of illnesses known as atopic diseases (14). If you suffer from one of these illnesses, it is more likely that you will suffer from another in the group (15). The term 'Atopic March' (16) can be used to describe the way one allergic disease leads on to another. It has been shown that levels of a substance called Immunoglobulin E (IgE) are raised in those undergoing an allergic reaction (17), be this eczema, asthma or hay fever. Immunoglobulin E is an antibody; these antibodies are released in response to an allergen, causing an allergic reaction.

The effects of the disease itself include itching, which can have negative psychological consequences, as can the effect on people's appearance and lack of sleep due to the persistence of the condition (18). There have been several studies showing that eczema is related to other later health outcomes. For example, it can cause or exacerbate later mental health problems such as depression, anxiety and obsessive compulsive disorder (OCD) due to excessive washing (19), (20). There is also some evidence that eczema, asthma and hay fever are associated with Attention Deficit Hyperactivity Disorder (ADHD) (21). Eczema can affect a person's ability to succeed academically at school and in their chosen job later in life (22) due to days of absence due to illness, lack of sleep due to itching, or psychological/social factors such as depression or embarrassment due to the appearance of the rash.

Treatments

Eczema causes inflammation of the skin, leaving it red, dry and cracked, and treatment centres upon soothing the rash and other symptoms, preventing the itch (15) and healing the skin. The areas most commonly affected are the hands and face, and the backs of knees and elbows (23). Current

treatments target the skin. Emollients (moisturisers) and topical corticosteroids can be used during flare ups (23). Also doctors can recommend that sufferers apply bandages to help heal the skin if it cracks and breaks during what is known as the itch scratch cycle (24). This is where a person begins to itch, so they scratch, which makes the itch worse. This can lead to bleeding and infection, and possibly scarring. As a result, a doctor may recommend taking antihistamines if the itching is particularly bad (23). Self-care can involve trying to resist the urge to scratch or wearing light clothes if this is difficult.

Financial impact

The impact of eczema on the NHS in the UK is profound. Costs of treating the disease, including clinic/hospital appointments, medication and psychological therapy can impact greatly on a country's finances. For example, the annual cost of eczema in the UK is approximately £465 million and in the United States $5.297 billion (in 2015) (25). This can include things like medication costs, special clothing, cosmetics or certain foods that a person can purchase to reduce the effects of the eczema rash.

1.1.2 Epidemiology

It is important to look at what affects the development of eczema and why, including environmental factors. Allergens such as pollen, animal hair, foods, fabrics and certain soaps and cosmetics can exacerbate the condition and cause an immune response in the body (26). Hormonal changes and stress can also have an influence (27). This is particularly the case with women. Women going through their menstrual period or the menopause usually find that the surge or drop in hormones in the body can have an effect on eczema (28). In addition, women who are pregnant can find that their eczema 'plays up' during pregnancy (29). Usually an increase in hormones leads to worse eczema (30). However, the menopause can also cause flare ups. Stress can also be a trigger because more stress can lead to eczema worsening. Not only can the hormones in the body make eczema worse, factors related to stress like depression, boredom, irritation, and obsessive and compulsive behaviour, can all exacerbate the disease (19) (20). The 'Hygiene Hypothesis' is one explanation for the cause of eczema (31). If an environment is too clean, the body will not be exposed to infectious agents or parasites. The body then needlessly reacts to allergens as if it is under attack. Eczema can also be hereditary (32), indicating a genetic component. It is known that eczema in children is more

common when the parent has eczema, with the heritability estimated at 70-80 % (33). This could also indicate an epigenetic component. Perhaps there is an element of interaction between the environment and genetics which surfaces as epigenetics. See section 1.1.2 for an explanation of epigenetics.

Incidence

Incidence refers to the number of new cases of eczema which emerge in a particular time period. The disease has been found to be increased by 2- to 3-fold during the past decades in industrialized countries (34). The International Study of Asthma and Allergies in Childhood (ISAAC) (35) has some of the best reported data on eczema. It is widely referred to as it is the largest study that has been carried out, including 2 million children in 100 countries (35). The latest available data (Phase Three of the ISAAC study) showed that while eczema incidence has reached a plateau in the UK and New Zealand, it continues to increase in countries where young children are combined with a low income.

Prevalence

Eczema is more common amongst children than adults, with 25% of school aged children and up to 10% of adults suffering from it (1). It is thought that 45% of eczema cases occur in the child's first six months, and 85% of cases before the age of five (36). Eczema is usually worse in hotter countries, largely because sweating, the body's natural way of keeping cool, can exacerbate eczema and its symptoms (37). Whilst 20% of children are affected (1) on average, this varies greatly in prevalence throughout the world. Beginning at age six/seven, the figures start at a presence of eczema of 0.9% in India increasing to 22.5% in Ecuador. For the age group 13-14 this went from 0.2% in China to 24.6% in Columbia (1).

Pattern over life course

Eczema is an illness which affects children, and usually by mid-childhood it has attenuated or disappeared. The prognosis depends to some extent on genetics and the heritability of the disease, as well as exposure to recognised allergens, such as pet hair, cosmetics and pollen etc. Paternoster

et al. (4) looked at the pattern of the progress of eczema throughout life, whether it strikes young or older, and whether it attenuates or progressively gets worse. Comparing two longitudinal birth cohort studies, ALSPAC (n = 9894) and Prevention and Incidence of Asthma and Mite Allergy (PIAMA) (n = 3652), from the Netherlands, they looked to identify different phenotypes using longitudinal latent class analysis (4). They found six overall classes which could be used to describe the longitudinal phenotype of eczema presence. The most common class was described as 'early-onset-early-resolving'. The two which were next most prevalent were 'early-onset-persistent' and 'early-onset-late-resolving' and had most in common with parental diagnosis of eczema. See figure 1.2:

Figure 1.2: Longitudinal classes identified using LLCA in 2 independent birth cohorts: A, ALSPAC (n = 9894) and B, PIAMA (n = 3652) (4)

* Reproduced with permission (4)

Sex differences in eczema

Mollerup *et al.* (38) found that hand eczema was more common in women than men. In fact, it is twice as frequent. Mollerup went on to explain that occupation may play a large role. Another study by Gerada *et al.* (39), The International Study of Asthma and Allergies in Childhood (ISAAC), the

largest worldwide epidemiological research ever undertaken on childhood allergies, found that young boys with allergies, including eczema, were clinically worse off than girls, with more symptoms and more problems encountered in daily life. However, in adolescence this changed, with allergic disease being more prevalent in girls. A reason for this may be due to different hormones in males and females, or different exposures to allergens depending on their gender.

Environmental/non-genetic risk factors

Eczema can be caused or exacerbated by allergens known colloquially as 'triggers'. Things which could trigger a reaction include soaps, certain fabrics, pollen, animal hair and even stress (40). Dietary changes can have an influence as an allergy towards certain foods could be causing or exacerbating their eczema (41). There are various ways which a person could reduce their risk of developing eczema. For example, cutting out cow's milk and eggs and perhaps other dairy products could possibly make a difference to the presentation of the disease (41). Children could also potentially be seen by a specialist who could suggest further treatments, like light therapy or counselling.

Current new treatments are being investigated, in terms of a preventative medication or a cure. People increasing their understanding of inflammation in atopic dermatitis is likely to have a positive impact (42), as is procedures like health care professionals repeatedly washing their hands in hospitals and on the wards. Other drugs are also in the pipeline.

Smoking risk factor

Smoking is thought to be a very influential variable in many association studies of eczema. When looking at smoking during pregnancy and growing up in a smoky environment, evidence has been found by looking at Kantor *et al.* (43), Singh *et al.* (44), Tanaka *et al.* (45) and Shrinde *et al.* (46). A systematic review and meta-analysis by Kantor *et al.* (43) looked at atopic dermatitis and exposure to tobacco smoke. Out of 5817 original manuscripts, 86 studies were included, 23 for smoking during pregnancy. No link was found between smoking during pregnancy and eczema (odds ratio 1.06, 95% confidence intervals 0.80, 1.40). Another study by Singh *et al.* (44) showed no evidence of an association between smoking during pregnancy and eczema. However another paper by Tanaka *et*

al. (45) looking at pre- and postnatal smoking exposure and atopy in Japanese children, found that pre-natal exposure (in the womb) was positively associated with eczema. The effect of a smoky environment on eczema was addressed by Shrinde *et al.* (46), who found that the chances of developing eczema were increased when a child grew up in a smoky home. However this paper did not look at DNA methylation.

Although Kantor *et al.* found no strong evidence of association between maternal smoking during pregnancy and offspring eczema, it was found that higher active (OR 1.87, 95% CI 1.32, 2.63) and passive (OR 1.18, 95% CI 1.01, 1.38) smoking were associated with increased eczema (43). In agreement with the systematic review, another study by Singh *et al.* (44) found that exposure to environmental tobacco smoke was associated with childhood eczema. Importantly the two studies did sub-analyses on slightly different smoking groups (smoking in pregnancy and smoking environment). In addition to these studies, the study by Tanaka *et al.* (45) showed that there was no relationship between a smoky environment and eczema. Overall there is potentially a link between a smoky environment and eczema because the two large systematic reviews by Kantor and Singh claimed as much. Table 1.1 provides a summary of the main literature on the risk factor smoking.

Table 1.1: Broad summary of literature on smoking

	Kantor (43)	Singh (44)	Tanaka (45)	Shrinde (46)	Overall
Smoking during pregnancy	NULL	NULL	POSITIVE	NULL	NULL
Smoky environment	POSITIVE	POSITIVE	NULL	POSITIVE	POSITIVE

Animal exposure risk factor

Pet ownership, particularly the presence of cats and dogs has been linked to eczema. There are two large systematic reviews on the topic of cat and dog exposure, by Fretzayas *et al.* (47) and Langan *et al.* (48), and three other main studies, by Epstein *et al.* (49), Wegienka *et al.* (50) and Pohlabeln *et al.* (51). Evidence showed that exposure to cats (pooled odds ratio, 0.76; 95% confidence intervals, 0.62-0.92) and dogs (pooled OR, 0.68; 95% CI, 0.53-0.87) is linked to a lower likelihood of eczema developing. Fretzayas *et al.* (47) and Langan *et al.* (48) found contradictory results when looking at cats, and a small negative association when looking at dog exposure and eczema. Epstein's study

(49) showed a positive association with cats, as did the study by Wegienka *et al.* (50). Epstein's study (49) showed a negative relationship with dogs. Overall it could be stated that dog exposure reduced the likelihood of developing eczema.

Fretzayas, a systematic review (47), found the results of exposure to cats contradictory. Several studies (52), (53), (54), (55), (56) were found to have no effect of cat exposure on eczema. However, some other studies (57), (58) found that cat exposure was associated with eczema, with teens who kept a cat as a pet in infancy showing lower cases of eczema. Fretzayas *et al.* (47) found that most studies concluded no relationship between dog ownership and eczema. Langan's systematic review (48), however, showed that exposure to dogs lowered the risk of developing eczema (pooled odds ratio, 0.68; 95% confidence intervals, 0.53, 0.87). Similarly, the study by Epstein *et al.* (49) found that dog ownership reduced the risk of developing eczema.

Overall studies by Epstein *et al.* (49) and Wegienka *et al.* (50) found that cat exposure increased the risk of developing eczema. Langan's systematic review (48), a study by Epstein *et al.* (49) and one by Pohlabeln *et al.* (51) found that dog ownership reduced the risk of developing eczema. The evidence is therefore conflicting. Table 1.2 provides a summary of the main literature on the risk factor animal exposure.

Table 1.2: Broad summary of literature on animal exposure

	Fretzayas (47)	Langan (48)	Epstein (59)	Wengenka (50)	Pohlabeln (51)	Overall
Cat exposure	CONTRADICTORY	CONTRADICTORY	POSITIVE	POSITIVE		POSITIVE
Dog exposure	NULL	NEGATIVE	NEGATIVE		NEGATIVE	NEGATIVE

<u>Breastfeeding risk factor</u>

There is much literature which claims that breastfeeding protects against eczema, some that claims a null (or not understood) effect, and some which argues that it exacerbates the disease. Overall breastfeeding is thought to reduce eczema risk. Investigating whether there is a link between breastfeeding and eczema, might help mothers make a more informed decision about whether to breastfeed.

Studies have been carried out by Heinrich *et al.* (60), Victora *et al.* (61), Elbert *et al.* (62), and Huang *et al.* (63), amongst others. A paper by Heinrich *et al.* (60) features a systematic review that was carried out by Victora *et al.* (61) which summarised all the literature and studies prior to 2 October 2014. 29 studies showed that exclusively breastfeeding for three-four months decreased the risk of developing eczema. It is important to consider whether exclusive or non-exclusive breastfeeding makes a difference. A paper by Elbert *et al.* (62) found that a shorter duration or non-exclusiveness of breastfeeding didn't make a difference to whether the child developed eczema. A cross-sectional study by Huang *et al.* (63) found that exclusive breastfeeding for more than six months reduced eczema. Breastfeeding does not only relate to how much breastfeeding is carried out (ie, a lot, if it is exclusive) but how much bottle (or other alternative) feeding is taking place. If it is thought that cow's milk contains allergens, it might not be the continuation and benefits of lots of breastfeeding but the avoidance of the allergens in cow's milk that is making the difference to the development of eczema. Generally, nowadays guidelines suggest that children under one are not given cow's milk. However, this may have been different in the 1990's. In future cow's milk and formula will be referred to as 'breast milk alternative'.

There is some scientific evidence on either side of the argument. For example, the case of breastfeeding/eczema in Japan (64) showed that exclusive breastfeeding for six months or more increased the risk of developing eczema (odds ratio 1.14; 95% confidence intervals 1.06, 1.23). However Heinrich *et al.* (60) reports on the randomised controlled trial carried out in Belarus (65). It found that there were less cases of eczema in the intervention group that included breastfeeding, (OR 0.54; 95% CI 0.31, 0.95). Table 1.3 provides a summary of the main literature on the risk factor breastfeeding.

Table 1.3: Broad summary of literature on breastfeeding

	Heinrich (60)	Victora (61)	Elbert (62)	Huang (63)	Overall
Breastfeeding	NEGATIVE	NEGATIVE	NULL	NEGATIVE	NEGATIVE

Genetic risk factors

Genetics also have a role to play in the cause of eczema (32). 31 genes in total have been found to be related to eczema (66), including the strongest risk factor, Filaggrin (67), (68). Variants identified

have both skin barrier and immune functions, but not all variants have known function. A genome-wide association study (GWAS) is a process whereby specific genes are linked to the presence of diseases or conditions, by studying the frequency of common genetic variants in large numbers of eczema cases compared to unaffected controls.

1.2 EPIGENETICS

1.2.1 Epigenetic factors

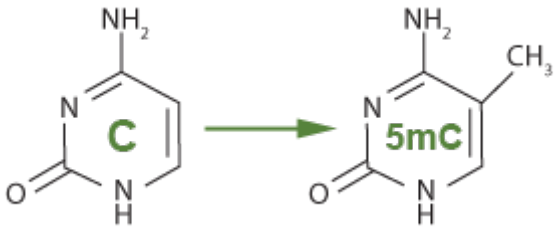
Genetic variation (naturally existing changes in DNA) is thought to be responsible for the development of many different diseases. Although epigenetics has been acknowledged for several decades, our understanding of it is relatively new, a route through which modifications are made to DNA without the sequence itself being affected. “Epi” refers to the fact that there are “additional” features which are sitting on top of the main DNA nucleotide chain and affecting expression of genes. It is the branch of genetics which determines which genes are turned on and off. The full DNA genome is still intact, but different genes will be expressed and others not expressed (69).

The term ‘epigenetics’ refers to a range of DNA modifications (70). Three examples are DNA methylation, histone modification (71) and miRNA’s, also known as microRNA’s (72). Histones are proteins which exist in the nucleus of cells. They are involved in the coiling up of the DNA double helix so that it fits in the nucleus. Whilst the main role of histones is to package the DNA into chromosomes, other processes such as transcriptional activation/inactivation also take place. Sometimes the coiling of the DNA becomes too tight or intricate for the DNA to be accessed and proteins therefore cannot be expressed. This is the influence of epigenetics.

1.2.2 DNA methylation

Methylation is a naturally occurring phenomenon and occurs when a methyl group (CH3), which is present in the nucleus, attaches to a cytosine base on the DNA chain, controlled by enzymes called methyltransferases (5-methyltransferase), to become 5-methylcytosine (see figure 1.3). This is known as a CpG (cytosine-phosphate-guanine) site, if there is a cytosine base next to a guanine base. DNA methyltransferase (DNMT’s) are a group of enzymes that facilitate the transfer of a methyl group to DNA. Demethylation occurs when the methyl group is removed. It is replaced by a hydrogen atom in the place of the methyl group. The Illumina Infinium HumanMethylation 450K (Illumina450K) has made it possible for DNA methylation to be measured at many individual CpG sites throughout the genome in a cost effective way on large numbers of individuals.

Figure 1.3: The process of methylation – a methyl group is attached to a cytosine base



CpG sites occur throughout the genome, but commonly occur in clusters close to the 5' ends of genes. These dense clusters of CpG sites are known as CpG islands, and they are comprised of over 50% cytosine and guanine bases in a >200 base pair region.

1.2.3 EWAS

Whilst there are many other methods, one main way of measuring DNA methylation is using methylation arrays such as the 450K Illumina Infinium Assay. This array is made up of 450K probes, each of which measure the proportion of methylation at a CpG site, either as unmethylated or methylated. The process starts with bisulphite conversion. This is a chemical process used to differentiate between methylated/unmethylated bases (73). When DNA is mixed with bisulphite, it converts unmethylated cytosine bases to uracil. Methylated cytosines (5-methylcytosine) remain as cytosine. This difference between cytosine and uracil helps identify which of the DNA CpG sites are methylated with a methyl group and which are not (74). Data generated from the Ilumina Infinium array is then used in an epigenome-wide association study (EWAS) analysis. Overall the EWAS looks at the association between methylation and eczema, and can identify loci where methylation may differ between cases and controls, or between an exposed or unexposed group in relation to a risk factor. However, it is not possible to infer anything about causality or the direction of effect.

1.2.4 Genetics of epigenetics

Epigenetics can be influenced by genetic factors, such genetic variants are called Methylation Quantitative Trait Loci (mQTL's) (75). This is where specific SNPs in the genome influence the methylation at CpG sites. It is thought that over 50% of CpG sites are affected by the presence of

mQTL's. There are two types of mQTL's: in-*cis* (which means very near to the CpG site, usually on the same chromosome) and in-*trans* (which can be physically further away and potentially on a different chromosome to where the CpG site is located) (75). *Cis* mQTL's usually have more of an effect on methylation at the CpG sites, as they are closer (76).

1.3 CURRENT LITERATURE OF EPIGENETICS OF ECZEMA

1.3.1 Search strategy

A search was carried out using PubMed. The following terms were used to search: “eczema”, “atopic dermatitis”, “EWAS”, “epigenetic” “methylation”. Studies were limited to human, rather than animal illness, and all those looked at included an abstract.

1.3.2 Systematically reviewing the literature

This study is looking at DNA methylation and eczema and whether there is a relationship between the two as demonstrated by EWAS analysis. Atopic diseases range across eczema, asthma and hayfever, amongst others. EWAS’s have been carried out on the other diseases (77, 78) and a GWAS on eczema (66), but no EWAS on eczema. There is therefore also no literature which links the risk factors, smoking, animal exposure and breastfeeding, with eczema, although there does exist EWAS on the risk factors and methylation. There is some evidence to say DNA methylation is associated with atopy, but there are few studies investigating DNA methylation associations with eczema. They were useful because other atopic diseases share similarities with eczema, but questions still exist, leading to why I conducted this study.

Of the literature out there looking at eczema, one study (79) looked at methylation in the Isle of Wight birth cohort (80), looking at DNA methylation at age 18 in the F1 generation of this study and cord blood methylation in the F2 generation. 88 CpG sites were associated with eczema, of which 41 were also seen in F2. However there is other literature on the topic, starting with EWAS’s on asthma (77) and Immunoglobulin E (7). After adjustment for cell count, some associations were observed but could be explained by elevated eosinophils which are a feature of the disease state in asthma (78). These are related because asthma is another atopic illness, and IgE is an antibody which is produced during an allergic reaction. IgE is a proxy for allergic disease.

1.4 MAIN QUESTIONS

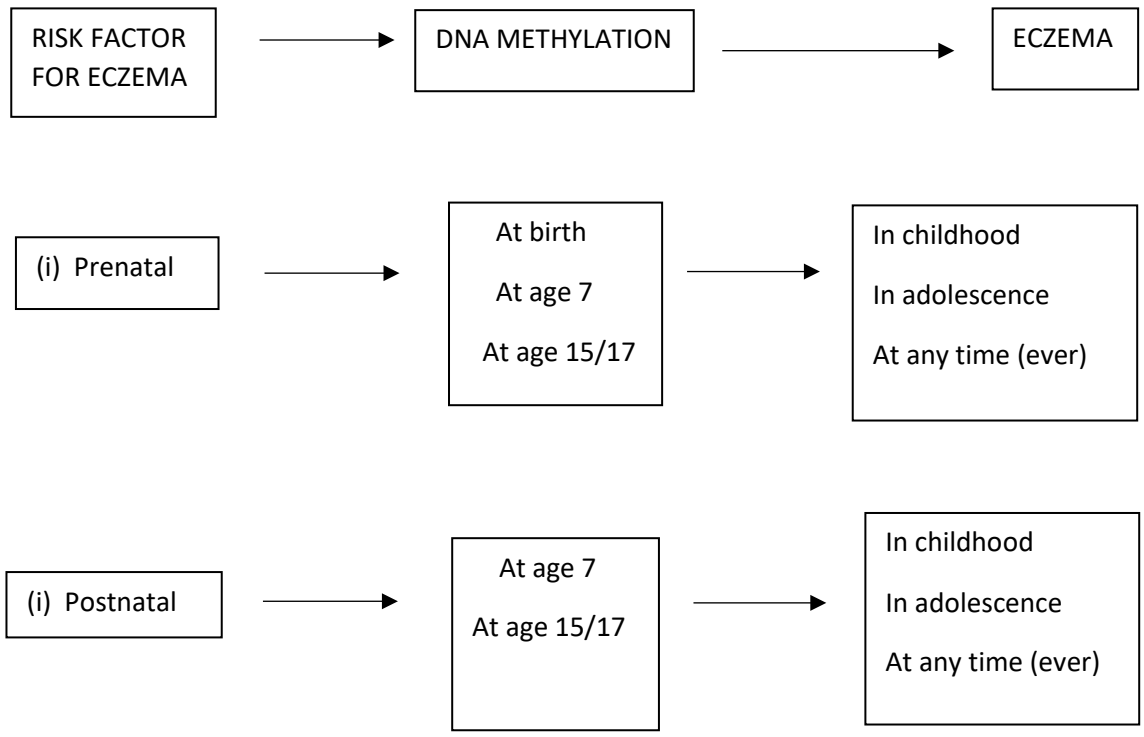
I will use ALSPAC (81) data to address the following questions: **1. Is childhood eczema associated with differences in DNA methylation?** I will compare cord blood methylation with whether a person has ever had eczema by the age of 15/17, or methylation at age seven and again, whether they have ever had eczema by the age of 15/17. In this situation, comparing cord blood may identify CpG difference present before birth which are associated with eczema. I will also conduct cross-sectional analysis, looking at methylation at ages seven and 15/17 and current eczema. This will give information on whether methylation in the blood can be affected by events or exposures in the last 12 months. The limitation of this is that reverse causation may be in play. Looking at methylation later on can demonstrate how plastic it is and how it has responded to changes in a person's life.

ALSPAC is richly phenotyped with eczema status throughout childhood, which is helpful because gaps remain in this field of research. Whilst much is known about eczema, the genetic predisposition of a person to develop it, and the environmental triggers of the disease (82), much is still unknown. A substantial EWAS has not been carried out on the disease, even though EWAS have been carried out for asthma and IgE. If DNA methylation is shown to vary between those with and without eczema, I can then look at the factors which affect the methylation itself and use both the exposure and the methylation as a predictor of risk factors for the disease. I can then see if DNA methylation acts as a mediator between certain risk factors and the development of eczema. There does exist the issue of reverse causation, which would imply that eczema itself causes alterations in DNA methylation rather than the other way around. Through examining cord blood, I can attempt to overcome this. However, there is still reason to look at later methylation time points. Cord blood is useful for looking at methylation. Methylation in a new born has still been exposed to inter-uterine exposures. DNA methylation measured at birth measures DNA methylation patterns in individuals unexposed to later life exposures. If I want to look at how an environmental exposure has affected methylation, for example passive smoking during childhood, it is helpful to look at later methylation, say at age seven or 15/17, and compare this to cord blood to identify changes. That is why the three timepoints identified, birth, age seven and age 15/17, are all utilised in this study.

2. Is there any evidence that methylation mediates the relationship between exposures and eczema? Factors such as smoking, animal exposure, and breastfeeding, may have an influence on the methylation profile of an individual, which in turn may result in eczema risk. See figure 1.4. Here

I will look at early life exposure on eczema development, so mainly looking at child based questionnaires up to the age of seven, and mother questionnaires up to the child age of seven (including during pregnancy). I will mainly look at methylation at age seven as this would have given time for the exposure to influence the child's DNA. The child based questionnaires would give an indication of the type of environment a child is growing up in and the risk factors they are exposed to. The mother's questionnaires would provide similar information about the child's environment, but also give details of risk factors the child was exposed to in utero. Here I am trying to determine if it is possible that DNA methylation mediated the relationship between exposure and eczema.

Figure 1.4: Prenatal and postnatal implications of the risk factor -> DNA methylation -> eczema process



CHAPTER 2. METHODS

2.1 POPULATION/STUDY

The data that will be used in this study comes from the Avon Longitudinal Study of Parents and Children (ALSPAC) (83) (84) and the Accessible Resource for Integrated Epigenomics Studies (ARIES) (85). ARIES is a subset of ALSPAC for which DNA methylation data has been generated at childhood timepoints birth, age 7, and age 15/17 years.

ALSPAC has been described in full (83, 84), but here I briefly summarise the recruitment and data collection. The ALSPAC study has been running for over 26 years and is based in Bristol. ALSPAC recruited 14,541 pregnant women resident in Avon, UK, with expected dates of delivery which fell between 1st April 1991 to 31st December 1992. 14,541 was the initial number of pregnancies for which the mother enrolled in the ALSPAC study and had either returned at least one questionnaire or attended a Children in Focus clinic by 19/07/99. Of these initial pregnancies, there was a total of 14,676 fetuses, resulting in 14,062 live births and 13,988 children who were alive at one year of age. When the oldest children were approximately seven years of age, an attempt was made to bolster the initial sample with eligible cases who had failed to join the study originally. The total sample size for analyses using any data collected after the age of seven is therefore 15,247 pregnancies, resulting in 15,458 fetuses. Of this total sample of 15,458 fetuses, 14,775 were live births and 14,701 were alive at one year of age. Excluded were those mothers who had moved out of the area or were lost to follow-up, and those partaking in another study of infant development in Avon (81). Questionnaires were given to children and mothers, as well as fathers, and clinics were held to collect samples of blood, nails and DNA, talk through questionnaires and anything else the participants wanted to discuss.

The ARIES (85) subset is a collection of 1,024 individuals from ALSPAC (81). They were specifically chosen from a selection of positive responders from the ALSPAC dataset. Although 15,445 mothers partook in ALSPAC, only 1,024 of these were included in ARIES. Those in ALSPAC took part in completing the questionnaires and providing some samples, but the 1,024 from ARIES all provided blood DNA samples in cord blood, at age 7 and at age 15/17. Blood from 1018 mother–child pairs (children at three time points and their mothers at two time points) were selected for analysis using the Illumina Infinium 450K array (76).

2.2.1 Illumina Infinium 450K Array

The Illumina Infinium 450K array was used to measure the methylome of the children. It has 450k probes, which each measure the percentage of methylation at a CpG site. Each probe targets a CpG site and measures the proportion of copies of DNA in the sample that are methylated or unmethylated. A beta value is generated which represents the proportion of methylation at each specific CpG site. Beta values range from 0 to 1. 0 = unmethylated and 1 = methylated. The assay is run in the following way: After DNA extraction, samples were bisulphite converted using the Zymo EZ DNA Methylation™ kit (Zymo, Irvine, CA, USA). Following conversion, genome-wide methylation was measured using the Illumina Infinium HumanMethylation450 (HM450) BeadChip. The arrays were scanned using an Illumina iScan, with initial quality review using GenomeStudio (76). ARIES was pre-processed and normalised using the meffil R package (86). In this study measurements will be taken at birth, from cord blood, and at ages seven and 15/17. ARIES consists of 5469 DNA methylation profiles obtained from 1022 mother-child pairs measured at five time points (three time points for children: birth, childhood and adolescence; and two for mothers: during pregnancy and at middle age).

2.2 EXTRACTING QUESTIONS, DEFINING AND TABULATING

2.2.1 Eczema

Extracting eczema questions

The Avon Longitudinal Study of Parents and Children (ALSPAC) study is comprised of 25 child based questionnaires, 25 child completed questionnaires, 21 mother questionnaires and one father questionnaire. Child based questionnaires are focussed on the child but completed by the mother, whereas child completed, which begin at approximately age 10, are filled in by the child themselves. In deriving the eczema variable, I used the child based and child completed questionnaires. Of these, 31 different questions relating to eczema and/or rash were available. Some were repeated in several questionnaires, others were stand alone questions. In total there were 189 questions. From these 189, 13 were selected, see table A1. Questions about simple 'rash', not specified to eczema, were not used. The three main fundamental questions chosen were based around "Has your child had eczema in the last 12 months?" (seven questions), "Has a doctor ever diagnosed your child with eczema?" (two questions) and "Has your child *ever* had eczema?" (four questions). See the Appendix for table A1 of the 13 eczema questions:

Eczema definition

In this section I generated three basic variables:

i) Have you ever had eczema? (up to 15/17 years of age)

A case (1 (Yes)) referred to a 'yes' to any question out of the 13 in table A1, and a control (0 (No)) was not a case and had answered 'no' to any question of 'Have you ever had eczema?' In ALSPAC it was found that 47.0% of children had eczema, and 53.0% didn't, at some point in their life by the timepoint of age 15/17 (see table 2.1). 47% of children in ALSPAC with eczema is quite high (when compared to the 25% of children affected mentioned earlier). This could be down to the distribution. A control is not a case and answered no to 'ever eczema?'. If a control was just, "not a case" the number of controls would be higher and the cases less.

Table 2.1: Ever had eczema in ALSPAC and ARIES

	ALSPAC		ARIES	
Ever eczema	**Freq.**	**Percent**	**Freq.**	**Percent**
0 (No)	4,891	53.0	487	49.1
1 (Yes)	4,336	47.0	505	50.9
Total	9,227	100.00	992	100.00

ii) Have you had eczema in the last 12 months at age 7?

A case (1 (Yes)) was defined as such if a person answered 'yes' (1) to option 1 (yes and saw a dr) or 2 (yes but didn't see a dr) of question kr042 ("Have you had eczema in the last 12 months?"). A control (0 (No)) was defined if a person answered 'no' (0) to option 3 (no). Here 16.3% of people had eczema in the past 12 months, whilst 83.7 didn't (see table 2.2)

Table 2.2: Ever had eczema at age 7 in ALSPAC and ARIES

	ALSPAC		ARIES	
Eczema in the last 12 months at age 7	**Freq.**	**Percent**	**Freq.**	**Percent**
0 (No)	6,866	83.7	757	81.3
1 (Yes)	1,340	16.3	174	18.7
Total	8,206	100.00	931	100.00

iii) Have you had eczema in the last 12 months at age 15/17?

A case (1 (Yes)) was a positive answer to option 1 (yes and saw a dr) or 2 (yes but didn't see a dr) of question ccs5023. A control (0 (No)) was a positive answer to option 3 (no). Question ccs5023 was "Have you had eczema in the last 12 months?" at 192 months. Here 15.3% of people had eczema in the past 12 months, whilst 84.7% didn't (see table 2.3)

Table 2.3: Ever had eczema at age 15/17 in ALSPAC and ARIES

	ALSPAC		ARIES	
Eczema in the last 12 months at age 15/17	**Freq.**	**Percent**	**Freq.**	**Percent**
0 (No)	4,233	84.7	640	86.3
1 (Yes)	762	15.3	102	13.8
Total	4,995	100.00	742	100.00

Here 15.3% of people had eczema in the past 12 months, whilst 84.7% didn't.

2.2.2 Smoking

Extracting smoking questions

There were a total of 80 questions related to smoking in the ALSPAC questionnaires, some from the mother questionnaires and some from the child based. The three "smoking during pregnancy" questions that were pulled from the questionnaires were taken from the mother questionnaires because they addressed the mother's behaviour during pregnancy. Specific questions are detailed in Appendix I, tables A2 and A3.

Smoking definition

Smoking during pregnancy

Current smokers were defined as those who reported smoking (question code b650) and reported that they had not stopped smoking (question code b659). Former smokers were defined as those who reported smoking (question code b650) and reported that they had stopped smoking (question code b659). Individuals were defined as never smokers if they were NOT either of the first two categories. In ALSPAC it was found that 56.6% of people never smoked, 25.4% smoked before pregnancy but stopped and 18.0% smoked during pregnancy. In the ARIES subset, the outcome was

similar in that the majority never smoked (62.40%), 26.6% smoked but gave up and 11.0% kept smoking (see table 2.4). Questions b650, b659 and b665 relate to the mother and baby at 18 weeks gestation. One possible problem here is that some mothers who didn't answer the above questions would have automatically been classed as belonging to the non-smoker group.

Table 2.4: Smoked during pregnancy in ALSPAC and ARIES

	ALSPAC		ARIES	
Smoked_during_pregnancy	**Freq.**	**Percent**	**Freq.**	**Percent**
0 (Never smoked)	8,741	56.6	639	62.3
1 (Smoked before pregnancy, but stopped)	3,918	25.4	272	26.6
2 (Yes)	2,786	18.0	113	11.0
Total	15,445	100.00	1,024	100.00

Smoky environment

To define smoky environment, two questions were taken from the child based questionnaires, as they related to the environment the child was growing up in during the first 6 months of life. They addressed how often during the day the baby is in a room or enclosed place where people are smoking, both on weekdays and weekends. Exposure to smoky environment was defined as an answer of either "all the time (1)", "more than 5 hours (2)", "3-5 hours (3)" and "1-2 hours (4)". 83.0% of children did not grow up spending a lot of time in a smoky room, and 17.0% did (see table 2.5).

Table 2.5: Smoky environment in ALSPAC and ARIES

	ALSPAC		ARIES	
Smoky_environment	**Freq.**	**Percent**	**Freq.**	**Percent**
0 (No)	12,827	83.1	890	86.9
1 (Yes)	2,618	17.0	134	13.1
Total	15,445	100.00	1,024	100.00

2.2.3 Animal exposure

Extracting animal exposure questions

There were a total of 113 questions relating to animal exposure in the ALSPAC questionnaires. These questions were mainly based around cats, dogs, rabbits, rodents and birds. Cat and dog exposure was investigated as they are the pets for which there is evidence of association with eczema. Questions asked if the study participants are in contact with the animals at least once a week at the age of 15 months. Full details of the study questions can be found in Appendix I, tables A4 and A5.

Animal exposure definition

Individuals were classified as being exposed to cats or dogs if they reported 'yes' to being in contact with the animals at least once a week in the home or elsewhere.

In ALSPAC fewer children are exposed cats (29.5%) than are unexposed (70.5%). In ARIES, the proportion of children with exposure to cats is higher than in the whole cohort (43% vs 29.5%) (see table 2.6).

Table 2.6: Cat exposure in ALSPAC and ARIES

	ALSPAC		ARIES	
Cat_exposure	**Freq.**	**Percent**	**Freq.**	**Percent**
0 (No)	10,895	70.5	584	57.0
1 (Yes)	4,550	29.5	440	43.0
Total	15,445	100.00	1,024	100.00

Similarly, fewer children are exposed to dogs (30.3%) than are unexposued (69.7%). In ARIES, the proportion of children exposed to dogs is slightly higher than that found in the overall cohort (35.6% vs 30.3%) (see table 2.7).

Table 2.7: Dog exposure in ALSPAC and ARIES

	ALSPAC		ARIES	
Dog_exposure	**Freq.**	**Percent**	**Freq.**	**Percent**
0 (No)	10,764	69.7	659	64.4
1 (Yes)	4,681	30.3	365	35.6
Total	15,445	100.00	1,024	100.00

2.2.4 Breastfeeding

Extracting breastfeeding questions

The ALSPAC questionnaires contain 21 questions relating to maternal breastfeeding. There were 19 child based questions which focussed on whether the child was breast fed and for how long. Seven questions were chosen, at the age of 4 weeks. These asked how the baby was fed during: 1) First 24 hours, 2) 1st week, 3) 2nd week, 4) 3rd week and 5) 4th week. Most studies look at breastfeeding for a longer duration than this, but 4 weeks was chosen; any longer and mothers may have finished breastfeeding; any shorter and the effects of being breastfed might not materialise. There were three main variable categories for each question: Breast only (1), Bottle only (2), and Breast and bottle (3). See Appendix I table A6 for full details of questionnaire data used.

Breastfeeding definition

A “breast only” category included those babies who were fed by breast in all questions posed up to the age of 4 weeks. Conversely, a “bottle only” category included those babies who were fed by bottle only in all questions posed. By creating a by creating a “breast only” category and a “bottle only” category, everything else was categorised as “breast and bottle”.

In ALSPAC most of the mothers never breastfed (51.1%), followed by 36.0% always breastfeeding (see table 2.8). In ARIES the proportion of babies who were never breastfed was much lower than in the whole cohort (17.4% vs 51.1%) and a much larger proportion were always breastfed up to 4 weeks of age (63.5% vs 36.0%).

Table 2.8: Breastfeeding in ALSPAC in the first 4 weeks of life

	ALSPAC		ARIES	
Breastfeeding	**Freq.**	**Percent**	**Freq.**	**Percent**
0 (Never breastfed)	5,709	51.1	130	17.4
1 (Breast and bottle fed)	1,440	12.9	142	19.0
2 (Always breastfed)	4,023	36.0	474	63.5
Total	11,172	100.00	746	100.00

Breast and bottle fed had the smallest amount of people with just 12.9%. However, in ARIES this looks a little different. Most mothers always breastfed (63.5%), and the lowest number of mothers never breastfed (17.4%). Those mothers who mixed breastfeeding and bottle feeding had a value between the other two, although it was closer to "never breastfed" (19.0%). There is a difference between ALSPAC and ARIES with more people breastfeeding in ARIES, which may be because the mothers who were chosen for ARIES were shown to have strong compliance with the study and dedication to giving data and samples. Therefore, it might follow that these people could be said to be more conscientious with all things health related, like breastfeeding.

2.2.5 Why socioeconomic status?

In this study I focus on confounders based around socioeconomic status. Papers have been published which demonstrate that, generally, people in higher social classes compared to those in lower classes are more likely to develop eczema. Socioeconomic status has been defined by maternal and paternal social class based on occupation variables C755 and C765. The occupation was coded using job codes, and the social class categorisation was derived. Each question had the answers: 'Missing', I, II, III (non-manual), III (manual), IV, V and 'Armed Forces'. These were dichotomised into 'non-manual' and 'manual'. The former included I, II, and III (non-manual), and the latter included III (manual), IV and V. In the next chapter, I look at confounding, several confounders being based around socioeconomic status, which is why their definition is highlighted here.

Table 2.9: Tabulation of socioeconomic status in ALSPAC and ARIES

	ALSPAC		**ARIES**	
	Freq.	**Percent**	**Freq.**	**Percent**
0. High social class (Non-manual)	5,885	51.04	584	59.84
1. Low social class (Manual)	5,646	48.96	392	40.16
Total	11,531	100	976	100

As can be seen in table 2.9 there is a higher percentage of people in a higher class in ARIES than in ALSPAC.

2.3 STATISTICAL ANALYSIS

Statistical analyses were run using STATA (15MP) and R (3.5.1).

2.3.1 Exploratory analysis

I used STATA to create datasets and definitions and run logistic regressions between potential confounders and eczema status.

2.3.2 Epigenome wide association studies (EWAS)

I then used R to run a set of EWAS's. The package I used was *meffil* . Multiple testing was dealt with by using a Bonferroni correction. This reduces type 1 errors that are the result of multiple testing and a significant result appearing simply by random error. Methylation patterns differ by cell type thus posing an issue in analyses conducted on heterogeneous cell populations. If cell composition is related to the phenotype of interest, cell composition can be a confounder in the association between methylation and the phenotype. In ARIES methylation levels are measured in blood cells, such as Eosinophils, a type of white blood cell that is increasingly activated in patients with the disease (6), which produces Immunoglobulin E, an antibody (85). It is important to adjust for cell count when looking at eczema because you may get different amounts of different cells due to severity of the illness. Each of these cells will have different methylation which could skew the results making levels seem higher or lower. Adjusting for the cell count or composition when an EWAS is run helps get around this problem of confounding by cell count and gives results not confounded by cell composition. Cell composition was estimated from methylation data using a method developed by Houseman *et al.* (87)and Bakulski *et al.* (88). Surrogate variable analysis was conducted using *sva* (implemented in *meffil*) and to remove unknown sources of variation. Plots were generated using *meffil* and ggplot2. Code used to run EWAS analysis is shown in Appendix II.

EWAS analyses were iterative models with increasing complexity. Below in table 2.10 I am showing the models (i) to (vi) testing the association between cord blood DNA methylation (the predictor) with ever eczema (the outcome). To avoid repetition, the other three sets of analyses follow the same six models, but look at methylation at age 7 and whether a person has ever had eczema, methylation at age 7 and whether a person has had eczema in the last 12 months and methylation at

age 15/17 and whether a person has ever had eczema in last the 12 months. In analyses of cord blood methylation, smoky environment and breastfeeding are not included in the models as they occur after the DNA methylation is measured.

2.3.3 Makeup of models from (i) to (vi)

Table 2.10: Explanation of models (i) to (vi) and what they each adjust for

Model (i)	Adjusted for sex and surrogate variables
Model (ii)	As for model (i), plus two socioeconomic status variables (social class and child ethnic background)
Model (iii)	As for model (ii), plus maternal history
Model (iv)	As for model (iii), plus cell counts
Model (v)	As for model (iv), plus risk factors and minus cell counts
Model (vi)	As for model (v), plus cell counts

CHAPTER 3. DATA PREPARATION

3.1 INVESTIGATING MISSINGNESS

This chapter addresses the data taken from ALSPAC and how it was prepared for analysis. To start, I looked at missingness, which questions were unanswered and by whom. Participants not answering certain questions can introduce bias to the study, in both directions, and the reasons as to why they don't answer a question are important and can reveal much about the study design. Secondly, I explored confounding. Confounding is where a variable is associated with both the input and output variables and can therefore leading to misleading results. By adjusting for confounders, we can partially remove this bias. However, it is important to select the confounders carefully, as I do in section 3.2.

3.1.1 Missing data

In statistical analysis there are three types of missing data: 'missing completely at random', 'missing at random' and 'missing not at random'. 'Missing completely at random' are when the missing data is independent of the variables and completely random. 'Missing at random' is when missingness is not random but can be accounted for. For example, people of lower SES are more likely to be missing. 'Missing not at random' is when, for example, someone can't answer a question on how ill they are *because* they are too ill. These definitions can be used when looking at the questions asked about eczema and risk factors.

Four variables were looked at: eczema, smoking, animal exposure and breastfeeding. When looking at ALSPAC data, the category with most missing values is eczema. Here, the answers to questions used in the 'eczema' definition were missing around 50% of the time. In ARIES however this dropped to below just 10% of the time. Two questions in particular that were badly answered were questions ccs023 and cct4055, asking whether or not the child had had eczema in the past 12 months at age 192 months, and whether a child had ever had eczema at 216 months. This would push the conclusion towards less eczema being present in the child population because a missing answer to a question might be categorized as a 'no' to whether a person has had eczema in the last 12 months or ever. For the smoking variable, as well as the animal exposure and breastfeeding variables, the answers to the questions in the ALSPAC and ARIES dataset showed a much lower degree of missingness. Question kb550, which asked the parent to indicate how often during the day the baby

is in a room or enclosed place where people are smoking on weekends, at the age of 6 months, was least well answered. This may be because parents are more aware of their children's environment on the week days than weekends. Or they may feel that their answers to the questions relating to week days and weekends have similar answers, if the child spends his/her time in a similar way throughout the week. Question kc370 and kc371 asked how much time the child spends with cats and dogs. And questions ka031, ka032 and ka033, which asked how the child was fed in the 1st, 2nd and 3rd week of life, were also answered least. This might be because the first three weeks of a child's life are so stressful that it may be difficult for the mother to remember how they were fed. Or they may answer incorrectly, introducing bias in both directions, because, as is shown so much in guidelines and literature, they were embarrassed or ashamed about whether or not they breastfed.

There are many varied reasons as to why a question is not answered, and these can bias the results in either direction. Reasons for missingness might include questions relating to a private subject, information on which the mother or child did not want to divulge. The child or mother might not understand what eczema is and therefore being unable to report it even if it has occurred. Other reasons could include forgetting to answer a question, and language or a disability being a barrier. Since eczema is unlikely to cause death in an individual, it would not be the case that people are dying and thus not able to follow up their responses by questionnaire. Answering a questionnaire fully, and then leaving a follow-up question blank could be because the person's circumstances may have changed (illness, income, etc) and they may no longer be able to partake in the study. I think the drop in numbers in people answering questions between birth and up to the age of 15/17 is due to loss to follow up in ALSPAC.

Table 3.1 shows how many answers to questions were missing out of the four key questions asked about eczema, smoking, animal exposure and breastfeeding.

Table 3.1: Missingness present in eczema, smoking, animal exposure and breastfeeding questions

Question	Age	ALSPAC (15,445)		ARIES (1,024)	
		Missing	Non-missing	Missing	Non-missing
kq035	81 months	6946 (45.0%)	**8499 (55.0%)**	98 (9.6%)	**926 (90.4%)**
kr042	91 months	7239 (46.9%)	**8206 (53.1%)**	93 (9.1%)	**931 (90.9%)**
ks1042	103 months	7225 (46.8%)	**8220 (53.2%)**	108 (10.6%)	**916 (89.5%)**
kv1060	128 months	7661 (49.6%)	**7784 (50.4%)**	122 (11.9%)	**902 (88.1%)**
ta1030	157 months	8444 (54.7%)	**7001 (45.3%)**	146 (14.3%)	**878 (85.7%)**
tb1060	166 months	8397 (54.4%)	**7048 (45.6%)**	138 (13.5%)	**886 (86.5%)**
ccs5023	192 months	10450 (67.7%)	**4995 (32.3%)**	282 (27.5%)	**742 (72.5%)**
kv1070	128 months	7631 (49.4%)	**7814 (50.6%)**	125 (12.2%)	**899 (87.8%)**
tb1070	166 months	8373 (54.2%)	**7072 (45.8)**	137 (13.4%)	**887 (86.6%)**
kv1122	128 months	7880 (51%)	**7565 (49%)**	149 (14.6%)	**875 (85.4%)**
tb1122	166 months	8624 (55.8%)	**6821 (44.2%)**	176 (17.2%)	**848 (82.8%)**
tc6110	198 months	9852 (63.8%)	**5593 (36.2%)**	208 (20.3%)	**816 (79.7%)**
cct4055	216 months	12,113 (78.4%)	**3332 (21.6%)**	506 (49.4%)	**518 (50.6%)**
b650	18 weeks gest	2196 (14.2%)	**13,249 (85.8%)**	25 (2.4%)	**999 (97.6%)**
b659	18 weeks gest	78 (1.2%)	**6661 (98.8%)**	6 (1.5%)	**383 (98.5%)**
b665	18 weeks gest	2086 (13.5%)	**13,359 (86.5%)**	17 (1.7%)	**1007 (98.3%)**
kb548	6 months	4114 (26.6%)	**11,331 (73.4%)**	43 (4.2%)	**981 (95.8%)**
kb550	6 months	4207 (27.2%)	**11,238 (72.8%)**	49 (4.8%)	**975 (95.2%)**
kc370	15 months	4522 (29.3%)	**10,923 (70.7%)**	51 (5.0%)	**973 (95.0%)**
kc371	15 months	4522 (29.3%)	**10,923 (70.7%)**	51 (5.0%)	**973 (95.0%)**
ka030	4 weeks	3531 (22.9%)	**11,914 (77.1%)**	44 (4.3%)	**980 (95.7%)**
ka031	4 weeks	3579 (23.2%)	**11,866 (76.8%)**	44 (4.3%)	**980 (95.7%)**
ka032	4 weeks	3580 (23.2%)	**11,865 (76.8%)**	42 (4.1%)	**982 (95.9%)**
ka033	4 weeks	3581 (23.2%)	**11,864 (76.8%)**	38 (3.7%)	**986 (96.3%)**
ka034	4 weeks	3458 (22.4%)	**11,987 (77.6%)**	33 (3.2%)	**991 (96.8%)**
ka061	4 weeks	3164 (20.5%)	**12,281 (79.5%)**	27 (2.6%)	**997 (97.4%)**
ka094	4 weeks	3170 (20.5%)	**12,275 (79.5%)**	29 (2.8%)	**995 (97.2%)**

3.2 CONFOUNDERS

It is important to adjust for obvious confounders to prevent influencing the results. Otherwise there may appear to be an association when in fact there is not. I decided to start by using a paper by Granell *et al.* (89) which put forward a list of 10 potential confounders associated with environmental variables and asthma. As an atopic illness, these confounders may be relevant to the study of eczema. I therefore decided to start with the 10 confounders and reduce them to a succinct list to use in this study. I did this through logistic regressions analysis to see whether there is in fact a relationship between eczema and the potential confounder, and to see the independence of the relationship between the confounders and socioeconomic status. I wouldn't want to adjust for confounders that have a very similar effect.

3.2.1 Is there a relationship between eczema and each potential confounder?

A confounder is a variable which can affect both the independent and dependent variable simultaneously. In this case it is a variable which affects both methylation and eczema. This can be environmental (for example, smoking) or genetic (mQTLs). If the confounder affects both the exposure and the outcome there may not be any real association between the exposure and outcome, they could both just be being affected by confounding. I carried out a series of logistic regressions between each confounder and eczema, see table 3.2. Based on p-values, those relationships with strong evidence for association (for example, "Maternal history of asthma or allergy") are highlighted red. Those with moderate evidence for association are highlighted orange. Those left blank indicate a high p-value. I could potentially drop two of the socioeconomic variables, ("Home ownership status" and "Single mother"), from the SES variables, as well as "Low birth weight", "Maternal age at delivery" and "Preterm", as these do not look to have an association with eczema. One of the variables that I will include is maternal history of allergy as there is longstanding evidence that allergic disorders can be heritable (90), and evidence here shows that it is associated with eczema in our data set. In future studies it may be beneficial to look at paternal history as well as maternal, but seeing as though there is only one father questionnaire compared to 21 mother questionnaires, I shall leave this for now. In table 3.2, the confounders are listed in order of size of p-value, from smallest to largest. For more information on the confounders chosen, see table 3.3.

Table 3.2: Relationship between 10 confounders and eczema, in order of increasing p-value

CONFOUNDERS	ECZEMA	
	ODDS RATIO (Confidence Intervals)	P-values
MATERNAL HISTORY OF ASTHMA OR ALLERGY	1.41 (1.29, 1.54)	6.4e-15
SOCIAL CLASS	0.79 (0.73, 0.87)	3.0e-7
CROWDING INDEX	0.90 (0.81, 1.00)	0.049
CHILD ETHNIC BACKGROUND	1.37 (1.10, 1.70)	0.004
SEX	1.27 (1.18, 1.39)	<0.001
PRETERM DELIVERY (<37 WEEKS)	0.85 (0.71, 1.03)	0.095
HOME OWNERSHIP STATUS	1.01 (0.90, 1.14)	0.834
SINGLE MOTHER	1.02 (0.91, 1.13)	0.764
MATERNAL AGE AT DELIVERY	1.01 (0.93, 1.10)	0.765
LOW BIRTH WEIGHT (<2500gr)	0.93 (0.76, 1.13)	0.464

* SES variables are in green

** Red highlighting indicates very low p-values

*** Orange highlighting indicates moderately low p-values

3.2.2 Is each confounder independent?

I think that social class is the most important SES variable. I conducted a regression analysis between social class (SES defined by manual/non-manual work) and ever eczema with other SES confounders included as covariates. I looked at the effect of each confounder and concluded that 'home ownership status', 'single mother' and 'crowding index' variables were not independent of social class. I therefore did not include those confounders in my EWAS models.

I also tested for correlation between social class and home ownership status (as one example). The results show that the two variables are highly correlated (a very small p-value of $P<0.05$). Because of

that and the tests above I therefore don't think it is necessary to include all of the SES related variables.

3.2.3 Confounders included in EWAS analysis

Table 3.3: Reasons to include/exclude the 10 confounders:

VARIABLE	Include as a covariate in EWAS analyses (Yes / No)	WHY?
Sex	Yes	Associated with eczema, table 3.2. Important confounder included in most studies
Social class (91) (92)	Yes	Associated with eczema, table 3.2. (91) shows upper class can make atopy worse
Child ethnic background (92) (93) (94)	Yes	Associated with eczema, table 3.2. (94) shows eczema is more prevalent in Asian/Black than Whites
Maternal history of allergy	Yes	Associated with eczema, table 3.2. Known that eczema is hereditary (90)
Low birth weight	No	Not associated with eczema, table 3.2. Little literature to support association between low birth weight and eczema
Maternal age at delivery	No	Not associated, table 3.2. Little literature to support association between maternal age at delivery and eczema
Preterm	No	Not associated, table 3.2 Little literature to support association between preterm birth and eczema
Home ownership status	No	Not associated, table 3.2. Not independent of "socioeconomic status", (shown by Chi2 in 3.2.3)
Single mother	No	Not associated with eczema, table 3.2. Not independent of "socioeconomic status", (shown by Chi2 in 3.2.3)
Crowding index	No	Slight association with eczema, table 3.2. However, not independent of "socioeconomic status", (shown by Chi2 in 2.3.2)

CHAPTER 4. ECZEMA EWAS'S

4.1 INTRODUCTION

The main question is: **Is childhood eczema associated with differences in DNA methylation in blood?** I looked at The Avon Longitudinal Study of Parents and Children (ALSPAC) (81) and the Accessible Resource for Integrated Epigenomics Studies (ARIES) (85) to address the question. Using questionnaire data, whether a child has ever had eczema and/or whether they have had it in the past 12 months were identified. DNA methylation was measured at birth in cord blood and thereafter at ages seven and 15/17. Since this is a childhood analysis, I considered cord blood and childhood blood methylation. Methylation in cord blood is influenced by genetics and environmental factors in utero. Looking at cord blood also avoids the issue of reverse causation because eczema will not have been able to influence the methylation. ALSPAC is well suited to answer key questions. The data set is large and comprehensive, and covers mothers from when they are pregnant, right up to when the child is an adult. There are different questionnaires, so each period of life is ensured to be covered. For example, during childhood, a mother will fill out child based as well as mother completed questionnaires and the child will fill out child completed questionnaires. Information on a large number of topics, phenotypes and circumstances has been collected, which can be investigated as a study or added as confounders to make a study/analysis more accurate and realistic. ARIES is a useful resource which measures methylation at three time points; birth, age seven and age 15/17. This gives information on methylation caused by in utero effects before the child is born, methylation in childhood and methylation at the point of transition into adulthood. Blood is being used rather than skin tissue because blood is the available sample in ARIES. Skin tissue would be more relevant in measuring eczema because eczema directly affects the skin.

Epigenetics can act as a mediator between risk factors and eczema via DNA methylation. It may also identify biomarkers of future disease risk, highlighting 'at risk' individuals better than using SNPs alone. If methylation at CpG sites is influencing the development of eczema, it may be possible in the future to develop a treatment that targets this methylation. Carrying out an EWAS is unprecedented and therefore important. The set of EWAS's being conducted will cover 450K probes but will not be targeted to selected probes. Rather, the analysis will be open to all probes which may be significantly associated with the variable being tested (risk factor or eczema in this study).

4.2 AIMS

1. **Investigate whether there is a relationship between cord blood methylation and ever eczema.** This longitudinal analysis involves looking at methylation which has occurred in utero during pregnancy (which avoids the situation of reverse causation) and whether a person has ever had eczema in their life by the age of 15/17.

2. **Investigate whether there is a relationship between blood methylation at age seven and ever eczema.** Measuring methylation at age seven, again in a longitudinal analysis, and looking at whether a person has ever had eczema by the age of 15/17 provides a great deal of information regarding early childhood exposures.

3. **Investigate whether there is a relationship between blood methylation at age seven and eczema in the last 12 months.** This is a cross-sectional analysis looking at whether eczema develops during a limited 12-month long period at the age of seven. Looking at methylation at age seven has the same advantages as question 2 in that you can look at the effect of variables during the child's first seven years of life.

4. **Investigate whether there is a relationship between blood methylation at age 15/17 and eczema in the last 12 months.** This second cross-sectional analysis again looks at eczema in the last 12 months, but this time looks at methylation at age 15/17.

4.3 METHODS

For each question, six models analysed using regression will be carried out with various confounders and variables added or removed. These are:

(i) Sex and surrogate variables

(ii) Sex, surrogate variables and two socioeconomic status variables (social class and child ethnic background)

(iii) Sex, surrogate variables, two socioeconomic status variables (social class and child ethnic background) and maternal history

(iv) Sex, surrogate variables, two socioeconomic status variables (social class and child ethnic background), maternal history and cell count

(v) Sex, surrogate variables and two socioeconomic status variables (social class and child ethnic background), maternal history and risk factors

(vi) Sex, surrogate variables and two socioeconomic status variables (social class and child ethnic background), maternal history, risk factors and cell count

The statistical method being used is logistic regressions between methylation at CpG sites and eczema. The analysis was conducted in the software package R. When analysing cord blood methylation as exposure, in models (v) and (vi) the risk factors adjusted for will be smoked during pregnancy, cat exposure and dog exposure. For methylation measured later (at 7 and 15/17), the risk factors will also include a smoky environment and breastfeeding. This is because a smoky environment during childhood and breastfeeding could not have influenced cord blood methylation. Overall I will focus on models (iv) and (vi). The only difference between the two is that model (vi) adjusts for risk factors whereas model (iv) doesn't. These were chosen because they include the maximum amount of confounders and will therefore yield the most accurate results.

4.4 EWAS RESULTS

4.4.1 Overview of all EWAS analyses

Out of all the results, none met the stricter threshold I used at P<1x10-8. However there are 25 sites across model (iv) of the four questions, which can be identified as 'suggestively associated' at the p-value P<0.05. In question 4, the associations were all at P>0.05. In this chapter I will focus on model (iv) which is a regression model testing for the association between eczema and methylation (the outcome) adjusting for sex, surrogate variables, two socioeconomic status variables (social class and child ethnic background), maternal history and cell count. This gives the number of associations in each question at P<0.05 as 2, 4, 1 and 18. The total number of associations in each question, the sum of the six models, is 20, 44, 11 and 96 respecitvely at the P<0.05 threshold. The fourth question, looking at methylation at age 15/17 and whether a person has had eczema in the last 12 months, yielded the most results. This could be because this is the only analysis looking at methylation at age 15/17, so a child has had longer for methylation to occur. To calculate Bonferroni I needed to divide the p-value by the total number of probes tested. The number of probes measured was 450K, so Bonferroni was calculated as 0.05/450,000 and 0.05/(450,000 x 4).

Table 4.1: How many associations were detected at p-value thresholds in different EWAS models. Each model tests association between methylation and eczema. I) to vi) show the covariates in each model

		(i) Sex + SVs	(ii) Sex + SVs + two SES	(iii) Sex + SVs + two SES + mat hist	(iv) Sex + SVs + two SES + mat hist + cells	(v) Sex + SVs + two SES + mat hist + risk factors	(vi) Sex + SVs + two SES + mat hist + risk factors + cells	TOTAL
		i	ii	iii	iv	v	vi	
Cord blood, ever eczema 408 cases, 616 controls	P<1x10-8	0	0	0	0	0	0	0
	P<1x10-7	0	0	0	0	0	0	0
	P<0.05	5	5	3	2	3	2	20
Methylation at seven, ever eczema 327 cases, 697 controls	P<1x10-8	0	0	0	0	0	0	0
	P<1x10-7	0	0	0	0	0	0	0
	P<0.05	11	6	8	4	8	7	44
Methylation at seven, eczema in last 12 months 110 cases, 914 controls	P<1x10-8	0	0	0	0	0	0	0
	P<1x10-7	0	0	0	0	0	0	0
	P<0.05	2	2	2	1	2	2	11
Methylation at 15/17, eczema in last 12 months 63 cases, 961 controls	P<1x10-8	0	0	0	0	0	0	0
	P<1x10-7	0	0	0	0	0	1	1
	P<0.05	11	14	17	18	18	18	96

* Orange highlighting indicates the model I focussed on.

**Yellow highlighting shows the p-value cut off of P<0.05

Code:

P<1x10-8 = Bonferroni - Total number of tests (4 x number of tests on array)

P<1x10-7 = Bonferroni - Number of tests on array

P<0.05 = Suggestive associations

4.4.2 Is there a relationship between cord blood methylation and ever eczema?

The analysis here adjusts for 'Sex, surrogate variables, two socioeconomic status variables (social class and child ethnic background), maternal history and cell count'. The manhattan and volcano plots, 4.2 and 4.3, show that there are no CpG's with strong evidence for association with eczema, but two sites that show weak evidence for an association at P<0.05. On average, babies who go on to have eczema had more methylation at the sites cg04804139 and cg09418000 in cord blood, with coefficients of 0.007 and 0.013. This means they have 0.7% and 1.3% more methylation respectively at these sites, the highest coefficient being 0.013. This coefficient relates to the effect size of the relationship, and similarly to the other models I shall come onto. Overall, the sum of the six models, is 20 at the P<0.05 threshold. The data below on the manhattan plot shows the chromosome on the x-axis and p-value on the y-axis. When risk factors are adjusted for, the results don't attenuate with coefficients of 0.006725 and 0.01270 respectively. See table 4.2 for the results found.

Figure 4.1: Manhattan plot - Is there a relationship between cord blood methylation and ever eczema?

Each dot is a CpG site, and red lines show the cut off point over/under which sites are deemed associated. Here this is at the point P=1.1x10-7. The Y-axis has a positive and negative scale to display hyper and hypomethylation. This shows when there is more or less methylation at each CpG site.

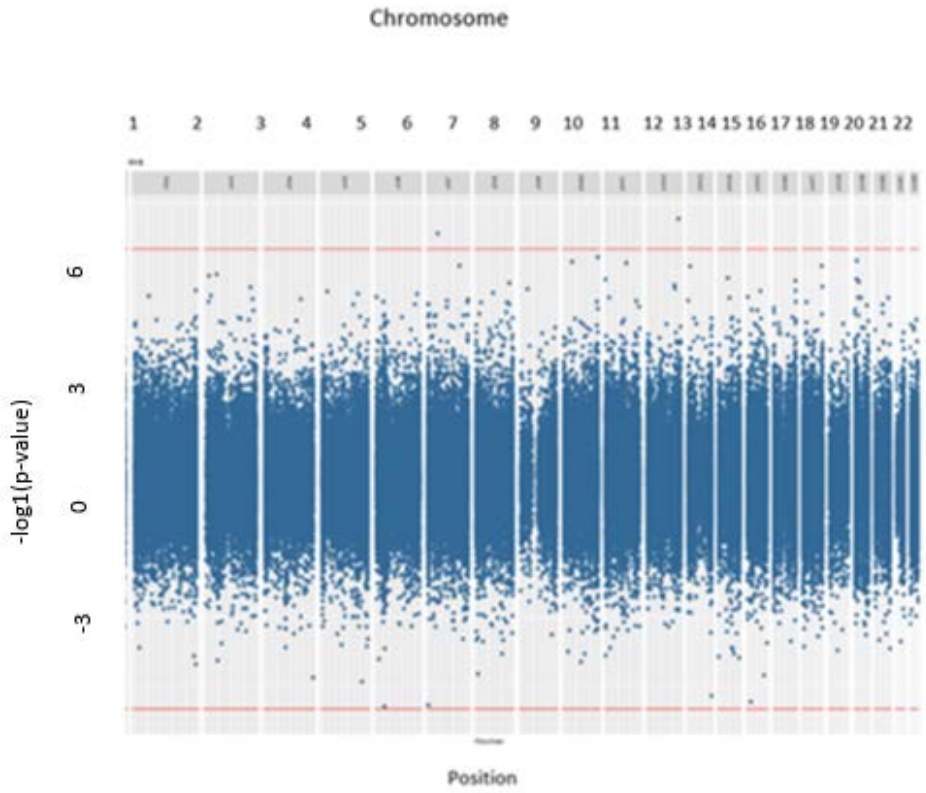
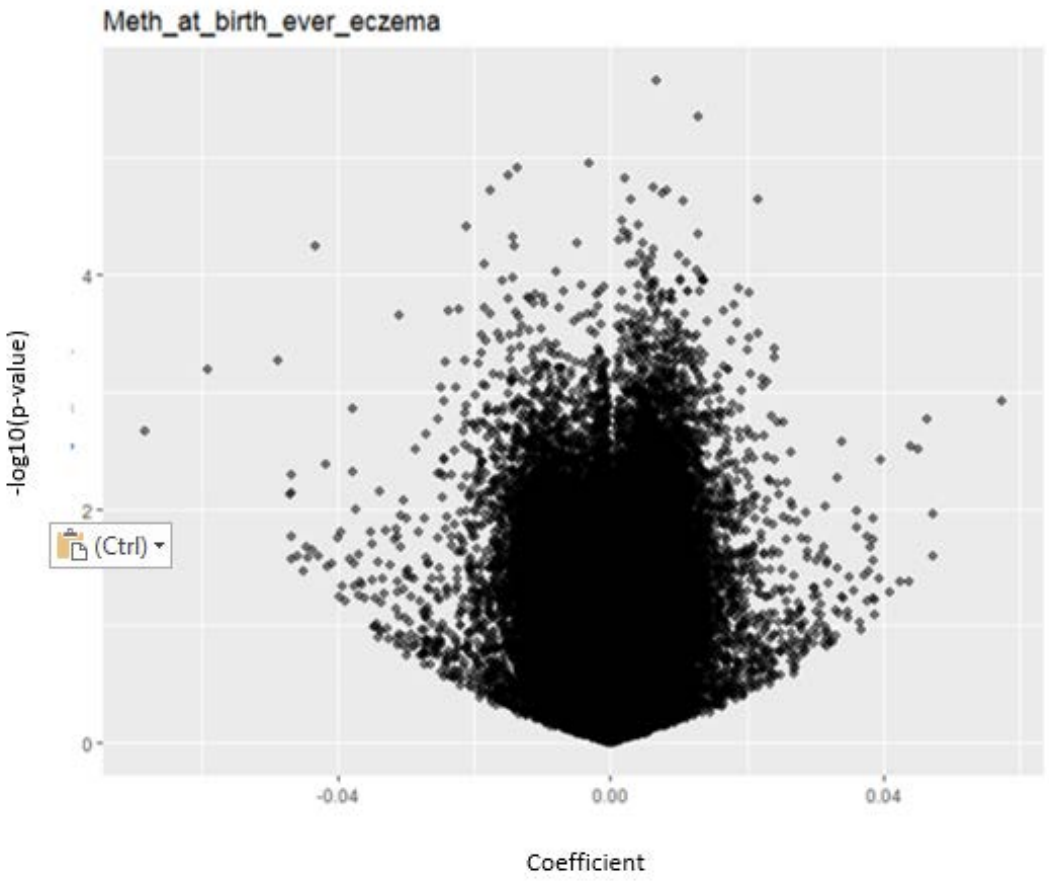


Figure 4.2: Volcano plot - Is there a relationship between cord blood methylation and ever eczema?

The X-axis shows the coefficient, and the Y-axis the –log10 p value. Each point is a coefficient from EWAS plotted against its respective –log10 p value. In this volcano plot, the coefficient is the effect size comparing eczema cases vs controls in EWAS analysis adjusting for covariates sex, social class, child ethnic background and maternal history of allergy, and surrogate variables computed using sva.



4.4.3 Is there a relationship between blood methylation at age seven and ever eczema?

There is no evidence for any strong associations in the analysis looking at methylation at seven and ever eczema at P<0.05. In figures 4.3 and 4.4 there are four suggestive associations. In this question people with eczema had a reduced amount of methylation at the site cg07166235 (-0.002) but an increased amount at cg24211994 (0.007), cg14511273 (0.008) and cg26368024 (0.002) compared to controls. The highest coefficient here is for CpG site cg14511273, which shows a large relationship.

When risk factors are adjusted for, the results attenuate with coefficients of 0.002, 0.002 and -0.0009 respectively, suggesting that the relationship between methylation and eczema may not be genuine, but biased or influenced by adjusting and not adjusting for risk factors. The sum of the six models is 44 at the P<0.05 threshold. Here we are looking at the point P=0.05.

Figure 4.3: Manhattan-plot - Is there a relationship between blood methylation at age seven and ever eczema?

On this graph, each dot is a cpg site. The red lines show the cut off point over/under which sites are deemed associated.

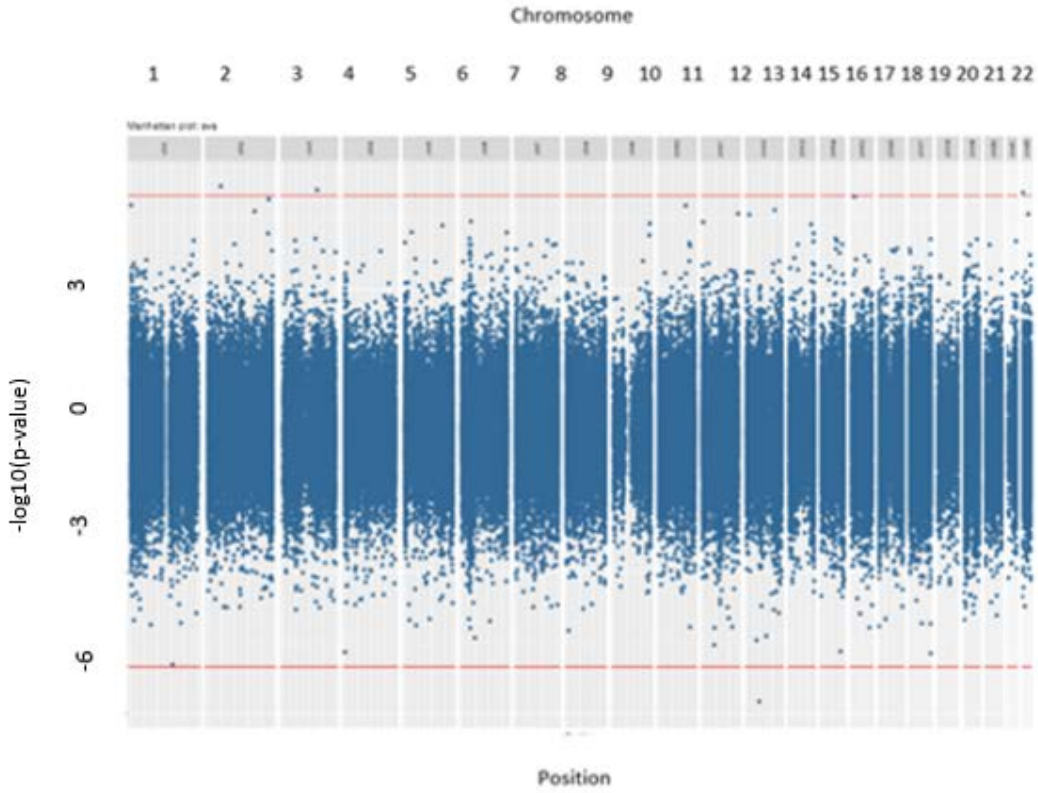
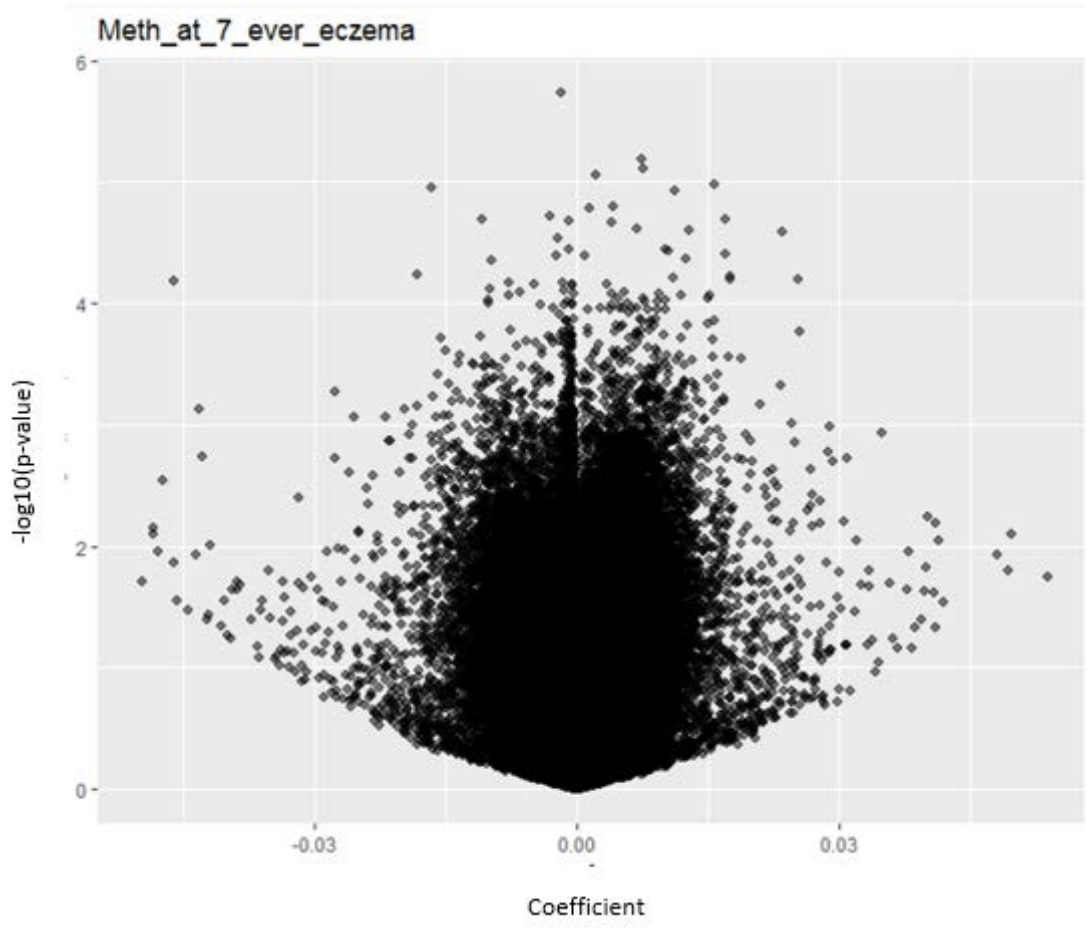


Figure 4.4: Volcano-plot - Is there a relationship between blood methylation at age seven and ever eczema?

The X-axis shows the coefficient, and the Y-axis the –log10 p value. Each point is a coefficient from EWAS plotted against its respective –log10 p value.



4.4.4 Is there a relationship between blood methylation at age seven and eczema in the last 12 months?

Here there was evidence for a weak association at P<0.05. Figures 4.7 and 4.8 show the least number of associations out of all the questions. There was just one suggestive association at P<0.05, CpG site cg23673397. The volcano plot here reaches out to the left-hand side, indicating CpG sites with more negative coefficients and hypomethylation when looking at eczema in the last 12 months at age seven. Generally cases tend to have lower levels of methylation than controls. In this

question people with eczema had a reduced amount of methylation at the site cg23673397 (-0.011). When risk factors are adjusted for, the results attenuate with coffieicints of 0.003. Here, the sum of the six models is 11 at the P<0.05 threshold.

Figure 4.5: Manhattan-plot: Is there a relationship between blood methylation at age seven and eczema in the last 12 months?

Each dot is a cpg site, and red lines show the cut off point over/under which sites are deemed associated.

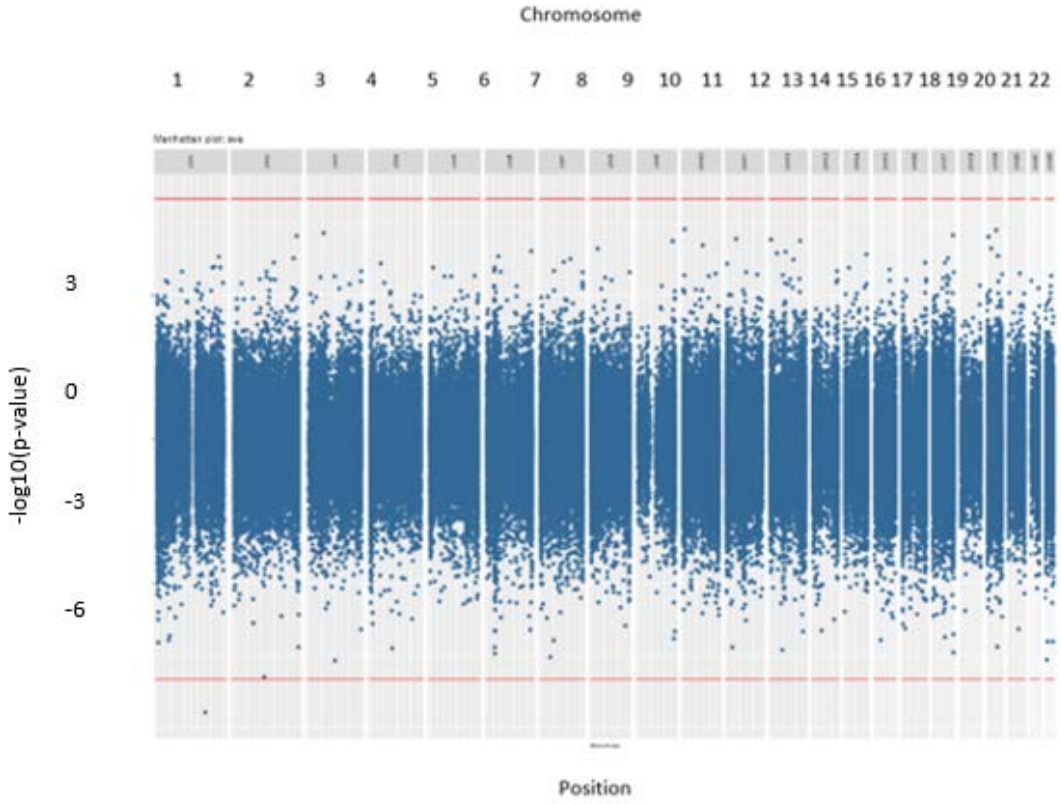
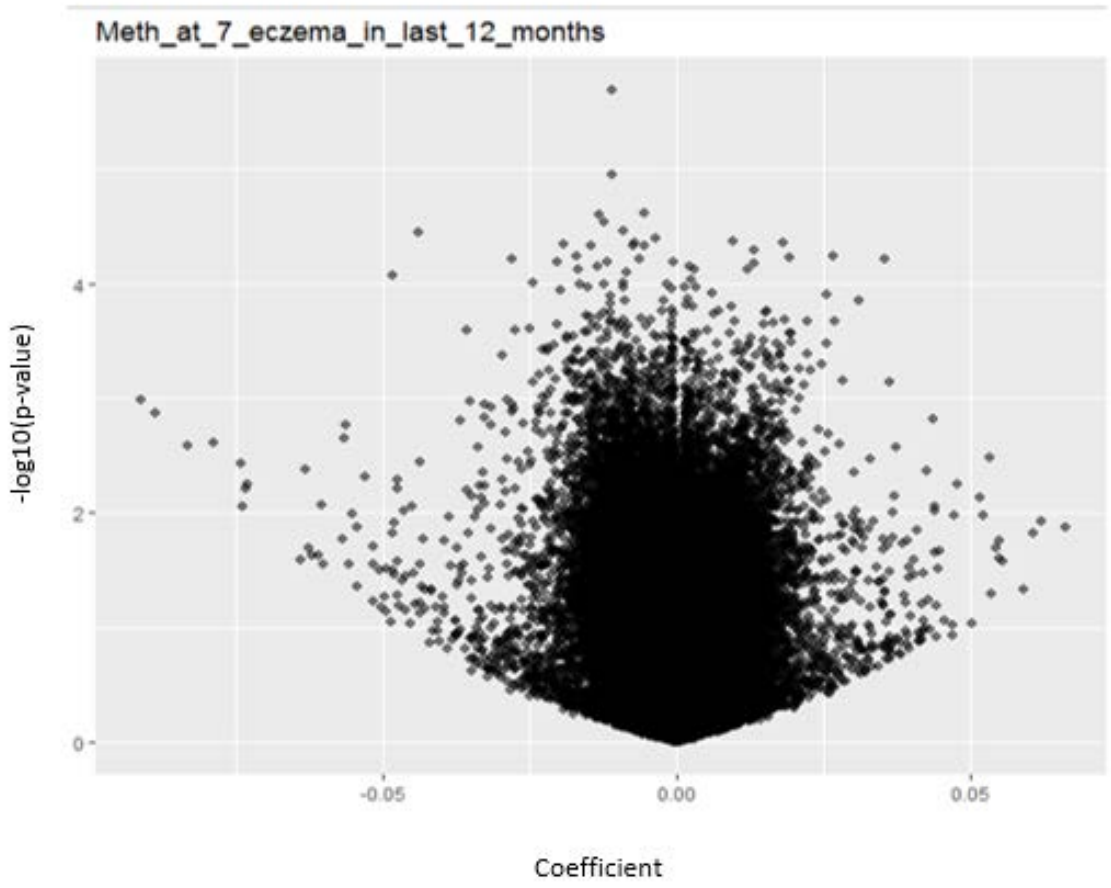


Figure 4.6: Volcano-plot: Is there a relationship between blood methylation at age seven and eczema in the last 12 months?

The X-axis shows the coefficient, and the Y-axis the –log10 p value. Each point is a coefficient from EWAS plotted against its respective –log10 p value.



4.4.5 Is there a relationship between blood methylation at age 15/17 and eczema in the last 12 months?

There were 18 suggestive CpG associations in the analysis looking at methylation at 15/17 and eczema in last 12 months at P<0.05. Figures 4.9 and 4.10 show 18 suggestive associations which is the most out of all the questions. There are three associations which fall at the point of 1x10-7, cg03220363, cg07721777 or cg26716834. There are also 15 at 1x10-6. The highest coefficient is at site CpG cg26716834 with -0.026. As you can see from table 4.1, one CpG site becomes significant when you adjust for other risk factors when looking at the p-value of P<1x10-7. This is not included in the 18 already reported. The sum of all the six models is 96 at the P<0.05 threshold.

Figure 4.7: Manhattan-plot: Is there a relationship between blood methylation at age 15/17 and eczema in the last 12 months?

Each dot is a cpg site, and red lines show the cut off point over/under which sites are deemed associated.

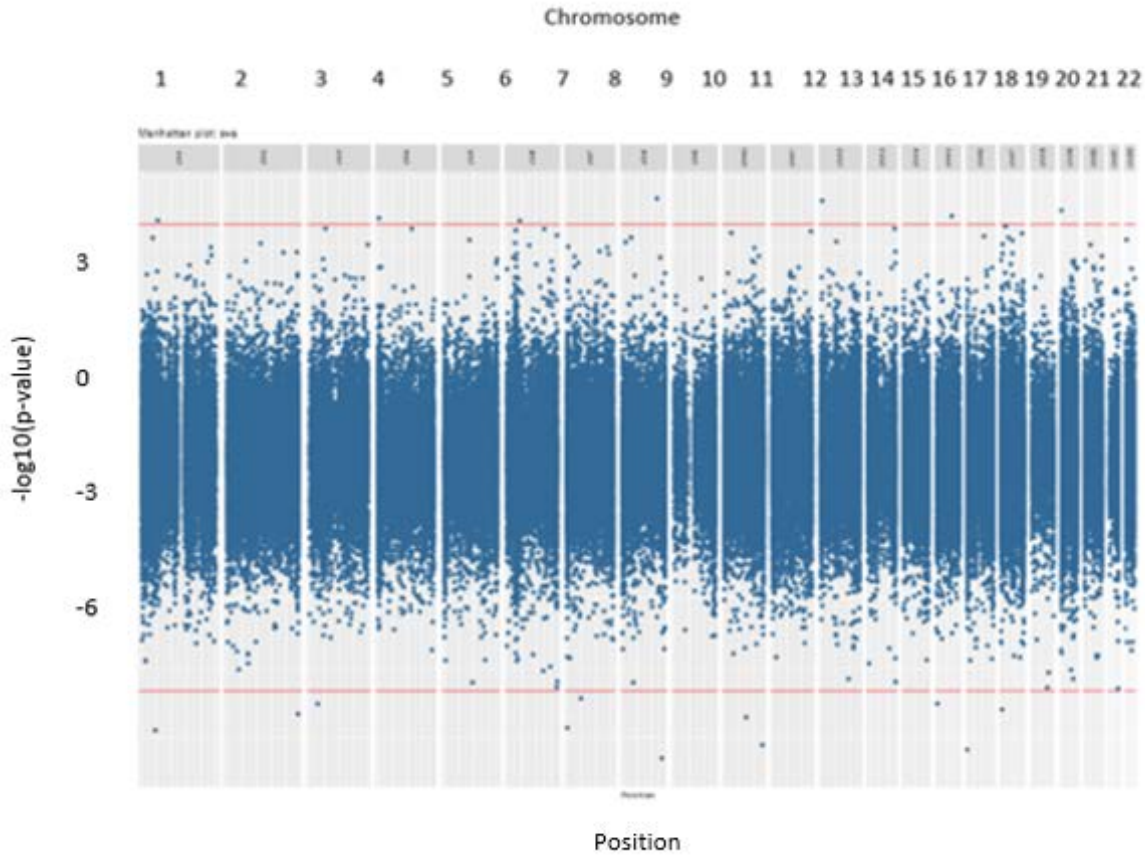
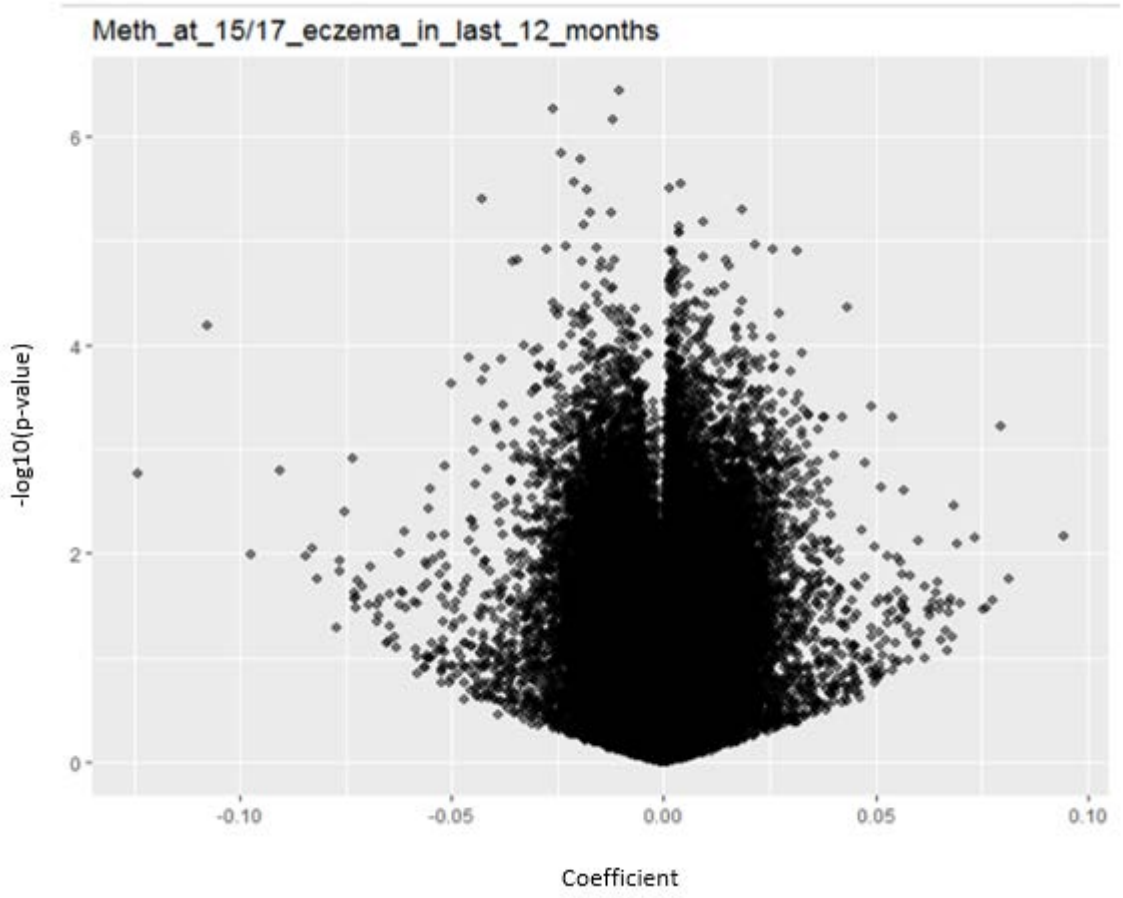


Figure 4.8: Volcano-plot: Is there a relationship between blood methylation at age 15/17 and eczema in the last 12 months?

The X-axis shows the coefficient, and the Y-axis the –log10 p value. Each point is a coefficient from EWAS plotted against its respective –log10 p value.



4.4.6 Summary table of findings

Table 4.2: The 25 suggestive associations found in the four eczema EWAS questions that were asked, looking at model (iv) and their putative functional role of nearest gene locus (table spread over two pages) at P<0.05

Question	CpG site	Chromosome	Mean (Standard deviation)	Coefficient	P-value	95% confidence intervals	Nearest gene (location of CpG site)	Gene function (95)
1	cg04804139	chr12	0.875 (0.052)	0.007	2.2x10-6	0.004, 0.01	NCOR2 (Main gene body)	Associated with certain cancers
1	cg09418000	chr7	0.914 (0.033)	0.013	4.6x10-6	0.007, 0.018	POU6F2 (Main gene body)	Tumor suppressor involved in Wilms tumor (WT) predispositio n
2	cg07166235	chr12	0.033 (0.009)	-0.002	1.8x10-6	-0.002, -0.0001	ADCY6 (gene region TSS200)	Adenylyl Cyclases (AC) are a group of enzymes
2	cg24211994	Chr2	0.871 (0.028)	0.007	6.4x10-6	0.004, 1.011	UNKNOWN	
2	cg14511273	Chr3	0.269 (0.043)	0.008	7.7x10-6	0.004, 1.011	CCDC48 (Main gene body)	Protein Coding gene
2	cg26368024	Chr22	0.043 (0.013)	0.002	8.7x10-6	0.001, 0.003	DGCR14 (gene region TSS200)	UNKNOWN
3	cg23673397	chr1	0.867 (0.032)	-0.011	2.0x10-6	-0.016, -0.006	FLJ23867;Q SOX1 (TSS1500;B ody)	Uncharacteri zed Protein
4	cg03220363	chr8	0.872 (0.024)	-0.011	3.6x10-7	-0.015, -0.007	TG;SLA (Main gene body)	Glycoprotein homodimer produced by thryroid gland (95)
4	cg26716834	chr16	0.852 (0.052)	-0.026	5.6x10-7	-0.036, -0.016	SOX8 (3'UTR)	Regulation of embryonic development (95)
4	cg07721777	chr10	0.862 (0.027)	-0.012	7.0x10-7	-0.016, -0.007	NONE	
4	cg15388975	chr1	0.835 (0.047)	-0.024	1.5x10-6	-0.034, -0.014	TMEM53 (3'UTR)	Transmembr ane Protein (95)
4	cg10415664	chr7	0.799 (0.041)	-0.020	1.7x10-6	-0.027, -0.012	MAD1L1 (Main gene body)	Component of mitotic spindle-assembly checkpoint (95)

4	cg03284839	chr10	0.756 (0.063)	-0.021	2.8x10-6	-0.030, -0.012	P4HA1 (Main gene body)	Encodes component of prolyl 4-hydroxylase (95)
4	cg04770165	chr8	0.059 (0.028)	0.004	2.8x10-6	0.002, 0.006	EIF3H (TSS200)	Component of eukaryotic translation initiation factor 3 (95)
4	cg09770904	chr12	0.019 (0.003)	0.001	3.2x10-6	0.001, 0.002	CD9 (1stExon;5' UTR)	Cell surface glycoproteins with four transmembrane domains (95)
4	cg10313065	Chr2	0.850 (0.050)	-0.018	3.2x10-6	-0.026, -0.011	NONE	
4	cg03955767	chr17	0.466 (0.083)	-0.043	4.0x10-6	-0.061, -0.025	NONE	
4	cg19653589	chr19	0.141 (0.042)	0.019	5.1x10-6	0.011, 0.026	GNG7 (5'UTR)	Guanine nucleotide-binding proteins (95)
4	cg16971668	chr3	0.099 (0.036)	-0.012	5.4x10-6	-0.018, -0.007	EOMES (TSS1500)	Belongs to the TBR1 sub-family of T-box genes (95)
4	cg03799387	chr15	0.141 (0.042)	-0.017	5.4x10-6	-0.025, -0.01	NONE	
4	cg01182386	chr15	0.085 (0.020)	0.009	6.7x10-6	0.005, 0.013	MPI (TSS200)	MPI (Mannose Phosphate Isomerase) (95)
4	cg00645664	chr17	0.827 (0.045)	-0.019	7.0x10-6	-0.027, -0.011	ABCA13 (Body)	Diseases include Ovarian Serous Carcinoma
4	cg11864499	chr14	0.035 (0.010)	0.004	7.4x10-6	0.002, 0.005	C4orf10;MFSD10 (TSS1500;TSS200;TSS1500)	Member of major facilitator superfamily
4	cg21849289	chr1	0.030 (0.012)	0.004	8.4x10-6	0.002, 0.005	C1orf123 (TSS1500)	Protein Coding gene
4	cg07274194	chr6	0.048 (0.009)	0.004	8.5x10-6	0.002, 0.005	DLK2 (5'UTR; TSS200)	Calcium ion binding

* Yellow highlighting indicates a low p-value at P<1x10-5

Those with the smallest p-values at 1x10-6 have relatively large coefficients at cg03220363, cg26716834 and cg07721777 with -0.011, -0.026 and -0.012 respecively. The highest coefficient is at site CpG cg26716834 with -0.026. The smallest coefficients generally come at the beginning of each question. For example, the first site of question 1 (with the smallest p-value) is cg04804139; the second question has the smallest coefficient at cg07166235, etc. For all of the 25 CpG sites, I noted the gene they were present in and then looked this up to see if the genes had anything to do with eczema and atopy. However there were no genes which fulfilled this criteria.

4.4.7 Comparison of p-values in all analyses

Here I compare p-values between questions. For example, if there was strong evidence of an association between a CpG site and eczema, would this continue to other questions/models? I carried out the following comparisons:

- Ever eczema (either in cord blood or at age 7) and methylation at age 7 and whether a person has had eczema in the last 12 months.

- Ever eczema (either in cord blood or at age 7) and methylation at age 15/17 and whether a person has had eczema in the last 12 months.

- Eczema when methylation is measured at 15/17 and whether methylation measured earlier at age 7.

Table 4.3: The comparison of different p-values and coefficients between all four questions, looking at model (iv) (table spread over three pages)

Question	CpG site	Ever eczema, cord blood			Ever eczema, age 7			Eczema in last 12 months, age 7			Eczema in last 12 months, age 15/17		
		Coeff.	P-value	95% confidence intervals	Coeff.	P-value	95% confidence intervals	Coeff.	P-value	95% confidence intervals	Coeff.	P-value	95% confidence intervals
1	cg04804139	0.007	2.18x10-6	0.004, 0.01	-0.001	0.329	-0.002, 0.001	-0.001	0.348	-0.004, 0.0013	-0.005	0.009	-0.009, -0.001
1	cg09418000	0.013	4.55x10-6	0.007, 0.018	-0.001	0.600	-0.005, 0.03	-0.004	0.084	-0.009, -0.0006	-0.001	0.813	-0.008, -0.007
2	cg07166235	-0.0007	0.106	-1.52x10-3, 0.0001	-0.002	01.8x10-6	-0.002, -0.0001	-0.001	0.003	-0.002, -0.0005	0.001	0.061	-6.259x10-5 0.003
2	cg24211994	0.002	0.312	-0.002, 0.006	0.007	4.6x10-6	0.004, 1.011	0.002	0.301	0.002, 0.0006	0.009	0.019	-0.001, 0.016
2	cg14511273	0.002	0.433	-0.002, 0.005	0.008	7.7x10-6	0.004, 1.011	0.003	0.239	-0.002, 0.007	0.005	0.065	-0.003, -0.011
2	cg26368024	-0.001	0.140	-0.002, 0.0002	0.002	8.7x10-6	0.001, 0.003	0.002	0.010	0.0004, 0.003	0.001	0.553	-0.001, 0.003
3	cg23673397	0.003	0.271	-0.002, 0.0007	-0.003	0.076	-0.007, 0.0003	-0.011	2.0x10-6	-0.016, -0.006	-0.002	0.552	-0.009, 0.005

4	cg03220363	0.002	0.221	-0.001, 0.047	4.7x10-5	0.960	-0.002, 0.002	-0.001	0.311	-0.004, 0.0012	-0.011	3.6x10-7	-0.015, -0.007
4	cg26716834	-0.001	0.814	-0.0008, 0.006	0.001	0.547	-0.003, 0.006	0.001	0.688	-0.005, 0.0008	-0.026	5.6x10-7	-0.036, -0.016
4	cg07721777	0.002	0.156	-0.0009, 0.005	0.0006	0.643	-0.002, 0.003	-0.003	0.079	-0.006, 0.0003	-0.012	7.0x10-7	-0.016, -0.007
4	cg15388975	0.002	0.571	-0.005, 0.009	-0.005	0.063	-0.01, 0.0003	-0.007	0.050	-0.014, 6.33x10-6	-0.024	1.5x10-6	-0.034, -0.014
4	cg10415664	0.004	0.142	-0.001, 0.01	0.003	0.239	-0.002, 0.007	0.002	0.441	-0.003, 0.006	-0.020	1.7x10-6	0.027, -0.012
4	cg03284839	0.002	0.497	-0.004, 0.007	0.002	0.403	-0.002, 0.005	0.001	0.547	-0.003, 0.0006	-0.021	2.8x10-6	-0.030, -0.012
4	cg04770165	0.001	0.305	-0.0008, 0.003	0.0008	0.649	-0.003, 0.004	0.0003	0.700	-0.001, 0.0002	0.004	2.8x10-6	0.002, 0.006
4	cg09770904	0.0004	0.078	-3.941x10-5, 0.0007	-0.0002	0.379	-0.0005, 0.0002	-0.0001	0.616	-0.0006, 0.0003	0.001	3.2x10-6	0.001, 0.002
4	cg10313065	0.0005	0.843	-0.005, 0.006	-0.0002	0.729	-0.001, 0.001	-0.001	0.538	-0.006, 0.0003	-0.018	3.2x10-6	-0.026, -0.011
4	cg03955767	-0.004	0.432	-0.002, 0.007	-0.008	0.084	-0.018, 0.001	-0.011	0.088	-0.023, 0.002	-0.043	4.0x10-6	-0.061,

													-0.025
4	cg19653589	0.001	0.743	-0.004, 0.006	3.1x10-5	0.99	-0.004, 0.004	0.0004	0.863	-0.004, 0.0005	0.019	5.1x10-6	0.011, 0.026
4	cg16971668	-0.0004	0.728	-0.003, 0.002	0.001	0.511	-0.002, 0.003	0.001	0.545	-0.002, 0.004	-0.012	5.4x10-6	-0.018, -0.007
4	cg03799387	0.004	0.182	-0.002, 0.009	0.003	0.225	-0.002, 0.007	0.0002	0.958	-0.005, 0.006	-0.017	5.4x10-6	-0.025, -0.01
4	cg01182386	-0.002	0.209	-0.004, 0.001	0.0004	0.700	-0.002, 0.002	0.003	0.035	0.0002, 0.005	0.009	6.7x10-6	0.005, 0.013
4	cg00645664	0.004	0.142	-0.001, 0.01	0.001	0.377	-0.002, 0.005	0.002	0.441	-0.003, 0.006	-0.019	7.0x10-6	-0.027, -0.011
4	cg11864499	-0.0009	0.0910	-0.002, 0.0001	0.0002	0.686	-0.0007, 0.001	0.006	0.289	-0.005, 0.002	0.004	7.4x10-6	0.002, 0.005
4	cg21849289	-0.001	0.110	-0.002, 0.002	0.0006	0.154	-0.0002, 0.001	0.001	0.085	-0.0001, 0.002	0.004	8.4x10-6	0.002, 0.005
4	cg07274194	-6.88x10-5	0.908	-0.001, 0.001	0.0003	0.561	-0.0007, 0.001	0.0005	0.441	-0.0008, 0.002	0.004	8.5x10-6	0.002, 0.005

* Question 1 – Looking at the relationship between cord blood methylation and ever eczema.

Question 2 - Looking at the relationship between blood methylation at age seven and ever eczema.

Question 3 - Looking at the relationship between blood methylation at age seven and eczema in the last 12 months.

Question 4 – Looking at the relationship between blood methylation at age 15/17 and eczema in the last 12 months.

** Pale green highlighting indicates cross tabulation. For example, when looking at a CpG site found as a result of the analysis of one question, the green highlighting shows the data for that question (ie, cg04804139 and 'ever eczema, cord blood')

*** Yellow highlighting indicates a low p-value at $P<0.05$

Overall, I can therefore conclude that five CpG sites (highlighted in yellow) showed similarly small p-values at P<0.05 when comparing between questions. There were two sites, cg07166235 and cg26368024, which had small p-values when looking at ever eczema (either in cord blood or at age 7) and also when looking at question 3, methylation at age 7 and whether a person has had eczema in the last 12 months. There were another two, cg04804139 and cg24211994 which again had small p-values at ever eczema and at question 4, methylation at age 15/17 and whether a person had had eczema in the last 12 months. Lastly only 1 of the 18 CpG sites associated with eczema when methylation is measured at 15/17, cg01182386, showed any evidence for association when methylation measured earlier at age 7 (and given multiple testing may be false positive). All of these associations also showed relatively large coefficients, with CpG site cg01182386 showing the largest at 0.003. This continuation between models indicates a more robust set of results. If there is a CpG site with significant p-values in one model and similarly small p-values in another, this would strengthen the relationship between methylation and those certain CpG sites.

4.5 GWAS RESULTS

4.5.1 mQTL's

Using the mQTL database (96) I was able to lookup the top 25 associated CpGs to see if any had mQTLs (methylation quantitative trait loci), and whether any of these SNPs were associated with eczema. mQTLs occur when a SNP has an influence over whether a CpG site is methylated or unmethylated. I looked only at the top mQTL defined as the one with the smallest p-value for each CpG:

Table 4.4: Comparing whether 25 suggestive CpG associations are linked to any mQTLs (table spread over two pages)

CpG site	mQTL	Chromosome	Distance between mQTL and CpG	A1	Beta value (between CpG and SNP)	P-value	95% confidence intervals
cg04804139	rs7838880	8	102238092	T	0.629	9.47e-08	0.004, 0.01
cg09418000	NONE						
cg07166235	NONE						
cg24211994	NONE						
cg14511273	rs185502486	4	90267468	T	0.897	4.07e-09	0.004, 1.011
cg26368024	rs8077781	17	53745908	G	0.517	8.16e-08	0.001, 0.003
cg23673397	NONE						
cg03220363	NONE						
cg26716834	NONE						
cg07721777	rs10052472	5	17445751	C	-0.311	2.79e-08	-0.016, -0.007
cg15388975	NONE						
cg10415664	NONE						
cg03284839	NONE						
cg04770165	rs73071166	12	97138230	C	-0.770	7.74e-08	0.002, 0.006
cg09770904	rs13106548	4	94937020	A	-0.374	8.24e-08	0.001, 0.002
cg10313065	NONE						

cg03955767	rs79278020	4	52775165	T	-1.014	8.96e-08	-0.061, -0.025
cg19653589	rs147992452	19	18914	T	0.667	1.15e-08	0.011, 0.026
cg16971668	NONE						
cg03799387	NONE						
cg01182386	rs12051677	17	3917634	A	-0.232	1.45e-09	0.005, 0.013
cg00645664	NONE						
cg11864499	rs79817130	11	125570948	G	0.491	8.38e-08	0.002, 0.005
cg21849289	NONE						
cg07274194	NONE						

Overall there were nine CpG sites which had mQTLs and these are shown in table 4.4. CpG site cg14511273 had the largest beta value when compared to its mQTL at 0.897, and also had a small p-value. I looked to see whether any of the mQTLs linked to DNA methylation also linked to eczema in the GWAS carried out by Paternoster *et al.* (66). However there were none.

4.6 CONCLUSION

Overall there is a lack of any strong association between DNA methylation and eczema. There are 25 sites which could be 'suggestive' at P<0.05. For cord blood there are two associations, for methylation at age 7 and ever eczema there are 4, and for methylation at age 7 and eczema in the last 12 months there is 1. There are 18 which come from the analysis looking at methylation at age 15/17 and whether a person has had eczema in the past 12 months.

Table 4.5: The replication of associated CpG sites between two models (questions)

CpG SITES	Mean (Standard deviation)	BASELINE MODEL				MODEL COMPARED TO			
		MODEL	Coefficient	P-value	Confidence intervals	MODEL	Coefficient	P-value	Confidence intervals
cg07166235	0.033 (0.009)	Ever eczema (cord or age 7) (Q2)	-0.002	1.8x10-6	-0.002, -0.0001	Methylation at 7, eczema in last 12 months (Q3)	-0.001	0.003	-0.002, -0.0005
cg26368024	0.043 (0.013)	Ever eczema (cord or age 7) (Q2)	0.002	8.7x10-6	0.001, 0.003	Methylation at 7, eczema in last 12 months (Q3)	0.002	0.010	0.0004, 0.003
cg04804139	0.875 (0.052)	Ever eczema (cord or age 7) (Q1)	0.007	2.18x10-6	0.004, 0.01	Methylation at 15/17, eczema in last 12 months (Q4)	-0.005	0.009	-0.009, -0.001
cg24211994	0.871 (0.028)	Ever eczema (cord or age 7) (Q2)	0.007	6.4x10-6	0.004, 1.011	Methylation at 15/17, eczema in last 12 months (Q4)	0.009	0.019	-0.001, 0.016
cg01182386	0.085 (0.020)	Methylation at 15/17, eczema in last 12 months (Q4)	0.009	6.7x10-6	0.005, 0.013	Methylation at 7, eczema in last 12 months (Q3)	0.003	0.035	0.0002, 0.005

All had small p-values but CpG sites cg24211994 and cg01182386 had the largest coefficients at 0.009. In this study I am looking at the effect of methylation on the development of eczema. However, it can also be hypothesised that the eczema itself was causing the differences in

methylation observed. This could be the case for questions 2, 3 and 4, which all involve methylation at age seven or age 15/17. The weak associations here could potentially be *caused* by eczema influencing the methylome. However, one question manages to address this issue by sidestepping the process of eczema affecting methylation. This is question 1. Here methylation must have come first because it was measured in cord blood, which means at birth. It would not have been possible for the child to develop eczema in utero, which means methylation is more likely to be causal. So, I know that the two CpG sites associated with question 1 and model (iv) must have been more methylated compared to controls before the development of eczema.

For this study, the data has already been collected by the Children of the 90's consortium, which is an advantage, and questions cover a wide range of topics which makes confounding easier. There is evidence that many CpG sites show an association at different time points (three associations to be precise) which strengthens our claim that DNA methylation at these CpG sites have an association with eczema long term. However, as already explained, the associations whilst in existence are weak. As covered in the 'Discussion' chapter later, perhaps the next step is to increase sample size and/or replication. Disadvantages include that the questions in ALSPAC are self-reported, power is low and batch effects can be a problem. EWAS also does not cover all of the CpG sites, only approximately 2%.

CHAPTER 5. RISK FACTORS

5.1 INTRODUCTION

I have looked at the relationship between DNA methylation and eczema in children. Next, I wanted to look at methylation as a mediator between environmental risk factors and eczema. The three environmental risk factors I chose to look at were: 1) smoking, 2) animal exposure (cats and dogs) and 3) breastfeeding. The reasons for choosing these three risk factors are explained in following sub sections. They have all been shown to be associated with eczema. I either carried out EWAS's myself or referred to them in the literature to see if there was any relationship between the risk factors and DNA methylation. Once the EWAS's had been carried out, I took the CpG sites identified as suggestively associated at $P<0.05$ and looked to see whether the p-values were similarly small when looking at different questions, ie. methylation at another time point or a different definition of eczema. I also looked at coefficient size. I then either looked at or created EWAS which were to do with the three risk factors mentioned above and looked to see whether any CpG's were associated with both eczema and the risk factor, thus implying a mediating role of DNA methylation between the two.

5.2 ENVIRONMENTAL RISK FACTORS

The data that will be used in this study comes from the Avon Longitudinal Study of Parents and Children (ALSPAC) (83) (84) and the Accessible Resource for Integrated Epigenomics Studies (ARIES) (85). Questionnaires were completed as part of ALSPAC , and DNA methylation measured. The Illumina Infinium 450K arrays were used to measure the methylome of the children. It has 450K probes, which each measure the percentage of methylation at a CpG site. Each probe targets a CpG site and measures the proportion of copies of DNA the sample that are methylated or unmethylated. In this study measurements will be taken at birth, from cord blood, and at ages seven and 15/17. They will then be compared to smoking, animal exposure and breastfeeding.

5.2.1 Association between smoking and eczema

In table 5.1 I carried out a logistic regression to see if there was a relationship between eczema and whether a mother smoked during pregnancy or brought their child up in a smoky environment, but the p-value was 0.234. In ALSPAC there were 2,511 cases in the category '0 (Never smoked)', (136 in ARIES) 1210 in the category '1 (Smoked before pregnancy, but stopped)' (54 in ARIES) and 615 in the category '2 (Yes)' (505 in ARIES). The rather large p-value would indicate that there is not much evidence to reject the null hypothesis (that there is no relationship between eczema and smoking). Table 5.2 does the same for a smoky environment. Here there were 3,640 cases in the '0 (No)' category (443 in ARIES) and 696 in the '1 (Yes)' category (62 in ARIES). Here however, the p-value is much smaller at 0.002. This would indicate that there is a relationship between eczema and a smoky environment. There is more eczema in a smoky environment but not with smoking during pregnancy.

Table 5.1: Tabulation of whether a person has ever had eczema and whether their mother smoked during pregnancy

	Mother smoked during pregnancy							
	ALSPAC				**ARIES**			
Eczema ever in child reported at age 15/17	**0 (Never smoked)**	**1 (Smoked before pregnancy, but stopped)**	**2 (Yes)**	**Total**	**0 (Never smoked)**	**1 (Smoked before pregnancy, but stopped)**	**2 (Yes)**	**Total**
0 (No)	2,766	1,373	752	4,891	305	131	51	487
1 (Yes)	2,511	1,210	615	4,336	315	136	54	505
Total	5,277	2,583	1,367	9,227	620	267	105	992

Table 5.2: Tabulation of whether a person has ever had eczema and whether they spent their early years in a smoky environment

	A smoky environment					
	ALSPAC			**ARIES**		
Eczema ever in child reported at age 15/17	**0 (No)**	**1 (Yes)**	**Total**	**0 (No)**	**1 (Yes)**	**Total**
0 (No)	3,984	907	4,891	421	66	487
1 (Yes)	3,640	696	4,336	443	62	505
Total	7,624	1,603	9,227	864	128	992

Logistic regressions were carried out to look at the relationship between smoking during pregnancy and eczema, and a smoky environment and eczema, both with and without adjustment for the other three risk factors, cat exposure, dog exposure and breastfeeding.

Table 5.3: Logistic regressions of smoking against eczema, with and without adjustment for other risk factors (cat exposure, dog exposure and breastfeeding)

	ECZEMA					
	NO ADJUSTMENT			**ADJUSTMENT FOR OTHER RISK FACTORS (CAT EXPOSURE, DOG EXPOSURE AND BREASTFEEDING)**		
	Odds Ratio	P-value	Confidence Intervals	Odds Ratio	P-value	Confidence Intervals
SMOKED DURING PREGNANCY						
(Smoked but stopped)	0.972	0.552	0.884, 1.068	0.951	0.366	0.854-1.060
(Yes)	0.902	0.090	0.800, 1.016	1.056	0.472	0.911-1.224
SMOKY ENVIRONMENT						
(Yes)	0.840	0.002	0.754, 0.937	0.918	0.201	0.806-1.046

In the unadjusted models, the odds ratios were 0.972 when comparing "1- Smoked before pregnancy, but stopped" with the baseline, "0 – Never smoked" (95% confidence intervals 0.88, 1.07), a reduction of 3%. This means people are 3% less likely to have eczema if they smoked but stopped rather than if they never smoked. The OR when comparing "2 – Yes" with "0 – Never smoked" was 0.902 (95% CI 0.80, 1.02), a decrease of 10%. However, the confidence intervals both cross one, which would indicate little evidence for any effect. The p-values are 0.552 and 0.090 respectively. That is the logistic regressions when looking at an unadjusted model. When adjusting for risk factors, the OR go from 0.972 to 0.951 ("1- Smoked before pregnancy, but stopped" with the baseline "0 – Never smoked") and 0.902 to 1.056 ("2 – Yes" with "0 – Never smoked"). The important difference seems to be that although the result of the OR for comparing smoked versus didn't smoke during pregnancy changes to OR>1 after adjustment, there are still wide confidence intervals. Another logistic regression was carried out looking into whether a smoky environment during a child's early years was associated with the later development of eczema. The OR was 0.84 (95% CI 0.75-0.94). There is a 16% decreased risk of developing eczema associated with being brought up in a smoky environment than if you are brought up in a cleaner environment. This may

be to do with the 'Hygiene hypothesis', which explains that a child growing up in an environment more full of allergens will be used to them and will not have such an extreme reaction when they come into contact with them. It could also be that people growing up in a dirtier environment might not be so conscientious about going to the doctors if they notice a rash and thus letting the eczema get noticed. The p value was 0.002, demonstrating good evidence for this association. When adjusting for the other risk factors the OR rises to 0.918, suggesting that there is less of an effect between eczema and a smoky environment as the result is attenuating towards 1. The fact that adjusting for risk factors (confounders) causes the result to approach 1 means that there may not be a true relationship between a smoky environment and eczema, it may be being biased. This indicates that inhaling smoke during early childhood does not necessarily cause problems that lead on to eczema.

EWAS

There were 25 CpG sites weakly associated with eczema which I previously identified in Chapter 4. I wanted to compare what the coefficient/p-values were in a smoking EWAS with those same CpG sites. An EWAS for smoking has already been carried out, looking at current smokers, previous smokers and never smokers (97). I extracted the 25 eczema-associated CpG results from the smoking paper and examined if any showed any evidence for association with smoking. This showed FDR significant findings for CpG sites associated with smoking. Only CpG sites with association p-values $P<1\times10\text{-}7$ were available from the supplementary material.

Methylation at CpG's associated with smoking – Look-up in eczema CpG's

An EWAS using ALSPAC and ARIES data was not conducted because there already exists an EWAS in the literature. Joubert *et al.* (98) meta-analysed the effect of smoking in pregnancy on offspring methylation in the 13 cohorts in the Pregnancy And Childhood Epigenetics (PACE) consortium. 6000 CpG sites were differentially methylated when looking at maternal smoking during pregnancy and newborn methylation. This is the biggest EWAS investigating maternal smoking with respect to offspring methylation, and earlier (and smaller) studies show similar effects. Joehanes *et al.* (97) conducted a meta-analysis of those who currently smoke against those who have never smoked. DNA methylation at 2623 CpG sites were identified as being related to smoking, which were annotated to 1405 genes, at the Bonferroni threshold of $P<1x10\text{-}7$. Joehanes is the largest meta

analysis of own smoking on own methylation. In this section, I shall refer to the data set from Joehanes *et al.* (97) as, being a meta-analysis, this has the most amount of subjects out of each paper, and has 2623 signfiicant DNA methylation sites.

I took the top associations between methylation at CpG sites and smoking as defined by their coefficient and p-values in Joehanes *et al.* (97). Out of 2623, I took the top sites where P<1x10-8, and merged this CpG list with the eczema EWAS dataset (question 1, ever eczema, cord blood and question 2, ever eczema, methylation at age 7) and then subset the results to show which CpG sites had small p-values at a size of P<0.05 for both 'smoking and methylation' and 'eczema and methylation', indicating DNA methylation acting as a mediator. There were two sets of evidence, looking at cord blood and looking at methylation at age 7. This is important because the cord blood can give us information about smoking during pregnancy, and the second study could look at growing up in a smoky environment. Of these I got a total of 50 CpG sites with small p-values in both when looking at question 1, cord blood methylation and ever eczema. From this I decided to take the top nine results as these were P<0.01. I also got a total of 56 CpG sites with small p-values in both when looking at question 2, methylation at age 7 and ever eczema, and for these I decided to take the top 14 sites. Therefore the total number of small p-values, indicating a mediating role of methylation were 9 and 14 respectively (although there are more, 50/56 in fact, if we relaxed the Bonferroni threshold). We could therefore suggest that there may be an association between smoking and eczema, at 23 CpG sites. This is really interesting because it suggests that smoking during pregnancy does not have as much of an effect on eczema development as growing up in smoky environment. For utilising the eczema variable, I used questions 1 and 2, which is methylation in cord blood and methylation at age seven and whether a child has ever had eczema by the age of 15/17. This was the most suitable model, because you are looking at methylation mid way through a child's life, but comparing it to whether a child has ever had eczema at some point during childhood. The model I used was model (iv) which adjusts for sex, surrogate variables, two socioeconomic status variables (social class and child ethnic background), maternal history and cell counts, with and without risk factors (and in model vi) cat exposure, dog exposure and breastfeeding). Coefficient values were also taken into account as well as p-values, with the CpG site with the largest coefficient being cg01294327 at -0.040. It has this larger value when looking at ever eczema and methylation at age seven.

Methylation at CpG's associated with eczema – Look-up in smoking EWAS

I looked up 2623 CpG associations in the Joehanes *et al*. paper (97) for smoking at Bonferroni threshold of P<1×10-7. I then measured this against eczema when methylation was measured at age 7. I could then take the smoking data, which had been extracted and look at the top associations for eczema. The important thing here is whether there is any evidence for the same CpGs being associated with both smoking and eczema. When looking in this other direction, all I can conclude after referring to the paper 'Epigenetic Signatures of Cigarette Smoking' by Joehanes *et al.* was that one CpG site, cg19653589, has an association with both eczema and smoking.

Table 5.4: All CpG sites identified as having an association with both smoking and eczema EWAS 'ever eczema' (on next page)

Risk factor	Question	CpG site	Chromosome	Gene	SMOKING		ECZEMA EVER CORD BLOOD No risk factors			EVER ECZEMA AGE SEVEN No risk factors		
					Coeffient	P-value	Coeffient	P-value	95% confidence intervals	Coeffient	P-value	95% confidence intervals
SMOKING	1	cg27332104	N/A	N/A	0.012	5.1x10-22	0.005	0.001	0.002, 0.008	0.003	0.34	-0.003, 0.01
	1	cg13855261	17	HS3ST3B1	-0.008	4.3x10-17	-0.008	0.002	-0.013, 0.003	-0.001	0.810	-0.005, 0.004
	1	cg09206294	15	MAPKBP1	0.005	2.5x10-11	0.007	0.002	0.002, 0.011	-0.001	0.794	-0.005, 0.004
	1	cg13689560	19	C19orf36	-0.008	2.8x10-12	-0.011	0.003	-0.017, -0.004	-0.003	0.028	-0.005, -0.0003
	1	cg15187398	19	MOBKL2A	-0.029	1.9x10-13	-0.008	0.006	-0.014, -0.002	0.003	0.363	-0.003, 0.008
	1	cg19635644	13	UBAC2	-0.006	2.7x10-12	0.005	0.007	0.001, 0.009	0.002	0.498	-0.004, 0.008
	1	cg06434490	6	N/A	-0.009	9.6x10-13	0.009	0.008	0.002, 0.016	-0.001	0.587	-0.006, 0.004
	1	cg26707709	2	SNED1	0.03	5.0x10-14	-0.004	0.008	-0.007, -0.001	0.0006	0.838	-0.005, 0.006
	1	cg09570614	2	N/A	0.013	1.2x10-22	0.008	0.009	0.002, 0.013	0.006	0.028	0.0007, 0.012
	2	cg11405655	1	N/A	-0.014	1.7x10-12	-0.003	0.175	-0.008, 0.001	-0.007	0.001	-0.010, -0.003
	2	cg01993576	6	SLC29A1	-0.008	1.1x10-15	-0.002	0.341	-0.007, 0.003	0.007	0.001	0.003, 0.011
	2	cg18451588	7	PRR15	-0.019	9.3x10-18	-0.001	0.374	-0.002, 0.001	-0.003	0.002	-0.004, -0.0009
	2	cg11445634	15	N/A	-0.013	1.8x10-16	0.0006	0.858	-0.005, 0.007	0.01	0.003	0.003, 0.016

	2	cg22957360	10	N/A	-0.007	1.1x10-10	-0.001	0.781	-0.008, 0.006	0.008	0.003	0.003, 0.013
	2	cg25503804	11	F2	-0.01	8.7x10-14	-0.001	0.76	-0.008, 0.006	0.007	0.003	0.002, 0.012
	2	cg05593667	6	N/A	-0.012	5.5x10-20	0.0004	0.833	-0.003, 0.004	-0.006	0.004	-0.011, -0.002
	2	cg20722088	12	DUSP6	-0.007	3.3x10-12	0.001	0.751	-0.003, 0.004	0.004	0.005	0.001, 0.007
	2	cg01561259	17	RPTOR	0.01	5.4x10-11	-0.001	0.849	-0.01, 0.009	0.012	0.005	0.004, 0.02
	2	cg20152539	17	HS3ST3B1	-0.015	4.1x10-13	-0.005	0.200	-0.013, 0.003	-0.011	0.007	-0.02, -0.003
	2	cg00501876	3	CSRNP1	-0.022	2.5x10-15	-0.001	0.700	-0.005, 0.003	-0.005	0.007	-0.009, -0.001
	2	cg01294327	19	LINGO3	-0.040	2.0x10-13	-0.003	0.292	-0.01, 0.003	-0.012	0.009	-0.022, -0.003
	2	cg22871253	6	EZR	-0.011	1.3x10-15	-0.001	0.723	-0.01, 0.007	-0.009	0.009	-0.016, -0.002
	2	cg01692968	9	N/A	-0.033	7.4x10-16	-0.002	0.573	-0.01, 0.006	0.01	0.01	0.003, 0.018

* Yellow highlighting indicates a low p-value at P<0.05

The CpG site with the largest coefficient is cg01294327 at -0.040. It has a similarly large value when compared to ever eczema and methylation at age seven. None of the genes identified as containing the CpG site had an association with eczema or atopy.

Overlap of CpGs

Table 5.5: Overlap of CpG sites associated with smoking and eczema, between cord blood and methylation age seven

CpG sites	Chromosome	Gene	EVER ECZEMA, CORD BLOOD			EVER ECZEMA AGE SEVEN		
			Coeffient	P-value	95% confidence intervals	Coeffient	P-value	95% confidence intervals
cg09570614	2	N/A	0.008	0.009	1.922, 0.013	0.006	0.03	0.0007, 0.011
cg13689560	19	C19orf36	-0.011	0.003	-0.02, -0.004	-0.003	0.03	-0.005, -0.0003
cg26728709	2	SNED1	-0.004	0.008	-0.007, -0.001	0.005	0.04	0.002, 0.01

As can be seen from table 5.5, there are three CpG sites, cg09570614, cg13689560 and cg26728709, which overlap when looking at cord blood methylation and whether a child has ever had eczema, and methylation at age seven and whether a child has ever had eczema. The first category (cord blood methylation and whether a child has ever had eczema) provides information on smoking during pregnancy, whereas the second category (methylation at age seven and whether a child has ever had eczema) is more interesting when looking at a child growing up in a smoky environment. The fact that these CpG sites appear in both analyses strengthens the associations because they are appearing in both. The CpG site cg09570614 has the largest coefficients as well as low p-values. The Appendix V has the full table of all 50/56 results (see table A7).

5.2.2 Association between animal exposure and eczema

In table 5.6 I carried out a logistic regression in ALSPAC looking at the relationship between eczema and cat exposure in ALSPAC, the p-value was 0.288. Table 5.6 does the same for dog exposure in ALSPAC. Here, however, the p-value is much smaller at 0.001 (odds ratio 0.866, 95% confidence intervals 0.795-0.943) This would indicate that there is some relationship between eczema and dog exposure in that dog exposure causes less eczema. Overall the literature claims that cats can make eczema worse, whereas dogs can help protect against it.

Table 5.6: Tabulation of whether a person has ever had eczema and whether they spent their early years exposed to cats

	Exposure to cats					
	ALSPAC			**ARIES**		
Eczema ever in child reported at age 15/17	**0 (No)**	**1 (Yes)**	**Total**	**0 (No)**	**1 (Yes)**	**Total**
0 (No)	3,098	1,793	4,891	77	410	487
1 (Yes)	2,701	1,635	4,336	82	423	505
Total	5,799	3,428	9,227	159	833	992

Table 5.7: Tabulation of whether a person has ever had eczema and whether they spent their early years exposed to dogs

	Exposure to dogs					
	ALSPAC			**ARIES**		
Eczema ever in child reported at age 15/17	**0 (No)**	**1 (Yes)**	**Total**	**0 (No)**	**1 (Yes)**	**Total**
0 (No)	3,036	1,855	4,891	252	235	487
1 (Yes)	2,837	1,499	4,336	255	250	505
Total	5,873	3,354	9,227	507	485	992

Table 5.8: Logistic regressions of risk factors against eczema, with and without other risk factors (smoking during pregnancy, a smoky environment and breastfeeding)

	ECZEMA					
	WITHOUT RISK FACTORS			WITH RISK FACTORS		
	Odds Ratio	P-value	Confidence Intervals	Odds Ratio	P-value	Confidence Intervals
CAT EXPOSURE	1.047	0.288	0.962-1.139	1.031	0.531	0.936-1.137
DOG EXPOSURE	0.866	0.001	0.795-0.943	0.902	0.043	0.817-0.997

For cats, there is no real evidence in this analysis that there is a relationship between cat exposure and eczema. For dog exposure, the odds ratio was 0.87. This suggests an association between dog exposure and the development of eczema. Generally, eczema will be less likely if the child is around dogs, and more likely if they aren't. The p value is 0.001, which indicates an association between dog exposure and eczema, with 95% confidence intervals 0.79, 0.94. After adjusting for other risk factors the OR increases to 0.90 and the p-value to 0.043. Though there is some attenuation, there is still evidence for a negative association after adjustment.

EWAS

Cat exposure

I investigated the association for cat exposure in 1,024 individuals with available data in ARIES (adjusting for sex, surrogate variables, two socioeconomic status variables (social class and child ethnic background), maternal history and cell counts) between DNA methylation at age 7 and whether a child has ever had eczema. No CpG sites were associated with cat exposure below the Bonferroni threshold of $P<1x10-7$. Five CpG sites showed evidence of association with cat exposure when using a relaxed p-value threshold of $P<0.05$. Results are shown in the manhattan and volcano plots.

Table 5.9: Results of EWAS of cat exposure without and with risk factors

CpG sites **UNADJUSTED**	Chromosome	Coefficient	P-value	95% confidence intervals	Gene	Gene function (taken from genecards (95))
cg10824810	16	-0.022	2.723x10-6	-0.031, -0.013	IRX5	Diseases associated with IRX5 include Hamamy Syndrome and Griscelli Syndrome, Type 3.
cg14643686	12	-0.01	3.506x10-6	-0.014, -0.005	UNKNOWN	
cg09114441	19	-0.002	3.68x10-6	-0.003, -0.001	ZNF546	a Protein Coding gene. Among its related pathways are Gene Expression
cg21758133	3	0.003	6.817x10-6	0.002, 0.005	FOXP1	Diseases associated with FOXP1 include Mental Retardation With Language Impairment And With Or Without Autistic Features
cg13790288	2	0.003	8.902x10-6	0.002, 0.005	CD28	The protein encoded by this gene is essential for T-cell proliferation and survival, cytokine production, and T-helper type-2 development
ADJUSTED						
cg26491624	16	0.008	2.867x10-5	0.004, 0.01	ABCC1	The protein encoded by this gene is a member of the superfamily of ATP-binding cassette (ABC) transporters
cg07264682	10	-0.017	6.959x10-6	-0.024 - 0.009	UNKNOWN	

Out of these, the CpG site with the largest coefficient is cg10824810 at -0.022.

Figure 5.1: Manhattan -plot - Is there a relationship between methylation at age 15/17 and cat exposure in model (iv)?

Each dot is a cpg site, and red lines show the cut off point over/under which sites are deemed associated (either hypo/hypermethylated).

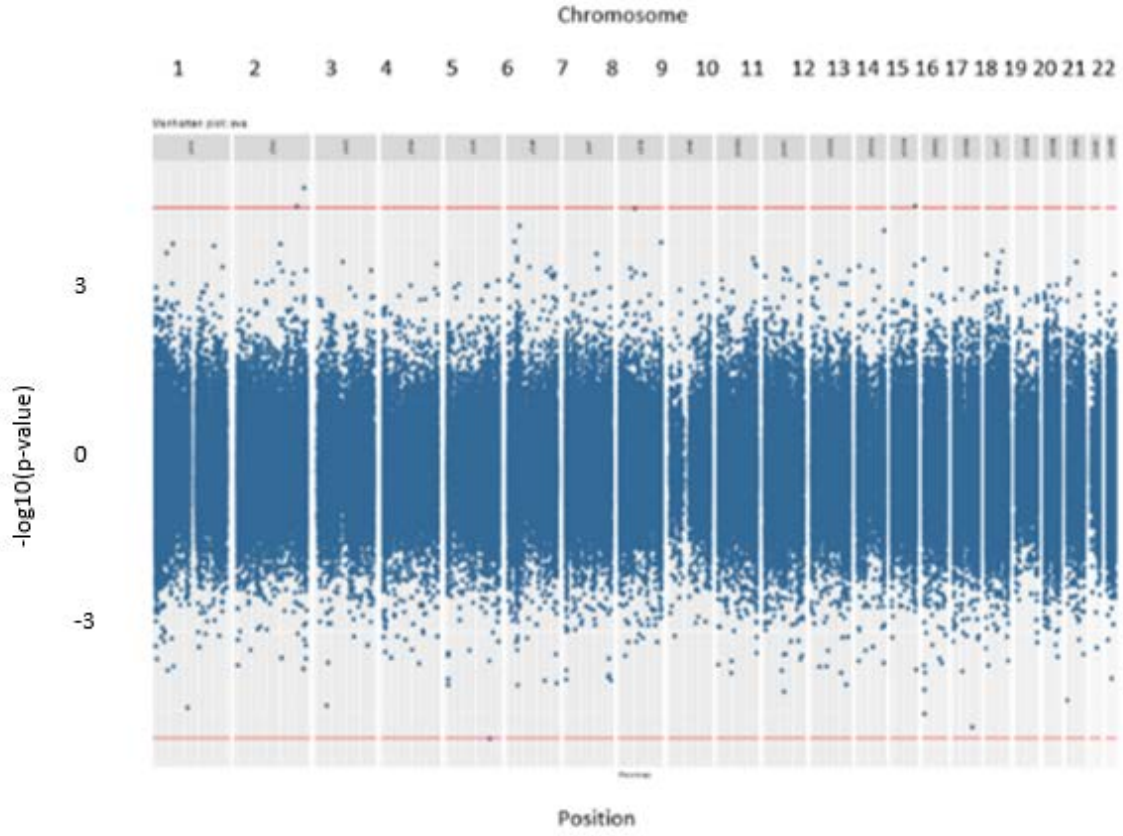
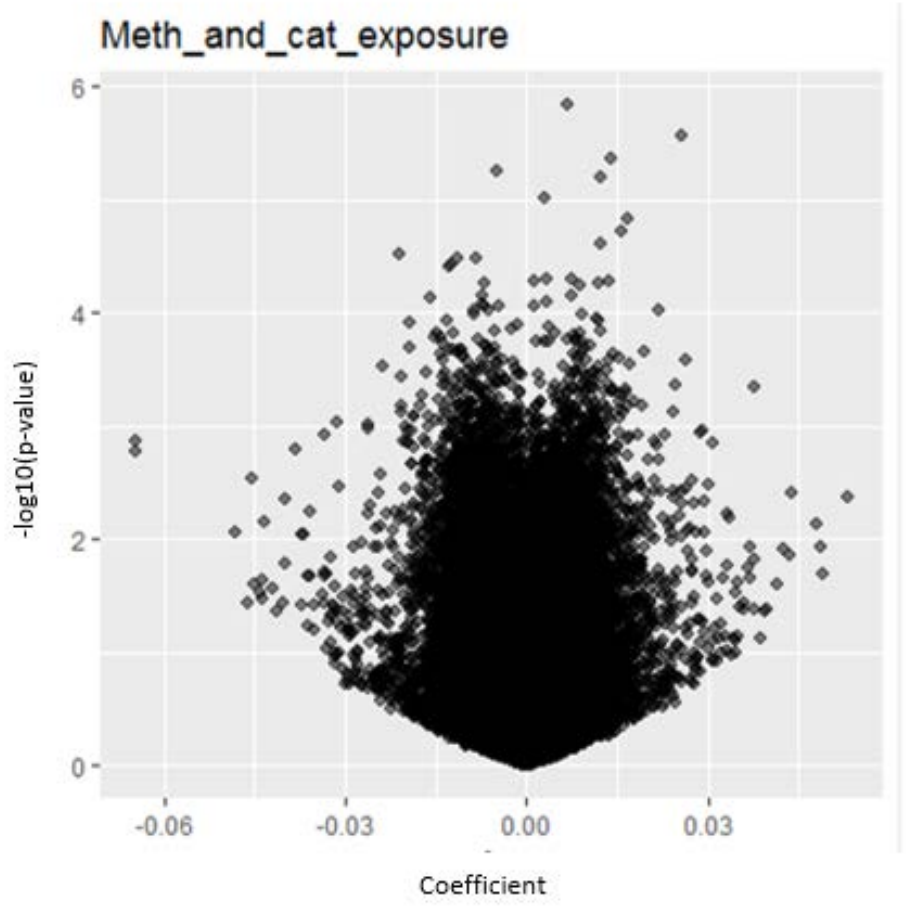


Figure 5.2: Volcano-plot - Is there a relationship between methylation at age 15/17 and cat exposure in model (iv)?

The X-axis shows the coefficient, and the Y-axis the –log10 p value. Each point is a coefficient from EWAS plotted against its respective –log10 p value.



Dog exposure

I investigated the association of dog exposure in 1,024 individuals with available data in ARIES (adjusting for sex, surrogate variables, two socioeconomic status variables (social class and child ethnic background), maternal history and cell counts) between DNA methylation at age 7 and whether a child has ever had eczema. There is evidence for four suggestive associations in the analysis looking at methylation at 15/17 and dog exposure at $P<0.05$. Again, there were no strong associations highlighted in this analysis defined by a P-value of $P<1x10-7$.

Table 5.10: Results of EWAS of dog exposure without and with risk factors

CpG site UNADJUSTED	Chromosome	Coefficient	P-value	95% confidence intervals	Gene	Gene function (taken from genecards (95))
cg27252766	2	0.005	7.576x10-6	0.003, 0.007	UNKNOWN	
cg03276401	17	0.015	9.994x10-6	0.008, 0.022	TEX2	TEX2 (Testis Expressed 2) is a Protein Coding gene
ADJUSTED						
cg03639185	4	-0.1	7.115x10-7	-0.134, -0.059	UNKNOWN	
cg25703541	22	0.007	8.234x10-6	0.004, 0.01	LOC391322	(D-Dopachrome Tautomerase-Like) is a Protein Coding gene

The largest coefficient is of CpG site cg03276401, with 0.015, showing the largest relationship.

Figure 5.3 Manhattan -plot - Is there a relationship between methylation at age 15/17 and dog exposure in model (iv)?

Each dot is a cpg site, and red lines show the cut off point over/under which sites are deemed associated.

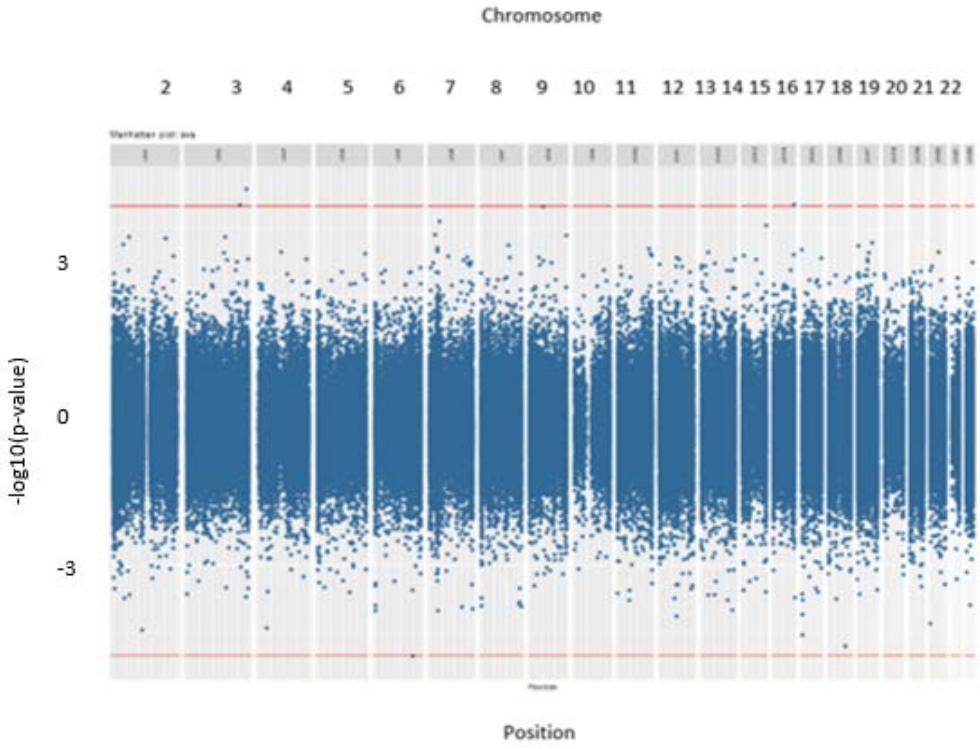
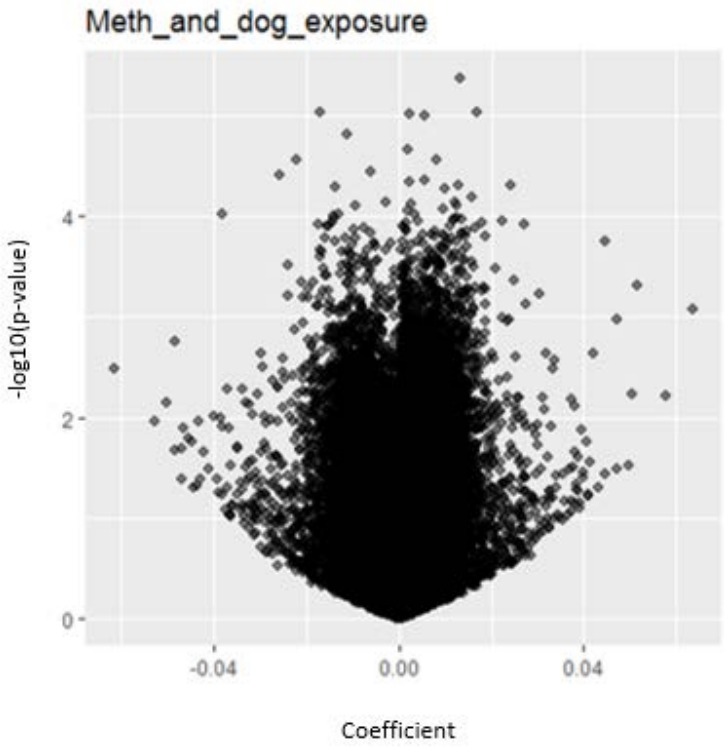


Figure 5.4: Volcano-plot - Is there a relationship between methylation at age 15/17 and dog exposure in model (iv)?

The X-axis shows the coefficient, and the Y-axis the –log10 p value. Each point is a coefficient from EWAS plotted against its respective –log10 p value.



Methylation at CpG's associated with animal exposure – Look-up in eczema EWAS

According to table 5.10, for cat exposure there are five CpG significant sites when no risk factors are adjusted for and two when risk factors are adjusted for (note, this relates to other risk factors apart from the one I am focussing on). The risk factors include smoking during pregnancy, a smoky environment and breastfeeding, as well as dog exposure in the cat exposure analysis, and cat exposure in the dog analysis. For dog exposure, there are two significant CpG sites when risk factors are not adjusted for and two when they are. Obviously, all of these have p-values of x10-6 or less. The coefficients range from -0.002 to 0.005 for cats, and -0.017 to 0.015 for dogs. However overall there are no CpG sites associated with both cat/dog exposure and eczema.

Table 5.11: All CpG sites identified as having an association with animal exposure, tabulated against the EWAS results of associations between methylation and eczema

		CpG	Chromosome	Gene	CATS No risk factors			DOGS No risk factors			EVER ECZEMA No risk factors		
					Coef	P-value	95% confidence intervals	Coef	P-value	95% confidence intervals	Coef	P-value	95% confidence intervals
CATS	NO RISK FACTORS (MODEL 4)	cg10824810	16	IRX5	-0.002	2.723x10-6	-0.031, -0.013	-0.001	0.814	-0.01, 0.009	0.0002	0.962	-0.009, 0.0094
		cg14643686	12	N/A	-0.0001	3.506x10-6	-0.014, -0.006	0.0001	0.962	-0.004, 0.005	0.0009	0.685	-0.003, 0.005
		cg09114441	19	ZNF546	-0.002	3.680x10-6	-0.003, -0.001	-0.0007	0.247	-0.002, 0.005	-0.0004	0.433	-0.001, 0.0006
		cg21758133	3	FOXP1	0.003	6.817x10-6	0.002, 0.005	0.0008	0.318	0.0008, 0.002	0.0004	0.619	-0.001, 0.002
		cg13790288	2	CD28	0.0003	8.902x10-6	0.002, 0.005	0.0004	0.624	-0.001, 0.002	-0.001	0.095	-0.003, 0.0002
DOGS	NO RISK FACTORS (MODEL 4)	cg27252766	2	N/A	-4.15x10-5	0.965	-0.02, 0.002	0.005	7.576x10-6	0.003, 0.007	0.0005	0.593	-0.001, 0.002
		cg03276401	17	TEX2	-0.007	0.020	-0.013, 0.002	0.015	9.994x10-6	0.01, 0.02	-0.002	0.585	-0.008, 0.005

* Pale green squares highlight cat/dog CpG sites up against cat/dog analysis, without adjustment for risk factors

** Yellow highlighting indicates a low p-value at P<0.05

The largest coefficients are at CpG site cg03276401 in cat exposure, dog exposure and eczema, showing a large effect size. None of the genes had anything to do with eczema.

Methylation at CpG's associated with eczema – Look-up in animal exposure EWAS

EWAS showed children who had been exposed to cats had an increased amount of methylation at the sites cg21758133 (0.3%), and cg13790288 (0.3%), and a decreased amount of methylation at site cg10824810 (-2%), cg14643686 (-1%) and cg09114441 (-0.2%). When risk factors were adjusted for, this changed to two suggestive associations. Children who had been exposed to cats had an increased amount of methylation at the site cg26491624 (0.8%) but a decreased amount at cg07264682 (-0.2%). When looking at the EWAS of dog exposure, children who had spent time with dogs had an increased amount of methylation at the site cg27252766 (0.5%) and cg03276401 (1.5%). However, after adjusting for risk factors, there were two suggestive associations. Children had an increased amount of methylation at cg25703541 (0.7%) and a decreased amount of methylation at the site cg03639185 (-10%). None of the CpG sites identified were associated with eczema, inflammation or rash in the gene lookups.

When looking at the relationship between animal exposure and eczema, the CpG associated p-values which were suggestive at $P<0.05$ when looking at the unadjusted model for cats attenuated slightly when looking at the adjusted models. When looking at the second table on animal exposure and eczema, it can be seen that there are no associations between the CpG sites and eczema, as all of the p-values are quite large and the confidence intervals quite often straddle 0.

Table 5.12: All CpG sites identified as having an association with eczema, tabulated up against the EWAS results of associations between methylation and animal exposure (table spread over three pages)

CpG		Chromosome	Gene	ECZEMA No risk factors			CATS No risk factors			DOGS No risk factors		
				Coefficient	P-value	95% confidence intervals	Coefficient	P-value	95% confidence intervals	Coefficient	P-value	95% confidence intervals
1) A Cord blood, ever eczema	cg04804139	12	NCOR2	-0.001	0.329	-0.002, 0.001	-0.001	0.279	-0.003, 0.0008	-0.002	0.12	-0.004, 0.0004
	cg09418000	7	POU6F2	-0.001	0.600	-0.005, 0.03	0.004	0.048	0.00003, 0.008	0.0006	0.79	-0.004, 0.005
1) B Methylation at age seven, ever eczema	cg07166235	12	ADCY6	-0.002	1.8x10-6	-0.002, -0.0001	-0.0003	0.414	-0.001, 0.0004	0.00002	0.95	-0.001, 0.0001
	cg24211994	2	N/A	0.007	6.4x10-6	0.004, 1.011	-0.0002	0.912	-0.003, 0.003	-0.002	0.27	-0.005, 0.002
	cg14511273	3	CCDC48	0.008	7.7x10-6	0.004, 1.011	-0.0008	0.618	-0.004, 0.002	-0.00003	0.99	-0.004, 0.004
	cg26368024	22	DGCR14	0.002	8.7x10-6	0.001, 0.003	0.0001	0.812	-0.000, 0.001	0.0005	0.266	-0.0004/ 0.002
2) A Methylation at age seven, eczema in	cg23673397	1	FLJ23867	-0.003	0.076	-0.007, 0.0003	-0.001	0.69	-0.004, 0.003	0.002	0.351	-0.006, 0.002

last 12 months												
2) B Methylation at age 15/17, eczema in last 12 months	cg03220363	8	TG;SLA	4.7x10-5	0.960	-0.002, 0.002	0.00005	0.955	-0.002, 0.002	0.002	0.070	-0.0001, 0.004
	cg26716834	16	SOX8	0.001	0.547	-0.003, 0.006	-0.003	0.195	-0.008, 0.002	-0.003	0.245	-0.008, 0.002
	cg07721777	10	N/A	0.0006	0.643	-0.002, 0.003	-0.001	0.687	-0.003, 0.002	-0.002	0.22	-0.004, 0.001
	cg15388975	1	TMEM53	-0.005	0.063	-0.01, 0.0003	0.00005	0.985	-0.005, 0.005	-0.006	0.029	-0.012, -0.001
	cg10415664	7	MAD1L1	0.003	0.239	-0.002, 0.007	-0.0001	0.952	-0.005, 0.004	0.001	0.524	-0.002, 0.005
	cg03284839	10	P4HA1	0.002	0.403	-0.002, 0.005	-0.0003	0.873	-0.004, 0.003	-0.0002	0.924	-0.004, 0.004
	cg04770165	8	EIF3H	0.0008	0.649	-0.003, 0.004	0.0003	0.585	-0.001, 0.001	-0.0001	0.871	-0.001, 0.001
	cg09770904	12	CD9	-0.0002	0.379	-0.0005, 0.0002	0.0002	0.267	-0.0002, 0.0006	0.00008	0.676	-0.0003, 0.0005
	cg10313065	2	N/A	-0.0002	0.729	-0.001, 0.001	0.0016	0.38	-0.002, 0.005	0.001	0.533	-0.003, 0.005
	cg03955767	17	N/A	-0.008	0.084	-0.018, 0.001	0.004	0.379	-0.005, 0.013	0.001	0.808	-0.009, 0.011
	cg19653589	19	GNG7	-3.054x10-5	0.99	-0.004, 0.004	-0.002	0.238	-0.006, 0.0014	-0.002	0.336	-0.006, 0.002
	cg16971668	3	EOMES	0.001	0.511	-0.002, 0.003	0.001	0.352	-0.001, 0.004	0.0007	0.612	-0.002, 0.003
	cg03799387	15	N/A	0.003	0.225	-0.002, 0.007	0.002	0.264	-0.002, 0.007	0.0006	0.803	-0.004, 0.005

	cg01182386	15	MPI	0.0004	0.700	-0.002, 0.002	0.0004	0.665	-0.002, 0.002	0.001	0.324	-0.001, 0.003
	cg00645664	17	ABCA13	0.001	0.377	-0.002, 0.005	-0.0005	0.750	-0.004, 0.003	0.001	0.524	-0.002, 0.005
	cg11864499	14	C4orf10	0.0002	0.686	-0.0007, 0.001	0.0007	0.102	-0.0001, 0.001	0.0003	0.473	-0.001, 0.001
	cg21849289	1	C1orf123	0.0006	0.154	-0.0002, 0.001	-0.00002	0.969	-0.0008, 0.0008	0.0002	0.565	-0.001, 0.001
	cg07274194	6	DLK2	0.0003	0.561	-0.0007, 0.001	0.0005	0.318	-0.0005, 0.0014	0.0006	0.272	-0.005, 0.002

* Yellow highlighting indicates a low p-value at P<0.05

The largest coefficient is 0.004 for cats at CpG site cg09418000 and cg03955767 and -0.006 for dogs at CpG site cg15388975. As earlier in the document, no genes found here had an association with anything to do with eczema or atopy.

5.2.3 Association between breastfeeding and eczema

In table 5.13, the tabulation shows that nearly 3000 mothers always breastfed (split roughly between eczema and non-eczema). Next approximately 2000 never breastfed. A mixture of breast and bottle fed had the lowest amount of people with approximately 1000 mothers.

Table 5.13: Tabulation of whether a person has ever had eczema and whether they were breast or bottle fed

	Breastfeeding							
	ALSPAC				**ARIES**			
Eczema ever in child reported at age 15/17	**0 (Never breastfed)**	**1 (Breast and bottle fed)**	**2 (Always breastfed)**	**Total**	**0 (Never breastfed)**	**1 (Breast and bottle fed)**	**2 (Always breastfed)**	**Total**
0 (No)	1,217	550	1,550	3,317	73	64	218	355
1 (Yes)	894	504	1,565	2,963	50	72	248	370
Total	2,111	1,054	3,115	6,280	123	136	466	725

Table 5.14: Logistic regressions of risk factors against eczema, with and without other risk factors

	ECZEMA					
	WITHOUT RISK FACTORS			**WITH RISK FACTORS**		
	Odds Ratio	**P-value**	**Confidence Intervals**	**Odds Ratio**	**P-value**	**Confidence Intervals**
BREASTFEEDING						
(Breast and bottle)	1.250	0.003	1.078	1.172	0.092	0.975-1.410
(Breastfed)	1.377	<0.001	1.232-1.540	1.283	0.001	1.104-1.491

A logistic regression was carried out to investigate the effect of maternal breastfeeding on childhood eczema. The odds ratios were 1.25 when comparing “1 - Breast and bottle fed” with “0 – Never

breastfed" (95% confidence intervals = 1.08, 1.45; p-value = 0.003), and 1.38 when comparing "2 – Always breastfed" with "0 – Never breastfed" (95% CI = 1.23, 1.54; p-value = <0.001). Neither of these confidence intervals overlap one. There was good evidence for a positive association between breastfeeding and eczema, which is the opposite to what one would expect. There was some evidence of attenuation when adjusting for risk factors, however some evidence for these positive associations remained (p=0.092 and 0.001).

EWAS

The same analysis was carried out as in section 5.2.1 and 5.2.2. For the breastfeeding data, I analysed CpG sites previously reported to be associated with breastfeeding. Here I am using breastfeeding associated CpGs identified by Hartwig *et al.* (99). In this paper there were seven CpG sites listed as having a relationship with breastfeeding. For utilising the eczema variable, I used question 2, which is methylation at age seven and whether a child has ever had eczema by the age of 15/17. The model I used was model (iv) which adjusts for sex, surrogate variables, two socioeconomic status variables (social class and child ethnic background), maternal history and cell counts, with and without risk factors (smoking during pregnancy, a smoky environment, cat exposure and dog exposure).

Methylation at CpG's associated with breastfeeding – Look-up in eczema CpG's

Here I took the top associated CpG sites for breastfeeding (Hartwig *et al.* (99)) and looked them up in the eczema database. For each CpG site results are shown for methylation at age seven and whether a child has ever had eczema.

Table 5.15: CpG sites identified as having an association with breastfeeding, tabulated up against the EWAS results of associations between methylation and eczema

	CpG	Chromosome	Gene	Time-point	BREASTFEEDING		ECZEMA EVER		
							No risk factors		
					Coefficient	P-value	Coefficient	P-value	95% confidence intervals
BREASTFEEDING	cg11414913	1	N/A	Age 7	-3.19	5.2x10-8	-0.002	0.546	-0.013, 0.002
	cg00234095	17	N/A		-1.74	4.9x10-7	0.002	0.490	-0.004, 0.001
	cg04722177	19	N/A		-2.90	2.7x10-6	-0.005	0.191	-0.013, 0.002
	cg03945777	7	PTPRN2		-0.84	3.2x10-6	-0.001	0.623	-0.01, 0.005
	cg17052885	17	RPTOR		1.79	4.9x10-6	-0.0003	0.871	-0.005, 0.004
	cg05800082	6	DST		1.05	5.8x10-6	-0.001	0.372	-0.004, 0.001
	cg24134845	10	HPSE2		0.23	3.3x10-5	-0.0001	0.756	-0.0009, 0.0007

Out of the four genes that were associated with the CpG sites, none of these had links or an relationship with eczema or any other type of atopic illness.

Methylation at CpG's associated with eczema – Look-up in breastfeeding EWAS

The breastfeeding EWAS paper (99) only reports results for the seven CpG sites identified as being associated with breastfeeding. Unfortunately, I was unable to locate the full data set to look up the eczema significant CpG's in the breastfeeding data (99). None of the identified CpG sites associated with eczema were also strongly associated with breastfeeding. However there may be some that are more weakly associated that I could not test. The largest coefficient was -0.005.

5.4 CONCLUSION

In this chapter, I moved on to whether DNA methylation is a mediator by looking at risk factors and eczema. Three environmental risk factors were looked at: smoking, animal exposure (namely, cats and dogs) and breastfeeding.

5.4.1 Smoking

In ALSPAC there was some evidence for an association between smoking and childhood eczema. A smoky environment was expected to show a protective association, but this attenuated after adjustment for other risk factors, suggesting that there was no real association between a smoky environment and eczema. This is different to most of the literature, which claims that smoking during pregnancy or a smoky environment can cause eczema or make it worse. Our study may be underpowered and thus not giving a similar outcome.

I found a study by Joehanes *et al.* (97) with 2623 CpG sites associated with smoking. First of all I trimmed these sites to only those where P<1x10-8. Of these, I merged the CpG sites from the Joehanes paper with my eczema ever variables (question 1, cord blood and ever eczema and question 2, methylation at age 7 and ever eczema) and then subsetted these so I would only have data which had a p-value of P<0.05 in both the "smoking and methylation" data and "eczema and methylation" data. Of these I got a total of 50 CpG sites with small p-values in both, of which I decided to take the top nine results in the first study as these were P<0.01, and 56 CpG sites, of which I decided to take the top 14 results, in the second. This was then reversed in the other direction, taking CpG's which were important in eczema to see if they were also significant in an EWAS of smoking. This way I could see if eczema and risk factors had associations with the same CpG sites, indicating an association between the two variables. When looking in the other direction, all I can conclude after referring to the paper 'Epigenetic Signatures of Cigarette Smoking' by Joehanes *et al.* was that one CpG site, cg19653589, has an association with both eczema and smoking. The CpG site with the largest coefficient is cg01294327 at -0.040. It has a similarly large value when compared to ever eczema and methylation at age seven.

Lastly table 5.5 shows that there are three CpG sites, cg09570614, cg13689560 and cg26728709, which overlap when looking at cord blood methylation and whether a child has ever had eczema, and methylation at age seven and whether a child has ever had eczema. These strengthen the

association between smoking and cord blood/ever eczema and smoking and methylation at seven/ever eczema because they continue and appear in both analyses.

5.4.2 Animal exposure

As explained above, EWAS showed that children who had been exposed to cats in early life had a increased amount of methylation at the sites cg21758133 (0.3%), and cg13790288 (0.3%), and a decreased amount of methylation at site cg10824810 (-2%), cg14643686 (-1%) and cg09114441 (-0.2%). When risk factors were adjusted for, this changed to two suggestive associations. Children who had been exposed to cats had an increased amount of methylation at the site cg26491624 (0.8%) but a decreased amount at cg07264682 (-0.2%). When looking at the EWAS of dog exposure, children who had spent time with dogs had an increased amount of methylation at the site cg27252766 (0.5%) and cg03276401 (1.5%). However, after adjusting for risk factors, there were two suggestive associations. Children had an increased amount of methylation at cg25703541 (0.7%) and a decreased amount of methylation at the site cg03639185 (-10%). None of the CpG sites identified were associated with eczema, inflammation or rash in the gene lookups. Overall there are no results which show an association with ecema.

5.4.3 Breastfeeding

In ALSPAC breastfeeding was associated with increased risk of eczema, which attenuates slightly, but not completely, after adjustment for confounders. This is not consistent with literature. Generally, it is thought that breastfeeding provides nutrients and hormones which can protect the growing child. It might be assumed that breastfeeding lowers the risk of eczema, not exacerbates it. A reason for this difference may be that exclusiveness of breastfeeding was not investigated, suggesting that its not the benefit of breastfeeding that's having an effect but the avoidance of cow's milk or potentially formula milk. Or breastfeeding could be carried out over a longer period of time. Four weeks is not really very long given babies are recommended to start eating solids at six months. Other reasons may be residual confounding. These factors could have an impact.

Breastfeeding was also investigated in terms of an EWAS. When looking at breastfeeding and eczema, there was one CpG site associated with both variables which had small p-values, cg04722177. This had a relatively large coefficient of -0.005.

However since these were in cord blood there is no association because breastfeeding cannot affect cord blood. No CpGs associated with breastfeeding were also associated with eczema and no CpGs associated with eczema were amongst the most significantly associated with breastfeeding in the Hartwig *et al.* paper (99), but I was unable to test all CpGs, as the full results were not available.

5.4.4 Summary

Overall there were 23 CpG sites which had small p-values when looking at "smoking and methylation" and "eczema and methylation". This could indicate that there was a relationship between smoking during pregnancy and a smoky environment and eczema. In particular, there was evidence for methylation at CpG sites for a mother smoking during pregnancy (50 CpGs, 9 selected) and methylation and a smoky environment (56 CpGs, 14 selected). There was also one CpG site which might have a weak association between smoking and eczema, cg19653589. This was found using the Joehanes *et al.* (97) paper. I took the 2623 sites from the paper and whittled them down to those which p-values had $P<0.01$. I also found three CpG sites that showed overlap when looking at smoking and eczema, between smoking and cord blood methylation/ever eczema and smoking and methylation at age seven/ever eczema. The CpG site cg09570614 has the largest coefficients as well as low p-values. Whilst some suggestive evidence was found here, no robust associations were found for animal exposure and breastfeeding. Potentially DNA methylation could be acting as a mediator in the smoking/eczema relationship, but this could not said to be occurring for the other two risk factors. The limitations of this are that the result was taken from a paper and not carried out myself. The next step would be to replicate the study in another cohort or validate the findings in ALSPAC using another method, eg, pyrosequencing.

Figure 5.5: Risk factors associated with eczema (smoking during pregnancy)

Smoking is associated with eczema using ALSPAC data (tick). 9 CpG sites were associated with smoking (Joehanes *et al.*). n=9 comes from the most strongly associated CpG's with smoking during pregnancy and eczema. Methylation was associated with eczema at 25 CpG sites in ALSPAC. Overall 1/25 sites was associated with eczema and smoking.

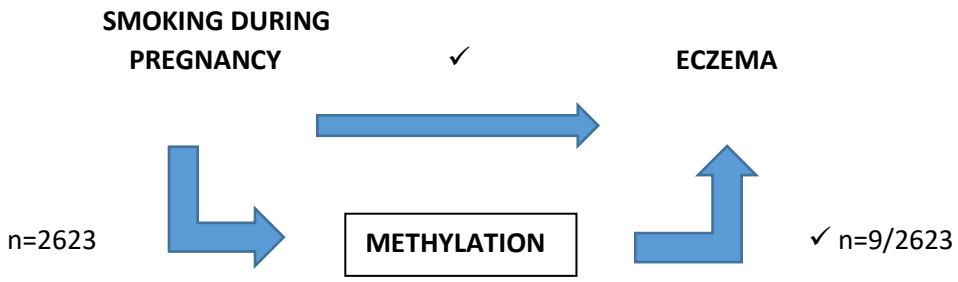


Figure 5.6: Risk factors associated with eczema (a smoky environment)

Smoking is associated with eczema using ALSPAC data (tick). 14 CpG sites were associated with smoking (Joehanes *et al.*). n=14 comes from the most strongly associated CpG's with a smoky environment and eczema. Methylation was associated with eczema at 25 CpG sites in ALSPAC. Overall 1/25 sites was associated with eczema and smoking.

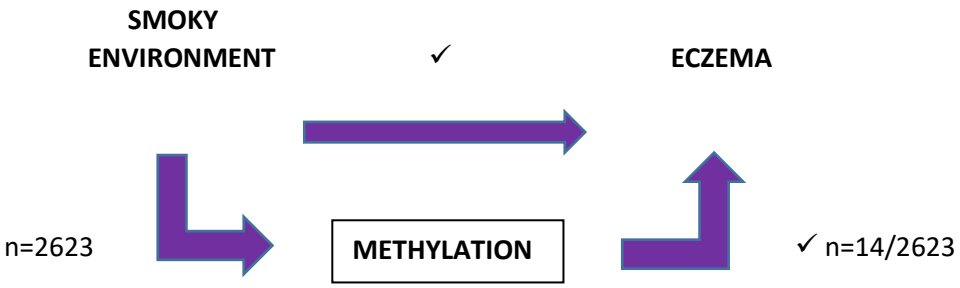


Figure 5.7: Risk factors associated with eczema (cat exposure)

Cat exposure is associated with eczema using ALSPAC data (tick). 6 CpG sites is the number of suggestive CpG sites found from the cat EWAS. Methylation was associated with eczema at 25 CpG sites in ALSPAC. Overall 0/6 sites were associated with eczema and cat exposure.

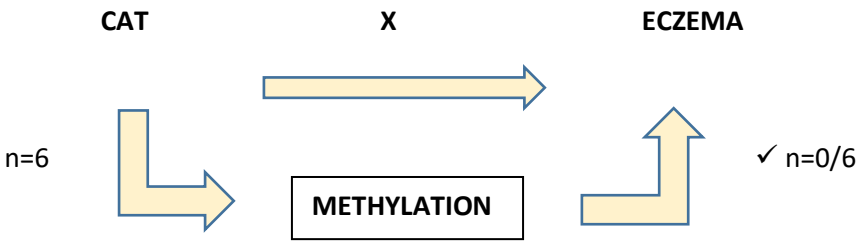


Figure 5.8: Risk factors associated with eczema (dog exposure)

Dog exposure is associated with eczema using ALSPAC data (tick). 4 CpG sites is the number of suggestive CpG sites found from the dog EWAS. Methylation was associated with eczema at 25 CpG sites in ALSPAC. Overall 0/4 sites were associated with eczema and dog exposure.

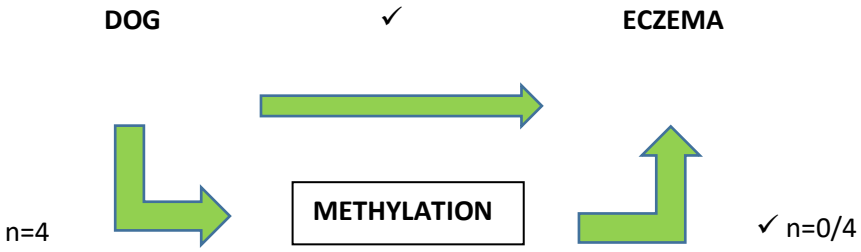
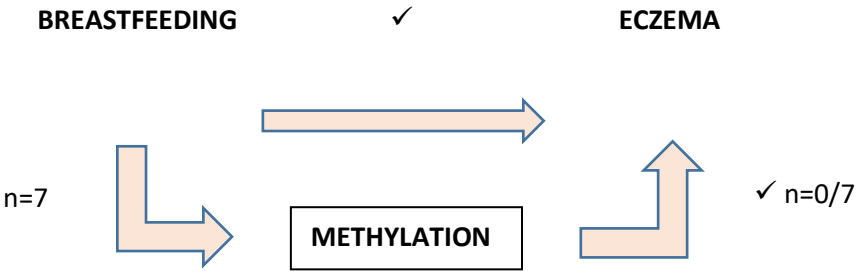


Figure 5.9: Risk factors associated with eczema (breastfeeding)

Breastfeeding is associated with eczema using ALSPAC data (tick). 7 CpG sites were associated with breastfeeding (Hartwig *et al.*). n=7 comes from the most strongly associated CpG's with breastfeeding. Methylation was associated with eczema at 25 CpG sites in ALSPAC. Overall 0/7 sites were associated with eczema and breastfeeding.



* These show the overlap between the risk factor and methylation, and between the methylation and eczema

Table 5.16: A summary of the literature, logistic regressions, and EWAS's that were carried out

	Smoking during pregnancy	**Smoky environment**	**Cat exposure**	**Dog exposure**	**Breastfeeding**
What the literature says?	Worse	Worse	Worse	Better	Better
Logistic regressions	Null	Better	Null	Better	Worse
EWAS plan	Already carried out, Johannes paper	Already carried out, Johannes paper	I carried out EWAS (5 unadjusted associations, 2 adjusted)	I carried out EWAS (2 unadjusted associations, 2 adjusted)	Already carried out, Hartwig paper
RF -> Eczema	See pg 86. Joehanes 2623, minimised to those <1x10-8, merged. Under 0.05 came up with 50/56 which was reduced to those less than 0.01, 50 and 56	See pg 86. Joehanes 2623, minimised to those <1x10-8, merged. Under 0.05 came up with 50/56 which was reduced to those less than 0.01, 50 and 56	Cat data, looked up eczema data	Dog data, looked up eczema data	Hartwig paper, nothing was found in eczema data
Eczema -> RF	There was 1 lookup of eczema CpG in RF data (smoking)	There was 1 lookup of eczema CpG in RF data (smoking)	Eczema data, looked up cat data	Eczema data, looked up dog data	Could not locate as these were taken from a paper

CHAPTER 6. DISCUSSION

6.1 OVERVIEW

I chose to carry out an EWAS on eczema because this is unprecedented, asides from several smaller studies. There have been EWAS's carried out in other atopic illnesses, which indicates that an EWAS of eczema is important and could yield interesting results. Eczema is a vital area for research. Currently the annual cost of eczema in the UK is approximately £465 million. There are also many other comorbidities that occur with eczema. It can cause or exacerbate later mental health problems such as depression, anxiety, obsessive compulsive disorder (OCD) due to excessive washing (19), (20) and Attention Deficit Hyperactivity Disorder (ADHD) (21). The 'Atopic March' also entails other atopic diseases, such as asthma and hay fever, to develop alongside, before or after eczema. By carrying out a series of EWAS's I aimed to find one or more CpG sites with an association between DNA methylation and eczema which are suggestively associated at $P<0.05$ in cord blood and later blood collection, and whether a person ever has eczema or has it in the last 12 months. I then looked at three risk factors to try to establish whether methylation is acting as a mediator between the risk factor and eczema.

Throughout this thesis I have looked at several epigenome-wide association studies of eczema. An EWAS looks to see if there is a difference in DNA methylation levels between DNA measured in cases and controls of eczema at particular CpG sites. In chapter 5, I identified associations between methylation and smoking, animal exposure (cat and dog) and breastfeeding, either by extracting data from available literature or by conducting additional EWAS's in ARIES. In this chapter I compared results from EWAS's testing associations in the two relationships: i) methylation and eczema, and ii) methylation and risk factors, to see whether there was overlap between these, indicating a link between the risk factor and eczema, with methylation as a potential mediator. In order to identify if methylation is mediating the relationship between risk factor (exposure) and eczema, more analyses are required. These could be formal mediation analyses (eg, using multivariable regression according to the method outlined by Baron and Kenny (100)) or using Mendelian Randomisation (MR) (101).

The reasons for choosing each of the three risk factors are detailed in Chapter 5. Smoking, pet exposure and breastfeeding are all thought to have an association with eczema. If the results identify a link between one of these risk factors, methylation and thereafter eczema, this would potentially make a big contribution to health worldwide. This is because guidelines could be brought

out which advise mothers on smoking and animal exposure during pregnancy and early life, and breastfeeding after birth. In addition, treatments or preventiative measures could be put in place to treat or prevent the disease at an early age.

6.2 SUMMARY

In this study I looked at four situations based around whether **childhood eczema is associated with differences in DNA methylation.** I compared cord blood methylation with whether a person has ever had eczema by the age of 15/17, or methylation at age seven and again, whether they have ever had eczema by the age of 15/17. I also conducted cross-sectional analysis, looking at methylation at ages seven and 15/17 and current eczema. Overall I found 25 sites where there was a suggestive association between DNA methylation and eczema. However the genes containing the CpG sites, as looked up in genecards (95), did not in any way relate to eczema or atopic disease.

The second section of the study looked at environmental risk factors, such as smoking, animal exposure, breastfeeding. Here I looked to see whether methylation acts as a mediator between a risk factor and eczema, although no causal inference can be implied, merely association. The environmental risk factors were chosen after reviewing the current literature and assessing the factors that would fill a gap in the research. Smoking is well studied and several EWAS's have been carried out between smoking and methylation (102) (103) (104). However little work has been done on the relationship between smoking and eczema. Animal exposure is interesting because I wanted to see whether there is a possible mediator between animal exposure and eczema in the literature. There is also evidence that breastfeeding is associated with eczema, therefore I was interested in looking at it with respect to methylation. Overall I found some associations between smoking and eczema (23, split into 9 in cord blood which would indicate smoking during pregnancy and 14 at age seven which would indicate a smoky environment), as well as an association at cg19653589, but none when looking at animal exposure and breastfeeding (105). There is potentially evidence therefore, that methylation at those 23 CpG sites could be mediating smoking during pregnancy/a smoky environment and the development of eczema. However I did not find that any of the associated genes were linked to eczema.

I also found three CpG sites, cg09570614, cg13689560 and cg26728709, which overlap when looking at cord blood methylation and whether a child has ever had eczema, and methylation at age seven and whether a child has ever had eczema, indicating a more robust association. These all had relatively large coefficients, with the highest being cg09570614.

6.3 ECZEMA AND METHYLATION

6.3.1 Results

When looking at the first longitudinal question, methylation at birth and whether a person ever has eczema by age 15/17, there was no evidence for any associations below the Bonferroni threshold of P<1x10-7 but there were two suggestive associations at P<0.05. On average, babies who go on to have eczema had more methylation at the sites cg04804139 (0.7%) and cg09418000 (1.3%) in cord blood, with coefficients of 0.007 and 0.013. Secondly, looking at methylation at age 7 and whether a person has ever had eczema by age 15/17, there is evidence for four suggestive associations in the analysis looking at methylation at seven and ever eczema at P<0.05. In this question people with eczema had a reduced amount of methylation at the site cg07166235 (0.2%) but an increased amount at cg24211994 (0.7%), cg14511273 (0.8%) and cg26368024 (0.2%). Thirdly, looking at associations cross sectionally with methylation at age 7, but eczema only in the last 12 months, there was evidence for a weak association at P<0.05, CpG site cg23673397 (1.1%) in a positive direction. Lastly, there were 18 suggestive CpG associations in the analysis looking at methylation at 15/17 and eczema in last 12 months at P<0.05. None pass Bonferroni correction. Of the 25 there are three associations with P<1x10-7 and 15 at P<1x10-6, those at P<1x10-7 being cg03220363 (-1.1%), cg07721777 (-2.6%) or cg26716834 (-1.2%) (all with methylation measured at 15/17).

The number of associations in each question at P<0.05 is 2, 4, 1 and 18. The total number of CpG sites associated in each question, the total of the six models, is 20, 44, 11 and 96 respecitvely at the P<0.05 threshold. This is not what would be expected, as one might think that the number of associations would decrease as you go from looking at methylation in cord blood and whether a person has ever had eczema, to methylation at age 15/17 and whether a person has had eczema in the last 12 months. This is because according to table 4.1, the number of cases goes down from 408 in the first question (cord blood, ever eczema), to 327 (methylation age seven, ever eczema), to 110 (methylation at age seven, eczema in last 12 months) and lastly down to 63 (methylation age 15/17, eczema in last 12 months). However results suggest more associations between methylation and eczema as methylation is measured later. This does not mean that eczema alters DNA methylation, but rather that methylation has had more time to occur. This is good because it provides more evidence to be able to reliably say that there is a relationship between methylation and eczema.

Next I looked for associations between DNA methylation and eczema, whether the methylation be in cord blood or at age 7 or 15/17, and whether eczema has ever occurred in childhood or occurred in the last 12 months. The relationship between DNA methylation and eczema was investigated in Chapter 4, which carried out EWAS's between methylation and eczema.

When looking at whether there was any methylation effect which persisted across timepoints, two CpGs, cg07166235 and cg26368024, had small p-values at P<0.05 in 'ever eczema' (both in cord blood and methylation at age seven), and methylation at age 7 and eczema in the last 12 months. Two CpGs, cg04804139 and cg24211994, had small p-values at P<0.05 in 'ever eczema' (both in cord blood and methylation at age seven), and methylation at age 15/17 and eczema in the last 12 months. Lastly, the methylation effect persisted across methylation at age 15/17 and eczema in last 12 months, and methylation at 7 and eczema in the last 12 months. The CpG site cg01182386 has a similarly small p-value when looking at both. The largest coefficients across all of these were at CpG site cg24211994 with coefficients at 0.007 and 0.009 when comparing ever eczema at cord blood or age 7, and methylation at 15/17 and methylation in the last 12 months.

When looking at the results from the four questions, the combined 25 associations for each of the four analyses using model (iv) are all only suggestive. In the model adjusting for sex, surrogate variables, two socioeconomic status variables (social class and child ethnic background), maternal history and cell count', there are two, four, one and 18 CpG sites respectively which show a difference in DNA methylation levels between DNA measured in cases and controls at particular CpG sites. An issue worth considering is reverse causation. Could it be in fact that eczema causes methylation rather than the other way around? The only sure way of making sure you avoid the problem of reverse causation is to use the results, the associated CpG sites, found from the question 1 analysis, which is where DNA methylation is measured in cord blood. Or conduct additional analyses, eg using Mendelian Randomization.

6.3.2 Advantages and limitations

Phenotypic data in ALSPAC

ALSPAC is a study which contains data from mothers during pregnancy and throughout the child's life (currently up to age 25 years). Questionnaires cover the period from maternal pregnancy to late teens of the offspring. There are many questions about eczema throughout childhood, not just at

one point, which means we have access to a range of data including when eczema starts to develop, its life course/pattern and when it exacerbates or goes away. Bias could be introduced by the fact that there are self-reported symptoms. People who do not understand about the disease could under or overreport suffering from it. The effect of misreporting means that there would appear to be less eczema in the population than there actually is. An error such as this in the outcome data would lower power and effect sizes would be likely to be attenuated.

ARIES data

Methylation data is taken at birth (in cord blood), at age 7 and at age 15/17. So, both longitudinal and cross-sectional analysis can be carried out. The data produced is genome-wide, which means the study can be conducted with a hypothesis-free approach. Basically that means that we are seeing what results emerge, rather than putting forward a hypothesis and targeting certain CpG sites.

Power

The ALSPAC cohort has a relatively large number of participants, but it is important to look at power. ARIES only includes 1,024 people so power is low. Power is defined as the probability that you can reject the null hypothesis (that there is no association). It is the likelihood of identifying an effect when there is something there which can potentially be identified. When power is high there is less chance of making a Type II error. As well as sample size, it depends on significance-value threshold and the size of effect in a population. It is important to carry out a power calculation to find the amount of trials needed to detect an effect, or to avoid incorrectly rejecting the null hypothesis. Reasons to carry out a power calculation are to find the minimum number of subjects necessary to include (important if the study/researchers are on a low budget), or to determine power of an existing resource.

I used an online calculator to calculate the number of people required to detect a significant effect. With a power of 80%, and an alpha level of 1.1x10-7, I estimated I could identify a mean methylation difference of 1% between cases and controls assuming a standard deviation of 2.
An estimated 303 cases and 303 controls were required. This is less than the number used in my analysis indicating the sample size was likely to have been sufficient.

Confounders

As well as questions on eczema and lifestyle factors, details on other variables are available, so that confounders can be adjusted for. Cell count also needs to be considered as a potential confounder. I adjusted for this when running my set of EWAS's. One way of explaining this is that DNA methylation differs by the type of cell being looked at. In ARIES methylation levels are measured in blood cells, such as eosinophils. Not adjusting for cell count means the results you get might be due to cell proportions rather than actual methylation changes. Cell count could be an issue if cell proportions differ by case/control status. In addition, I have dealt with batch affects by looking at conducting surrogate variable analysis and including surrogate variables as confounders in my analyses. This captures other latent sources of variation, not just batch. It relies on detecting patterns in the data that are unrelated to variables in your model so also likely to be an imperfect approximation. The pros of this are that it can correct for any potential bias due to batch. However it cannot cover all variation, so should not be relied upon completely.

EWAS

Illumina Infinium beadchips cover approximately 2% of CpG sites so offer relatively low genome-wide coverage. However, CpG sites are annotated to 99% of all RefSeq genes so most coding regions of the genome have some level of coverage. In addition, results from array-based methods should be confirmed using a separate assay. They can be replicated and just repeated in the same way. This is replication, the action of copying or reproducing something. They can also be validated, which checks or proves the accuracy of an analysis. Ideally you would want to test the exact same CpG site but by using a different method or potentially using a different group or population.

Tissue specificity

In this study, ARIES data looks at DNA methylation in blood cell types. But this might not be the relevant tissue for eczema. In eczema, studying skin tissue would be better. One study in the literature (79) looked at atopic dermatitis and methylation in different tissue types and found no genome-wide significant differences in methylation when looking at whole blood, but a difference between cases and controls when looking at lesional epidermis of healthy subjects and patients. An advantage would be that if skin tissue testing yielded different levels of DNA methylation between

cases and controls this could be more revealing than measuring blood samples alone. However eczema implicated pathways include immune function, that are likely to be present in the blood. So for some CpGs blood might be the relevant tissue.

6.3.3 Wider context

It is important to place the results that have been found into a broader picture. As already known, a GWAS was carried out by Paternoster *et al.* (66). Of the 31 genes identified as related to eczema (66), (including the strongest risk factor, Filaggrin (67)), the study by Paternoster *et al.* (66) found 10 new genetic loci to add onto the 21 already discovered. However, a full EWAS had yet to be carried out. One collection of studies that has done work in the area of atopic diseases, particularly asthma, is The Pregnancy And Childhood Epigenetics (PACE) consortium (98) - the epigenetic involvement of early childhood environmental exposures on disease development. Other literature shows that EWAS's have been carried out, but on atopic illnesses asthma (77) and antibodies like Immunoglobulin E (7). In 2017 Arathimos *et al.* found all significant links between methylation and asthma attenuated after adjustment for confounding, so there were no associations (77). Xu *et al.* found 14 CpG sites associated with asthma (78). An EWAS is vitally important because it will help identify CpG sites associated with eczema and increase our understanding of what causes the disease.

6.3.4 Implications

Not as a direct result of this research, but in future an increased awareness of the relationship between methylation and eczema may make it possible to develop a treatment that targets methylation and has the potential to reverse the disease or prevent the methylation/disease developing in the first place. Individuals could get their DNA tested and if they were susceptible, pre-eczema development treatment could be offered to stop the disease occurring or getting worse (although this is far in the future at this moment). Currently our understanding of epigenetics and the mecahnisms of underlying disease is quite reduced. A better understanding will eventually lead to better management and treatment, or perhaps even prevention.

6.3.5 Future work

There are many ways in which the study can be improved. I could take the results I have discovered and try to replicate and validate these findings. This could be done by approaching another study to see if through their different eczema questions and separate methylation measurements yielded results which supported the same CpG sites associated with eczema. To start, the EWAS's could be carried out again using the same data and see if the results are the same. This would get around batch effect, for example in case the code in STATA was run with typos or some random errors occurred. After that, results could be validated by using other data from a separate study to see if there were similar results. In terms of conducting analysis in skin, I could potentially look at publically available data with cross tissue comparisons of the CpG sites of interest in blood and skin to see if they are likely to correlate. This might indicate studies in skin are likely to be worthwhile.

Carrying out a meta-analysis to increase the number of individuals in the study would be advantageous. In addition, rather than defining all eczema cases together, the study could look at the life-course of disease. ie. whether it develops young then disappears, develops young and remains until adulthood, or develops older and continues until old age, etc. This would be useful for trying to ascertain the pattern of eczema and whether methylation changes at birth lead onto a short spell of the disease, or a long-term development of eczema that lasts until adulthood. Potentially some lab work could be carried out to measure DNA methylation in more relevant tissues, such as skin.

6.4 RISK FACTORS AND METHYLATION

6.4.1 Results

Here I introduced three risk factors, smoking, animal exposure and breastfeeding. From the literature I could see that smoking and breastfeeding have been analysed using pre-existing EWAS's. However, I decided to run four EWAS's on animal exposure (cats and dogs), with and without adjusting for other risk factors for each.

In Chapter 4 I looked at the different relationships between DNA methylation and eczema through four questions. Chapter 5 focused on the risk factors and what they showed, and whether they were related to eczema. The main point here is to see whether methylation is mediating the relationship between the risk factor and eczema. To work this out, we can look at the EWAS's between risk factors and methylation, and methylation and eczema, to see if there is an overlap, whether increased/decreased methylation in cases and controls occurs at the same CpG sites in both analyses. There were 23 CpG sites associated with both smoking and eczema, as well as one more CpG site cg19653589 which had small p-values and a large change in the size of the coefficient. There is potentially evidence of a relationship here at these 23 sites between smoking and eczema. These results were taken from the Joehanes paper, which listed 2623 as associated with smoking and methylation. I then took the top 9/14 where $P<0.01$, before merged and subsetting the data to find out which CpG sites had small p-values for both.

Three CpG sites, cg09570614, cg13689560 and cg26728709, were found to overlap when looking at smoking and eczema, and comparing the two different timepoints of when methylation was measured, cord blood and age seven. These all had relatively large coefficients, with the highest being cg09570614.

Next, I looked at whether there was a relationship between animal exposure and eczema, with methylation as a mediator. However there were none. When considering breastfeeding, again there were two associations at cg04722177 and cg03945777 but since these were in cord blood they will not be considered.

6.4.2 Advantages and limitations

Phenotypic data in ALSPAC

One advantage is that the measuring of these variables continues throughout childhood into late teens. ALSPAC has a lot of questions on variables, making it easier to look at factors like smoking, animal exposure and breastfeeding. 'Smoking' looks at both smoking during pregnancy (at 18 weeks gestation) and a smoky environment during childhood (at six months). Cat and dog exposure are recorded at 15 months, and breastfeeding at four weeks. Looking at smoking during pregnancy, 18 weeks is approximately half way through the pregnancy, so would be a good point to record whether a mother is smoking. The six months postnatally for a smoky environment gives a long enough duration for children to be exposed to the smoke, as with the cats and dogs. Breastfeeding at 4 weeks is quite early on, so catches mothers who only breastfeed for a very short while. However when looking at the possible breastfeeding scenarios and how I have analysed the data , if I was to take *any* breastfeeding, then it might not have been possible to measure any of the beneficial effects of drinking formula milk.

The three variables are measured before birth (ie, smoking) and during childhood (a smoky environment, animal exposure and breastfeeding) and are covered by both child based and mother questionnaires. One limitation is that for the smoking variable, the data used doesn't cover how many cigarettes a day were smoked. However this could be said for all the variables. They are categorised as 'yes' or 'no' (in answer to whether the have eczema or they smoke, breastfeed, exposure their child to animals), however there will always be different degrees of exposure. There may even be a relationship between the amount of exposure and the severity of eczema, for example.

Secondly, the breastfeeding variable doesn't cover bottled breast milk. However, babies being exclusively fed expressed milk may be different (on average) to those breast fed directly (eg, they probably more likely to be pre-term). In addition, mothers giving expressed milk would probably define their feeding category as breast fed, not bottle fed. So they would look at whether they were giving breast milk or formula, rather than the route by which it is delivered. Exclusivity of breastfeeding is important because it may not be the intake of mother's milk that is having the benefits, it may be the avoidance of cow's milk which may contain allergens. That's why we have created the 'always breastfeed' category. This could be accounted for in ALSPAC by producing a

questionnaire which asks mothers whether they feed by breast, and if they answer 'yes', whether this is exclusively breast milk, or do they alternate or include cow's milk. According to research, the UK doesn't have a particularly high breastfeeding rate, meaning that it may be more difficult to find women who are both in ALSPAC and ARIES and whether they exclusively feed breast milk. However, in my study I have found that actually levels of women who feed are actually quite large. Another big issue to look at, rather than exclusivity, is investigate mothers breastfeeding for a longer period of time.

Power

It has been shown that some exposures have quite large effects on methylation. For example, smoking has some very significant effect sizes (eg, AHRR) of approximately 25% reduced methylation in people who smoke compared to those who have never smoked. Therefore to detect some associated CpG sites, the sample size needed would be very small. However, with animal exposure and breastfeeding, these could have smaller effect sizes, meaning that many more people are needed to be included in the study to yield some meaningful results.

6.4.3 Wider context

According to the literature smoking during pregnancy and a smoky environment make eczema worse. Cat exposure also makes it worse, and dog exposure and breastfeeding improve the outcome. Looking into it in more detail, two large systematic reviews find no evidence for a relationship between smoking during pregnancy and eczema (43) (44) and one study shows a positive association (45). Two reviews show a positive association between a smoky environment and eczema (43), (44) and another shows no association (45). Fretzayas (47) and Langan (48) found contradictory results when looking at cats, and a small negative association when looking at dog exposure and eczema. Overall more cat exposure was associated with more eczema, but dog exposure reduced the likelihood of developing eczema. Overall breastfeeding is thought to reduce eczema risk. Methylation has been put forward as a mediating factor which sits in the middle of the risk factor and eczema. I do not know for sure whether this is a causal pathway, all we can say is that there is an association. Studies looking into whether there is a mediating role of DNA methylation between smoking, animal exposure and breastfeeding, and eczema is unprecedented. If we find that

there is overlap between the CpG sites associated with eczema, and the different risk factors, this could be an indication of methylation as a mediator, a result which could be investigated further.

DNA methylation has been proposed as a mediator between the risk factor and eczema. However, there are other hypotheses. It may just be that the risk factors, especially smoking and animal exposure, are allergens which cause problems with the skin barrier, completely sidestepping the possibility of methylation being a mediator and playing a role. My analysis does not completely answer this question. All we know is that there are several CpG sites which seem to mediate the relationship between smoking exposure and eczema. To test for causality we could look to see whether the CpG site being addressed is similarly associated at different timepoints in the child's life. If results are repeated this may be an indication that there is a relationship between the risk factor and eczema mediated by methylation. To rule out reverse causation we could focus our interest on looking at samples from cord blood. However this won't work for postnatal exposures.

6.4.4 Implications

There was only a small amount of CpG sites which had a weak association with eczema in my study. However the analyses helped identify risk factors that are also associated with methylation. I then looked to see if there was a relationship between the risk factor and methylation, and then between methylation and eczema. If similar CpG sites were found to be associated in both EWAS's it could be potentially be that DNA methylation is mediating this relationship.

6.4.5 Future improvements

This study has made some progress in identifying some CpGs that might potentially be mediating the relationship between smoking and eczema. I did not find any evidence for CpGs mediating the relationships between animal exposure or breastfeeding and eczema. However further work is required to investigate what causal relationships might explain these observations. Formal mediation analysis would be a good next step to follow up with, although you can only do this when there is evidence of overlap. Looking in the reverse direction, methylation might come after the onset of eczema (rather than before). Or perhaps methylation could cause someone to smoke or breastfeed? In this situation, methylation in cord blood could go on to make someone take up smoking/breastfeeding. If there was a difference of an increase or decrease in methylation

measured in cord blood and this coincided with cases and controls of eczema, smoking, animal exposure and breastfeeding, this could be evidence that the methylation was coming first.

One improvement that could be made is to carry out a better powered EWAS of the risk factors or eczema, to identify an increased amount of CpG sites associated with them. Or alternatively, we could look at the mechanisms themselves that may be mediating the relationships in question. Looking into the actual amount of smoking someone does, ie, the number of cigarettes, rather than just 'Yes I smoke', 'No I don't smoke', could be insightful. It may also be important to investigate the difference between cats and dogs, why are there different outcomes of exposure to each? And why did pet exposure and breastfeeding have no overlap in their CpG sites with eczema? Hypothetically this could be because there was no strong relationship between animal exposure/breastfeeding and eczema, mediated by methylation. Lastly it could be interesting to look at duration and exclusivity of breastfeeding. It may not be the benefit from breast milk that is having the protective effect on eczema, it could be the absence of cow's milk, which contains allergens. Therefore, avoiding cow's milk by feeding breast milk could be a way of avoiding these allergens, and thus reducing eczema.

6.5 CONCLUSION

25 CpG sites associated with eczema were found after carrying out four main EWAS's. Following on from this, two CpG sites, cg04804139 and cg24211994, which had small p-values at $P<0.05$ when looking at methylation in cord blood and at age seven respectively and whether a person has ever had eczema (at age 15/17), also had small p-values when methylation was measured at 15/17 and eczema in the last 12 months. The same was found for CpG sites cg07166235 and cg26368024, except this time the comparison was between methylation at age seven and ever eczema, and methylation at age seven and whether a person has ever had eczema. There is also a link between methylation at the ages of 15/17 and 7 in the last 12 months, there is a CpG site, cg01182386, which has similarly small p-values in both. The coefficient values were split with some larger than others. The largest were at CpG site cg24211994.

There were 23 CpG sites which were associated with both smoking and eczema. In addition, after referring to the paper 'Epigenetic Signatures of Cigarette Smoking' by Joehanes *et al.* one CpG site, cg19653589, had an association with both eczema and smoking. This may suggest these CpGs mediate the relationship between smoking and eczema, but causal analyses or formal mediation analyses would be required to confirm this. In addition, the p-values are not very low and do not meet the Bonferonni threshold. The largest coefficient is cg01294327 at -0.040. CpG sites cg09570614, cg13689560 and cg26728709, showed overlap when looking at cord blood methylation and whether a child has ever had eczema, and methylation at age seven and whether a child has ever had eczema, which provides stronger evidence that there is an association between smoking and eczema.

Looking at animal exposure, whilst I found associations between cat/dog exposure and DNA methylation (see tables 5.9 and 5.10), there were no associations between cat and dog exposure and methylation, and methylation and eczema. Breastfeeding showed two associations as well, at cg04722177 and cg03945777, although these are not important as they occurred in cord blood. A larger EWAS for the risk factors and eczema might be required in order to detect such CpGs by increasing power. Out of all of the CpG sites, the genes they were contained within did not have an association with eczema or atopy that is known of, so we were unable to draw conclusions in this way.

The study could be improved, as already discussed: For example, a meta-analysis to increase power, an improved study of the existing risk factors, comparing mothers smoking only during pregnancy or

only during childhood to see which is more influential, etc, or using CpG sites associated with smoking at specific timepoints. Identification of specific CpGs that influence eczema (and mediate the relationship between risk factors and eczema) would lead to better understanding of the underlying mechanisms of this condition and could lead to better treatment to prevent disease or treat exacerbation during flare ups. If strong associations with methylation were identified, it may be possible to identify babies or children who were more susceptible to developing eczema.

In the future, in order to strengthen this study, it would first be beneficial to repeat the EWAS's to check the results and rule out bias resulting from batch effect. Another study could then be carried out by another person, with a different data set and software to see if we can validate the results. The power in this study is quite low, so a way around this would be to increase the sample size, either by carrying out more data collection or by carrying out a meta-analysis. Of course it must be considered that methylation might not be mediating the relationships at all. It may be some other factor involved, or it may be reverse causation.

REFERENCES

1. Brown SJ. Atopic eczema. Clinical Medicine. 2016;16:66-9.
2. Gilles S, Akdis C, Lauener R, Schmid-Grendelmeier P. The role of environmental factors in allergy: A critical reappraisal. Experimental Dermatology. 2018;27:1193-200.
3. Cabana MD, McKean M, Caughey AB, Fong L. Early Probiotic Supplementation for Eczema and Asthma Prevention: A Randomized Controlled Trial. Pediatrics. 2017.
4. Paternoster L, Savenije OEM, Heron J, Evans DM. Identification of atopic dermatitis subgroups in children from 2 longitudinal birth cohorts. Journal of Allergy and Clinical Immunology. 2018;141:964-71.
5. Rożalski M, Rudnicka L, Samochocki Z. Atopic and Non-atopic Eczema. Acta Dermatovenerologica Croatica. 2016;24:110-5.
6. Rossberg S, Gerhold K, Geske T, Zimmermann K, Menke G. Elevated blood eosinophils in early infancy are predictive of atopic dermatitis in children with risk for atopy. Pediatric Allergy and Immunology. 2016;27:702-208.
7. Liang L, Willis-Owen S, Laprise C, Wong K, Davies G. An epigenome-wide association study of total serum immunoglobulin E concentration. Nature. 2015;520:670-4.
8. Takadama S, Nakaminami H, Aoki S, Akashi M. Prevalence of skin infections caused by Panton-Valentine leukocidin-positive methicillin-resistant Staphylococcus aureus in Japan, particularly in Ishigaki, Okinawa. Journal of Infection and Chemotherapy. 2017;23:800-3.
9. Clebak KT, Malone MA. Skin Infections. Primary Care. 2018;45:433-54.
10. Sala-Cunill A, Lazaro M, Herráez L, Quiñones MD. Basic Skin Care and Topical Therapies for Atopic Dermatitis: Beyond Essential Approaches. Journal of Investigational Allergology and Clinical Immunology. 2018;28:379-91.
11. Grant T, Aloe C, Perzanowski M, Phipatanakul W. Mouse Sensitization and Exposure Are Associated with Asthma Severity in Urban Children. Journal of Allergy and Clinical Immunology: In Practice. 2016;5:1008-14.
12. Takemura M, Inoue D, Takamatsu K, Itotani R. Co-existence and seasonal variation in rhinitis and asthma symptoms in patients with asthma. Respiratory Investigation. 2016;54:320-6.
13. Hoyte FCL, Nelson HS. Recent advances in allergic rhinitis. F1000Research. 2018;7:1333.
14. Stone KD. Atopic diseases of childhood. Current Opinion in Pediatrics. 2002;14:634-46.
15. Leslie TA, Greaves MW, Yosipovitch G. Current topical and systemic therapies for itch. Handbook of Experimental Pharmacology. 2015;226:337-56.
16. Tezza G, Mazzei F, Boner A. Epigenetics of allergy. Early Human Development. 2013;89(S1):S20-S1.
17. Amarasekera M. Immunoglobulin E in health and disease. Asia Pacific Allergy. 2011;1:12-5.
18. Schmitt J, Romanos M. Lack of studies investigating the association of childhood eczema, sleeping problems, and attention-deficit/hyperactivity disorder. Pediatric Allergy and Immunology. 2009;20:299-300.
19. Kouris A, Armyra K, Christodoulou C, Katoulis A. Quality of life, anxiety, depression and obsessive-compulsive tendencies in patients with chronic hand eczema. Contact Dermatitis. 2015;72:367-70.
20. Hammer-Helmich L, Linneberg A, Obel C, Thomsen SF. Mental health associations with eczema, asthma and hay fever in children: a cross-sectional survey. British Medical Journal. 2016;6:e012637.
21. van der Schans J, Pleiter JC, de Vries TW, Schuiling-Veninga CC. Association between medication prescription for atopic diseases and attention-deficit/hyperactivity disorder. Annals of Allergy, Asthma & Immunology. 2016;117:186-91.
22. Musgrove K, Morgan JK. Infantile eczema: A long-term follow-up study. British Journal of Dermatology. 1976;95:365-72.

23. www.nhs.uk.
24. Murota H, Katayama I. Exacerbating factors of itch in atopic dermatitis. Allergology International. 2017;66:8-13.
25. Drucker AM, Wang AR, Li WQ, Sevetson E. The Burden of Atopic Dermatitis: Summary of a Report for the National Eczema Association. Journal of Investigative Dermatology. 2016;137:26-30.
26. Caubet JC, Eigenmann PA. Allergic triggers in atopic dermatitis. Immunology And Allergy Clinics of North America. 2010;30:289-307.
27. Lonne-Rahm SB, Rickberg H, El-Nour H, Mårin P. Neuroimmune mechanisms in patients with atopic dermatitis during chronic stress. Journal of the European Academy of Dermatology and Venereology. 2008;22:11-8.
28. Newman GT. Atopic dermatitis and severe menopausal syndrome in young castrate. Journal of the American Medical Women's Association. 1953;8:97.
29. Weatherhead SR, S. C, Reynolds NJ. Eczema in pregnancy. British Medical Journal. 2007;335:152-4.
30. Shankar AU, Kumar BR, Gupta R, Agarwal P. Hand Eczema. Indian Journal of Dermatology. 2014;59:213-24.
31. Shunsheng Han C. A specific hygiene hypothesis. Medical Hypotheses. 2016;93:146-9.
32. Rodríguez E, Weidinger S. Genetics of atopic eczema. Hautarzt. 2015;66:84-9.
33. Rodriguez E, Weidinger S. [Genetics of atopic eczema : An update.]2015.
34. Nutten S. Atopic Dermatitis: Global Epidemiology and Risk Factors. Annals of Nutrition and Metabolism. 2015;66(S1):8-16.
35. Odhiambo JA, Williams HC, Clayton TO, Robertson CF. Global variations in prevalence of eczema symptoms in children from ISAAC Phase Three. Journal of Allergy and Clinical Immunology. 2009;124:1251-8.
36. Kay J, Gawkrodger DJ, Mortimer MJ. The prevalence of childhood atopic eczema in a general population. Journal of the American Academy of Dermatology. 1994;30:35-9.
37. Lawton S. Sun, sea, holidays and eczema. Journal of Family Health. 2015;Supp:4-6.
38. Mollerup A, Veien NK, Johansen JD. An analysis of gender differences in patients with hand eczema - everyday exposures, severity, and consequences. Contact Dermatitis. 2014;71:21-30.
39. Gerada E, Muscat HA, Camilleri L. Gender differences in the prevalence and severity of wheezing, rhinitis and eczema in 5- to 8- year old and 12- to 15- year old Maltese children (ISAAC-Malta). European Respiratory Journal. 2015;46(S59):1337.
40. Zirwas MJ. Contact Dermatitis to Cosmetics. Journal of Allergy and Clinical Immunology. 2018;Epub ahead of print.
41. Khanna R, Shifrin N, Nektalova T. Diet and dermatology: Google search results for acne, psoriasis, and eczema. Cutis. 2018;102:44-6, 8.
42. www.nationaleczema.org.
43. Kantor R, Kim A, Thyssen J, Silverberg J. Association of atopic dermatitis with smoking: A systematic review and meta-analysis. Journal of the American Academy of Dermatology. 2016;75:1119-25.
44. Singh S, Sharma B, Salvi S, Chhatwal J. Allergic rhinitis, rhinoconjunctivitis, and eczema: prevalence and associated factors in children. The Clinical Respiratory Journal. 2016;12:547-56.
45. Tanaka K, Miyake Y, Furukawa S, Arakawa M. Pre- and postnatal smoking exposure and risk of atopic eczema in young Japanese children: a prospective pre-birth cohort study. Nicotine & Tobacco Research. 2016;19:804-9.
46. Shirinde J, Wichmann J, Voyi K. Environmental tobacco smoke and the risk of eczema symptoms among school children in South Africa: a cross-sectional study. Public Health Research. 2015;5:e008234.
47. Fretzayas A, Kotzia D, Moustaki M. Controversial role of pets in the development of atopy in children. World Journal of Pediatrics. 2013;9:112-9.

48. Langan S, Flohr C, Williams H. The role of furry pets in eczema: a systematic review. Archives of Dermatology Journal. 2007;143:1570-7.
49. Epstein T, Bernstein D, Levin L, Khurana Hershey G. Opposing effects of cat and dog ownership and allergic sensitization on eczema in an atopic birth cohort. Journal of Pediatrics. 2011;158:265-71.
50. Wegienka G, Johnson CC, Havstad S, Ownby DR. Lifetime dog and cat exposure and dog- and cat-specific sensitization at age 18 years. Clinical & Experimental Allergy. 2011;41:979-86.
51. Pohlabeln H, Jacobs S, J B. Exposure to pets and the risk of allergic symptoms during the first 2 years of life. Journal of Investigational Allergology & Clinical Immunology. 2007;17:302-8.
52. Arshad SH, Tariq SM, Matthews S, Hakim E. Sensitization to common allergens and its association with allergic disorders at age 4 years: a whole population birth cohort study. Pediatrics. 2001;108:E33.
53. Ownby DR, Johnson CC, Peterson EL. Exposure to dogs and cats in the fi rst year of life and risk of allergic sensitization at 6 to 7 years of age. JAMA – The Latest Medical Research, Reviews, and Guidelines. 2002;288:963-72.
54. Roost HP, Künzli N, Schindler C, Jarvis D. Exposure to indoor allergens in early infancy and sensitization. Journal of Allergy and Clinical Immunology. 1999;97:70221-5.
55. Ritz BR, Hoelscher B, Frye C, Meyer I. Allergic sensitization owing to 'second-hand' cat exposure in schools. Allergy. 2002;57:357-61.
56. Mandhane PJ, Sears MR, Poulton R, Greene JM. Cats and dogs and the risk of atopy in childhood and adulthood. Journal of Allergy and Clinical Immunology. 2009;124:745-50.
57. Roost HP, Künzli N, Schindler C, Jarvis D. Role of current and childhood exposure to cat and atopic sensitization. Journal of Allergy and Clinical Immunology. 1999;104:941-7.
58. Hesselmar B, Aberg N, Aberg B, Eriksson B. Does early exposure to cat or dog protect against later allergy development? Clinical & Experimental Allergy. 1999;29(611-617).
59. Epstein TG, Bernstein DI, Levin L, Khurana Hershey GK. Opposing effects of cat and dog ownership and allergic sensitization on eczema in an atopic birth cohort. Journal of Pediatrics. 2011;158:265-71.
60. Heinrich J. Modulation of allergy risk by breast feeding. Current Opinion in Clinical Nutrition and Metabolic Care. 2017;20:217-21.
61. Victora CG, Bahl R, Barros AJ, França GV. Breastfeeding in the 21st century: epidemiology, mechanisms, and lifelong effect. The Lancet. 2016;387:475-90.
62. Elbert N, van Meel E, den Dekker H, de Jong N. Duration and exclusiveness of breastfeeding and risk of childhood atopic diseases. Allergy. 2017;72:1936-43.
63. Huang C, Liu W, Cai J, Weschler L. Breastfeeding and timing of first dietary introduction in relation to childhood asthma, allergies, and airway diseases: A cross-sectional study. Journal of Asthma. 2017;54:488-97.
64. Lodge CJ, Tan DJ, Lau MX, Dai X. Breastfeeding and asthma and allergies: a systematic review and meta-analysis. Acta Paediatrica. 2015;104:38-53.
65. Patel R, Oken E, Bogdanovich N, Matush L. Cohort profile: The promotion of breastfeeding intervention trial (PROBIT). International Journal of Epidemiology. 2014;43:679-90.
66. Paternoster L, Standl M, Waage J, Baurecht H. Multi-ancestry genome-wide association study of 21,000 cases and 95,000 controls identifies new risk loci for atopic dermatitis. Nature Genetics. 2015;47:1449-56.
67. Sandilands A, Sutherland C, Irvine AD, McLean WH. Filaggrin in the frontline: role in skin barrier function and disease. Journal of Cell Science. 2009;122:1285-328.
68. Liang Y, Chang C, Lu Q. The Genetics and Epigenetics of Atopic Dermatitis-Filaggrin and Other Polymorphisms. Clinical Reviews in Allergy & Immunology. 2016;51:315-28.
69. Carey N. The Epigenetics Revolution.140:e20163000.
70. Liu W. Epigenetics in Schistosomes: What We Know and What We Need Know. Frontiers in Cellular and Infection Microbiology. 2016;6:149.

71. www.whatisepigenetics.com.
72. Richardson BC, Patel DR. Epigenetics in 2013. DNA methylation and miRNA: key roles in systemic autoimmunity. Nature Reviews Rheumatology. 2014;10:72-4.
73. Patterson K, Molloy L, Qu W. DNA methylation: bisulphite modification and analysis. Journal of Visualized Experiments. 2011;56:e3170.
74. Bibikova M, Barnes B, Tsan C, Ho V. High density DNA methylation array with single CpG site resolution. Genomics. 2011;98:288-95.
75. Gao X, Thomsen H, Zhang Y. The impact of methylation quantitative trait loci (mQTLs) on active smoking-related DNA methylation changes. Clinical Epigenetics. 2017;9:87.
76. Gaunt TR, Shihab HA, Hemani G, Min JL. Systematic identification of genetic influences on methylation across the human life course. Genome Biology. 2016;17:61.
77. Arathimos R, Suderman M, Sharp GC. Epigenome-wide association study of asthma and wheeze in childhood and adolescence. Clinical Epigenetics. 2017;9:112.
78. Xu CJ, Söderhäll C, Bustamante M. DNA methylation in childhood asthma: an epigenome-wide meta-analysis. The Lancet Respiratory Medicine. 2018;6:379-88.
79. Rodríguez E, Baurecht H, Wahn A, Kretschmer A, Hotze M. An integrated epigenetic and transcriptomic analysis reveals distinct tissue-specific patterns of DNA methylation associated with atopic dermatitis. Journal of Investigative Dermatology. 2014;134(7):1873-83.
80. Quraishi B, Zhang H, Everson T, Ray M, Lockett G. Identifying CpG sites associated with eczema via random forest screening of epigenome-scale DNA methylation. Clinical Epigenetics. 2015;7:68.
81. CHECKLIST FOR PAPERS ON THE AVON LONGITUDINAL STUDY OF PARENTS AND CHILDREN (ALSPAC). wwwbristolacuk/media-library/sites/alspac/documents/researchers/ALSPAC-publications-checklist.
82. Sabounchi S, Bollyky J, Nadeau K. Review of Environmental Impact on the Epigenetic Regulation of Atopic Diseases. Current Allergy and Asthma Reports. 2015;15:33.
83. Boyd A, Golding J, Macleod J, Lawlor DA, Fraser A, Henderson J, et al. Cohort Profile: The 'Children of the 90s'-the index offspring of the Avon Longitudinal Study of Parents and Children. International Journal of Epidemiology. 2013;42(1):111-27.
84. Fraser A, Macdonald-Wallis C, Tilling K, Boyd A, Golding J, Smith GD, et al. Cohort Profile: The Avon Longitudinal Study of Parents and Children: ALSPAC mothers cohort. International Journal of Epidemiology. 2013;42(1):97-110.
85. Accessible Resource for Integrated Epigenomics Studies (ARIES). wwwbrisacuk/social-community-medicine/people/project.
86. Min J, Hemani G, Davey Smith G. Meffil: efficient normalisation and analysis of very large DNA methylation samples. Bioinformatics. 2018.
87. Houseman EA, Accomando WP, Koestler DC, Christensen BC. DNA methylation arrays as surrogate measures of cell mixture distribution. BMC Bioinformatics. 2012.
88. Bakalski KM, Feinberg JI, Andrews SV, Yang J. DNA methylation of cord blood cell types: Applications for mixed cell birth studies. Epigenetics. 2016;11:354-62.
89. Granell R, Sterne JA, Henderson J. Associations of different phenotypes of wheezing illness in early childhood with environmental variables implicated in the aetiology of asthma. PLOS ONE. 2012;7:e48359.
90. Ferreira MA, Vonk JM, Baurecht H. Shared genetic origin of asthma, hay fever and eczema elucidates allergic disease biology. Nature Genetics. 2017;49:1752-7.
91. Hamid F, Wahyuni S, van Leeuwen A, van Ree R. Allergic disorders and socio-economic status: a study of schoolchildren in an urban area of Makassar, Indonesia. Clinical & Experimental Allergy. 2015;45:1226-36.
92. Petherick E, Pearce N, Sunyer J. Ethnic and socio-economic differences in the prevalence of wheeze, severe wheeze, asthma, eczema and medication usage at 4 years of age: Findings from the Born in Bradford birth cohort. Respiratory Medicine. 2016;119:122-9.

93. Mei-Yen Yong A, Tay YK. Atopic Dermatitis: Racial and Ethnic Differences. Dermatologic Clinics. 2017;35:395-402.
94. Kaufman B, Guttman-Yassky E, Alexis A. Atopic Dermatitis in Diverse Racial and Ethnic Groups - Variations in Epidemiology, Genetics, Clinical Presentation, and Treatment. Experimental Dermatology. 2018;27:340-57.
95. www.genecards.org.
96. www.mqtldb.org.
97. Joehanes R, Just AC, Marioni RE, Pilling LC. Epigenetic Signatures of Cigarette Smoking. Circulation: Cardiovascular Genetics. 2016;9:436-47.
98. Joubert BR, Felix JF, Yousefi P, Bakulski KM. DNA Methylation in Newborns and Maternal Smoking in Pregnancy: Genome-wide Consortium Meta-analysis. The American Journal of Human Genetics. 2016;98:680-96.
99. Hartwig FP, Davey Smith G, Simpkin AJ. Association between breastfeeding and DNA methylatio over the life course: findings from the Avon Longitudinal Study of Parents and Children (ALSPAC). 2018;Unpublished.
100. Baron RM, Kenny DA. The Moderator-Mediator Variable Distinction in Social Psychological Research – Conceptual, Strategic, and Statistical Considerations. Journal of Personality and Social Psychology. 1986;51:1173-82.
101. Richmond RC, Hemani G, Tilling K, Davey Smith G, Relton CL. Challenges and novel approaches for investigating molecular mediation. Human Molecular Genetics. 2016;25:R149-56.
102. Markunas CA, Xu Z, Harlid S, Wade PA. Identification of DNA methylation changes in newborns related to maternal smoking during pregnancy. Environmental Health Perspectives. 2014;122:1147-53.
103. Teschendorff AE, Yang Z, Wong A, Pipinikas CP. Correlation of Smoking-Associated DNA Methylation Changes in Buccal Cells With DNA Methylation Changes in Epithelial Cancer. JAMA Oncology. 2015;1:476-85.
104. Tsaprouni LG, Yang TP, Bell J, Dick KJ. Cigarette smoking reduces DNA methylation levels at multiple genomic loci but the effect is partially reversible upon cessation. Epigenetics. 2014;9:1382-96.
105. Hartwig FP, Loret de Mola C, Davies NM, Victora CG. Breastfeeding effects on DNA methylation in the offspring: A systematic literature review. PLOS ONE. 2017;12:e0173070.

APPENDICES

Appendix I: Extracting questions

Table A1: 13 questions used in analysis relating to eczema drawn from ALSPAC questionnaires

	Questionnaire /clinic	Age (months)	Question code	Question	Answer
1	CHILD BASED	81	KQ035	In the last year, eczema?	1-Yes, saw Dr 2-Yes, no Dr 3-No
2	CHILD BASED	91	KR042	In the past 12 months: Eczema?	1-Yes, saw Dr 2-Yes, no Dr 3-No
3	CHILD BASED	103	KS1042	In the past 12 months: Eczema?	1-Yes, saw Dr 2-Yes, no Dr 3-No
4	CHILD BASED	128	KV1060	In the past 12 months: Eczema?	1-Yes, saw Dr 2-Yes, no Dr 3-No
5	CHILD BASED	157	TA1030	Has he had any of the following in the past 12 months? Eczema?	1-Yes, saw Dr 2-Yes, no Dr 3-No
6	CHILD BASED	166	TB1060	In the past 12 months: Eczema?	1-Yes, saw Dr 2-Yes, no Dr 3-No
7	CHILD COMPLETED	192	CCS5023	In the past 12 months: Eczema?	1-Yes, saw Dr 2-Yes, no Dr 3-No
8	CHILD BASED	128	KV1070	Has a doctor ever actually said that your study child has asthma or eczema?	1-Yes asthma 2-Yes eczema 3-Yes asthma and eczema 4-No
9	CHILD BASED	166	TB1070	Has a doctor ever actually said that your study child has asthma or eczema?	1-Yes asthma 2-Yes eczema 3-Yes asthma and eczema 4-No
10	CHILD BASED	128	KV1122	Has she ever had eczema?	1-Yes 2-No
11	CHILD BASED	166	TB1122	Has she ever had eczema?	1-Yes 2-No
12	CHILD BASED	198	TC6110	Has she ever had eczema?	1-Yes 2-No
13	CHILD COMPLETED	216	CCT4055	Has she ever had eczema?	1-Yes 2-No

Table A2: Three questions used in analysis relating to smoking during pregnancy pulled from ALSPAC questionnaires

	Questionnaire/ clinic	Age (months)	Question code	Question	Answer
1	MOTHER	18 weeks gest	b650	Have you ever been a smoker?	1-Yes 2-No
2	MOTHER	18 weeks gest	b659	Have you now stopped smoking?	1-Yes 2-No
3	MOTHER	18 weeks gest	b665	Did you smoke regularly at any of the following times in the last 9 months? First 3 months of pregnancy	1-No 2-Yes, cigarettes 3-Yes, cigars 4-Yes, pipe 5-Yes, other

Table A3: Two questions used in analysis relating to a smoky environment pulled from ALSPAC questionnaires

	Questionnaire/ clinic	Age (months)	Question code	Question	Answer
4	CHILD B ASED	6	kb548	Please indicate how often during the day the baby is in a room or enclosed place where people are smoking: Weekdays?	1-all the time 2-more than 5 hours 3-3-5 hours 4-1-2 hours 5-less than 1 hour 6-not at all
5	CHILD B ASED	6	kb550	Please indicate how often during the day the baby is in a room or enclosed place where people are smoking: Weekends?	1-all the time 2-more than 5 hours 3-3-5 hours 4-1-2 hours 5-less than 1 hour 6-not at all

Table A4: One question used in analysis relating to cat exposure pulled from ALSPAC questionnaires

	Questionnaire/ clinic	Age (months)	Question code	Question	Answer
1	CHILD BASED	15	kc370	Which pets is he in contact with at least once a week either in your home or elsewhere? Cat	1-Yes 2-No

Table A5: One question used in analysis relating to dog exposure pulled from ALSPAC questionnaires

	Questionnaire/ clinic	Age (months)	Question code	Question	Answer
2	CHILD BASED	15	kc371	Which pets is he in contact with at least once a week either in your home or elsewhere? Dog	1-Yes 2-No

Table A6: Seven questions used in analysis relating to breastfeeding pulled from ALSPAC questionnaires (table spread over two pages)

	Questionnaire/ clinic	Age (months)	Question code	Question	Answer
1	CHILD BASED	4 weeks	ka030	How have you fed your baby since he was born? First 24 hours	1-Breast only 2-Bottle only 3-Breast&bottle 4-Other
2	CHILD BASED	4 weeks	ka031	How have you fed your baby since he was born? Rest of 1st week	1-Breast only 2-Bottle only 3-Breast&bottle 4-Other
3	CHILD BASED	4 weeks	ka032	How have you fed your baby since he was born? 2nd week	1-Breast only 2-Bottle only 3-Breast&bottle 4-Other
4	CHILD BASED	4 weeks	ka033	How have you fed your baby since he was born? 3rd week	1-Breast only 2-Bottle only 3-Breast&bottle 4-Other

5	CHILD BASED	4 weeks	ka034	How have you fed your baby since he was born? 4th week	1-Breast only 2-Bottle only 3-Breast&bottle 4-Other
6	CHILD BASED	4 weeks	ka061	How is your baby being fed at the moment?	1-Breast only 2-Bottle only 3-Breast&bottle 4-Other
7	CHILD BASED	4 weeks	ka094	How often is your baby fed in the following ways: breast fed	1-Always 2-Often 3-Sometimes 4-Never 9-Don’t know

Appendix II: STATA code

Eczema do file

********** EXTRACTING DATA FROM ALSPAC **********

*** Syntax template for direct users preparing datasets using child and adult based datasets.

* This version created 29th October 2014 - always create a datafile using the most up to date template.

* This template is based on that used by the data buddy team and they include a number of variables by default.

* To ensure the file works we suggest you keep those in and just add any relevant variables that you need for your project.

****

* To add data other than that included by default you will need to add the relevant files and pathnames in each of the match commands below.

* There is a separate command for mothers, partner, mothers providing data on the child and data provided by the child themselves.

* each has different withdrawal of consent issues so they must be considered separately.

* You will need to replace 'YOUR PATHNAME' in each section with your working directory pathname.

****

* Child BASED files - in this section the following files need to be placed:

* Mother completed Qs about YP

* ALWAYS KEEP THIS SECTION EVEN IF ONLY CHILD COMPLETED REQUESTED, although you will need to remove the *****

********** CREATING 'CHILD BASED' DATA SET **********

```
set maxvar 20000
use "\\ads.bris.ac.uk\filestore\SSCM ALSPAC\Data\Current\Other\Sample Definition\kz_5b.dta", clear
sort aln qlet
gen in_kz=1
merge 1:1 aln qlet using "\\ads.bris.ac.uk\filestore\SSCM ALSPAC\Data\Current\Other\cohort profile\cp_r1a.dta", nogen
merge 1:1 aln qlet using "\\ads.bris.ac.uk\filestore\SSCM ALSPAC\Data\Current\Quest\Child Based\ka_4c.dta", nogen
merge 1:1 aln qlet using "\\ads.bris.ac.uk\filestore\SSCM ALSPAC\Data\Current\Quest\Child Based\kb_6d.dta", nogen
merge 1:1 aln qlet using "\\ads.bris.ac.uk\filestore\SSCM ALSPAC\Data\Current\Quest\Child Based\kd_4b.dta", nogen
merge 1:1 aln qlet using "\\ads.bris.ac.uk\filestore\SSCM ALSPAC\Data\Current\Quest\Child Based\kf_7c.dta", nogen
merge 1:1 aln qlet using "\\ads.bris.ac.uk\filestore\SSCM ALSPAC\Data\Current\Quest\Child Based\kj_6c.dta", nogen
merge 1:1 aln qlet using "\\ads.bris.ac.uk\filestore\SSCM ALSPAC\Data\Current\Quest\Child Based\kk_2c.dta", nogen
merge 1:1 aln qlet using "\\ads.bris.ac.uk\filestore\SSCM ALSPAC\Data\Current\Quest\Child Based\kl_r1b.dta", nogen
merge 1:1 aln qlet using "\\ads.bris.ac.uk\filestore\SSCM ALSPAC\Data\Current\Quest\Child Based\km_r1b.dta", nogen
merge 1:1 aln qlet using "\\ads.bris.ac.uk\filestore\SSCM ALSPAC\Data\Current\Quest\Child Based\kn_r1b.dta", nogen
merge 1:1 aln qlet using "\\ads.bris.ac.uk\filestore\SSCM ALSPAC\Data\Current\Quest\Child Based\kq_3a.dta", nogen
merge 1:1 aln qlet using "\\ads.bris.ac.uk\filestore\SSCM ALSPAC\Data\Current\Quest\Child Based\kr_2a.dta", nogen
merge 1:1 aln qlet using "\\ads.bris.ac.uk\filestore\SSCM ALSPAC\Data\Current\Quest\Child Based\ks_r1b.dta", nogen
merge 1:1 aln qlet using "\\ads.bris.ac.uk\filestore\SSCM ALSPAC\Data\Current\Quest\Child Based\kv_2a.dta", nogen
merge 1:1 aln qlet using "\\ads.bris.ac.uk\filestore\SSCM ALSPAC\Data\Current\Quest\Child Based\kw_r2b.dta", nogen
merge 1:1 aln qlet using "\\ads.bris.ac.uk\filestore\SSCM ALSPAC\Data\Current\Quest\Child Based\ta_2a.dta", nogen
merge 1:1 aln qlet using "\\ads.bris.ac.uk\filestore\SSCM ALSPAC\Data\Current\Quest\Child Based\tb_2a.dta", nogen
```

merge 1:1 aln qlet using "\\ads.bris.ac.uk\filestore\SSCM ALSPAC\Data\Current\Quest\Child Based\tc_2a.dta", nogen

keep aln qlet kz011b kz021 kz030 /// KZ

in_core in_alsp in_phase2 in_phase3 tripquad /// Cohort profile

ka249 ka250 ka251 ka252 ka253 ka254 ka255 ka256 ka257 ka263 /// KA

kb051 kb052 kb086 kb087 kb088 kb089 kb090 kb091 kb092 kb093 kb094 /// KB

kd050a kd086 kd087 kd088 kd089 kd090 kd085 kd050b /// KD

kf060 kf110 kf111 kf112 kf113 kf114 kf115 kf116 kf118 kf119 /// KF

kj040 kj100 kj101 kj102 kj104 kj105 kj106 kj107 kj108 kj109 /// KJ

kk265 kk268 /// KK

kl030 kl100 kl101 kl102 kl103 kl104 kl110 kl111 /// KL

km2203 km2241 /// KM

kn1030 kn1120 kn1121 kn1122 kn1123 kn1124 kn1130 kn1131 kn1133 kn1140 /// KN

kq023 kq035 kq090 kq091 kq092 kq093 kq094 kq100 kq101 kq102 kq103 kq105 kq195 kq198 kq228 kq234 /// KQ

kr030 kr042 kr076 /// KR

ks1030 ks1042 ks1170 ks1280 ks1281 ks1282 ks1283 ks1284 ks1290 ks1291 ks1293 ks1294 ks1300 ks1303 ks3000 ks3003 ks3050 ks3053 ///KS

kv1049 kv1060 kv1070 kv1110 kv1111 kv1112 kv1113 kv1114 kv1115 kv1116 kv1120 kv1121 kv1122 kv1170 kv2210 /// KV

kw1080 kw1280 kw1281 kw1282 kw1283 kw1284 kw1285 kw1290 kw1291 kw1293 kw1294 kw5210 kw5213 /// KW

ta1019 ta1030 ta1113 ta4100 ta4160 ta4170 ta4180 /// TA

tb1049 tb1060 tb1070 tb1110 tb1111 tb1112 tb1113 tb1114 tb1115 tb1116 tb1120 tb1121 tb1122 tb1170 tb2213 /// TB

tc6100 tc6101 tc6102 tc6103 tc6104 tc6105 tc6106 tc6110 // TC

* Dealing with withdrawal of consent: For this to work additional variables required have to be inserted before in_core, so replace the ***** line with additional variables.

* If none are required remember to delete the ***** line.

* An additional do file is called in to set those withdrawing consent to missing so that this is always up to date whenever you run this do file

order aln qlet ka250, first

order in_alsp tripquad, last

do "\\ads.bris.ac.uk\filestore\SSCM ALSPAC\Data\Syntax\Withdrawal of consent\child_based_WoC.do"

save "M:projects\ieu2\p5\002\working\data\results\PHD\2ND YEAR\ECZEMA VARIABLE\DATA SETS\childB.dta", replace

********** CREATING 'CHILD COMPLETED' DATA SET **********

* Child COMPLETED files - in this section the following files need to be placed:

* YP completed Qs

* Puberty Qs

* Child clinic data

* Child biosamples data

* If there are no child completed files, this section can be starred out.

* NOTE: having to keep kz021 tripquad just to make the withdrawal of consent work - these are dropped for this file as the ones in the child BASED file are the important ones and should take priority

```
use "\\ads.bris.ac.uk\filestore\SSCM ALSPAC\Data\Current\Other\Sample Definition\kz_5b.dta", clear
sort aln qlet
merge 1:1 aln qlet using "\\ads.bris.ac.uk\filestore\SSCM ALSPAC\Data\Current\Other\cohort profile\cp_r1a.dta", nogen
merge 1:1 aln qlet using "\\ads.bris.ac.uk\filestore\SSCM ALSPAC\Data\Current\Quest\Child Completed\ccg_r1c.dta", nogen
merge 1:1 aln qlet using "\\ads.bris.ac.uk\filestore\SSCM ALSPAC\Data\Current\Quest\Child Completed\cck_r1b.dta", nogen
merge 1:1 aln qlet using "\\ads.bris.ac.uk\filestore\SSCM ALSPAC\Data\Current\Quest\Child Completed\ccs_r1b.dta", nogen
merge 1:1 aln qlet using "\\ads.bris.ac.uk\filestore\SSCM ALSPAC\Data\Current\Quest\Child Completed\cct_1a.dta", nogen
merge 1:1 aln qlet using "\\ads.bris.ac.uk\filestore\SSCM ALSPAC\Data\Current\Clinic\Child\cif_7a.dta", nogen
merge 1:1 aln qlet using "\\ads.bris.ac.uk\filestore\SSCM ALSPAC\Data\Current\Clinic\Child\tf1_r2a.dta", nogen
merge 1:1 aln qlet using "\\ads.bris.ac.uk\filestore\SSCM ALSPAC\Data\Current\Clinic\Child\tf2_r2a.dta", nogen
keep aln qlet kz021 ///
                        ccg260 ccg261 ccg262 ccg263 ccg266 ccg290 ccg320 /// CCG
                        cck170 /// CCK
                        ccs5023 ccs5080 ccs5081 ccs5082 ccs5083 ccs5084 ccs5085 ccs5086 ccs5087 ccs5088 ccs5090 ccs5091 /// CCS
                        cct4055 cct4056 cct4057 cct4058 /// CCT
                        cf240 cf280 /// CIF
                        ff2060 ff2061 ff2062 ff2063 ff2064 ff2065 ff2066 ff2067 /// TF1
```

fg3160 fg3162 fg3163 fg3164 fg3165 fg3166 /// TF2

tripquad //

* Dealing with withdrawal of consent: For this to work additional variables required have to be inserted before tripquad, so replace the ***** line with additional variables.

* An additional do file is called in to set those withdrawing consent to missing so that this is always up to date whenever you run this do file

order aln qlet kz021, first

order tripquad, last

do "\\ads.bris.ac.uk\filestore\SSCM ALSPAC\Data\Syntax\Withdrawal of consent\child_completed_WoC.do"

drop kz021 tripquad

save "M:projects\ieu2\p5\002\working\data\results\PHD\2ND YEAR\ECZEMA VARIABLE\DATA SETS\childC.dta", replace

********** MERGING THE TWO DATA SETS TO MAKE ONE OVERALL DATA SET **********

** Matching all data together and saving out the final file*.

* NOTE: any linkage data should be added here*.

```
use "M:projects\ieu2\p5\002\working\data\results\PHD\2ND YEAR\ECZEMA VARIABLE\DATA SETS\childB.dta", clear
merge 1:1 aln qlet using "M:projects\ieu2\p5\002\working\data\results\PHD\2ND YEAR\ECZEMA VARIABLE\DATA SETS\childC.dta", nogen
* IF partner data is required please unstar the following line
/* merge m:1 aln using "YOUR PATHWAY\partner.dta", nogen */

* Remove non-alspac children.
drop if in_alsp!=1.

* Remove trips and quads.
drop if tripquad==1

drop in_alsp tripquad
save "M:projects\ieu2\p5\002\working\data\results\PHD\2ND YEAR\ECZEMA VARIABLE\DATA SETS\Data_extracted_from_ALSPAC.dta", replace

*************************************************************************************************************************

* QC checks*
use "M:projects\ieu2\p5\002\working\data\results\PHD\2ND YEAR\ECZEMA VARIABLE\DATA SETS\Data_extracted_from_ALSPAC.dta", clear

* Check that there are 15445 records.
count
```

********** USE/OPEN THE DATA SET **********

use "M:projects\ieu2\p5\002\working\data\results\PHD\2ND YEAR\ECZEMA VARIABLE\DATA SETS\Data_extracted_from_ALSPAC.dta", clear

********** REPLACING MINUS NUMBERS BY 'MISSING' **********

* To count how many variables are present

count if ka249 > 0 & ka249 !=.

* Replacing minus numbers by the . symbol

replace ccs5023 =. if ccs5023<0

* Replacing all the minus numbers by the . symbol

* Create a variable of all the 200 variables, then do the function on each 'var'

local vars="aln kz011b kz021 kz030 in_core in_phase2 in_phase3 ka249 ka250 ka251 ka252 ka253 ka254 ka255 ka256 ka257 ka263 kb051 kb052 kb086 kb087 kb088 kb089 kb090 kb091 kb092 kb093 kb094 kd050a kd086 kd087 kd088 kd089 kd090 kd085 kd050b kf060 kf110 kf111 kf112 kf113 kf114 kf115 kf116 kf118 kf119 kj040 kj100 kj101 kj102 kj104 kj105 kj106 kj107 kj108 kj109 kk265 kk268 kl030 kl100 kl101 kl102 kl103 kl104 kl110 kl111 km2203 km2241 kn1030 kn1120 kn1121 kn1122 kn1123 kn1124 kn1130 kn1131 kn1133 kn1140 kq023 kq035 kq090 kq091 kq092 kq093 kq094 kq100 kq101 kq102 kq103 kq105 kq195 kq198 kq228 kq234 kr030 kr042 kr076 ks1030 ks1042 ks1170 ks1280 ks1281 ks1282 ks1283 ks1284 ks1290 ks1291 ks1293 ks1294 ks1300 ks1303 ks3000 ks3003 ks3050 ks3053 kv1049 kv1060 kv1070 kv1110 kv1111 kv1112 kv1113 kv1114 kv1115 kv1116 kv1120 kv1121 kv1122 kv1170 kv2210 kw1080 kw1280 kw1281 kw1282 kw1283 kw1284 kw1285 kw1290 kw1291 kw1293 kw1294 kw5210 kw5213 ta1019 ta1030 ta1113 ta4100 ta4160 ta4170 ta4180 tb1049 tb1060 tb1070 tb1110 tb1111 tb1112 tb1113 tb1114 tb1115 tb1116 tb1120 tb1121 tb1122 tb1170 tb2213 tc6100 tc6101 tc6102 tc6103 tc6104 tc6105 tc6106 tc6110 ccg260 ccg261 ccg262 ccg263 ccg266 ccg290 ccg320 cck170 ccs5023 ccs5080 ccs5081 ccs5082 ccs5083 ccs5084 ccs5085 ccs5086 ccs5087 ccs5088 ccs5090 ccs5091 cct4055 cct4056 cct4057 cct4058 cf240 cf280 ff2060 ff2061 ff2062 ff2063 ff2064 ff2065 ff2066 ff2067 fg3160 fg3162 fg3163 fg3164 fg3165 fg3166"

foreach var in `vars' {

```
        replace `var'=. if `var'<0
}
********************************************** EVER_ECZEMA DEFINITIONS **********************************************
**************** CASES ***************
generate ever_eczema = 1 if kq035 == 1 | kq035 == 2 | kr042 == 1 | kr042 == 2 | ks1042 == 1 | ks1042 == 2 | kv1060 == 1 | kv1060 == 2 | ta1030 == 1 |
ta1030 == 2 | tb1060 == 1 | tb1060 == 2 | ccs5023 == 1 | ccs5023 == 2 ///
| kv1070 == 2 | kv1070 == 3 | tb1070 == 2 | tb1070 == 3 ///
| kv1122 == 1 | tb1122 == 1 | tc6110 == 1 | cct4055 ==

**************** CONTROL 1 - NO TO EVERYTHING ***************
generate ever_eczema_no_to_all = 0 if kq035 == 3 & kr042 == 3 & ks1042 == 3 & kv1060 == 3 & ta1030 == 3 & tb1060 == 3 & ccs5023 == 3 ///
& (kv1070 == 1 | kv1070 == 4) & (tb1070 == 1 | tb1070 == 4) ///
& kv1122 == 2 & tb1122 == 2 & tc6110 == 2 & cct4055 == 2

**************** CONTROL 2 - NO TO EVERYTHING, 1 MISSING ***************
generate ever_eczema_one_missing = 0 if kq035 == 3 & kr042 == 3 & ks1042 == 3 & kv1060 == 3 & ta1030 == 3 & tb1060 == 3 & ccs5023 == 3 ///
& (kv1070 == 1 | kv1070 == 4) & (tb1070 == 1 | tb1070 == 4) ///
& kv1122 == 2 & tb1122 == 2 & tc6110 == 2 & cct4055 == 2

replace ever_eczema_one_missing = 0 if kq035 == . & kr042 == 3 & ks1042 == 3 & kv1060 == 3 & ta1030 == 3 & tb1060 == 3 & ccs5023 == 3 ///
& (kv1070 == 1 | kv1070 == 4) & (tb1070 == 1 | tb1070 == 4) ///
```

```
& kv1122 == 2 & tb1122 == 2 & tc6110 == 2 & cct4055 == 2

replace ever_eczema_one_missing = 0 if kq035 == 3 & kr042 == . & ks1042 == 3 & kv1060 == 3 & ta1030 == 3 & tb1060 == 3 & ccs5023 == 3 ///
& (kv1070 == 1 | kv1070 == 4) & (tb1070 == 1 | tb1070 == 4) ///
& kv1122 == 2 & tb1122 == 2 & tc6110 == 2 & cct4055 == 2

replace ever_eczema_one_missing = 0 if kq035 == 3 & kr042 == 3 & ks1042 == . & kv1060 == 3 & ta1030 == 3 & tb1060 == 3 & ccs5023 == 3 ///
& (kv1070 == 1 | kv1070 == 4) & (tb1070 == 1 | tb1070 == 4) ///
& kv1122 == 2 & tb1122 == 2 & tc6110 == 2 & cct4055 == 2

replace ever_eczema_one_missing = 0 if kq035 == 3 & kr042 == 3 & ks1042 == 3 & kv1060 == . & ta1030 == 3 & tb1060 == 3 & ccs5023 == 3 ///
& (kv1070 == 1 | kv1070 == 4) & (tb1070 == 1 | tb1070 == 4) ///
& kv1122 == 2 & tb1122 == 2 & tc6110 == 2 & cct4055 == 2

replace ever_eczema_one_missing = 0 if kq035 == 3 & kr042 == 3 & ks1042 == 3 & kv1060 == 3 & ta1030 == . & tb1060 == 3 & ccs5023 == 3 ///
& (kv1070 == 1 | kv1070 == 4) & (tb1070 == 1 | tb1070 == 4) ///
& kv1122 == 2 & tb1122 == 2 & tc6110 == 2 & cct4055 == 2

replace ever_eczema_one_missing = 0 if kq035 == 3 & kr042 == 3 & ks1042 == 3 & kv1060 == 3 & ta1030 == 3 & tb1060 == . & ccs5023 == 3 ///
& (kv1070 == 1 | kv1070 == 4) & (tb1070 == 1 | tb1070 == 4) ///
& kv1122 == 2 & tb1122 == 2 & tc6110 == 2 & cct4055 == 2
```

```
replace ever_eczema_one_missing = 0 if kq035 == 3 & kr042 == 3 & ks1042 == 3 & kv1060 == 3 & ta1030 == 3 & tb1060 == 3 & ccs5023 == . ///
& (kv1070 == 1 | kv1070 == 4) & (tb1070 == 1 | tb1070 == 4) ///
& kv1122 == 2 & tb1122 == 2 & tc6110 == 2 & cct4055 == 2

replace ever_eczema_one_missing = 0 if kq035 == 3 & kr042 == 3 & ks1042 == 3 & kv1060 == 3 & ta1030 == 3 & tb1060 == 3 & ccs5023 == 3 ///
& (kv1070 == .) & (tb1070 == 1 | tb1070 == 4) ///
& kv1122 == 2 & tb1122 == 2 & tc6110 == 2 & cct4055 == 2

replace ever_eczema_one_missing = 0 if kq035 == 3 & kr042 == 3 & ks1042 == 3 & kv1060 == 3 & ta1030 == 3 & tb1060 == 3 & ccs5023 == . ///
& (kv1070 == 1 | kv1070 == 4) & (tb1070 == .) ///
& kv1122 == 2 & tb1122 == 2 & tc6110 == 2 & cct4055 == 2

replace ever_eczema_one_missing = 0 if kq035 == 3 & kr042 == 3 & ks1042 == 3 & kv1060 == 3 & ta1030 == 3 & tb1060 == 3 & ccs5023 == . ///
& (kv1070 == 1 | kv1070 == 4) & (tb1070 == 1 | tb1070 == 4) ///
& kv1122 == . & tb1122 == 2 & tc6110 == 2 & cct4055 == 2

replace ever_eczema_one_missing = 0 if kq035 == 3 & kr042 == 3 & ks1042 == 3 & kv1060 == 3 & ta1030 == 3 & tb1060 == 3 & ccs5023 == . ///
& (kv1070 == 1 | kv1070 == 4) & (tb1070 == 1 | tb1070 == 4) ///
& kv1122 == 2 & tb1122 == . & tc6110 == 2 & cct4055 == 2
```

```
replace ever_eczema_one_missing = 0 if kq035 == 3 & kr042 == 3 & ks1042 == 3 & kv1060 == 3 & ta1030 == 3 & tb1060 == 3 & ccs5023 == . ///
& (kv1070 == 1 | kv1070 == 4) & (tb1070 == 1 | tb1070 == 4) ///
& kv1122 == 2 & tb1122 == 2 & tc6110 == . & cct4055 == 2

replace ever_eczema_one_missing = 0 if kq035 == 3 & kr042 == 3 & ks1042 == 3 & kv1060 == 3 & ta1030 == 3 & tb1060 == 3 & ccs5023 == . ///
& (kv1070 == 1 | kv1070 == 4) & (tb1070 == 1 | tb1070 == 4) ///
& kv1122 == 2 & tb1122 == 2 & tc6110 == 2 & cct4055 == .

********** CONTROL 3 - NO TO EVERYTHING, ANY MISSING **********

generate ever_eczema_any_missing = 0 if (kq035 == 3 | kq035 == .) & (kr042 == 3 | kr042 == .) & (ks1042 == 3 | ks1042 == .) & (kv1060 == 3 | kv1060 == .)
& (ta1030 == 3 | ta1030 == .) & (tb1060 == 3 | tb1060 == .) & (ccs5023 == 3 | ccs5023 == .) ///
& (kv1070 == 1 | kv1070 == 4 | kv1070 == .) & (tb1070 == 1 | tb1070 == 4 | tb1070 == .) ///
& (kv1122 == 2 | kv1122 == .) & (tb1122 == 2 | tb1122 == .) & (tc6110 == 2 | tc6110 == .) & (cct4055 == 2 | cct4055 == .)
********** CONTROL 4 - IF NOT A CASE, AND ANSWERED 'NO' TO 'EVER HAD ECZEMA?' **********

replace ever_eczema = 0 if ever_eczema != 1 & (kv1122 == 2 | tb1122 == 2 | tc6110 == 2 | cct4055 == 2)

********** DR_DIAGNOSIS DEFINITIONS **********

********** CASES **********
```

generate dr_diagnosis = 1 if (kv1070 == 2 | kv1070 == 3) | (tb1070 == 2 | tb1070 == 3) | kq035 == 1 | kr042 == 1 | ks1042 == 1 | kv1060 == 1 | ta1030 == 1 | tb1060 == 1 | ccs5023 == 1

********** CONTROL 1 - NO TO EVERYTHING **********

generate dr_diagnosis_no_to_all = 0 if (kv1070 == 1 | kv1070 == 4) & (tb1070 == 1 | tb1070 == 4) & (kq035 == 2 | kq035 ==3) & (kr042 == 2 | kr042 == 3) & (ks1042 == 2 | ks1042 == 3) & (kv1060 == 2 | kv1060 == 3) & (ta1030 == 2 | ta1030 == 3) & (tb1060 == 2 | tb1060 == 3) & (ccs5023 == 2 | ccs5023 == 3)

********** CONTROL 2 - NO TO EVERYTHING, 1 MISSING **********

generate dr_diagnosis_one_missing = 0 if (kv1070 == 1 | kv1070 == 4) & (tb1070 == 1 | tb1070 == 4) & (kq035 == 2 | kq035 ==3) & (kr042 == 2 | kr042 == 3) & (ks1042 == 2 | ks1042 == 3) & (kv1060 == 2 | kv1060 == 3) & (ta1030 == 2 | ta1030 == 3) & (tb1060 == 2 | tb1060 == 3) & (ccs5023 == 2 | ccs5023 == 3)

replace dr_diagnosis_one_missing = 0 if (kv1070 ==.) & (tb1070 == 1 | tb1070 == 4) & (kq035 == 2 | kq035 ==3) & (kr042 == 2 | kr042 == 3) & (ks1042 == 2 | ks1042 == 3) & (kv1060 == 2 | kv1060 == 3) & (ta1030 == 2 | ta1030 == 3) & (tb1060 == 2 | tb1060 == 3) & (ccs5023 == 2 | ccs5023 == 3)

replace dr_diagnosis_one_missing = 0 if (kv1070 == 1 | kv1070 == 4) & (tb1070 ==.) & (kq035 == 2 | kq035 ==3) & (kr042 == 2 | kr042 == 3) & (ks1042 == 2 | ks1042 == 3) & (kv1060 == 2 | kv1060 == 3) & (ta1030 == 2 | ta1030 == 3) & (tb1060 == 2 | tb1060 == 3) & (ccs5023 == 2 | ccs5023 == 3)

replace dr_diagnosis_one_missing = 0 if (kv1070 == 1 | kv1070 == 4) & (tb1070 == 1 | tb1070 == 4) & (kq035 ==.) & (kr042 == 2 | kr042 == 3) & (ks1042 == 2 | ks1042 == 3) & (kv1060 == 2 | kv1060 == 3) & (ta1030 == 2 | ta1030 == 3) & (tb1060 == 2 | tb1060 == 3) & (ccs5023 == 2 | ccs5023 == 3)

replace dr_diagnosis_one_missing = 0 if (kv1070 == 1 | kv1070 == 4) & (tb1070 == 1 | tb1070 == 4) & (kq035 == 2 | kq035 ==3) & (kr042 ==.) & (ks1042 == 2 | ks1042 == 3) & (kv1060 == 2 | kv1060 == 3) & (ta1030 == 2 | ta1030 == 3) & (tb1060 == 2 | tb1060 == 3) & (ccs5023 == 2 | ccs5023 == 3)

replace dr_diagnosis_one_missing = 0 if (kv1070 == 1 | kv1070 == 4) & (tb1070 == 1 | tb1070 == 4) & (kq035 == 2 | kq035 ==3) & (kr042 == 2 | kr042 == 3) & (ks1042 ==.) & (kv1060 == 2 | kv1060 == 3) & (ta1030 == 2 | ta1030 == 3) & (tb1060 == 2 | tb1060 == 3) & (ccs5023 == 2 | ccs5023 == 3)

replace dr_diagnosis_one_missing = 0 if (kv1070 == 1 | kv1070 == 4) & (tb1070 == 1 | tb1070 == 4) & (kq035 == 2 | kq035 ==3) & (kr042 == 2 | kr042 == 3) & (ks1042 == 2 | ks1042 == 3) & (kv1060 ==.) & (ta1030 == 2 | ta1030 == 3) & (tb1060 == 2 | tb1060 == 3) & (ccs5023 == 2 | ccs5023 == 3)

replace dr_diagnosis_one_missing = 0 if (kv1070 == 1 | kv1070 == 4) & (tb1070 == 1 | tb1070 == 4) & (kq035 == 2 | kq035 ==3) & (kr042 == 2 | kr042 == 3) & (ks1042 == 2 | ks1042 == 3) & (kv1060 == 2 | kv1060 == 3) & (ta1030 ==.) & (tb1060 == 2 | tb1060 == 3) & (ccs5023 == 2 | ccs5023 == 3)

replace dr_diagnosis_one_missing = 0 if (kv1070 == 1 | kv1070 == 4) & (tb1070 == 1 | tb1070 == 4) & (kq035 == 2 | kq035 ==3) & (kr042 == 2 | kr042 == 3) & (ks1042 == 2 | ks1042 == 3) & (kv1060 == 2 | kv1060 == 3) & (ta1030 == 2 | ta1030 == 3) & (tb1060 ==.) & (ccs5023 == 2 | ccs5023 == 3)

replace dr_diagnosis_one_missing = 0 if (kv1070 == 1 | kv1070 == 4) & (tb1070 == 1 | tb1070 == 4) & (kq035 == 2 | kq035 ==3) & (kr042 == 2 | kr042 == 3) & (ks1042 == 2 | ks1042 == 3) & (kv1060 == 2 | kv1060 == 3) & (ta1030 == 2 | ta1030 == 3) & (tb1060 == 2 | tb1060 == 3) & (ccs5023 ==.)

********** CONTROL 3 - NO TO EVERYTHING, ANY MISSING **********

generate dr_diagnosis_any_missing = 0 if (kv1070 == 1 | kv1070 == 4) & (tb1070 == 1 | tb1070 == 4) & (kq035 == 2 | kq035 ==3) & (kr042 == 2 | kr042 == 3) & (ks1042 == 2 | ks1042 == 3) & (kv1060 == 2 | kv1060 == 3) & (ta1030 == 2 | ta1030 == 3) & (tb1060 == 2 | tb1060 == 3) & (ccs5023 == 2 | ccs5023 == 3)

replace dr_diagnosis_any_missing = 0 if (kv1070 == 1 | kv1070 == 4 | kv1070 ==.) & (tb1070 == 1 | tb1070 == 4 | tb1070 ==.) & (kq035 == 2 | kq035 ==3 | kq035 ==.) & (kr042 == 2 | kr042 == 3 | kr042 ==.) & (ks1042 == 2 | ks1042 == 3 | ks1042 ==.) & (kv1060 == 2 | kv1060 == 3 | kv1060 ==.) & (ta1030 == 2 | ta1030 == 3 | ta1030 ==.) & (tb1060 == 2 | tb1060 == 3 | tb1060 ==.) & (ccs5023 == 2 | ccs5023 == 3 | ccs5023 ==.)

```
********** CONTROL 4 - IF NOT A CASE, AND ANSWERED 'NO' TO 'EVER HAD ECZEMA?' **********

replace dr_diagnosis = 0 if dr_diagnosis != 1 & (tb1070 == 1 | tb1070 == 4)

********** SAVING THE 'EVER_ECZEMA' AND 'DR_DIAGNOSIS' DEFINITIONS **********
keep aln qlet ever_eczema ///
ever_eczema_no_to_all ///
ever_eczema_one_missing ///
ever_eczema_any_missing ///
dr_diagnosis ///
dr_diagnosis_no_to_all ///
dr_diagnosis_one_missing ///
dr_diagnosis_any_missing //
saveold "M:projects\ieu2\p5\002\working\data\results\PHD\2ND YEAR\ECZEMA VARIABLE\DATA SETS\final_eczema_definitions.dta"

********** LIMIT TO ARIES DATASET **********

* Merge the aln and qlet data from ALSPAC data set and ARIES data set
merge 1:1 aln qlet using "\\ads.bris.ac.uk\filestore\SSCM ALSPAC\Data\Useful_data\ARIES\YP_in_ARIES.dta"
```

```
* Get rid of any values that aren't in the ARIES dataset
drop if in_ARIES == .

* Drop the ARIES data set itself
drop in_ARIES

****************************************** TABULATION OF 'EVER_ECZEMA' AND 'DR_DIAGNOSIS' ******************************************
tab ever_eczema dr_diagnosis, column row
```

Smoking do file

```
********************************************** EXTRACTING DATA FROM ALSPAC **********************************************

*** Syntax template for direct users preparing datasets using child and adult based datasets.
* This version created 29th October 2014 - always create a datafile using the most up to date template.
* This template is based on that used by the data buddy team and they include a number of variables by default.
* To ensure the file works we suggest you keep those in and just add any relevant variables that you need for your project.

*****************************************************************************************************************************
```

```
* To add data other than that included by default you will need to add the relvant files and pathnames in each of the match commands below.
* There is a separate command for mothers, partner, mothers providing data on the child and data provided by the child themselves.
* each has different withdrawal of consent issues so they must be considered separately.
* You will need to replace 'YOUR PATHNAME' in each section with your working directory pathname.

********************************************** SMOKING ************************************************************

* Child BASED files - in this section the following files need to be placed:
* Mother completed Qs about YP

* ALWAYS KEEP THIS SECTION EVEN IF ONLY CHILD COMPLETED REQUESTED, although you will need to remove the *****

********************************************** CREATING 'CHILD BASED' DATA SET ************************************************

set maxvar 20000
use "\\ads.bris.ac.uk\filestore\SSCM ALSPAC\Data\Current\Other\Sample Definition\kz_5b.dta", clear
sort aln qlet
gen in_kz=1
merge 1:1 aln qlet using "\\ads.bris.ac.uk\filestore\SSCM ALSPAC\Data\Current\Other\cohort profile\cp_r1a.dta", nogen
merge 1:1 aln qlet using "\\ads.bris.ac.uk\filestore\SSCM ALSPAC\Data\Current\Quest\Child Based\kb_6d.dta", nogen
merge 1:1 aln qlet using "\\ads.bris.ac.uk\filestore\SSCM ALSPAC\Data\Current\Quest\Child Based\kc_5b.dta", nogen
```

```
merge 1:1 aln qlet using "\\ads.bris.ac.uk\filestore\SSCM ALSPAC\Data\Current\Quest\Child Based\ke_5c.dta", nogen
merge 1:1 aln qlet using "\\ads.bris.ac.uk\filestore\SSCM ALSPAC\Data\Current\Quest\Child Based\kg_4b.dta", nogen
merge 1:1 aln qlet using "\\ads.bris.ac.uk\filestore\SSCM ALSPAC\Data\Current\Quest\Child Based\kk_2c.dta", nogen
merge 1:1 aln qlet using "\\ads.bris.ac.uk\filestore\SSCM ALSPAC\Data\Current\Quest\Child Based\km_r1b.dta", nogen
merge 1:1 aln qlet using "\\ads.bris.ac.uk\filestore\SSCM ALSPAC\Data\Current\Quest\Child Based\kp_r1b.dta", nogen
keep aln qlet   kz011b kz021 kz030 /// KZ
in_core in_alsp in_phase2 in_phase3 tripquad /// Cohort profile
kb548 kb550 /// KB
kc360 kc361 /// KC
ke195 ke196 /// KE
kg172 kg173 /// KG
kk311a kk311b /// KK
km3030 km3031 /// KM
kp5090 kp5091 /// KP
tripquad //
```

* Dealing with withdrawal of consent: For this to work additional variables required have to be inserted before in_core, so replace the ***** line with additional variables.

* If none are required remember to delete the ***** line.

* An additional do file is called in to set those withdrawing consent to missing so that this is always up to date whenever you run this do file

```
order aln qlet kb548, first
order in_alsp tripquad, last
do "\\ads.bris.ac.uk\filestore\SSCM ALSPAC\Data\Syntax\Withdrawal of consent\child_based_WoC.do"
save "M:projects\ieu2\p5\002\working\data\results\PHD\2ND YEAR\RISK FACTOR VARIABLES\SMOKING\DATA SETS\childB.dta", replace

********************************************** CREATING 'MOTHER' DATA SET **********************************************

* MOTHER completed files
use "\\ads.bris.ac.uk\filestore\SSCM ALSPAC\Data\Current\Other\Sample Definition\mz_5a.dta", clear
sort aln
gen in_mz=1
merge 1:1 aln using "\\ads.bris.ac.uk\filestore\SSCM ALSPAC\Data\Current\Quest\Mother\a_3c.dta", nogen
merge 1:1 aln using "\\ads.bris.ac.uk\filestore\SSCM ALSPAC\Data\Current\Quest\Mother\b_4d.dta", nogen
merge 1:1 aln using "\\ads.bris.ac.uk\filestore\SSCM ALSPAC\Data\Current\Quest\Mother\c_7d.dta", nogen
merge 1:1 aln using "\\ads.bris.ac.uk\filestore\SSCM ALSPAC\Data\Current\Quest\Mother\e_4d.dta", nogen
merge 1:1 aln using "\\ads.bris.ac.uk\filestore\SSCM ALSPAC\Data\Current\Quest\Mother\f_2b.dta", nogen
merge 1:1 aln using "\\ads.bris.ac.uk\filestore\SSCM ALSPAC\Data\Current\Quest\Mother\g_5b.dta", nogen
merge 1:1 aln using "\\ads.bris.ac.uk\filestore\SSCM ALSPAC\Data\Current\Quest\Mother\h_6d.dta", nogen
merge 1:1 aln using "\\ads.bris.ac.uk\filestore\SSCM ALSPAC\Data\Current\Quest\Mother\j_5a.dta", nogen
merge 1:1 aln using "\\ads.bris.ac.uk\filestore\SSCM ALSPAC\Data\Current\Quest\Mother\k_r1b.dta", nogen
merge 1:1 aln using "\\ads.bris.ac.uk\filestore\SSCM ALSPAC\Data\Current\Quest\Mother\l_r1b.dta", nogen
```

```
merge 1:1 aln using "\\ads.bris.ac.uk\filestore\SSCM ALSPAC\Data\Useful_data\bestgest\bestgest.dta", nogen

keep aln mz001 mz010 mz010a mz013 mz014 mz028b ///
a200 a214 /// A
b522 b650 b651 b653 b654 b655 b656 b658 b659 b660 b663 b665 b667 b669 b670 b671 b679 b683 b685 b690 b695 /// B
c480 c481 c482 /// C
e170 e171 e172 e173 e174 e175 e176 e177 e178 e179 e181 e185 e186 e187 /// E
f560 f620 /// F
g515 g648 g820 /// G
h385 h525 h720 h845 h846 /// H
j368 j369 j630 j735 j736 j737 /// J
k5200 k6180 /// K
l5040 l5041 l5050 l5051 l6070 l6071 l6072 l6073 /// L
bestgest //

* Dealing with withdrawal of consent: For this to work additional variables required have to be inserted before tripquad, so replace the ***** line with additional variables.
* An additional do file is called in to set those withdrawing consent to missing so that this is always up to date whenever you run this do file
order aln mz010, first
order bestgest, last
do "\\ads.bris.ac.uk\filestore\SSCM ALSPAC\Data\Syntax\Withdrawal of consent\mother.do"
```

```
save "M:projects\ieu2\p5\002\working\data\results\PHD\2ND YEAR\RISK FACTOR VARIABLES\SMOKING\DATA SETS\mother.dta", replace

********** MERGING THE TWO DATA SETS TO MAKE ONE OVERALL DATA SET **********

** Matching all data together and saving out the final file*.

* NOTE: any linkage data should be added here*.

use "M:projects\ieu2\p5\002\working\data\results\PHD\2ND YEAR\RISK FACTOR VARIABLES\SMOKING\DATA SETS\childB.dta", clear

merge m:1 aln using "M:projects\ieu2\p5\002\working\data\results\PHD\2ND YEAR\RISK FACTOR VARIABLES\SMOKING\DATA SETS\mother.dta", nogen

* Remove non-alspac children.

drop if in_alsp!=1.

* Remove trips and quads.

drop if tripquad==1

drop in_alsp tripquad

save "M:projects\ieu2\p5\002\working\data\results\PHD\2ND YEAR\RISK FACTOR VARIABLES\SMOKING\DATA SETS\Data_extracted_from_ALSPAC-SMOKING.dta", replace

**********

* QC checks*

use "M:projects\ieu2\p5\002\working\data\results\PHD\2ND YEAR\RISK FACTOR VARIABLES\SMOKING\DATA SETS\Data_extracted_from_ALSPAC-SMOKING.dta", clear
```

```
* Check that there are 15445 records.
count
********************************************** USE/OPEN THE DATA SET **********************************************

use "M:projects\ieu2\p5\002\working\data\results\PHD\2ND YEAR\RISK FACTOR VARIABLES\SMOKING\DATA SETS\Data_extracted_from_ALSPAC-SMOKING.dta"

**********************************************REPLACING MINUS NUMBERS BY 'MISSING' ****************************************

* Replacing all the minus numbers by the . symbol
* Create a variable of all the 200 variables, then do the function on each 'var'
local vars="aln kz011b kz021 kz030 in_core in_phase2 in_phase3 kb548 kb550 b650 b659 b665"
foreach var in `vars' {
        replace `var'=. if `var'<0
}
********************************************** SMOKING DEFINITIONS **********************************************

*************** SMOKED DURING PREGNANCY ***************

******** YES SMOKED DURING PREGNANCY **********
```

```
generate smoked_during_pregnancy = 2 if (b650 == 1 & b659 == 2) | b665 == 2 | b665 == 3 | b665 == 4 | b665 == 5

******** YES BUT STOPPED BEFORE PREGNANCY **********

replace smoked_during_pregnancy = 1 if (b650 == 1 & b659 == 1)

******** NEVER SMOKED **********

replace smoked_during_pregnancy = 0 if smoked_during_pregnancy != 2 & smoked_during_pregnancy != 1

********** SMOKY ENVIRONMENT **********

******** YES SMOKY ENVIRONMENT *********

generate smoky_environment = 1 if kb548 == 1 | kb548 == 2 | kb548 == 3 | kb548 == 4 | kb550 == 1 | kb550 == 2 | kb550 == 3 | kb550 == 4

******** NOT SMOKY ENVIRONMENT *********

replace smoky_environment = 0 if smoky_environment != 1
```

```
************************************ CHANGING VARIABLE FROM CONTINUOUS TO CATEGORICAL ************************************

label define smoked_during_pregnancy_lb 0"didn't smoke" 1"smoked but stopped" 2"yes"

label values smoked_during_pregnancy smoked_during_pregnancy_lb

label define smoky_environment_lb 0"no" 1"yes"

label values smoky_environment smoky_environment_lb

************************************************ SAVING THE DEFINITIONS ************************************************

keep aln qlet smoked_during_pregnancy ///

smoky_environment //

save "M:projects\ieu2\p5\002\working\data\results\PHD\2ND YEAR\RISK FACTOR VARIABLES\SMOKING\DATA SETS\Real_data_extracted_from_ALSPAC-SMOKING.dta", replace

********************************************** MERGE TWO DATASETS **********************************************

use "M:projects\ieu2\p5\002\working\data\results\PHD\2ND YEAR\RISK FACTOR VARIABLES\SMOKING\DATA SETS\Real_data_extracted_from_ALSPAC-SMOKING.dta", clear

merge 1:1 aln qlet using "M:projects\ieu2\p5\002\working\data\results\PHD\2ND YEAR\ECZEMA VARIABLE\DATA SETS\final_eczema_definitions.dta", nogen
```

```
merge 1:1 aln qlet using "M:projects\ieu2\p5\002\working\data\results\PHD\2ND YEAR\RISK FACTOR VARIABLES\CONFOUNDING\SES\DATA SETS\motherchildB.dta", nogen

merge 1:1 aln qlet using "M:projects\ieu2\p5\002\working\data\results\PHD\2ND YEAR\RISK FACTOR VARIABLES\potential_early_risk_factors_RG.dta", nogen

************************************************** LIMIT TO ARIES DATASET **************************************************

* Merge the aln and qlet data from ALSPAC data set and ARIES data set

merge 1:1 aln qlet using "\\ads.bris.ac.uk\filestore\SSCM ALSPAC\Data\Useful_data\ARIES\YP_in_ARIES.dta"

* Get rid of any values that aren't in the ARIES dataset

drop if in_ARIES == .

* Drop the ARIES data set itself

drop in_ARIES

****************************************** TABULATION OF 'EVER_ECZEMA' AND 'SMOKING' ******************************************

tab ever_eczema smoked_during_pregnancy, column row

tab ever_eczema smoky_environment, column row

************************************** LOGISTIC REGRESSION - SMOKED DURING PREGNANCY **************************************

logistic ever_eczema i.smoked_during_pregnancy
```

********** LOGISTIC REGRESSION - SMOKY ENVIRONMENT **********

logistic ever_eczema smoky_environment

********** LOGISTIC REGRESSION WITH CONFOUNDING - SES **********

logistic ever_eczema i.smoked_during_pregnancy cmb_sc_grp

logistic ever_eczema smoky_environment cmb_sc_grp

********** LOGISTIC REGRESSION WITH 7 CONFOUNDERS **********

logistic ever_eczema i.smoked_during_pregnancy kz021 cmb_sc_grp ch_ethni mum_hist pet_cat pet_dog freq_breast6

logistic ever_eczema smoky_environment kz021 cmb_sc_grp ch_ethni mum_hist pet_cat pet_dog freq_breast6

Animal exposure do file

********** EXTRACTING DATA FROM ALSPAC **********

*** Syntax template for direct users preparing datasets using child and adult based datasets.

* This version created 29th October 2014 - always create a datafile using the most up to date template.

* This template is based on that used by the data buddy team and they include a number of variables by default.

* To ensure the file works we suggest you keep those in and just add any relevant variables that you need for your project.

****

* To add data other than that included by default you will need to add the relvant files and pathnames in each of the match commands below.

* There is a separate command for mothers, partner, mothers providing data on the child and data provided by the child themselves.

* each has different withdrawal of consent issues so they must be considered separately.

* You will need to replace 'YOUR PATHNAME' in each section with your working directory pathname.

********** ANIMALS **********

* Child BASED files - in this section the following files need to be placed:

* Mother completed Qs about YP

* ALWAYS KEEP THIS SECTION EVEN IF ONLY CHILD COMPLETED REQUESTED, although you will need to remove the *****

********** CREATING 'CHILD BASED' DATA SET **********

```
set maxvar 20000
use "\\ads.bris.ac.uk\filestore\SSCM ALSPAC\Data\Current\Other\Sample Definition\kz_5b.dta", clear
sort aln qlet
gen in_kz=1
```

```
merge 1:1 aln qlet using "\\ads.bris.ac.uk\filestore\SSCM ALSPAC\Data\Current\Other\cohort profile\cp_r1a.dta", nogen
merge 1:1 aln qlet using "\\ads.bris.ac.uk\filestore\SSCM ALSPAC\Data\Current\Quest\Child Based\kc_5b.dta", nogen
merge 1:1 aln qlet using "\\ads.bris.ac.uk\filestore\SSCM ALSPAC\Data\Current\Quest\Child Based\ke_5c.dta", nogen
merge 1:1 aln qlet using "\\ads.bris.ac.uk\filestore\SSCM ALSPAC\Data\Current\Quest\Child Based\kg_4b.dta", nogen
merge 1:1 aln qlet using "\\ads.bris.ac.uk\filestore\SSCM ALSPAC\Data\Current\Quest\Child Based\kk_2c.dta", nogen
merge 1:1 aln qlet using "\\ads.bris.ac.uk\filestore\SSCM ALSPAC\Data\Current\Quest\Child Based\km_r1b.dta", nogen
merge 1:1 aln qlet using "\\ads.bris.ac.uk\filestore\SSCM ALSPAC\Data\Current\Quest\Child Based\kp_r1b.dta", nogen
merge 1:1 aln qlet using "\\ads.bris.ac.uk\filestore\SSCM ALSPAC\Data\Current\Quest\Child Based\kq_3a.dta", nogen

keep aln qlet   kz011b kz021 kz030 /// KZ
in_core in_alsp in_phase2 in_phase3 tripquad /// Cohort profile
kc370 kc371 kc372 kc373 /// KC
ke200 ke201 ke202 ke203 /// KE
kg150 kg151 kg152 kg153 kg154 /// KG
kk285 kk287 kk288 kk300 kk301 kk302 kk303 kk304 /// KK
km2230 km2232 km2233 km3000 km3001 km3002 km3003 km3004 /// KM
kp5000 kp5001 kp5002 kp5003 /// KP
kq216 kq218 kq219 /// KQ
tripquad //
```

* Dealing with withdrawal of consent: For this to work additional variables required have to be inserted before in_core, so replace the ***** line with additional variables.

* If none are required remember to delete the ***** line.

* An additional do file is called in to set those withdrawing consent to missing so that this is always up to date whenever you run this do file

order aln qlet kc370, first

order in_alsp tripquad, last

do "\\ads.bris.ac.uk\filestore\SSCM ALSPAC\Data\Syntax\Withdrawal of consent\child_based_WoC_020715.do"

save "M:projects\ieu2\p5\002\working\data\results\PHD\2ND YEAR\RISK FACTOR VARIABLES\ANIMAL EXPOSURE\DATA SETS\childB.dta", replace

********** CREATING 'MOTHER' DATA SET **********

* MOTHER completed files

set maxvar 32767

use "\\ads.bris.ac.uk\filestore\SSCM ALSPAC\Data\Current\Other\Sample Definition\mz_5a.dta", clear

sort aln

gen in_mz=1

merge 1:1 aln using "\\ads.bris.ac.uk\filestore\SSCM ALSPAC\Data\Current\Quest\Mother\a_3c.dta", nogen

merge 1:1 aln using "\\ads.bris.ac.uk\filestore\SSCM ALSPAC\Data\Current\Quest\Mother\f_2b.dta", nogen

merge 1:1 aln using "\\ads.bris.ac.uk\filestore\SSCM ALSPAC\Data\Current\Quest\Mother\g_5b.dta", nogen

merge 1:1 aln using "\\ads.bris.ac.uk\filestore\SSCM ALSPAC\Data\Current\Quest\Mother\h_6b.dta", nogen

merge 1:1 aln using "\\ads.bris.ac.uk\filestore\SSCM ALSPAC\Data\Current\Quest\Mother\j_4b.dta", nogen

merge 1:1 aln using "\\ads.bris.ac.uk\filestore\SSCM ALSPAC\Data\Useful_data\bestgest\bestgest.dta", nogen

keep aln mz001 mz010 mz010a mz013 mz014 mz028b ///

a061 a062 a063 a064 a065 a066 a067 a070 a071 a072 a073 a074 a075 a076 a077 /// A

f380 f381 f382 f383 f384 f385 f386 f390 f391 f392 f393 f394 f395 f396 f397 f398 /// F

g515 g550 g551 g552 g553 g554 g556 g557 g558 g570 g571 g572 g573 g574 g575 g576 g577 g578 /// G

h442 h443 h444 h445 h446 h447 h448 h449 h450 h460 h461 h462 h463 h464 h465 h466 h467 h468 /// H

j395 j396 j397 j398 j399 j400 j401 j402 j403 j405 /// J

bestgest //

* Dealing with withdrawal of consent: For this to work additional variables required have to be inserted before tripquad, so replace the ***** line with additional variables.

* An additional do file is called in to set those withdrawing consent to missing so that this is always up to date whenever you run this do file

order aln mz010, first

order bestgest, last

do "\\ads.bris.ac.uk\filestore\SSCM ALSPAC\Data\Syntax\Withdrawal of consent\mother_WoC_020715.do"

save "M:projects\ieu2\p5\002\working\data\results\PHD\2ND YEAR\RISK FACTOR VARIABLES\ANIMAL EXPOSURE\DATA SETS\mother.dta", replace

********** MERGING THE TWO DATA SETS TO MAKE ONE OVERALL DATA SET **********

** Matching all data together and saving out the final file*.

* NOTE: any linkage data should be added here*.

use "M:projects\ieu2\p5\002\working\data\results\PHD\2ND YEAR\RISK FACTOR VARIABLES\ANIMAL EXPOSURE\DATA SETS\childB.dta", clear

merge m:1 aln using "M:projects\ieu2\p5\002\working\data\results\PHD\2ND YEAR\RISK FACTOR VARIABLES\ANIMAL EXPOSURE\DATA SETS\mother.dta", nogen

* Remove non-alspac children.

drop if in_alsp!=1.

* Remove trips and quads.

drop if tripquad==1

drop in_alsp tripquad

save "M:projects\ieu2\p5\002\working\data\results\PHD\2ND YEAR\RISK FACTOR VARIABLES\ANIMAL EXPOSURE\DATA SETS\Data extracted from ALSPAC - ANIMAL.dta", replace

****

* QC checks*

use "M:projects\ieu2\p5\002\working\data\results\PHD\2ND YEAR\RISK FACTOR VARIABLES\ANIMAL EXPOSURE\DATA SETS\Data extracted from ALSPAC - ANIMAL.dta", clear

* Check that there are 15445 records.

count

********** USE/OPEN THE DATA SET **********

use "M:projects\ieu2\p5\002\working\data\results\PHD\2ND YEAR\RISK FACTOR VARIABLES\ANIMAL EXPOSURE\DATA SETS\Data extracted from ALSPAC - ANIMAL.dta"

********** MERGE IN ECZEMA VARIABLE DATA **********

* Add eczema data

merge 1:1 aln qlet using "M:projects\ieu2\p5\002\working\data\results\PHD\2ND YEAR\ECZEMA VARIABLE\DATA SETS\final_eczema_definitions.dta", nogen

* Add CONFOUNDING - SES data

merge 1:1 aln qlet using "M:projects\ieu2\p5\002\working\data\results\PHD\2ND YEAR\RISK FACTOR VARIABLES\CONFOUNDING\SES\DATA SETS\motherchildB.dta", nogen

* Add other confounders

merge 1:1 aln qlet using "M:projects\ieu2\p5\002\working\data\results\PHD\2ND YEAR\RISK FACTOR VARIABLES\potential_early_risk_factors_RG", nogen

**********REPLACING MINUS NUMBERS BY 'MISSING' **********

* Replacing all the minus numbers by the . symbol

* Create a variable of all the 200 variables, then do the function on each 'var'

```
local vars="aln kz011b kz021 kz030 in_core in_phase2 in_phase3 kc370 kc371 a073 a076 f393 f396"
foreach var in `vars' {
        replace `var'=. if `var'<0
}

********************************************** ANIMAL EXPOSURE DEFINITIONS **********************************************

************* CAT EXPOSURE ***************

generate cat_exposure = 1 if a073 == 1 | a073 == 2 | f393 == 1 | f393 == 2 | kc370 == 1

replace cat_exposure = 0 if cat_exposure != 1

************* DOG EXPOSURE ***************

generate dog_exposure = 1 if a076 == 1 | a076 == 2 | f396 == 1 | f396 == 2 | kc371 == 1
replace dog_exposure = 0 if dog_exposure != 1

********************************************** LIMIT TO ARIES DATASET **********************************************

* Merge the aln and qlet data from ALSPAC data set and ARIES data set
```

```
merge 1:1 aln qlet using "\\ads.bris.ac.uk\filestore\SSCM ALSPAC\Data\Useful_data\ARIES\YP_in_ARIES.dta"

* Get rid of any values that aren't in the ARIES dataset
drop if in_ARIES == .

* Drop the ARIES data set itself
drop in_ARIES

************************************** TABULATION OF 'EVER_ECZEMA' AND 'ANIMAL EXPOSURE' **************************************

tab ever_eczema cat_exposure, column row
tab ever_eczema dog_exposure, column row

****************************************** LOGISTIC REGRESSION - CAT EXPOSURE ******************************************

logistic ever_eczema cat_exposure

****************************************** LOGISTIC REGRESSION - DOG EXPOSURE ******************************************

logistic ever_eczema dog_exposure
```

********** LOGISTIC REGRESSION WITH CONFOUNDING - SES **********

logistic ever_eczema cat_exposure cmb_sc_grp

logistic ever_eczema dog_exposure cmb_sc_grp

********** LOGISTIC REGRESSION WITH 8 CONFOUNDERS **********

logistic ever_eczema cat_exposure kz021 cmb_sc_grp ch_ethni mum_hist preg_smk post_smk8m pet_dog freq_breast6

logistic ever_eczema dog_exposure kz021 cmb_sc_grp ch_ethni mum_hist preg_smk post_smk8m pet_cat freq_breast6

Breastfeeding do file

********** EXTRACTING DATA FROM ALSPAC **********

*** Syntax template for direct users preparing datasets using child and adult based datasets.

* This version created 29th October 2014 - always create a datafile using the most up to date template.

* This template is based on that used by the data buddy team and they include a number of variables by default.

* To ensure the file works we suggest you keep those in and just add any relevant variables that you need for your project.

****

```
* To add data other than that included by default you will need to add the relvant files and pathnames in each of the match commands below.
* There is a separate command for mothers, partner, mothers providing data on the child and data provided by the child themselves.
* each has different withdrawal of consent issues so they must be considered separately.
* You will need to replace 'YOUR PATHNAME' in each section with your working directory pathname.

********************************************** BREASTFEEDING ************************************************************

* Child BASED files - in this section the following files need to be placed:
* Mother completed Qs about YP
* ALWAYS KEEP THIS SECTION EVEN IF ONLY CHILD COMPLETED REQUESTED, although you will need to remove the *****

********************************************** CREATING 'CHILD BASED' DATA SET ********************************************

use "\\ads.bris.ac.uk\filestore\SSCM ALSPAC\Data\Current\Other\Sample Definition\kz_5b.dta", clear
sort aln qlet
gen in_kz=1
merge 1:1 aln qlet using "\\ads.bris.ac.uk\filestore\SSCM ALSPAC\Data\Current\Other\cohort profile\cp_r1a.dta", nogen
merge 1:1 aln qlet using "\\ads.bris.ac.uk\filestore\SSCM ALSPAC\Data\Current\Quest\Child Based\ka_4c.dta", nogen
merge 1:1 aln qlet using "\\ads.bris.ac.uk\filestore\SSCM ALSPAC\Data\Current\Quest\Child Based\kb_6d.dta", nogen
merge 1:1 aln qlet using "\\ads.bris.ac.uk\filestore\SSCM ALSPAC\Data\Current\Quest\Child Based\kc_5b.dta", nogen
merge 1:1 aln qlet using "\\ads.bris.ac.uk\filestore\SSCM ALSPAC\Data\Current\Quest\Child Based\kd_4b.dta", nogen
```

```
merge 1:1 aln qlet using "\\ads.bris.ac.uk\filestore\SSCM ALSPAC\Data\Current\Quest\Child Based\ke_5c.dta", nogen
merge 1:1 aln qlet using "\\ads.bris.ac.uk\filestore\SSCM ALSPAC\Data\Current\Quest\Child Based\kg_4b.dta", nogen
merge 1:1 aln qlet using "\\ads.bris.ac.uk\filestore\SSCM ALSPAC\Data\Current\Quest\Child Based\kk_2c.dta", nogen

keep aln qlet   kz011b kz021 kz030 /// KZ
in_core in_alsp in_phase2 in_phase3 tripquad /// Cohort profile
ka030 ka031 ka032 ka033 ka034 ka061 ka094 /// KA
kb275 kb276 kb277 kb421 /// KB
kc401 kc402 kc403 /// KC
kd243a /// KD
ke240 ke241 /// KE
kg404 /// KG
kk636 /// KK
tripquad //

* Dealing with withdrawal of consent: For this to work additional variables required have to be inserted before in_core, so replace the ***** line with additional variables.
* If none are required remember to delete the ***** line.
* An additional do file is called in to set those withdrawing consent to missing so that this is always up to date whenever you run this do file

order aln qlet ka030, first
```

```
order in_alsp tripquad, last
do "\\ads.bris.ac.uk\filestore\SSCM ALSPAC\Data\Syntax\Withdrawal of consent\child_based_WoC.do"
save "M:projects\ieu2\p5\002\working\data\results\PHD\2ND YEAR\RISK FACTOR VARIABLES\BREASTFEEDING\DATA SETS\childB.dta", replace

********************************************** CREATING 'MOTHER' DATA SET **********************************************

* MOTHER completed files
use "\\ads.bris.ac.uk\filestore\SSCM ALSPAC\Data\Current\Other\Sample Definition\mz_5a.dta", clear
sort aln
gen in_mz=1
merge 1:1 aln using "\\ads.bris.ac.uk\filestore\SSCM ALSPAC\Data\Current\Quest\Mother\b_4d.dta", nogen
merge 1:1 aln using "\\ads.bris.ac.uk\filestore\SSCM ALSPAC\Data\Useful_data\bestgest\bestgest.dta", nogen
keep aln mz001 mz010 mz010a mz013 mz014 mz028b ///
b028 b029 /// B
bestgest //

* Dealing with withdrawal of consent: For this to work additional variables required have to be inserted before tripquad, so replace the ***** line with
additional variables.
* An additional do file is called in to set those withdrawing consent to missing so that this is always up to date whenever you run this do file
order aln mz010, first
order bestgest, last
```

```
do "\\ads.bris.ac.uk\filestore\SSCM ALSPAC\Data\Syntax\Withdrawal of consent\mother_clinic_WoC.do"

save "M:projects\ieu2\p5\002\working\data\results\PHD\2ND YEAR\RISK FACTOR VARIABLES\BREASTFEEDING\DATA SETS\mother.dta", replace

********** MERGING THE TWO DATA SETS TO MAKE ONE OVERALL DATA SET **********

** Matching all data together and saving out the final file*.

* NOTE: any linkage data should be added here*.

use "M:projects\ieu2\p5\002\working\data\results\PHD\2ND YEAR\RISK FACTOR VARIABLES\BREASTFEEDING\DATA SETS\childB.dta", clear

merge m:1 aln using "M:projects\ieu2\p5\002\working\data\results\PHD\2ND YEAR\RISK FACTOR VARIABLES\BREASTFEEDING\DATA SETS\mother.dta",
nogen

* Remove non-alspac children.

drop if in_alsp!=1.

* Remove trips and quads.

drop if tripquad==1

drop in_alsp tripquad

save "M:projects\ieu2\p5\002\working\data\results\PHD\2ND YEAR\RISK FACTOR VARIABLES\BREASTFEEDING\DATA SETS\Data_extracted_from_ALSPAC-
BREASTFEEDING.dta", replace

**********
```

```
* QC checks*

use "M:projects\ieu2\p5\002\working\data\results\PHD\2ND YEAR\RISK FACTOR VARIABLES\BREASTFEEDING\DATA SETS\Data_extracted_from_ALSPAC-BREASTFEEDING.dta", clear


* Check that there are 15445 records.

count


************************************************** USE/OPEN THE DATA SET **************************************************


use "M:projects\ieu2\p5\002\working\data\results\PHD\2ND YEAR\RISK FACTOR VARIABLES\BREASTFEEDING\DATA SETS\Data_extracted_from_ALSPAC-BREASTFEEDING.dta"


****************************************** REPLACING MINUS NUMBERS BY 'MISSING' ******************************************


* Replacing all the minus numbers by the . symbol

* Create a variable of all the 200 variables, then do the function on each 'var'

local vars="aln kz011b kz021 kz030 in_core in_phase2 in_phase3 ka030 ka031 ka032 ka033 ka034 ka061 ka094"


foreach var in `vars' {
        replace `var'=. if `var'<0
}
```

********** BREASTFEEDING DEFINITIONS **********

******** BREASTFEEDING ONLY **********

```
generate breastfeeding = 2 if (ka030 == 1 | ka030 == .) & (ka031 == 1 | ka031 == .) & (ka032 == 1 | ka032 == .) & (ka033 == 1 | ka033 == .) & (ka034 == 1 | ka034 == .) & (ka061 == 1 | ka061 == .) & (ka094 == 1 | ka094 == .)
```

******** BREASTFEEDING NEVER **********

```
replace breastfeeding = 0 if (ka030 == 2 | ka030 == .) & (ka031 == 2 | ka031 == .) & (ka032 == 2 | ka032 == .) & (ka033 == 2 | ka033 == .) & (ka034 == 2 | ka034 == .) & (ka061 == 2 | ka061 == .) & (ka094 == 4 | ka094 == .)
```

******** BREASTFEEDING AND BOTTLE **********

```
replace breastfeeding = 1 if breastfeeding != 2 & breastfeeding != 0 & (ka094 == 2 | ka094 == 3 | ka094 == .)
```

********** CHANGING VARIABLE FROM CONTINUOUS TO CATEGORICAL **********

```
label define breastfeeding_lb 0"never breastfed" 1"breast and bottle fed" 2"breastfed"
label values breastfeeding breastfeeding_lb
```

********** SAVING THE DEFINITIONS **********

```
keep aln qlet breastfeeding //

save "M:projects\ieu2\p5\002\working\data\results\PHD\2ND YEAR\RISK FACTOR VARIABLES\BREASTFEEDING\DATA SETS\Real_data_extracted_from_ALSPAC-BREASTFEEDING.dta", replace

************************************************** MERGE TWO DATASETS **************************************************

use "M:projects\ieu2\p5\002\working\data\results\PHD\2ND YEAR\RISK FACTOR VARIABLES\BREASTFEEDING\DATA SETS\Real_data_extracted_from_ALSPAC-BREASTFEEDING.dta", clear

merge 1:1 aln qlet using "M:projects\ieu2\p5\002\working\data\results\PHD\2ND YEAR\ECZEMA VARIABLE\DATA SETS\final_eczema_definitions.dta", nogen

merge 1:1 aln qlet using "M:projects\ieu2\p5\002\working\data\results\PHD\2ND YEAR\RISK FACTOR VARIABLES\CONFOUNDING\SES\DATA SETS\motherchildB.dta", nogen

merge 1:1 aln qlet using "M:projects\ieu2\p5\002\working\data\results\PHD\2ND YEAR\RISK FACTOR VARIABLES\potential_early_risk_factors_RG.dta", nogen

************************************************** LIMIT TO ARIES DATASET **************************************************

* Merge the aln and qlet data from ALSPAC data set and ARIES data set

merge 1:1 aln qlet using "\\ads.bris.ac.uk\filestore\SSCM ALSPAC\Data\Useful_data\ARIES\YP_in_ARIES.dta"

* Get rid of any values that aren't in the ARIES dataset

drop if in_ARIES == .
```

```
* Drop the ARIES data set itself
drop in_ARIES

****************************************** TABULATION OF 'EVER_ECZEMA' AND 'BREASTFEEDING' ******************************************
tab ever_eczema breastfeeding, column row

********************************************** LOGISTIC REGRESSION - BREASTFEEDING **********************************************
logistic ever_eczema i.breastfeeding
****************************************** LOGISTIC REGRESSION WITH CONFOUNDING - SES ******************************************
logistic ever_eczema i.breastfeeding cmb_sc_grp

****************************************** LOGISTIC REGRESSION WITH 15 CONFOUNDERS ******************************************

logistic ever_eczema i. breastfeeding cmb_sc_grp ch_ethni crowi_bi low_bw mum_age_grp mum_hist preterm rented_bi single preg_smk post_smk8m pet
pet_cat pet_dog
```

Appendix III: QQ-plots for each eczema EWAS

Figure A1: Q-Q plot - Is there a relationship between cord blood methylation and ever eczema?

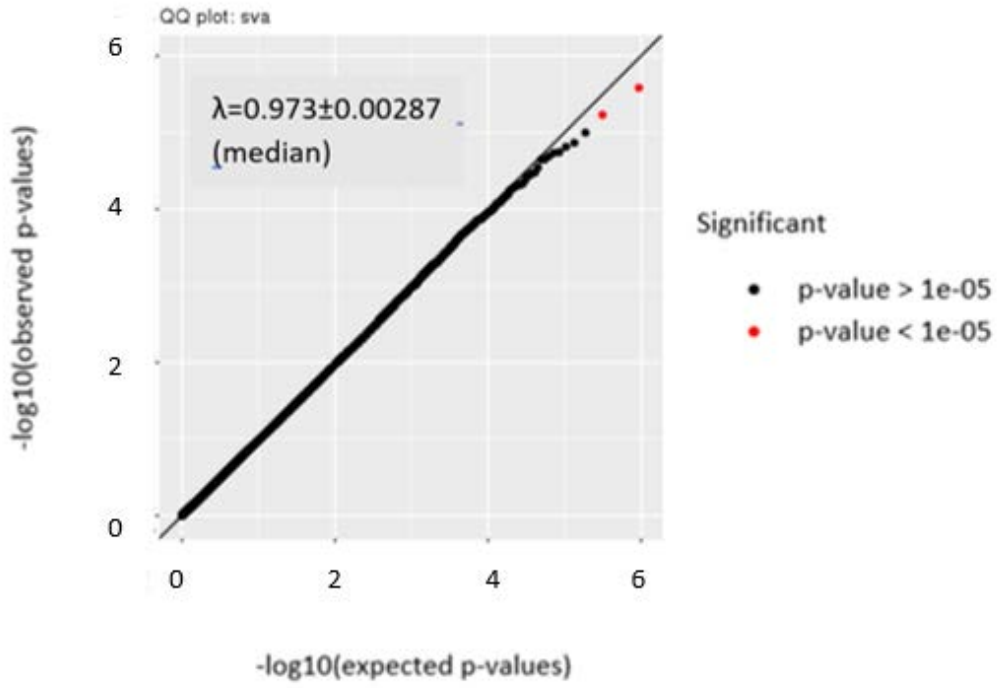


Figure A2: Q-Q plot - Is there a relationship between blood methylation at age seven and ever eczema?

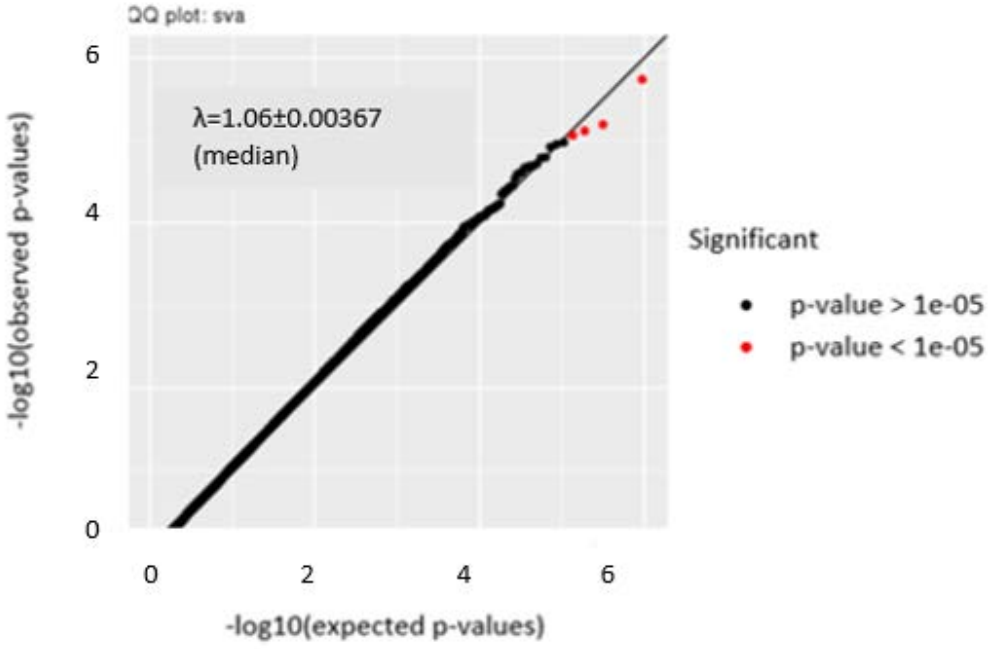


Figure A3: Q-Q plot: Is there a relationship between blood methylation at age seven and eczema in the last 12 months?

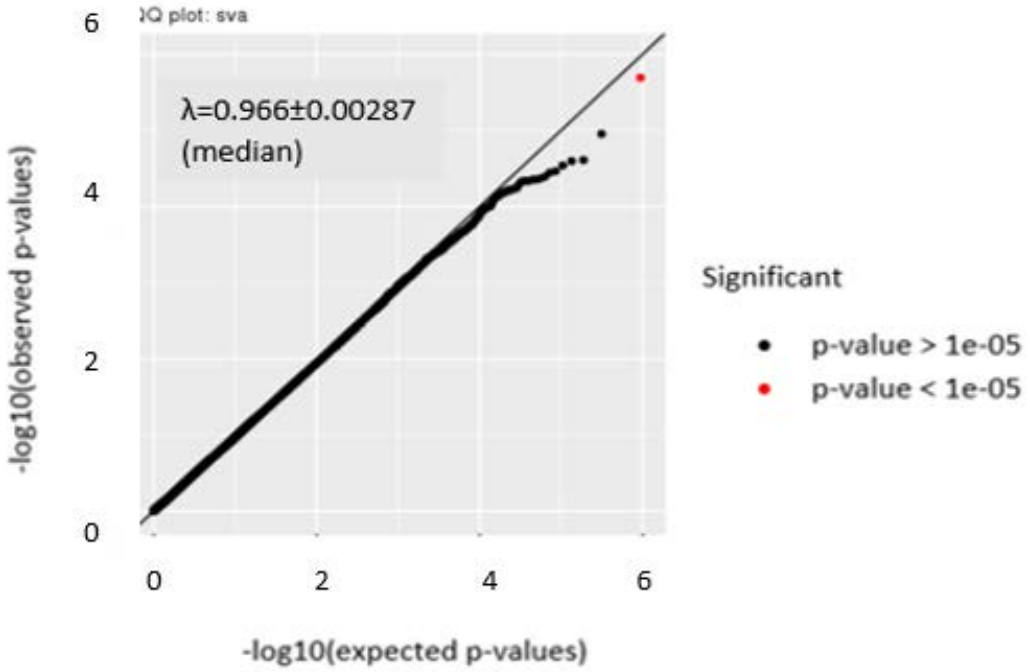
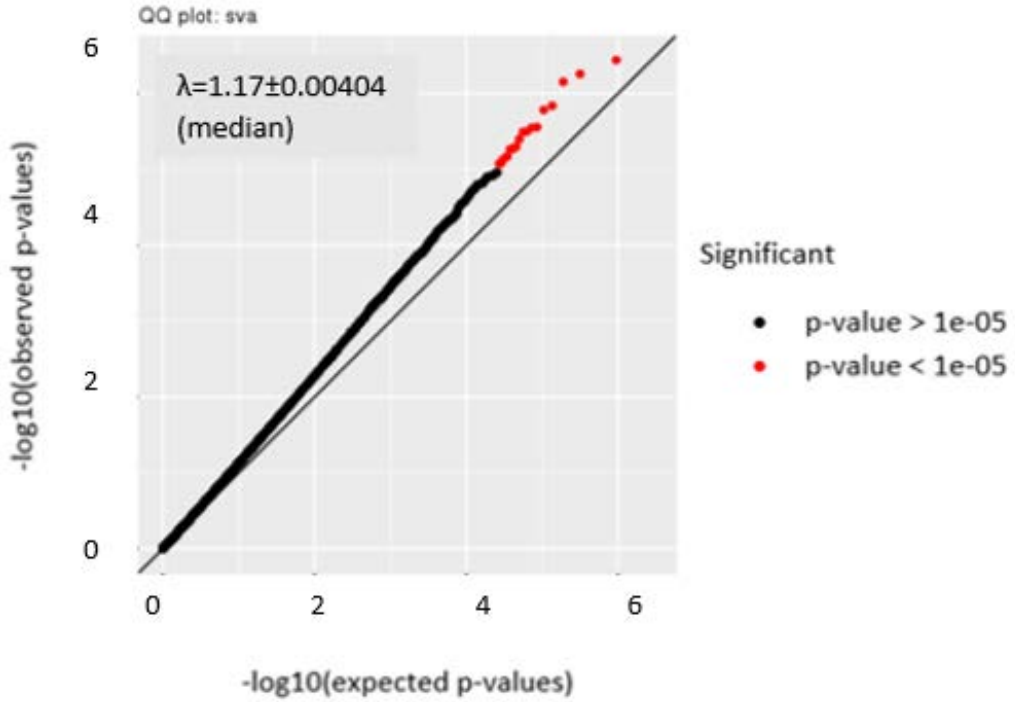


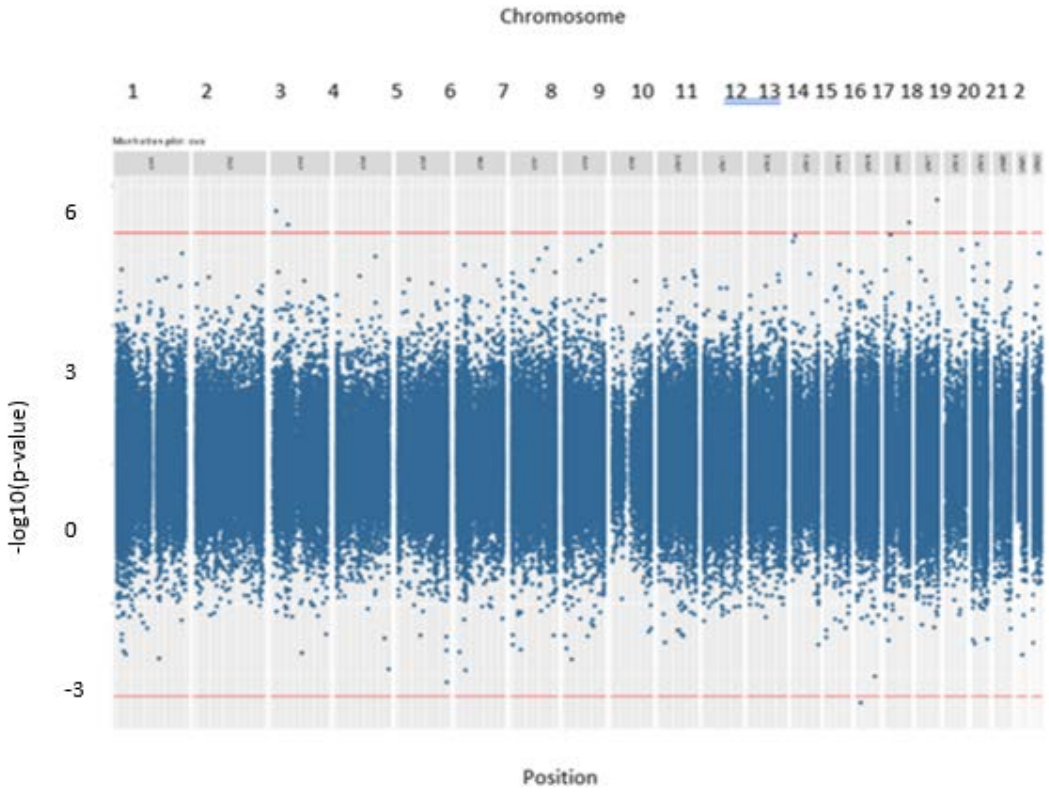
Figure A4: Q-Q plot: Is there a relationship between blood methylation at age 15/17 and eczema in the last 12 months?



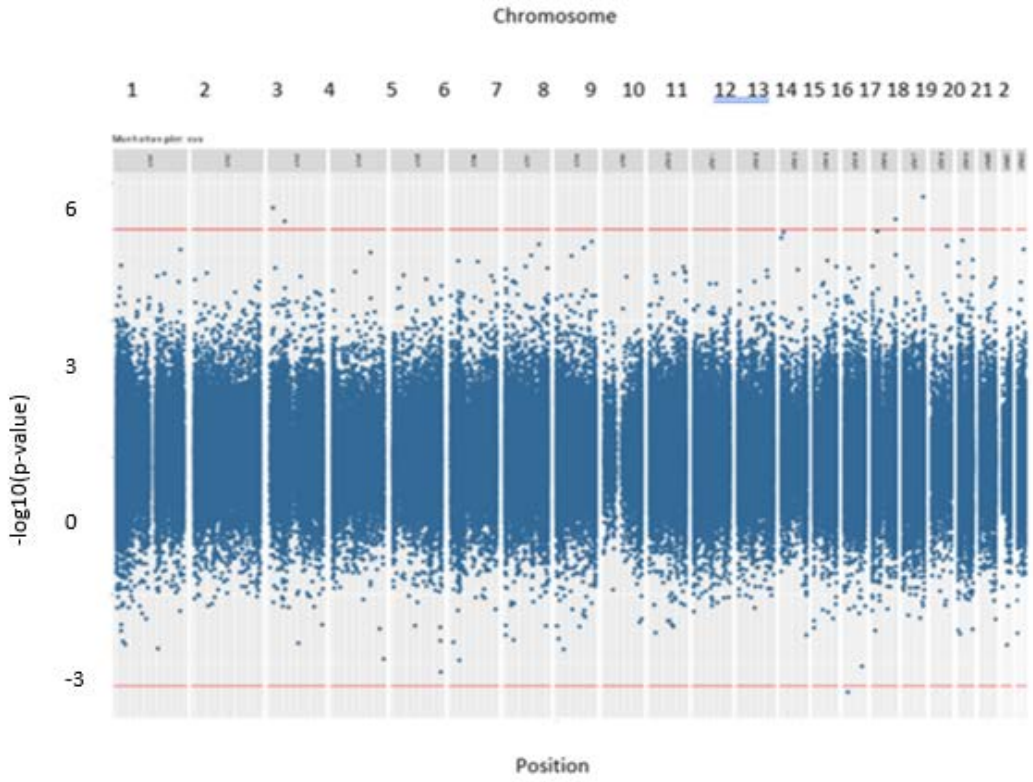
Appendix IV: Manhattan-plots for each analysis, from unadjusted to fully adjusted

CORD BLOOD METHYLATION, EVER ECZEMA

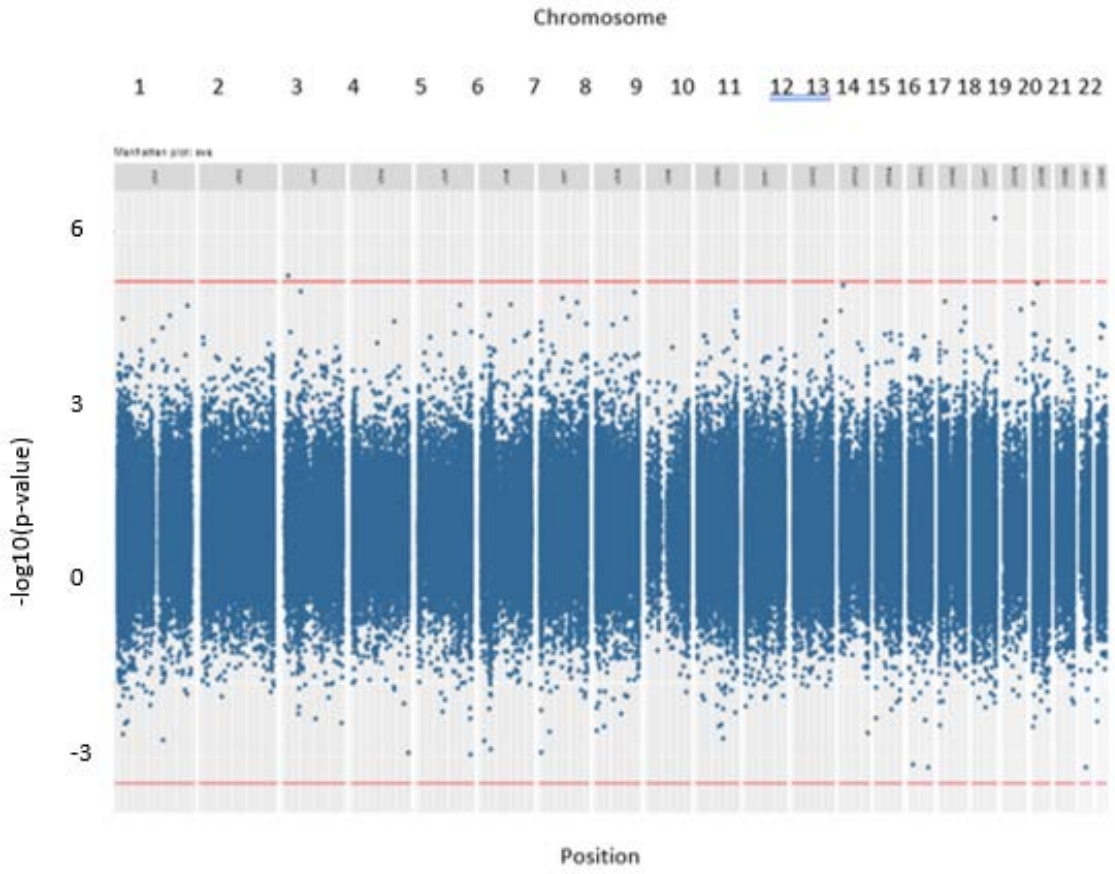
1) i) Cord blood, ever eczema (only adjusting for sex and surrogate variables)



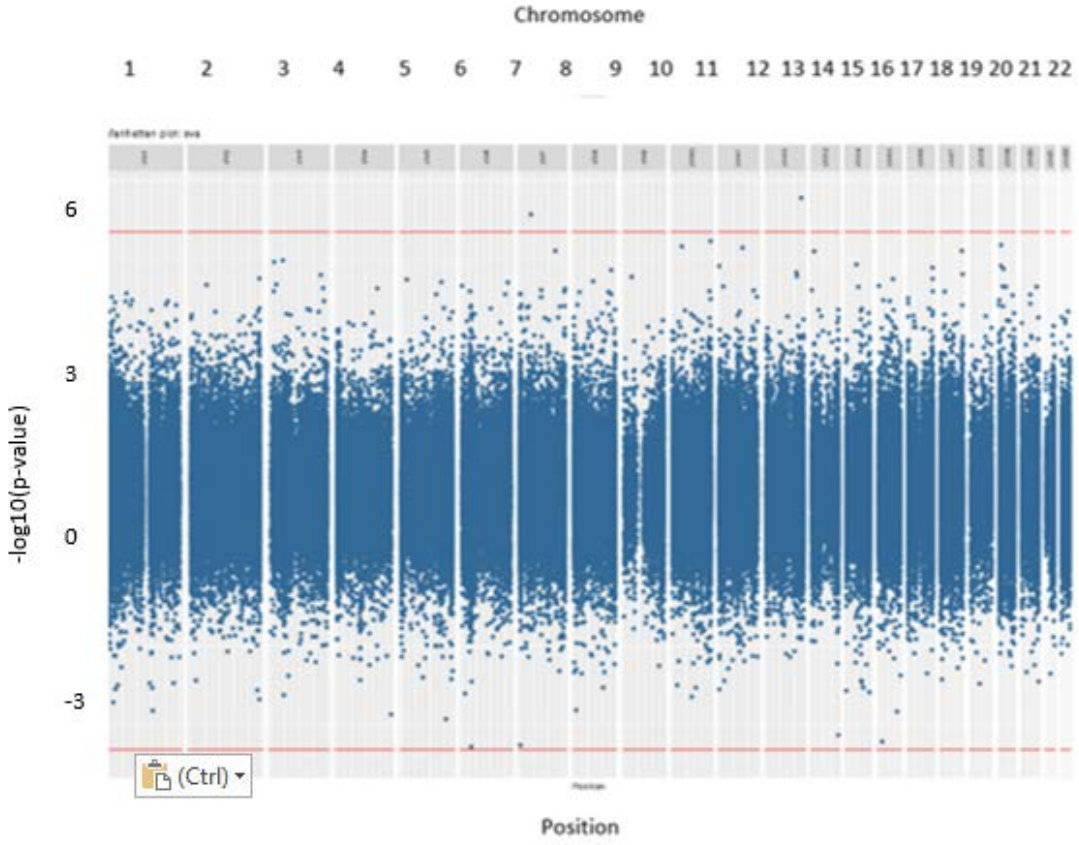
1) ii) Cord blood, ever eczema (adjusting for sex, surrogate variables and two socioeconomic status variables)



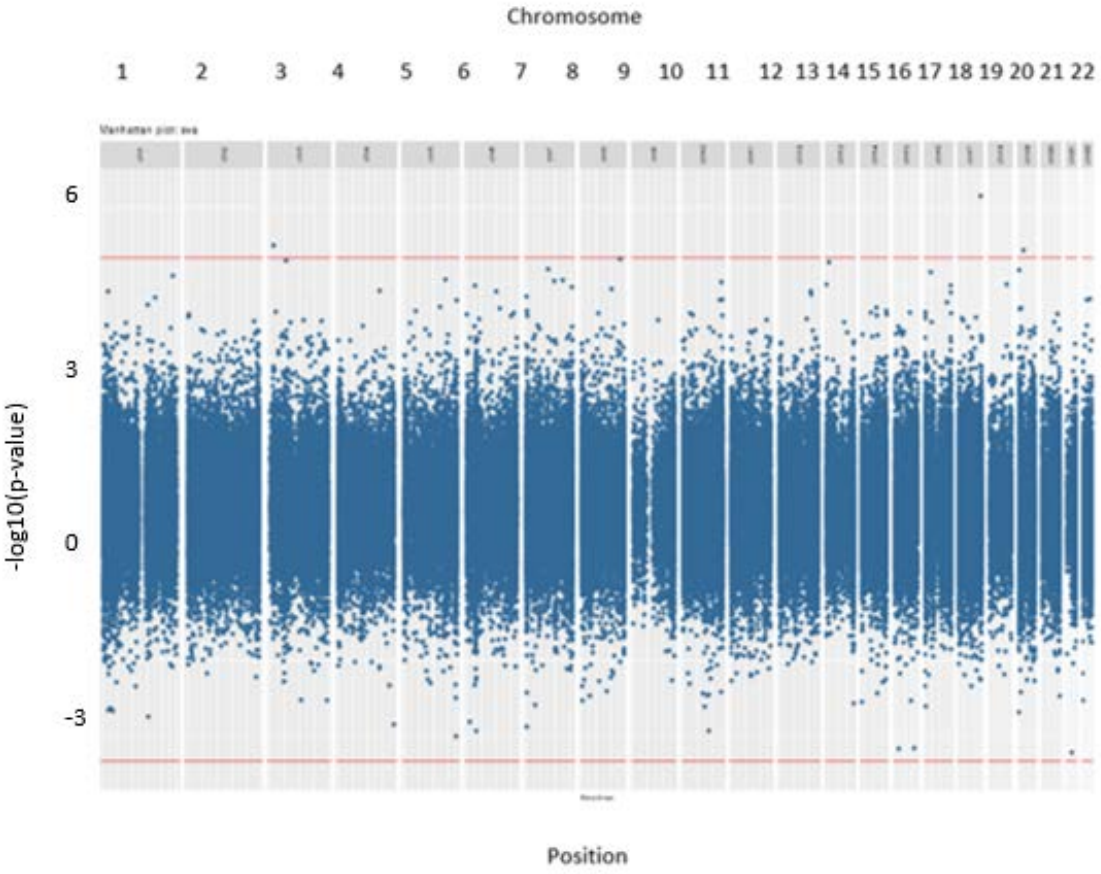
1) iii) Cord blood, ever eczema (adjusting for sex, surrogate variables, two socioeconomic status variables and maternal history)



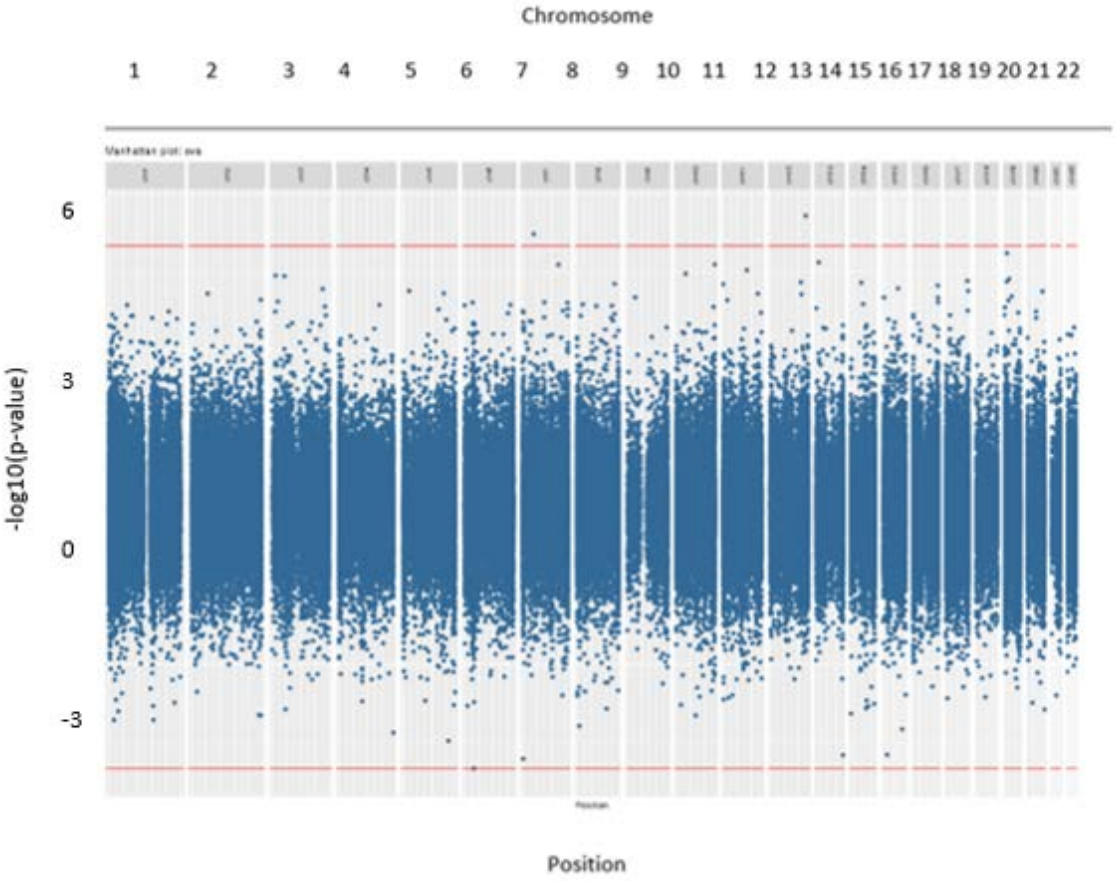
1) iv) Cord blood, ever eczema (adjusting for sex, surrogate variables, two socioeconomic status variables, maternal history and cell counts)



1) v) Cord blood, ever eczema (adjusting for sex, surrogate variables, two socioeconomic status variables, maternal history and three risk factors)

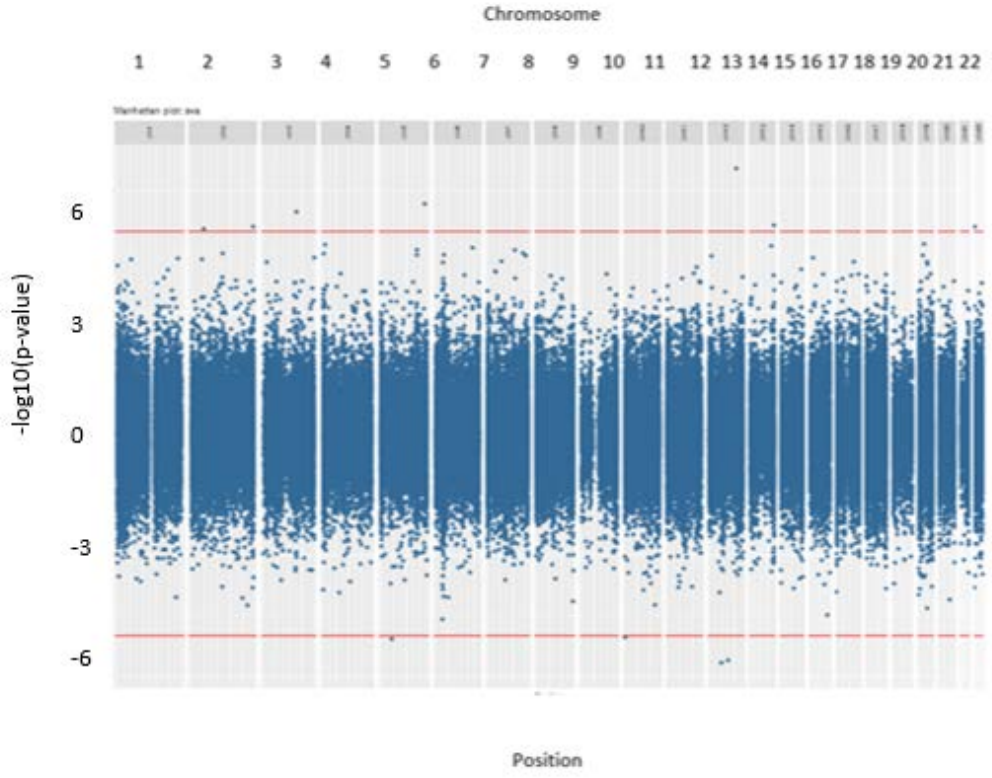


1) vi) Cord blood, ever eczema (adjusting for sex, surrogate variables, two socioeconomic status variables, maternal history, cell counts and three risk factors)

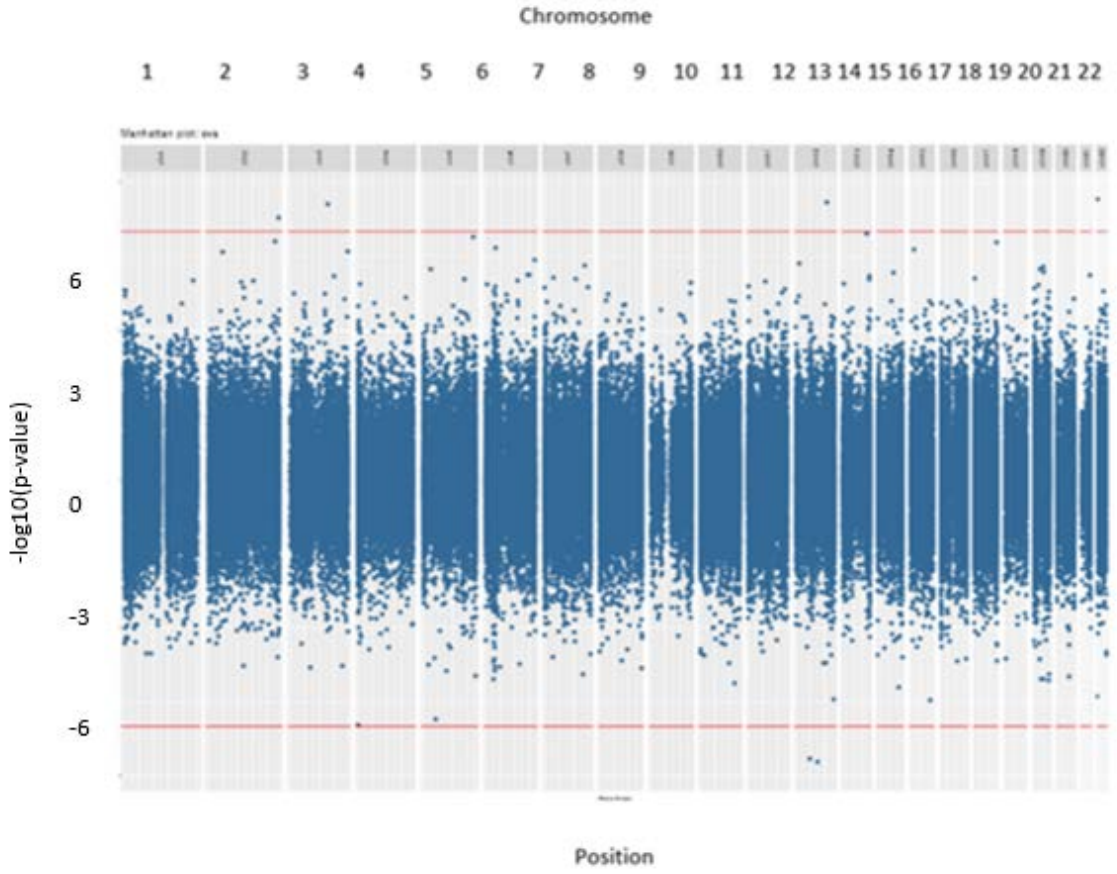


METHYLATION AGE 7, EVER ECZEMA

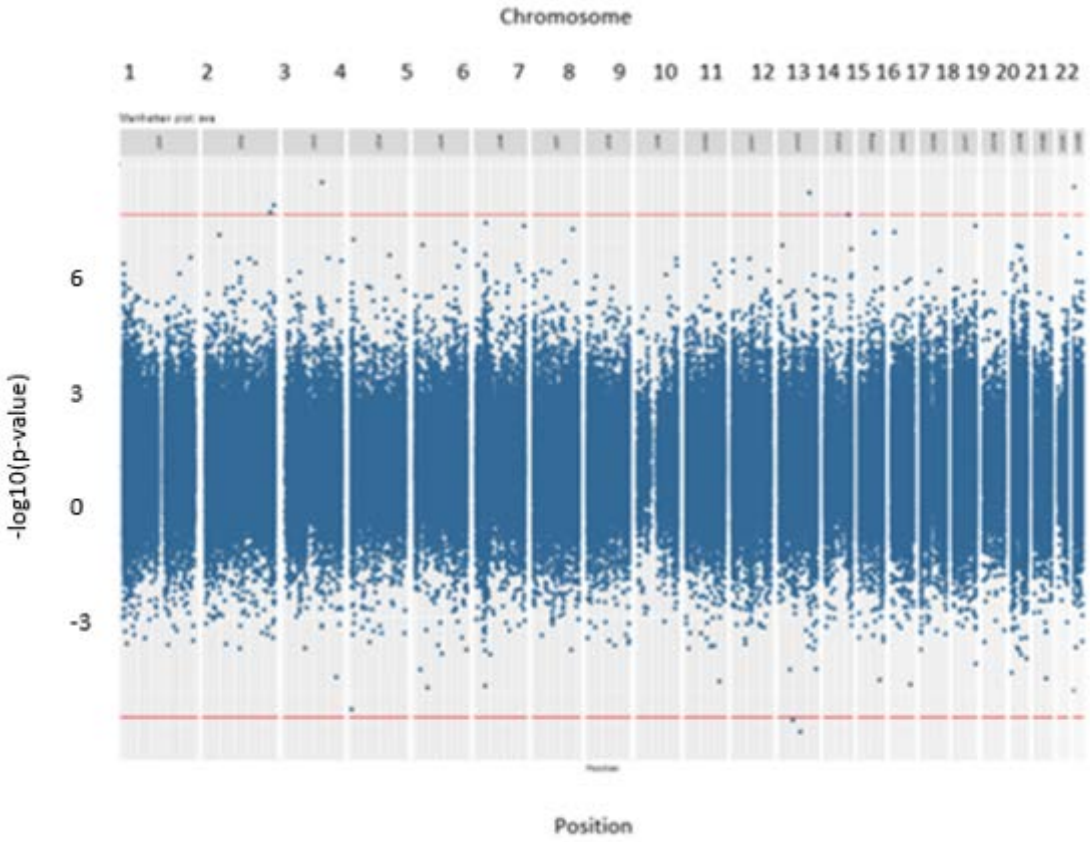
2) i) Methylation age 7, ever eczema (only adjusting for sex and surrogate variables)



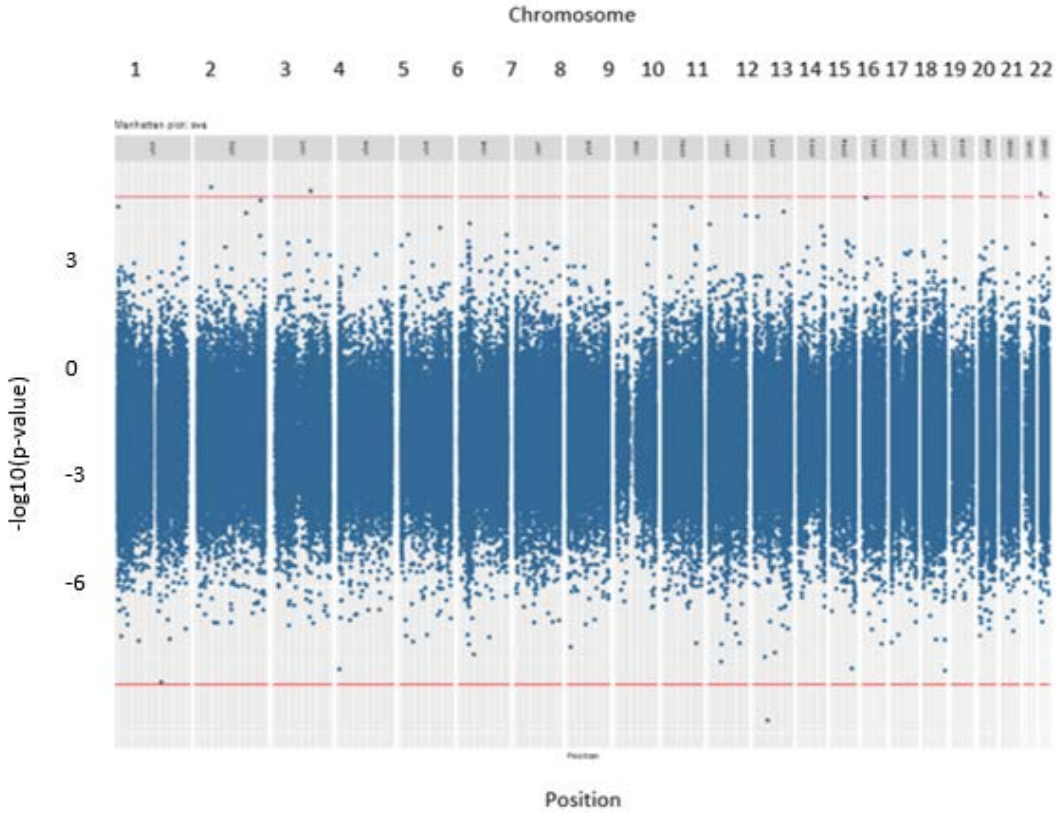
2) ii) Methylation age 7, ever eczema (adjusting for sex, surrogate variables and two socioeconomic status variables)



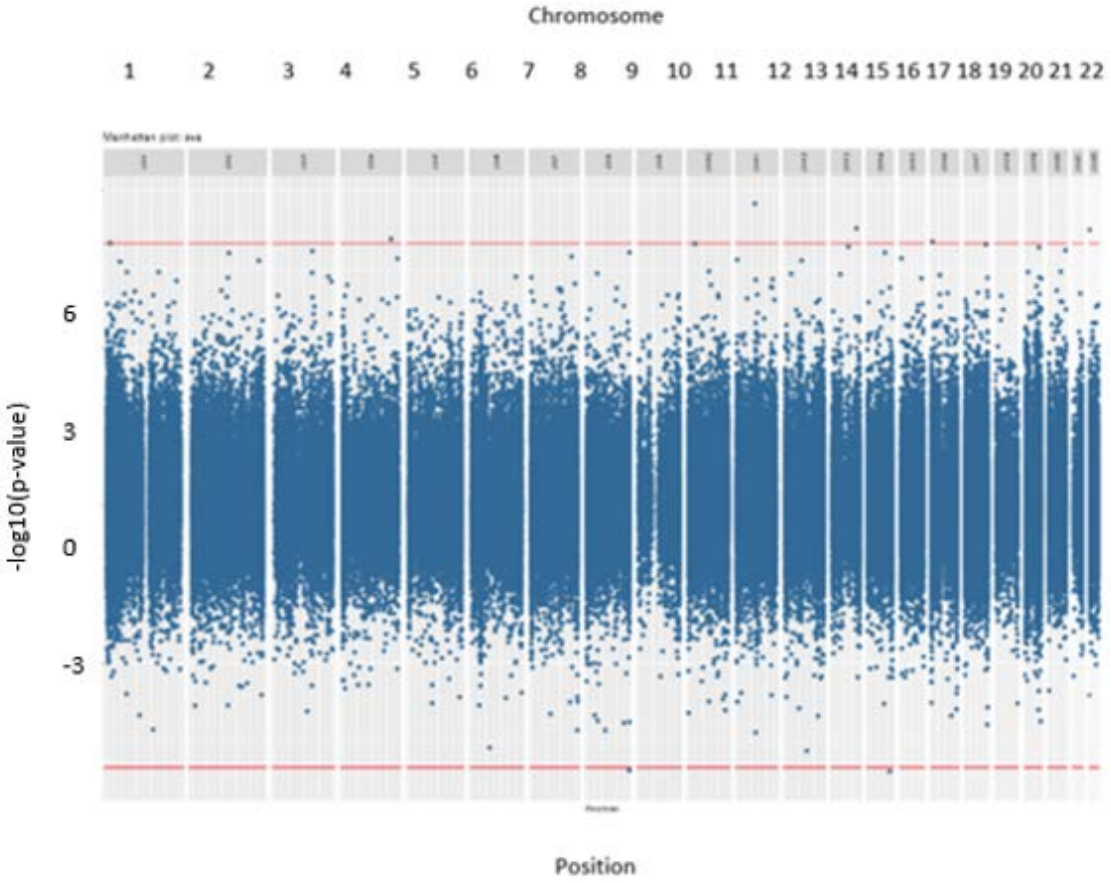
2) iii) Methylation age 7, ever eczema (adjusting for sex, surrogate variables, two socioeconomic status variables and maternal history)



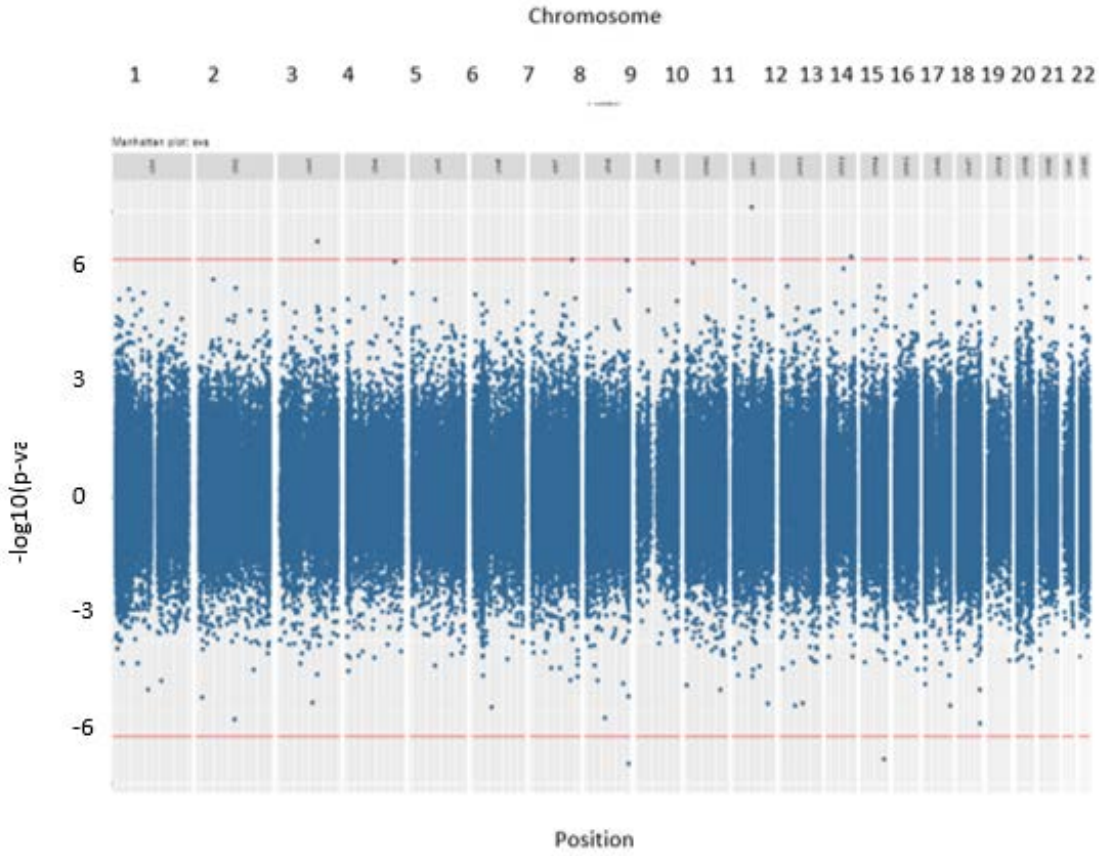
2) iv) Methylation age 7, *ever* eczema (adjusting for sex, surrogate variables, two socioeconomic status variables, maternal history and cell counts)



2) v) Methylation age 7, ever eczema (adjusting for sex, surrogate variables, two socioeconomic status variables, maternal history and three risk factors)

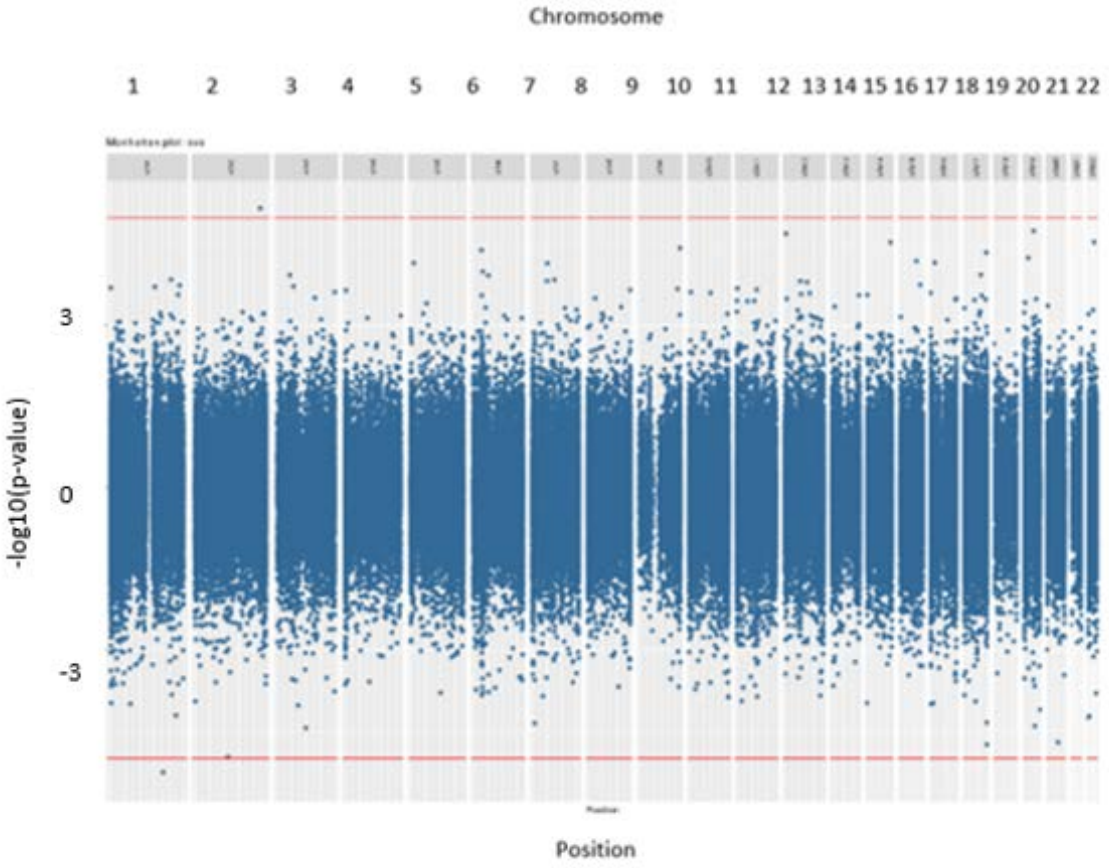


2) vi) Methylation age 7, ever eczema (adjusting for sex, surrogate variables, two socioeconomic status variables, maternal history, cell counts and three risk factors)

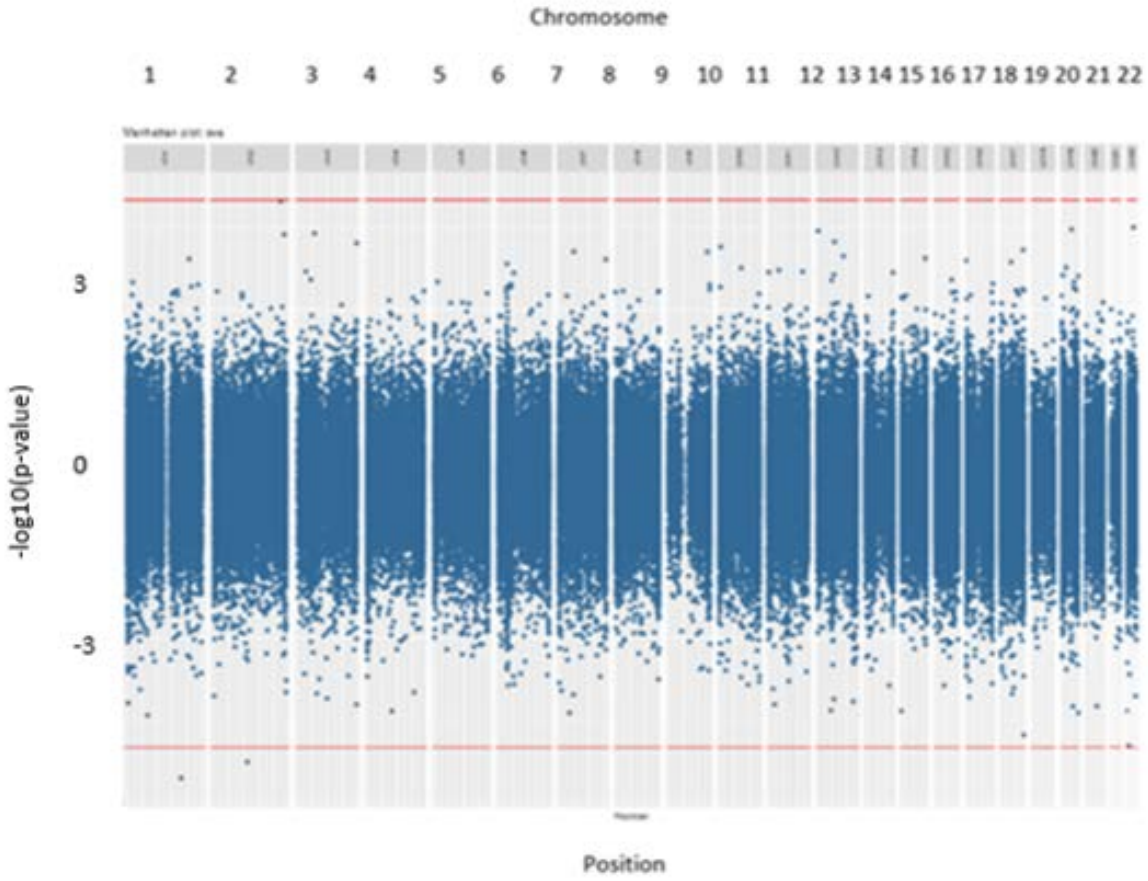


AGE 7 METHYLATION, ECZEMA IN LAST 12 MONTHS

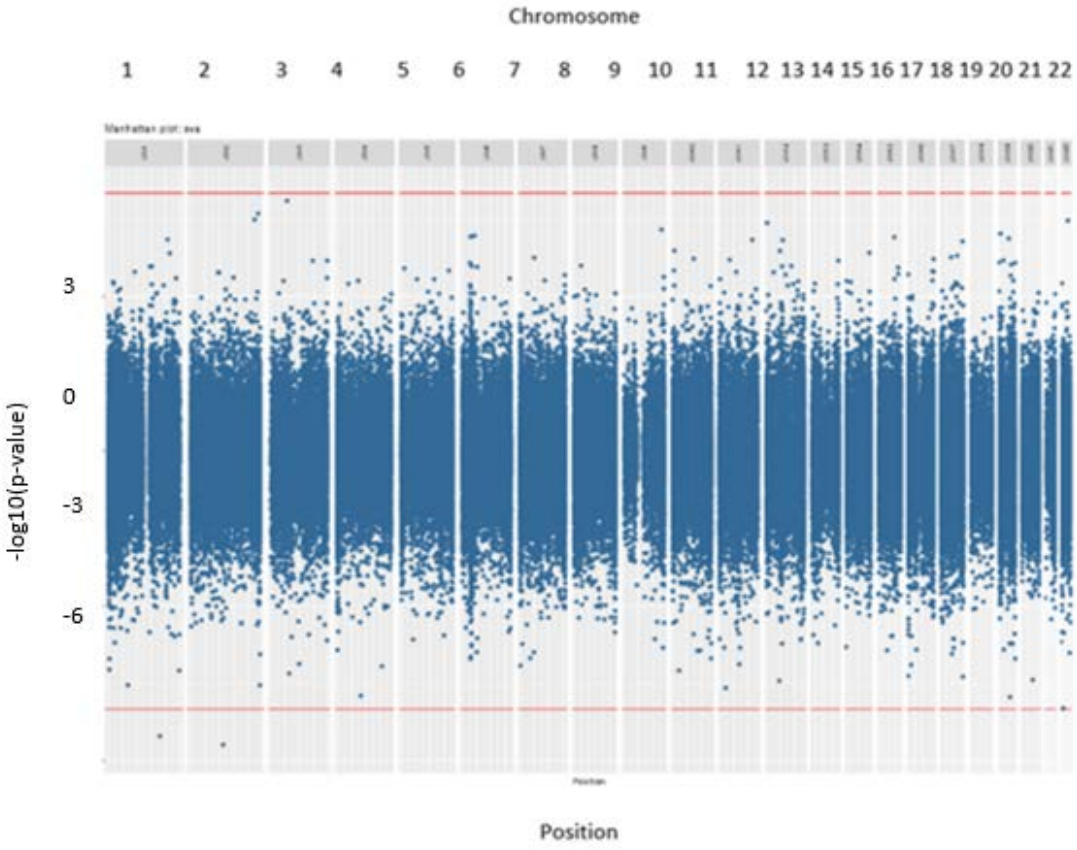
3 i) Methylation at 7, eczema in the last 12 months at age 7 (only adjusting for sex and surrogate variables)



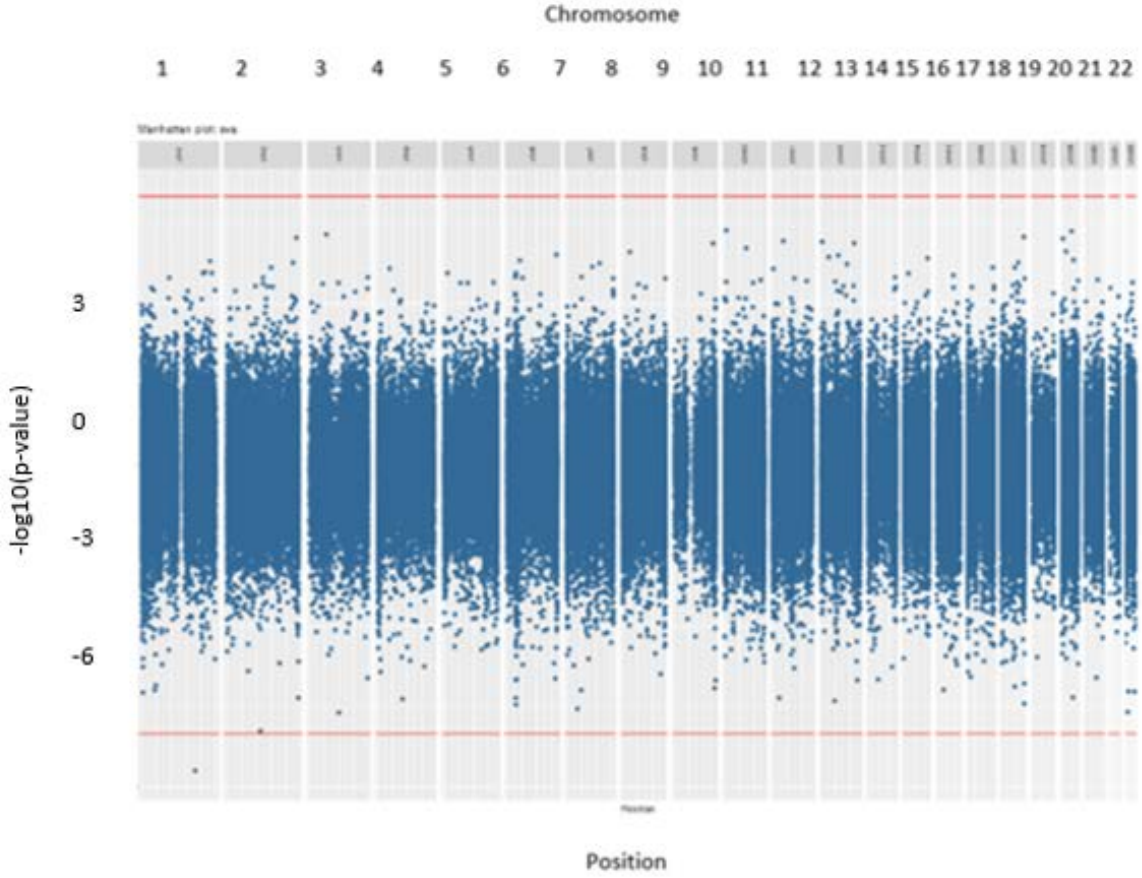
3 ii) Methylation at 7, eczema in the last 12 months at age 7 (adjusting for sex, surrogate variables and two socioeconomic status variables)



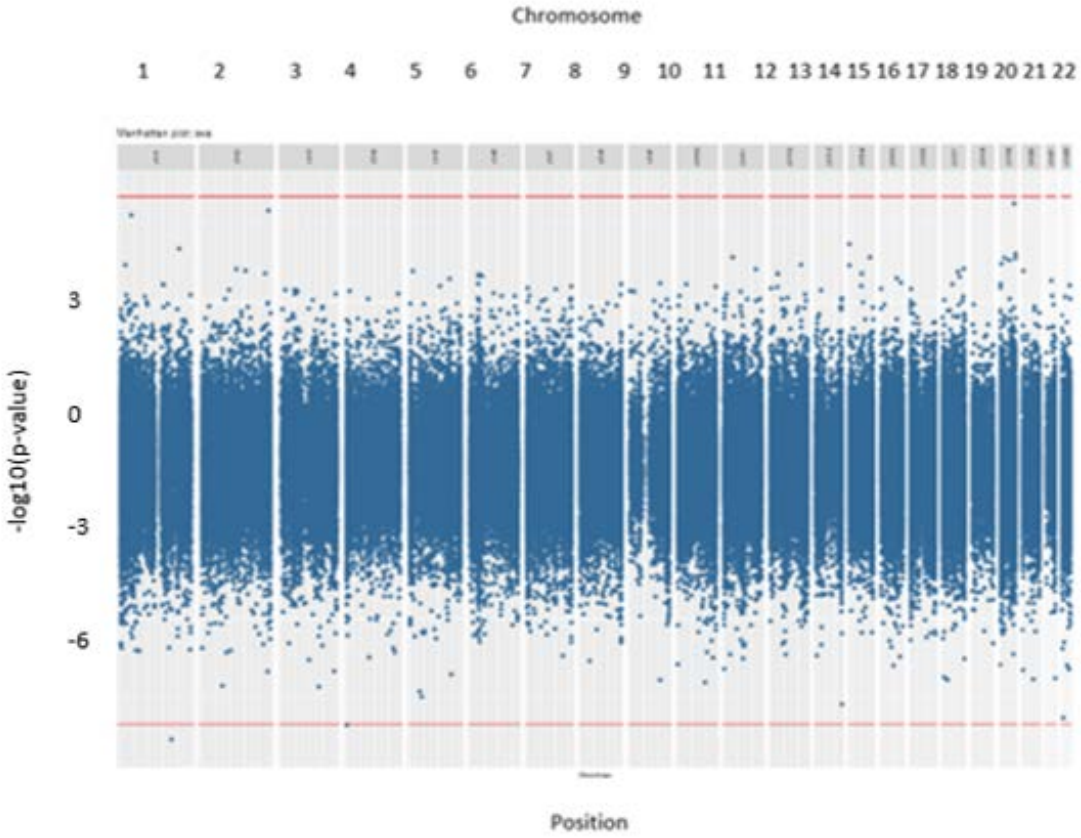
3) iii) Methylation at 7, eczema in the last 12 months at age 7 (adjusting for sex, surrogate variables, two socioeconomic status variables and maternal history)



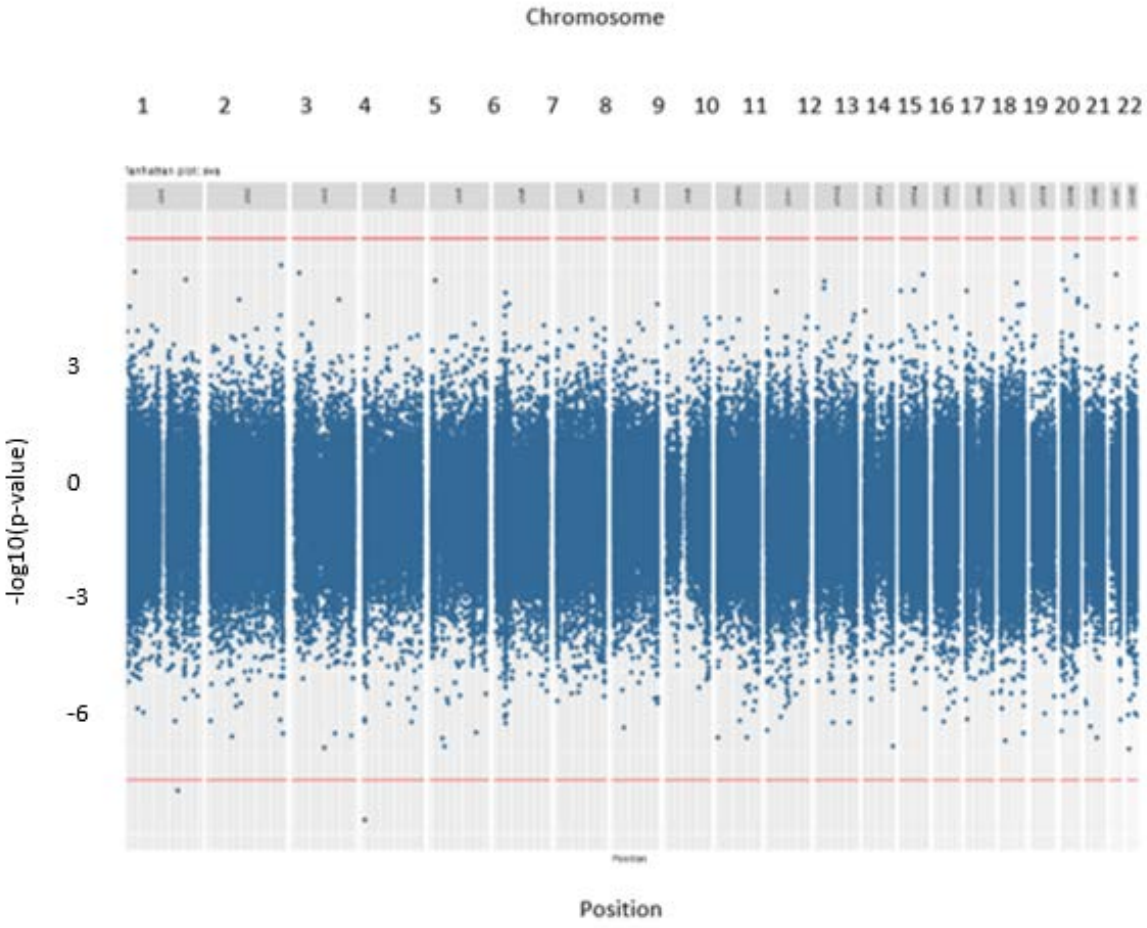
3) iv) Methylation at 7, eczema in the last 12 months at age 7 (adjusting for sex, surrogate variables, two socioeconomic status variables, maternal history and cell counts)



3) v) Methylation at 7, eczema in the last 12 months at age 7 (adjusting for sex, surrogate variables, two socioeconomic status variables, maternal history and three risk factors)

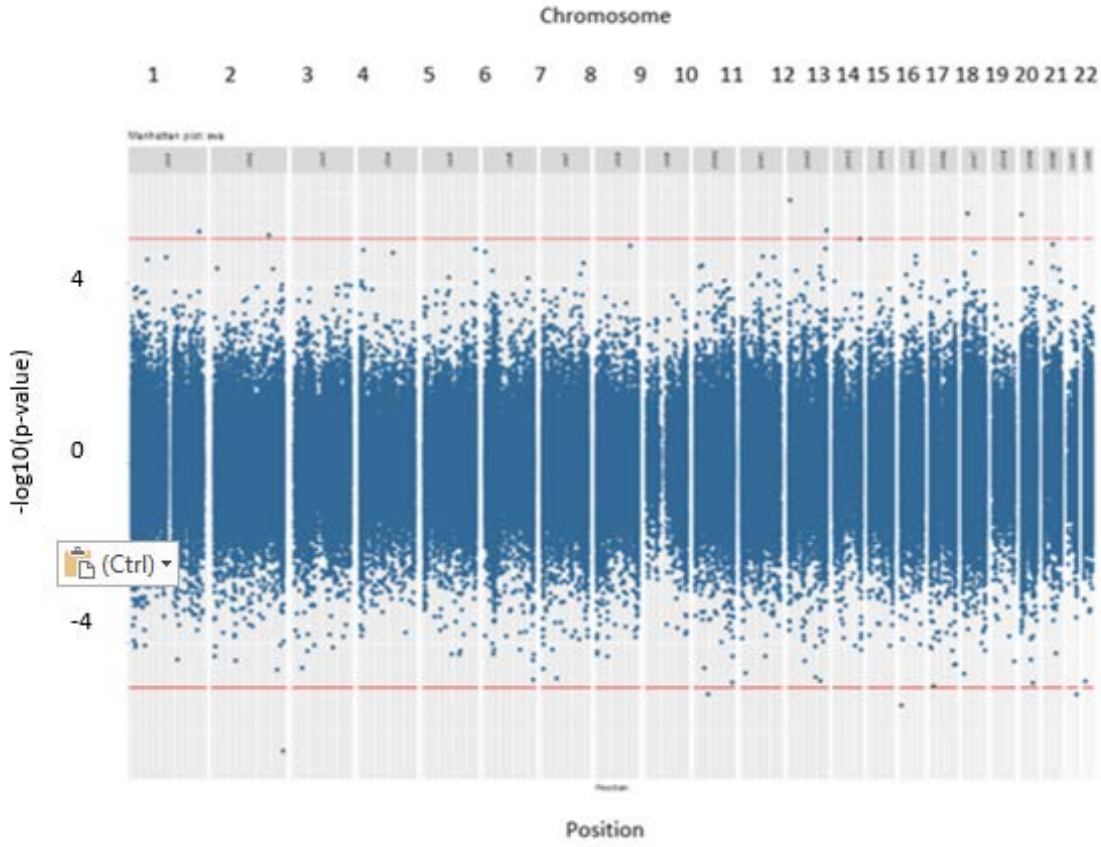


3) vi) Methylation at 7, eczema in the last 12 months at age 7 (adjusting for sex, surrogate variables, two socioeconomic status variables, maternal history, cell counts and three risk factors)

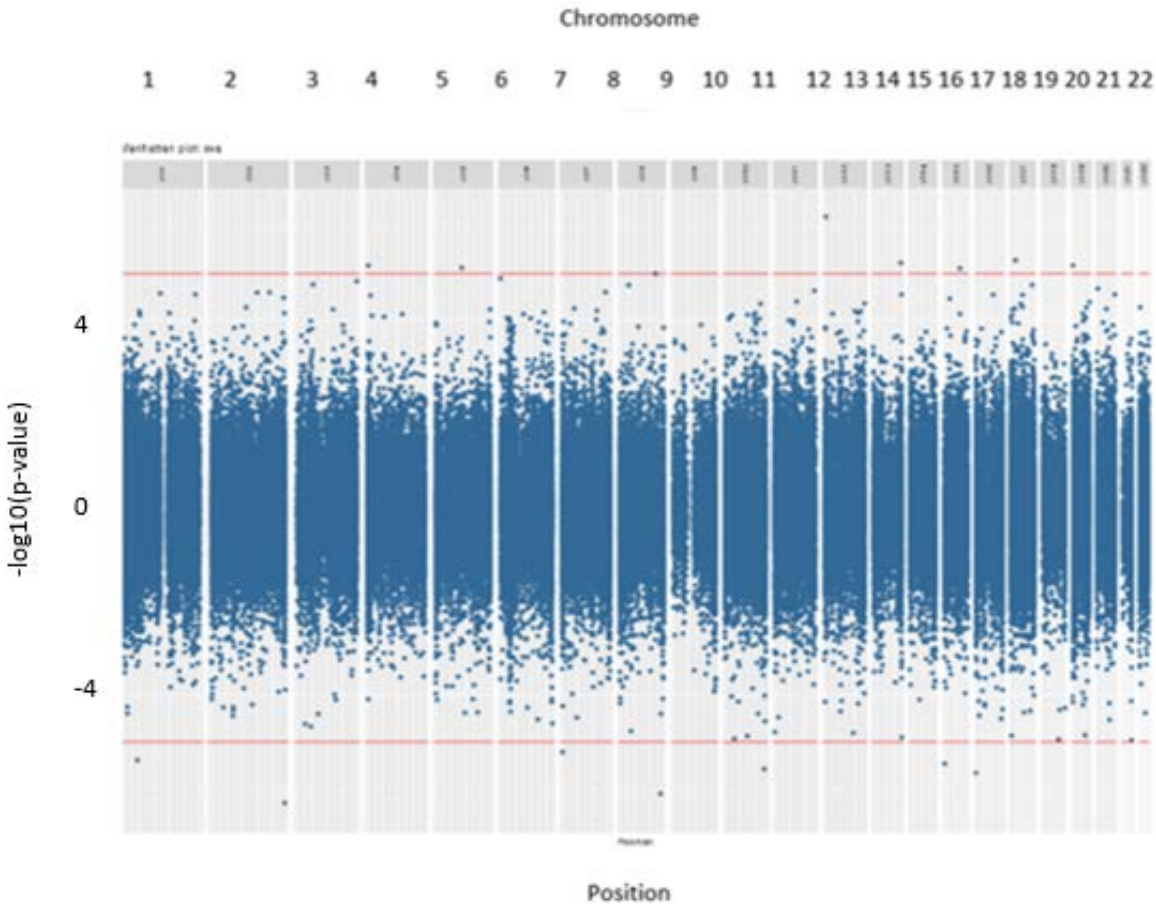


AGE 15/17 METHYLATION, ECZEMA IN LAST 12 MONTHS

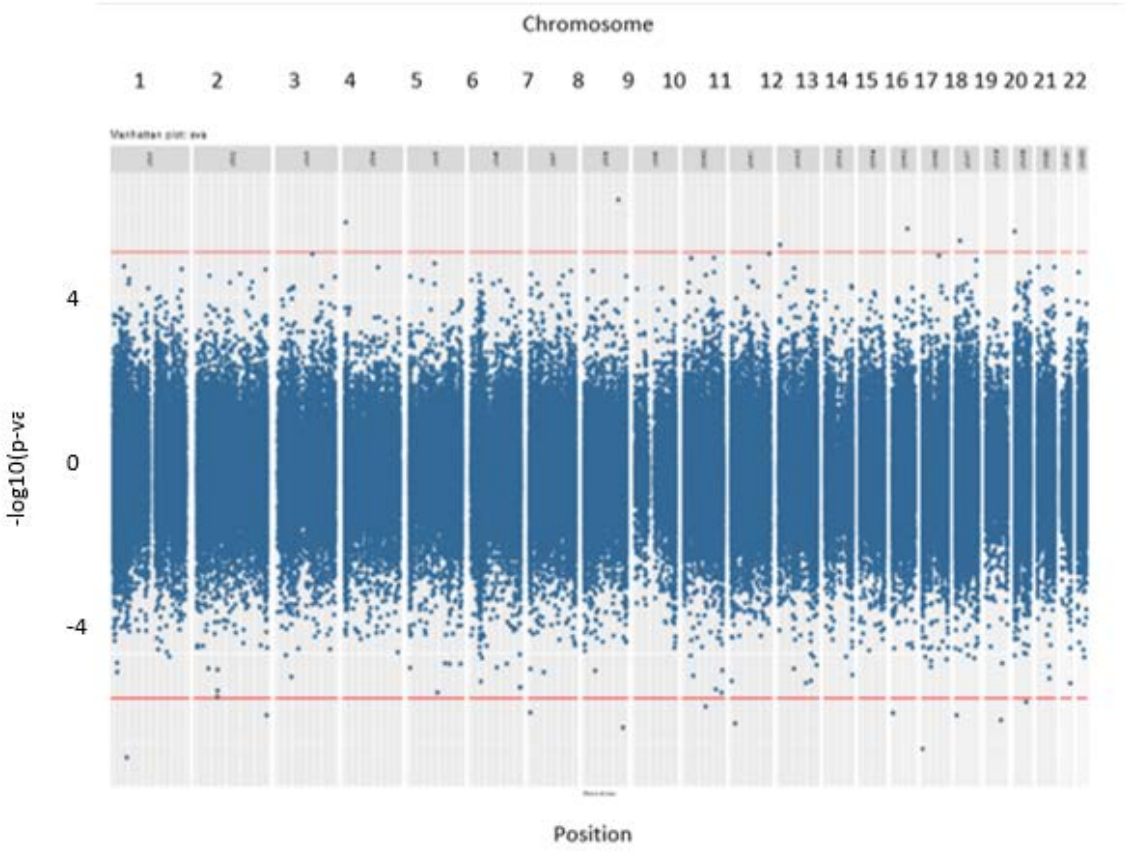
4) i) Methylation at 15/17, eczema in the last 12 months at age 15/17 (only adjusting for sex and surrogate variables)



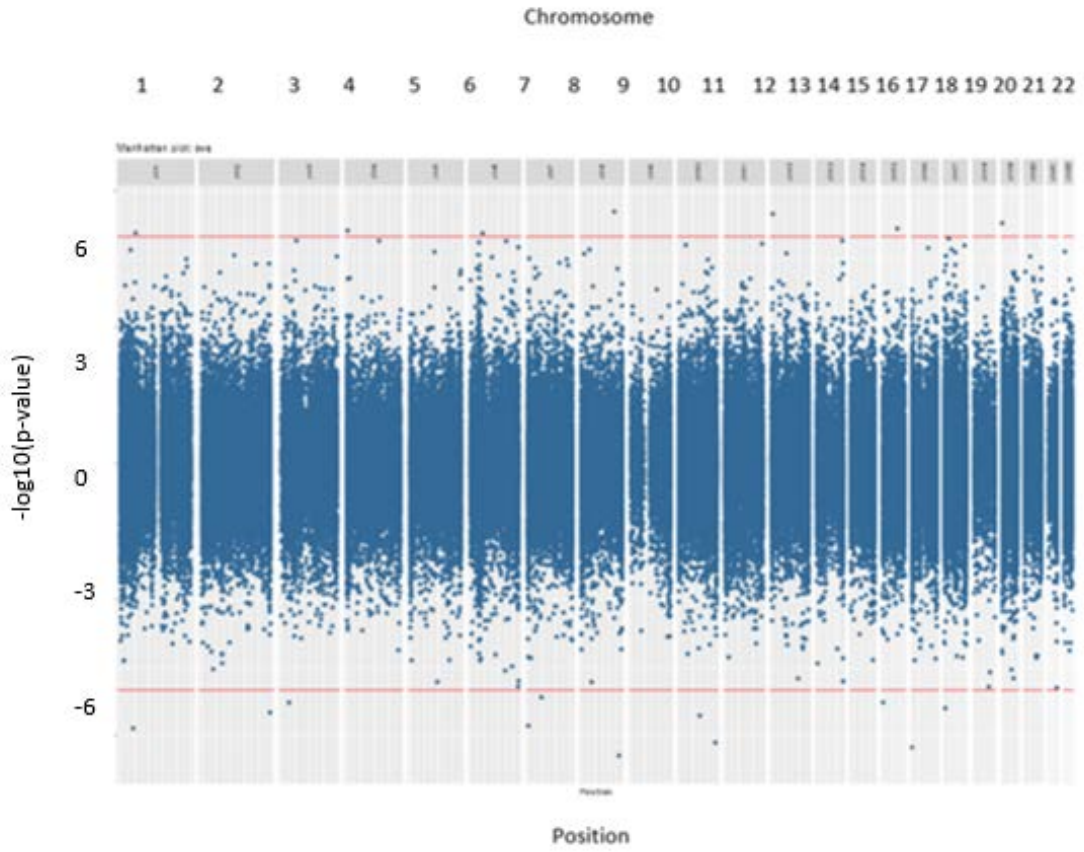
4) ii) Methylation at 15/17, eczema in the last 12 months at age 15/17 (adjusting for sex, surrogate variables and two socioeconomic status variables)



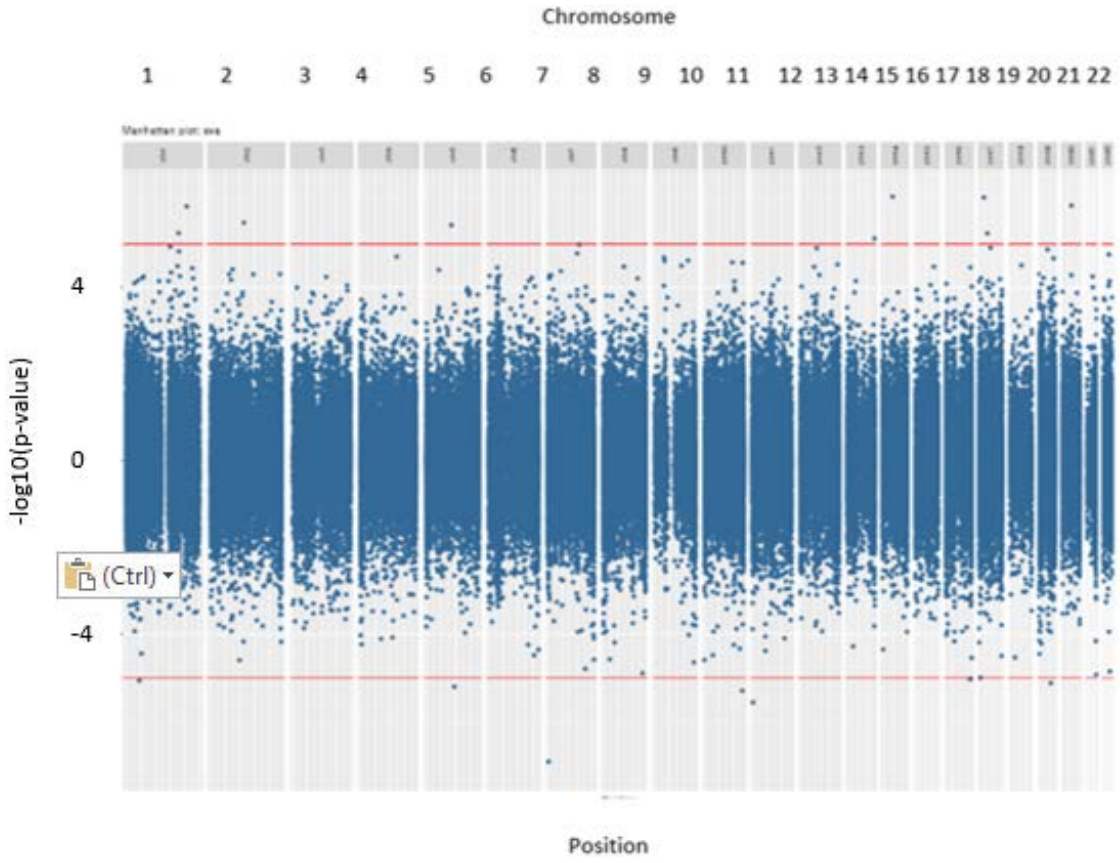
4) iii) Methylation at 15/17, eczema in the last 12 months at age 15/17 (adjusting for sex, surrogate variables, two socioeconomic status variables and maternal history)



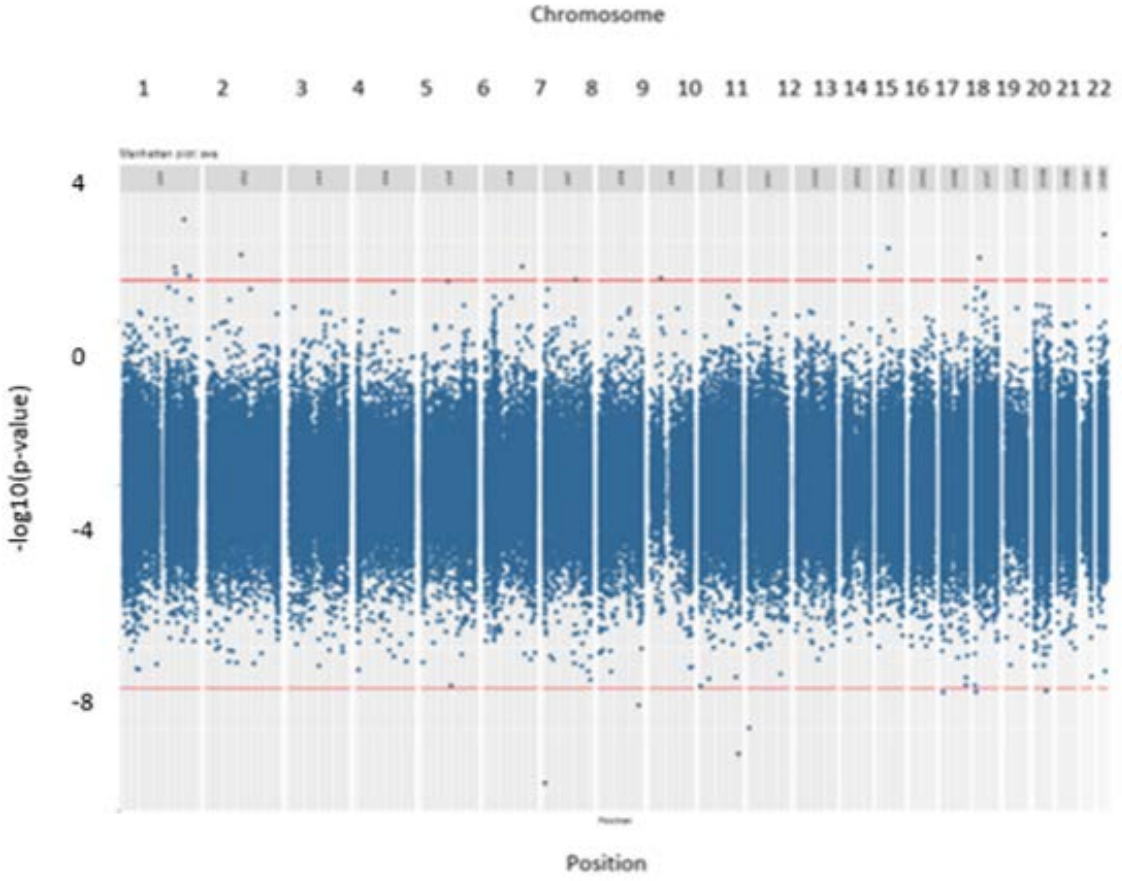
4) iv) Methylation at 15/17, eczema in the last 12 months at age 15/17 (adjusting for sex, surrogate variables, two socioeconomic status variables, maternal history and cell counts)



4) v) Methylation at 15/17, eczema in the last 12 months at age 15/17 (adjusting for sex, surrogate variables, two socioeconomic status variables, maternal history and three risk factors)



4) vi) Methylation at 15/17, eczema in the last 12 months at age 15/17 (adjusting for sex, surrogate variables, two socioeconomic status variables, maternal history, cell counts and three risk factors)



Appendix V: Overlap of CpG sites between smoking and eczema, looking at cord blood (smoking during pregnancy) and methylation at age seven (a smoky environment)

Table A7: Overlap of smoking CpG sites with eczema (table spread over three pages)

Taken from the Joehanes paper (97), there were 2623 CpG sites associated with smoking and methylation. I then took the top CpG sites that were less than P<1x10-8. I merged my eczema EWAS dataset with the smoking CpG sites and this gave 50 sites which had low p-values (P<0.05) for both variables. There were 56 when looking at methylation at age seven. Of these 50/56 I then took those with p-values less than 0.01 and this gave me 9 and 14 sites respectively, which is what I focused on in section 5.2.1.

CpG sites in Q1 (cord blood, ever eczema)					**CpG sites in Q2 (methylation at age 7, ever eczema)**				**Overlap of CpG sites**
CpG site	**Coefficient**	**P-value**	**95% confidence intervals**		**CpG site**	**Coefficient**	**P-value**	**95% confidence intervals**	
cg09570614	0.008	0.009	1.922, 0.013		**cg09570614**	0.006	0.03	0.0007, 0.011	✓
cg13689560	-0.011	0.003	-0.02, -0.004		**cg13689560**	-0.00	0.03	-0.005, -0.0003	✓
cg26728709	-0.004	0.008	-0.007, -0.001		**cg26728709**	0.005	0.04	0.002, 0.01	✓
cg00071265	0.006	0.043	0.0002, 0.01		**cg00496272**	-0.003	0.04	-0.006, -0.0002	x
cg00152041	-0.007	0.02	-0.013, -0.001		**cg00501876**	-0.005	0.007	-0.009, -0.001	x
cg00741986	-0.009	0.018	-0.016, -0.002		**cg00526336**	0.004	0.05	6.77x10-5, 0.007	x
cg01127300	-0.01	0.04	-0.02, -0.0004		**cg00639656**	-0.005	0.029	-0.009, -0.0005	x
cg01940273	-0.005	0.025	-0.009, -0.001		**cg01138448**	-0.006	0.03	-0.012, -0.001	x
cg02639359	-0.01	0.041	-0.02, -0.0004		**cg01294327**	-0.012	0.009	-0.022, -0.003	x
cg04263702	-0.008	0.012	-0.014, -0.002		**cg01561259**	0.012	0.005	0.004, 0.02	x
cg04361126	-0.011	0.013	-0.019, -0.002		**cg01692968**	0.0103	0.01	0.003, 0.018	x
cg04939496	0.008	0.05	0.0001, 0.015		**cg01744331**	-0.003	0.037	-0.006, -0.0002	x
cg05951221	-0.004	0.021	-0.007, -0.001		**cg01766850**	0.006	0.036	0.0004, 0.011	x
cg06434490	0.009	0.008	0.002, 0.016		**cg01839993**	-0.004	0.038	-0.009, -0.0002	x
cg06635952	0.011	0.03	0.001,		**cg01993576**	0.007	0.001	0.003,	x

			0.02					0.011	
cg07362537	0.01	0.014	0.002, 0.017		**cg03059073**	-0.007	0.03	-0.014, -0.001	x
cg09012001	0.005	0.04	0.0004, 0.01		**cg05593667**	-0.006	0.004	-0.011, -0.002	x
cg09022230	0.007	0.05	7.29x10-5, 0.013		**cg06130714**	-0.002	0.017	-0.004, -0.0004	x
cg09206294	0.007	0.002	0.002, 0.0112		**cg06439941**	-0.004	0.017	-0.008, -0.0008	x
cg09465703	0.009	0.029	0.001, 0.018		**cg07632771**	0.006	0.011	0.001, 0.01	x
cg09726654	0.004	0.027	0.0004, 0.007		**cg09219877**	-0.005	0.021	-0.009, -0.0007	x
cg10130088	0.006	0.036	0.0004, 0.013		**cg10151367**	-0.003	0.032	-0.006, -0.0003	x
cg10750182	-0.005	0.05	-0.01, -1.8x10-5		**cg11261850**	0.012	0.011	0.003, 0.021	x
cg11962640	0.003	0.05	6.79x10-5, 0.006		**cg11405655**	-0.007	0.001	-0.01, -0.003	x
cg12462247	0.009	0.012	0.002, 0.016		**cg11445634**	0.01	0.002	0.003, 0.016	x
cg12619504	-0.008	0.012	-0.014, -0.002		**cg13708645**	-0.007	0.036	-0.014, -0.0005	x
cg12803068	-0.02	0.03	-0.03, -0.001		**cg14404418**	0.007	0.025	0.001, 0.013	x
cg12877335	0.006	0.05	5.63x10-5, 0.011		**cg14614490**	0.005	0.041	0.0002, 0.01	x
cg13758913	-0.011	0.017	-0.02, -0.002		**cg14656043**	-0.004	0.028	-0.008, -0.001	x
cg13855261	-0.008	0.002	-0.013, -0.003		**cg15342087**	-0.004	0.04	-0.008, -0.0002	x
cg13914531	0.004	0.05	8.2x10-5, 0.009		**cg16611234**	0.009	0.034	0.0007, 0.016	x
cg15187398	-0.008	0.006	-0.014, -0.002		**cg17230002**	-0.007	0.038	-0.013, -0.0004	x
cg15417641	0.004	0.043	0.0001, 0.008		**cg17232357**	-0.011	0.013	-0.019 -0.002	x
cg16201146	0.005	0.023	0.001, 0.01		**cg17390562**	-0.004	0.049	-0.008 -2.8x10-5	x
cg16401465	-0.01	0.011	-0.018, -0.002		**cg18451588**	-0.003	0.002	-0.004 -0.001	x
cg16608652	0.004	0.028	0.0004, 0.007		**cg18704527**	-0.008	0.033	-0.014 -0.001	x
cg18158306	0.012	0.019	0.002, 0.022		**cg19211853**	0.004	0.049	1.9x10-5 0.007	x
cg18625627	-0.008	0.033	-0.015, -0.001		**cg19372602**	0.01	0.024	0.001 0.02	x
cg18961281	0.004	0.02	0.001, 0.008		**cg19593285**	0.007	0.011	0.002 0.013	x
cg19635644	0.005	0.007	0.001,		**cg20152539**	-0.011	0.007	-0.02	x

			0.009					-0.003	
cg23059461	-0.005	0.034	-0.01, -0.0004		**cg20722088**	0.0043	0.005	0.001 0.007	x
cg23771366	0.006	0.041	0.0003, 0.012		**cg20724032**	0.006	0.017	0.001 0.011	x
cg24859433	0.007	0.011	0.002, 0.012		**cg20902353**	0.006	0.047	9.14x10-5 0.011	x
cg25006998	-0.008	0.033	-0.016, -0.001		**cg21404980**	-0.005	0.028	-0.009 -0.0005	x
cg25560398	-0.01	0.032	-0.023, -0.001		**cg21446172**	-0.006	0.048	-0.012 -4.8x10-5	x
cg25953130	0.012	0.031	0.001, 0.023		**cg21869609**	-0.013	0.022	-0.024 -0.002	x
cg26077378	-0.002	0.035	-0.004, -0.0002		**cg22871253**	-0.009	0.009	-0.02 -0.002	x
cg26510500	-0.006	0.03	-0.012, -0.001		**cg22957360**	0.008	0.003	0.003 0.013	x
cg26908328	-0.007	0.01	-0.011, -0.002		**cg23621097**	-0.004	0.04	-0.007 -0.0002	x
cg27332104	0.005	0.001	0.002, 0.008		**cg23648810**	-0.006	0.013	-0.01 -0.001	x
					cg23842572	0.008	0.016	0.002 0.015	x
					cg24947694	-0.006	0.029	-0.012 -0.001	x
					cg25189904	0.012	0.018	0.002 0.02	x
					cg25503804	0.007	0.003	0.002 0.012	x
					cg26950531	-0.012	0.05	-0.023 -3.9x10-5	x
					cg27646484	-0.006	0.031	-0.012 -0.001	x

Appendix VI: QQ-plots for cat and dog exposure EWAS

Figure A5: Q-Q plot - Is there a relationship between methylation at age 15/17 and cat exposure in model (iv)?

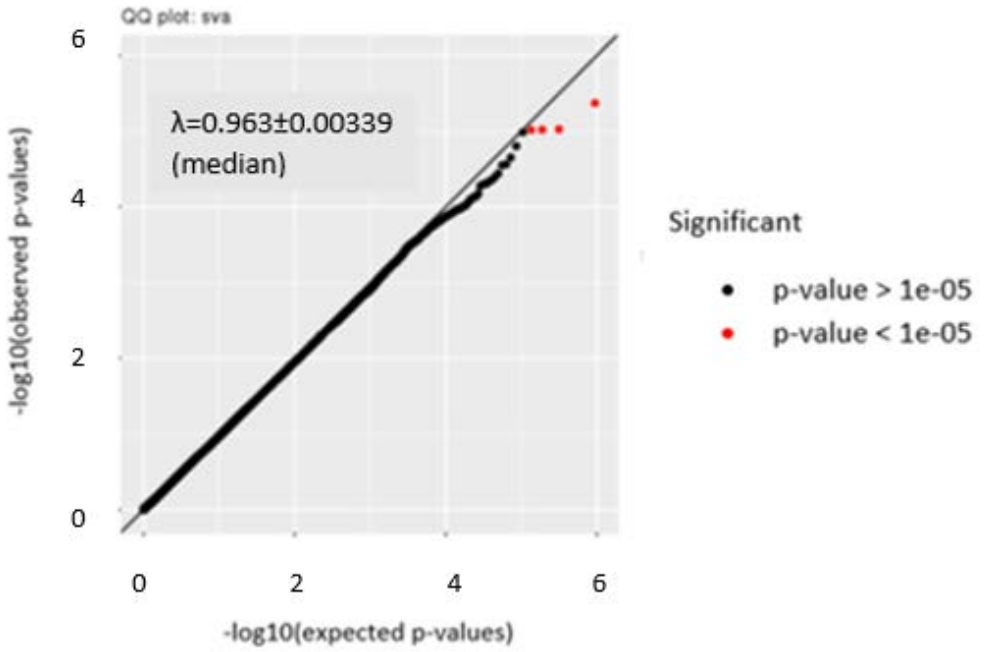


Figure A6: Q-Q plot - Is there a relationship between methylation at age 15/17 and cat exposure in model (vi)?

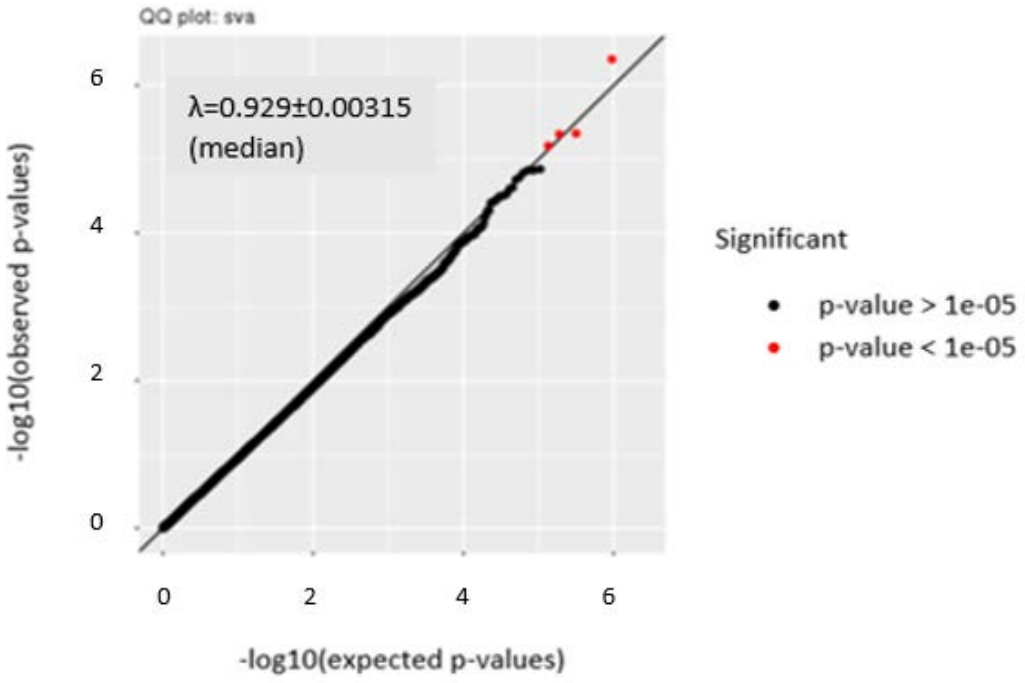


Figure A7: Q-Q plot - Is there a relationship between methylation at age 15/17 and dog exposure in model (iv)?

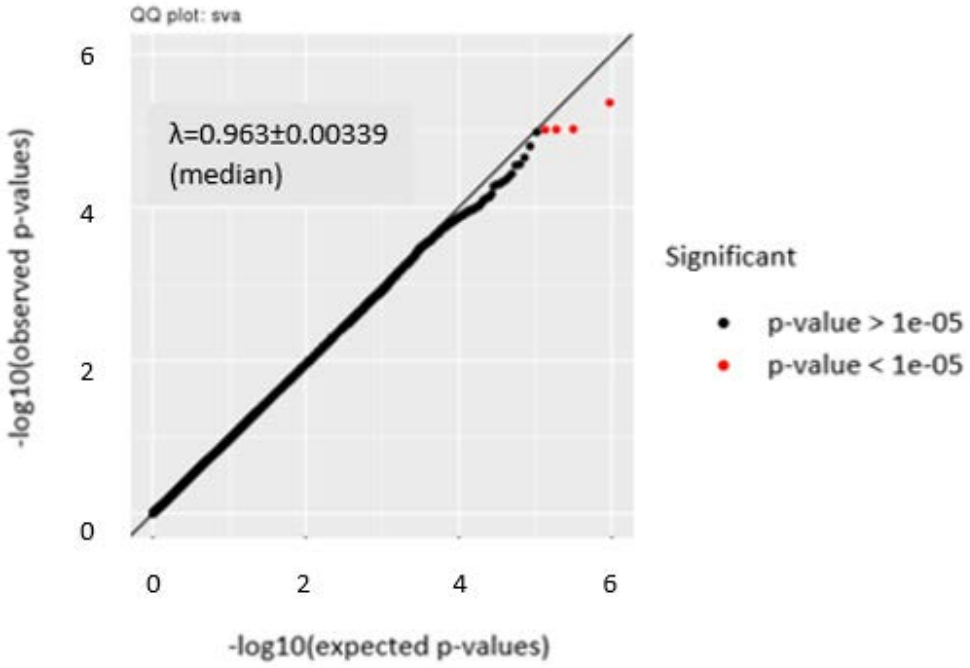
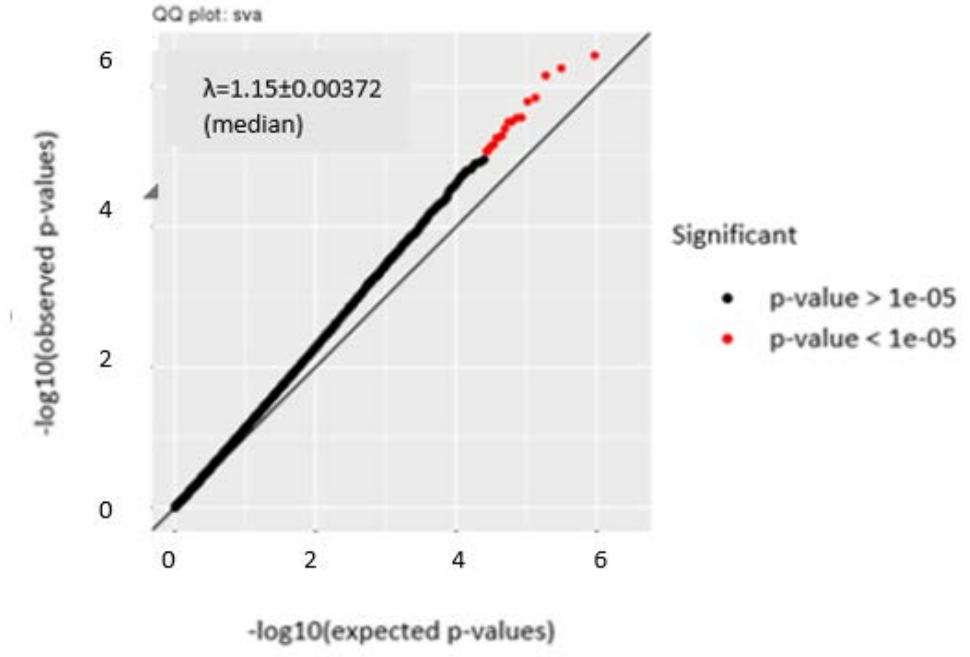


Figure A8: Q-Q plot - Is there a relationship between methylation at age 15/17 and dog exposure in model (vi)?



Appendix VII: Summary of findings

Table A8: Summary of important findings from each chapter

CHAPTERS	MAIN FINDINGS
4. Eczema EWAS	• There was evidence for 25 weakly associated GpG sites showing a relationship between DNA methylation and eczema. • There was a stronger association at the sites cg07721777 and cg26716834, when looking at methylation age 15/17 and eczema in the last 12 months. • When looking at continuation of associations through the four study questions, it was found that there was a similarly small p-value at P<0.05 at CpG sites cg26368024 and cg07166235 when looking at 'ever eczema' (both in cord blood and methylation at age seven) and methylation at age seven and eczema in the past 12 months. • Another two, cg04804139 and cg24211994 had small p-values in 'ever eczema' (both in cord blood and methylation at age seven), and methylation at age 15/17 and eczema in the last 12 months. • The CpG site cg01182386 has a similarly small p-value at P<0.05 when looking at methylation at age 15/17 and eczema in last 12 months, and methylation at 7 and eczema in the last 12 months. • None of the 25 CpG sites were in genes that had an association with eczema or other atopic illness.
5. Risk factors	• When looking at the risk factor smoking, there were 23 CpG sites with a potential association between smoking and eczema. 9 of these (out of 50) related to smoking during pregnancy and 14 (out of 56) related to a smoky environment. In addition one CpG site, cg19653589, which has an association with both eczema and smoking. We can hypothesise that DNA methylation acts as the mediator in this relationship.

	• Table 5.5 shows that there are three CpG sites, cg09570614, cg13689560 and cg26728709, which overlap when looking at two different timepoints, cord blood methylation and whether a child has ever had eczema, and methylation at age seven and whether a child has ever had eczema. These strengthen the association between smoking and eczema. • There were no associations between animal exposure (cat and dog) and eczema . • Looking at breastfeeding and eczema, there were two associations at cg04722177 and cg03945777, but these were not applicable because they occurred in cord blood. • None of the CpG sites identified had links with eczema or atopic disease.
6. Discussion	• There were 25 weakly associated GpG sites between methylation and eczema. Four sites in particular were shown to be particularly related, which occurred in question 4 of the study. • There were five CpG sites which showed similarly small p-values between different questions: there were two when looking at 'ever eczema' and methylation at age 7 and eczema in the past 12 months. There were another two looking at 'ever eczema' and methylation at age 15/17 and eczema again in the past 12 months. Lastly there was an association between a CpG site with small p-values when looking at eczema in the past 12 months and methylation at ages 7 and 15/17. • There were 23 CpG sites which showed a low p-values in both 'smoking and methylation' and 'methylation and eczema', and one link at CpG site cg19653589 between smoking and eczema, potentially mediated by methylation. Three CpG sites, cg09570614, cg13689560 and cg26728709, overlapped when looking at smoking and eczema. The genes these CpG sites were in did not have an association with eczema or atopy. • There were no associations between animal exposure (cat and dog) and eczema . • There were two associations between breastfeeding and eczema at cg04722177 and cg03945777 but because these occurred in cord blood they are not significant.