Investigating the DNA methylation profiles of children with oligoarticular juvenile idiopathic arthritis (JIA).

Raul Antonio Chavez Valencia

Submitted in total fulfilment of the requirements of the degree of

Doctor of Philosophy

Department of Paediatrics

Faculty of Medicine, Dentistry and Health Sciences

The University of Melbourne

December, 2019

Abstract

Juvenile idiopathic arthritis (JIA) is a complex autoimmune disease affecting children aged between 6 months and 16 years. JIA represents a group of 7 subtypes of disease, with the most common being oligoarticular JIA (oJIA). Despite a prevalence of up to 1 in 400, rates similar to those in T1D, JIA research is relatively sparse. Research into disease pathogenesis has largely focussed on genetic risk factors, and has also identified CD4+ T-cells as likely to mediate the autoimmune process. However, research is particularly needed regarding diagnosis and prognosis of disease and its outcomes. Currently, diagnosis is almost entirely dependent on clinical observation and history, with little in the way of biomarkers to classify patients or to guide clinical management.

Epigenetics represent biological modifications to DNA and chromatin that control gene expression and chromatin structure. DNA methylation is perhaps the most accessible modification available for study, and is known to modulate immune cell function particularly amongst CD4+ T-cell subsets. A number of autoimmune diseases have reported significant DNAm associations, and have also provided intriguing data on the potential of DNAm to predict clinical outcomes. This study hypothesised that DNAm is important in oJIA pathogenesis, and potentially provides a biological basis for the diagnosis and prognosis of disease. This study utilised CD4+ T-cells and a case-control study design to analyse the associations between DNAm and oJIA, with data generated from the Illumina Infinium HumanMethylation450 BeadChip array. Cases were matched with controls according to age and sex. Further, cases were subtyped according to current diagnostic criteria and had active disease, both of which attempted to ensure all cases were clinically homogeneous.

The first aim was to profile DNAm in oJIA cases compared to controls. Processing of data through analysis pipelines resulted in high quality data. Differential methylation analysis suggested that oJIA cases and controls could be segregated in cluster analysis using DNAm data, despite no genome wide significant hits being produced. Immune system pathways analysis suggested the top hits were relevant to disease, being enriched for receptor binding of cytokines such as IL6, IL17 as well as MHC class II. In addition, a number of top ranking probes were enriched within cell death and survival functions. Indeed, gene expression data suggested genes within those pathways were also correlated with DNAm. Technical validation of a selection of probes was highly successful, with all probes validating. A small replication study, however, was not able to reproduce these findings. Of particular note, a

wide distribution of DNAm values was observed for many of the validated probes. Since technical validation was so successful, this DNAm heterogeneity potentially derived from sample group heterogeneity, which may well have played a part in difficulties replicating data.

Therefore, biological sources of heterogeneity were explored in chapter 5, focussing primarily on the genetic associations with DNAm. Probes utilised for technical validation were analysed for genetic associations associating with either mean or variable DNAm. Both analyses suggested that the most robust associations were for known mQTLs and enhancer SNPs. Indeed, DNAm differences according to genotype were up to 13% and 27% for 2 probes analysed, representing a many-fold difference over case-control differences (typically ~5%). Combined with an intermediate level of minor allele frequency for many of these robustly associated SNPs, these mQTLs represent a likely source of biological variation contributing to oJIA DNAm variation. These minor allele frequencies increase the likelihood of inadvertent sampling bias, potentially resulting in difficulties in replicating DNAm data. Deeper analysis provided some initial indication that these mQTLs may also be potential oJIA risk loci, with the most significant associations again coming from known mQTL or enhancer SNPs. This also suggested DNAm data may well identify regions of interest for genetic risk loci discovery.

The final chapter hypothesised that sources of potential clinical heterogeneity not captured within current classification criteria may well lead to DNAm heterogeneity, as could recognised subgroups within oJIA. Of primary focus, age of disease diagnosis was assessed for associations with DNAm. This study found that case-control analyses of older diagnosed samples (≥6 years) resulted in case-control clustering using far fewer probes. Indeed, the reduction of probes required for clustering was more pronounced in the analysis of younger diagnosed samples (<6 years of age), and also resulted in a genome wide significant hit. These subgroups represented 2 highly divergent populations, since top ranking probes from each subgroup had virtually no overlapping probes. This data suggested that age subgroups in oJIA represent sources of sample heterogeneity, leading to DNAm heterogeneity. Technical validation for a large majority of the select probes from the younger-diagnosed analysis was also successful. However, a small replication study could not reproduce these initial findings. In light of the potential for mQTLs to have pronounced effects on DNAm, as explored in chapter 5, larger replication groups (or, indeed, discovery groups) will likely be needed to mitigate the risk of sampling error to enable reproduction of findings.

OJIA heterogeneity was also explored by looking at known subgroups, Persistent vs Extended disease. A number of oJIA cases would go on to develop extended disease, and the possibility existed for DNAm signatures to identify these cases prior to disease extension. This was indeed the case, with an exploratory analysis suggesting a number of probes can cluster persistent cases from extended-to-be cases. Further, these probes were able to produce a highly sensitive and specific test to predict disease extension, thereby providing a proof of principle for a prediction test using DNAm data.

This study is the largest case-control analysis of JIA DNAm to date, and provided insights into the potential for DNAm to identify pathogenic pathways, identify sources of oJIA heterogeneity, and opened the possibility for biological markers of disease to be used in clinical management. The findings regarding the pronounced effect of mQTLs on DNAm also suggest that genetics is a large source of DNAm variability, far larger than group differences typically found in a complex diseases (such as oJIA). The identification of subgroup specific differences, even with a clinically homogeneous subtype, warrants further investigation to explore potential differences in pathogenesis between age groups and the use of DNAm as biomarkers for classification or disease management.

Declaration

This is to certify that:

1. The thesis comprises only my original work towards the Doctor of Philosophy except where indicated in the Preface;
2. Due acknowledgement has been made in the text to all other material used;
3. The thesis is less than 100,000 words in length, exclusive of tables, figures, bibliographies and appendices.

Raul Antonio Chavez Valencia
Department of Paediatrics
The University of Melbourne
Parkville, Victoria 3052
Australia

Preface

Contributions

The author performed all the work presented in this thesis except where contributions from others have been acknowledged below.

Associate Professor Justine Ellis and Associate Professor Jane Munro founded and led CLARITY (ChiLdhood Arthritis Risk factor Identification study) and the resulting biobank, from which samples for this study were drawn from and were the foundation for this study.

The team within the Genes, Environment and Complex Disease group at the Murdoch Children's Research Institute (MCRI) recruited all patients to CLARITY, collected clinical information and performed blood collection, all of which were used as part of this thesis.

The MCRI Biobanking Facility staff, who performed blood sample processing, cryopreservation of mononuclear cells and managed the biobank on behalf of CLARITY. Samples in this biobank were used as part of this thesis.

Dr Rachel Chiaroni-Clarke generated gene expression data used in the discovery phase of this EWAS study.

Dr Braydon Meyer generated some of the cell sorting data utilised in the optimisation chapter of this study.

Dr David Martino and myself generated the vitamin D exposure EWAS data prior to the commencement of this study, the data from which was used for development and optimisation of the EWAS probe selection criteria within the optimisation chapter of this study.

Dr Benjamin Ong (MCRI) acquired spectra for all MALDI-TOF mass spectrometry experiments used in this thesis.

Dr Matthew Burton and Mr Paul Lau (MCRI) performed flow cytometry on all FACS sorted samples used in this thesis.

Editorial assistance for Chapter 1 of this thesis, consisting of copyediting, was provided by Johanna O'Farrell.

Publications

Parts of work presented within this thesis has contributed to the following peer-reviewed journal articles:

The DNA methylation landscape of CD4+ T cells in oligoarticular juvenile idiopathic arthritis. Chavez-Valencia RA, Chiaroni-Clarke RC, Martino DJ, Munro JE, Allen RC, Akikusa JD, Ponsonby AL, Craig JM, Saffery R, Ellis JA. J Autoimmun. 2018 Jan;86:29-38. (Parts of this publication are included in Chapter 4 and chapter 6, and is presented in the appendix).

In vitro exposure of human blood mononuclear cells to active vitamin D does not induce substantial change to DNA methylation on a genome-scale. Chavez Valencia RA, Martino DJ, Saffery R, Ellis JA. J Steroid Biochem Mol Biol. 2014 May; 141:144-9. (Chapter 3 of this thesis contributed to this publication, and is included in the appendix).

DNA methylation at IL32 in juvenile idiopathic arthritis. Meyer B, Chavez RA, Munro JE, Chiaroni-Clarke RC, Akikusa JD, Allen RC, Craig JM, Ponsonby AL, Saffery R, Ellis JA. Sci Rep. 2015 Jun 9;5:11063. (Chapter 3 of this thesis contributed to this publication, and is included in the appendix).

Short courses and internships

2016 MCRI Business development internship program

2017 Davies Collison Cave Patent law internship program

2017 Biomedical Research Victoria & Cancer CTx CRC STEMM bootcamp

Funding

During my PhD I was fortunate enough to be supported by the following competitive funding schemes:

- 2015 Arthritis Australia Marion A Simpson Grant
- 2014 Arthritis Australia Scholarship (The Arthritis Australia State and Territory Affiliate grant, funded by Arthritis South Australia).
- 2016 Student Conference Support Scheme, Murdoch Children's Research Institute.
- Henry and Rachael Ackman Travelling Scholarship, Department of Pediatrics, The University of Melbourne (2016).
- Graduate Research Travel Scholarship, Department of Paediatrics, The University of Melbourne (June 2016).
- 2013 Bright Sparks PhD pitch finalist. Amplify festival

The CLARITY study was supported by grants from the Australian National Health and Medical Research Council (NHMRC #APP1026349 to Dr Justine Ellis) and the Victorian State Government Operational Infrastructure Program, along with a number of philanthropic organisations including Arthritis Australia and the Rebecca Cooper Foundation.

Ethics Clearance

This study was approved by the Royal Children's Hospital Human Research Ethics Committee (HREC 27127Q), and Monash Medical Centre's Research Ethics Committee (HREC 12377).

Acknowledgements

This work is a testament to the support given to me while I completed this PhD during some of the most dynamic, wonderful and trying times of my life which included welcoming 3 new children into the world, multiple times of personal loss and heartache and everything else life throws at us.

To my Primary supervisor, Associate Professor. Jeff Craig, thank you for being a constant support over my project. From your insights on the science, regular feedback on chapters/abstracts/papers to the personal support, I deeply appreciate it all. A special mention to your invaluable encouragement as I explored career options, which allowed me to drive my career in a direction that is rarely explored and even less so encouraged within academia.

To Justine, thank you for believing in me and supporting me enter this phase of my career, from financial support to scientific guidance and discussions, pushing when I needed it and support over many of the trials.

Thank you Richard for your input into the epigenetic aspects of my project, and your encouragement early in my PhD.

To David Martino, who mentored me through the EWAS analysis and guided me as I learned to code. Thank you for your patience, constant encouragement and advice as I climbed that steep learning curve. You also acted as a career mentor, whether you knew it or not! It was great to see how you go about things and have your friendship over this time.

To Rachel Chiaroni-Clarke, thanks for being part of this journey with me. It was just me in the lab for a long time so having someone to share all the up's and down's and what if's and exciting moments was a real blessing. Thank you and I pray for nothing but joy, peace and a wonderful journey ahead for you.

To the ladies of the Genes, Environment and Complex disease group – Betty, Hannah, Kate, Amanda, Susan, Anna C and Anna D – your work was tremendous and it was a pleasure to turn up and see you all. From sharing cake to venting about frustrations, sharing Christmas parties together and working on this research together, you showed yourselves to be the

salt of the earth, the spice of life, and the light of the world. Anna C, thank you for always looking out for us students and making sure we were fed, amongst many other things! It was in those little moments I really felt cared for.

To the Business development office guys, particularly Lishia, Jo and James. Thank you for the opportunity to participate in the internship program, and your patience as an eager student made many mistakes! This internship opened up many doors for me, and allowed me to explore an industry that ultimately became the focus of my career. I'm deeply grateful for the opportunity.

To the MCRI community, the lunchtime soccer guys, the IT guys, Dr Ben Ong, other PhD students, especially members of the Orange Pod and the Fosang group. Thanks for being great neighbours, and being so open to sharing the space with me and our team!

Thank you to the Dr. Ben Ong and the members of the tissue culture and biobanking team, especially Anushka, Lucy and Ronita, who were ever so ready to help with accessing samples for this study, always with a smile. I really enjoyed working with you, next to you and all the lovely chats we had over the years. Dr. Ben, it was a lot of work but a lot of fun over the years. Thanks for showing me the ropes, introducing me to many MCRI people, all the laughs and all your support. You always do things with a smile and are always focussed on supporting researchers.

Finally, to my wife and family. I'm sorry. Johanna, I couldn't have asked for better love and support. The volume of trials that we endured in such a short period of time early in my PhD was only bearable because we were together. Especially in this last period of time to submission, you took so much on and kept going even when you were exhausted, just to allow me the time and space to work. Our kids' ability to love and help and keep having fun is because you kept making the space and time for them to flourish. I thank God every day for you. I will tell all and all will know your name.

Thanks be to God, who filled our hearts with His presence at the worst of times, and continued to call me closer during my most selfish of times. Few will see this work, but You have seen it all the way through. Let this be a gift and a prayer back to You, however small it is, that You may be present for all students and families undertaking this challenge.

Table of Contents

List of Figures

List of Tables

Abbreviations

Term	Description
A	Adenine
AGRF	Australia Genome Research Facility
Alu	Arthrobacter luteus (Alu) restriction endonuclease sensitive stretch of DNA, usually short in length
ARSA	arylsulfatase A
bp	Base pair
C	Cytosine
CD14	Cluster of differentiation 14
CD19	Cluster of differentiation 19
CD3	Cluster of differentiation 3
CD4	Cluster of differentiation 4
CD8	Cluster of differentiation 8
CI	Confidence interval
CLARITY	ChiLdhood Arthritis Risk factor Identification sTudY
CpG	Cytosine-phosphodiester bond-guanine
CYP1A1	Cytochrome P450, family 1, subfamily A, polypeptide 1
DAPI	4',6-diamidino-2-phenylindole
DMP	Differentially methylated probe
DMSO	Dimethyl sulfoxide
DNA	Deoxyribonucleic acid
DNMT	DNA methyltransferase
DNase	deoxyribonuclease
dNTP	Deoxyribonucleotide triphosphate
EDTA	Ethylenediaminetetraacetic acid
EWAS	Epigenome wide association study
FBS	Fetal bovine serum
FDR	False discovery rate
G	guanine
GWAS	Genome wide association study
HLA	Human leukocyte antigen
ILAR	International League of Associations for Rheumatology
kb	kilobase
LINE	Long interspersed nuclear elements
MCRI	Murdoch Children's Research Institute
MDS	Multi-dimensional scaling

MHC	Major histocompatibility complex
mL	milliLitre
mM	milliMolar
ng	nanogram
NHMRC	National Health and Medical Research Council
nM	nanoMolar
PBMC	Peripheral blood mononuclear cell
PBS	Phosphate buffered saline
PCR	Polymerase chain reaction
POU2AF1	POU Class 2 Homeobox Associating Factor 1
POU5F1	POU Domain, Class 5, Transcription Factor 1
RGS12	Regulator Of G Protein Signaling 12
RNA	Ribonucleic acid
RUNX2	Runt-related transcription factor 2
SNP	Single nucleotide polymorphism
T	thymine
T1D	Type 1 diabetes
TACSTD2	Tumor Associated Calcium Signal Transducer 2
Treg	T regulatory cell
Tss1500	200–1500 bases upstream of the Transcription start site
µL	microLitre

Chapter 1. **Introduction**

1.1 What is JIA? Life begets Art begets life

Juvenile arthritis (JIA) is the most common form of childhood rheumatic disease. It is an autoimmune disease largely affecting the joints where there is an abnormal immune response to self. Currently, JIA diagnosis is based on the exclusion of known causes of childhood arthritis [1]. Therefore it includes all forms of childhood arthritis having an unknown cause, beginning before age 16 and persisting for more than 6 weeks [1, 2]. JIA can cause serious short term and long term disability in the developing body [1]. Whilst treatments have vastly improved over the last 10 years with the introduction of antibody based drugs targeting cytokines and their receptors, these are focussed on symptom relief and are not curative [1].

Historically, JIA was defined first by either Cornil in 1864, Charcot in 1853 or possibly Herbeden in 1803 [3]. However, JIA's existence in the population potentially spans over 500 years. Life with JIA is thought to be depicted through Carravaggio's "Il Amore Dormient" (Sleeping Cupid) in 1608 and in Botticelli's "Portrait of a Youth" in 1483 [3-6]. More recently in 1997, JIA was defined as a group of diseases subtyped into 7 categories, with further refinement into fewer subtypes likely in the near future [7] [2].

Figure 1-1 Potential early depiction of juvenile arthritis in Caravaggio's 'Sleeping Cupid'.

JIA can potentially be observed through swollen fingers, a lump on the wrist akin to a synovial cyst, distension of the abdomen and potentially shortened thigh amongst other consequences of long term disease [4].

Estimates of JIA prevalence vary between 0.1 in 1000 to 4 in 1000, depending on ethnicity and study design [2]. Despite a similar prevalence to Type 1 diabetes, another childhood autoimmune disease, comparing PubMed search results suggests JIA publications lag behind T1D publications ~7:1, and even more so compared to rheumatoid arthritis (10:1). This relative lack of JIA research exists despite a long history of JIA in art and clinical history.

Currently, one of the central challenges in JIA is the distinct inability to diagnose and classify the disease quickly and accurately. Indeed, this challenge extended into debate on the correct diagnosis of the child in 'Sleeping Cupid' i.e. juvenile arthritis versus severe rickets [4, 5]. The difficulty in diagnosis and the lack of biomarkers means a diagnosis is not definitive at initial presentation, and is only made after 6 months of clinical observation post presentation [8]. This six month mark is particularly crucial for the subtypes of the two most common subtypes, oligoarticular JIA and polyarticular JIA (described below) [8]. Once a JIA diagnosis is made, classification of a patient into a subtype is also completed. Yet, subtyping of patients remains a controversial topic due to the known heterogeneity of patients within many of the subtypes [2]. This, again, is likely due to a lack biomarkers and clinical nature of the classification criteria [8]. Clinically based criteria is insufficient for classification purposes, since superficial clinical manifestations may be the result of a number of different underlying biological processes. This underlying biology is critical since it may define response to treatment [9]. Therefore, a better classification criteria is needed, ideally with biomarkers which subtype patients based on similar underlying biology.

In this chapter the current classification of JIA, heterogeneity inherent to these subtypes and immunopathology of JIA is discussed to identify potential biological markers that may address clear problems faced in the classification criteria.

1.2 Classification of JIA: Subtypes

Art mirrors life in the JIA field. Whether considering juvenile arthritis in images such as Caravaggio's "Il dormiente amore", or JIA classification systems in the research arena, both seem to be centred on the theme of debated diagnoses. Criteria for the classification of JIA into subtypes have been internationally agreed upon, where 7 subtypes of JIA are defined according to expert clinical consensus and in fact were designed to facilitate research [8, 10, 11]. However, this subtyping is controversial due to heterogeneity within subtypes [9, 11]. Whilst some subtypes do appear to be more homogenous than others, other subtypes likely

represent a heterogeneous group of diseases, potentially based on prognosis, age, response to treatment or severity of disease [9, 12-18].

Treatments are currently targeted towards these subtypes, despite the existence of non-responding subgroups due to patient heterogeneity [2, 9, 19, 20]. However, biomarkers for treatment response are not yet clinically available [9]. Obtaining homogenous subtypes based on biomarkers of underlying biology and/or response to treatment will have definitive positive benefits for patients. This includes better clinical trial design, improved trial outcomes due to more homogenous treatment groups and subsequent reduction of futile and burdensome treatment in non-responders [9]. Predictive biological markers may also allow earlier treatment and better protection of joint function [9]. The subtypes are outlined below, highlighting the inherent clinical variation and need to modify the existing classification criteria. These demonstrate that even within clinically homogenous subtypes, clinically relevant heterogeneity exists that may be elucidated through biological markers. Further, a number of studies have begun to identify potential biomarkers associated with clinical outcomes or response to treatment, highly relevant to the clinical management and welfare of JIA patients.

1.2.1 Subtypes with a clinical basis for classification

1.2.1.1 Systemic Arthritis

The subtype of systemic arthritis (sJIA) originally bore the name Still's Disease, after George Frederic Still who is credited with its early description [21]. It occurs in about 5-15% of North American and European JIA patients and affects males and females equally [21] [1]. It is perhaps the most severe of all the subtypes, with 10-20% of patients suffering from macrophage activation syndrome in which patients present with high, non-remitting fever, central nervous system dysfunction and can result in life threatening multi-organ failure [1, 22]. This disease has clinical hallmarks of autoinflammatory syndromes, as well as strong HLA class II associations consistent with an autoimmune disease [23] [24] [21]. Heterogeneity in response to treatment in clear in this subtype, with a subset of sJIA patients being non-responsive to antibody treatment to cytokines such as IL-1, IL-6 or their receptors [19, 25, 26]. The clinical or biological characteristics of non-responders are not yet clear [2, 19]. However, promising early research suggests a biomarker may identify responsive subgroups using MRP8/14 levels in serum [9, 27, 28]. In addition, this may also predict relapse after treatment [28].

1.2.1.2 Psoriatic Arthritis

Psoriatic arthritis (PsA) occurs in females more often than males, and can be a difficult subtype to diagnose [1, 29]. Diagnostic criteria tends to overlap with other subtypes but over time the exclusion criteria, as well as revisions to the inclusion criteria, have facilitated diagnosis [1, 29]. This criteria includes: presence of dactylitis or nail pitting or psoriasis or family history in a first degree relative of any of these, [2, 29]. Age of onset appears to mirror other subtypes such as oligoarticular JIA and polyarticular JIA, showing a bimodal distribution with peaks at 2-4years and 9-11 years of age [1]. Heterogeneity in this subtype can potentially result in subgrouping patients according to anti-nuclear antibody (ANA) positivity or age at onset, similar to reports in other JIA subtypes [1, 30-32]. It has been proposed that ANA positivity may help define a subgroup of PsA patients together with another JIA subtype, oligoarticular JIA, since the clinical features tend to overlap [2, 30]. Another potential classifier, age at onset, grouped patients with polyarticular JIA as well, particularly when considering IgG glycosylation levels [30]. Whilst ANA status was not useful when interpreting IgG glycosylation levels, using age at onset alone did appear to show JIA patients with far more variability in glycosylation than age matched controls [30]. In another study, chart review of PsA and oligoarticular JIA patients concluded that in addition to similarities in age at onset, sex ratio and ANA status, younger PsA patients more typically displayed hand and wrist joint involvement, mirroring oligoarticular JIA patients with a younger age of onset [31, 32]. Further, these younger PsA patients needed almost twice the length of time for treatment to achieve clinical remission [31]. Thus, heterogeneity based on age of onset in PsA mirrors that seen in other JIA subtypes and may be a variable important for better classification systems.

1.2.1.3 Undifferentiated Arthritis

This is a subtype that includes all diagnoses that do not fit any other subtype criteria. Naturally, this subtype will be heterogeneous in the sense that it can incorporate patients that incompletely fit multiple subtypes [1]. Before the ILAR criteria were revised for JIA in 2004, studies suggested many patients classified into a single subtype could also fit into the undifferentiated subtype [1, 8]. This was amended in subsequent revisions [1]. However, clinical experience is that many patients sit in this subtype category until revision of diagnoses after 6 months or later, whereupon a definite singular classification is made (personal communication, Dr Roger Allen, Royal Children's Hospital, Melbourne).

1.2.2 Subtypes with established biological bases for classification

1.2.2.1 Polyarticular JIA

Polyarticular arthritis (polyJIA), is a subtype of JIA affecting 5 or more joints in the first 6 months of disease [8]. This subtype makes up approximately 15-25% of all JIA diagnoses in North American and European studies [33]. Reports exist of 50% prevalence of polyJIA diagnosis in countries such as Kuwait, India and the African continent, however this is potentially biased by late presentation of patients years after disease onset [34, 35].

This subtype has established biomarkers which enable patients to be diagnostically split by IgM antibody reactivity to rheumatoid factor, either positive (RF+) or negative (RF-) [8]. PolyJIA patients who are RF+ appear to be more homogenous than those who are RF- and may be a subtype of its own, regardless of joint count [2, 36]. Prognosis of polyJIA patients may be poor, particularly if there is RF positivity, hip, cervical spine or early hand involvement amongst other factors [2, 36]. The short and long term effects of treatment in polyJIA are varied. Whilst poly-JIA patients on bDMARDs (biological disease modifying anti-rheumatic drugs) are free of disease flares during clinical trials, serious adverse reactions are significantly increased compared to those treated with first line treatments (methotrexate) [37-39]. Prognosis for polyJIA patients is relatively poor, with only 34% reporting no active joints during the last year of Vidqvist *et al's.* study, despite nearly 50% of polyJIA patients receiving bDMARD treatment at some stage [39].

1.2.2.2 Poly-articular Arthritis: Rheumatoid Factor positive

This subgroup of polyJIA is mainly seen in adolescent females and is thought to be very similar to adult RF-positive rheumatoid arthritis [2, 33]. RF positivity in this polyJIA subgroup is likely to define a more homogenous disease [2, 33]. These patients are observed to be have reduced rates of remission, higher proportions on steroids and bDMARDs and are more likely to have poorer outcomes [40]. These patients typically have symmetrical involvement of the small joints in the hand and feet [1]. Of particular note, this subgroup can present with rheumatoid nodules (small lumps of fibrous tissue) in as many as 1/3 of patients [1]. The rarity of these nodules in other subtypes, and the poor clinical outcomes also point to rheumatoid factor positivity as a distinct subtype of JIA.

1.2.2.3 Enthesitis related arthritis

Similar to sJIA, this subtype appears to be a different disease compared to other JIA subtypes. It is also a disease primarily affecting the lower joints, and is the only subtype

where males present more commonly that females [1, 41]. Enthesitis related arthritis (ERA) is thought to be a form of spondyloarthropathy with sacro-iliac joint progression occurring in a large proportion of cases, unlike other types of JIA [2, 41]. As is common in spondyloarthropathies, antibody reactivity to an antigen presenting molecule, HLA-B27, is found in most patients and constitutes part of the diagnostic criteria [1]. This subtype represents an example of the benefits of a biological marker for classification purposes, which correlate with clinical progression towards specific joint involvement. Classification criteria is based on identifying at least 2 of a number of different clinically apparent manifestations, including enthesitis, dactylitis, HLA-B27 positivity, being male >6years of age, nail pitting and a number of others, suggesting that this subgroup may contain patients with diverse underlying biological processes [8]. Whether these clinical differences reflect differences in biological processes is unknown.

1.2.3 Subtypes with potential for biologically based subgroups

These subtypes, as with others, are classified according to clinically apparent similarities upon initial presentation [2, 42, 43]. However, literature suggests that heterogeneity in clinical outcomes and response to disease may be reflected or even pre-empted within biological measures [2, 13, 15-17, 32, 42-51]. These could be combined with clinical presentation to enable better clinical management of patients [2, 13, 15-17, 32, 42-51]. The subtypes with known patient heterogeneity, together with mounting data regarding associated or predictive biological markers, are presented below.

1.2.3.1 Poly-articular arthritis: Rheumatoid factor negative

Patients negative for rheumatoid factor (RF-) occur in approximately 85% of all polyJIA cases, but this is known to vary according to ethnicity [33]. Patients with RF- polyJIA are distributed bimodally according to age, peaking at 1-3yrs and again in late childhood [33]. In contrast to RF+ polyJIA patients, subgroups within RF- polyJIA patients have been identified based on age at disease onset. In one example, gene expression data suggests biological processes related to B cells cluster RF-polyJIA with similar age at onset [12]. This grouping, based on age at disease onset, also clustered patients from another subtype of JIA, suggesting biological similarities across subtypes [12]. Biological similarities amongst RF- polyJIA females with older age at onset is also suggested, with an even stronger female preponderance compared to patients with young age at onset (10:1 vs 4:1 respectively, female to male ratio) [33]. Age at onset is again highlighted in Ravelli *et al.* 2011, where identification of antibody reactivity to nuclear proteins (ANA+) appears to cluster patients

with common clinical and disease progression characteristics such as age at onset, fewer overall joints affected and higher risk for eye disease [17]. These differences are seen by some as being able to group RF-polyJIA patients into at least 2 subgroups based on age of onset or RF positivity [2, 16] . These observations emphasise the heterogeneity inherent to the current classification criteria, particularly in the RF- polyJIA subgroup, and the need for biological determinants of diagnosis in addition to clinical markers [2].

1.2.3.2 Oligo-articular arthritis

Oligoarticular JIA (oJIA) is the most common presentation amongst all the JIA subtypes, making up ~27-56% of all JIA diagnoses [1]. This proportion does vary according to geographical region where the polyJIA subtype can sometimes make up the larger proportion of diagnoses (perhaps confounded by delay in diagnosis) e.g. in African or Sub-continental Asian regions [34, 35] [52, 53]. It typically has an early age of onset, with a peak age of onset of 2-4 years, with females affected in larger proportions than males [1]. Patients often have asymmetrical joints affected, with large joint involvement e.g. knees. Early studies suggested that different HLA alleles associated with different JIA subtypes, yet more powered studies suggested that HLA-DRB1 position 13 was the key HLA allele in JIA [54] [14].

1.2.3.3 Anti-nuclear antibodies potentially define oJIA subgroups

OJIA appears have further patient heterogeneity. Prakken and colleagues have argued that biological markers such as ANA positivity (along with age at onset) may better describe a group of patients, from both oJIA and polyJIA subtypes, with a similar disease course [15, 17] [47, 49]. Using this, some patients once separated into oJIA or polyJIA subtypes would be re-classified together into a new subtype e.g. <6 years age & ANA+ [50]. Later studies from Ravelli *et al.* extended the application of this new group to patients from further subtypes, including patients from psoriatic disease and the undifferentiated subtype [17]. Whilst numbers were smaller in these other subtypes, trends suggested that ANA status is a potential source of heterogeneity [17]. In support of Ravelli 2011, a paper from Stoll *et al.* mirrors the Psoriatic subtype finding [32]. Using this new classification, patients were much more likely to develop eye disease (iridocyclitis) [17, 50]. Therefore, these patients are now directed to regular ophthalmologist assessment to detect and treat eye disease early, thus demonstrating the potential benefits of identifying patient subgroups.

However, debate over the adoption of a new subtype based on age and ANA status continues. Further, the test for ANA is not without significant shortcomings due to lack of

specificity in JIA [55, 56]. Regarding the current gold standard ANA test, using indirect immunofluorescence (IIF), not one but a large spectrum of different antigens are captured in this test [57]. This is relevant when considering data showing that ~75-80% of oJIA patients with +ve ANA status do not go onto develop uveitis [17, 49]. This suggests a subgroup of ANA+ oJIA patients exists that will develop uveitis, perhaps grouping by specific antigens. This lack of specificity is currently leading to large numbers of needless ophthalmological follow ups, for which only ~20-25% will gain benefits.

The potential for uveitis specific ANA antigens is supported by data from ELISA based ANA testing. Currently, ELISA can only test against a subset of ANA antigens compared to IIF [58] [59]. This potentially excludes uveitis specific antigens. This appears to be supported in Nordal *et al.*'s data (based on ELISA methods) which found no relationship between uveitis and ANA status, suggesting the antigen subsets associated with uveitis were not present on the ELISA test [60].

The potential of ANA antigen subsets in JIA and uveitis is further supported when looking at concordance data between IIF and ELISA. Studies in JIA demonstrate only 21% & 36% concordance between IIF and ELISA, suggesting that antigen subsets resulting in ANA positivity in IIF were not completely present in ELISA tests [55, 56]. This inability for ELISA to detect sera with specific ANA's has been reported [61]. Indeed, this may be an issue specific to JIA, since many previous studies largely show concordance in IIF and ELISA results from patients with SLE and other systemic rheumatologic diseases [62-65].

Taken together, this data suggests using ANA as biomarker currently sends a large proportion of patients towards needless ophthalmological follow up. This is likely due to the unspecific nature of IIF tests for ANA. Unfortunately, ELISA based tests potentially fail to test for specific ANA antigens relevant to JIA or uveitis [55, 56]. Since no better tests exists for predicting uveitis onset, IIF based ANA testing for identification of JIA patients at risk of uveitis remains necessary. However, to prevent health system wastage and burden on families through years of futile ophthalmological testing, a serious need exists in defining a better biomarker. That is, a test which detects the specific ANA antigens or that uses other clinical or biological markers relevant to JIA associated uveitis.

1.2.3.4 Persistent vs Extended disease subgroups

OJIA is categorised according to 2 subgroups, persistent and extended disease (oJIA-pers and oJIA-ext respectively). OJIA-pers is defined as patients having <5 joints affected in the first 6 months [8], whilst oJIA extended disease (oJIA-ext) is based on observation of the disease after the first six months, and is characterised by 5 or more joints affected in total [53].

oJIA-ext can occur in up to half of those diagnosed with oJIA [66-68]. With a more polyarticular disease course, outcomes tend to be worse for patients with oJIA-ext. These patients are less likely to achieve clinical remission off medication, more likely to flare once in remission, have more time in active disease state, have lower functionality scores and higher chance of disability [67, 69, 70].

The heterogeneity in clinical outcomes between persistent and extended disease may be mirrored across a number of biological measures. Immunologically, oJIA-ext is characterised by altered CD4:CD8 ratios, higher levels of Th17 cells in the joint and reduced numbers of Tregs cells in the periphery and within the joint [44, 45, 48]. In addition, gene expression data was able to separate persistent and extended-to-be cases using just 8 vs 13 samples in synovial fluid mononuclear cells [45]. Initial models to predict disease extension (worsening of disease) at 1 year produced promising proof of principle data suggesting biological data may be able to identify patients at risk of poorer outcomes at 1 year [45]. These observations may be specific to synovial fluid or cells highly enriched in inflamed cells, since results were not mirrored in peripheral blood cells [45, 46]. This lack of correlation may be explained by the vast variation in cell types and numbers found in peripheral blood. Relapse after treatment withdrawal may have biological markers of value to the clinician. Using DNA methylation data, proof of concept data is oJIA-ext and polyJIA suggested increased disease activity (flare or failure to maintain inactive disease) can be identified up to 8 months prior [51].

Eng *et al.* presented data utilising machine learning to combine 14 indicators and produce 5 all new categories, based on initial presentation [13]. This promising work utilises biological and clinical data to identify more homogeneous patient groups at presentation. These groupings also describe highly homogeneous disease trajectories according to disease activity at 6 months [13]. However this time span is still relatively short, and a fuller picture of clinically relevant disease activity was not articulated, with disease activity only identified by haemoglobin and platelet count and included sex and ANA status. Once clinically

actionable disease activity measures at 6 or 12 months (and beyond) are used, this would be a powerful tool to continue developing for clinical management.

Whilst a number of these studies potentially suffer from small sample sizes, mounting data suggests disease heterogeneity can be reduced by identifying subgroups in biologically based experimental data as opposed to clinical data. Importantly, data from many of these studies originated from immune cells present in the joint or in blood, cells which are readily accessible and facilitate further development of biomarkers.

Whilst the ILAR definitions define clinically homogenous groups based on an early clinical presentation, it is clear that ILAR criteria remain controversial due to the inability to correlate subtypes with clinical outcomes and not merely initial clinical presentation [2, 11, 42, 43]. The need for biological markers in disease classification is apparent in the number of JIA subtypes showing apparent heterogeneity within groups, as outlined above. For patients, this means that disease course, response to treatment and prognosis are difficult to predict [43]. Of note, the most interesting examples of subtype heterogeneity are those to do with polyJIA and oJIA, with data from immune cells suggesting molecular markers may reside within these cell types. Therefore, an understanding of the immunopathology is necessary to identify sources of potential biomarkers for further study.

1.3 Pathogenesis of Juvenile Idiopathic Arthritis

1.3.1 Adaptive immune system

Autoimmune disease is marked by antigen specific response to self-antigens [71]. The adaptive immune system is critical to the specificity of this response. Therefore it is important to understand immunological cells involved in this response to better identify the immune basis of autoimmune disease.

The adaptive immune response is a highly specific response to almost any potential pathogen, with the aim of destroying the invader [72]. These responses can be so specific, even optical isomers and differences of a single amino acid can be distinguished by the immune system [72]. The cells involved in this type of response are lymphocytes, namely B and T-cells. Whilst both interact during an immune response, differing mechanisms of pathogen destruction are employed.

Two classes of adaptive immune mechanisms are present 1) the humoral response which includes antibodies, cytokines and chemokines and 2) a cell mediated response [73]. Highly specific antibody responses are typical of the B cell compartment, whereas T-cells mediate the antigen specific cell response. Specificity in B and T cell responses is obtained through antigen recognition receptors, which can be modified by genetic recombination of gene segments named variable (V), diversity (D) and joining (J). Further, point mutations in the variable regions as well as nucleotide insertion and deletion mechanisms at segment junctions can modulate specificity, providing specificity for 5×10^{13} different antigens [73-75]. In addition to destroying the pathogen via highly specific responses, the adaptive immune system also generates an immunological memory of past responses, allowing rapid response to reinfection with a known pathogen [73].

The antibody response to particular antigens is carried out by a number of B cells subsets, which enable a specific response to particular antigen types. These subsets can produce long or short lasting antibodies, with high, low or a poly-affinity for antigens [76].

Cell mediated immune responses are directed by CD4+ and CD8+ T-cells, the helper T-cells (Th) in the CD4+ T cell compartment. These are very important to adaptive immunity and are particularly relevant to autoimmunity [72]). As well as activating BCR stimulated B cells, CD4+ T-cells respond to intra and extracellular bacterial pathogens, viruses, parasites and provide mucosal immune protection [77-82]. Further, these stimulate cytotoxic T-cells to destroy cells infected by a pathogen [83]. Suppression of immune responses and maintenance of tolerance to self-antigens occurs via CD4 T-cells also, a process which is likely perturbed in autoimmune disease [84].

A number of subset of T-helper cells exist, centred on the classical Type 1 and Type 2 T-helper cell (Th1/Th2) paradigm [85]. In addition, Th17, Tfh, Th22, Th9 and regulatory T (Treg) cells are known to exist, all of which can be differentiated and identified by specific cytokine profiles [85, 86]. Imbalance in the Th1/Th2 axis, as well as between Th17 and Tregs, are thought to play important roles in autoimmunity [86, 87].

Th2 cells are important for immunity against multicellular parasites, and are present mainly in the gastrointestinal tract as well as the lung epithelium [88]. Th1 cells are involved in defense against intracellular infections and cell mediated immunity [89]. Cytokines produced

by this cell type have been associated with autoimmune diseases for some time, particularly through the potentially pro-inflammatory effects of IFN-γ secretion [86, 90].

Th17 cells are present at mucosal surfaces and provide immunity to microbes overlapping with Th1 and Th2 responses [82]. Cytokines from Th17 cells inhibit regulatory T cell development, and vice versa [84, 91]. Therefore these two cell types work in balance to check immune responses during infection in order to avoid tissue damage [92]. Th17 cells are now seen as particularly important in autoimmunity, with data from animal models suggesting the combination of Th17 upregulation and Treg downregulation (by the same cytokine signals) can trigger autoimmunity (reviewed in [87]).

Treg cells are involved in suppression of the immune system, particularly self-reactive T-cells, and are a well-recognised subset largely identified by FOXP3 expression [93, 94]. To carry out this suppressive function, Tregs target T-helper cells. Mechanisms include cell-cell contact that may initiate killing of self-reactive cells via induced apoptosis, cytokine release to cause inhibition, induction of cytolysis, or dendritic cells can be modulated to inhibit T-cells [95]. Tregs have likely roles in autoimmune diseases derived from their important functions in maintenance of self tolerance and controlling the immune response [96].

1.3.2 Immunopathology

The immunopathology of JIA is known to involve the adaptive immune system. Much like rheumatoid arthritis, there is mass infiltration of immune cells into the sub lining layer of the synovium within the joint, including monocytes, B cells as well as T-cells [97-99]. Immunohistochemical characterisation of immune cells in the synovium identified CD4+ and CD8+ T-cells, which were more activated in certain subtypes [100]. In addition, there is a higher ratio of CD4+:CD8+ T-cells in the synovial fluid [2]. The accumulation of CD4+ memory T-cells in the inflamed joint is also symptomatic of JIA [101]. Effector CD4+ T-cells in the joint appear to be Th1 skewed, expressing high concentrations of CCR5 [101, 102]. Of note, some subpopulations of T-cells may be specific for certain subtypes [103]. What is not clear is whether these subtype differences or Th1 skew is a result of long term disease processes in the joint, since participants in these studies were sampled a number of years post diagnosis. Nonetheless, CD4+ T-cells appear strongly correlated to JIA pathogenesis.

Of note, the Treg subset of CD4+ T-cells are known to be associated with autoimmune diseases, characterised by higher Treg numbers [104]. However, data in JIA is conflicting,

with reports of higher numbers of Tregs in opposing disease states i.e. remitting and active disease [44, 48, 105]. This apparent contradiction may be explained by Treg cell subsets. Bending *et al.* showed that remitting disease was associated with lower proportions of a Treg subset with strong FOXP3 expression [106]. Conversely, the subset of low FOXP3 expressing Tregs associate with a worse disease course, and may explain Wu *et al.*'s findings i.e. Wu *et al.*'s findings may reflect a higher proportion of low FOXP3 expressing Tregs [105, 106].

Immunopathology involving Tregs in JIA may be driven by genetic variants. Tregs found in joints from JIA patients contain T cell receptors encoded by autoimmunity associated variants [106]. Thus immunological evidence pointing to the importance of CD4+ T-cells in JIA may have genetic variants driving this association.

Another CD4+ T cell subset of relevance to JIA is Th17 cells. Th17 cells in JIA are found to be enriched in the synovial fluid of JIA patients compared to peripheral blood across a number of different studies [48, 107, 108]. Th17 cells are observed to exist in a reciprocal relationship with Tregs in JIA [109] [48, 105, 108]. Further, Th17 cells appear increased in the subgroup with worse disease i.e. oJIA extended [48]. One study found a contrasting relationship compared to the above results [110], which suggested Th17 cells are not elevated in synovial compared to peripheral blood. Whilst that study interrogated higher numbers of patients and avoided confounding due to medication exposure, the study's robustness was diminished due to the use of a less specific measure of Th17 cells (%Th17 of total mononuclear cells). This likely reduced the sensitivity to observing changes in subsets of cells, since it was against a background/denominator of very large cell numbers (total mononuclear cells). Using the more specific % CD4 T-cells or % T-cells (as the other studies did) would have mitigated this diminished sensitivity, and better reflects the balance with Tregs that Th17 cells cooperate in within the T cell compartment.

Interestingly, most studies suggest that there are no differences in peripheral Th17 cells of JIA patients compared to controls [48, 105, 107]. However, when comparing between JIA cases only, those who have longer remission days or active disease are more likely to have increased Th17 cells [48, 105]. Therefore, Th17 cells appear to be important in the pathogenic process of more profound disease as observed in both synovial fluid and peripheral blood.

Thus it is clear that dysregulation of CD4+ T-cells in number, subset expansion and function are relevant to JIA pathology. The importance of CD4+ T-cells in JIA is further highlighted by the early discovery of strong genetic associations, particularly that of the HLA region.

1.4 Genetic Risk factors in JIA

1.4.1 Introduction

JIA is a complex disease, with both genetic and environmental factors contributing to disease causation [2]. Three billion bases make up the genetic code, and variations to the sequence can have functional or regulatory effects on the genome. Estimations of variation occurrence have suggested as many as 1 in 1250 base pairs may differ between individuals, though this can vary across certain genomic regions [111].

Genetics may contribute to disease through to variations in DNA sequence that may affect protein coding, DNA structure, gene regulation and expression. Indeed, many disease associated sequence variants are located in non-coding functional regions [112]. Whilst only 2.9% of the genome is protein coding, 80.4% of the genome is involved in a biochemical events including protein coding and functional non-coding regions [112].

Variations in the sequence occur throughout the genome, spanning single or multiple bases, can occur as translocations or repeats of small and large genome segments [113]. These variants are largely responsible for the phenotypic differences seen between individuals, and also contribute to an individual's likelihood for developing complex diseases such as autoimmune diseases [113].

Complex diseases are likely to fit into the model of 'common disease, common variant' (CD-CV) [114]. This refers to a model where common variants (with a minor allele frequency >0.05) provide only a modest propensity to disease, enabling transmission of these pathogenic alleles to further generations. The accumulation of these variants with modest effects is what likely leads to disease. As a complex disease, JIA is well suited to research under the CD-CV model to understand the contribution of genetics to disease pathogenesis and clinical characteristics [115].

1.4.2 Genetic studies in JIA

Research into the biological determinants of JIA have largely focussed on genetic predisposition. Early studies suggested a complex relationship with disease and one not likely to follow typical Mendelian inheritance [116-120]. The importance of genetic factors was suggested in early studies of families with JIA affected children. These suggested that there was higher occurrence of autoimmunity in 1st and 2nd degree family members [121, 122]. Furthermore, monozygotic (MZ) twin concordance rates of 25-40% are observed [116, 123, 124]. Disease concordant MZ twins have also been found to be concordant for age of onset and disease course [125, 126]. In addition, concordant MZ twins developed JIA much closer in time compared to concordant non-twin siblings (5.5 vs 37 months apart, respectively) [126]. Together, these studies formed an early evidence base for a genetic contribution to JIA.

Concurrent to early family studies, HLA associations were amongst the first genetic associations to be reported in JIA. Indeed, *HLA-B27*, *HLA-DR4* and *DR1* associations were based on family data, and began to lay an evidence base for the role of CD4+ T-cells in JIA due to unique presence of HLA class II molecules on these cells [127-129]. The HLA region has also been identified as the strongest genetic risk factor for JIA [130, 131]. Early studies helped to validate the HLA region's genetic risk contribution to JIA, particularly the *HLA-DRB1*08* allele, and this finding was validated by meta-analyses [132, 133]. This region was found to be particularly relevant in oJIA and polyJIA [132].

Fine mapping efforts of Hollenbach *et al.* suggested more precise haplotypes and alleles responsible for JIA susceptibility, again focussing on common JIA subtypes oJIA and polyJIA [14]. Higher precision HLA typing found specific HLA risk haplotypes which overlapped with Thomson *et al.*'s findings, such as *HLA DRB1-DQA1-DQB1* haplotype [133] [14]. This haplotype association was again replicated in one of the largest genetics studies in JIA [130]. To further validate the importance of the *HLA-DRB1* region in oJIA and RF- polyJIA susceptibility, Hinks and Bowes *et al.* suggest approximately half of the known heritability resulting from the entire HLA region (8-10%) is explained by the *DRB1* alleles [130, 131]. These findings strengthen the evidence base for CD4+ T-cells in the pathogenesis of JIA.

1.4.3 Heterogeneity in HLA associations according to JIA subtype and subgroups

HLA associations according to subtype have so far provided conflicting results, either for or against the notion of heterogeneity in JIA subtypes. Hinks & Bowes *et al.* finely mapped the HLA region in a large JIA case-control study [54]. This study converted HLA genetic data to HLA amino acid residues. They observed multiple independent effects in the *HLA-DRB1* region, which ultimately originated from amino acid 13 and 67 positions [54]. The amino acid 13 position was the primary HLA signal associated with not only oJIA and polyJIA, indeed with 5 of the 7 JIA subtypes [54]. This suggests a potential genetic and biological homogeneity in disease processes contributing to most JIA subtypes.

However, this homogeneity contrasts with debates and studies that have identified subgroups within defined subtypes i.e. identifying a lack of homogeneity in clinical, molecular, HLA genetic and immunological data [12-14, 17, 42, 50]. Regarding HLA associations, Hollenbach *et al.* observed associations in age subgroups within JIA subtypes, as well as subtype specific HLA risk alleles [14]. Their observed *HLA-DRB1*08* association was specific to poly-JIA cases aged older than 6 years at disease onset [14]. Further, a young age of onset effect (<6 years age) in oJIA cases was also observed in *HLA-DRB1*11:03/11:04* [14]. This finding highlighted the heterogeneity in JIA classification due to age and mirrored early data showing HLA associations according to age of onset, and also mirrored previous JIA classification criteria that incorporated age of onset [134] [135, 136]. Hollenbach *et al.* also reported oJIA subgroup differences in persistent vs extended disease related to HLA class I haplotypes. Specifically, *HLA-A* associations appear to protect from extended disease and *HLA-C* predisposes to persistent disease [14]. Hints of uveitis and temporo-mandibular joint arthritis associated HLA alleles also exist [137-139]. These studies suggest *HLA* risk alleles may not be homogeneous across JIA subtypes or associated with clinical outcomes.

In addition to the mentioned contrasting studies, 1) disease specific *HLA* associations are found across paediatric autoimmune diseases (potentially mirroring JIA subtypes differences) and 2) the contrasting genetic studies (e.g. Hollenbach *et al.* and Hinks Bowes *et al.*) suffered from small sample sizes in some subtypes [14, 54, 131]. Therefore the potential for homogeneous *HLA* risk alleles suggested in Hinks *et al.* in JIA subtypes needs further exploration [54].

Alternatively, subtype or oJIA subgroup differences may potentially be due to genes outside of the *HLA* region. Hinks & Bowes *et al.*'s data suggests most subtypes have ~50% of the heritability originating from outside the *HLA* region [54]. Further studies on biological differences between JIA patients may help determine a biological basis for a new JIA classification system that could include genetic, immunological or other molecular markers. In addition, more recent and larger *HLA* genetic studies have unfortunately not reported age of onset data. Mining those datasets would also be important to determining the genetic basis of subtype heterogeneity.

1.4.4 Non-HLA genetic associations in JIA

Whilst a number of robust *HLA* associations have been identified in JIA, this is likely due to the strong effect size of these alleles (with odds ratios of approximately 6) [54, 130]. Yet few non-*HLA* genetic associations have been robustly reported, likely owing to the complex nature of JIA including the likely interplay of environment in aetiology, the large number of subtypes, the likely subgroups existing within subtypes, and the low prevalence of this disease [122].

However, prior to 2015, only 6 non-*HLA* associated sites have met curation criteria for NHGRI/EBI GWAS catalogue from 5 GWAS' reported in JIA [131, 140] [141-144]. These studies typically suffered from small samples sizes, reduced genome coverage and even the largest study suffered issues with population stratification in the lead association [140]. Since 2016, larger GWAS' and meta-analyses have begun adding more robust associations to the body of knowledge (Table 1-1). Li *et al.*'s large paediatric autoimmune disease meta-analysis produced 46 loci passing NHGRI/EBI curation criteria Table 1-1 [131]. However, of those only 7 loci were driven by JIA [131]. Among these 7 loci, 3 were novel and 4 replicated associations from the literature including *IL2RA* and *PTPN22* [131]. McIntosh *et al.* reported 45 novel JIA loci in the largest GWAS meta-analysis to date, 4 of which overlap regions of known associations to other autoimmune diseases [145]. Subtype specific GWAS have also been published more recently, including for sJIA and RF+ polyJIA subtypes [23, 146]. Sample sizes for these studies were relatively small and limited the number of robust associations, mainly reporting the *HLA region* (sJIA and RF+: 2 and 1 association $< 5 \times 10^{-8}$ from 770 and 340 samples respectively) [23, 146]. However, a number of suggestive associations are reported (23 and 13 associations respectively) indicating loci that could be validated in larger or more specific candidate gene studies.

An example of a powerful candidate gene approach has been the 'immunochip', a microarray interrogating a collation of known immune related SNPs from multiple immune diseases [147]. Hinks & Cobb *et al.* used a high powered study using immunochip which focussed on oJIA and RF- polyJIA [130]. They reported 14 robust JIA susceptibility loci at GWAS significance level, many of which were outside the HLA region and 7 of which replicated published variants including *PTPN22*, *PTPN2*, *STAT4* [130]. Of the 11 regions with a suggestive association, four had supportive evidence from the literature [130]. Six loci were independently replicated in a study by this group including *ERAP2-LNPEP*, *RUNX1* and *UBE2L3* [148]. The unreplicated loci may indicate a lack of true association or the relatively small number of samples in the replication cohort [148]. Further, McIntosh *et al.*'s large meta-analysis replicated 12 loci observed in Hinks & Cobb *et al.* [145]. Altogether, Hinks & Cobb *et al.*'s Immunochip study was perhaps the most robust and outcome rich of all genetic studies in JIA. However it was limited to oJIA and RF- polyJIA subtypes, and follow up studies on other subtypes have been limited in sample size or focussed the *MHC* region only [23, 54, 146]. Therefore, higher powered studies JIA in subtypes including oligo- and RF- polyJIA and others are needed to better define genetic susceptibility factors. These studies may point towards potential mechanisms including immunological or epigenetic.

Interestingly, a comprehensive study has been reported of risk variants for autoimmunity which suggest potential epigenetic and immunological mechanisms. The Epigenome roadmap consortium utilised risk variants identified specifically from immunochip based studies, including the JIA study from Hinks and Cobb *et al.* [149]. This consortium was able to identify potentially causal SNPs in autoimmune diseases, noting that the lead SNPs from Immunochip studies were often a median of 14kb away [149]. Of note, these potentially causal SNPs often mapped to regulatory regions, enriched at immune cell enhancers [149]. In JIA, this coincided with regulatory elements for Th17 cells, Th1 and Th2 cells [149]. This finding mirrored ENCODE's reporting of GWAS hits enriched at gene regulatory sites, particularly enhancers [112]. Thus, the enrichment of disease associated genetic variants at non-coding sites, particularly at epigenetically marked enhancer sites, suggests that the convergence of genetic and epigenetic data is becoming more and more relevant, as is the need for research of complex diseases combining these two mechanisms. This is likely to be relevant for JIA research due to the potential for interplay between genetic heritability and environmental factors mediated by epigenetics.

Table 1-1 Summary table of GWAS in JIA. Data sourced from NHGRI/EBI GWAS catalogue [152].

Author	Analysis	Discovery sample number	Replication sample number	No. assoc. ($p<1 \times 10^{-5}$)	Top 3 ranking SNPs
Behrens *et al.*, 2008 [144]	JIA	67 cases 1952 controls	-	1	rs2395148, *TSBP1-AS1*
Hinks e*t al.*, 2009 [142]	JIA	279 cases 184 controls	321 cases, 2024 controls	1	rs12046117, *VTCN1*
Thompson *et al.*, 2012 [143]	JIA	814 cases 658 controls	1744 cases 7010 controls	4	rs4688011, *TIMMDC1* rs6479891, *JMJD1C* rs12411988, *REEP3*
Hinks *et al.*, 2013 [130]	JIA (oligoarticular or rheumatoid factor-negative polyarticular)	2,816 cases 13,056 controls	-	31	rs7775055, *HLA-DQB1/MTCO3P1* rs6679677, *RSBN1/PHTF1* rs10174238, *STAT4*
Aydın-Son et al., 2015 [150]	JIA	26 cases 52 controls	-	0	-
Li *et al.*, 2015 [131]	Paediatric autoimmune diseases inc. JIA	1112 cases 27131 controls	-	46	rs2066363, *ADGRL2* rs114846446, *LINC01250* rs7672495, *CYTL1/RN7SKP113*
Finkel *et al.*, 2016 [140]	JIA	1166 cases 9500 controls	480 cases 480 controls	1	rs953387, *CXCR4*
Ombrello *et al.*, 2016 [23]	Systemic JIA	982 cases 9010 controls	-	25	rs41291794, *HLA-DRB9/HLA-DRA* rs72632736, *LINC01777/EEF1DP6* rs1823549, *AC114485.1/AC099567.1*
Haasnoot *et al.*, 2018 [151]	Uveitis in JIA compared to JIA patients with no uveitis	214 JIA cases with uveitis vs 362 JIA cases w/o uveitis	-	2	HLA-DRB1*11 HLA-DRB1*08
Haasnoot *et al.*, 2018 [151]	Uveitis in JIA compared to controls	214 JIA cases with uveitis 398 controls	-	0	-
Haasnoot *et al.*, 2018 [151]	JIA	576 cases 398 controls	-	0	-

1.5 Environmental influences on JIA

JIA is recognised as a complex disease, with genetics and the environment both likely contributing to aetiology. Noteworthy is the attributable portion to each factor, with estimates varying widely. Li *et al.* suggest ~80% of JIA is genetically driven with just 20% explained by the environment, which appears to support population based heritability estimates [131] [122, 127]. However both estimates may be potentially inflated by gene-environment interactions or the presence of subtypes with stronger heritability (e.g. ERA or oJIA-ext as seen in Hink & Bowes *et al.*) [131]. The potential influence of environment is highlighted in Hinks & Bowes *et al.*, who suggest the immunochip genes contributes just 13% to heritability [54]. This is striking since their immunochip study is the richest source of validated JIA susceptibility loci. A significant potential environmental component is also suggested from monozygotic twin concordance rates, estimated only to be 25-40%, which are well below the full concordance expected of Mendelian trait [118, 126, 153]. Therefore, environment and interplay between genes and environment are important to consider in JIA aetiology [126, 154].

1.5.1 Neonatal and family life influences

A number of environmental exposures have been linked to JIA, but debate ensues on the robustness of these reports. An environmental exposure with inconsistent findings is breastfeeding. A lack of breastfeeding may be a form of stress on the child e.g. reduced bonding time. The molecular basis for this effect has been suggested through epigenetic studies in children of anxious mothers, nurturing deficient rats and more recently in nurturing deficient macaques [155-157]. However, epidemiological studies have been conflicting in JIA showing null, protective or risk associations [158-160]. Papers reporting a risk suffered from low sample numbers and/or selection bias of controls (controls were known to the JIA cases), whilst the protective report did not adjust for multiple testing [154, 158-160]. The CLARITY project also suggests breastfeeding as a risk factor, yet adjustment for socio-economic status removed this association [158]. Clearly study designs are an issue and whilst the CLARITY study is building numbers to help answer this question, lack of other environmental studies plus the lingering threat of recall bias indicate that further work needs to be done in JIA to elucidate the risks or protective effects of breastfeeding.

Sibling exposure has recently been reported to be protective [161]. The protection seemed to be higher with greater sibling year exposure, and higher numbers of siblings [161].

Despite the reduced sample sizes at these higher exposures, confidence intervals remained narrow suggesting a real effect. In addition, the study added robustness with validation in a second control group [161]. This report replicated findings from a registry based study that also found protection from JIA with any sibling exposure compared to an only child [162]. However, another registry based study reported that siblings had no protective effect [163]. The discrepancies in the two registry based reports may be due to sample selection with Carlens *et al.* selecting a 30 year window compared to Neilsen *et al.'s* 3 year window. However, despite a larger study by Carlens *et al.*, a potential exists for confounding due to difficulties in accurately identify JIA 30 years ago. Carlens *et al.'s* data was collected at a time when JIA was even less recognised, where multiple classification criteria existed for JIA and diagnoses may have included a number of other diseases with rheumatic manifestations [11].

Despite a need for further studies on environmental exposures, hypotheses from other immune related diseases provide another perspective e.g. the hygiene hypothesis and the role of exposure to early life infection may be pertinent to JIA [154, 162, 164]. This has found some support particularly since reports suggest farm living and sibling exposure (therefore infection exposure) are protective, and high incomes a risk factor [161, 162].

1.5.2 Microbial and antibiotic exposure

Indeed, the role of infections in JIA have long been reported, however studies on a number of viruses are conflicting. Studies into rubella initially suggested a role for this in JIA aetiology, whilst 2 later studies (one with a similar analysis) suggested otherwise [165-167].

Parvovirus B19 infection is perhaps the most contentious. An association with this virus and JIA was reported early on and supported by two other reports [168-170]. However, multiple conflicting studies have since been published, with the largest study utilising at-diagnosis samples reporting no association. This suggests earlier positive results were potentially a consequence of long term disease [171, 172].

Other studies into Epstein-Barr virus and Human Cytomegalo virus were negative, however positive associations with Influenza A was reported [173, 174]. This positive association with Influenza A was translated into a small clinical trial treating the viral infection, with positive outcomes reported [175]. Akin to anti-retroviral treatment, treatment of Streptococcus infection was related to more cases of remission compared to control patients [176].

However, all these studies have shortcomings, particularly due to small sample sizes or control groups, differences in samples types analysed (IgM responses vs virus isolated from blood) and difficulties identifying cause vs consequence due to sampling times (incident vs prevalent cases). These make robust conclusions and comparisons difficult.

Related to microbial exposure is antibiotic use. Studies showing an increase in JIA with antibiotic use appear contrary to studies on the increased risk of JIA due to Streptococcus [176-178]. However, Horton *et al.* suggests that deleterious effects on the gut microbiome may better explain the mechanism behind this association. Alterations in gut microflora, and increased gut permeability are observed in JIA and potentially support subclinical gut inflammation as important contributors to pathogenesis [179-181].

1.5.3 Smoking

Another environmental factor of relevance is smoking, as it is known to be a risk factor for adult rheumatoid arthritis. Data in JIA is still sparse and controversial, with 2 studies showing protection, 1 no effect and 1 other study suggesting smoking as a risk factor [158, 163, 182, 183]. Regarding the protective effects of smoking in pregnancy observed in 2 separate studies [158, 183], the results of the Ellis *et al.* study appear robust as the effect remains across a number of different measures of smoking exposure. A third study reported smoking as a risk factor for JIA [182]. Trends towards this are observed upon closer inspection of data from Shenoi *et al.*, particularly when looking at smoking only in pregnancy [182, 183]. Yet the numbers in Jaakkola *et al.* may be too small to draw firm conclusions from.

The molecular effects of smoking are well reported, particularly in DNA methylation (reviewed in [184]). Importantly, the amount of smoking exposure is an important variable for molecular effects of smoking to take place. Specifically, maternal smoking through a large part of pregnancy, not just at any time in gestation, was important to this effect [184]. Amongst JIA studies reporting smoking data, the specific duration of smoking exposure was not captured or reported well. Thus, studies to further elucidate the true effect of smoking in JIA require higher numbers, more refined questions or better study designs to capture this smoking data accurately. This will be important as the effects of smoking at the epigenetic level are dependent on the timing and duration of the exposure.

It is clear that epigenetic effects due to an environmental exposure like smoking can be observed. Since epigenetic marks like DNA methylation can act as the conduit between

genes and the environment, these are likely to be important in a complex disease such as JIA. It is therefore important to understand epigenetics and DNA methylation in particular.

1.6 Introduction to Epigenetics

Epigenetics can be broadly described as the structural adaptation of chromosomal regions so as to register, signal or perpetuate altered activity states [185]. These adaptations are regularly cited as mitotically heritable, covalent modifications of DNA and histones, which are intimately involved in the control of gene expression [186]. Epigenetic modifications to DNA affect the accessibility of DNA to transcription machinery thereby regulating expression of genes [187]. These also largely determine cell fate and function, and together form a pattern that is distinct for any cell type [188].

In addition to being mitotically heritable, epigenetics may act as the conduit between genes and environmental exposures [189]. This possibility heightens during times of epigenetic change in early development, which could lead to 'epimutations' [188]. These can be mitotically inherited by daughter cells potentially leading to disease states resulting from aberrant gene expression [189].

DNA methylation and histone modifications are the two major types of covalent modifications studied. These mechanisms are well conserved amongst plant and animals, as are mechanisms for its maintenance [190]. DNA methylation is the most widely studies epigenetic modification, and is particularly relevant to immunology. Therefore, DNA methylation will be the focus of this study, with mechanisms and associations with immunology and disease expanded upon below.

1.6.1 DNA methylation

In mammals, DNA methylation (DNAm) of symmetrical CpG dinucleotides are the most common type of DNA methylation observed, but it is also known to be found in different contexts [191, 192]. In CpG dinucleotides, covalent modification occurs by addition of a methyl group to the 5th carbon of a cytosine nucleotide. Of the ~28 million CpGs in the genome, known as the methylome, a large majority of CpGs in the genome are methylated [193]. Yet 10% are usually unmethylated (often occurring at transcription start sites), creating a bimodal distribution of CpG methylation in the human genome [193, 194].

CpG dinucleotides at transcription start sites tend to occur as CpG islands, regions of genomic sequence highly enriched for CpG dinucleotides that can span up to 2Kb in length [193, 195]. These are known to be associated with open chromatin structures and transcription, and DNAm modification of these is associated with dysregulation of gene expression in cancer [195]. Logistically, DNAm overall is likely to be important in maintaining gene expression patterns across mitotic cycles due to its stability [194]. Mitosis involves the disruption of most other factors and structures for gene expression whilst this modification remains stable, ideally suited as a mechanism for gene expression maintenance across multiple mitotic generations [194].

The regulatory enzymes important in maintenance and establishment of DNAm are named DNA methyltransferases (DNMTs) [194, 196, 197]. Maintenance of DNAm it achieved with DNMT1, particularly during cell replication. In this scenario, DNMT1 uses the undisrupted DNAm on the original strand to maintain DNAm patterns across cell divisions [194]. It does this due to its high affinity to the DNA replication machinery during replication and its high specificity for hemi-methylated DNA, as would be found during a newly synthesised DNA strand [198, 199]. De novo activity occurs via DNMT3A and DNMT3B, which set the bimodal distribution of DNA methylation in the newly fertilised developing human embryo during the implantation stage [194, 197, 200, 201].

Both at local and global levels, the potential of genetics and environment to modulate methylation are potentially relevant for complex disease pathogenesis. Below I outline some potential genetic and environmental factors important for methylation regulation.

1.6.2 Regulation of DNA methylation

1.6.2.1 Environment and DNA methylation

Windows of opportunity in methylome vulnerability are suggested to occur during times of rapid epigenetic change, such as reprogramming of the zygote methylome [202]. A window may also exist in early life, as it is suspected of being home to a changing and environmentally sensitive methylome in studies of immune cells, rats and association studies in humans [157, 203-205]. An early life environmental insult cued at the earliest of embryonic development stages may set abnormal epigenetic patterns resulting in abnormal phenotypes [205]. The phenotype may not even present immediately, but the self-sustaining mechanisms inherent to epigenetics and DNAm perpetuate the effects and may result in long term changes [194, 205].

Slight differences in intra-uterine environments are suspected to explain monozygotic twins' discordance in methylomes [202, 206]. Diverging methylomes throughout life, known as epigenetic drift, is seen in individuals and monozygotic twins in the early years of life and into adulthood [207-209]. Thus a breathing, pliable methylome apparent in early development and early childhood allows time and space for environmental cues to influence and possibly perturb [202, 205]. Environmental cues such as smoking, metal exposure, and stress due to famine or maternal anxiety are reported to have such effects [155, 184, 204, 210-214]. An environmental cue of known importance is smoking.

Smoking effects on the methylome are now well replicated, in particular the changes seen in the *AHRR* gene, encoding an enzyme involved in the detoxification of chemicals found in tobacco smoke (reviewed in [184]). Maternal smoking affected umbilical cord blood methylation in multiple cohorts, all affecting the same probe [215-217]. The exposure to smoking throughout most of gestation is important to this effect [184]. In addition, prenatal exposure effects can be detected even after 18 months. However, this may be reversible since DNAm levels in adult former smokers quickly approaches levels of never smokers within the first few years of smoking cessation [217, 218]. DNAm changes at the gene level may also be affected the genetic structure, influencing how much of the environmental insult is able to change local methylation. Genetic regulation of DNAm is therefore important to understand.

1.6.2.2 Genetic regulation of DNA methylation

Genetic modulation of DNAm through methylation quantitative trait loci (mQTL) is a well-studied phenomenon, with pivotal studies in brains identifying 5-10% of CpGs have an mQTLs associated [219, 220]. Further studies, across different cell types and interrogating larger numbers of CpGs, validated these initial observations and suggested the percentage of CpGs with an mQTL associated may more likely be between 2-10% [221-229].

Published mQTLs are commonly found in close proximity to the CpG of interest. Whilst a large proportion of mQTL appear to directly affect the CpG of interest [219, 228], mQTLs are highly enriched in regions ~10Kb from the CpG, with most mQTLs often reported an average under +/-100Kb of the CpG of interest [221-229]. This average distance of 100kb changes little despite a number of publications searching up to 1Mb from the CpG [220, 224, 226]. Two seminal studies from Liu *et al*. and Gibbs *et al*., despite looking for mQTL associations

genome wide, still suggested an overwhelming proportion of mQTLs (~80%) were under ~150kb and ~120Kb from the CpG respectively [219, 222].

MQTLs are likely to be involved in the modulation of multiple CpG sites, as reported by Liu *et al.* [222]. These CpG sites may be contiguous, potentially following patterns similar to Linkage disequilibrium in genetics but over much smaller regions, and may be spread over hundreds of kilobases [222].

Ethnicity is an important factor influencing DNAm. It is clear that ethnicity accounts for a large proportion of DNAm differences between individuals, as observed in early investigations of HapMap cell lines [230-233]. Following genetic differences and similarities between ethnicities, DNAm in Americans of African descent were clearly different from Caucasian and Asian Americans, who tended to cluster together [230-232]. As well as having global effects on DNAm, ethnicity specific DNAm is observed through region specific enrichment e.g. Caucasian (and Caucasian enriched) specific CpG sites are more prominent at transcriptionally active or enhancer sites upstream of the gene, whilst depleted amongst African Yoruba samples [231, 232].

The driver of these ethnic differences appears to heavily favour genetic sources i.e. local mQTLs [230-232]. Interestingly, data also suggests that these mQTLs are population specific. That is, DNAm differences according to genotype are often limited to a single ethnic group such that MQTLs in one ethnicity are often not observed in another [230-232]. This SNP data mirrors CpG data, where ethnicity specific genomic region enrichment is observed [231, 232]. This correlation makes sense, if one considers the local effect of ethnicity specific mQTLs perhaps leading to region specificity in DNAm.

The remaining ethnic differences are likely due to gene x environment (GxE) and gene x gene (GxG) interactions, potentially favouring GxE as observed in Teh *et al.* and Czamara *et al.* [234, 235]. This complex modulation of DNAm again suggests DNAm is relevant to study in complex diseases [236]. Concordantly, many of the SNPs behind these ethnicity mQTLs were also associated with complex traits that have ethnicity specific preponderance [231, 232].

Therefore, whilst ethnicity is an important component of DNAm studies, arguably the specific, local genetic drivers near any one CpG site are going to be an equally defining factor to take into account during DNAm studies. This may be particularly so of complex disease

studies, where differences are likely to be small and potentially have mQTLs contributing to disease risk.

1.6.2.3 MQTLs are potentially relevant to complex diseases

MQTL are known to be relevant to disease and may be relevant to oJIA. The link between mQTLs and disease susceptibility was hypothesised by Feinberg *et al*. early in the GWAS era and appeared to pre-empt the direction of research post-GWAS era [237]. GWAS in complex diseases resulted in a number of top ranking SNPs residing within non-protein coding regions of the genome [112, 238-240]. Studies of mQTLs, in part, have been used in attempts to overcome a lack of simple, protein modification based mechanistic explanations of GWAS hits [241]. Consequently, a number of mQTL studies suggest GWAS signals are indeed enriched within mQTLs, particularly in autoimmune diseases, cancer, chronic inflammatory diseases, haematological parameters and neurological diseases such as schizophrenia [222, 226, 242-244]. This relationship to autoimmune diseases, and potentially to oJIA, may be relevant to this study. In particular, mQTLs may be a source of oJIA heterogeneity affecting DNAm. MQTLs may also identify new susceptibility risk loci in JIA.

1.6.2.4 DNAm variation modulated by genetics

Whilst DNA methylation is likely to be environmentally modifiable, the genetic aspects contributing to the extent of variability at any CpG is a question well worth considering. Feinberg and Irizarry in 2010 proposed a particular model, suggesting genetics may indeed play this role in modulating epigenetic variation [236]. They suggest these genetic changes would affect the distribution of loci specific DNAm in a population [236]. In simulations, this variation of DNAm allows for increased fitness, since the ability to vary DNAm can be a mechanism for adapting to larger range of environmental changes [236]. To Feinberg and Irizarry, this allows natural selection to use variability in epigenetics to adapt to changing environments [236]. This variation would occur at regions they termed *variably methylated regions (VMRs)*.

In order to validate this hypothesis, Feinberg and Irizarry re-examined previously published DNAm data and identified VMRs i.e. regions with high DNAm variation between individuals [236]. VMRs were observed across tissues, and often associated with tissue specific DNAm. The connection of VMRs to genetics was identified, with VMRs correlated with differences in the underlying sequence [236]. This demonstrated to Feinberg and Irizarry a proof of principle supporting their hypothesis. i.e. a genetic basis for methylation variation.

Indeed this hypothesis bore specific fruits, in those found in Feinberg's 2013 paper linking genetics and DNAm in Rheumatoid Arthritis. A number of those genetic variants linked to DNAm were highly associated with variation in methylation [245]. Genetic variants associated with this variation were shown to be approximately 80Kb upstream of the probe itself [245].

Further, Teh *et al.* and Czamara *et al*. both also identified this association between DNAm variation and genetics, also validating early hypotheses from Feinberg *et al*. [234, 235]. Firstly, both papers validated VMRs are strongly associated with genetics. Secondly, the genetics x environment link, noted as a mechanism for environmental adaptation by Feinberg *et al*., was one of the most significant findings in these papers [234, 235]. That is, 62-88% of VMRs were best explained by a model integrating genetics and environment. Environment alone often explained less than 1% of VMRs. Thirdly, genetic variants associating with VMRs were also enriched for complex disease risk loci (identified from publicly available GWAS data) [234]. Together, these 3 key papers validate the strong potential for genetics to play a role in modulating DNAm variation, and the potential for these to influence risk to complex diseases [234, 245, 246].

1.6.2.5 Enhancer regions may be relevant to mQTLs and disease

Feinberg and Irizarry propose regions that are tissue specific and localise near a gene may be sufficient for genetics to influence DNAm variation. Regulatory regions such as enhancer sequences fit into their description of possible genetic mechanisms for a number of reasons. Enhancers are known to be tissue specific, can be proximal or distal to a gene whilst still having effects on the gene, have recently been associated with differential methylation, and also harbour SNPs associated with methylation differences. These regions may well harbour more as yet unidentified genetic variants associating with methylation and disease.

Genetic variants at enhancer regions are also known to be important for methylation. Gutierrez-Arcelus *et al.* in 2013 found enhancers are enriched for genetic variants associated with methylation (mQTLs). In addition and corroborating Feinberg's hypothesis, these enhancer mQTLs were also associated with tissue differentiation [247]. Enhancer associated mQTLs are also associated with disease, including in breast cancer, colorectal cancer and glioblastoma [248-250]. Indeed, enhancer mQTLs are associated with clinical outcomes including glioblastoma survival [250]. In breast cancer, much of the tumour heterogeneity was accounted for by mQTLs at enhancers [248]. This variation is analogous to Feinberg's

hypothesis regarding regulatory mQTLs and increased variation between individuals. In addition, Czamara *et al*. also identified mQTLs of highly variable regions enriched for regulatory regions such as transcription or enhancer regions [234]. This data lends support to the importance of genetics in affecting DNAm, particularly at regulatory regions such as enhancers.

The disease relevance of enhancer regions were further exemplified in large meta-analyses of disease risk variants. GWAS hits for multiple diseases were enriched at enhancers in multiple cell types in ENCODE data, including immune like lymphoblastoid cell lines [112]. A meta-analysis of Immunochip studies in multiple autoimmune diseases identified genetic variants enriched at immune cell specific enhancer sites [112, 149]. Of particular interest to this study, Gutierrez-Arcelus *et al.* found GWAS hits for diseases amongst enhancer located SNPs associated with DNAm [251].

Therefore, enhancer regions are likely to be highly relevant to disease genetics as well as DNAm, and are potential regions where SNPs may modulate DNAm variation as postulated Feinberg *et al.* [236].

1.6.3 DNA methylation in the immune system and autoimmunity

1.6.3.1 DNA methylation in immune cells

Methylation is known to be important in later stages of lymphoid cell development [193]. *DNMT1*$^{-/-}$ early stage thymocytes are skewed in numbers and are less abundant in secondary lymphoid tissues [252]. Naïve T-cells with *DNMT1* deletions have disrupted proliferative capacity, along with impaired maturation [252]. Further, despite impaired replication, cytokine levels were increased in these knockouts [252].

Moving to peripheral immune cells, numbers of memory T-cells were also affected by *DNMT1* mutations [252]. In addition, effector CD4 and CD8 T-cells also misregulate cytokine profiles for the Th2 lineage [253]. This occurs without misregulating other lineage specific cytokines, which the author suggested may implicate DNA methylation control of the Th2 program, particularly in CD8 T-cells [253].

Although these effects may be due to gross changes to the phenotype that come with highly disruptive gene knockouts. Whilst Makar *et al.* did not look more specifically into the role DNA methylation had in controlling the Th2 cytokine program at the gene level, Lee *et al.* did

find a correlation of CD8 demethylation and expression with a bias of CD8+ early stage thymocytes [252, 253]. Whether this caused the skewing of early stage thymocytes, or is just an association is not clear.

The importance of DNA methylation to CD4+ T-cells is well characterised, particularly in subset formation and lineage maintenance. With the importance of CD4+ T-cells to autoimmunity, DNA methylation's role in regulating normal immune function can give insights to its potential dysregulation in disease.

The classical methylation dependent control of immune cell lineage is that of Treg cells. Demethylation of the *FOXP3* enhancer region is important for maintaining FoxP3 expression and suppressive activity in Tregs [254-256]. The enhancer region is now known as the Treg specific demethylation region (TSDR) [254-256].

However, the role of methylation is not limited to Tregs in the CD4+ population. At a global level, in addition to earlier studies on *DNMT1*'s effects on CD4+ T-cell cytokines, *DNMT3a* knockouts affect the lineage speciation of CD4+ T-cells. These knockouts allow cytokine expression of both Th1 and Th2 lineages even under Th1 culture conditions [257]. Expression of these cytokines is concurrent with demethylation of the *IFNG* and *IL4* loci [257].

Effects of *DNMT3a* deletion was not restricted to Th1 conditions, but Th2, Th17 and iTregs also expressing Th1 cytokine IFNγ [258]. This further suggests that de novo methylation helps restricts lineage plasticity, therefore serving as a controller to maintain the identity of a T-cell lineage [258].

Methylation changes in T-cell differentiation have been investigated through genome wide DNAm studies in mice. Using memory, naïve and effector T-cells, memory T-cells undergo larger changes in methylation than effector T-cells and naïve T-cells [259]. In response to antigen, hypomethylation is enriched at subset relevant genes, and many of these sites display enhancer activity in reporter assays [259].

Further, genome scale DNA methylation data was able to identify T-cell subsets, particularly the naïve and activated cells from both CD4+ Treg and conventional CD4+ T-cells [260]. In Tregs, activation induced changes in DNA methylation at hundreds of loci, enriched at sites for FOXP3 binding [260]. This suggested DNA methylation is not just an upstream controller

of Treg function via *FOXP3* TSDR regulation, but an important controller of downstream FOXP3 function in Tregs at many FOXP3 regulated loci [260].

Continuing with immune cell differentiation, Th1 and Th2 cell differentiation is accompanied by methylation changes in loci of cytokines of the respective lineages. In Th1 cells, early studies suggested Th1 centric *IFN-γ* hypomethylation in Th1 cells and not Th2 cells [261]. This was reproduced in studies using early stage thymocytes, naïve T-cells and CD8 T-cells where the *IFN-γ* locus was unmethylated, and the Th2 program appears to have active silencing of the *IFN-γ* site [262-264]. This coincides with Makar *et al.*'s suggestion of methylation's role in control of the Th2 program. During the Th2 program, IL4 expression can be induced by demethylating agents, validated by data showing chromatin remodelling at the *IL4* loci is accompanied by DNA hypomethylation [265-267]. Kim *et al.* shows that this is STAT-6 dependent, as did Bird *et al.* [256]. Yet they differ in their suggestions on cell cycle dependence. Kim *et al.* states this is an active process independent cell cycle [256]. Yet the more comprehensive cell cycle blocking experiments from Bird *et al.* suggest that cell cycle progression is actually necessary for IL4 expression. It's interesting to note that Bird *et al.* did not look at IL4 expressing cells at day 0 to compare to day 1, where Kim *et al.* draw their conclusions from. Nonetheless, in addition to *IL4*, *IL2* demethylation corresponds with open chromatin at the *IL2* locus and higher expression in CD4+ effector T-cells [268].

The role of DNA methylation in T-cell differentiation and function is becoming clearer and better defined. Thus the dysregulation of DNA methylation is therefore likely to be implicated in immune dysregulation as a mediating factor. By natural extension, a potential outcome of this could be autoimmunity. In JIA, this has begun to be explored in CD4+ T-cells. A particularly subset of Tregs at the site of inflammation have methylation of the Treg specific demethylated region (TSDR) reported to be decoupled from FOXP3 expression, suggesting DNAm may not be a reliable marker of abnormal Treg function in JIA [106]. Whilst the authors state a de-coupling, a complete de-coupling may not necessarily be the case.

Closer inspection of their data suggests that methylation is only slightly dysregulated in different subsets of Tregs as compared to classical Treg populations. Therefore DNAm may not be completely decoupled from FOXP3 expression. Supporting this, their data shows non-classical CD25low Tregs as well as FOXP3low Tregs with abnormal methylation profiles coupled with abnormal expression. This coupled dysregulation of mid-low DNAm with expression appears to mirror observations of Tregs in mice. In the FOXP3low Treg population

in Floess *et al.*, expression and DNAm again appear to be coupled, where the amplicons of the TSDR region are only partially demethylated whilst maintaining some FOXP3 expression [254]. Therefore, DNAm in Tregs may still be a reliable marker of abnormal Treg function.

Both datasets, taken together, suggest that methylation is still likely to be important in regulation of FOXP3 expression in Tregs in diseases like JIA. Thus at sites of inflammation, dysregulated methylation may indeed be linked to autoimmunity. Bending *et al.* describe a link to another autoimmune disease in addition to JIA, noting that the CD25low population has also been found in lupus patients [106].

JIA research into epigenetics from our group has also focussed on the CD4+ T-cell group and found a number of significant associations, as well as genotypic and expression relationships to methylation in one of the significant findings [269]. This underscores the importance of the CD4+ T cell compartment to epigenetic research.

The early JIA research into DNAm was based on microarray technology, as has much of DNAm work in autoimmune disease research particularly in the period since 2010. Yet, autoimmune disease associations with DNA methylation date further back to as early as 1990, when epigenetic research was still in its relative infancy. Despite the field's infancy, strong associations were still found and are described below.

1.7 DNA methylation in autoimmune diseases

Whilst the imperative for epigenetic research in autoimmunity is clear, the potential for this line of investigation had precedents as far as 30 years ago. Inhibitors of DNA methyltransferases, enzymes that act to maintain or create de novo methylation, were used to induce autoimmune like states in T-cells. Self-reactivity was induced in in CD4+ T-cells, and mice developed lupus like symptoms after such treatments [270-279]. Genomic hypomethylation was a feature of Systemic Lupus Erythematosus (SLE) and Rheumatoid arthritis (RA) T-cells [270, 271]. Once normal human T-cells were treated with DNA methyltransferase inhibitors, they incorporated phenotypic and functional similarities to T-cells from lupus patients [275, 277-279]. This strongly suggested DNA methylation was key to immune cell function and possibly involved in autoimmunity. Further, early methylation studies into RA had already indicated the potential of methylation research in identifying therapeutic targets. IL6 was found to be differentially methylated, a molecule that later

became a target of antibody therapies in RA and JIA [280]. However, much like Cornil, Heberden's or even Caravaggio's early clinical and artistic descriptions of JIA, these early signs of epigenetic disturbances in autoimmune diseases (AID) were not developed through extensive research for some time.

Further research did eventually come to pass, particularly when technology for genetic variant interrogation (through microarray technologies) translated to the epigenetic space to allow interrogation of hundreds of CpGs. Using this technology, early studies on AID methylomes with phenotypic similarities did have some success [281]. Once the HumanMethylation27 beadchip microarray made large scale methylome studies accessible and affordable (followed by the updated HumanMethylation450 microarray and other array technologies), AID studies in DNAm significantly increased with a number of publications in the years following [282-286] [245, 287-312].

Early genome scale studies focussed on (RA), SLE, Type 1 diabetes (T1D) and inflammatory bowel disease (IBD). Seminal EWAS' in autoimmune diseases are described below.

Table 1-2 Summary table of EWAS studies in 4 key autoimmune diseases, plus JIA

Disease	Cell type used	Discovery cohort size	EWAS method	Significant probes	Validation sample numbers (independent)	Validation cell type	Validation cells notes	No. CpGs validated in indep. samples	Reference
SLE	White blood cells (PBL's)	5 MZ discordant twins, 2 matched controls/twin pair	Goldengate	49	Discovery + 1 MZ twin, 4 DZ twins, 7 non twin sibling pairs. All discordant	PBL's	Pyrosequencing	2/2 CpGs	Javierre *et al.* 2010 [281]
SLE	CD4+ T cells	12 cases vs 12 controls	Infinium HM27k	341	90	CD4+ T-cells	Pyrosequencing	5 tested, correlation between EWAS and pyrosequencing =0.83	Jeffries *et al.* 2011 [296]
SLE	White blood cells (PBL's)	12 cases vs 12 controls	Infinium HM27k	2165 genes	66 cases, 102 controls, 12 RA cases	White blood cells (PBL's)	Pyrosequencing and Methylation-sensitive single-nucleotide primer extension	2 of 2 CpGs	Lin *et al.* 2012 [297]
SLE	CD4+ T cells, CD19+ B cells and CD14+ monocytes	49 cases, 58 controls (monocytes:27 cases-control pairs)	Infinium HM450k	166 CpGs in B-cells, 97 CpGs in monocytes, and 1,033 CpGs in T-cells	26 cases, 18 controls	CD4+ T cells (Naïve, memory and T reg)	Infinium HM450k	76.8% of 1031 sites. 33-37% of 1031 sites in other cells types	Absher *et al.* 2013 [287]
SLE	CD4+ Naïve T cells used	18 cases, 18 controls	Infinium HM450k	86	18 cases, 18 controls	CD4+ naïve T cells	HM27k	Validated using Jeffries *et al.* 2011, r^2=0.86	Coit *et al.* 2013 [289]
SLE	PBMC	n = 186 (94 anti-dsDNA positive and 92 anti-dsDNA negative SLE cases)	Infinium HM450k	16 CpG sites at p<1.07 E-7	replication (n = 140. 62 anti-dsDNA positive and 78 anti-dsDNA negative SLE cases	PBMC	Infinium HM450k	16 of 17 CpG sites	Chung *et al.* 2015 [306]
SLE	PBMC	80 patients with SLE having nephritis and 245 patients with SLE not having nephritis	Infinium HM450k	19	56 (28 with kidney involvement, 28 without), 56 controls	CD4+ T cells	Infinium HM450k	4 of 19 CpGs	Mok *et al.* 2016 [309]
SLE	whole blood	347 cases, 400 controls. Treatment naïve n =132	Infinium HM450k	7625 in overall analysis. 5321 associated with medication in SLE.	SLE (n=201) and controls (n=187). In treatment naïve study replication n=89	Whole blood	Infinium HM450k	7245 CpG sites, in treatment naïve :3295 of 7425	Imgenberg-Kreuz *et al.* 2018 [307]
SLE	PBMC	333 cases	Infinium HM EPIC	256 CpG sites in 124 genes across all 3 groups, 247 between mild and Severe 2 subgroup, Severe 2 vs Severe 1 subgroups: 18 CpGs, Severe 1 vs Mild subgroups: 53 CpGs (all FDR 0.1)	302	PBMC	Infinium HM450k	132/158 CpGs	Lanata *et al.* 2019 [308]
IBD	Intestinal tissue	28 (5 CD, 9 UC, 14 paired controls)	Goldengate	50	12 (4 CD, 8 UC, 6 paired controls)	Intestinal tissue	Goldengate	50 of 50 CpGs	Lin *et al.* 2011 [298]

Table 1.2 Continued. Summary table of EWAS studies in 4 key autoimmune diseases, plus JIA

Disease	Cell type used	Discovery cohort size	EWAS method	Significant probes	Validation sample numbers (independent)	Validation cell type	Validation cells notes	No. CpGs validated in indep. samples	Reference
IBD	Rectal biopsies	CD: 16, UC:16, 8 controls	Infinium HM27k	NA	CD:16, UC 16, 8 controls	Rectal biopsies , including separation into Epithelial and non-Epithelial (immune cell enriched) fractions	pyrosequencing	3 of 3 CpGs	Cooke *et al.* 2012 [290]
IBD	White blood cells (PBL's)	MSAM: 4 CD discordant + 4 concordant MZ twins, 7 UC discordant + 3 concordant MZ twins. 450k: 3 discordant MZ twins (UC & CD), 14 paediatric CD, 8 paediatric UC, 14 controls	MSAM (Methylation specific amplification microarray) and 450k	CD vs controls: 1 .UC vs control (adults): 0. UC vs controls (children):6	-	-	-	-	Harris *et al.*2012 [294]
IBD	Intestinal biopsies	20 discordant UC twins	HM27k & Methylated DNA Immunoprecipitation 385k array	61	50 UC cases	Intestinal biopsies	Pyrosequencing	148-fold more MVPs in proximity to the 10 validated loci, p= 9.11 × 10−44	Hasler *et al.* 2012 [295]
IBD	Whole blood	21 CD cases vs 19 controls + 16 Paediatric CD cases	Infinium HM27k	1117	-	-	-	-	Nimmo *et al.* 2012 [300]
IBD	Leukocytes	36 paediatric CD (18 treatment naïve, 18 established CD) cohorts and 3 controls.	Infinium HM450k	65	87 adults with CD and 85 healthy controls	Whole-blood (adults)	pyrosequencing	20/20 replicated	Adams *et al.* 2014 [313]
IBD	PBMC	149 IBD cases (61 UC, 88 CD) and 39 controls	Infinium HM450k	3196 probes between CD cases and controls (p<0.05),1481 probes between UC cases and controls	Treatment naïve paediatric IBD (15 CD and 9 UC) and 22 controls.	Colonic mucosa tissue samples		2 of 2 CpGs	McDermott *et al.* 2016 [283]
IBD	PBMC	240 newly-diagnosed IBD cases (121 CD and 119 UC) and 190 controls	Infinium HM450k	439, and 5 DMRs. 412 DMPs when comparing CD to controls, 203 when comparing UC to controls. No differences between UC and CD	240 established IBD cases, 98 controls. Also 36 paediatric CD, 36 controls from Adams 2014.	PBMC	pyrosequencing and HM450k	Top DMP is adults, r=.77 in children of top 5,000 DMPs (CD vs controls).	Ventham *et al.* 2016 [284]

Table 1.2 Continued. Summary table of EWAS studies in 4 key autoimmune diseases, plus JIA

Disease	Cell type used	Discovery cohort size	EWAS method	Significant probes	Validation sample numbers (independent)	Validation cell type	Validation cells notes	No. CpGs validated in indep. samples	Reference
IBD	Mucosal biopsies	66 treatment-naïve children newly diagnosed with IBD (43 with Crohn's disease [CD], 23 with ulcerative colitis [UC]), and 30 children without IBD (controls)	Infinium HM450k	3569 in CD vs control (Ileum), 641 CD vs UC in ileum. 0 in UC vs controls in ileum. 1743 in CD vs controls in colon cells.0 in CD vs UC in colon cells. 51613 in UC vs control in colon cells.	CD (n=5) and matched healthy controls (n=7)	Mucosal biopsies	pyrosequencing for top DMPs.	3 of 3 top CpGs	Howell *et al.* 2018 [282]
IBD	Whole blood	164 paediatric patients (1-17 years old). 74 controls.	Infinium HM EPIC	1189 b/w diagnosis and controls. 717 unique genes	None. But looked at 3 years later in same samples	-	Almost all probes changed to normal levels.	-	Somineni *et al.* 2019 [286]
T1D	Whole blood	96 nephropathy T1D cases, 96 non-nephropathy T1D controls	Infinium HM27k	19	-	-	-	-	Bell *et al.* 2010 [288]
T1D	CD14+ monocytes	15 discordant MZ twins	Infinium HM27k	132	1. Discovery + 5 discordant MZ twins 2. Discovery + 4 discordant MZ twins 3. 7 cases, pre and post disease onset 4. 4 cases (Ab+, no disease) vs 5 controls (discordant twin pairs from above)	CD14+ monocytes	pyrosequencing & HM27k	1. 0/13 via pyroseq. 2. P=0.0375 showing expected direction (in 132 CpGs). 3. P=0.015 showing expected direction (in 132 CpGs), i.e. 71% and 66% of pre vs control AND post vs control. 4. P=0.0023, showing expected direction (in 132 CpGs), i.e. 67% of T1D CpGs showed same directionality	Rakyan *et al.* 2011 [301]
T1D	Three immune effector cell types. CD4+ T cells, CD19+ B cells and CD14+CD16− monocytes	52 MZ twin pairs discordant for T1D	Infinium HM450k	1 probe diff in mean DNAm in T cells. DVPs: 10,548 DVPs in B cells, 4,314 in T cells and 6,508 in monocytes	12 T1D-discordant MZ twin. CD14+ and CD4+ cells derived from 201 and 139 unrelated, healthy individuals. 98 newborns of whom 50 had progressed to overt T1D during childhood, while 48 did not.	cord blood of newborns who progress to overt T1D	HM27k: 5 year follow up samples. H450k: unrelated controls analysis + newborns analysis	All DVPs were also hypervariable in CD4+ and CD14+ cells compared to healthy unrelated controls. DVPs FDR<0.001 also present 5 years later (CD14+ monocytes P=7.7 × 10−5 and CD4+ cells P=7.8 × 10−3). Zero (0) found in cord blood pre-disease.	Paul *et al.* 2016 [314]
T1D	whole blood	7 long-term disease-discordant MZ twin pairs and five pairs of HLA-identical, disease-discordant non-twin siblings (NTS)	Infinium HM450k	20,760 in discordant MZ twins. In HLA identical NTS analysis 12,362 CpGs	6 additional T1D-discordant MZ twin pairs		Technically validated via pyrosequencing of five candidate CpG	0 of 5 CpGs	Elboudware *et al.* 2016 [315]
RA	White blood cells (PBL's)	5 MZ discordant twins, 2 matched controls/twin pair	Goldengate	0	-	-	-	-	Javierre *et al.* 2010 [281]
RA	Fibroblast like synoviocytes (FLS)	5 RA, 6 Osteoarthritis (OA)	Infinium HM450k	1859 RA vs OA	5 RA, 6 OA, 6 normal	FLS	HM450k	7 (gene expression), Top 20 loci clustered unseen samples correctly	Nakano *et al.* 2013 [299]
RA	Fibroblast like synoviocytes (FLS) @ 3rd, 5th and 7th passage	3 RA, 3 OA, 3 normal	Infinium HM450k	2375 genes vs OA & Normal FLS	-	-	-	-	Whitaker *et al.* 2013 [302]

Table 1.2 Continued. Summary table of EWAS studies in 4 key autoimmune diseases, plus JIA

Disease	Cell type used	Discovery cohort size	EWAS method	Significant probes	Validation sample numbers (independent)	Validation cell type	Validation cells notes	No. CpGs validated in indep. samples	Reference
RA	Fibroblast like synoviocytes (FLS)	6 RA, 6 OA	Infinium HM450k	2571	-	-	-	3/3 CpGs	De La Rica *et al.* 2013 [291]
RA	Whole blood	354 cases, 337 controls	Infinium HM450k	51,476	12 cases, 12 controls	CD14+ monocytes	HM450k	3/3 CpGs	Liu *et al.* 2013 [245]
RA	B cells	50 cases and N = 75 healthy controls	Infinium HM450k	64 significant CpG sites	1st. N = 14 patients and N = 15 controls. 2nd. N = 11 cases and N = 11 controls. 3rd 56 SLE and 47 control B-cell samples	SLE: B cells	HM450k	10/64 CpG sites in first replication cohort. 4/64 in second replication cohort. In 3rd (SLE) cohort, all show same direction	Julià 2017 [310]
RA	CD14+ monocytes, CD19+ B cells, CD4+ naïve and memory T cells	63 Female patients, 57 were seropositive for rheumatoid factor and/or anti-cyclic citrullinated. 31 controls	Infinium HM450k	Using only ~5000 FLS probes, 1056 were also significant in naïve CD4 cells. Genome wide analysis: 0 associations.	-	-	-	-	Rhead *et al.* 2017 [316]
RA	T cells	46 treatment-naive cases	Infinium HM450k	six sites predict treatment response	-	-	pyrosequencing	-	Glossop *et al.* 2017 [317]
RA	CD14+ monocytes, CD19+ B cells, CD4+ naive T cells, and CD4+ memory T cells	63 FEMALE patients, 57 were seropositive for rheumatoid factor and/or anti-cyclic citrullinated	Infinium HM450k	27 CpG sites with active disease. Erosive disease identified 84 CpG sites mostly in B cells (74%). RAAD score: 89 CpG sites mostly in monocytes. Sharp score: 68 CpG sites spread amongst cells types. A handful were shared b/w all cell types.	-	-	-	-	Mok *et al.* 2018 [311]
RA	Monocytes	33 RA (n=10 second visit samples), healthy controls n=17	Infinium HM EPIC	133	-	-	-	-	Rodríguez-Ubreva *et al.* 2019 [312]
RA	PBMCs	2 RA subjects and 2 normal donors	methyl-seq	clustered RA and controls separately	18 CD cases, 25 controls	Whole blood	HM450k	10 CpGs	Tseng *et al.* 2019 [318]
RA	Whole blood	137 subjects, of whom 63 were ACPA-positive, 66 were ACPA-negative, and 8 had self-reported RA	bisulfite sequencing	1303 DMRs, 1909 DMPs	9 RA treatment naïve, 13 controls. DMRs: 8 self reported RA cases	CD4 T cells from RA samples (all female)	15.6% of DMPs, 80 (38%) of 208 RA-specific DMRs were replicated	15.6% of DMPs. 80 (38%) of 208 RA-specific DMRs were replicated	Shao *et al.* 2019 [319]
JIA	CD4+ T cells	13 cases vs 13 controls	Infinium HM27k	145, 11 in MTX naïve only analysis	12 cases, 12 controls	CD4+ T cells	EpiTYPER	2/2 CpGs	Ellis *et al.* 2012 [292]
JIA	CD4+ T cells	68 poly-articular and extended oligo-articular JIA patients	Infinium HM450k	5 CpG modules	-	-	-	-	Spreafico *et al.* 2016 [51]

Table includes information on significant findings and biological validation studies performed. PBL: Peripheral blood lymphocytes. SLE: systemic lupus erythematosus. IBD: Inflammatory bowel disease. T1D: Type 1 diabetes. RA: rheumatoid arthritis. MZ: monozygotic. DZ: dizygotic. ACPA: anti-citrullinated protein antibodies. HM27k/450k/EPIC: Human Methylation27/450/EPIC beadchip array. DMR: differentially methylation region. DVP: differential variable methylation position.

1.7.1 Systemic Lupus Erythematosus

SLE studies are generally grouped into 2 types of studies, those looking at sub group analyses and those comparing general SLE differences. The common theme amongst all is the presence of interferon related genes and pathways as important to SLE generally and to subgroups with worsening or active disease.

Subgroup differences in SLE appear to be robustly identifiable, and may only require a small number of CpGs. Auto-antibody presence and nephritis are associated with merely 16 and 19 differentially methylated probes (DMPs) respectively, with both findings replicated in independent samples [309]. However, these used *a priori* clinical classifications, which may contain complex and heterogeneous biological mechanisms. However, Lanata *et al.* relied on biology (using DNAm as a proxy) to determine the subgroups [308]. Their study was able to robustly identify 3 unique subgroups according to severity of disease [308]. Again, only a small number of CpGs are needed to identify groups, with the largest number of CpGs associations belonging to comparisons of the most clinically disparate groups (mild disease vs very severe disease).

Case-control studies generally identified interferon pathways as significantly associated with SLE. This finding was reported in small early, case-control SLE EWAS' [287, 289]. Since then, the enrichment of genes in the interferon pathway has been replicated a number of times, suggesting a robust association with SLE [285, 306-309]. Indeed, it may be present in and slightly more prominent in active or more severe disease [307, 308]. Interestingly, this is in contrast to earlier studies that did not find associations with disease activity [287, 289]. However, this may reflect an epigenetic poising of interferon genes, potentially facilitating interferon expression prior to or during disease flare [287, 289].

1.7.2 Inflammatory Bowel Disease

Studies of genome scale DNAm in inflammatory bowel disease (IBD) typically involved use of mucosal biopsy studies, particularly in earlier studies, whereas later studies favoured investigations of immune cell DNAm [282-284, 286, 290, 294, 295, 298, 300, 313]. Studies usually utilised both Crohn's disease (CD) and Ulcerative colitis (UC) samples. More recent studies in IBD also utilised larger sample sizes, similar to the patterns observed in SLE [282-284, 286, 313]. The general findings tended to replicate across studies, that is, UC and CD DNAm profiles are similar both in PBMCs and colon cells [282-284]. These similarities between UC and CD are suggested to result from an all-encompassing IBD signature [282].

Also generally observed are the differences compared to controls, where CD displays numerous differences unlike the limited differences observed in UC vs controls [282-284].

Whilst inflammation may have contributed to some of the significant associations in IBD, the possibility of disease specific DMPs in IBD was supported by evidence from Howell *et al.* [282]. Use of non-inflamed controls allowed identification of disease specific associations, many of which were stable over time and in culture [282]. The stability of DNAm associations over time are similar to early findings in SLE and perhaps serve a similar purpose hypothesised in SLE. That is, a potential mechanism for priming a cell for inflammation, either by impairing function or poising inflammatory pathways for disease flare [287, 289, 320].

1.7.3 Inflammatory Bowel Disease DNA methylation links with genetics

The enrichment of IBD DNAm associations near known IBD risk loci is a common finding. Enrichment near risk loci is reported to be as close as 10Kb and strengthening when approaching 250kb from DMPs [284]. DMPs correlated with expression also suggest enrichment of risk loci for other complex immune diseases such as T1D and Multiple Sclerosis [282].

Whilst one study was not able to replicate these general findings, differences in enrichment analyses (windows surrounding DMPs vs annotated genes of DMPs) and fewer risk loci used (47 vs ~200) likely contributed to discrepancies [295]. These included use of fewer risk loci (due to fewer known at the time of publication), as well as a limited enrichment analysis which used only annotated genes of DMPs compared to windows surrounding DMPs in other studies.

Genetic risk loci acting through DNAm have also been identified [284, 286]. A SNP near the IBD risk loci *RPS6AK2* [286], was identified as causal of IBD mediated by DNAm, whilst there was suggestive evidence for a similar effect of a *VMP1* SNP [284].

Somineni *et al.*'s genetic analyses suggested previously identified IBD DMPs may be a consequence of disease [284, 286, 313]. However, Somineni *et al.* analysed children, contrasting with analyses of adults in previous studies. Interestingly, analysis of childhood disease may be a powerful method to understand causal biology, as childhood disease may have a stronger genetic risk compared to adult onset, due to the early onset of disease. This

is perhaps evidenced by the children's study being able to detect causal genetic risk loci using 50% fewer samples than the adult study by Ventham *et al.*

Study of childhood disease may facilitate identification of genetically driven (thus potentially causal) DNAm differences, as well as facilitating isolation of DNAm associations consequential to disease.

Nonetheless, it is clear that genetics is likely to play a role in IBD EWAS findings, and may be further elucidated through larger studies.

1.7.4 Rheumatoid Arthritis

EWAS in Rheumatoid arthritis were commonly undertaken employing a case-control design, however most were smaller in size than IBD or SLE EWAS'. The smaller RA EWAS' looked at specific cell types e.g. B cells or monocytes, which may have reduced the potential confounding in studies of whole blood [310-312]. Despite the differences in cell types used, these cell specific studies identified pathways involved in cytokine signalling as well as monocyte or macrophage signalling or differentiation [310, 312]. In addition, pathways for interferon signalling were found in both high disease activity and remission analyses, findings similar though contentious in SLE [287, 289, 306-309].

Liu *et al.*'s EWAS is the most powerful RA study to date, interrogating vastly higher number of samples and integrating GWAS data [245]. This EWAS was also able to identify genetic risk loci for RA acting through DNAm, mainly in the MHC region. They also suggest the monocyte fraction as the pathogenic cell type, since DMPs were robustly replicated in monocytes and contained larger methylation differences potentially masked in whole blood analyses [245]. Variable DNAm according to genotype was also observed, similar to reports from Rodriguez *et al.* [312]. However, Liu *et al.* observed these within the *MHC* region, whilst Rodriguez *et al.*'s variable DMPs were not enriched for the *MHC* region, suggesting that variable DNAm maybe a more widespread phenomenon in RA [245, 312].

Interestingly, out of the four case-control studies, most identified limited numbers of DMPs [245, 310-312]. The largest study to date of over 300 RA samples (Lui *et al.*) reported over 51,000 DMPs, compared to an average of 65 DMPs from other smaller studies averaging 48 RA samples [245, 310-312]. Indeed, one study found no genome wide significant DMPs [311]. Whilst Liu *et al.*'s findings in whole blood were adjusted for cell type, differences in

specific subsets within cell types (which are well known to exist) may have inflated the number of DMPs identified. Further, the cell mixture correction method developed by Houseman *et al.*'s (utilised by Liu *et al.*) was developed from only a maximum of 11 samples that may not may have captured the variability in a population, or the genetic differences well known between Scandinavian and other western European populations [321, 322]. Since Liu *et al.*'s data were sourced from a Scandinavian country these potential deficiencies in Houseman *et al.*'s method may be pertinent, as DMPs may be confounded by cell type differences not fully corrected for.

RA EWAS were also carried out using knee joint synovial fibroblasts (FLS). Nakano *et al.* described just 20 CpG sites able to differentiate RA from Osteo-arthritis (OA) and normal FLS [299]. Through this and follow up studies, it became clear that that RA and OA FLS are very clearly differentiated at the DNA methylation level, as well as gene expression and miRNA expression levels, and that some of these significant DMPs are also reflected in peripheral blood cells [291, 299, 302, 316].

1.7.5 Type 1 Diabetes

Monozygotic twins discordant for disease were primarily used for T1D EWAS, which attempted to remove the confounding effects of differing intra-uterine and early childhood environments as well as genetics. This is particularly important in diseases arising in childhood such as T1D.

Two studies from the same group were able to detect both mean differences and higher DNAm variability in T1D twins [301, 314]. Mean DNAm differences in monocytes were observed in 132 loci, unconfounded by complications of diabetes such as kidney or heart disease. A number of these replicated in independent samples, however none were observed in a smaller study of similar design [315]. Rakyan *et al.*'s DMPs appear to be temporally stable marks, with similar associations observed prior to disease diagnosis, detecting the same direction of effect in 71% of DMPs in pre-T1D samples [301]. These DMPs may also represent disease specific associations, since T1D risk loci and auto-antigen genes (*HLA-DQB1* and *GAD*) were amongst top ranking probes [301].

Increased DNAm variability (differentially variable probes- DVPs) between discordant twins has also been reported, in far larger numbers than DMPs [301, 314]. Increased variability appear to be stable throughout disease, with similar associations observed after a further 5

years of disease [314]. Although mean DNAm differences were observed pre-disease, DNAm variability was not found in pre-disease cord blood [314]. However, DNAm variability may yet manifest after birth and prior to T1D onset [314].

Differences in DMP findings between studies (Rakyan *et al.* and Elboudwarej *et al.*) is likely due to differences in cell types used [314, 315]. The latter study's use of heterogeneous blood samples contrasts with the single cell type (monocytes) used in the earlier study reporting DMPs [315]. Even though a cell mixture correction algorithm was used, residual cell type differences may have contributed to these differences.

Much like in IBD, SLE and RA, T1D DNAm differences may yet be due to an underlying inflammatory or immune process, since disease duration in both studies averaged 11 and ~20 years [301, 314]. Inflammatory processes may also have influenced pre-disease DNAm findings, since pre-disease samples displayed signs of autoimmunity i.e. presence of known T1D associated antibodies [301]. Whilst DMPs precede disease diagnosis, it may not precede pre-clinical disease processes such as the autoimmune inflammatory process, as alluded to by the authors [301]. Further, the effect of inflammation associated with prolonged autoimmunity and potentially poorly managed glucose levels may potentially explain why 29% of DMPs and no DVPs were detected prior to disease. As done for IBD studies, use of known inflammation associated probes may help disentangle diseases specific findings from generalised inflammation derived DMPs.

1.7.6 Juvenile Idiopathic Arthritis

EWAS reports in JIA are limited, with just 2 studies published prior to this investigation. An early study by this group reported a number of associations with oJIA and polyJIA with pathway analysis identifying enrichment around TNF, a current therapeutic target, suggesting findings were disease relevant [292]. Interestingly, many associations were no longer present once methotrexate (MTX) naïve only samples were used, correlating with other studies in RA. The potential effect of MTX on DNAm is plausible, due to the effects on the folate pathway which feed the de novo DNAm and DNAm maintenance mechanisms [323].

Clinical outcomes in JIA may be predictable using DNAm. A more recent study, similarly using CD4+ T-cells, looked at clinical outcomes (post treatment withdrawal) and suggested DNAm markers can predict those who will maintain inactive disease and those who will flare [51]. T-

cell activation pathways appeared to be enriched within top hits, correlating with gene expression findings in CD4+ T-cells. Oddly, DNA methylation patterns were similar between patients who had disease flare, and patients maintaining inactive disease. This may potentially be explained by low levels of inflammation that may exist in patients that do not achieve inactive disease. As such, correcting the analysis for clinical markers of inflammation may help untangle the effects of general inflammation from disease specific mechanisms.

1.8 Risks and Opportunities in Autoimmune disease DNA methylation studies

1.8.1 Effects of medication on DNA methylation

Medications have the potential to influence the methylomes of AID patients. In a number of genome scale studies in T1D, IBD, SLE and RA (disease activity, joint erosion, long term damage medications have been controlled for in statistical analyses and often report no differences on DMPs once these are accounted for [284, 295, 311, 314].

Some studies have attempted preliminary analyses looking at effects of specific medications on DNAm. These studies are usually small, and have reported variable and sometimes conflicting results. Glucocorticoids and azathioprine may have the most robust associations with DNAm, with expansive differences due to glucocorticoids observed in SLE [324]. Anti-TNF treatment has also been associated with DNAm in RA samples but not in IBD [286, 311], whilst statins, thyroxin, immunomodulator's and chloroquine's show no associations [314, 324].

Methotrexate (MTX) affects the folate pathway, important for DNAm maintenance [323]. Four studies in JIA and RA suggest effects of MTX on DNAm [311, 325, 326]. In JIA, significant associations in JIA were significantly reduced in number upon use of an MTX naïve analysis (JIA), whilst analyses lumping methotrexate, chloroquine's and leflunomide also suggested significant associations in long established RA [292, 311]. However, 2 other studies (similar in size to the JIA study - n=33 and n=8 respectively) suggest no genome scale effects of MTX in RA and SLE [307, 312]. Mok *et al.* suggest that effects on DNAm may be cell type specific, where anti-TNF treatment effects naïve CD4+ T-cells and DMARDs preferentially affecting monocytes [311]. The variable susceptibility of cell types for medication induced DNAm differences may potentially explaining conflicting findings. However, all studies are small and may require larger sample sizes.

Studies have attempted to avoid issue of medication by utilising treatment naïve samples. Treatment clearly reduces inflammation, which is known to affect DNAm, but other specific effects are not deeply understood. Whilst some medications may not have direct effects on the methylome, the anti-inflammatory effects and subsequent effects on the methylome are likely to affect many DMPs [286]. These are discussed below.

1.8.2 Inflammation in AID DNA methylation

DMPs identified in autoimmune disease may be consequences of inflammatory processes. It is known that low grade inflammation does indeed modify the methylome, as observed by DMPs associating with C-reactive protein (CRP) [327]. In addition, further associations with CRP, as well as ESR (Erythrocyte sedimentation rate, another clinical measure of inflammation) and histologically identified inflammation have been reported [282, 284, 286, 312]. Indeed, many IBD DMPs overlap with CRP associated DNAm [284, 286]. This is further supported in both IBD and RA, where DMPs revert to control levels in in remission samples or post treatment [286, 312]. Autoimmune specific inflammatory processes may also underlie DMPs, as observed in RA and IBD [312] [295]. In RA, cytokines increased during high disease activity recapitulate DMPs in cultured cells [312]. Autoimmune specific inflammation might also be observable in RA joint cells. Osteoarthritis (OA) has been recently recognised as having inflammatory processes involved [328]. However, OA joint cells (synoviocytes) differ from autoimmune RA synoviocytes at the DNAm level, suggesting autoimmune specific processes associating with RA DMPs [291, 299, 302].

Mitigation of inflammation effects on DNAm have been attempted by use of control samples. In IBD, histologically non-inflamed IBD samples were used to identify inflammation associated DMPs, finding 80-92% of DMPs may be thus associated [282]. Hasler *et al.* could not replicate IBD DMPs in non-IBD inflammatory samples, suggesting IBD specific DMPs were identified [295]. Studies in SLE and RA may need to further address generalised inflammation as a source of confounding amongst DMPs, perhaps by using CRP associated DMPs as performed in IBD studies.

Identifying causal DMPs and mitigating DMPs consequential to inflammation may also be performed using genetics. Causal inference tests to identify causal SNPs acting through DMPs has been completed in IBD and RA, and have identified a small number of CpGs in *RPS6AK2* and potentially *VMP1* (IBD), *USP35* and the *HLA* region (RA & SLE) likely to be causative of disease and not consequential of disease processes like inflammation [245, 284,

286, 308]. However, these reports have been accomplished only in studies with larger sample sizes. Future EWAS incorporating genetics may require even larger studies (as has been completed in GWAS) to identify robust causative SNPs mediated by DNAm.

1.8.3 Clinical outcome associations in AID EWAS

The potential for DNAm to associate with clinical outcomes or measures has been investigated in a number of AIDs. Indeed, the variability in RA DNAm was suggested to be related to disease activity by Rodriguez *et al.* Similar to findings in IBD, RA DNAm was found to be associated with ESR and CRP, both clinical measures of inflammation [312]. Further, disease activity, in which greater inflammation is present, was also associated with DMPs at thousands of CpGs [312]. Similar associations to disease activity in naïve CD4+ cells have been reported in another RA study, and both studies also reported interferon pathways as enriched amongst DMPs, mirroring findings in SLE and IBD [309, 312]. This suggests that DNAm may reflect common inflammatory pathways upregulated during AID disease activity, or that Interferon may be a strong modulator of DNAm.

In SLE, DNAm is able to identify clinically relevant subgroups according to disease severity, outside of conventional clinical grouping [308]. Indeed, the interferon pathway associates with differences between mild and the most severe subgroups of disease. However, whether these differences underlie pathogenesis in more severe disease or is just an effect of disease, is unknown.

There's potential for DNAm to predict clinical outcomes in IBD, with two reports suggesting DNAm can identify either need for surgery, future use of biological (antibody) therapy or those needing therapy escalation [282, 284], or response to medication in RA [317]. Responders to disease modifying anti-rheumatic drugs can be identified with 90% specificity, however specificity and sensitivity is lower and not yet of clinical utility for other predictive algorithms in AID [282, 284] [317]. Prospective cohort studies and larger numbers will be needed to identify highly sensitive and specific algorithms based on DMPs. Further, even in studies with high specificity, a significant risk of overfitting exists (i.e. data being reliable only for the original dataset due to the small and heterogeneous sample) [317]. However, blood DNAm is an accessible and readily measurable marker which facilitates further clinical development of biomarker tests.

A clear precedent of strong findings from AID EWAS exists, suggesting DNAm may well harbour identifiable differences associated with autoimmune disease pathogenesis, disease activity and clinical outcome. Whilst studies in SLE, IBD and RA are more mature and robust, a gap exists for larger and more robust studies in JIA. Further, proof of concept studies in clinical outcome prediction provide rationale for larger studies to enable successful translation of these research findings to the clinic.

1.9 Hypothesis and Aims

The primary hypothesis in this study is that there are key differences in the CD4+ T-cell DNA methylation profile of JIA cases compared to healthy controls, matched for age and sex.

This study utilised a case-control design focussing on a single subtype, oligoarticular JIA, naïve from disease modifying anti-rheumatic drugs and identified at or near diagnosis. This aimed to identify genes or pathways aberrant early in disease.

The approach to investigating DNAm was a genome scale microarray analysis, integrated with genetic analyses. The following chapters describe development of the approach, as well as insights into the potential role of DNAm in oJIA.

Chapter 2. **Materials and Methods**

2.1 Participant recruitment

Participants were recruited through the Royal Children's Hospital (RCH) and Monash Medical Centre in Melbourne, Victoria in Australia as part of CLARITY (ChiLdhood Arthritis Risk factor Identification sTudY) [158]. At the time of sample selection, 604 JIA cases and 786 controls had been recruited, incorporating all 7 ILAR subtypes of JIA. Cases were recruited from rheumatology clinics, and a blood sample collected. Questionnaires regarding early life environmental exposures were given to parents or guardians to complete. Controls recruited were otherwise healthy children admitted for minor surgical procedures such as tonsillectomy, pinning of ears or removal of grommets. Blood was collected during the procedure, and parents/guardians also filled in questionnaires on early life exposures. This project was approved by the Royal Children's Hospital and Monash Medical Centre human research ethics committees, HREC #27127 and HREC #12377 respectively.

2.2 Blood processing

Up to 9 mL of peripheral blood was collected from each participant into a tube containing K_3EDTA as the anticoagulant (Greiner Bio-one, Kremsmunster Austria) and delivered to the MCRI Biobanking facility. Within 2 hours of collection, plasma was removed and stored. Within 24 hours of collection, PBMCs (peripheral blood mononuclear cells) were isolated using Ficoll gradient procedures (GE Healthcare, Illinois USA) and leucosep tubes (Greiner Bio-one, Kremsmunster Austria). Ficoll separation of blood cells allows isolation and separation of mononuclear cells from red blood cells and granulocytes based on density differences between cell types [329]. In order to do this, blood was layered onto the ficoll layer, and centrifuged for 10 minutes at 1000 x g. This step separates cell fractions into layers based on density. The PBMC layer is removed and placed into a new tube, as is the granulocyte layer. Granulocytes are stored at -80˚C. In order to remove platelets from the PBMCs isolated, cells are twice mixed with PBS and centrifuged at 250 x g. To store PBMCs viably, cells are frozen using freeze mix at concentrations ~5-10 x10^7 cell/mL. Freeze mix consisted of FBS and DMSO at 10%v/v. Samples are frozen to -80˚C using a slow cool container (cooling at a rate of 1˚C/min) (Thermo Fisher, Waltham, Massachusetts USA), before being transported to vapour phase liquid nitrogen storage.

PMBCs are a highly diverse complement of immune cells. This is particularly relevant to epigenetics, since DNAm differences according to cell type is a well-known phenomenon

[254-264, 268]. However, these differences may also mask or potentially confound findings from DNAm studies. Importantly, specific cell types are known to be strongly associated with autoimmune diseases e.g. JIA (described in the Introduction, Chapter 1.3.2). As such, investigation of specific cell types may more likely identify robust and potentially pathogenic signatures associated with disease. Specifically, JIA has many known links with CD4+ T-cells, including immunological, histological, molecular and genetic [14, 101-103, 130]. Therefore, both to reduce the cellular heterogeneity that may potentially obscure disease relevant associations, and to facilitate the potential identification of pathogenic associations, PBMC's were further processed to isolate CD4+ T-cells for DNAm analyses. The isolation of CD4+ T-cells are further described in Chapter 3.2.5.

2.3 DNA methylation analysis

2.3.1 DNA extraction

DNA for EWAS work (Chapter 4) was extracted using the Qiagen Flexigene DNA kit (Qiagen, Hilden, Germany) using the cultured cells protocol. This kit is a salt and alcohol precipitation based method using microtubes for the DNA isolation. Cells are lysed in buffer FG1, before having nuclei homogenized and proteins and proteases degraded using buffer FG2 and protease solution. After incubation at 65˚C, DNA was precipitated with isopropanol at a 1:1 ratio with the homogenized nuclei solution. DNA was then washed with 70% Ethanol (Merck, Darmstadt Germany) to remove excess salts, before allowing excess ethanol to evaporate at room temperature. DNA was eluted in volumes ~50 µL.

DNA samples for EWAS were quantified using picogreen as part of quality control checks carried out by a commercial service provider prior to application to methylation beadchip arrays (ServiceXS, Leiden, The Netherlands). A subset were also quantified using the Nanodrop 1000 in house (Thermo Fisher Scientific, Waltham, USA), and results were very similar (data not shown). Therefore, DNA samples used for other projects described in this thesis were confidently quantified using the Nanodrop 1000.

2.3.2 Bisulphite conversion

The DNA bisulphite modification technique was developed to convert unmethylated cytosines to uracil, thus effecting a nucleotide base change that can be detected similarly to the alleles of a single nucleotide polymorphism (SNP) (Figure 2-1) [330]. In this way, the methylation status of C bases within CpG motifs can be measured.

The process for conversion is as follows. DNA is fragmented at 95˚C, then bisulphite converted by addition of sodium bisulphite to deaminate cytosine residues in fragmented DNA. This process adds a sulphite moiety, which is removed (desulphonation) by incubation at high pH to yield uracil. The methylated cytosines are protected from the process, allowing differentiation of methylated vs unmethylated residues by a base change of cytosine to uracil Figure 2-1 [331].

Samples for EWAS were bisulphite converted by the commercial service providers (ServiceXS, Leiden, The Netherlands), utilizing the Zymo EZ DNA Methylation gold kit (Zymo Research, Irvine USA). Other bisulphite conversions required for technical validation and replication studies described in this thesis also utilized this kit to maintain consistency and avoid differences in methylation due to conversion methods. This kit employs a column based method. 250-500 ng of gDNA was denatured and simultaneously bisulphite converted in a 1.5 mL microtube by a 98˚C incubation for 2.5 hours, followed by 64˚C incubation for 2.5 hours and 4˚C hold overnight (for all samples). The sample was mixed with M-binding buffer, applied to the column, and bound to the membrane by a 20 second centrifugation at 18,000 x g. The bound converted DNA was then washed with M-binding buffer and the centrifugation repeated. M-Desulphonation buffer was then applied to the column and incubated for 15-20 minutes at room temperature. Repeating the centrifugation step as above, the sample was then washed using M-Wash buffer twice. Converted DNA was then eluted from the membrane using M-elution buffer, incubating for 10 minutes prior to the centrifugation step as previously described. DNA was then stored at -20˚C.

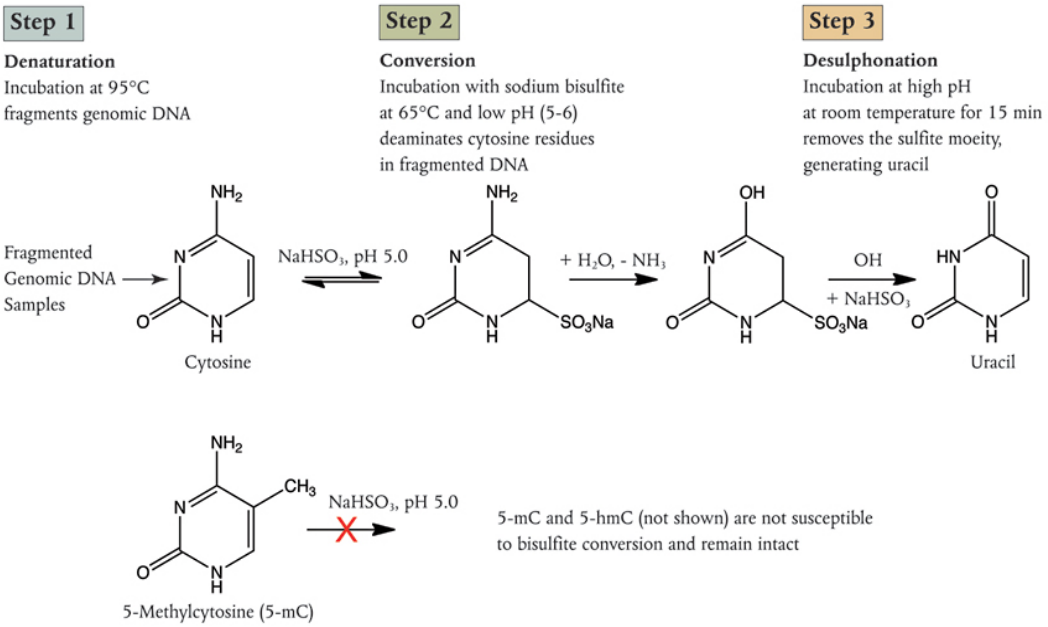

Figure 2-1 Bisulphite conversion of genomic DNA.

Chemical reaction steps involved in converting a cytosine with an unmethylated carbon at the 5 position (5-C) to a uracil. Cytosine with a methylated carbon at the 5 position (5-mC) are protected from this conversion and remain a cytosine. Therefore, differences in methylation can be detected as a nucleotide differences [332].

2.3.3 Genome scale analysis of DNA methylation

Illumina Infinium HumanMethylation450 BeadChips (HM450K) (Illumina, San Diego, USA) were used for genome scale DNA methylation analysis. The arrays measure DNA methylation at over 485,000 CpG sites throughout the genome covering 99% of RefSeq genes, and averaging ~17 CpGs per gene region [333]. Regions covered include gene bodies, regulatory regions as well as intergenic regions. Samples were distributed randomly across chips to avoid confounding due to inter-chip variability.

2.3.3.1 Preprocessing

Once data had been generated, preprocessing of data attempted to remove systematic technical errors inherent to the assay. In this context, background fluorescence of probes and differences in colour intensities between probes needs to be adjusted for. Further, since the HM450K arrays employ 2 different probe chemistries to interrogate DNAm (Type I and Type II), differences in DNAm distribution due to probe types also needs to be normalised [333].

Data was analysed using R studio (www.rproject.org), and packages used for analysis were downloaded from Bioconductor (www.bioconductor.org). The Minfi package was utilized to carry out quality control and preprocessing steps [334]. Samples were visually inspected for deviations from the expected bimodal distribution of DNAm, utilising the 'qcReport' function in Minfi. To adjust for background fluorescence and differences in colour intensities, the 'genome studio method' was implemented. Probe type differences need to be adjusted for in order to avoid these technical differences affecting DNAm measurements. To do this, SWAN was applied [335].

Next, removal of non-informative and low quality probes was carried out. X and Y chromosome probes were removed to reduce the potential for sex based differences to confound the DNAm analysis. Probes with a detection p-value >0.01, i.e. probes that were not significantly different from background levels, were removed. This left 464,144 probes remaining. Probes with SNPs at the single base extension site with MAF>0.05 were also removed, to avoid common genetic variants directly affecting methylation readings. Probes known to bind non-specifically, listed as part of the Minfi package, were also removed. A total of 430,191 probes were available for analysis. Signal intensities were then converted to beta and M values. Beta values represent the ratio of the methylated intensity over the total intensity (of methylated and unmethylated intensities), ranging from 0-1 and is a more

biologically intuitive value. However, the use of M-values is more statistically valid for differential DNAm analysis, since beta values violate a key assumption of many modelling techniques used for high throughput microarray analysis i.e. variation in standard deviations across the methylation range [336]. This is resolved by the use of M values, which represents the log2 ratio of the methylated and unmethylated intensities, which homogenises standard deviations.

2.3.4 Data exploration

To identify data structures inherent to the dataset, principal components analysis was carried out using the prcomp function within R studio. Principal components were correlated and analysed against potential covariates such as group (case/control), age, sex and chip using Pearson's correlations and student's asymptotic p-values. Data was visualized as a heat map using the WGCNA package in R studio [337].

2.3.5 Differential methylation analysis: RUV

The MissMethyl package was used to identify DMPs via Bayesian linear regression, implementing the RUV4 package in a case-control design [338]. RUV (applied as the RUV-inverse function) attempts to use the data itself to capture factors that need to be adjusted for in linear modelling. It attempts to identify both technical and biological variation not related to the factor of interest (see

Figure 2-2). Illumina negative control probes (INCs) on the arrays enable technical error to be estimated and adjusted for (Stage 1). A differential DNAm (DM) analysis was carried out using INCs to adjust for technical errors. From this analysis a ranked list of probes was generated. This ranked list served to provide an empirically derived probe list. Bottom ranked probes (empirical control probes - ECPs) could then be used to identify covariates for a second DM analysis. These probes represent non-disease associated probes. However, since biological confounders will affect all probes, the bottom ranked probes (ECPs) can be used to identify potential biological confounders. The number of ECPs to use for the analysis was decided by use of MDS plots to visualise clustering of samples. When visual clustering of samples according to case/control status was no longer present, this identified the cutoff of ranked probes to use as ECPs. Once a set of ECPs was identified, a new DM analysis was carried out (Stage 2).

The cycle of DM analysis - ECP generation can continue (Stage 2), as ECPs change after each DM iteration. But ECPs eventually stabilize to achieve relatively constant number. The probe IDs of ECPs were compared between each iteration. Once the number of matching probes stabilized and approached ~ 90% ECP overlap between iterations, n – 1 iterations were used for the final analysis. Numbers of probes and iterations differed between analyses throughout this thesis. For information specific to these analyses, see Materials and Methods sections in the respective experimental chapters.

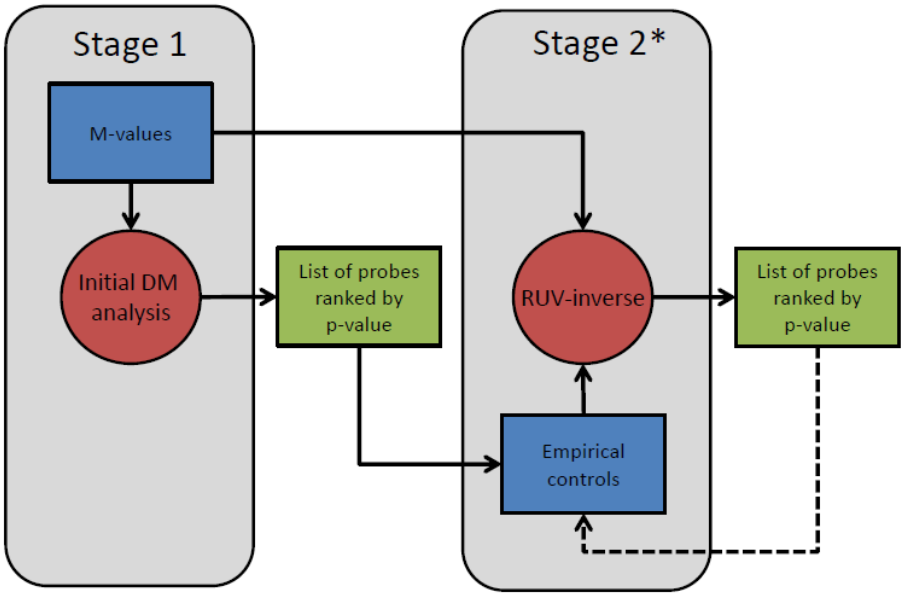

Figure 2-2 Diagram of RUV pipeline for differential methylation analysis.

This is a 2 stage process, where stage 1 identifies a set of probes not associated with the factor of interest. These bottom ranked probes are then used in stage 2 as 'Empirical controls'. These control probes are used to identify unwanted sources of confounding and used in a second DMP analysis as co-factors. Image adapted from Maksimovic *et al.* 2015 [339].

2.3.6 Locus specific DNA methylation analysis: Agena EpiTYPER

The EpiTYPER workflow (Agena Bioscience, San Diego USA) was used for locus specific validation and replication studies of selected CpGs of interest from the EWAS. This was conducted on the MassArray system utilising mass spectrometry (Agena Bioscience, San Diego USA). EpiTYPER is able to detect a sequence change resulting from bisulphite conversion and PCR of methylated and unmethylated CpGs [340]. This difference in methylation states is detected as a mass change of 16 daltons (Figure 2-3).

To detect this mass change, the EpiTYPER system firstly amplifies the bisulphite treated gDNA using a forward and a T7 tagged reverse primer. Dephosphorylation of unincorporated deoxynucleotide triphosphates occurs using shrimp alkaline phosphatase. The reverse strand is then *in vitro* reverse transcribed using the T7 tagged reverse primer, followed by a cleavage reaction. Here, cleavage occurs at a specific base (rUTP- T cleavage) using RNase A. The cleavage products are the substrates of mass spectrometry analysis. The base differences resulting from bisulphite conversion confers a mass difference detectable by MALDI-TOF MS (**M**atrix **A**ssisted **L**aser **D**esorption/**I**onisation – **T**ime **o**f **F**light **m**ass **s**pectrometry). MALDI-TOF mass spectrometry analysis identifies cleavage products which may contain 1 CpG or sometimes more than 1 CpG site. Cleavage products from 2 unique sites may sometimes have identical masses. These are not analyzable as it is not possible to determine the contribution of each [340, 341].

EpiTYPER software v1.2 then calculates unanalyzable fragments and methylation status of fragments. Methylation status is calculated as the ratio of unmethylated (C->T) to methylated (C) signals. If a fragment contains multiple CpG sites, the methylation status will then be an average of all the CpG sites within the fragment.

2.3.7 EpiTYPER analysis: Assay design

Assays were designed to amplify ~200-400 bp of bisulphite converted DNA from selected regions. Regions selected and the primer sequences are outlined in the respective chapters. To design these assays, a region +/-500 bp from the selected CpG was downloaded using UCSC genome browser (https://genome.ucsc.edu/) into a Microsoft word document. Assays were designed manually by performing an *in silico* bisulphite conversion of all non-CpG cytosines to thymines (C->T). The presence of hairpins and dimers were checked using the oligoAnalyzer online tool, as were annealing temperatures (http://sg.idtdna.com/calc/analyzer). Once a primer pair was designed, the amplicon was

analysed for RNAse A cleavage patterns *in silico* using the R package MassArray [342]. This allows visualization of CpGs likely to be analyzable. Amplicons which contained an analyzable CpG of interest selected from the EWAS were selected. To facilitate reverse transcription, a T7 promoter sequence (5´-GTTTAGGTATGTGTTTAGGTGGTGG-3´) was added to the reverse primer, and a 10-mer sequence (5´-AGGAAGAGAG-3´) was added to the forward sequence to aid in balancing the Tm of the T7-promoter tag.

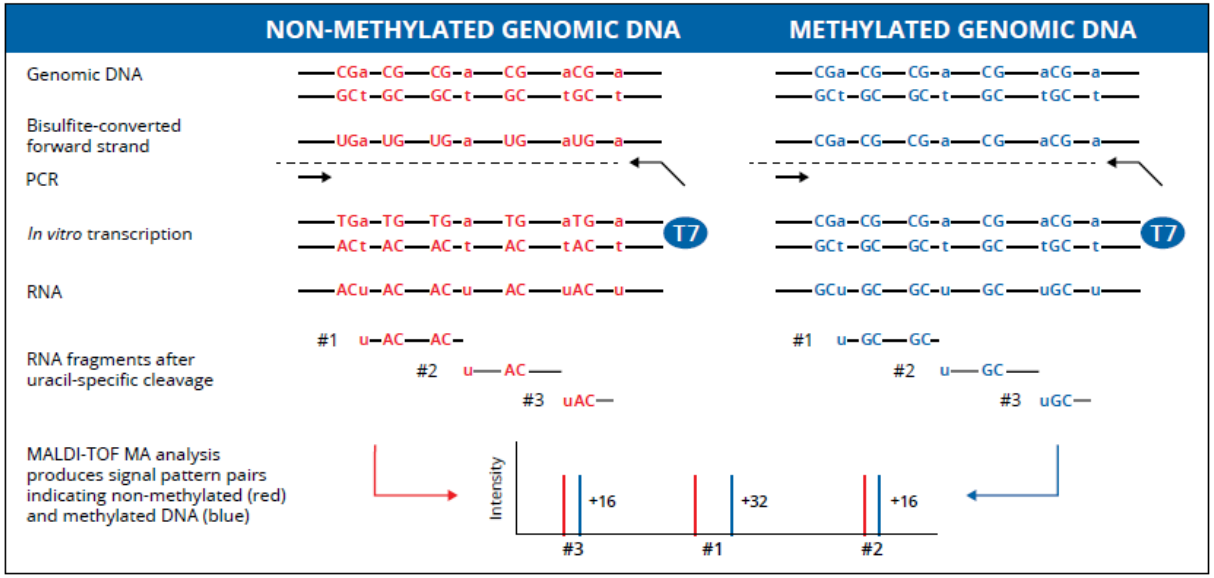

Figure 2-3 EpiTYPER Workflow.

Taking bisulphite converted DNA, the EpiTYPER protocol enables detection of unmethylated (red) and methylated (blue) DNA at single CpG sites (#2 &#3) or multiple CpG sites (#1) via MALDI-TOF. For a single CpG site, this is detectable as a 16Da difference between Adenine (unmethylated) and Guanine (methylated), For 2 or more CpG sites, these are detectable as multiples of 16Da e.g. 2 CpGs→ 32Da difference between methylated and unmethylated sites. (Image sourced from Agena Bioscience website [343]).

2.3.8 EpiTYPER: PCR of bisulphite converted DNA

In order to amplify bisulphite converted DNA, Fastart 2x PCR Mastermix (Roche, Basel, Switzerland) containing 50 units/mL Thermus aquaticus (Taq) polymerase was used. The mastermix contained buffers necessary for PCR reaction, including 400 µM each of dATP, dCTP, dGTP, dTTP, and 3 mM of magnesium chloride as a co-factor for Taq polymerase. A supermix containing this and PCR primers was made, and 15 µL aliquoted using an Eppendorf multipipette used with a 0.1 mL combitip (Eppendorf, Hamburg Germany). PCR volumes, concentrations and thermocycling conditions are outlined below (Table 2-1 & Table 2-2). PCR was completed in a 384 well plate (Thermo Fisher, Waltham, USA) and cycled in a 384 well dual block GeneAmp 9700 system (Thermo Fisher, Waltham, USA). Samples were run in triplicates. PCR cycling conditions were adapted from Coolen et al. 2007 [341]. Modifications to annealing temperatures were made and were assay specific, see below for specific conditions used.

2.3.9 Agarose gel electrophoresis

In order to check whether PCR amplification specific for bisulphite converted DNA had occurred, amplification products was visually checked using agarose gel electrophoresis. Non-converted (gDNA) and converted amplification products were compared. Since primers are specific for converted DNA, gDNA samples were expected to show no bands at the expected size. A 1-1.5% w/v agarose (Bioline, London, UK) was made, with 10 µL of GelRed (Biotium, Fremont, USA). Samples were run for 45 minutes-1 hour at 100V in TBE buffer, and visualized under UV light using the G:Box and Genesys software (Synoptics, Cambridge, UK). PCR products were sizes compared to a 1Kb ladder (NEB, Ipswich, MA USA) and compared to the expected product size.

2.3.10 Dephosphorylation of unincorporated nucleotides

Shrimp alkaline phosphatase (SAP) was used to inactivate unincorporated nucleotides. Reagent volumes for the SAP mix were added as per Table 2-3. 2 µL of SAP mix per sample was added to a new 384 well plate using a multipipette together with a 0.1 mL combitip (Eppendorf, Hamburg Germany). Following PCR reactions, 5 µL of PCR product was transferred from the PCR plate to the 384 well plate containing the SAP mix. This was automated using the Sequenom matrix liquid handler (Agena Bioscience, San Diego, USA). This plate was then cycled in a 384 dual block GeneAmp 9700 system (Thermo Fisher, Waltham, USA) at 37˚C for 20 minutes, followed by 85˚C for 5 minutes, and a 4˚C

temperature hold or placed at -20˚C until needed. This dephosphorylated all unincorporated dNTPs rendering them ineffective for further base extensions.

2.3.11 In-vitro reverse transcription and cleavage reaction

To produce cleavage fragments for mass spectrometry analysis, *in vitro* reverse transcription and T cleavage reactions were completed using reagents and volumes listed in Table 2-4. 5 µL of this mix was added per well of a new 384 well plate using an Eppendorf multipipette with a 0.1 mL combitip (Eppendorf, Hamburg Germany). 2 µL of SAP treated PCR products were transferred to the new reaction plate using the matrix liquid handler (Agena Bioscience, San Diego, USA). This plate was incubated at 37˚C for 3 hours, followed by a temperature hold of 4˚C.

Table 2-1 List of reagents and volumes used for the EpiTYPER PCR amplification step.

Reagent	Conc. in 15 µL	Volume for 15 uL reaction
2x Fastart PCR mastermix including:	-	7.5
Taq Polymerase	0.375 U	
dNTPs	200 µM each	
$MgCl_2$	1.5 mM	
Reverse primer (10 µM)	400 mM	0.6
Forward primer (10 µM)	400 mM	0.6
ddH_2O	-	5.3
Bisulphite converted gDNA	10-20 ng/reaction	1.0
Total		15.0

Table 2-2 EpiTYPER thermocycling conditions used for PCR amplification of bisulphite converted gDNA

Step	Temperature (C°)	Time (min:sec)	Cycles
Denaturation	95	10:00	1
Denaturation	95	00:20	5
Primer Annealing	58-62 (variable- assay dependent)	00:30	
Primer Extension	72	01:30	
Denaturation	95	00:20	40
Primer Annealing	62	00:30	
Primer Extension	72	01:30	
Primer Extension	72	07:00	1
Hold	4	Hold	1

Table 2-3 Reagents and volumes for dephosphorylation of unincorporated nucleotides during EpiTYPER DNA methylation analysis

Reagent	Conc. in 2 µL	Volume (µL) per reaction
Shrimp Alkaline phosphatase (SAP)	0.56 U	0.3
ddH2O	-	1.7
Total		2.0

Table 2-4 Reagents and volumes for T cleavage reaction

Reagent	Conc. in 5 µL	Volume per reaction (µL)
RNAse-free ddH_2O	NA	3.15
5x T7 polymerase buffer	0.64x	0.89
T cleavage mix	NA	0.24
DTT (100 mM)	3.14 mM	0.22
T7 RNA/DNA polymerase	22 Units/reaction	0.44
RNAse A	0.09 mg/mL	0.06
Total		**5.0**

2.3.12 Water and resin addition

Following reverse transcription and T-cleavage reactions, samples were topped up to a volume sufficient for analysis with 20 µL of dH_2O. This was automated using the matrix liquid handler (Agena Bioscience, San Diego, USA). A reservoir of dH_2O and the T-cleavage plate was placed into the liquid handler, and 20 µL was automatically dispensed into each well. To remove salts from the reaction, ~6 mg of resin was added to the sample using the dimpled resin plate (Agena Bioscience, San Diego, USA). Resin and sample were mixed thoroughly by constant rotation for 5-120 minutes at room temperature. Sample plates were then centrifuged for 5 minutes at 3200 x g to pellet the resin.

2.3.13 Spotting and MALDI-TOF analysis

Samples were transferred from a 384 well plate to a 384 spot SpectroCHIP using the Nanodispenser RS1000 (Agena Bioscience, San Diego, USA). Approximately 10 nL of sample was dispensed onto the chip. This was then loaded into the MALDI-TOF MS for fragment analysis, which was performed on site by MassArray facility manager Dr. Ben Ong.

2.3.14 EpiTYPER 1.2 program

EpiTYPER v1.2 program was used to prepare data files for MALDI-TOF analysis. 'Plate Editor' mode was used to initiate a new project for each experiment. Into this, a .txt file was uploaded containing amplicon data including amplicon name, primer lengths and target gDNA sequence. A separate .txt file was used to upload sample names. The 'Analyzer' mode was used to download EpiTYPER DNAm data as a .csv file, using the 'export grid' function.

2.3.15 EpiTYPER DNA methylation data analysis

Three replicates from each sample were available for analysis. All replicates were used if methylation values were +/-10% of the median value. Any samples falling outside of this range were discarded. The average value of all remaining replicates was used to calculate the DNAm value for each sample. Student's t-test was used to assess statistical significance of differences between group means.

2.4 DNA genotyping

2.4.1 iPLEX SNP genotyping using MassArray: Assay overview

SNP genotyping using the MassArray system (Agena Bioscience, San Diego, USA) is a method of medium-throughput detection of insertions, deletions, substitutions and other

polymorphisms in PCR amplified DNA. The method utilizes MALDI-TOF mass spectrometry. It is based on extending a primer by a single base over the SNP of interest using mass-modified ddNTPs. Mass modification allows for more convenient identification of different bases/extension products by mass. Since different bases have different masses, polymorphisms are identified by mass differences of single base extended products. The process is outlined in Figure 2-4.

Firstly, assays are designed using the online assay design software Assay design suite V2.0 (https://www.agenacx.com/Home). Once assays are designed, PCR of gDNA amplifies DNA. After unincorporated dNTPs are dephosphorylated using Shrimp alkaline phosphatase, the iPLEX reaction can take place. This is the single base extension step, as described above. Mass modified ddNTPs terminate the reaction after one base extension. Mass modification of ddNTPs also increases the mass of the single base extension product, thereby allowing detection via the MassArray MALDI-TOF MS. The available window of analysis (4500 dalton (Da) to 8000 Da) means that reactions can be multiplexed with theoretical limits of 40 unique SNPs assayable [344, 345].

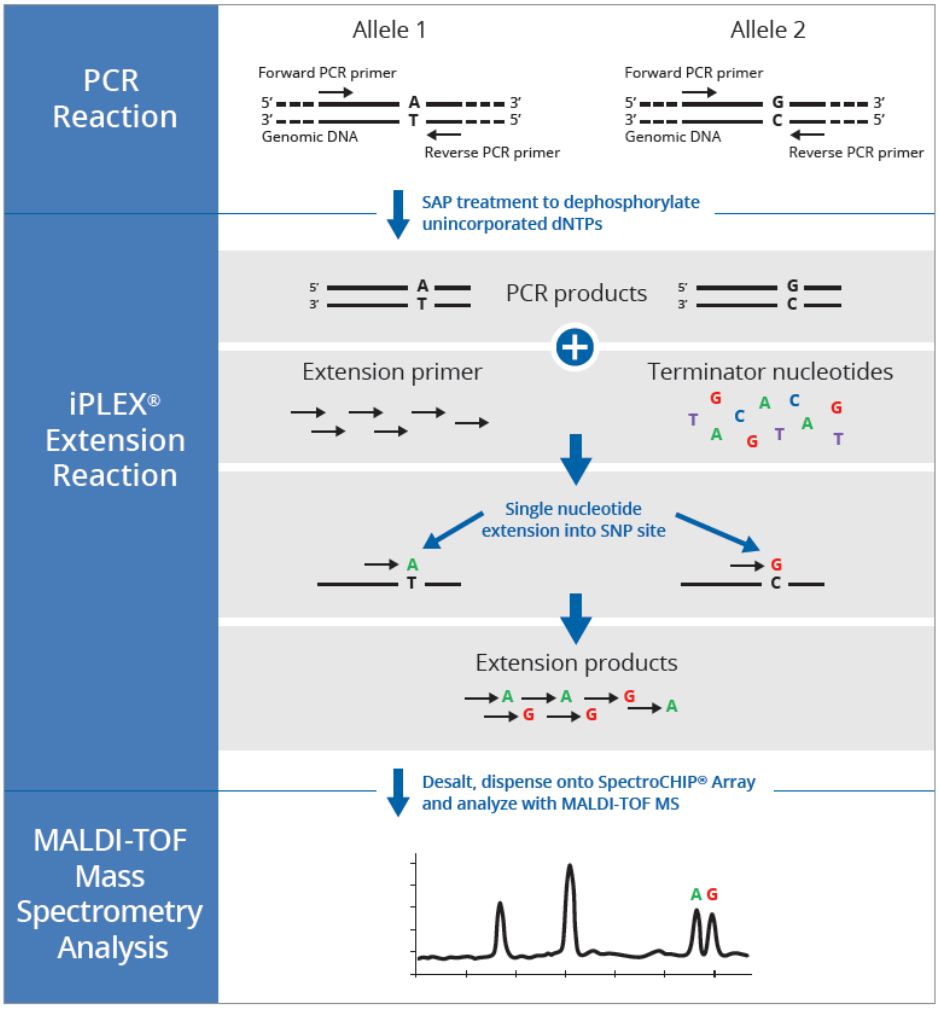

Figure 2-4 Overview of MassArray iPLEX genotyping process.

Beginning with genomic DNA, a SNP is able to be interrogated by use of a single base extension (iPLEX) reaction. The resulting extension product is measured by MALDI-TOF to detect single base differences directly related to the SNP of interest [346].

2.4.2 SNP selection

SNPs surrounding probes of interest from Chapter 4 were investigated for potential correlations to 1) mean methylation levels or 2) methylation variance according to genotype.

Selection of SNPs for analysis of genetic effects on methylation revolved around 3 themes: 1) known methylation quantitative trait loci (mQTLs) 2) Regional SNPs (+/-10 kb) and 3) Distal SNPs located with enhancer regions upto +/-100 kb. Specific regions (i.e. +/-10 kb or specific enhancer sites) were selected since the task of genotyping all SNPs across large genomic regions (+/-100 kb) was not feasible due cost and resource restrictions. Therefore 3 themes for SNP selection were chosen and are described below.

2.4.2.1 Known mQTLs

Known mQTLs were selected from the published literature. Papers were selected on the basis of:

- Tissue type: only data originating from immune related cell types were used
- EWAS data type: data must have originated from either Illumina Infinium HumanMethylation27 or HM450K beadchip arrays.

Accordingly, 2 papers were selected. Liu et al 2014 [222], which looked at whole blood samples collected from rheumatoid arthritis patients and controls. This paper analysed methylation data using HM450K arrays, and was a more comprehensive study of SNP associations with methylation data than this groups previously published data (Liu et al 2013 [245]). Therefore, the Liu et al 2014 paper was selected. The second paper, Guitierrez-Arcelus et al 2013, investigated methylation x SNP interactions in a selection of cell types from healthy newborn babies, including T cells and lymphoblastoid cell lines [247]. The mQTLs associated with T cells were used.

2.4.2.2 Regional SNPs

To identify any new potential mQTLs in the region immediately surrounding the probe of interest, a region of +/- 10kb was used from the probe location to identify SNPs. SNPs were identified using the HapMap database (Phase 3), and all SNPs with minor allele frequency (MAF) of 0.05 or greater were downloaded. A tag SNP approach was used for selecting the

final list of SNPs for a particular probe. The tag SNP approach takes advantage of high linkage disequilibrium (LD), that is, SNPs that are likely to be inherited together. A tag SNP approach selects a representative SNP from a region with high LD, which enables fewer SNPs to be genotyped whilst representing larger regions of the genome [347].

Haploview was used to capture all SNPs within the region with a high level of linkage disequilibrium using r^2 >0.8, plus a MAF >0.05 to produce a final list of tag SNPs [348].

2.4.2.3 Distal SNPs within enhancers

Tissue specific promoters such as enhancers may be home to genetic variants that may modify DNAm, as proposed by Feinberg et al 2010 [236]. To identify associations of genetic enhancer regions with DNA methylation, enhancers surrounding the selected probes were visualized using UCSC genome browser (https://genome.ucsc.edu/). Data from the BLUEPRINT epigenomics resource (www.blueprint-epigenome.eu/) was used to identify immune cell related enhancer sites. Cell types selected were CD4+ T cells, CD8+ T cells, CD8+ memory T cells, B cells and monocytes. Enhancers +/-100kb from the selected probes were visualized, since a number of papers have found that a large majority of mQTLs can be found within the 100kb window from probes of interest [220, 349].

SNPs present in the enhancer region were selected for genotyping. SNPs from these regions were downloaded from HapMap Phase 3. Haploview was used to select tag SNPs based on r^2>0.8 and MAF >0.05 [348].

2.4.3 Assay design

Once SNPs of interest were chosen, assays were designed using the online assay design software (https://www.agenacx.com/Home), using high multiplexing settings. Assays were designed to be 80-120 bp in length for optimal PCR amplification. For each region, SNPs of interest were uploaded and sequences were retrieved using the 'Retrieve and Format sequences' function. This also retrieves proximal SNPs in the region. The primer design algorithms attempt to avoid proximal SNPs during the design process ('Identify Optimal Primer areas' function). Any proximal SNPs identified that were identified to be interruptive of primer locations were ignored if that SNP did not have any population data or MAF<0.05. The 'Design Assays' function completed the process. This outputs primer pairs for each SNP of interest, plus a single extension primer with which the single base extension reaction will occur.

Identified PCR primer pairs had 10-mer tags (5'-ACGTTGGATG-3') added to the 5' ends to increase their masses to those outside of the detectable range in MALDI-TOF MS (4500 Da to 8000 Da).

Primers were sourced from IDT (IDT Technologies, Coralville, USA). These were shipped as lyophilized pellets, and reconstituted at 100 µM for PCR primers, and 500 µM for extension primers.

A PCR primers pool was made for each multiplex reaction. The multiplex PCR primer mix contained all forward and reverse primers at a final concentration of 0.5 µM.

2.4.4 DNA amplification using Polymerase chain reaction

All reactions for the genotyping process were undertaken using the iPLEX GOLD complete genotyping kit (Agena Bioscience, San Diego, USA). To amplify regions of interest, PCR was performed. Volumes of reagents per reaction are listed in ***Table 2-5***, and added using a multipipette with a 0.1 mL combitip (Eppendorf, Hamburg Germany). Genomic DNA at 5-15 ng/µL was used and added using an 8 well multichannel pipette. The reactions were performed in a 384 well PCR plate (Thermo Fisher, Waltham, USA), and cycled on a GeneAmp thermocyler 9700 system with a 384 well block (Thermo Fisher, Waltham, USA). Cycling conditions are listed in ***Table 2-6***.

2.4.5 Dephosphorylation of unincorporated nucleotides

Shrimp alkaline phosphatase (SAP) was used to inactivate unincorporated nucleotides. Volumes of Shrimp alkaline phosphatase and RNAse free ddH_2O were added as per Table 2-3. 2 µL per sample was added to the PCR product in the existing 384 well plate using a multipipette together with a 0.1 mL combitip (Eppendorf, Hamburg Germany). This plate was then cycled in a 384 dual block GeneAmp 9700 system (Thermo Fisher, Waltham, USA) at 37˚C for 45 minutes, followed by 85˚C for 5 minutes, and a 4˚C temperature hold. This dephosphorylated all unincorporated dNTPs rendering them ineffective in further base extensions.

2.4.6 Linear primer adjustment of unextended primers

For multiplex reactions, a pool of unextended primers was made. Due to the inverse relationship between signal and increasing molecular mass of primers, adjustment of primer concentrations within the mix is needed to obtain sufficient signal. This is particularly

important for primers at the higher mass end of the spectrum. To do this, extension primers for the unextended primer mix were adjusted with water according to the linear primer adjustment method, such that primer concentrations within the reaction will vary from 0.5-1.5 µM. This regression method is implemented in a spreadsheet provided by Agena (with thanks to Dr. Darryl Irwin, Agena Bioscience Asia Pacific). To calculate the volume of water to add to stock primers, the mass of the primer, primer stock concentrations, final volume of primer mix and the number of primers are taken into account. Then, equal volumes of the adjusted primers are pooled, constituting the primer mix for iPLEX extension reactions.

2.4.7 iPLEX reaction

Once dinucleotides had been dephosphorylated, the single base extension reaction was carried out using the iPLEX GOLD reagents (Agena Bioscience, San Diego USA). Volumes and reagents are as listed in Table 2-7. Volumes of reagents vary according to the multiplex level, differing between low plex reactions i.e. reactions with < 18 PCR assays, and high plex reactions (>18 assays). These variations are listed in the tables. 2 µL of the iPLEX mix was added to the SAP reaction plate using a multipipette (Eppendorf, Hamburg Germany). Reaction plates were then thermocycled according to Table 2-8, using the GeneAmp 9700 system (Thermo Fisher, Waltham USA).

2.4.8 Water addition and reaction cleanup.

To top up the volume of the reaction sufficient for analysis, 16 µL of dH_2O was added. This was automated using the matrix liquid handler (Agena Bioscience, San Diego, USA). A reservoir of dH_2O and the iPLEX reaction plate was placed into the liquid handler, and 16 µL was automatically dispensed into each well. To remove salts from the reaction, resin was used as per 2.3.12.

2.4.9 Spotting and MALDI-TOF analysis

Samples were transferred from a 384 well plate to a 384 spot SpectroChip, as noted in 2.3.13. Typer 4.0 software was used to create plate files in preparation for analysis by MALDI-TOF MS (Agena bioscience, San Diego USA). The ‘Plate Editor’ mode was used to define a new plate of samples. Sample lists were uploaded .txt files, and applied to experimental plates. Assays were uploaded into ‘Assay Editor’ as text-tab delimited files in .txt format, and then applied to the experimental wells.

Once MALDI-TOF analysis was complete, data was visualized and quality controlled also undertaken as described below using the Typer 4.0 software (Ageno Bioscience, San Diego, USA). Assays that had <90% calls made were removed. Samples with ‹90% calls were removed. No SNPs were out of Hardy-Weinberg equilibrium and were usable for further analysis. Samples were also visualized using cluster plots of genotypic groups. Clustering of data was defined by cluster settings, using thresholds of 0.1 (low mass heterozygote), 0.2 (homozygote), 0.1 (high mass heterozygote). Samples falling outside of the defined cluster fields were visually inspected and manually adjusted to the appropriate genotype group, or deemed as a ‘No call’ if data was ambiguous.

Table 2-5 Reagents and volumes used for PCR amplification steps for genotyping using the iPLEX method.

Reagents	Conc. in 5 µL	Volume (µL) Low Plex	Conc. in 5 µL	Volume (µL) High Plex (>27plex)
dH_2O	-	1.9	-	1.8
10x PCR buffer (inc. 20 mM $MgCl_2$)	1X (2 mM $MgCl_2$)	0.5	1X (2 mM $MgCl_2$)	0.5
$MgCl_2$ (25 mM)	2 mM	0.4	2 mM	0.4
dNTP mix (25 mM each)	500 µM	0.1	500 µM	0.1
PCR primer mix (500 nM each)	100 nM	1.0	100 nM	1.0
PCR enzyme (5 U/µL)	0.5 U/reaction	0.1	1 U/reaction	0.2
gDNA (5-20 ng/µL)	5-20 ng/µL	1.0	5-20 ng/µL	1.0
TOTAL		5.0		5.0

Volumes differ according to plex level, i.e. low vs high plex. High plex reactions include assays that have >27 SNPs interrogated.

Table 2-6 Thermocycling conditions used for PCR amplification of gDNA for the iPLEX genotyping process

Step	Temperature (C˚)	Time (min:sec)	Cycles
Denaturation	94	04:00	1
Denaturation	94	00:20	45
Primer Annealing	56	00:30	
Primer Extension	72	01:00	
Primer Extension	72	03:00	1
Hold	4	Hold	1

Table 2-7 Reagents and volumes used for iPLEX single base extension reaction.

Reagents	Conc. in 2 µL	Volume per reaction -µL (low plex)	Conc. in 2 µL	Volume per reaction (µL) (high plex)
H_2O	-	0.7395	-	0.619
iPLEX buffer	0.222 X	0.2	0.222 X	0.2
Termination mix (ddNTPs)	0.5 X	0.1	1 X	0.2
Unextended primer mix	0.5-1.5 µM	0.94	0.5-1.5 µM	0.94
iPLEX enzyme	0.5 X	0.0205	1 X	0.041
Total		2.0		2.0

Reaction volumes differ according to plex level. High plex reactions include assays that have >18 SNPs interrogated. This differs to the high plex value used in PCR reactions. Concentrations of unextended primer mix are noted as a range and are calculated based on the assay and primer specific details as noted in 2.4.6.

Table 2-8 Thermocycling conditions for iPLEX single base extension reaction.

Step	Temperature (˚C)	Time (min:sec)	Cycles
Denaturing	94	00:30	1
Denaturing	94	00:05	40 cycles
Annealing	52	00:05	5 cycles
Extension	80	00:05	
Extension	72	03:00	1
Hold	4	forever	1

This is a nested PCR reaction, wherein the a single denaturation step is followed by the single base extension reaction occurring as repeating 5 second annealing and extension reactions. This process is then repeated over 40 denaturation steps.

2.4.10 SNP association analysis

2.4.10.1 Methylation quantitative trait loci analysis

To investigate the relationships between genetics and methylation, SNP genotype and methylation data from the same loci were analysed. The EWAS loci are noted in chapter 4.3.4. All analyses were carried out in R studio.

To analyse the difference in mean methylation between allele groups, normally distributed data was analysed using linear regression modelling. Non-normally distributed data was analysed using the Kruskal-Wallis test.

Post-Hoc tests were completed for significant associations observed from Kruskal-Wallis analyses. To do this, the Dunn test was used to assess associations between pairs of genotypes.

To assess the variation in methylation between allele groups, normally distributed data was analysed using Bartlett's test. For non-normally distributed data, Levene's test was used.

Sensitivity analyses were carried out to understand potential confounders in the data and understand the robustness of the initial findings. These analyses were conducted in controls only or females only (as these groups represented the largest sample groups and largest potential source of confounding).

Methylation distribution was visualized using density plots within the ggPlot2 package [350].

2.4.10.2 Allelic association analysis

To assess whether there were associations of SNPs with oJIA status, an allelic association analysis in a case-control design was implemented using PLINK [351]. Odds ratios (ORs) and 95% confidence intervals (CIs) as well as p-values were calculated, with a threshold of significance set at 0.05.

2.5 Functional analyses

2.5.1 RNA isolation

Gene expression data was generated to complement DNA methylation data, and assess the EWAS data for functional relevance to disease, immune or epigenetic processes (See chapter

4). To do this, RNA was extracted using the miRNeasy kit (Qiagen, Hilden Germany). Previously extracted RNA from manual TRIzol extractions was also purified using this kit, to provide high quality RNA for RNAseq studies. This work was partially completed by Rachel Chiaroni-Clarke. The kit is a column based method, where RNA is bound to a silica based matrix. Buffers are placed into the matrix and the columns spun to wash the column of cellular matter, proteins and salts, before RNA is eluted from the matrix. The kit uses a phenol and guanidine thiocyanate solution (either TRIzol or Qiagen's own QIAzol solution (Qiagen, Hilden Germany). This solution (700 µL) is added to lyse and homogenise cells, and also to protect the sample from RNAse activity. Then 140 µL of chloroform is added. This is vigorously mixed and left at room temperature for 2-3 minutes.

To isolate RNA, the sample is centrifuged at 12,000 x g for 15 minutes at 4˚C. The aqueous layer is transferred to a new tube, and 1.5 volumes of 100% ethanol is mixed with the aqueous layer (Merck, Darmstadt Germany). Up to 700 µL is placed into the miRNeasy mini spin column, and bound to the column by centrifugation using at least 8,000 x g for 15 seconds at room temperature.

The column is then washed with buffer RWT, a proprietary blend stringent washing buffer containing guanidine salts and ethanol. It is then washed twice with another proprietary buffer, RPE, which also contains ethanol to remove traces of salts. RNA is then eluted using RNAse free water.

A subset of samples had RNA extracted using a modified TRIzol based protocol (completed by me), followed by clean-up using the miRNeasy kit (completed by Rachel Chiaroni-Clarke).

2.5.1.1 DNA removal from TRIzol extracted RNA samples

As part of the TRIzol modified protocol, DNA was removed using the DNA-free kit (Thermo Fisher Scientific, Waltham USA.). DNase I and cation containing buffer is added to the RNA solution, then incubated at 37˚C for 20-30 minutes. Since divalent cations can also catalyze RNA degradation in a heated solution, these need to be removed. Using 0.1 volume of the RNA sample, the DNase inactivation resin is added to inactivate DNases and remove cations. This is incubated for 2 minutes with occasional mixing. The resin is pelleted at 10,000 x g for 90 seconds, after which the supernatant containing DNA-free RNA is removed into a new tube. RNA was then stored at -80˚C.

RNA samples for RNAseq were quantified using the Qubit system (Thermo Fisher, Waltham USA) by in house service providers at Murdoch Children's Research institute.

2.5.2 RNAseq

To generate gene expression data, RNA was sequenced by in house sequencing service providers, to a depth of 20x. RNAseq data was preprocessed by Rachel Chiaroni-Clark, and is described below.

Processing, normalisation and analysis of read counts was performed using statistical methods from the RUVseq, EDASeq and edgeR packages [352-354]. 75 base pair paired end reads were produce by RNA-sequencing. Fastq files containing raw paired-end reads underwent quality control checks using FASTQC. Trimmomatic was then used to trim reads, and STAR was used for alignment to a reference genome. BAM files were then created containing the mapped reads using SAMtools. To count reads to a gene regions (annotated by Entrez gene identifiers), featureCounts was used. Entrez gene identifiers were also annotated using Ensembl gene ID conversion (ensembl.org). Log(fold change) of cases vs control data was extracted using LIMMA [355].

To obtain information on overlaps between DNA methylation data and RNAseq data from the same genes, DNA methylation probes was also annotated using Ensembl gene ID tool (www.ensembl.org). For a subset of samples that had both DNAm and gene expression data, data was paired using the matching Ensembl gene ID. Data was visualized using Prism 7 software (Graphpad, La Jolla, USA).

Chapter 3. **Optimisation and Validation of methods**

3.1 Introduction

Genome scale DNA methylation studies have their foundation in the use of microarrays, particularly the Illumina HumanMethylation450 BeadChip. Service providers of these assays have minimum requirements for sample quality and quantity, below which data quality cannot be guaranteed. In addition, PBMC samples from the CLARITY cohort were limited. Therefore, optimisation of yields from cell sorting protocols, DNA and RNA extraction protocols were necessary to obtain sufficient numbers and sufficient sample for high quality data.

In addition, optimisation of a criteria for the selection of probes to validate was necessary, owing primarily to the limited amount of DNA available reducing the number of validation experiments possible. Further, the small differences in DNAm previously observed in complex diseases or traits [245, 356], and the detection limitations of the technology to be used for validation [340, 341] also suggested that development of a selection criteria more likely to identify true DNAm differences was important.

3.2 Methods

3.2.1 DNA extraction

DNA and RNA can be isolated simultaneously from samples using the AllPrep DNA/RNA mini kit (Qiagen, Hilden, Germany) according to manufacturer's instructions. Briefly, to isolate DNA, cells were pelleted at 300 x *g,* and supernatant was removed. To lyse cells, RLT buffer was added, and mixed with pipetting to initially mix cells then vortexing for 1 minute. The homogenised lysate was then transferred to an AllPrep DNA spin column, and DNA bound to the column by centrifugation for 30 seconds at 8000 x *g*. The flowthrough containing RNA was then placed aside in -80˚C.

To continue DNA isolation, the bound DNA was washed with buffer AW1 (500 µL) by centrifugation at 10,000 x *g* for 30 seconds. A further wash of the bound DNA was completed using 500 µL of buffer AW2, spinning the column at 18,000 x *g* for 2 minutes. Extra drying of the column completed with another centrifugation step at 10,000 x *g* for 30 seconds. To elute the DNA, 100 µL of EB buffer at room temperature was placed onto the column and incubated for 1 minute, followed by centrifugation of the column for 1 minute at 10,000 x *g*.

To optimise this extraction method, modifications to the protocol were performed as described below:

- Rebinding of the flowthrough post DNA binding: To allow any residual DNA to be bound to the column.
- Use of EB buffer at 50˚C: To increase the amount of DNA eluted from the column.
- Incubation of EB buffer for 10 minutes: To increase the amount of DNA eluted from the column.

In addition, the following step was added to increase DNA elution from the column:

- Second elution: Another volume (100 µL) of EB buffer at 50˚C was incubated over the column for 10 minute.

DNA extraction efficiencies were calculated based on the expected DNA amount (i.e. 6 x 10^{-12} grams DNA per cell x no. cells). A ratio of actual vs expected was then used to calculate the efficiency of DNA extraction.

3.2.2 DNA precipitation

During DNA isolation utilising the AllPrep mini kit (Qiagen, Hilden Germany), flowthrough from DNA binding and wash steps were normally discarded. However, these may contain residual DNA. In order to isolate DNA from flowthroughs, DNA was precipitated as described below.

Firstly, 3M sodium acetate (pH 5.2) was added (at 1/10 volume of flowthrough). Then, 100% ethanol was added at 2.5x the total volume (sample + sodium acetate). The sample was left to incubate for 20 minutes at room temperature. The DNA was pelleted by centrifugation at 18,000 x *g* for 10 minutes. Supernatant was removed and DNA was washed with 70% ethanol before centrifugation again at 18,000 x *g* for 10 minutes. DNA was dissolved using distilled water.

3.2.3 RNA extraction

After cell sorting, TRIzol solution was added to cells which lysed, protected and homogenised the cellular material. Samples were lysed using a 21-gauge needle and syringe to shear the cell membrane. Then, 200 µL of chloroform per 1 mL of TRIzol was added to this solution, shaken vigorously and incubated for 5 minutes at room temperature. Centrifugation at 4˚C at 12,000 x *g* then separates the solution into phases, where the uppermost aqueous layer contains RNA. This aqueous phase was removed and placed into a new tube. An equal volume of dH_2O was placed into the TRIzol tube, and the centrifugation for phase separation was repeated. The aqueous phase was again removed and placed together with the previously removed layer. 100 µg of glycogen (Invitrogen, California USA) was added to the aqueous phase, prior to precipitation of RNA using an equal volume of 100% isopropanol (Merck, Darmstadt Germany). Tubes were incubated at -80˚C for 2 hours to precipitate RNA. Following which, centrifugation (12,000 x *g* for 15 minutes at 4˚C) was performed. The pellet was then washed with 80% ethanol thrice, to remove contaminants prior to repeating the previous spin conditions. Ethanol was then removed and RNA dissolved using RNAse-free water. To fully dissolve the RNA, tubes were placed in a heat block at 50˚C for 10 minutes. Samples were then immediately treated with DNA removal reagents as described below.

Modifications to the TRIzol protocol were made which attempted to increase yields of RNA, including:

- Addition of 100 µg glycogen to the aqueous phase. This was done to enhance the precipitation of RNA
- Extra phase separation step: addition of RNAse-free dH_2O equal to the volume of aqueous phase removed. This was followed by repeating the phase separation spin and removal of the aqueous phase again. The reason for this was that complete removal of the aqueous phase cannot be done after the initial phase separation spin, as this would also capture Trizol (which contaminates the sample). To capture the remaining 10-20% of aqueous phase volume (containing DNA), the leftover RNA within the Trizol tube was diluted with RNAse-free dH_2O. Another phase separation step allows remaining RNA to be captured within the aqueous phase in a larger volume.
- 2 hour -80˚C isopropanol incubation step. This was to increase the amount of RNA that can be precipitated from the solution.

- 50˚C for 10 minutes incubation step at RNA elution, to enhance dissolving of the RNA into solution e.g. RNA stuck to plastic (Kelly Roeszler, personal communication).
- DNA removal step using DNA-free kit (Thermo Fisher Scientific, Waltham USA.)

RNA was assessed for quality and quantity using the TapeStation system (Agilent, Santa Clara USA). High quality RNA was defined as RINe measure of >7.

RNA was also extracted using the Directzol kit (Zymo Research, Irvine USA). Briefly, RNA in TRIzol was mixed with 100% ethanol and transferred to a spin column. The flow through was discarded. Washing of bound RNA was done by centrifuging 400 µL of RNA PreWash solution over the column at 16,000 x *g* for 30 seconds, followed by centrifugation of 700 µL of Wash buffer for 2 minute at 16,000 x *g*. RNA was eluted by placing RNAse free water onto the column and centrifuging for 30 seconds at 16,000 x *g*. RNA was stored at -80˚C.

3.2.4 DNA removal from RNA samples

RNA expression arrays require pure samples free from DNA. DNA-free (Thermo Fisher Scientific, Waltham USA.) was used to accomplish this. Briefly, 0.1 volume of 10X DNase I Buffer and 1 µL rDNase I were added to the RNA sample. Gentle mixing of the sample was proceeded by 20-30 minutes of incubation at 37˚C. DNAse inactivation reagent was added, also at 0.1 volumes, mixed well and incubated for 2 minutes at room temperature. Samples were mixed intermittently to prevent sedimentation of the inactivation reagent. The inactivation reagent was then pelleted by centrifugation at 10,000 x *g,* and the RNA (supernatant) carefully removed and placed in a fresh RNAse free tube.

3.2.5 Immune cell isolation

To isolate CD4+ T cells, liquid nitrogen frozen PBMC samples were thawed at 37˚C using ‘wake up’ media. This enables cells to resume cellular functions with the salts and nutrients needed. It was inevitable that some cells would lyse during the freezing and thawing process, releasing DNA. Cell free DNA binds cells together and facilitate clumping of live and dead cells, reducing yields from the thawing process [357]. Therefore, to enable optimal yields and provide nutrients for the thawing process, the ‘wake up’ media consisted of RPMI 1640 with DNAse (final concentration of 1 mg/mL). This was added to thawed cells and after gentle mixing, samples were incubated at room temperature for 5 minutes. Cells were washed in the wake up media by centrifuging for 7 minutes at 320 x *g*. The supernatant was removed, and cells were resuspended in 175 µL PBS. To isolate CD4+ T cells, cells were

stained for CD3 and CD4 for 15 minutes at 4˚C (antibodies used for flow cytometry: Becton Dickinson (San Jose, CA, USA) antibodies, CD4-PE cat# 340440, CD3-APC cat# 347327). Further, cells were stained using 1 µL of 1.5 mg/mL DAPI (Sigma-Aldrich, Missouri USA) to identify non-viable cells and enable selection of viable cells. CD4+ T cells were positively selected using flow cytometry using the following gates:

- Viable cells: using DAPI negative cells
- Single cells: cells which do not have another cell bound or attached
- Lymphocytes: cells were assessed for forward and side scatter parameters. Lymphocytes have reduced forward and side scatter due to the shape of their nuclei, which allows differentiation between monocytes, granulocytes and lymphocytes
- CD3+ cells: using the antibody stain
- CD4+ cells: using the antibody stain

Cells were collected into 1.5mL tubes prior to storage at -80˚C. Some of this work was completed by Braydon Meyer.

Modifications to the cell sorting protocol were attempted to optimise yields, and are described below:

- **Use of FBS:** To coat all pipette tips, pre-sort tubes and post sorting collection tubes. The aim was to prevent cell adhesion to the plastic surface of these consumables.
- **Washing of cryovials with media immediately after thawing**: To remove any cells bound to the walls of the cryovial.
- **Mixing of tubes to reduce cells sticking to tubes**: This occurred at any incubation step to again prevent adhesion of cells to the plastic walls. Gentle flicking of the tube suffices.
- **Doubling the sorting volume:** To help prevent loss of cells due to cell-cell adhesion during sorting. Sorting volumes would be ~200 µL compared to 100 µL in previous studies.

3.2.6 Bisulphite Conversion

DNA was bisulphite converted using the MethylEasy Xceed kit (Human Genetic signatures, Australia). DNA in 20 µL had 2.2 µL of 3M NaOH solution mixed, which was then incubated at 37˚C for 15 minutes. The mixture of reagent 1 and 2 was added (220 µL), and after gentle

pipetting was wrapped in aluminium foil and incubated at 80˚C for 45 minutes. Condensation was collected by short centrifuge pulse, 240 µL of reagent 3 was added to the mixture. The entire solution was placed into a purification column and centrifuged for 1 minute at 13,000 *x g* at room temperature. After discarding the flowthrough, 300 µL of reagent 4 was added to the column and centrifuged again as previously done. This step was repeated, and the column dried by centrifuging for 4 minutes at 13,000 *x g*. To elute samples, reagent 5 at 65˚C was placed onto the column and incubated at room temperature for 1 minute. The samples were then incubated at 95˚C for 20 minutes to complete the conversion, prior to storage at -20˚C. Leftover bisulphite converted DNA not used to generate EWAS data was used for technical validation work.

3.2.7 Vitamin D (calcitriol) exposure EWAS.

To develop a probe selection criteria, data was utilised from calcitriol exposure EWAS. Data from this EWAS was useful for developing a selection criteria since it was similar to the future JIA EWAS, as it also investigated immune cells, used HumanMethylation450 BeadChip arrays, and vitamin D was a potential environmental modulator of the methylome with DNAm differences expected to be similar in size as those in other autoimmune diseases [245, 358, 359]. The work in this chapter would contribute to the development of a publication [360].

Briefly, PBMC samples from 2 men and 2 women (aged 26-33) were exposed to calcitriol in culture. Two calcitriol exposure conditions were used, 10 mM or 100 mM in quadruplicates. Previous publications suggested exposures from nanomolar to micromolar scale could affect gene expression and DNAm. So we tested a range of concentrations up to 100 mM and found these 2 exposures affected DNA methyltransferase expression (a proxy for DNAm modulation) (data not shown). Exposure times were 0, 24, 96 and 120 hours, with cells harvested after each exposure time. At each concentration, samples from all exposure times were pooled to reduce the effects of individual blood compositions, as well as enrich for large effects of calcitriol on DNAm.

EWAS data was generated using HumanMethylation450 BeadChip arrays and pre-processed as per section 2.3.3, resulting in a total of 454,563 probes for analysis. Differential DNAm based on calcitriol dose was calculated using a fixed effect linear model in the presence of covariates including time in culture and array batch. Differential DNAm analysis was completed by Dr David Martino.

3.2.8 EpiTyper DNA methylation analysis

Probes of interest were identified for technical validation and replication studies using a selection criteria outlined below (Chapter 3.3.4.2). Primers were designed for 2 genes, *RUNX2* and *POU5F1*. Details of the primer design processes are provided in Chapter 2.3.7. Details of the primers designed are provided in Table 3.1. EpiTYPER analysis of DNA methylation was performed with the method in sections 2.3.8 – 2.3.15. Data was analysed as per the EWAS, where each concentration had time point data pooled, and DNAm differences between exposed and unexposed calculated. Student's T-test was used to assess statistical significance between group means.

Table 3-1 List of primer sequences designed for EpiTyper analysis of the Vitamin D exposure EWAS.

Probe ID	Pos (hg19)	Gene	Chr	Primer	Sequence (5' to 3')	Amplicon length	CpG Units*
cg09445935	31136913	*POU5F1*	6	Forward	GTTTTGGGGTTTTGGAATAAATATTGGTTT	223	4
				Reverse	CCAACCTCTTRGCAAAAAATCACTACTA		
cg21543859	45391645	RUNX2	6	Forward	GYGGGAGTTIGGGTIGTTTTTTTTAGT	211	2
				Reverse	CAACAAAAACCTCTCACCACCTTTCAAAAA		

* Usable CpG units. EpiTyper CpG units may cover more than 1 CpG site. Pos: Position. Chr: Chromosome.

3.3 Results

3.3.1 Optimisation of the cell sorting protocol

3.3.1.1 Introduction

The fundamental basis of this project lay in obtaining high quality data from genome scale DNA methylation microarrays i.e. HumanMethylation450 BeadChip. Furthermore, initial expectations for the project included using microarrays or exome sequencing for gene expression work. Thus, there was a pressing need for a high number of cells to supply sufficient DNA and RNA for these experiments. Previous experience in our group suggested a minimum yield of cells from sorting was ~300,000, which would not suffice for both DNA and RNA experiments (requiring approximately 300,000 cells for each extraction, based on our previous experience).

Hence, it was of great importance to have optimised a cell sorting protocol to maximise the cell yield. Our group had previously published DNAm data using cell sorted CD4+ T-cell DNA [292]. Therefore, a baseline for expected number of cells was known and able to be developed upon.

3.3.1.2 Aims

The aims were to modify the cell sorting protocol to increase yields of CD4+ T-cells. Modification of the cell sorting preparation protocol was attempted to avoid adhesion of cells to plastic consumables and to each other. Reduced adhesion to consumables would allow for more cells to enter cell sorting, increasing the potential yields of CD4+ T-cells.

3.3.1.3 Optimisation of cell yields

Samples from the CLARITY biobank were used and one aliquot (due to limited sample availability) of PBMCs was thawed. A set of 28 samples were sorted and compared to samples sorted using the baseline (old) protocol.

Table 3-2 CD4+ T cell yield improves with modification of the cell sorting protocol.

	Viability	CD4	
	(% average)	**Purity (%)**	**Cell count [10^6 (SD)]**
Old Protocol (n=21)	81	98	0.95 (0.59)
Modified Protocol (n=28)	86	97	1.23 (0.51)

Comparison of cell sorting yield, purity and cells count in sorted CD4+ T-cells between old and modified protocols. Averages for purity and cell count are shown. SD: standard deviation. CD4+ T cells were positive for both CD3 and CD4 markers.

Comparing CD3+CD4+ T cell count and viability between baseline and modified protocols, viability unexpectedly improved on average by 5% (Table 3-2). There was very little effect of the modified protocol on purity (modified protocol: 97% vs 98% in unmodified). Importantly, in accordance with the hypotheses, the modified protocol did increase yields of cells by an average of 280,000. Minor decreases in cell number variability also were observed in the modified protocol (0.59 vs 0.51 in modified protocol samples). The data suggested that an increase in the yield of viable cells during the sorting preparation protocol can be achieved. This was primarily through the use of strategies to prevent cell adhesion to plastic consumables, and potentially to other cells. Work performed here contributed to a publication “DNA methylation at IL32 in juvenile idiopathic arthritis”, published in Scientific Reports in June 2015. This is inserted as appendix I.

3.3.2 Development of methods for extraction of DNA

3.3.2.1 Introduction

EWAS and genome scale expression studies required 1500 ng of DNA (including in-house technical validation work) and 1000 ng of RNA according to service providers (Service XS, The Netherlands, email communication). Kits designed to extract both sample types simultaneously (the AllPrep kit) were previously utilised in our group. However, our experience with this kit using PBMCs suggested RNA and DNA yields would be insufficient. In our prior experience, using 750,000 cells resulted in approximately 78% loss of DNA (22% efficiency, max 1000 ng DNA, 250-300 ng RNA).

Safety and technical benefits abound with the use of the Qiagen AllPrep kit, which enabled the extraction of both DNA and RNA from a single sample without the use of cytotoxic chemicals, in a time efficient and less labour intensive manner. In contrast, using TRIzol for RNA extractions was time consuming and utilises cytotoxic reagents. Furthermore, to extract

DNA, an adjacent and time consuming protocol was required. However, using the more time consuming and technically more difficult DNA extraction protocol, only 300,000 cells were needed to extract DNA at just sufficient quantities and concentrations (~1.7 µg, >25 ng/µL).

Taking these issues into account, it was clear that the Qiagen AllPrep kit - a dual purpose kit - would provide for a safer, easier and more repeatable protocol at the cost of reduced yields of DNA and RNA. Having optimised the yields from cell sorting, it was important to optimise DNA and RNA yields to ensure the availability of sufficient sample for downstream work.

3.3.2.2 Aims and Hypotheses

The aim was to reach DNA yields required for HumanMethylation450 BeadChip methylation arrays of 1500 ng (which include DNA for validation) at >25ng/µL.

The hypotheses of these experiments were:

DNA loss happens post DNA binding, and was lost as either:

- Flow-through at DNA binding steps OR
- Flow-through at RNA binding steps

And that this lost DNA can be recovered by binding the DNA or RNA flowthrough over the DNA column.

To assess this, the following tests were done:

1. Precipitating the flowthrough from the DNA or RNA binding steps
2. Rebinding the DNA or RNA flowthrough over the DNA column

See Figure 3-1 for a schematic of the tests to be employed.

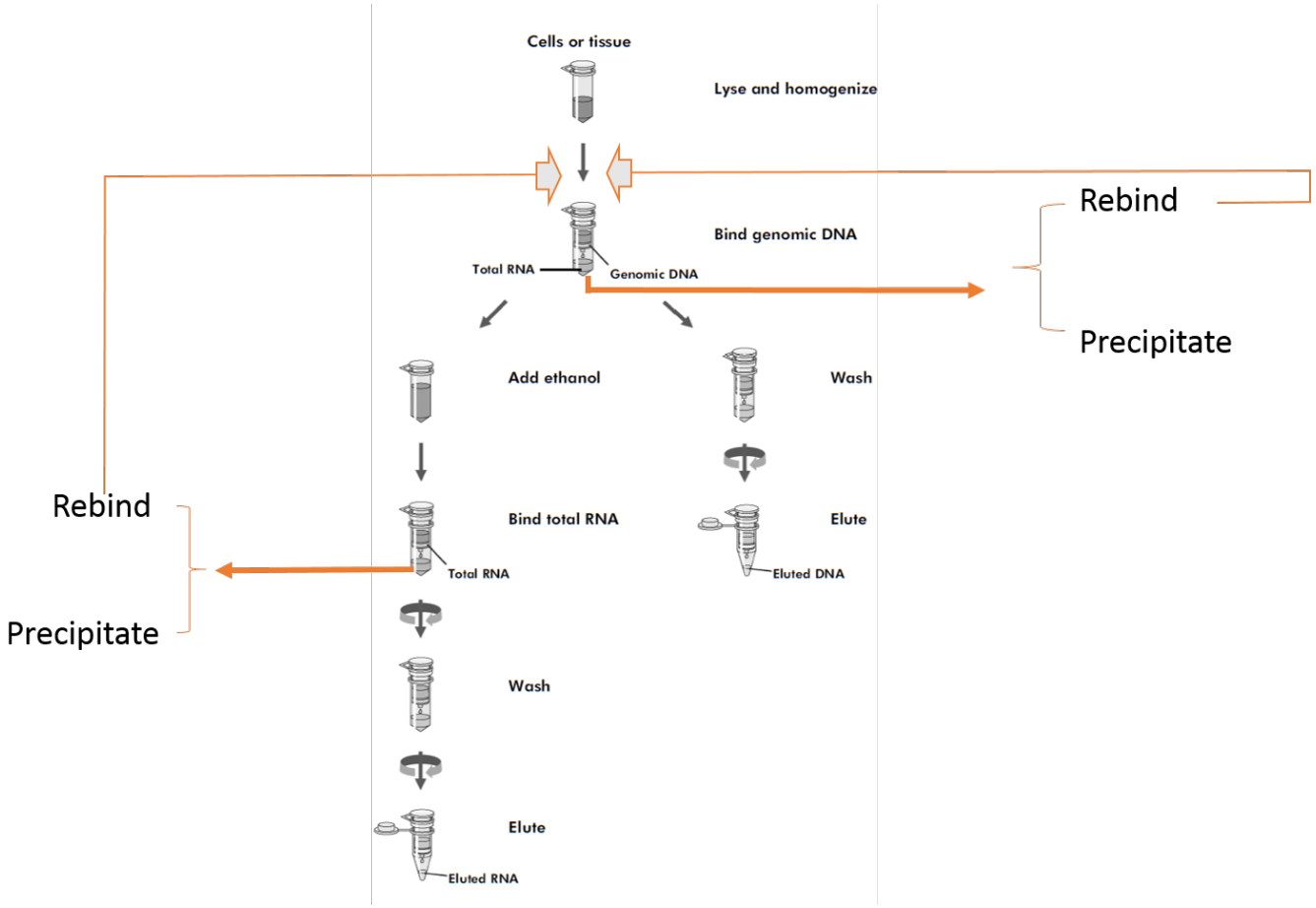

Figure 3-1 Schematic of the AllPrep protocol.

Orange lines and arrows originate from protocol points where DNA/RNA was hypothesised to be lost during use of the kit. These would be points where experiments would be conducted to attempt to obtain higher yields of DNA. Adapted from Qiagen AllPrep DNA/RNA mini handbook November 2005.

3.3.2.3 Dual DNA/RNA extractions using columns do not result in sufficient yields

To assess loss of DNA and attempt to recover it, 500,000 non-sorted PBMCs were used, as this was a number regularly achieved from cell sorting (from previous experience using the JIA cohort for cell CD4+ T-cell DNAm studies [292]). The minimum number of cells obtained from sorting was ~300,000, making the stated number almost double the minimum. DNA yields were compared to the Qiagen Flexigene kit, a method which was long, laborious and only extracts DNA. However, efficiencies of ~ 90% DNA extracted were commonly achieved.

Using the eluate from DNA binding step to rebind over the DNA binding column, yields for the AllPrep kit were slightly improved, with 39% more DNA collected compared to using the standard protocol (1150 ng vs 830 ng of total DNA) Figure 3-2).

Using a modified protocol together with a repeated DNA binding step added no extra value for increasing DNA yields. The efficiency of DNA extracted from both methods were higher compared to standard protocol (38% vs 28%; modified vs standard protocol, respectively), and even increased slightly with the use of extra elution volumes (52% vs 38%, modified vs standard protocol, respectively).

However, the DNA did not meet the required thresholds of total DNA and concentration (i.e. >1500 ng at >25ng/µL). The highest total DNA yields were from the modified + 2nd elution protocol, reaching a maximum quantity of 1546 ng total DNA (Figure 3-2). Despite the total DNA just reaching the requirements, concentrations were below 25 ng/µL. Indeed, no tests returned DNA at the concentrations required (>25 ng/µL).

As total DNA and concentrations were insufficient, RNA was not assessed. The method used for ongoing DNA extractions would be the Qiagen Flexigene kit and a parallel extraction of RNA was required.

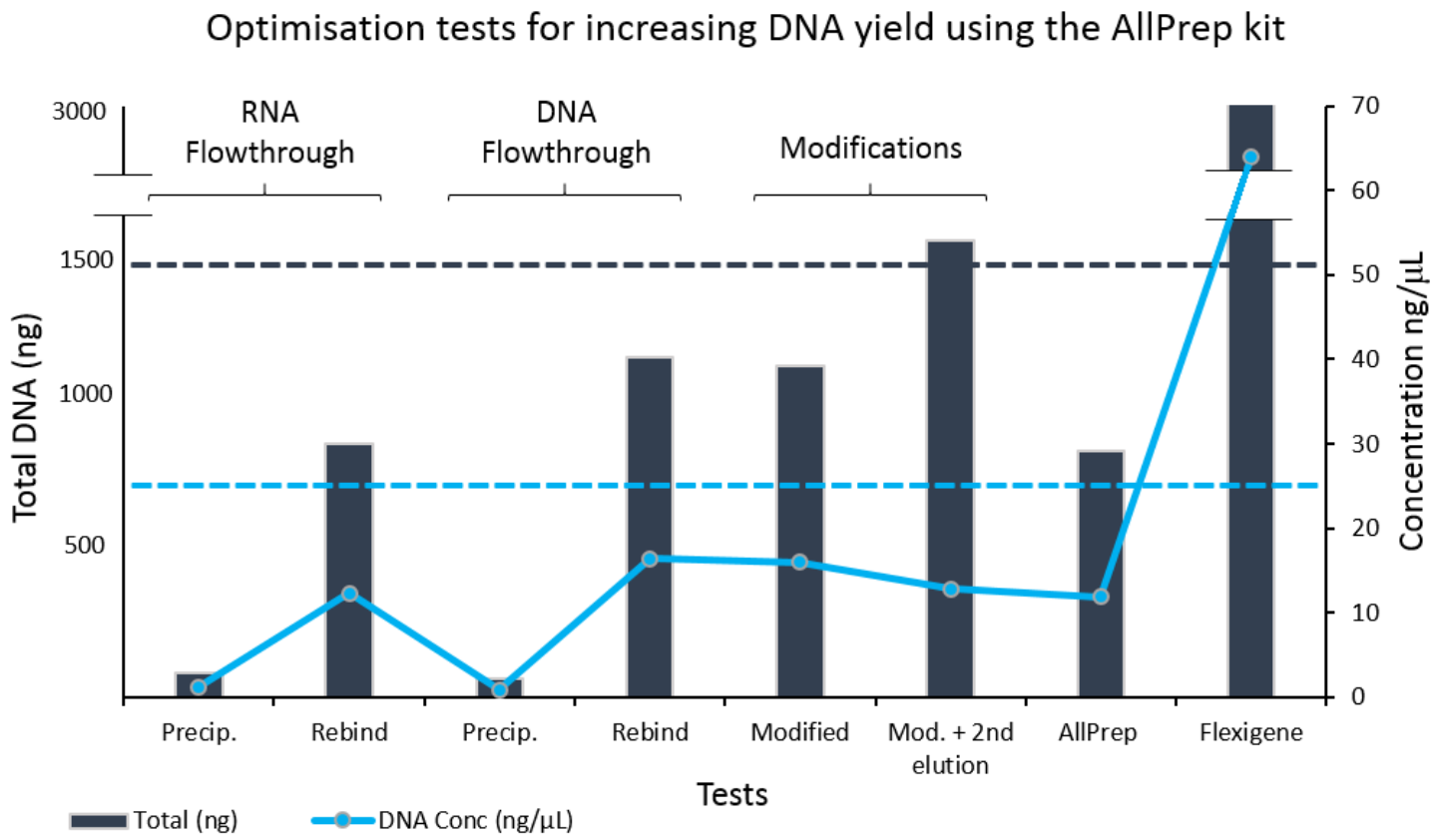

Figure 3-2 DNA extraction optimisation experiments using the AllPrep DNA extraction kit to increase yields.

Tests used either DNA or RNA flowthrough, or through modifications of the protocol. Modified protocols also used rebinding of DNA flowthrough. Dotted dark blue line was the threshold of total DNA that extractions needed to meet. Dotted light blue line was the threshold of DNA concentration that extractions needed to meet. Both thresholds needed to be met in order for DNA to be usable for EWAS study. N=2, averages shown.

3.3.3 Optimisation of RNA extractions

Utilising the aforementioned findings, DNA extractions were hence performed using the manual Flexigene method, therefore requiring a separate RNA extraction protocol. As such, optimisation of the RNA extraction method needed to occur. Similar to manual methods for DNA extractions, issues remained with in-house (TRIzol-based) protocol including: it being highly laborious, time consuming, requiring use of cytotoxic chemicals as well as issues with potential cross-contamination of reagents. In addition, RNA yields using this purified cell type were not known. Therefore, a need existed to define the RNA yields and quality of RNA from TRIzol extractions, and to explore a potential solution to the laborious nature of the TRIzol method for RNA extractions by assessing RNA output and quality.

The Directzol kit indicated faster, simpler RNA extractions, all of which addressed issues experienced in-house using the standard TRIzol RNA extraction protocol. Thus, the TRIzol method and the Directzol kit would be compared in these optimisation experiments.

3.3.3.1 Aims and Hypotheses

The aim was to obtain sufficient RNA yields at a high quality RNA, comparing 2 methods: The Directzol kit from Zymo and the standard TRIzol method. The minimum total RNA needed for gene expression arrays was 300 ng, with an equal amount needed for validation work. High quality was defined as samples with RINe ≥7.

3.3.3.2 RNA quality and yields were increased using a modified TRIzol method

Cell sorted samples would be split for separate DNA and RNA extractions. However, the minimum number of cells obtained from cell sorting was 300,000. This coincided with the minimum number of cells needed to obtain sufficient DNA. Consequently, cells sorted would be split 1:1 (DNA:RNA) only if ~300,000 cells were successfully obtained for each aliquot.

To test whether sufficient RNA could be obtained from approximately 300,000 cells, RNA extractions needed to be tested and optimised. Therefore, experiments utilised 400,000 PBMCs for RNA extractions.

Tests for both the standard TRIzol and Directzol methods resulted in borderline sufficient RNA for microarray gene expression work (366 ng and 385 ng, respectively). The low yield of RNA, and the lack of success in optimising silica column based extractions kits for higher yields (attempted in section 3.2.3 above, AllPrep kit optimisation), suggested the Directzol

kit was inappropriate for these studies. Consequently, the TRIzol method would be used and modified to increase RNA yields for microarray work and validation.

The manual TRIzol protocol was modified using the following changes typically used for increasing yields from this protocol, with the aim of increasing RNA yields compared to the standard TRIzol method.

- 100 µg glycogen added per sample.
- Extra phase separation step.
- 2 hr -80^{o}C isopropanol incubation step.
- 50^{o}C incubation at the elution step.
- DNA removal step using the Thermo DNA-free kit: Required for RNA arrays.

These modifications resulted in large gains in RNA yields, where total RNA more than doubled, as did RNA concentrations (Table 3-3). The quality of RNA was deemed as microarray quality, as assessed by a RINe value of 9. This RINe value was not affected by the protocol modifications and was well over the minimum of RINe 7 typically used as a threshold for samples destined for microarray work.

Therefore, the modified protocol of using TRIzol for RNA extractions was deemed the most appropriate in terms of RNA yield, concentration and quality.

Table 3-3 : Assessing RNA extraction methods for optimal yields and quality.

Test	Concentration (ng/µL)	Total RNA (ng)	RINe
TRIzol (unmodified)	14.6	366	9
Directzol	11	385	9
TRIzol (modified)	32	1120	9

All measures, including RINe, were measured using the TapeStation system.

3.3.4 Preliminary studies to develop a probe selection criteria for technical validation of EWAS data

Whilst protocols for optimal yields from DNA and RNA extraction were defined, the issue of limited DNA sample still remained for technical validation work. That is, it was necessary to develop a probe selection criteria that maximised the potential to validate array findings

while reducing DNA use on unlikely candidates. Compounding the need for an effective selection criteria, JIA is a complex disease where differences were not expected to be large (Δmethylation of 5-10%, compared to cancer studies where Δmethylation of ~30% are typically observed) [245, 356]. The expected smaller differences in DNAm i.e. around 5%, were less likely to be validated due to the technical limitations of the EpiTyper platform [340]. This heightened the need for effectively selecting probes with high likelihood of validation. Consequently, a small EWAS study using immune cells was used to test a selection criteria to validate EWAS data.

In order to select probes for in-house technical validation, genome scale data from vitamin D exposure experiments was used to assess the selection criteria (see for a description of this experiment, and Chavez-Valencia *et al.*[360]). Similar to the JIA EWAS experiments, the genome scale dataset would be from peripheral blood mononuclear cells (PBMCs) utilising a small sample set (n=4). As such, a selection criteria for validation of top hits would be transposable to the JIA EWAS. This data also provides an opportunity to identify shortfalls of the criteria prior to the JIA EWAS itself.

3.3.4.1 Aims and Hypotheses

The aim was to define a selection criteria for probes from genome scale data and validate the array data using the Agena EpiTyper system.

3.3.4.2 Development of a probe selection criteria utilising calcitriol exposure EWAS data

Data from vitamin D exposure studies were utilised for these experiments. Briefly, PBMCs were exposed to 2 different calcitriol concentrations, 10 mM and 100 mM. Calcitriol exposure occurred across four time points, for up to 120 hours in culture. To analyse DNAm through EWAS, DNA from all time points were pooled according to vitamin D exposure concentrations. The differential methylation analysis compared unexposed PBMCs to either pooled 10 mM or pooled 100 mM calcitriol exposed PBMCs.

In order to select probes for technical validation on the EpiTyper system, a selection criteria was arranged.

The selection criteria for probes to take forward to technical validation were:

- p-value of <0.05

- Difference in DNAm (∆beta) of >=5% (average across all time points) of either 10 mM or 100 mM exposure
- Functional relevance to vitamin D or immune system function or disease, demonstrated either through:
    - KEGG pathways
    - Ingenuity pathway analysis
- DNAm differences (∆beta) in the same direction with similar magnitude in both 10 mM and 100 mM concentrations (any p value).
- Further, either one of the following criteria:
    - Genetic association with vitamin D
    - CpGs in the region showing similar ∆beta values (any p-value)

Two probes were selected using this criteria. *POU5F1* (also known as *OCT4*) and *RUNX2*.

Figure 3.4 shows that *POU5F1* had a high ∆beta (11.5%) in 100 mM calcitriol exposure at p-value <0.05. This gene was also a published T1D GWAS hit from 2 separate studies, adding weight to its role in immune system [361, 362]. *POU5F1* was also published as being associated with vitamin D exposure [363]. A ∆beta of ~10% was also observed in the 10 mM exposure samples, and a regional CpG showed a similar direction of effect with a ∆beta approximating 4% in both exposure groups.

Figure 3.3 shows a *RUNX2* significantly associated probe in the 100 mM exposure, with a ∆beta of -7.7% (p<0.05). In the 10 mM exposed samples, this probe had a similar direction of effect with a ∆beta approximating -4.4%. This gene's function was known to be associated with vitamin D [364].

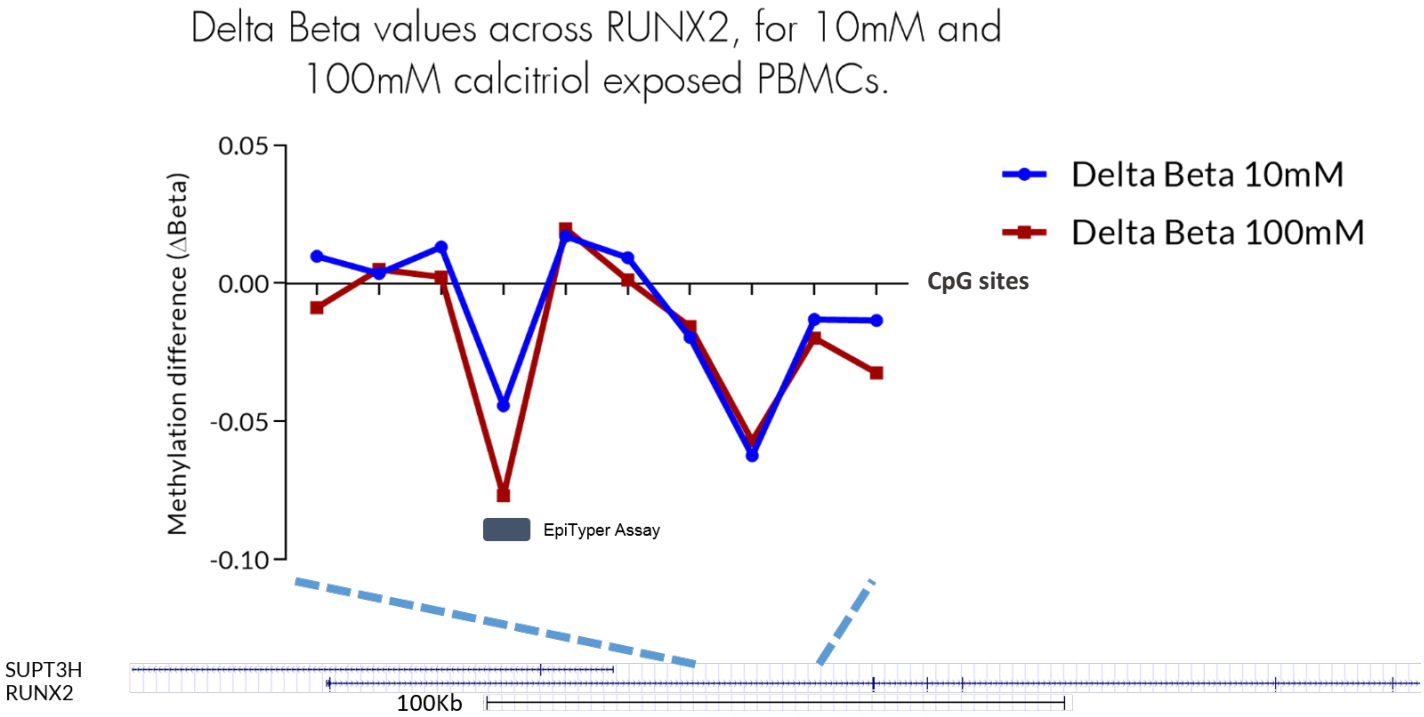

Figure 3-3 EWAS methylation differences in RUNX2 from calcitriol exposed PBMCs, in 10 mM and 100 mM conditions.

RUNX2 had a top ranked probe selected for technical validation via EpiTYPER, with the probe and assay denoted by the dark grey bar. Data from CpGs also interrogated and surrounding the probe of interest (region ~20Kb) are also shown.

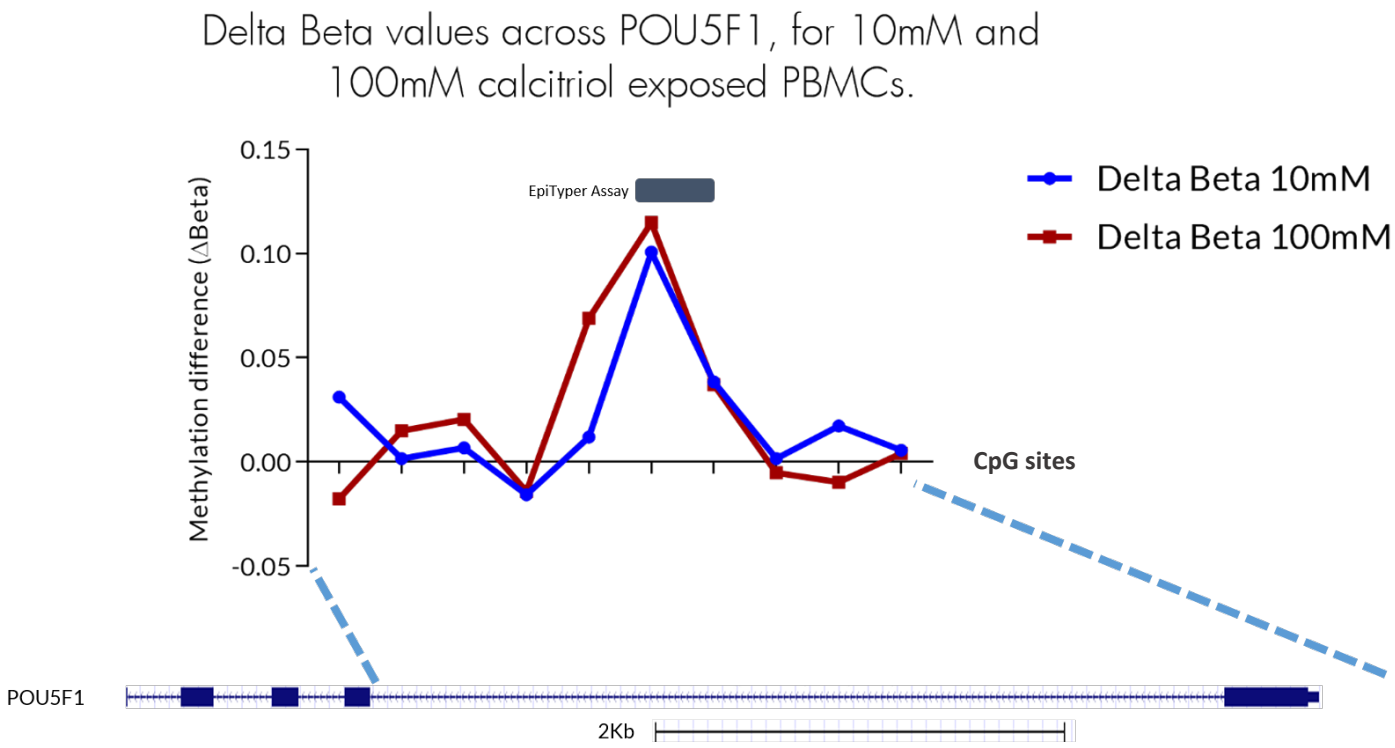

Figure 3-4 EWAS methylation differences in POU5F1 from calcitriol exposed PBMCs, in 10 mM and 100 mM conditions.

POU5F1 had a top ranked probe selected for technical validation via EpiTYPER, with the probe and assay denoted by the dark grey bar. Data from CpGs also interrogated and surrounding the probe of interest (region ~5Kb) are also shown.

3.3.4.3 Testing of Probe selection criteria through loci specific studies of top ranking EWAS probes

Technical validation of the array data was unsuccessful. *POU5F1* EpiTyper data from the same CpG site measured during the EWAS returned no differences in both 10 mM and 100 mM samples (CpG2: ∆beta = 0.0004 and 0.003, p=0.95 and 0.53 respectively). These findings were despite the *POU5F1* EWAS probe having one of the largest ∆beta values within the top significantly associated probes (11.5%). A CpG site near the EWAS probe was also interrogated, but again no or small differences in the opposite direction to the CpG of interest were observed (∆beta=-0.01 and -0.004, p=0.22 and 0.55, respectively) (Figure 3-5a).

Moving to the analyses of *RUNX2*, comparison of array data and EpiTyper data yielded similar results. EpiTyper analysis was unable to validate EWAS findings. The EWAS assayed CpG site, examined in EpiTyper, demonstrated a ∆beta in the opposite direction for both calcitriol exposures, (∆beta = 0.01 and 0.02, p=0.12 and 0.32, respectively) (Figure 3-5b). This was despite *RUNX2* array data suggesting the magnitude of ∆beta values similar in size to *POU5F1*, observed in both calcitriol exposure concentrations (Figure 3-3). A number of proximal CpG sites were also captured in the EpiTyper assay, all of which indicated ∆beta values in the opposite direction (compared to array data) or no DNAm differences (∆beta ranging from 0 to 0.04 across both calcitriol concentrations, with the lowest p-value being no less than p=0.12) (Figure 3-5b).

This work contributed to a publication in May 2014, “In vitro exposure of human blood mononuclear cells to active vitamin D does not induce substantial change to DNA methylation on a genome-scale”, published in the Journal of Steroid Biochemistry and Molecular Biology. It is inserted as appendix II.

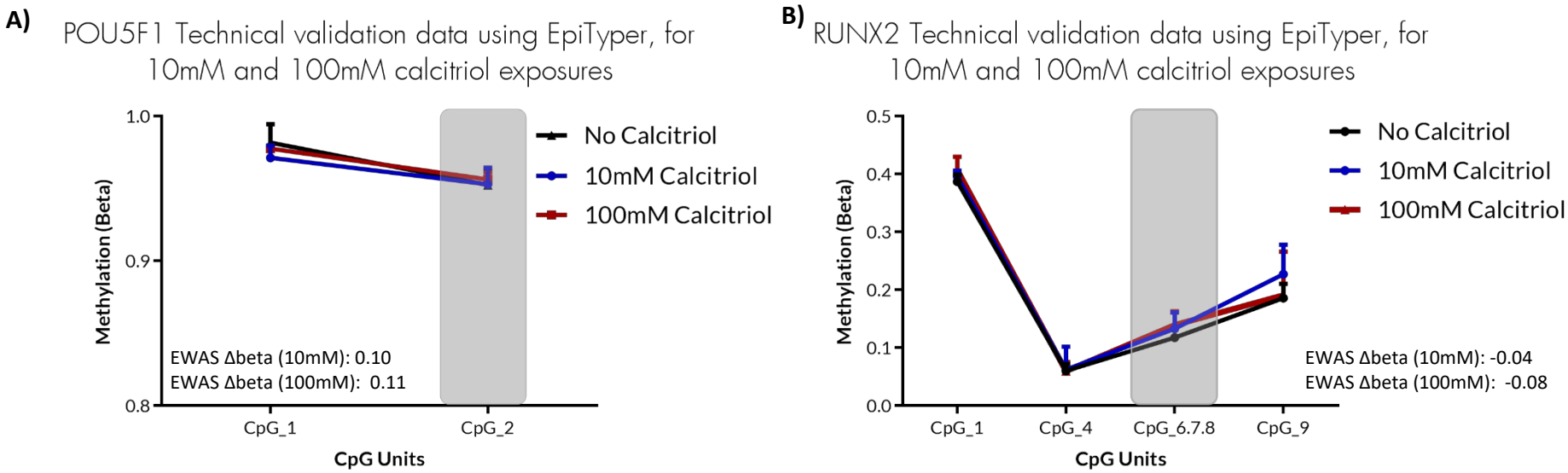

Figure 3-5 Technical validation data from EpiTyper did not validate DNAm differences observed in EWAS in POU5F1 and RUNX2.

a) POU5F1 and b) RUNX2. The EWAS CpG sites are highlighted in grey. EpiTyper was able to capture the EWAS interrogated CpG sites and surrounding CpGs. The RUNX2 EWAS CpG of interest was captured by CpG unit 6.7.8, which interrogated 3 CpGs simultaneously (highlighted by grey bar). EWAS delta beta values are shown, where Δbeta 100 mM or 10 mM represents differences in DNAm between 100 mM or 10 mM exposed and unexposed PBMCs, respectively. Error bars represent standard deviations.

3.4 Discussion

The requirement of sufficient sample for DNA and RNA work meant a need to increase yields of both by optimising protocols. This was particularly pertinent since patient samples were finite and limited volumes of sample from each patient were available for use.

Column based methods for extraction did not provide sufficient DNA and RNA yields. The numbers of cells tested here were nearly double the minimum number of cells that typically resulted from sorting. Yet these extractions, which included protocol modifications, could only output borderline sufficient DNA. Further, concentrations were insufficient, and any concentrating of DNA would result in further loss of DNA. These factors (i.e. the limited number of cells available, insufficient DNA concentrations and totals) meant the AllPrep kit was deemed inappropriate for EWAS purposes.

Alternatively, utilising the Flexigene approach, the efficiency of the manual DNA extraction method was 90%, with sufficient DNA obtainable from 300-400,000 cells. Together, these strongly suggested that separate extractions for RNA and DNA be undertaken. Sorted cells would be evenly split into 2 aliquots, one each for DNA and RNA extractions. Since DNA was the primary resource needed, cells would only be split if 300,000 cells per aliquot could be obtained. It was recognised that there would be a subset of samples that would not have sufficient cells and would thus lower the number of samples available for gene expression work. Hence, optimisation of cell sorting protocols was necessary.

Since the use of two parallel extractions needed to occur, increasing yields from cell sorting was important to reduce the number of samples with insufficient RNA for gene expression work. Optimisation of the cell sorting protocol was successful, providing an extra ~100,000 cells per DNA/RNA extraction. For DNA, this results in a potential extra 600 ng of DNA, which would supply sufficient sample for important technical validation work.

To aid in identification of candidates for technical validation, it was necessary to develop a probe selection criteria. This was of particular importance due to the limited DNA available for technical validation. Further, small beta value differences were expected (~5%) from a complex disease such as JIA, as observed in similar complex diseases or traits such as rheumatoid arthritis and vitamin D deficiency [245, 359]. Small differences in DNAm were less likely to be reliably detected due to the limitations in the EWAS array, particularly in

intermediate DNAm levels i.e. >0.2 and <0.8 [333, 365] [366]. For these reasons, it was important that the selection criteria attempted to reduce false positives being taken into validation studies.

However, the selection criteria did not appear to successfully identify true positives, as both selected probes did not validate using EpiTyper assays. This suggests that either the array or EpiTyper data was not accurate, or there was no real detectable effect of vitamin D exposure, or that the selection criteria was deficient. Whilst further analyses of the vitamin D exposure dataset would suggest that calcitriol does not substantially alter DNA methylation in PBMCs (see appendix II) [360], the selection criteria may have contributed to selecting probes less likely to be technically validated.

One potential shortcoming of the selection criteria was that it did not fully utilise known regional DNA methylation correlation to aid selection of probes to validate. It is known that regional correlation of DNAm reduces rapidly after 2Kb [367]. The selection criteria did not set a window of distances within which to search for correlated probes. Therefore, probes outside of this known correlation window were able to be utilised, which were less likely to have truly correlated with the probe of interest. Utilisation of a defined and recognised regional co-methylation window would be important, as this would utilise nearby probes where DNAm was more likely to have truly correlated, thereby increasing the likelihood of technical validation.

In addition, a regionally correlated probe did not need to be significantly associated, neither did Δbeta values need to cross a threshold (as per the probe of interest). The EpiTyper system was more likely to detect Δbeta values of 5% or greater, as differences below this sit within the known error in this system [340, 341]. Regarding p-values of regional probes, the use of non-significantly associated probes (either in the one of the exposure groups or from regionally correlated probes) may also have selected for false positives. Therefore it was important that in future criteria the Δbeta, as well as p-values of any probes of interest (including regionally correlated probes), meet pre-determined thresholds of significance in order to increase the likelihood of technical replication using the EpiTyper system.

Overall, more stringency could be employed by drawing from the available data to identify probes for technical validation. This could include using both p-value and Δbeta thresholds when identifying regionally correlated probes.

Focussing on the dataset itself will be more important than identifying functional relevance of a probe/gene of interest. DNAm probes of interest may serve as diagnostic or prognostic markers, where the value of the hits lies not in functional relevance but in the ability to differentiate between treatment responders and non-responders, or in predicting patients facing disease progression. As such, prioritising functional relevance for probe selection would not be necessary. The development of a shortlist of potential probes for validation would then inform future exploration of functional relevance.

Accordingly, future technical validation work would use an amended selection criteria. Based on the observations made in this chapter, the criteria would focus more on the dataset itself and less on the functional relevance of those hits.

The selection criteria would be

- P value cut-offs of $p<0.05$.
- Use of regional CpGs (+/-2kb) also within the p-value cut-offs.
- Δbeta thresholds of >5%
    - For both probes of interest and regionally correlated probes.
    - Δbeta in the same direction for all probes in the region of interest.

Chapter 4. **Investigating the DNA Methylation Profiles of Children with Oligoarticular Juvenile Idiopathic Arthritis Compared to Healthy Controls**

4.1 Introduction

JIA is a childhood autoimmune disease. It is known to be a complex disease consisting of genetic and environmental causes [154]. However, little research has been performed to identify the environmental components contributing to disease. DNA methylation (DNAm) is a well-recognised conduit between the genes and the environment. Further, DNAm is known to play important roles in regulating immune cell function and differentiation [253-255, 267]. This suggests it is important to investigate DNAm in autoimmune diseases such as JIA. A number of highly cited and high impact studies revealed strong associations of DNAm with autoimmune diseases, reinforcing the relevance of DNAm in immune function and dysfunction [245, 281, 287, 301].

CD4+ T-cells are known to be important in the pathogenesis of JIA, with a number of studies showing increased cell numbers and ratios of CD4:CD8 T cells, suggesting a selective accumulation of CD4+ T cells. Further immunological data suggests that increases in Th1 and Th17 CD4+ T-cells are also associated with JIA [2, 48, 101, 102, 105, 108]. In addition, the strong genetic association of the HLA class I and II with JIA, which are often present on CD4+ T cells, affirms the importance of this cell type to disease processes [14, 120, 127, 128].

Therefore, considering the important role of CD4+ T-cells to JIA and the potential relevance of DNAm to JIA, this study aimed to investigate the role of CD4+ T-cell DNAm in oligoarticular JIA (oJIA), a common subtype within JIA, using an epigenome wide association study (EWAS).

4.2 Methods

4.2.1 Samples

For DNAm studies, the diagnostic heterogeneity of JIA cases may obscure any potential findings. Additionally, it is recognised that the methylome is inherently variable due to factors such as disease type, cell type, environmental and medication exposure [269, 281, 368-371]. Therefore, it was important that a number of these sources of sample group heterogeneity and potential confounding factors were controlled for in sample selection for EWAS analysis.

To this end, cases were selected using a criterion designed to minimise the heterogeneity of the phenotype in the sample group. Details of the sample group are outlined in Table 4-1 and Table 4-2. This would ideally minimise the sample heterogeneity and variability in the methylome and increase the sensitivity of the analysis to detect differentially methylated probes (DMPs). This selection criteria included subjects that were:

- A single subtype of oJIA
- Unexposed to medications such methotrexate and biological disease modifying anti-rheumatic drugs that may affect DNA methylation
- Active disease at blood sampling

Cases were clinician-diagnosed as oJIA as guided by ILAR (International League of Associations for Rheumatology) criteria, which meant a clinically homogenous sample group was selected for EWAS analysis [8]. The oligoarticular subtype is also the least likely to be exposed to methotrexate (MTX) or biological disease modifying anti-rheumatic drugs (bDMARDs).

Controlling for exposure to medications such as MTX and biological disease modifying anti-rheumatic drugs (bDMARDs) was an important part of the criteria and particularly relevant to methylation studies. MTX directly affects the one-carbon metabolic (folate) pathway by inhibiting key enzymes and producing anti-inflammatory effects [372].

Importantly, the folate pathway is intimately involved in DNAm maintenance and *de novo* DNAm by providing methyl group donors for this process [372]. Inhibition of this pathway is known to lead to decreased DNAm [372]. Previous studies have suggested MTX may indeed affect the methylome in rheumatic diseases [292, 325]. Therefore it was very important to avoid cases exposed to MTX.

In addition, also controlling for bDMARD exposure was important as these may have effects on the methylome. A number of studies suggest that methylation can predict response to bDMARD treatment in rheumatic diseases, suggesting a link between bDMARD exposure and methylation [51, 373, 374]. For example, positive response to bDMARD treatment is associated with *LRPAP1* (Low Density Lipoprotein Receptor-Related Protein Associated Protein 1), and flaring of disease activity after bDMARD withdrawal is associated with T cell activation genes in JIA [51].

Further, bDMARDs target cytokines and their receptors which affect the activation states of immune cells [375]. This may directly affect the methylome since DNAm is known to vary according to the activation state of immune cells, particularly at FOXP3 binding sites in Tregs [260]. Further, upon stimulation by antigen presenting cells, the development of naïve CD4+ T cells to effector Th1 or Th2, as well as Treg development, is characterised by DNAm changes at a number of loci [252-254, 260, 262, 264, 267, 268, 376, 377]. Therefore it was also important to remove bDMARD exposure as a possible confounder in our study by use of the selection criteria. By prioritising the oJIA subtype for investigation, this also meant data from these samples were the least likely to confound results due to medication exposure as well as adding to the clinical heterogeneity of the cases.

Sample selection included only cases with active disease at time of blood sampling. As mentioned earlier, DNAm can vary according to activation state of immune cells. JIA cases with active disease are characterised by an active immune system, favouring inflammation. For example, a number of inflammatory markers are increased during active disease, such as S100 proteins derived from neutrophils [378]. The complement cascade factor C-reactive protein forms is also important in assessing and scoring disease activity levels in JIA [378, 379]. Therefore, the focus on only active cases homogenised the sample group by avoiding differences due to immune system activation levels.

Use of the oligoarticular subtype also allowed the dataset to be partly mitigated for cause versus effect differences. Since DNAm is an environmentally modifiable phenomenon, the potential exists for disease processes to affect the methylome [380, 381]. Therefore, cases with a long history of disease are more likely to display methylomes affected by disease processes. In these cases, significant DNAm associations may potentially result from effects of long term disease. To avoid this, the use of this oJIA subtype as part of the selection criteria meant selected cases were typically sampled very early in disease and sometimes at diagnosis. This reduces the likelihood that any potential DNAm differences seen result from disease. Potentially causative DNAm changes are more likely to be detected in a sample group such as this.

Based on the selection criteria, 57 cases (including 1 twin pair) were selected from the CLARITY project. Extra details on the CLARITY project are provided in Chapter 2.1. Average time from diagnosis to blood collection was 8.5 months, with only 5 cases exceeding 2.5

years post diagnosis. This average is close to the date of a confirmed diagnosis, which occurs 6 months following initial diagnosis.

The male to female ratio was approximately 1:4. This ratio is consistent with the higher number of females diagnosed with oJIA. Controls were selected from the CLARITY biobank. Healthy individuals were recruited from the Royal Children's Hospital from Day surgery clinics, as described in Chapter 2.1. Controls were selected to match for age to within 12 months and were also matched for sex. This matching occurred for all but 5 cases, as noted in Table 4-1 and Table 4-2. DNA samples from 1 case-control pair and another control were used as technical replicates to enable assessment of data reproducibility from the EWAS arrays. The basic characteristics of each 'discovery' case-control pair is shown in Table 4-2.

Matching to age and sex was done in order to control for any confounding effects both variables may have on DNAm differences [380]. The associations of age with DNAm are well established [382, 383], therefore age needs to be controlled for. Early studies reported sex differences in DNAm within repeat sequences such as LINE and Alu sequences in whole blood at Xq28 and 19q13.4 [384], as well as *CALCA*, *MGMT* and *MTHFR* [385]. Genome scale DNAm work later found large numbers of sex related differential DNAm spread across the autosomes [386, 387]. In addition, genetic and gene expression data suggest that sex differences likely exist [388-390]. Therefore, controlling for sex was also important.

The influence of ethnicity in DNAm is a known phenomenon [230-233]. Overall, 63% of the samples were Caucasian, as defined by a relatively strict definition of both parents reporting Caucasian ancestry. There were an equal number in both cases and controls (37). This is the largest ethnic group in Australia, and the largest in this sample set. As most other samples were of mixed ancestry and made up a minority of the overall samples, focussing on Caucasian status thereby enabled control for effects of ethnicity in this study.

Finally, restricting the DNAm analysis to CD4+ T-cells reduced the variability in DNAm inherent to cell type differences. Details of the blood processing and cell sorting for CD4+ T cells are provided in Chapter 2.2 and 3.2.5, respectively. CD4+ T cells are known to be important in JIA pathology, since this cell type is found in joint infiltrates and are selectively accumulated in the joints [101-103]. Further, CD4+ T cells have a high affinity for HLA class II, for which there are number of known genetics associations in JIA [14, 130].

Therefore, the final cohort was selected in order to minimise the heterogeneity in samples, reduce confounding and increase potential for finding early disease markers in JIA using DNAm.

Table 4-1. Samples sourced from the CLARITY cohort for EWAS using the HM450k arrays

Sample ID	Sample Group	Age	Sex	Subtype	ANA	Extended cases with 6m blood[a]	Matched Pairs	Time since diagnosis to blood collection (months)	Matched control status (matched/unmatched)
C00274	CONTROL	3	M	~	~	~	1	~	Matched
J00267	CASE	3	M	persistent	pos	~	1	12.0	Matched
C00178	CONTROL	6	F	~	~	~	2	~	Matched
J00018	CASE	6	F	extended	pos	YES	2	32 months prior to diagnosis	Matched
C00116	CONTROL	2	F	~	~	~	3	~	Matched
J00418	CASE	1	F	persistent	neg	~	3	0.5	Matched
C00723	CONTROL	13	F	~	~	~	4	~	Matched
J00488	CASE	12	F	extended	pos	~	4	1.2	Matched
C00579	CONTROL	4	F	~	~	~	5	~	Matched
J00219	CASE	4	F	extended	pos	YES	5	1.2	Matched
C00542	CONTROL	4	M	~	~	~	6	~	Matched
J00099	CASE	4	M	persistent	pos	~	6	18.8	Matched
C00132	CONTROL	16	F	~	~	~	7	~	Matched
J00492	CASE	16	F	persistent	pos	~	7	0.2	Matched
C00184	CONTROL	2	F	~	~	~	8	~	Matched
J00388	CASE	3	F	extended	pos	YES	8	1.1	Matched
C00120	CONTROL	2	F	~	~	~	9	~	Matched
J00410	CASE	2	F	persistent	neg	~	9	0.2	Matched
C00409	CONTROL	2	F	~	~	~	10	~	Matched

Table 4-1 Continued. Samples sourced from the CLARITY cohort for EWAS using the HM450k arrays

J00242	CASE	1	F	persistent	pos	~	10	0.2	Matched
C00209	CONTROL	3	F	~	~	~	11	~	Matched
J00462	CASE	2	F	persistent	pos	~	11	2.7	Matched
C00576	CONTROL	12	F	~	~	~	12	~	Matched
J00360	CASE	12	F	persistent	neg	~	12	1.1	Matched
C00385	CONTROL	8	M	~	~	~	13	~	Matched
J00180	CASE	7	M	persistent	pos	~	13	1.2	Matched
C00413	CONTROL	3	F	~	~	~	14	~	Matched
J00329	CASE	4	F	extended	NA	YES	14	2.3	Matched
C00306	CONTROL	10	F	~	~	~	15	~	Matched
J00181	CASE	10	F	persistent	pos	~	15	0.3	Matched
C00035	CONTROL	2	F	~	~	~	16	~	Matched
J00343	CASE	2	F	persistent	pos	~	16	0.9	Matched
C00370	CONTROL	9	M	~	~	~	17	~	Matched
J00346	CASE	9	M	extended	pos	YES	17	0.4	Matched
C00233	CONTROL	5	F	~	~	~	18	~	Matched
J00464	CASE	5	F	persistent	NA	~	18	0.2	Matched
C00402	CONTROL	2	F	~	~	~	19	~	Matched
J00486	CASE	2	F	persistent	pos	~	19	1.2	Matched
C00312	CONTROL	8	F	~	~	~	20	~	Matched
J00155	CASE	8	F	persistent	pos	~	20	30.0	Matched
C00555	CONTROL	7	F	~	~	~	21	~	Matched

Table 4.1 Continued. Samples sourced from the CLARITY cohort for EWAS using the HM450k arrays

J00399	CASE	8	F	persistent	neg	~	21	10 months prior to diagnosis	Matched
C00433	CONTROL	3	M	~	~	~	22	~	Matched
J00401	CASE	2	M	persistent	pos	~	22	0.3	Matched
C00407	CONTROL	3	F	~	~	~	23	~	Matched
J00387	CASE	3	F	persistent	pos	~	23	0.1	Matched
C00484	CONTROL	7	F	~	~	~	24	~	Matched
J00111	CASE	7	F	extended	pos	NO	24	44.6	Matched
C00334	CONTROL	10	F	~	~	~	25	~	Matched
J00008	CASE	11	F	extended	pos	NO	25	10.0	Matched
C00414	CONTROL	3	F	~	~	~	26	~	Matched
J00110	CASE	4	F	persistent	pos	~	26	24.3	Matched
C00536	CONTROL	9	M	~	~	~	27	~	Matched
J00390	CASE	9	M	persistent	pos	~	27	1.1	Matched
C00380	CONTROL	10	F	~	~	~	28	~	Matched
J00365	CASE	11	F	persistent	pos	~	28	0.0	Matched
C00275	CONTROL	5	F	~	~	~	29	~	Matched
J00100	CASE	5	F	extended	neg	NO	29	51.8	Matched
C00264	CONTROL	5	F	~	~	~	30	~	Matched
J00136	CASE	6	F	persistent	neg	YES	30	2.5	Matched
C00188	CONTROL	4	F	~	~	~	31	~	Matched
J00417	CASE	4	F	persistent	NA	~	31	0.9	Matched
C00157	CONTROL	2	M	~	~	~	32	~	Matched

Table 4.1 Continued. Samples sourced from the CLARITY cohort for EWAS using the HM450k arrays

J00321	CASE	2	M	persistent	pos	~	32	0.5	Matched
C00581	CONTROL	12	F	~	~	~	33	~	Matched
J00397	CASE	11	F	extended	NA	NO	33	60.0	Matched
C00553	CONTROL	4	F	~	~	~	34	~	Matched
J00467	CASE	4	F	persistent	NA	~	34	0.4	Matched
C00540	CONTROL	3	F	~	~	~	35	~	Matched
J00395	CASE	2	F	persistent	pos	~	35	1.9	Matched
C00322	CONTROL	8	F	~	~	~	36	~	Matched
J00443	CASE	8	F	persistent	neg	~	36	0.9	Matched
C00383	CONTROL	14	F	~	~	~	37	~	Matched
J00196	CASE	13	F	persistent	pos	~	37	0.1	Matched
C00197	CONTROL	2	M	~	~	~	38	~	Matched
J00460	CASE	3	M	persistent	neg	~	38	1 month prior to diagnosis	Matched
C00242	CONTROL	4	F	~	~	~	39	~	Matched
J00116	CASE	5	F	persistent	pos	~	39	24.0	Matched
C00261	CONTROL	9	F	~	~	~	40	~	Matched
J00296	CASE	8	F	persistent	pos	~	40	2.1	Matched
C00293	CONTROL	12	M	~	~	~	41	~	Matched
J00172	CASE	12	M	persistent	neg	YES	41	0.4	Matched
C00174	CONTROL	3	F	~	~	~	42	~	Matched
J00167	CASE	2	F	persistent	pos	~	42	0.0	Matched
C00728	CONTROL	13	F	~	~	~	43	~	Matched

Table 4-1 Continued. Samples sourced from the CLARITY cohort for EWAS using the HM450k arrays

J00133	CASE	12	F	extended	pos	NO	43	123.7	Matched
C00458	CONTROL	4	F	~	~	~	44	~	Matched
J00425	CASE	3	F	persistent	pos	~	44	1.2	Matched
C00295	CONTROL	6	F	~	~	~	45	~	Matched
J00164	CASE	7	F	extended	pos	YES	45	0.9	Matched
C00153	CONTROL	2	F	~	~	~	46	~	Matched
J00493	CASE	2	F	persistent	pos	~	46	1.2	Matched
C00440	CONTROL	3	F	~	~	~	47	~	Matched
J00412	CASE	4	F	persistent	pos	~	47	0.3	Matched
C00123	CONTROL	11	M	~	~	~	48	~	Matched
J00182	CASE	11	M	extended	neg	YES	48	0.1	Matched
C00366	CONTROL	12	F	~	~	~	49	~	Matched
J00004	CASE	11	F	persistent	pos	~	49	24.9	Matched
C00395	CONTROL	13	F	~	~	~	50	~	Matched
J00319	CASE	12	F	persistent	pos	~	50	0.0	Matched
C00466	CONTROL	11	M	~	~	~	51	~	Matched
J00372	CASE	13	M	persistent	pos	~	51	1.2	Matched
C00589	CONTROL	2	F	~	~	~	52	~	Matched
J00123	CASE	2	F	persistent	pos	~	52	2.8	Matched
C00133	CONTROL	3	M	~	~	~	~	~	Unmatched
C00170	CONTROL	2	M	~	~	~	~	~	Unmatched
C00228	CONTROL	3	M	~	~	~	~	~	Unmatched

Table 4-1 Continued. Samples sourced from the CLARITY cohort for EWAS using the HM450k arrays

C00582	CONTROL	13	F	~	~	~	~	~	Unmatched
J00132	CASE	2	F	extended	pos	YES	~	1.7	Unmatched
J00271	CASE	3	F	extended	pos	YES	~	5.0	Unmatched
J00369	CASE	3	F	persistent	NA	~	~	0.9	Unmatched
J00475	CASE	2	F	persistent	pos	~	~	1.2	Unmatched
C00310	CONTROL	12	F	~	~	~	~	~	Unmatched
J00020	CASE	5	M	extended	pos	NO	30	14.7	Unmatched

52 case control pairs are included, matched for age and sex. Samples that are not matched occurred due to sample replacement at service provider sample QC step. Replacement samples could not be matched for age and sex. ANA: Anti-Nuclear Antibody status (pos: positive, neg: negative).
[a]Cases diagnosed with extended disease which had a blood sample taken prior to diagnosis of extended disease.

Table 4-2 Sample group summary table for Discovery studies.

	Discovery		
	Case	**Control**	**Matched pairs[a]**
All Inclusive	56[b]	57	52
Mean age (SD)	6.1 (4.0)	6.4 (4.2)	
Female/Male	45/11	43/14	41/11
≤6yrs/>6yrs at diagnosis[c]	35/21	33/24	31/21

[a] Matched by age and sex. [b] Does not include one case removed due to inactive disease at time of blood collection. [c] Number of cases in each age-at-diagnosis category/number of controls in same category for age at recruitment. Mean age = age at blood collection

4.2.2 Statistical analyses – Genome scale methylation data

Genome scale DNA methylation data was generated using the Illumina Infinium HumanMethylation450 (HM450k) Beadchip kit (Illumina, San Diego USA). Details of the generation of this data as well as the quality control and data processing are provided in Chapter 2.3.3-2.3.5.

4.2.2.1 Differential DNA methylation analysis: LIMMA

The LIMMA package was used to identify differentially methylated probes (DMPs) [355]. This was done within the R statistical programming language. Using the M values of the dataset, a linear model was fitted and a moderated t-test was employed. Using the MDSplot function, clustering of samples was visualized. When probes best clustered samples according to case/control status (unadj.p<0.001, 1421 top ranking probes), these probes were retained. Beta values were obtained by logit transforming M values. Delta beta values for each probe were calculated by taking the average beta $value_{cases}$ – average beta $value_{controls}$.

4.2.2.2 Differential DNA methylation analysis: RUV

RUV was used for DMP analysis. For details on this method, see Chapter 2.3.5. Briefly, RUV is a method for removal of unwanted variation. It uses bottom ranked probes from an initial DMP analysis to identify covariates not related to the factor of interest (i.e. oJIA). These are then used as covariates in a subsequent DMP analyses.

Specific to this analysis, three iterations of the RUV DMP analysis were completed using the bottom 90% of ranked probes as empirical control probes (ECPs). P values were adjusted for multiple testing using the Benjamini-Hochberg method [391]. MDS plots were used to identify case control clusters using top ranking probes. Using top ranked probes with unadjusted p<0.01, case/control groups formed separate clusters and these probes (4166 probes) were the basis for further work. Beta values and delta beta values were calculated as above (4.2.2.1)

4.2.3 Functional analyses

4.2.3.1 Pathway analysis

To analyse biological pathways enriched in the EWAS dataset, Ingenuity pathway analysis (Qiagen, Hilden, Germany) was used. Genes associated with the 4166 RUV top ranked probes (with unadj.p<0.01) were used as 'data set' files for a 'Core analysis'. All variables

were left on default settings during analysis setup. For pathway analysis, the background reference genome was set to the ingenuity knowledge base. Following this analysis, 'Networks', 'Functions' and 'Canonical pathways' were reviewed and enrichment values plotted.

4.2.3.2 Gene ontology analysis

Gene ontology analysis was carried out using DAVID i.e. **D**atabase for **A**nnotation, **V**isualization and **I**ntegrated **D**iscovery (**DAVID**) Bioinformatics Resource (National Institute of Allergy and Infectious Diseases (NIAID), NIH; [392, 393]). Genes derived from the top ranked RUV DMPs (unadj.p<0.01, 4166 probes) were used as the input. Genes ID's were then converted to 'OFFICIAL GENE SYMBOL' using DAVID. The complete human genome was used as the reference genome and default settings used for analysis. Review of enriched gene ontologies was done using the 'Functional Annotation clustering'.

To visualise groupings of gene ontologies, the online tool REViGO (**Re**duce and **vi**sualise **g**ene **o**ntology) was used [394]. This tool reduces the dataset by gathering related ontologies under a single heading, then uses clustering algorithms to visualise ontologies. Ontologies that are more highly related will cluster together. Gene ontologies from DAVID Functional Annotation Clustering tool were used as input for this tool. Default settings were used, and the output cluster map from 'Molecular functions' was used.

4.2.4 EpiTYPER DNA methylation analysis

4.2.4.1 Assay design

Probes of interest were identified for technical validation and replication studies using a selection criteria outline below (Chapter 4.3.4). Primers were designed for 5 genes, *CHD5*, *CYP1A1*, *TACSTD2*, *SYNE1* and *VPS53* (listed in Table 4-3). Details of primer design are provided in Chapter 2.3.7. Details of the primers designed are provided in Table 4-3.

4.2.4.2 Polymerase Chain Reaction

PCR (polymerase chain reaction) was completed to amplify DNA fragments to sufficient quantities to allow for detection via Mass Spectrometry. This was completed as part the MassArray EpiTYPER system for loci specific DNA methylation analysis (Agena Bioscience, San Diego, USA). Up to 500ng of bisulphite converted genomic DNA was used for PCR reactions. Details of the bisulphite conversion process are given in Chapter 2.3.2. Ten nanograms (10ng) of bisulphite converted DNA per sample was used, and every sample was

run in triplicates for each assay. PCR was completed using FastStart mastermix (Roche, Basel, Switzerland) based on the protocol in Coolen *et al.* 2007 [341]. Table 4-4 outlines the thermocycling conditions used. To check for amplification, samples were visualized on an agarose gel (Chapter 2.3.9).

Table 4-3 List of primer sequences designed for EpiTYPER analysis of the JIA EWAS study.

Probe ID	Pos (hg19)	Gene	Chr	Primer	Sequence (5' to 3')	Amplicon length	CpG Units*
cg12101586	75019203	*CYP1A1*	15	Forward	AGG AAG AGA GGR GAG GAG AGG TAG TTT GTA TGT GTT	263	12
				Reverse	CAG TAA TAC GAC TCA CTA TAG GGA GAA GGC TAA CCT AAA CCA ACR ACR AAT AAT AAT ACC AC		
cg15205435	6187920	*CHD5*	1	Forward	AGG AAG AGA GAG TTG GGT AGG GAT TGA AAT TAG AT	363	14
				Reverse	CAG TAA TAC GAC TCA CTA TAG GGA GAA GGC TAA CCA ACC CTC CAA ACT ATA ACC		
cg17210938	59043173	*TACSTD2*	1	Forward	AGG AAG AGA GTA TAT AAA YGT GTG TTA TTG TGA GAA TTG GAT AAA G	253	10
				Reverse	CAG TAA TAC GAC TCA CTA TAG GGA GAA GGC TAA AAA AAC CRA TAC TCA CCT ATA AAC TTC TAA AAA CC		
cg04370829	406249	*VPS53*	17	Forward	AGG AAG AGA GGG GTA ATA GTA GTA TAG GTT ATA TTG AGT TAT TAG TGT	243	9
				Reverse	CAG TAA TAC GAC TCA CTA TAG GGA GAA GGC TRA AAA CCR AAA ACC ATT CAC CCC TA		
cg27316956	152958698	*SYNE1*	6	Forward	AGG AAG AGA GTT TGA GGG AGA TTG GTT AGG TTT TAT AG	341	3
				Reverse	CAG TAA TAC GAC TCA CTA TAG GGA GAA GGC TTA ACT ACA ATC ATC CTA TTA AAC TTA AAC TTC AC		

* EpiTYPER CpG units may cover more than 1 CpG site. Pos: Position. Chr: Chromosome

Table 4-4 PCR thermocycling conditions for assays CHD5, CYP1A1, TACSTD2, SYNE1 and VPS53.

Temperature (˚C)	Time	Cycles
95	10 min	5
95	10 sec	
58	30 sec	
72	1 min 30 sec	
95	10 sec	40
62	30 sec	
72	1 min 30 sec	
72	7 min	1
4	Hold	1

4.3 Results

4.3.1 Preprocessing of EWAS data

4.3.1.1 Data quality control suggests high quality data

To assess the DNAm profiles of oJIA cases compared to controls, genome scale methylation data was generated using the Illumina Infinium HumanMethylation450 beadchip arrays, and assessed methylation at over 485,000 CpG sites across the genome, covering 99% of RefSeq genes. For more details on generation of EWAS data, please refer to Chapter 2.3.3.

Quality control was performed using the Minfi package in R studio. To assess sample quality, the signal (above background levels) for each probe was analysed and p-values generated. Low p-values indicate significant differences compared to background signal. For each sample, p values from all probes were averaged and plotted in Figure 4-1A. Samples with p-values above 0.01 indicate low quality samples. The very large majority of samples were at least 1 order of magnitude below the cut-off, if not 1.5 orders of magnitude below. This suggested that all samples were usable for further data analysis.

To identify samples with technical or quality problems affecting overall DNAm distributions, the DNAm distributions of all samples was visualised. DNAm across the genome has a distinctive bimodal distribution with peaks near 0% and 100% methylation. This can be visualised using a density plot. As can be seen in Figure 4-1B, distributions from this EWAS dataset indicate that DNAm in all samples does indeed follow this bimodal distribution. These plots enable one to visualise any samples that may need to be flagged for investigation or removal due to quality issues e.g. if peaks are highly skewed, or have highly reduced densities at expected peak locations. This issue did not arise in this dataset. For any of the minor distribution shifts present in the plots, these are likely to be due to probe type differences, background or colour differences and will most likely be adjusted after further preprocessing steps.

To help assess the technical performance of the EWAS assay and gain insight into the effects of any potential technical errors, three technical replicates were used in the EWAS assay. Figure 4-2 shows an example of the correlation between technical replicates on the chips. Visually, these replicate samples follow a highly linear positive relationship and this is reflected in the correlation coefficient of 0.9971. Across 3 technical replicates throughout

the EWAS data, this correlation was observed to average 0.9971, which was higher compared to the average all pairwise correlations of biological replicates of 0.9937 (min 0.984 – max 0.996), excluding a twin pair and technical replicates. This high correlation is present in technical replicates despite positioning samples on different chips (thus adding the technical differences due to chips on top of assay preparation errors).

With no samples failing QC, consistent DNAm distributions following the expected bimodal spread and the high accuracy of data observed through high correlations, review of the dataset's sample quality, technical performance and data quality all suggested the data was of high quality. This gave sufficient confidence to proceed to data preprocessing prior to differential methylation analysis.

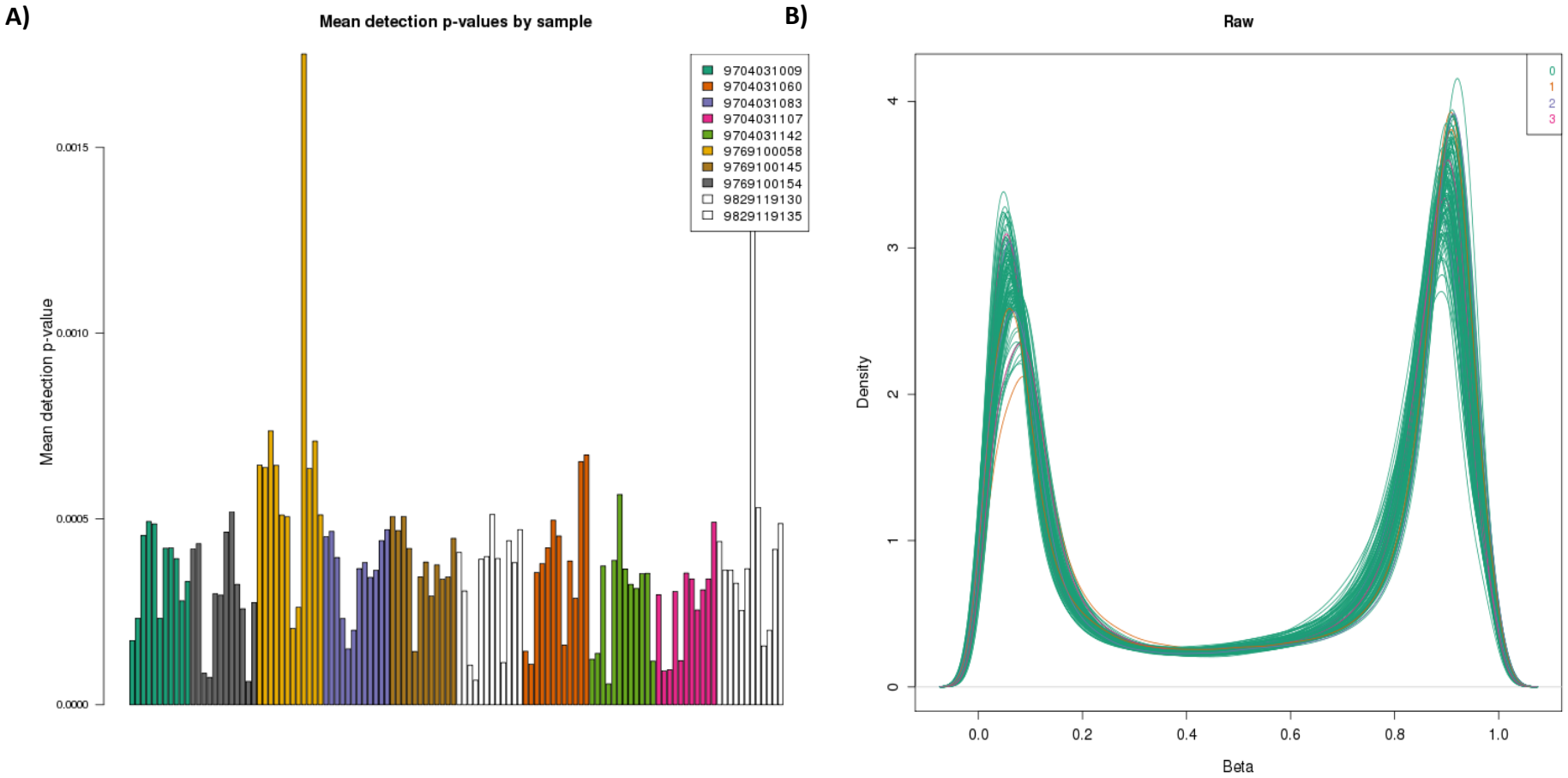

Figure 4-1 Preprocessing quality control across chips and individual samples.

A-B. (A) Average probe detection p values plotted for all samples. Each probe signal was tested for differences compared to negative control probes in each sample. P-values across all probes were then averaged for each sample. Samples are coloured according to chip, and each bar represents an individual samples on each chip (B) Density plot of beta values across all samples. Technical replicates, numbered 1-3, are coloured other than green

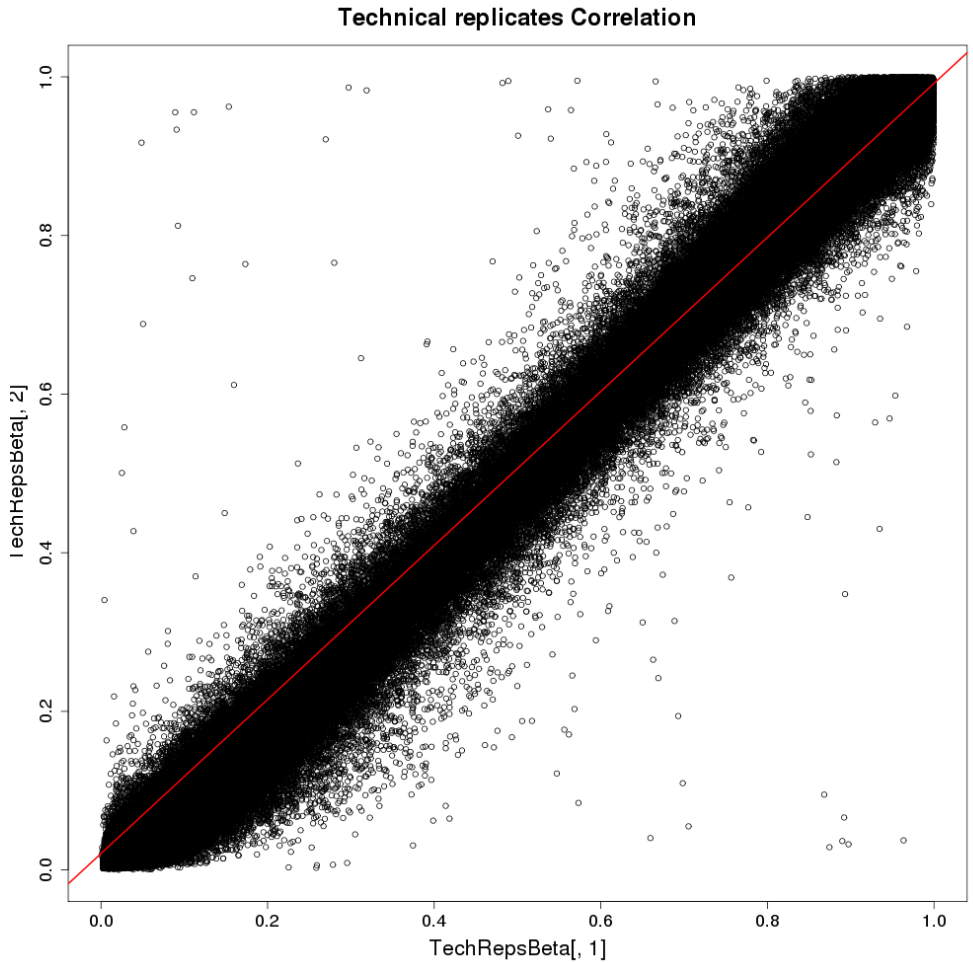

Figure 4-2 Scatter plot of all probes from technical replicate of sample ID J00123.

Beta values from technical replicates were plotted, and correlation calculated to be 0.9

4.3.1.2 Probe type normalisation: less manipulative methods are required

Confident in the quality of the data, pre-processing of the data was performed. Technical errors exist in the Infinium assay due to intensity differences from the 2 colours of fluorescent probes used, as well as the background signal. Therefore the Minfi package was used to correct for these factors. Chapter 2.3.3.1 details the steps taken. Corrected data (CorrData) was then used for further work.

As noted in Chapter 2.3.3.1, two different probes types are used to interrogate DNAm in this array. This leads to differences in DNAm distributions (appearing as small bumps at the tails of density plots of M values) that requires data adjustment (Figure 4-4E) [333]. To adjust for probe type differences, normalisation of values between these 2 probe types was implemented using the R package Minfi [334]. Within Minfi, there are 3 probe type normalisation methods, as there is no currently accepted gold standard method. Therefore, these three options were applied separately to identify a method which manipulated data the least whilst still adjusting for probe type differences. To identify an appropriate method, technical replicates were again used as a point of reference. Density and MDS plots were used to assess proximity of technical replicate samples and visualise the extent of dataset manipulation from the different methods. Ideally, the proximity of technical replicates should remain similar or improve after adjustment. In addition, density and MDS plots would ideally show clusters or distributions similar to CorrData i.e. where data is less manipulated whilst still normalising for probe type differences.

Raw data showed that no samples that have obviously distorted beta value distributions (Figure 4-1B). The high quality of the data therefore afforded itself to a probe type normalisation technique, one that would be less manipulative and disruptive of the dataset compared to techniques that visibly changed sample distributions in MDS and DNAm density plots.

Looking at MDS plots of CorrData compared to subset quantile and functional normalisation methods (Figure 4-3A-C respectively), both subset quantile and functional normalisation appear to bring technical replicates closer to each other compared to CorrData. Yet both methods also appear to change the overall appearance of the MDS plot (Figure 4-3B & C), where clusters and relationships between samples appear highly adjusted (to narrow and highly-defined clusters) compared to clusters in CorrData (Figure 4-3A). This suggests that these normalisation methods highly manipulate the data compared to the background and

colour-adjusted data, potentially removing biological differences as well as technical differences due to probe types. Focussing on the subset quantile within array normalisation (SWAN) method, the high amount of manipulation is also reflected in density plots. Compared to CorrData, the quantile normalised data shows a narrow and homogenous distribution across all samples (Figure 4-4A & B). Functional normalisation also manipulates data heavily, observable in density plots as variation in density at the 0.6-0.8 beta value range compared to CorrData (Figure 4-4A & C). In contrast to those methods, SWAN probe type normalisation (Figure 4-4D) appeared to affect DNAm distributions so that density plots show an increased consistency of peak heights across samples at both ends of the spectrum. Further, SWAN appeared to maintain a similar clustering pattern in MDS plots compared to CorrData (Figure 4-3D & A), unlike previous normalisation methods. In addition, technical replicates maintained proximity in MDS plots, similar to those seen in CorrData (Figure 4.4A & D), following the prescription for an ideal method outlined above.

To assess whether probe type differences had been removed, beta values were converted to M values (see chapter 2.3.3.1 for details), and density plots produced of CorrData vs SWAN adjusted data.

As mentioned earlier, unadjusted data contains characteristic small bumps at the tails of the M value distribution due to probe type differences, which SWAN does indeed remove (Figure 4-4E-F). This suggests that data manipulation is reduced whilst normalising samples for probe type differences. Over manipulation of data may lead to the unintended loss of biological differences together with the intended probe type normalisation. Considering the proximity of technical replicates and the reduced data manipulation needed to adjust for probe type differences, SWAN was chosen as a normalisation method.

Having fully considered potential sources of technical variation within the dataset, it was further refined by removing non-informative probes. These methods are detailed in chapter 2.3.3.1, but briefly, this involved removal of X and Y chromosome probes, probe SNPs with MAF>0.05 and probes known to bind non-specifically. Exploration of the correlations and compositions in the data were then carried out. This left a final dataset of 430,191 probes for further analysis.

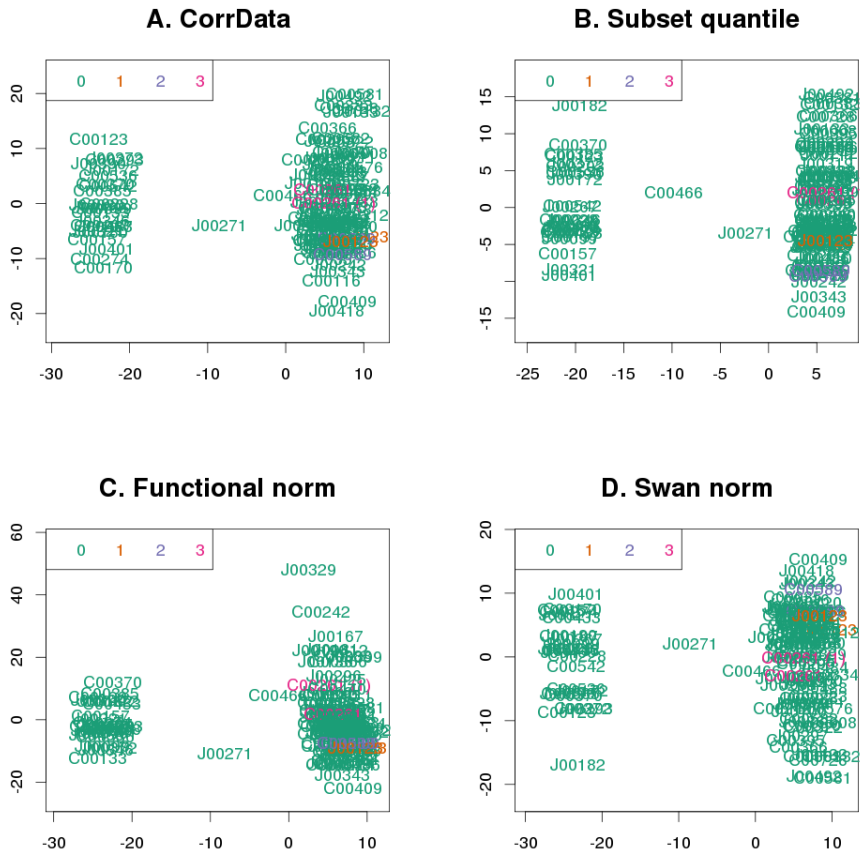

Figure 4-3 MDS plots of samples adjusted with different methods for probe type normalisation.

Three technical replicates are coloured differently for better visualisation, numbered 1, 2 and 3 in the legend. Samples labelled 0 in the legend are all biological replicates. A) Data post background and colour correction, CorrData B) Subset Quantile Normalisation C) Functional Normalisation D) SWAN normalisation. Technical replicates are visualised using different colours in all plots.

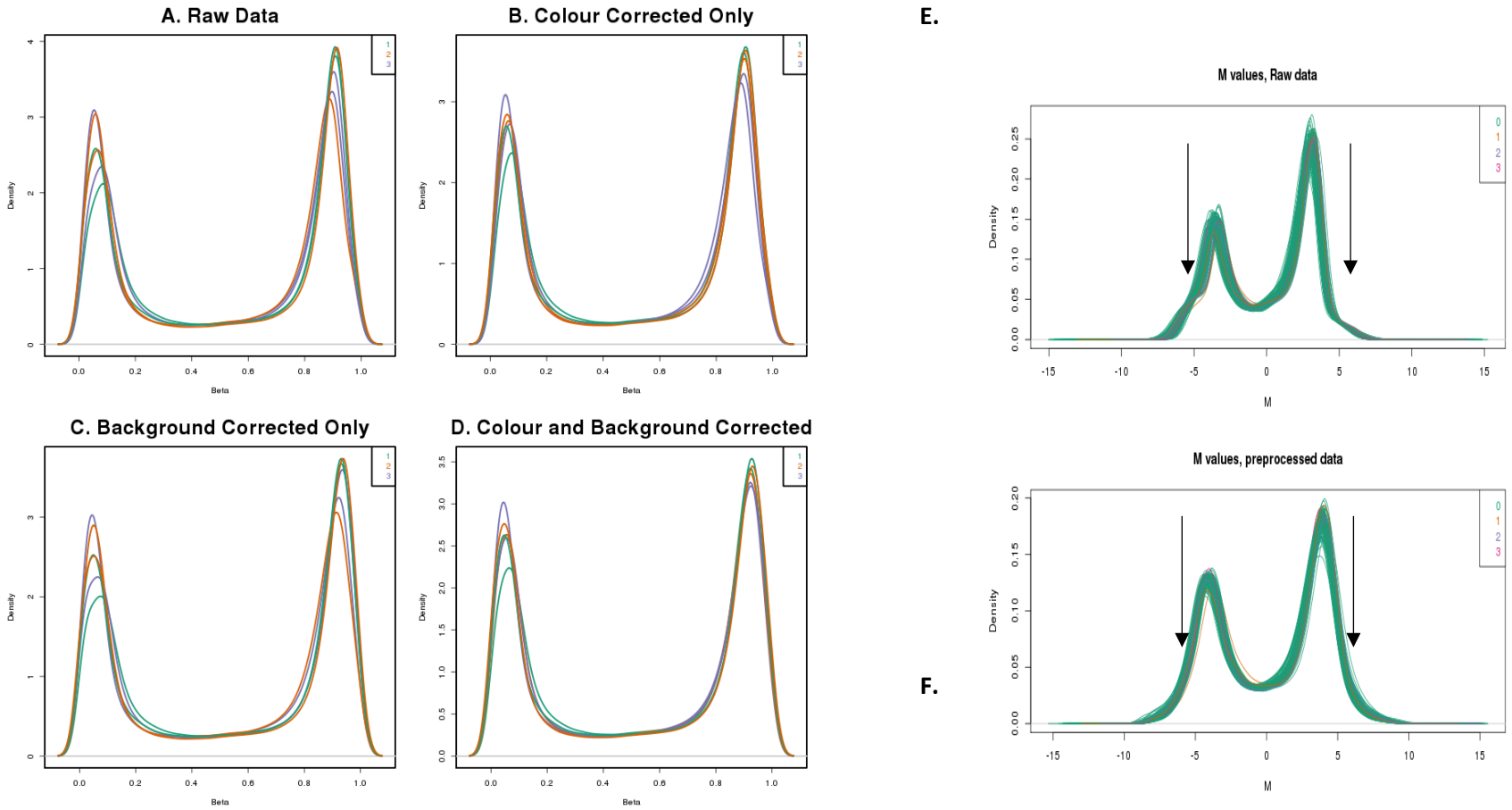

Figure 4-4 Density plots of samples adjusted with different methods for probe type normalisation.

Technical replicates (x3) are coloured differently for better visualisation. A) CorrData (background and colour adjusted B) Subset Quantile normalisation C) Functional Normalisation D) SWAN normalisation. Plots E & F are density plots using M values from all samples. E) Raw data, prior to probe type adjustment F) SWAN adjusted data. Technical replicates are plotted using different colours. All other samples are in green. Black arrows indicate probe type bias in raw data (E) and the removal of this bias using SWAN (F).

4.3.1.3 Data Exploration: Structures within the data validate the use of age- and sex-matched controls

Once general data quality control and pre-processing steps had occurred, the dataset was interrogated for any structures and patterns in the methylation variation.

To identify these data structures, principal components (PC) analysis was performed and correlations to potential covariates and confounders were calculated. The heatmap in Figure 4-5 shows that of the top 20 PCs of variation, age effects accounted for the top 2, as demonstrated by low p-values. Further, PCs 3 and 4 were also potentially explained by age, also showing very low p-values with slightly smaller Pearson's correlations. This validates the use of age matched case-control pairs, of which the vast majority of the sample is comprised.

Sex also contributed significantly to variability within the methylation dataset. This was despite having removed the sex chromosomes from the dataset during preprocessing (Chapter 2.3.3.1). The PCs associated with sex were 9, 15 and 17 and 20, representing relatively small fractions of the total variation. Despite this, having 3-4 PCs with strong significant correlations with sex (all $p < 6 \times 10^{-3}$) again validated the use of sex matched case-control pairs for a vast majority of the sample.

Samples are assayed using a 'chip' containing 12 samples and PC analysis revealed that 'chip' was also another large component of variation. Differences in data across chips may arise due to technical or assay preparation differences potentially leading to values that differ significant from true values. PCs 5 and 6 were observed to be strongly correlated ($p = 5 \times 10^{-16}$ and $p = 3 \times 10^{-5}$ respectively) with chip. Considering this, adjustment for chip effects would be used in differential methylation analysis. These chip related associations also validate the randomisation of sample placement across chips. Samples were randomised not only to avoid technical batch effects due to chip differences, but also to avoid confounding due clustering of samples with similar ages and sex. These effects were evidently prudent to consider prior to assay.

Ethnicity was also explored as a potential variable of relevance to this data. As noted earlier, Caucasians (defined as both parents stating Caucasian ancestry) constituted 63% of this sample set (section 4.2.1). This definition allowed for enrichment of DNAm differences to other groups. Caucasian vs non-Caucasian status was explored in the overall data. However,

it was not strongly associated with DNAm, particularly compared to age. Only principal component 12, with relatively high p-value (0.03) and low correlation (compared to age) was identified, far lower than those from the age associated principal components (data not shown). Further, the number of Caucasians in cases vs controls was equal, limiting any potential bias due to ethnicity. Since this highly enriched ethnic group did not associate with DNAm in this study, this was the only ethnic group explored as it was unlikely that much smaller and more heterogeneous ethnic groups would have substantial effects on DNAm in this study.

Therefore, this analysis identified chip as a factor to be adjusted for in differential DNAm analysis, and validated the use of age and sex matched controls. Due to the matching of controls by age, sex, and ethnicity for the large majority of samples, these factors will not be included as covariates in future differential methylation analyses.

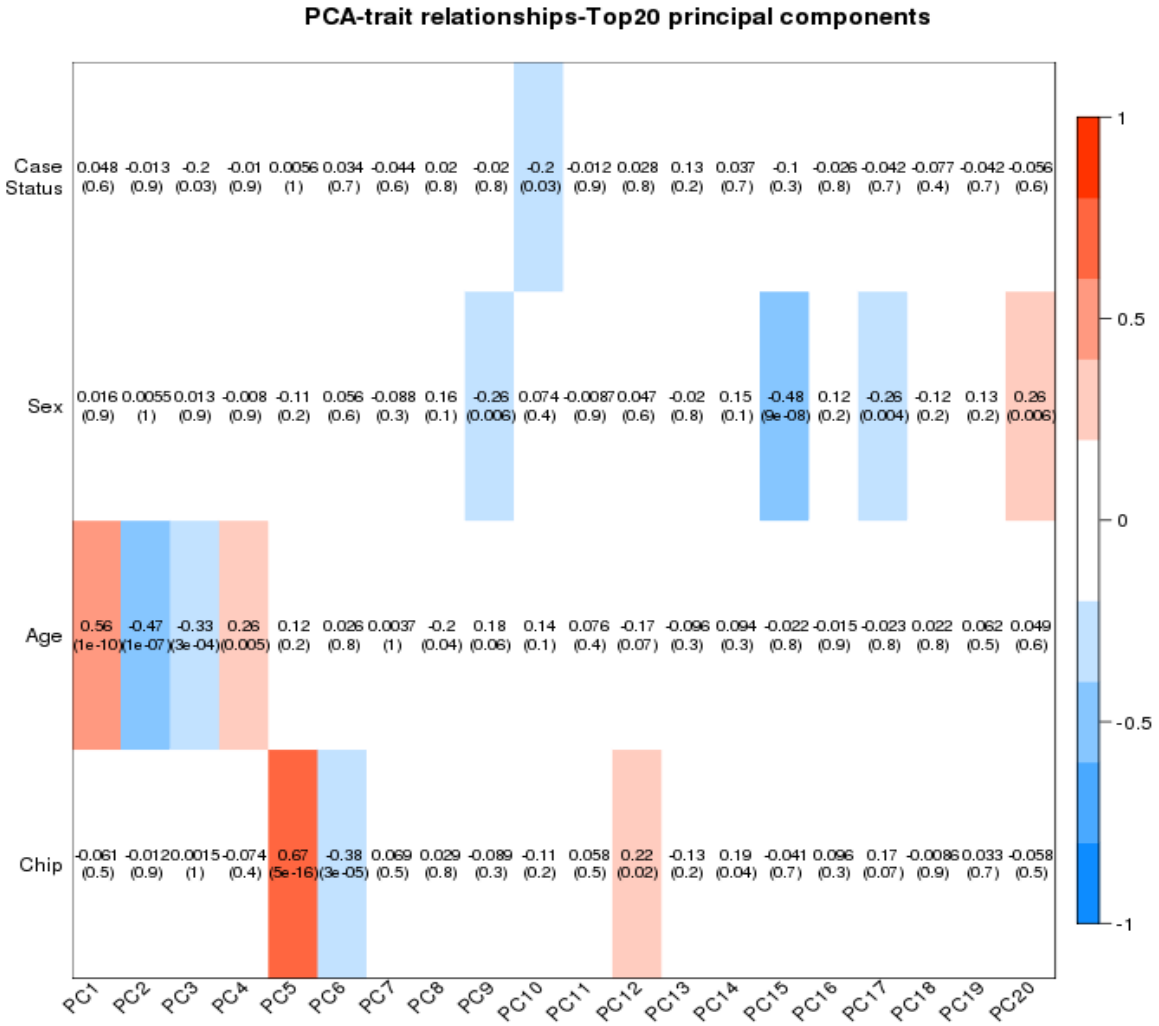

Figure 4-5 Heatmap of the top 20 Principal components.

Principal components (PC) listed on the x-axis, with variable of interest listed on the y-axis. For each PC-phenotype pair, Pearson's correlation is given with p-value below in parentheses. Pairs are coloured according to strength of correlations guide suggested be Evans (1996), as indicated on the colour bar on the far right [395]. Positive correlations are in red, and negative correlations are in blue.

4.3.2 Differential DNA methylation analysis: Use of LIMMA and RUV pipelines

To carry out differential methylation analysis, LIMMA was implemented to compare DNAm between JIA cases and controls. See Chapter 4.2.2.1 for more details on the method implemented.

No probes reached significance, defined as a Benjamini-Hochberg adjusted p-value<0.05. Since individual CpGs were not associated with oJIA, the potential for a DNAm signature related to oJIA was explored.

Figure 4-6B shows that using 1421 of the top ranked (unadj.p < $5x10^{-3}$) differentially methylated probes (DMPs) partially separates case/control clusters. At the same time, this plot also shows a strong divide within the clusters occurring along dimension 2 of the MDS plot. This source of variation is therefore defined by the 2nd largest principal component amongst these probes.

To attempt to identify the variable defining this division in clusters, MDS plots were labelled with a number of different variables. Figure 4-7 shows that this component was not able to be explained by a number of variables available for this study. Therefore, it was unknown whether the source of this divide represented a technical or biological factor. This variation may have confounded results, reducing the sensitivity of the analysis. Therefore, removing this unwanted variation was deemed necessary for further clarification of the data.

4.3.2.1 RUV analysis shows no significant DMPs, yet cluster patterns suggests DNA methylation can differentiate between JIA cases and controls

Having largely removed medication, disease heterogeneity, age and sex as sources of confounding through the experimental design and subsequent data exploration steps, these typical sources of confounding could be largely ruled out as the source of this unwanted variation. Additionally, other phenotypes available for the participants did not associate with the variation. Therefore, in order to attempt to adjust for this variation, a pipeline named RUV (embedded within the MissMethyl package), was implemented [338, 339].

RUV attempts to use the data itself to capture factors that need to be adjusted for in linear modelling. It attempts to identify both technical and biological variation not related to the factor of interest. More detail on RUV is presented in Chapter 4.2.2.2.

Both technical and biological variation unrelated to the comparison of interest were estimated and adjusted for using RUV. MDS plots using probes with unadjusted $p<0.01$ (Top ranked 4166 probes) visualise cases and controls clustering in separate, if slightly overlapping groups. This plot did not display the large splitting of the groups visible in the LIMMA analysis (Figure 4-8). This suggested that it was an unknown (biological or technical) variation identified by RUV that caused the splitting of clusters in the LIMMA analysis. RUV analysis did appear to have removed the large source of variation visible in MDS plots. Despite correcting for both technical and biological noise through various approaches, no significant DMPs were found. Interestingly, MDS plots of top ranking RUV probes also show that case-control clusters are dispersed across a large part of dimension 1 (Figure 4-8), but this did not appear to affect case control clustering.

Similar to MDS plots from the LIMMA analysis, top ranking RUV DMPs were also able to separate cases and controls. This was despite the lack of differentially methylated probes. Therefore, there was a preliminary suggestion that DNAm data contains patterns that can be defined by case or control status. Therefore the DMPs defining this case-control separation were explored further.

The results from the RUV analysis contributed to a publication in January 2018 in the Journal of Autoimmunity, “The DNA methylation landscape of CD4+ T cells in oligoarticular juvenile idiopathic arthritis”, and is listed in Appendix III.

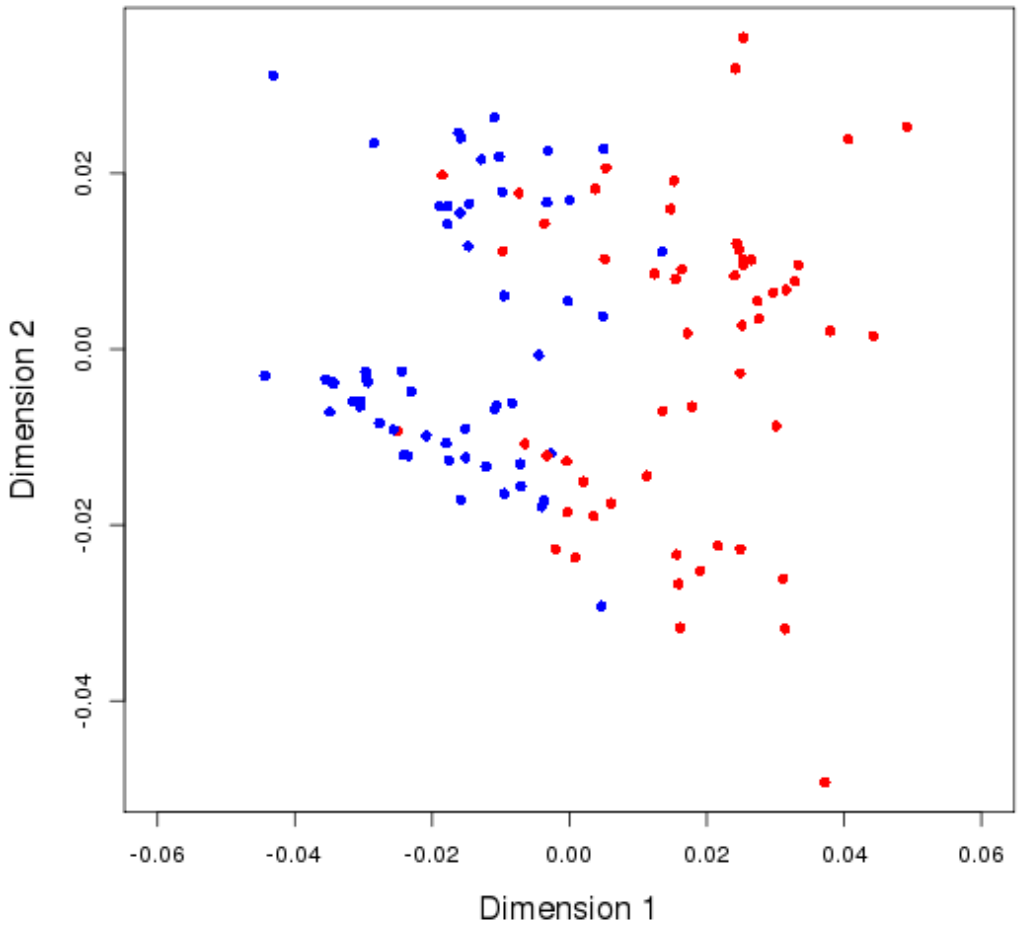

Figure 4-6 MDS plots of top ranking DMPs from LIMMA analysis with unadjusted p-value<0.005 (1421 probes).

Includes all samples, coloured by sample group (Case or Control).

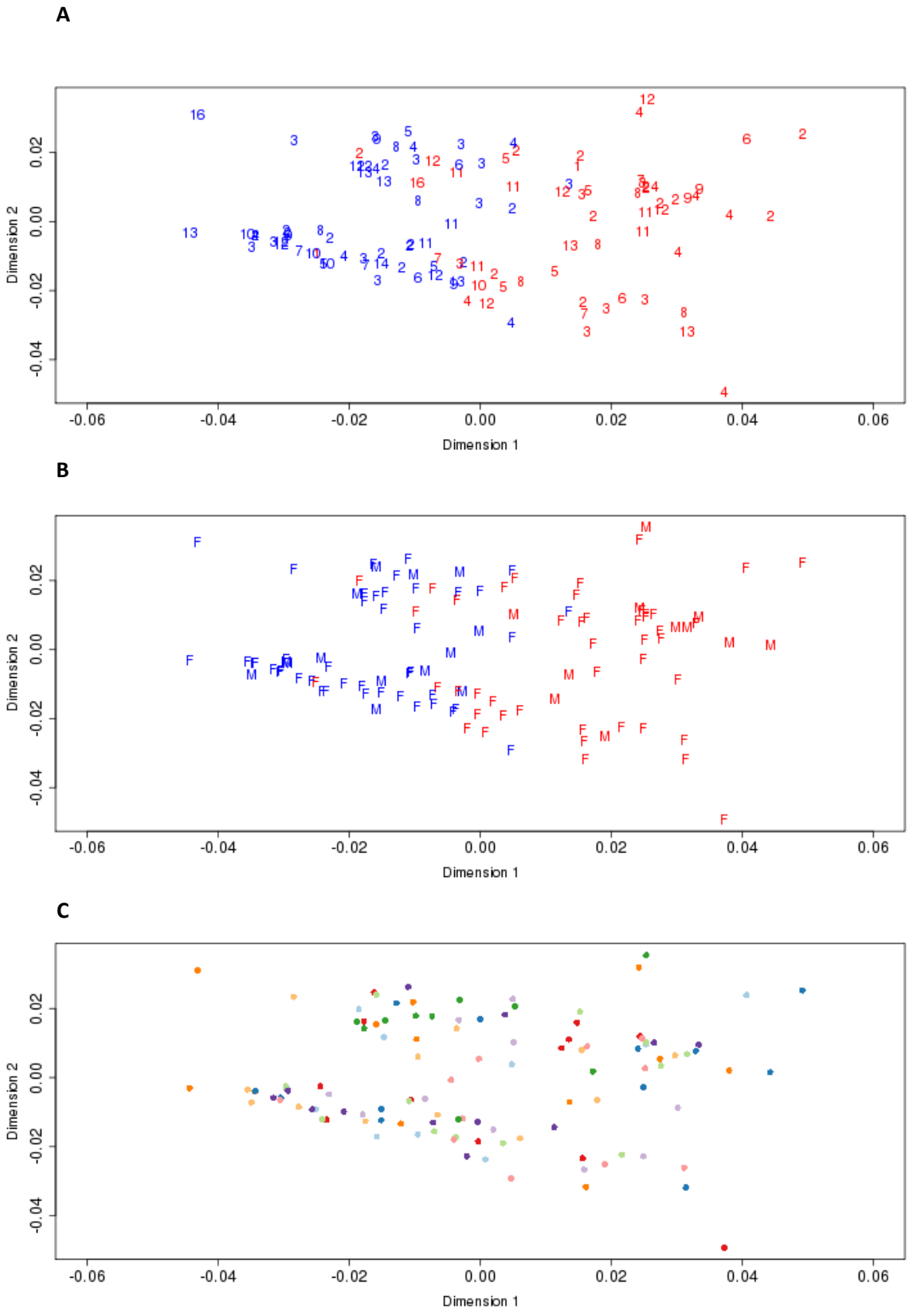

Figure 4-7 MDS plots of top ranks LIMMA DMPs, with unadjusted p-values <0.005.

Samples were labelled with different phenotypes to attempt to identify the source of the split in the clustering. A) Samples are labelled with Sex B) Samples are labelled with Ages of samples C) Samples are coloured according to Chip ID. Cases are red and controls are blue in A) and B).

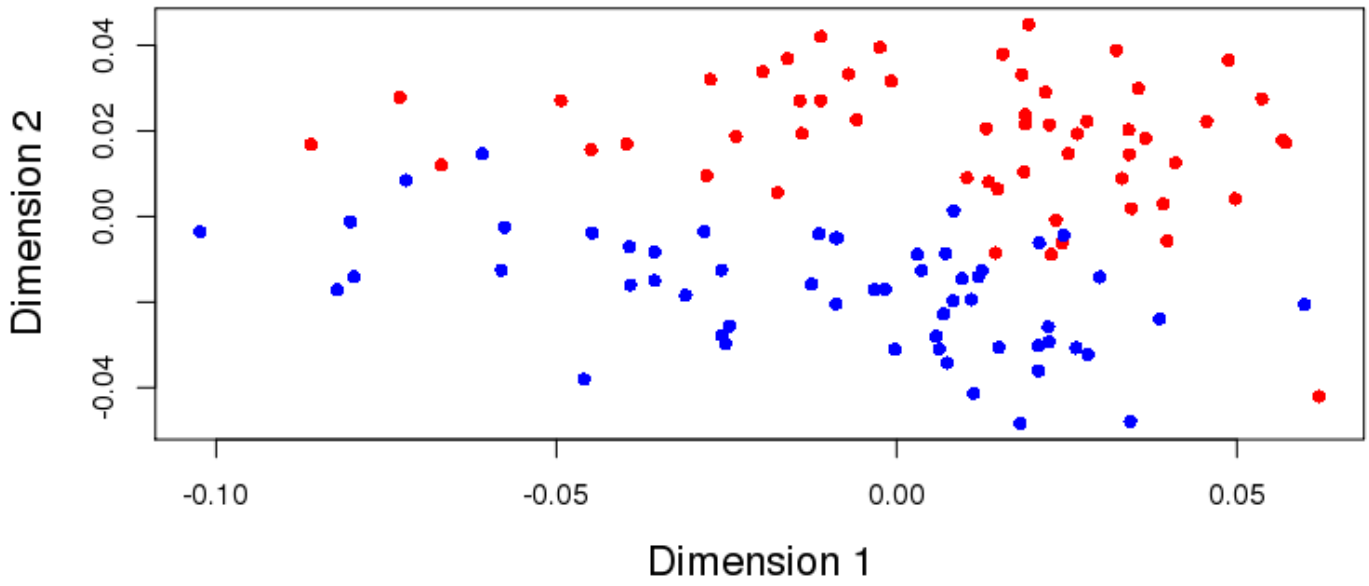

Figure 4-8 MDS plot of RUV analysis using the 4166 top ranked RUV DMPs.

DMPs were identified in the case: control analysis with unadjusted p<0.01. Cases are in red and controls in blue.

4.3.3 Technical validation studies from selected top RUV DMPs

Despite having an experimental design which reduced the heterogeneity of the oJIA sample group, RUV DMP analysis did not yield any significant associations at FDR 0.05 with JIA. However, MDS plots from both LIMMA and RUV suggested that case/control status may likely explain cluster patterns based on the top ranking loci. In light of this clustering, it is possible that probes used for cluster analysis (unadj.p<0.01, top ranked 4166 probes) may contain genuine case-control differences. In addition, a number probes with unadj.p<0.01 were at or near genes that had been previously associated with immune functions or epigenetic mechanisms (described below in Chapter 4.3.6). Therefore, to determine whether data from these probes had been accurately measured in this dataset, methylation was re-measured in the same samples using an independent analysis platform. This constitutes technical validation [396]. Similar results across two platforms would strengthen confidence in the suggestion of JIA related DNAm differences evidenced by MDS plot of RUV data (Figure 4-8).

Technical validation was completed using the Agena EpiTYPER system, which is a loci specific method using Mass spectrometry to quantitate methylation (see Chapter 2.3.6-2.3.15) for a detailed outline of the technique).

To select probes of interest, a selection criteria was implemented and is described below.

4.3.4 Rationale and criteria for selecting DMPs

In order to select DMPs for technical validation, selection criteria were implemented. This is important as measurements from methylation studies are susceptible to measurement error [397]. Making better use of data for selecting probes for technical validation may help reduce the impact of measurement error. Therefore, the selection criteria includes use of the known phenomenon of regionally correlated methylation. This involves the inclusion of a CpG of interest surrounded by CpGs with similar methylation levels [397]. This correlation helps identify genuine differences in methylation, which are not simply the result of outliers or technical errors. In addition, these correlated regions are generally associated with more functionally relevant findings [397]. The selection criteria also included a requirement for a plausible link to immune or epigenetic pathways.

The selection criteria used here included:

- Implementation of a P-value cut-off: using unadjusted p-values <0.01 (top ranked 4166 probes). These DMPs were able to cluster cases from controls using MDS analysis. These probes are therefore likely to contain true case-control differences.
- Beta value cut-offs: mean delta beta values of ~5% or greater between cases and controls. This was selected to take into account the technical limits of the EpiTYPER system, as discussed by Coolen *et al.* [341]. They found EpiTYPER could consistently detect differences of ~5%. As such, selecting for differences of this magnitude means these are more likely to be detected in the EpiTYPER system, and less likely to be overcome by the technical noise from the system [341].
- Regional correlation of methylation values within p-value cut-offs, and delta beta values ~5%: 2 or more CpGs within +/-2kb, delta beta in same direction. This takes advantage of regional correlation of methylation, where CpGs approximately +/- 2kb of any one CpG site are highly linked [397, 398]. Selecting regions where methylation is correlated minimises the likelihood that observed changes are due to technical effects.
- Functional relevance: the annotated gene has published data with associations with immune system or epigenetic mechanisms

4.3.5 DMPs linked to oJIA successfully validate in the same sample set

A total of 5 DMPs were interrogated using the EpiTYPER system. Namely, *CHD5, TACSTD2, VPS53, SYNE1* and *CYP1A1*. Details of the EWAS probe selected are listed in Table 4-5. The functional relevance of these is outlined briefly below.

CHD5 is known to form a chromatin remodelling complex (NuRD: Nucleosome remodelling histone deacetylase) to effect epigenetic changes and has been implicated in altering gene transcription. [399, 400]. There is also some evidence to suggest that it is involved in natural killer cell cytotoxicity and Toll Like Receptor signalling and interferon expression [401]. There are 3 DMPs within the p-value and delta beta thresholds spaced 200bp apart, located in a CpG island.

CYP1A1 has been associated with DNAm as a gene highly likely to be responsive to smoking exposure [402, 403]. This is of particular interest as debate exists around the importance of smoking exposure on JIA risk [154]. In addition, smoking exposure is a clear risk factor for

rheumatoid arthritis, an analogous disease in adults [404]. Further, there are suggestions it is involved in TLR signalling [401, 405]. Three other CpG sites falling within the p value and delta beta thresholds were also found. These are located in a CpG island and encompasses a transcription start site (TSS1500) for this gene.

TACSTD2 is known to be regulated by AIRE (**A**uto**I**mmune **RE**gulator), a known gene involved in with T cell tolerance and autoimmune disease prevention [406, 407]. Expression of TACSTD2 also drives the expression of NF-κB, STAT1 and STAT3, proteins known to be important in inflammation and immune responses [407]. Six other probes are also within the p value and delta beta cut-offs.

SYNE1 is known to be a cell cytoskeletal protein [408]. It is predicted to be involved in a number of immune cell pathways including chemokine signalling pathway through genetic and physical interactions with DISC1 (https://thebiogrid.org/, [401]). Disc1 is important particularly in the interaction with the immune system to produce schizophrenia behaviours in mice [409, 410]. Two other probes within the prescribed limits are nearby, and all are located ~500bp from a CpG island.

VPS53 contains 2 DMPs which are within the stated limits for regional correlation, and are located just upstream of a predicted mRNA for *VPS53*. This protein is able to move to the mitochondria and induce apoptosis, a function with hypothesised impairment in many autoimmune diseases [411, 412].

In contrast to results obtained in the previous vitamin D study (Chapter 3.3.4), all DMPs linked to oJIA in the discovery EWAS were successfully validated with the same direction of effect. Further, almost all EWAS DMPs interrogated on EpiTYPER were significant by students t-test at $p<0.05$ (Figure 4-9). This supports the decision to investigate non-FDR (0.05) significant DMPs.

In addition, data from all assays suggested the presence of differentially methylated regions. CpG sites proximal to the selected DMP but not assayed on the HM450k arrays were able to be interrogated using the EpiTYPER system. Many of these proximal sites were also found to be significant ($p<0.05$, students t-test) and showed delta beta values in the same direction. Overall, a large majority of EpiTYPER-interrogated CpG sites (33/37) across all 5 assays were

significantly associated with oJIA ($p<0.05$) and displayed case-control delta beta differences in the same direction as the EWAS DMP (Figure 4-9).

Table 4-5 Summary information of EWAS probes selected for technical validation (CHD5, CYP1A1, TACSTD2, SYNE1 and VPS53).

Probe name	Gene	Chromosome	Position (Hg19)	UCSC Refgene Group	delta beta	Unadj. P-value	Distance of nearby CpGs	Delta beta values of nearby CpGs
cg27316956	*SYNE1*	6	152958899	TSS1500	-0.05	0.000471	+275	-0.04
cg12101586	*CYP1A1*	15	75019203	TSS1500	-0.05	0.003592	-133, -60,+108,+140,+159,+233	-0.03,-0.05,-0.04, -0.05,-0.01
cg17210938	*TACSTD2*	1	59043173	TSS200	-0.06	0.005779	-242, +26, +35, +82, +107, +197	-0.03, -0.09, -0.08, -0.08, -0.08, -0.10, -0.04
cg15205435	*CHD5*	1	6187920	Body	-0.06	0.006249	-287, +200	-0.02, -0.04
cg04370829	upstream of predicted mRNA for *VPS53*	17	406249		0.07	0.005947	+252	0.07

Genomic locations are according to Genome Reference Consortium GRCh37 Hg19. Delta beta values reflect hypomethylation (-) or hypermethylation (+) in cases. Distance of nearby CpGs are upstream (-) or downstream (+) of EWAS probe position listed under 'Position (Hg19)'.

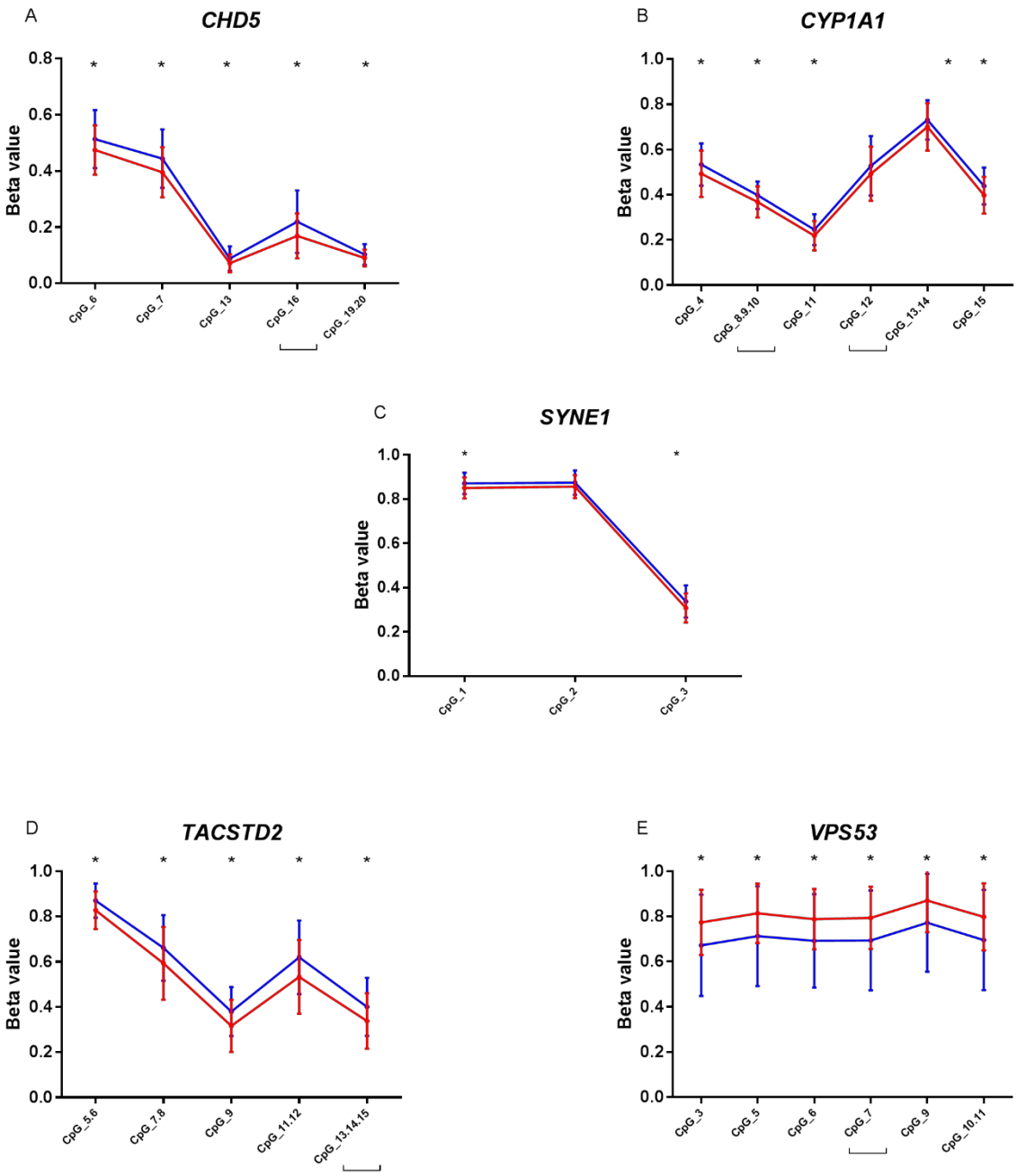

Figure 4-9 Methylation data from EpiTYPER technical validation studies from 5 selected genes.

CpG site interrogated on the EWAS array is underlined. EpiTYPER was able to interrogate multiple sites proximal to the selected CpG. A) CHD5 B) CYP1A1 C) SYNE1 D) TACSTD2 E) VPS53. Cases are in red and controls are in blue. * Denotes significant differences at $p<0.05$ (students t-test)

4.3.6 Investigation of the functional relevance of aberrant DNA methylation to oJIA

DNAm differences between cases and controls may have direct functional consequences on disease. Equally, they may represent a useful measurable biological characteristic correlating with disease (biomarker) [413]. To explore the potential functional effects DMPs, genes annotated to the top ranked 4166 RUV DMPs that were able to cluster cases from controls in an MDS analysis were subject to Ingenuity pathway analysis. This revealed a number of pathways related to the immune system, including canonical pathways '*SAPK/JNK* signalling', '*Wnt* signalling' and the 'Th1 pathway' (Figure 4-10) [414, 415]. Further, a large number of genes relating to 'Cell Death and Survival' in a number of different cell types (including immune cells) were also enriched amongst the top ranked DMPs (Table 4-6).

Gene ontology analysis using the online tool DAVID was also carried out, using the same genes [392, 393]. To facilitate interpretation of the many gene ontology terms, clustering of gene ontology terms with similar functions was completed and visualised using REViGO [394]. Functional clusters suggest that receptor binding and receptor activity were enriched ontologies (Figure 4-11). Within these clusters, a number of immune system related terms clustered into these groups.

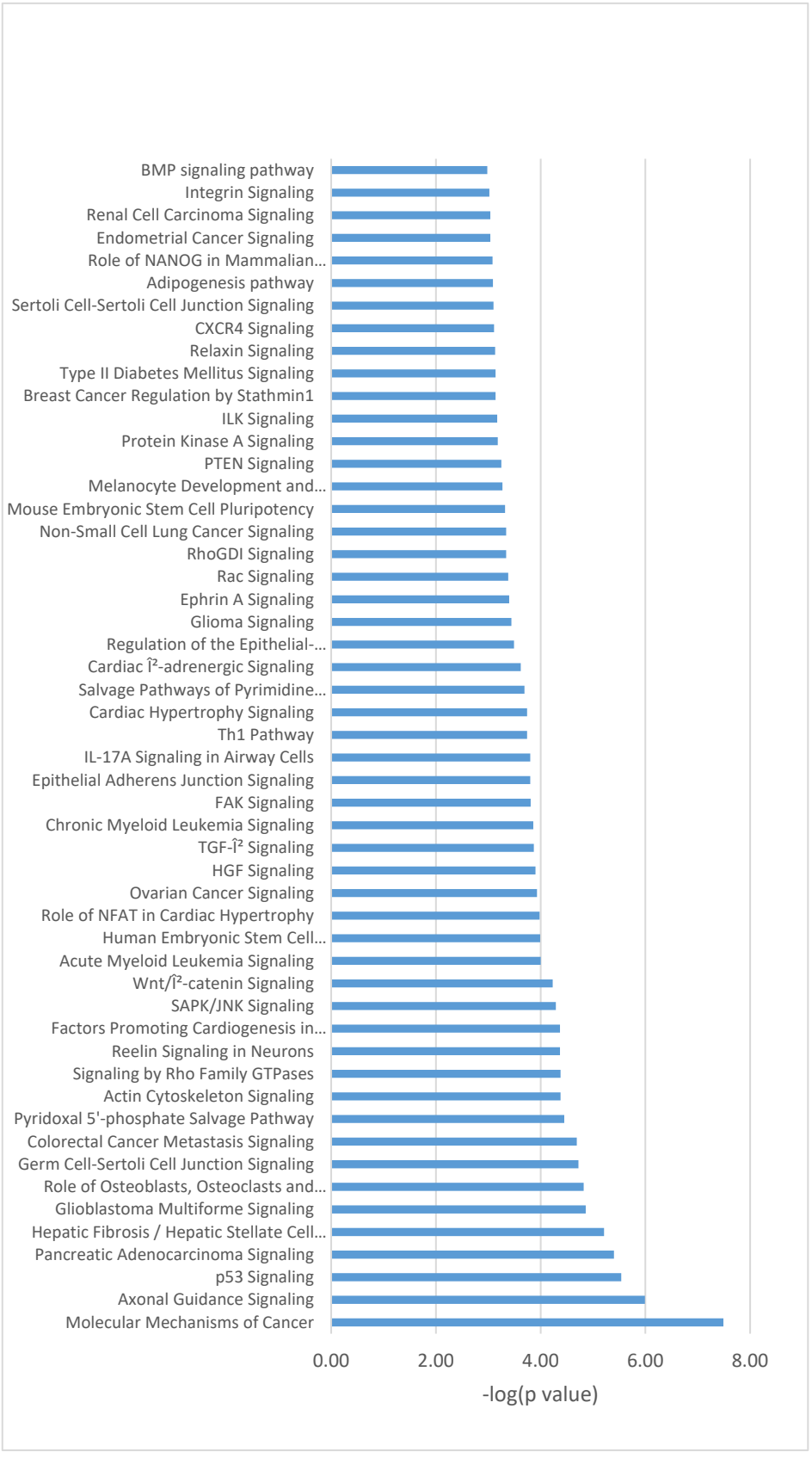

Figure 4-10 Canonical pathways enriched in the top ranking DMPs from RUV analysis.

Probes with unadj.p<0.01 were used as input for the IPA pathway analysis. Shown are the Canonical pathways that are enriched at p-value<0.001

Table 4-6 Top Molecular and Cellular Functions from IPA analysis

Molecular/Cellular Function	p-value range	No. Genes
Cell Morphology	5.34E-06 - 1.36E-24	639
Cellular Growth and Proliferation	6.37E-06 - 2.72E-22	991
Cell Death and Survival	6.08E-06 - 2.42E-19	823
Cellular Development	6.29E-06 - 5.38E-18	905
Cellular Movement	4.27E-06 - 8.63E-18	550

Genes from 4166 top ranked RUV DMPs (unadj.p<0.01) were used as input for this analysis. 'P-value range' refers to p-values of individual pathways sitting within these higher level functions. 'No. Genes' notes the total number of unique genes within the higher level functions.

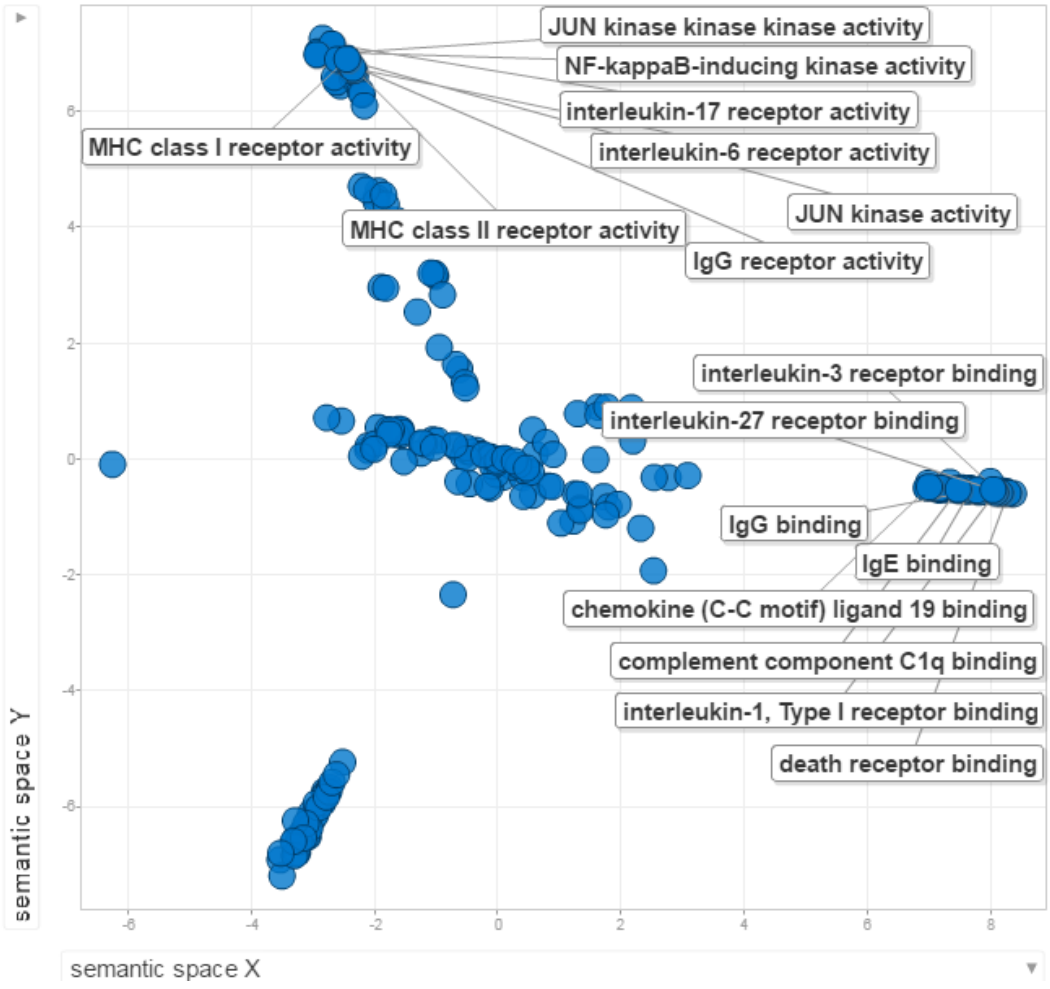

Figure 4-11 Visualisation of gene ontology groupings.

Genes from 4166 top ranked DMPs (unadj.p<0.01) were used to extract gene ontology information from DAVID. These were then used as input for REVIGO, to summarise and group gene ontology information. Bubbles represent gene ontology groups. Closely clustered bubbles represent groups that are more similar to each other. Ontologies related to immune system functions are labelled, which reside amongst 2 clusters with similar gene ontologies relating to receptor activity (uppermost cluster) or receptor binding (far right cluster).

4.3.7 Overlaps with gene expression data suggest a functional relevance of aberrant DNA methylation in JIA

Functional analysis was continued by exploring gene expression data from with the oJIA cases and controls. To do this, a small subset of matched case control samples from the EWAS sample group were analysed using RNA seq. This data was partially generated by Rachel Chiaroni-Clarke for a pilot study of sex specific differences in oJIA gene expression. Ten cases and ten controls were selected, age and sex matched as listed in Table 4-7. Samples stored in TRIzol were then used for RNA-seq data generation within the MCRI Translation Genomic Unit. Data was generated using an in house service provider (see Chapter 2.5.2). Differences in expression (log(FC)) between oJIA cases and controls were paired to delta beta values from each gene, focussing on the annotated genes from the top 4166 DMPs (unadj.p<0.01). Figure 4-12 highlights the 5 DMPs selected for validation, annotated as *CHD5*, *CYP1A1*, *TACSTD2*, *VPS53* and *SYNE1.*

All probes selected for validation but *CHD5* were located at or near a transcription start site. Therefore, gene expression for all but *CHD5* was expected to follow the canonical inverse relationship to DNAm. DMPs for *CHD5* were located in the gene body, where gene expression is positively correlated with methylation. Three of the 5 validated probes display expression fold changes of ~1.5, suggesting a correlation of methylation with gene expression. However, these changes did not follow these expected inverse relationship of methylation to gene expression [416]. A further 2 of the selected methylation loci have unchanged gene expression, suggesting these may be DMPs with no functional effect on the gene immediately proximal to the probe.

Figure 4-12 shows top-ranked DMPs that have gene expression fold changes of ~1.5 with DNAm delta beta values of 5%. A number of these loci relate to immune system functions such as antigen presentation (*HLA-DQA1*) and inflammation and immune response to parasites (*ITGBL1*) (Figure 4-8). The functional implications of the overlapping EWAS and RNAseq data suggested that some of the top EWAS DMPs may have some functional relevance. Not all genes from validated DMPs show expression changes, which may reflect a correlation with disease instead of a functional association. However, considering all probes with unadj.p-value<0.01 can cluster cases and controls (Figure 4-8), and the potential functional relevance observed (via gene ontology, pathway analysis and RNAseq), this supported efforts for an exploratory replication study to be undertaken.

Table 4-7 Samples used for RNAseq analysis.

Sample_Name	Sample_Group	Age	Sex	Subtype	ANA	Pairs
C00120	CONTROL	2	F	control	control	1
J00388	CASE	3	F	extended	pos	1
C00123	CONTROL	11	M	control	control	2
J00182	CASE	11	M	extended	neg	2
C00188	CONTROL	4	F	control	control	3
J00329	CASE	4	F	extended	NA	3
C00197	CONTROL	2	M	control	control	4
J00401	CASE	2	M	persistent	pos	4
C00366	CONTROL	12	F	control	control	5
J00008	CASE	11	F	extended	pos	5
C00385	CONTROL	8	M	control	control	6
J00390	CASE	9	M	persistent	pos	6
C00395	CONTROL	13	F	control	control	7
J00133	CASE	12	F	extended	pos	7
C00433	CONTROL	3	M	control	control	8
J00099	CASE	4	M	persistent	pos	8
C00440	CONTROL	3	F	control	control	9
J00410	CASE	2	F	persistent	neg	9
C00542	CONTROL	4	M	control	control	10
J00020	CASE	5	M	extended	pos	10
C00536	CONTROL	9	M	control	control	~
J00155	CASE	8	F	persistent	pos	~
J00372	CASE	13	M	persistent	pos	~

These samples are a subset of those used for HM450k analysis. Subtype: refers to a known subgroup of oJIA, noted as persistent (1-4 joints affected) or extended disease (>4 joints affected after the first 6 months). ANA: Anti-nuclear antibody status, with ‘pos’ denoting positive appearance of antibodies in the patients’ serum. Controls were not test for ANA. Pairs: Samples with the same number are matched for age and sex.

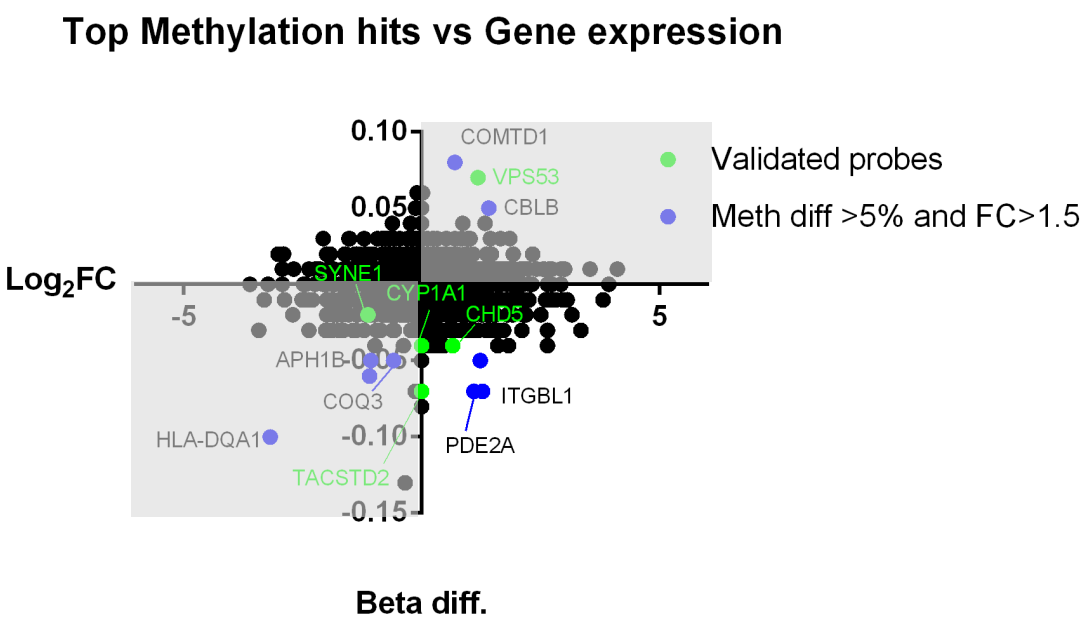

Figure 4-12 Top EWAS hits vs log(FC) of a subset of EWAS samples.

Genes from the top 4166 hits of the RUV analysis. Canonical inverse relationships between methylation and expression are denoted by grey quadrants Beta diff. refers to mean methylation differences between oJIA cases and controls.

Table 4-8 Top EWAS hits with 5% delta beta and ~1.5FC from RNAseq data, captured from a subset of the EWAS sample group.

Gene	Name	Pathways
HLA-DQA1	Major Histocompatibility Complex, Class II, DQ Alpha 1	Antigen Presentation
ITGBL1	Integrin, Beta-Like 1	Cell-endothelium adhesion (inflammation), immune response to parasites
PDE2A	Phosphodiesterase 2A	Mitochondria location
KCNK3	Potassium Channel, Two pore domain subfamily K, Member 3	PKA pathway (growth, development and metabolism)
VAMP5	Vesicle associated membrane protein 5	muscle development
COQ3	Coenzyme Q3 Methyltransferase	Mitochondria electron transfer
APH1B	APH1B Gamma secretase Subunit	Notch pathway (immune cell development) and NFKb signalling
CBLB	Cbl Proto-oncogene B, E3 Ubiquitin Protein ligase	Regulates TCR, BCR, and Ig epsilon receptor
COMTD1	Catechol-O-Methyltransferase Domain containing 1	putative rRNA methyltransferase

4.3.8 Replication studies with additional oJIA cases

Given the validity of oJIA DMPs demonstrated above, replication was conducted in a new set of samples from within the CLARITY cohort using the same selection criteria as for the discovery samples (see Chapter 4.2.1). Eleven oJIA cases with active disease, and no prior exposure to methotrexate or bDMARDs were chosen from within the CLARITY cohort. Controls were selected to match for age and sex.

Table 4-9 describe replication cohort samples; Table 4-10 displays side by side the overall characteristics of the discovery and replication samples. These were comparable in terms of age distribution, showing similar means and near equal spread. Whilst the female to male ratio in the discovery cohort was ~3:1, reflecting ratios observed in the clinic, the replication cohort sex ratio was slightly under this ratio (~ 2.6:1). As age defined the 2 largest sources of variation in the discovery dataset (and up to 4 principal components chapter 4.3.1.3), it was deemed most important to have a similar age distribution comparable to the discovery cohort.

Table 4-9 Samples sourced from the CLARITY cohort for replication studies.

Sample ID	Sample Group	Age	Sex	Matched Sample (age & sex) status (matched/unmatched)
J00037	CASE	3.2	F	matched
J00176	CASE	9.6	F	matched
J00128	CASE	12.8	F	matched
J00541	CASE	11.6	F	matched
J00115	CASE	1.9	F	matched
J00092	CASE	10.8	F	matched
J00411	CASE	3.5	F	matched
J00582	CASE	4.2	F	matched
J00178	CASE	14.2	M	matched
J00567	CASE	7.7	M	matched
J00097	CASE	1.8	M	matched
C00526	CONTROL	3.9	F	matched
C00260	CONTROL	9.2	F	matched
C00753	CONTROL	13.3	F	matched
C00431	CONTROL	11.3	F	matched
C00397	CONTROL	1.7	F	matched
C00543	CONTROL	10.5	F	matched
C00121	CONTROL	3.9	F	matched
C00593	CONTROL	4.1	F	matched
C00760	CONTROL	14.0	M	matched
C00289	CONTROL	7.8	M	matched
C00742	CONTROL	2.7	M	matched

Table 4-10 Discovery and Replication sample group summary tables

A

	Discovery		
	Case	**Control**	**Matched pairs[a]**
All Inclusive	56[c]	57	51
Mean age (SD)	6.1 (4.0)	6.4 (4.2)	
Female/Male	45/11	43/14	41/11
≤6yrs/>6yrs at diagnosis[b]	35/21	33/24	29/21

B

	Replication		
	Case	**Control**	**Matched pairs[a]**
All Inclusive replication	11	11	11
Mean age (SD)	7.4 (4.4)	7.5 (4.2)	
Female/Male	8/3	8/3	8/3
≤6yrs/>6yrs at diagnosis[b]	5/6	5/6	5/6

A.Discovery samples. B. Replication samples. [a] Matched by age and sex. Mean age = age at blood collection [b] Number of cases in each age-at-diagnosis category/number of controls in same category for age at recruitment. [c] Does not include one case removed due to inactive disease at time of blood collection

4.3.9 DNA methylation analysis of replication samples – an effect of age?

Four loci from technical validation steps with the largest delta beta values were taken to the replication group and assayed using Agena EpiTYPER. This included *VPS53*, *TACSTD2*, *CYP1A1*, *CHD5* and excluded SYNE1 (delta beta ~2% in technical validation experiments).

Surprisingly, despite earlier successful validation, replication of the data was not achieved. Figure 4-13 (A-D) show discovery, validation and replication data side by side for each loci taken to replication studies. In almost all loci, the opposite direction of delta beta values were observed compared to EWAS and technical validation data.

Notably, there was a large spread of values inherent to discovery and technical validation data. This was observed across many of the loci assayed, spreading from 30% of the possible methylation range to almost 90%. The large spread in the data may be indicative of a complex methylation landscape, potentially shaped by a number of factors that go beyond cases vs control status including age, genetics and stochastic variation [236, 397]. Whilst the methylation profile appears complex, there were hints regarding the importance of age to the discovery data (Figure 4-5). Therefore, age was explored as a potential source of complexity in the methylation landscape.

Age is known to be associated with DNAm data, with published data going so far as to propose the prediction of biological age from methylation data [417]. In this sample group, the age distribution was large and covered almost the entirety of childhood i.e. discovery samples age range spanned from 2-16 years of age. Figure 4-5 shows that age explains the first 2 components of variation. To assess if age could explain a large portion of the variation in the data within the top ranking DMPs, age labels were placed over the MDS plot which utilised the top ranking 4166 probes with unadjusted $p<0.01$ (Figure 4-14). The first component of variation (dimension 1) does indeed appear to be defined by age. In addition, the spread of ages appears to increase as age increases. That is, a majority of younger samples appear to cluster more tightly together between 0 and ~0.03 along dimension 1. This is compared to older samples, which appear to disperse along a larger part of dimension 1 between -0.1 and -0.025. This grouping does not appear to be limited to case status, as it is a phenomenon common to both cases and controls. These differences in the spread of older and younger samples suggests that there are DNAm differences within the top ranked probes based on age. Any potential differences within case and/or control groups may add to the complexity of the methylation landscape. That is, variation amongst samples may lead

to increased variation in methylation values at a given loci. Therefore, it may be important to take into account older and younger age subgroups in any analysis of the oJIA EWAS data. The observation of potential age subgroups in the DNAm data overlaps with observations of different clinical outcomes in JIA based on age subgroups. This include gene expression data finding age specific associations, network analysis identifying pathway differences in age subgroups and HLA associations identified as specific to younger age groups [12, 14, 17, 418]. This suggests analyses of age subgroups is important in DNAm in oJIA, and will be the focus of further investigations in Chapter 6.

The results from the technical validation and replication analysis contributed to a publication in January 2018 in the Journal of Autoimmunity, “The DNA methylation landscape of CD4+ T cells in oligoarticular juvenile idiopathic arthritis”, and is listed in Appendix III.

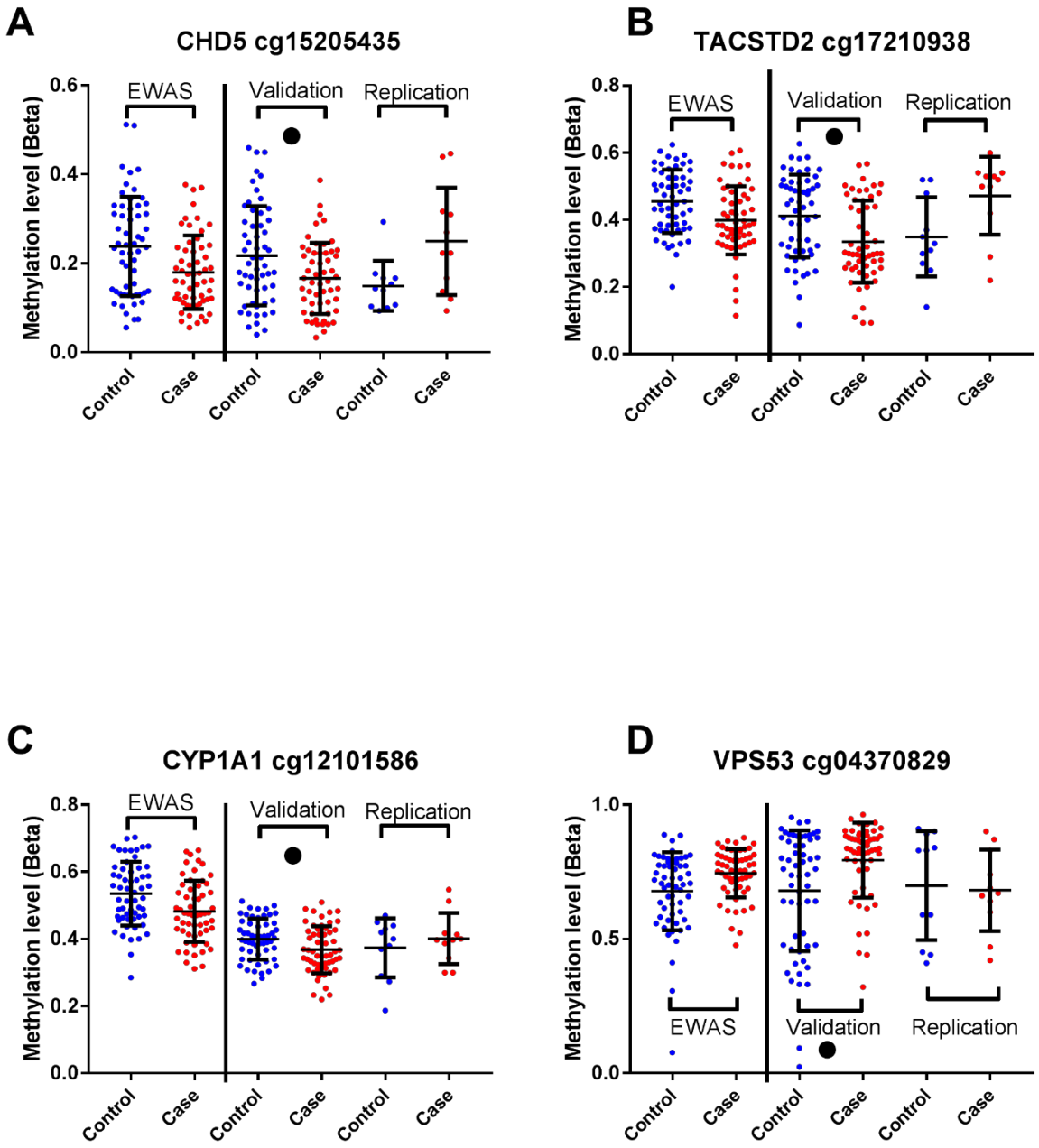

Figure 4-13 Scatterplots of CpG sites interrogated on both the EWAS and EpiTYPER systems.

Cases are in red, controls in blue. Whiskers represent SD and horizontal line represents means. Black dot represents significance at p<0.05 (students t-test).

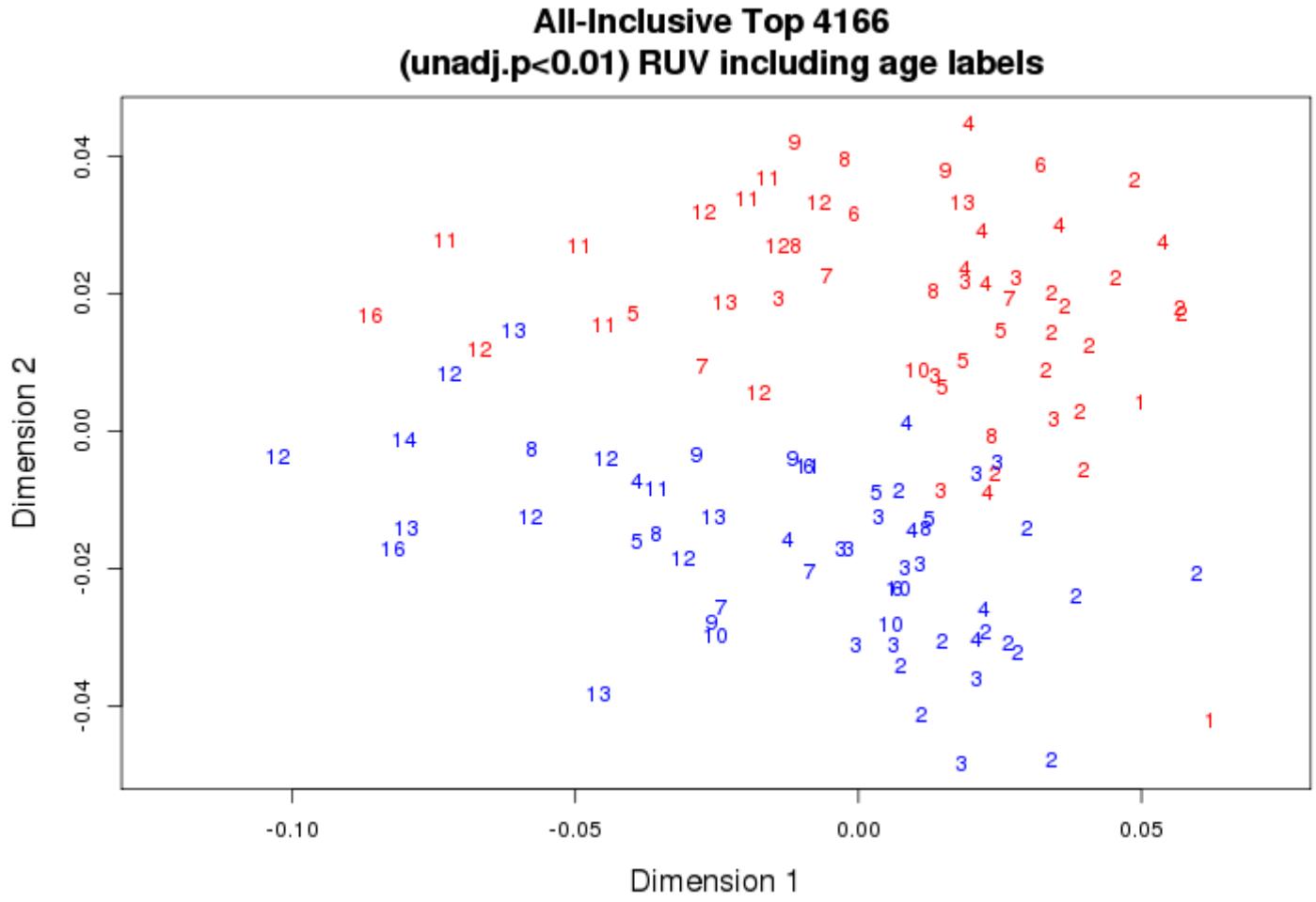

Figure 4-14 MDS plot of the top ranked probes with unadj.p value<0.01 (4166 probes) from RUV analysis.

Cases are in red and controls are in blue, and all samples are labelled with the age at blood collection.

4.4 Discussion

The results of this study of DNAm in oJIA cases with active disease was surprising in many respects. A number of EWAS publications have found large numbers of loci associated with autoimmune diseases such as SLE, T1D and RA have been highly cited according to NCBI citations (ranging from 58-405 citations) [245, 281, 287, 301]. Further, with the overlap in susceptibility loci, disease pathology and cell types of importance between RA and JIA [130], coupled with reports of significant aberrant methylation in RA, there were expectations of aberrant methylation at a genome-scale in oJIA also. Pilot work from this group also showed DMPs in a small sample group using the early genome scale technologies [292]. Therefore it was interesting that no statistically significant DMPs were identified. It was unexpected that these non-significant probes would be able to cluster samples according to case or control status, citing the lack of differentially methylated probes. However, using the top ranked 4166 probes (probes with unadjusted p values<0.01), this data was able to separate cases from controls in an MDS cluster analysis. These results appeared to reflect genuine differences in DNAm as a number of probes were successfully technically validated. Yet replication of the data in another set of samples from within the CLARITY cohort was not successful. Delta beta values in the replication data were in the opposite direction compared to discovery data. What was also observed was the large spread across the methylation range within the discovery data. This large variation in the discovery dataset, together with a small set of samples used for replication work, may have resulted in sampling error leading to replication delta beta values that were discordant with the discovery dataset.

Nonetheless, this study had a number of strengths compared to many previous autoimmune disease (AID) EWAS, and these strengths also shed light on potential reasons for the lack of significant DMPs. Indeed, it is perhaps the quality of the design that may best illuminate reasons for lack of DMPs detected.

OJIA being an autoimmune disease, it was expected that differences in methylation values were likely to reflect small differences found in other AID EWAS. As an indication, T1D differences published have been as small as delta beta 1% between discordant twins [301], using blood derived DNA. RA delta beta values were often published as ~1-5% [245], also using blood derived DNA. Therefore it was important to increase the homogeneity of the sample group, increase the specificity of the cell type investigated and reduce the confounding due to medication in order to increase the signal-noise ratio in the dataset. This

sample group reflects the concerted effort to address these issues, firstly by focussing on a particular subtype of JIA, oligoarticular disease.

This subtype allowed for a homogeneous sample group for a number of reasons. All cases were clinically phenotyped according to ILAR criteria by a paediatric rheumatologist, meaning that cases were clinically homogenous according to the gold standard classification system. Further, oJIA cases were less likely to be exposed to medications such as MTX and bDMARDs due to the reduced burden of disease at diagnosis. The selection criteria excluding cases compromised by these medications served to ensure cases were not exposed to MTX and bDMARDs. To further the case homogeneity, only cases with active disease were selected. Having only cases with active disease this served to reduce variation due to immune activation states [378, 379].

The selected criteria leading to these sample group characteristics attempted to diminish not only heterogeneity in the sample group, but also confounding due to these factors. Medication exposure to things such as cytokine inhibitors and methotrexate avoids potential effects on the methylome. MTX is a known a folate pathway/DNAm pathway inhibitor [325, 372]. BDMARDs serve to reduce immune system activation and inflammation and likely affect the methylome, which is known to vary according to inflammatory and activation states of immune cells, particularly during Th1 or Th2 differentiation or Treg activation [252-254, 258, 260, 264, 267, 268, 377]. Selecting only cases with active disease similarly avoids known changes in DNAm due to immune activation states. In addition, to reduce the confounding due to cell type differences in immune cell methylation, this study focussed on a particular T cell compartment, CD4+ T cells [253, 256, 260, 262, 267, 376, 419-421]. Not controlling for these factors may contribute to false positives i.e. cases – control differences in DNAm which would reflect medication exposure/immune system activation differences rather than disease causation/processes.

Indeed, a look at a number of AID EWAS' leads to the hypothesis that the sources of confounding this study design attempts to avoid may explain the large number of DMPs in previous AID EWAS', discussed in detail below.

Highlighting the potential for confounding, some papers even neglect to mention a number of these factors including medication [51, 287, 290, 291, 293, 295, 299], sample group heterogeneity[287, 289, 291, 299, 422], cell type specificity [288, 290, 300], disease duration

[287, 290, 291, 295, 299, 300], sex [293, 422] and age [289-291, 299]. In a number of these publications, thousands of differentially methylated sites have been reported from single publications. Therefore, the potential for confounding of results due to any of these factors remains an issue in many studies. That is to say, it remains a distinct possibility that many of the DMPs from previous studies are false positives due to the confounders named above. The number of genuine DMPs may indeed be few.

In this study, having addressed some of the risk of confounding through study design, it is possible that the majority of the signal expected may have been removed along with the confounders. Indeed, looking at papers that were able to control for a number of factors, the number of reported significant DMPs was far reduced compared to thousands of DMPs seen for many other studies. For example, papers from Liu *et al.*, Bell *et al.*, Rakyan *et al.* and Hasler *et al.* reported 9, 19, 132 and 703 DMPs in Rheumatoid arthritis, Type 1 diabetes and Inflammatory bowel disease respectively [245, 288, 295, 301]. Bell *et al.*, Rakyan and Hasler *et al.* all used a similar FDR cut-off to each other and to this study. The numbers of significant DMPs in these studies can be compared to studies where a number of confounders are likely to be present, as noted above. For example, SLE EWAS' reported by Lin *et al.* and Absher *et al.* have published 2165 and 1033 DMPs, respectively [287, 297]. Cooke *et al.* and Nimmo *et al.* with inflammatory bowel disease (IBD) EWAS' have published 3156 and 2714 DMPs, respectively, associated with disease [287, 290, 297, 300]. In addition, studies in RA by Nakano *et al.*, de la Rica *et al.* and Glossop *et al.* have reported 1859, 1240 and 2000 DMPs [291, 299, 423]. This lends support to the hypothesis that a majority of the signal from a number of studies may potentially be due to confounding factors.

4.4.1 Limitations

This study also had some limitations that may be unique to JIA, and perhaps immunological diseases of childhood, that may also explain the lack of DMPs and the difficulty in replicating discovery results.

Firstly, whilst this is one of the largest studies of JIA methylation to date, this study may still have been underpowered to detect small delta beta values. Some have proposed tentative suggestions for sample sizes for detection of small methylation differences, ranging from 100% power to detect 5% delta beta in 100 cases vs 100 controls, another suggests only 30% delta beta are detectable with 50 cases and 50 controls [424, 425]. In addition, biological variability differs between regions and this may also affect power calculations [236, 397]. Yet

there is no current accepted gold standard for calculating sample sizes for methylation studies. Therefore assessment of appropriate sample sizes are difficult, and this remains a limitation of this study.

Pathway analysis suggests a more powered study may detect disease relevant findings. Single CpG sites were not significantly associated with disease, yet data from gene ontology and pathway analyses suggests that top ranking probes are enriched for functions that are similarly enriched in other AID studies. 'Apoptotic pathway' and 'Wnt/Jnk signalling', enriched in this study, were also found to be a common dysregulated pathway in gene expression studies of AID including JIA [412]. In addition, an increase in apoptosis within lymphocytes and synovial cells has also been reported in JIA [426, 427]. Genes in these pathways may be genuine DMPs in this study, in which a more highly powered study may more likely detect and replicate.

4.4.1.1 Age

Another limitation is the possibility that oJIA is not a clinically homogeneous disease, despite the gold standard ILAR diagnosis criteria used to phenotype cases. Sample heterogeneity likely reduces study power and the effect sizes of any findings by diluting cases with 'non-cases' [428]. This heterogeneity is particularly relevant to age of diagnosis. Some differences exist in clinical course with younger aged JIA cases together with positive Anti-Nuclear Antibody (ANA) status [50]. ANA negative patients are thought to present with an older age of onset, together with a worse disease course, differing patterns of joints affected as well as higher rates of related eye disease [50]. Therefore age and ANA status has been placed at the centre of a debate on a new classification system for JIA [16, 17]. The potential relevance of age subgroups is highlighted by a number of other publications identifying associations with particular age groups. Specifically, using gene expression data found biological differences between younger onset and older onset oJIA and polyarticular JIA cases [12]. Network analysis also identified signalling pathways specific to older onset JIA cases [418]. In addition, HLA genetic associations specific to younger onset JIA have been reported [14].

Differences in age are also immunologically relevant. A number of cell types go through a change in abundance as age increases. In this respect, 6 years of age appears to be a relevant marker as after this age, cell numbers become highly variable. This includes T cells such as CD4+ effector, memory cells and naïve T cells [429]. In addition, memory Th1, Th2 and Th17 cells increase in number during childhood [430]. Further, there are also functional

differences according to age. Th2 responses appear to be stronger in childhood prior to age 6, after which the strength of Th1 responses increases [430]. This means that around ages 6 or below, the immunological cell repertoire and function is likely to be different than those of older children whose immune system has had time to develop and reach an adult-like state. Thus, CD4+ subsets appear to vary between these different age groups, and may contribute to the complexity and variability in methylation observed in this study. These differences may have confounded or reduced the sensitivity of the analysis in this study.

To support age as a source of sample heterogeneity, I observed a large source of variation within the top ranked DMPs of the RUV analysis. Using the top ranked DMPs from the RUV analysis with unadjusted p values <0.01, these probes were able to separate cases and controls. Yet the first component of variation explaining the data was age, not case/control status. In addition, there was a suggestion of clustering according to younger and older samples (

Figure 4-14). The observation of a strong age component in the data and the suggestion of age subgroups may reflect differing immunological or other biological processes according to age. To lessen the effect of sample heterogeneity on this study, stratified or age-specific analyses to reduce or remove age as a source of variation may be necessary and will be investigated in Chapter 6.

4.4.1.2 Genetics

Another limitation of this study is the inability to control for genetic factors affecting DNA methylation. The issue of the genetic contribution to DNA methylation variability is one currently being explored in the literature. The known mechanisms through which genetics may influence methylation is limited. Well recognised is allele specific methylation, where genetic variants may affect binding sites to effect DNAm differences [431]. In some instances, these are a number of CpG creating or destroying SNPs which may have a large effect on these [432]. While mechanisms are not entirely clear, there are a number of associations of methylation with local sequence up to 150kb from CpGs of interest, even up to 300kb in non-contiguous clusters [220, 222] [219].

The importance of genotype to methylation was highlighted by Liu *et al.* [245]. Liu *et al.* used a method to identify genotype dependent methylation in their Rheumatoid arthritis EWAS [245]. This data suggests that a list of thousands of significant DMPs can be reduced to just a

few hundred DMPs, mainly dependent on particular genotypes. This shows that genetics may indeed be a strong factor affecting methylation differences. Interestingly, controlling for genetics in AID EWAS' appears to affect the number of DMPs found. This is observed in publications where discordant twin pairs attempt to limit the effect of genetics on DNAm, which are summarised here. For example, earlier I'd noted the reduced number of DMPs found by publications where a number of confounders had been addressed, including Liu *et al.*, Bell *et al.*, Rakyan *et al.* and Hasler *et al.* (9, 19, 132 and 703 DMPs respectively) [245, 288, 295, 301]. Further, all of these studies compared twins discordant for disease or specified analyses that focussed on identifying genetics related differences in DNAm. The effect of genetics on methylation is lessened in these scenarios, and potentially also explains the reduced number of DMPs observed. Using IBD as an example where Hasler *et al.*'s discordant twin study is the comparator, studies from Cooke *et al.* and Nimmo *et al.* published 2714 and 3198 DMPs respectively compared to the 703 DMPs found in Hasler *et al.* [290, 295, 300]. Looking at Rheumatoid arthritis, Lui *et al.* controlled for genetic effects by limiting their study to genotype dependent DMPs, reducing the number of findings from thousands to just a handful of DMPs [245]. Their initial finding of thousands of DMPs mirrors the thousands found in other RA studies from de La Rica *et al.* and Glossop *et al.* ([291, 423]. These comparisons highlight the potential role genetics plays in determining methylation in AID, as well as potentially explaining many of the published findings in previous AID EWAS.

Genetics may also play a role in methylation variability. Liu *et al.*'s paper demonstrated that genotype may affect more than just differences in mean methylation. Liu *et al.* identified VMRs, or variably methylated regions, where the range of DNAm at particular genomic regions changes according to a particular genotype. Therefore, variation of DNAm at a particular region may be a result of underlying genetics. Genetics, then, is an important variable that may need particular attention in future EWAS studies. Further, it may explain the DNAm variation observed in this study's discovery data, and the inability to replicate data. The sampling error in a small replication cohort may have been biased towards a particular genotype, leading to differences in methylation between discovery and replication sample groups.

While many DMPs in EWAS may be explained by genetics, this does not exclude the methylation differences as irrelevant. It may mean genetics will need to be assessed as a potential causal factor in EWAS findings, where methylation may mediate the relationship between SNP and disease, as Liu *et al.* have discussed [245]. In addition, analyses may need

to adopt new ways of integrating genetic data into analysis of epigenetic data, or perhaps the design of epigenetics experiments. This is the suggestion of Liu *et al.* in 2015, who identified large clusters of correlated DNAm driven by genetics [222]. They suggest SNP driven correlated methylation blocks they've identified could be used to combine GWAS with targeted epigenome analysis, or be used in a similar fashion to LD blocks in SNP analyses for assisting custom array designs for EWAS.

Consequently, the potential genetic differences between groups and the potential variability in the oJIA sample group according to age, likely form part of a matrix of factors combining to suggest that the DNAm landscape is complex in oJIA. In order understand the complexity of DNAm at the particular sites targeted for validation, and to begin unravelling reasons for a failure to replicate the validated data, an investigation integrating genetic and epigenetic data of regions surrounding oJIA DMPs is prudent. Therefore, the following chapter will investigate this point.

Chapter 5. **Genetic associations with DNA methylation in oligoarticular JIA**

5.1 Introduction

EWAS findings in this study were not able to be replicated in a second, small sample group from the larger CLARITY study, despite successful technical validation. Also observed was heterogeneity in DNAm data, for example, large distributions of DNAm at individual CpGs and reversal of ∆beta values in replication studies. This suggested a complexity behind DNAm in oJIA. The successful technical validation suggests this complexity likely resides in biological factors, of which genetics is a potential source.

Genetic modulation of DNAm through methylation quantitative trait loci (mQTL) is a well-studied phenomenon, with pivotal studies in brains identifying 5-10% of CpGs having mQTLs associated [219, 220]. Further studies, across different cell types and interrogating larger numbers of CpGs, validated these initial observations and suggested the percentage of CpGs with an mQTL associated may more likely be between 2-10% [221-229].

MQTLs are known to be relevant to disease and may be relevant to oJIA. The link between mQTLs and disease susceptibility was hypothesised by Feinberg *et al*. early in the GWAS era and appeared to pre-empt the direction of research post-GWAS era [237]. GWAS in complex diseases resulted in a number of top ranking SNPs residing within non-protein coding regions of the genome [112, 238-240]. Studies of mQTLs, in part, have been used in attempts to overcome a lack of simple, protein modification based mechanistic explanations of GWAS hits [241]. Consequently, a number of mQTL studies suggest GWAS signals are indeed enriched within mQTLs, particularly in autoimmune diseases, cancer, chronic inflammatory diseases and neurological diseases such as schizophrenia [222, 226, 242-244]. This relationship to autoimmune diseases, and potentially to oJIA, may be relevant to this study. In particular, mQTLs may be a source of oJIA sample heterogeneity affecting DNAm, potentially making replication of DNAm findings difficult in this study. MQTLs may also identify new susceptibility risk loci in JIA.

5.1.1 Enrichment of mQTLs near CpGs.

Potential mQTLs will likely be in close proximity to CpGs of interest [221-229]. Whilst a large proportion of mQTLs appear to directly affect the CpG of interest [219, 228], mQTLs are highly enriched in regions ~10Kb from the CpG with most mQTLs often reported an average under +/-100Kb of the CpG of interest [221-229]. This average distance of 100kb changes little despite a number of publications searching up to 1Mb from the CpG [220, 224, 226].

Further, two seminal studies from Liu *et al*. and Gibbs *et al*., despite looking for mQTL associations genome wide, still suggested an overwhelming proportion of mQTLs (~80%) were under ~150kb and ~120Kb from the CpG respectively [219, 222]. Therefore interrogating CpGs for mQTLs within a window +/-100Kb will be important in this study.

5.1.2 Enhancers are highly relevant to DNA methylation and disease.

Specific regions of interest to interrogate for mQTLs are enhancers, which are known to be associated with DNAm and disease. Enhancers are particularly appropriate to interrogate due to associations both with autoimmune diseases and enrichment for mQTLs.

Enhancer regions are reported to be enriched for mQTLs across a number of tissue types, including T-cells from neonates [247]. Concerning disease, enhancers contain mQTLs correlated with breast cancer risk loci, colorectal cancer risk and are associated with glioblastoma survival [248-250]. Further, enhancer mQTLs are more strongly correlated with gene expression in breast cancer cell lines [248].

Of specific relevance to oJIA, enhancers may be especially pertinent to complex diseases and particularly autoimmune diseases. Data from Ernst *et al*. and ENCODE strongly suggested GWAS associations in complex diseases are enriched in enhancer regions [112, 149, 433]. Complementing these findings, Farh *et al*. identified enhancer regions as over-represented amongst top ranking SNPs from a large selection of complex autoimmune diseases [149]. Susceptibility loci from oJIA immunochip data itself appeared to be enriched for CD4+ T-cell enhancer regions [149]. Enhancers are therefore relevant to mQTLs, disease and particularly to autoimmune diseases, making them an ideal region to explore genetic sources of DNAm complexity in oJIA.

This chapter will explore SNPs very proximal to technically validated EWAS CpGs, as well as explore enhancer located SNPs within key genomic windows identified from the literature (+/-100kb from the CpG of interest). The aim will be to understand potential genetic sources of DNAm modulation, which may explain the DNAm heterogeneity observed and the difficulty in replicating EWAS data.

5.2 Methods

5.2.1 SNP selection

SNPs were investigated for effects on DNAm at CpGs identified in the EWAS (*CYP1A1*, *CHD5*, *TACSTD2* and *VPS53*- identified in Chapter 4.3.4). The three themes used to select SNPs for genotyping are described in Chapter 2.4.2. To outline, SNP were selected based on 3 themes: 1) known methylation quantitative trait loci (mQTLs) from the literature 2) Regional SNPs (+/-10kb) and 3) Distal SNPs located with enhancer regions.

Known mQTL were selected based on the literature described in section 2.4.2.1, as is regional SNP selection in section 2.4.2.2. Below is described the selection of enhancer SNPs.

To visualize enhancers surrounding the selected probes, UCSC genome browser (https://genome.ucsc.edu/) was used in conjunction with data from the BLUEPRINT epigenomics data track (www.blueprint-epigenome.eu/). Immune cell related enhancers were visualized including CD4+ T cells, CD8+ T cells, CD8+ memory T cells, B cells and monocytes.

Enhancer regions were selected if:

- H3K4me1 or H3K27ac enhancer marks were present AND
- Above marks present without the presence of transcription start sites marked by H3K4me3
- Present in CD4+ T cells AND
- Enriched in immune cells when comparing with non-immune cells. Non immune cells data was sourced using chromatin state data from ENCODE/BROAD, and included:
    - human umbilical vein endothelial cells (HUVEC)
    - mammary epithelial cells (HMEC)
    - skeletal muscle myoblasts (HSMM)
    - epidermal keratinocytes (NHEK) and
    - lung fibroblasts (NHLF) cell types

For each EWAS probe an enhancer was selected to genotype. Below are figures (figures 5.1-5.4) of regions surrounding the selected EWAS probes, highlighting the selected enhancer and data from the different cell types taken into account.

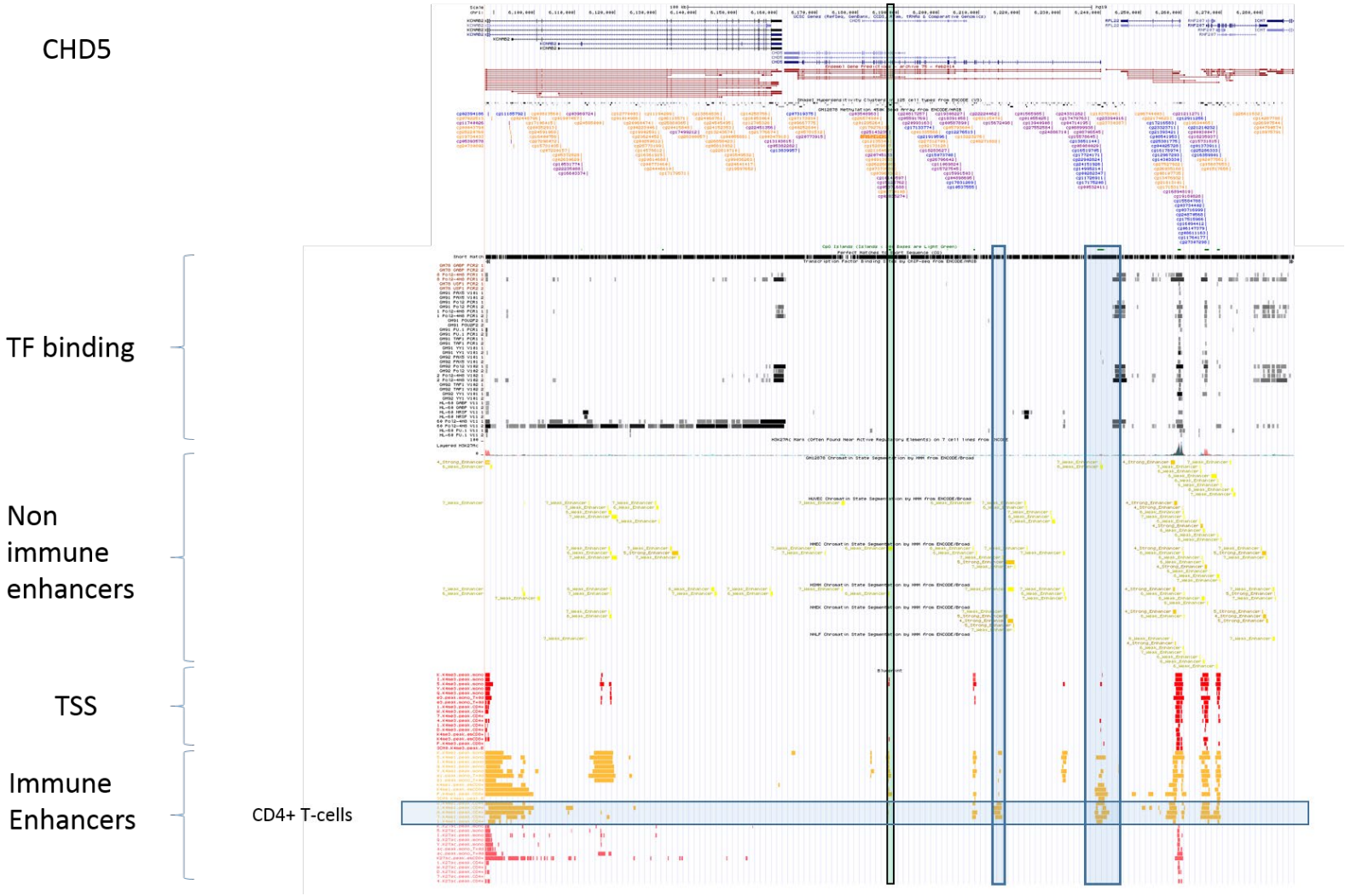

Figure 5-1 Visualisation of genomic region surrounding CHD5.

Region plotted is ~+/-100Kb from the probe of interest (green vertical narrow box). Blue vertical boxes show enhancers selected to interrogate for potential mQTLs. Enhancers for immune and non-immune cells are shown, and enhancers specific to CD4+ T cells are highlighted by the horizontal blue box. Immune cell enhancers are further divided into H3K4me1 (yellow) and H3K27ac (pink), as well as including transcription start sites (H3K4me3-red). Transcription factor binding sites (TF binding) are also shown. Data sourced from UCSC genome browser **[434]**.

Figure 5-2 Visualisation of genomic region surrounding CYP1A1.

Region plotted is ~+/-100Kb from the probe of interest (green vertical narrow box). Blue vertical boxes show enhancers selected to interrogate for potential mQTLs. Enhancers for immune and non-immune cells are shown, and enhancers specific to CD4+ T cells are highlighted by the horizontal blue box. Immune cell enhancers are further divided into H3K4me1 (yellow) and H3K27ac (pink), as well as including transcription start sites (H3K4me3-red). Transcription factor binding sites (TF binding) are also shown. Data sourced from UCSC genome browser **[434]**.

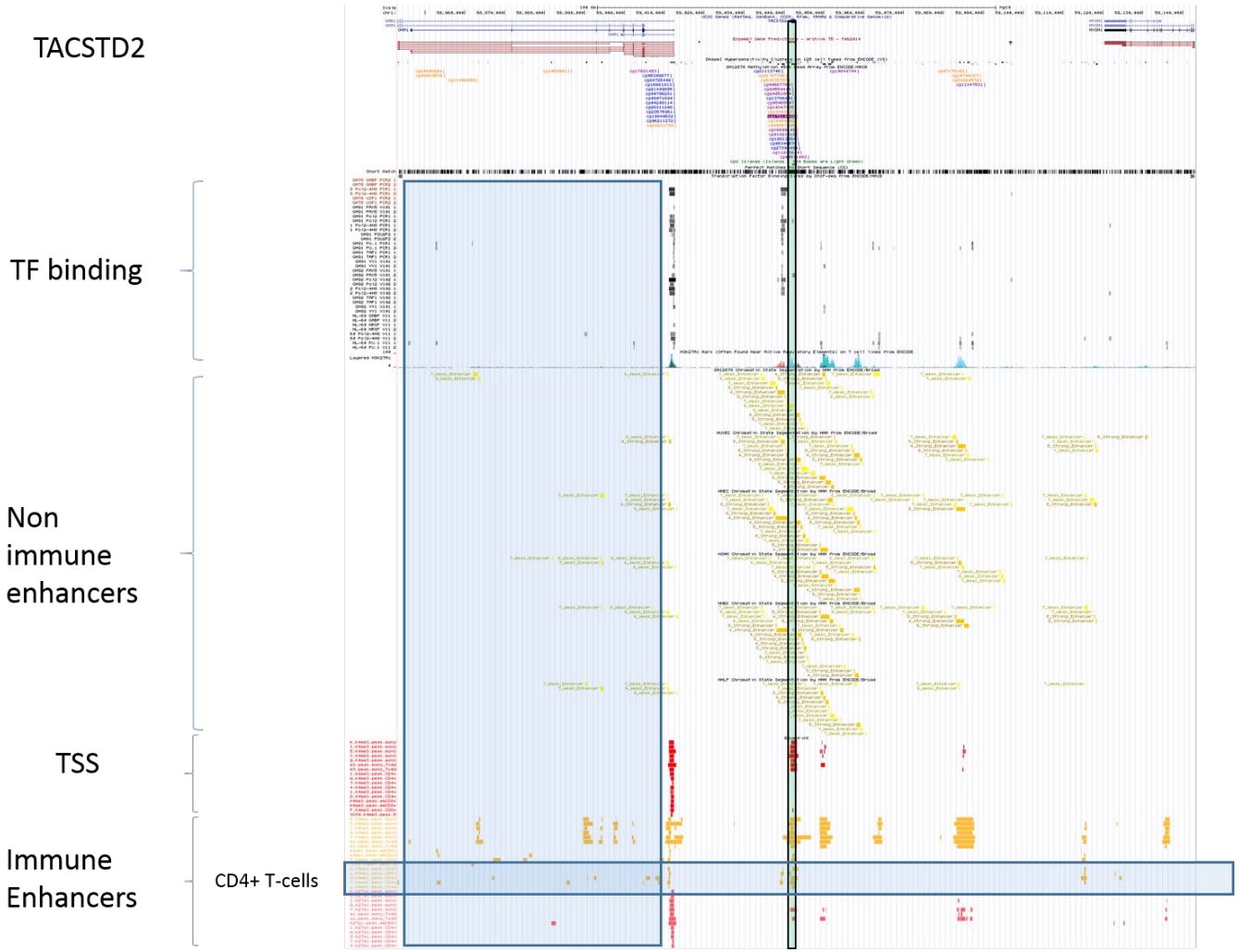

Figure 5-3 Visualisation of genomic region surrounding TACSTD2.

Region plotted is ~+/-100Kb from the probe of interest (green vertical narrow box). Blue vertical boxes show enhancers selected to interrogate for potential mQTLs. Enhancers for immune and non-immune cells are shown, and enhancers specific to CD4+ T cells are highlighted by the horizontal blue box. Immune cell enhancers are further divided into H3K4me1 (yellow) and H3K27ac (pink), as well as including transcription start sites (H3K4me3-red). Transcription factor binding sites (TF binding) are also shown. Data sourced from UCSC genome browser **[434]**.

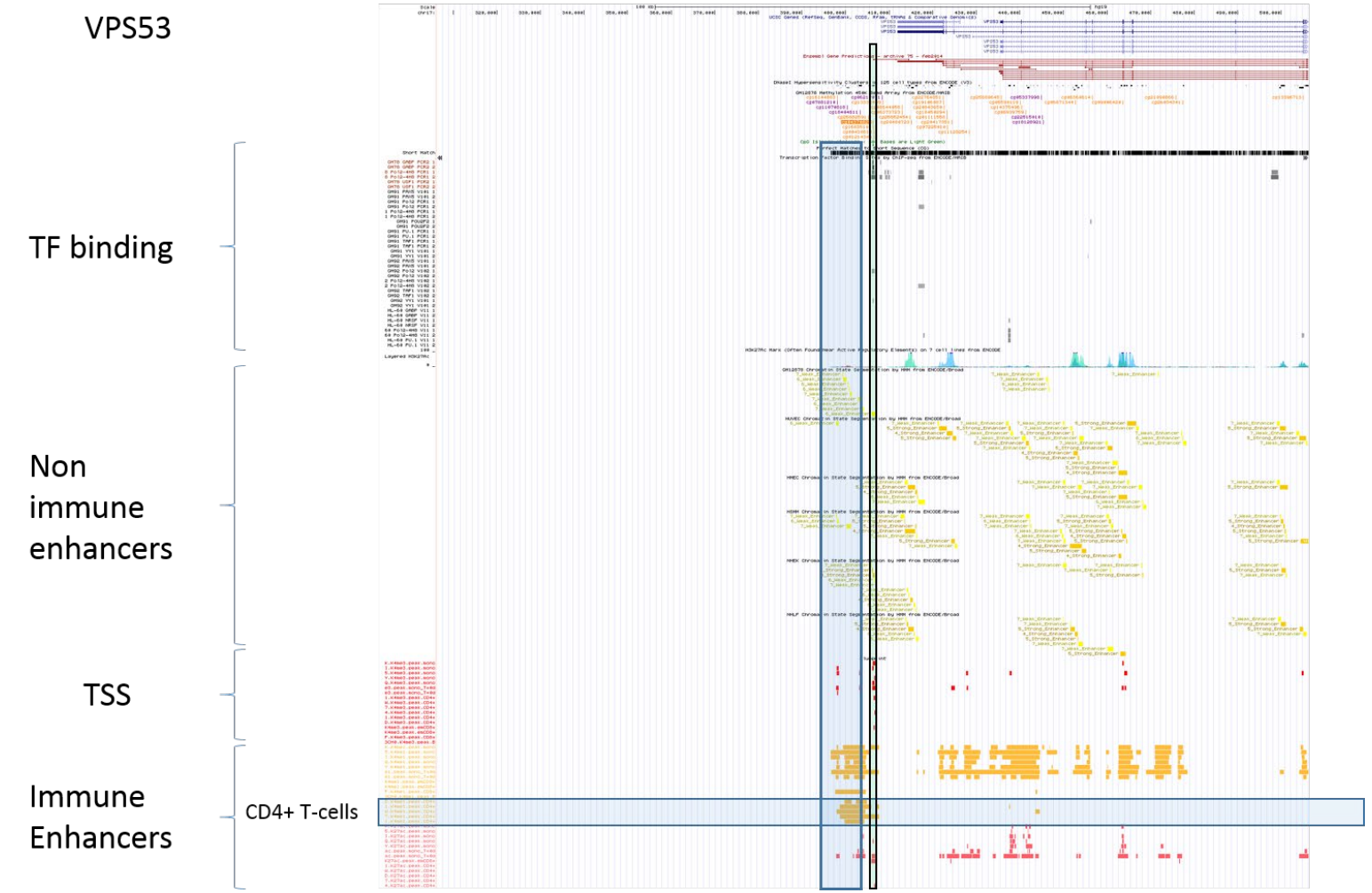

Figure 5-4 Visualisation of genomic region surrounding VPS53.

Region plotted is ~+/-100Kb from the probe of interest (green vertical narrow box). Blue vertical boxes show enhancers selected to interrogate for potential mQTLs. Enhancers for immune and non-immune cells are shown, and enhancers specific to CD4+ T cells are highlighted by the horizontal blue box. Immune cell enhancers are further divided into H3K4me1 (yellow) and H3K27ac (pink), as well as including transcription start sites (H3K4me3-red). Transcription factor binding sites (TF binding) are also shown. Data sourced from UCSC genome browser **[434]**.

5.3 Results

5.3.1 Known mQTLs are also observed in this study

To investigate potential sources of heterogeneity in oJIA DNAm, published mQTL associating with any of the 4 probes in this study (covering genes *CHD5*, *VPS53*, *TACSTD2* and *CYP1A1*-probes with the highest delta beta differences) were searched for in the literature. Published mQTLs were identified for two probes in this study (*CHD5* and *TACSTD2*) (Table 5-1). Source papers utilised EWAS platforms and cells types similar to this study i.e. HumanMethylation450 Bead chip arrays and use of immune cells. All samples including cases and controls were combined for mQTL analyses. In addition to investigating genetic sources of DNAm heterogeneity, these mQTLs would also be valuable for assessing the robustness of EWAS data from this study compared to published data. Since this study was limited in sample numbers to analysis, correction for multiple testing was not carried out. Instead, sensitivity analyses were performed to screen for the more robust associations. Controls only and females only were groups utilised for these sensitivity analyses.

All published mQTLs for *CHD5* (cg15205435) and *TACSTD2* (cg17210938) were found to have significant differences in DNAm between genotype groups, with strength of associations invariably below $p=2.8 \times 10^{-4}$. These SNPs are listed in Table 5-1.

Due to a limited sample size, sensitivity analyses were conducted to gain a better understanding of relationships between mQTLs and disease status or sex. These sensitivity analyses focussed on analysing the mQTL association in either controls only or females only. All mQTLs for *CHD5* were found to be associated in both sensitivity analyses (Table 5-1). Of note, two *CHD5* mQTLs with the most robust statistical significance in these tests were rs1883764 and rs2273041 ($\chi2(2)=27.1$ and 25.9, $p=1.32 \times 10^{-6}$ and 2.32×10^{-6} respectively in full sample analyses).

Regarding *TACSTD2* mQTLs, two SNPs stood out as strongly associating with DNAm through sensitivity analyses, rs11581395 and rs232847 ($\chi2(2)=71.4$ and 75.7, $p=3.1 \times 10^{-16}$ and 3.7×10^{-17} respectively).

Of note, known mQTLs were associated in the control (i.e. non-diseased) population, analogous to the study population of the source papers [222, 247]. Interestingly, p-values

increased when cases were a part of the sample group. This suggested the mQTL effect may be more pronounced in cases, to a point where these mQTLs may be susceptibility loci.

Post hoc analysis of the *CHD5* SNPs noted (rs1883764 and rs2273041) suggests it is the minor allele homozygotes that drive the association, with average DNAm reducing 13% compared to major allele homozygotes (Figure 5-5C-D). The global minor allele frequencies (MAF) are in the intermediate range (MAF=0.37 and 0.36 respectively). The presence of mQTLs at these frequencies may have impacted the EWAS replication study due to genotype sampling differences between replication and discovery sample.

Post-hoc analysis of robustly associated *TACSTD2* SNPs, rs11581395 and rs232847, suggest mean DNAm 23% and 27% respectively when minor allele homozygotes are compared to major allele homozygotes (Figure 5-5A-B), whilst the strongest association within each mQTL analysis belonged to the comparison of the heterozygote and major allele homozygotes. The high prevalence of the minor allele (MAF=0.29 and 0.25 respectively) again suggests these known mQTLs may have made replication of EWAS data difficult due to sampling bias affecting DNAm.

Table 5-1 Published mQTLs sourced from the literature and tested for association with DNAm in this study.

Gene	Probe	SNP	p-value	p-value Females	p-value Controls	Paper source
CHD5	cg12135344**	rs1883764	1.32×10^{-6}	4.53×10^{-5}	5.69×10^{-3}	Liu, Li *et al.* 2014 [222]
	cg15205435*	rs1883766	9.11×10^{-5}	9.87×10^{-4}	0.03	Gutierrez-Arcelus *et al.* 2013 [247]
	cg12135344**	rs2250358	6.06×10^{-6}	2.09×10^{-4}	0.01	Liu, Li *et al.* 2014 [222]
	cg12135344**	rs2273041	2.32×10^{-6}	5.62×10^{-5}	9.40×10^{-3}	Liu, Li *et al.* 2014 [222]
	cg17927613**	rs6697522	9.11×10^{-5}	9.87×10^{-4}	0.03	Gutierrez-Arcelus *et al.* 2013 [247]
	cg12135344**	rs9435080	3.34×10^{-6}	1.29×10^{-4}	0.01	Liu, Li *et al.* 2014 [222]
TACSTD2	cg17210938	rs10889097	**2.26×10^{-4}**	**1.08×10^{-3}**	0.42	Liu, Li *et al.* 2014 [222]
	cg17210938	rs11581395	3.10×10^{-16}	2.07×10^{-13}	1.74×10^{-7}	Liu, Li *et al.* 2014 [222]
	cg17210938	rs232847	3.65×10^{-17}	4.90×10^{-14}	1.13×10^{-7}	Gutierrez-Arcelus *et al.* 2013 [247]
	cg17210938	rs338228	5.85×10^{-13}	4.55×10^{-10}	6.67×10^{-5}	Liu, Li *et al.* 2014 [222]
	cg17210938	rs338230	**2.02×10^{-5}**	**4.00×10^{-4}**	0.18	Liu, Li *et al.* 2014 [222]
	cg17210938	rs6587830	2.84×10^{-4}	1.87×10^{-3}	0.04	Liu, Li *et al.* 2014 [222]
	cg17210938	rs6667614	8.50×10^{-13}	1.18×10^{-10}	1.07×10^{-5}	Liu, Li *et al.* 2014 [222]

Tests were performed using the Kruskal-Wallis method, with Chi-squared statistic used for p-value calculation. Grey highlighted SNPs represent significantly associated SNPs, which were also associated across sensitivity analyses. Bolded SNPs represent SNPs associated in the whole sample analysis but not replicated in both sensitivity analyses. Females from both cases and controls were used for the female only sensitivity analysis (P-value females only). Controls only sensitivity analysis contained both males and females, and excluded cases (P-value Controls only). * denotes the lead probe identified and validated in the EWAS analysis. ** denotes secondary probes identified within the same region in the EWAS

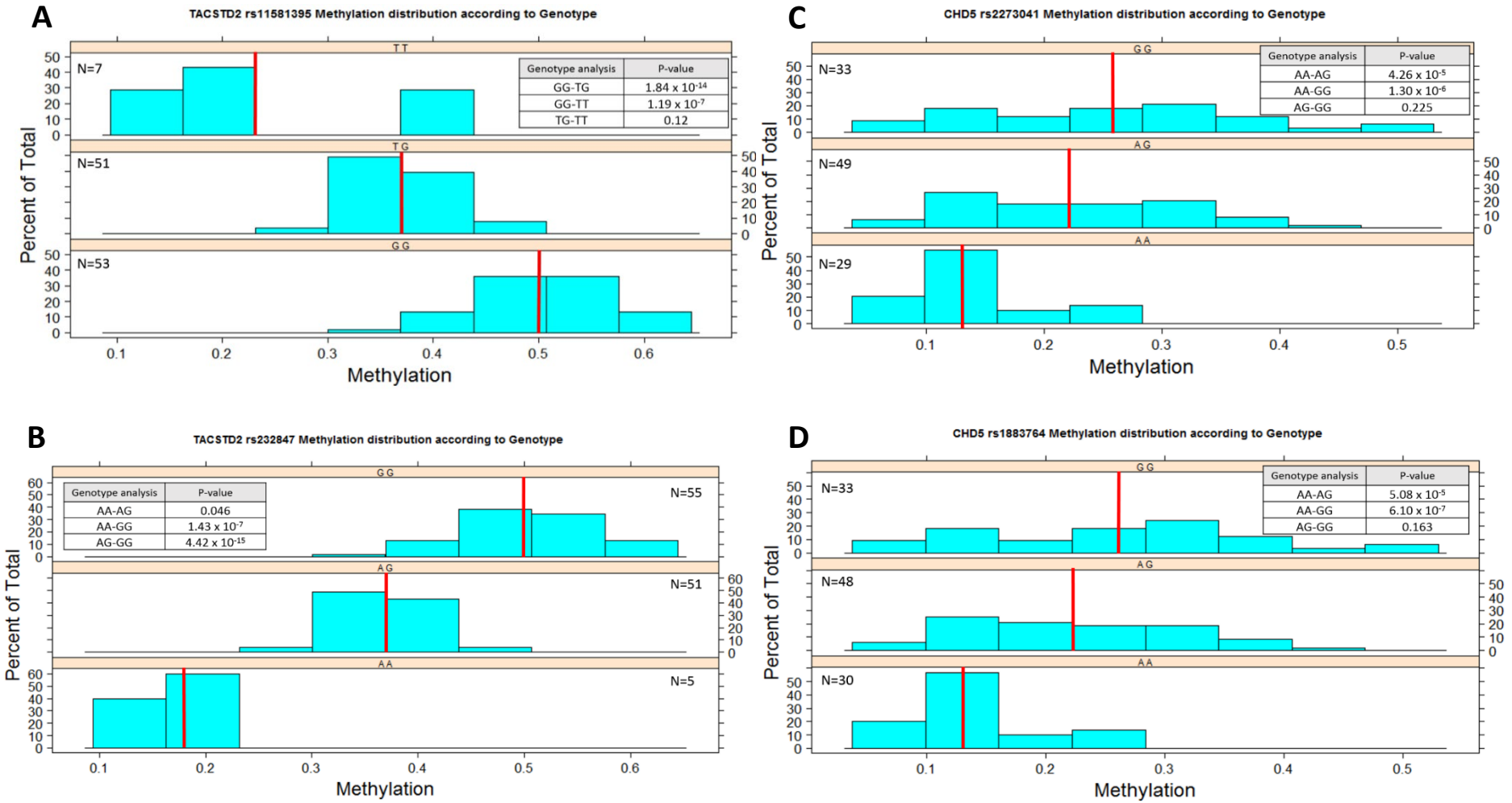

Figure 5-5 Histograms of DNAm distribution according to genotype.

The most robustly associated known mQTLs for TACSTD2 (a, b) and CHD5 (c, d) are presented. Each line containing a distribution of DNAm represents data from 1 genotype (e.g. homozygotes, or heterozygotes). Genotypes are noted in the cream coloured section. X-axis represents DNAm, separated into bins. The y-axis represents percent of sample data (within a bin). Red lines indicates the mean DNAm for that genotype. N= No. of samples within that genotype. Post-Hoc analysis of individual genotypes (e.g. genotype 1 vs genotype 2) are shown in the table within each histogram. P-value within the table refers to Dunn's test.

5.3.2 New mQTLs are commonly observed in enhancers.

5.3.2.1 Introduction

Known mQTLs were observed as a potential source of DNAm heterogeneity for 2 probes of interest; however, these and the remaining probes may yet be associated with unidentified mQTLs and potentially specific to this population or enriched within oJIA cases.

In particular, SNPs from other genetic regions may also contain mQTLs. Indeed, Feinberg *et al.* in 2010 alluded to regulatory sites housing mQTLs [236]. Evidence does indeed suggest SNPs modulating DNAm may located or even enriched within regulatory regions, particularly enhancers [247-250]. Enhancer located SNPs may partly explain the DNAm heterogeneity observed in this study, and identify new mQTLs relevant to oJIA.

Therefore, enhancers enriched within immune cells were investigated (as described in the methods section 5.2.1). Whilst mQTLs are enriched in proximal areas, almost all mQTLs can be observed +/-100kb from the CpG [219, 221-229]. Since enhancers can exert effects from large distances, enhancer SNPs within this window were selected for mQTL analysis.

In addition, since mQTLs are known to be highly enriched at regions very proximal to CpGs, SNPs within +/-10Kb were selected for investigation [221-229].

5.3.2.2 CHD5

Evidence for new mQTLs was found within enhancers for all probes except probe cg12101586 (*CYP1A1*). An enhancer 30kb upstream of *CHD5* was interrogated, from which 7 SNPs were associated with DNAm (Table 5-2). To assess the robustness of associations, sensitivity analyses were carried out utilising only controls or only females (the groups with largest sample sizes). All 7 SNPs also survived sensitivity analyses. The most robust associations belonged to rs11121295, rs9434662 and rs12565328 ($p= 1.0 \times 10^{-6}$, $p= 1.0 \times 10^{-6}$, $p= 1.6 \times 10^{-14}$ respectively in full sample analyses).

Post-Hoc analyses of the 3 most robustly associated SNPs suggest it is the gain of minor alleles that results in higher average DNAm (Table 5-3). Every minor allele increases average DNAm between 3-12%. The biggest change in DNAm occurred when a single minor allele was added to the major allele (rs11121295, rs9434662 and rs12565328 p-values are $p= 5.1 \times 10^{-7}$ (9.8% change), 5.1×10^{-7} (9.8% change) and 5.3×10^{-11} (12.6% change) respectively for tests comparing major allele homozygotes and heterozygotes). For all SNPs in post Hoc

analysis, the presence of 2 minor alleles increased standard deviations of DNAm more so than average DNAm (minor allele homozygotes 14% vs. 9% in major allele homozygotes). This suggests increased variance may be another feature of these mQTL associations.

In contrast to enhancer located SNPs, SNPs located within a 10Kb region of the probe were not associated with DNAm. This suggests that the association with DNAm is not mediated by the proximity of SNPs to the probe, rather the location of SNPs within the enhancer region itself.

5.3.2.3 VPS53

A region nearby the probe cg04370829 (*VPS53*) was also interrogated, capturing an enhancer ~20Kb 5' of the probe, as well as interrogating proximal SNPs +/-10Kb of the probe. A single SNP, rs4968169, associated with DNAm and remained so throughout sensitivity analyses (p= 0.004). Mirroring results from the previous probes interrogated, this SNP was also located within an enhancer region. Another SNP within the enhancer was also close to passing sensitivity analyses (rs11247557). However, SNPs outside the enhancer did not associate with DNAm, suggesting it is the location within the enhancer that may be driving the mQTL association. Indeed, since not all SNPs in the enhancer associate with DNAm, it may be specific areas within the enhancer that may drive this association and modulate DNAm.

Post-hoc analyses of genotype pairs suggested associations in rs4968169 were driven by addition of single minor allele (p= 0.004, heterozygote vs. major allele homozygotes). A single minor allele appeared to reduce average DNAm by ~7%. The addition of an extra minor allele did not have any extra additive effect on mean DNAm. However, as with *CHD5* SNPs, it appears that the presence of 2 minor alleles (minor allele homozygotes) appear to increase the variance of DNAm between individuals with the same genotype (1 standard deviation = 21% in minor allele homozygotes vs. 8% in major allele homozygotes).

5.3.2.4 TACSTD2

To investigate SNPs relating to *TACSTD2* probe cg17210938, SNPs from 100kb upstream to 10Kb downstream were analysed. The upstream region contains a number of interspersed enhancer regions, and due to the high linkage disequilibrium (LD- a measure of co-inheritance) in the region it was possible to select a minimal number of SNPs to interrogate the whole region.

Two SNPs were associated with DNAm, rs6666304 and rs6694699, and these associations remained through two sensitivity analyses ($p= 1.5 \times 10^{-11}$, p= 0.01 respectively in full sample analyses) (Table *5-2*). Two other SNPs also showed strong evidence for DNAm differences according to genotype, rs232828 and rs11581844 (p= 0.01 and $p= 2.6 \times 10^{-5}$ respectively). However, these associations did not remain for the control only analysis, suggesting cases are potentially driving this association.

Interestingly, similar to *CHD5* and *VPS53*, *TACSTD2* SNPs were also linked to enhancer regions. Known *TACSTD2* mQTLs, also strongly associated in this study, were located within the enhancer rich region and tagged many SNPs throughout this region due to high LD. In addition, two of the above SNPs, r66694699 and rs11581844 were either located in an enhancer or nestled within an enhancer rich area, respectively.

In contrast to enhancer linked SNPs, SNPs proximal to the CpG of interest and associating with DNAm were limited. This also mirrored the findings in *VPS53* and *CHD5*.

Post Hoc analyses of pairwise comparisons suggests it is addition of minor alleles that hypomethylates DNA. The strongest associations belonged to the heterozygote vs. major allele homozygotes of rs6666304 and rs6694699 ($p= 2.5 \times 10^{-11}$ and p = 0.01 respectively), where average DNAm is 13% and 7% lower in heterozygotes. However, minor allele homozygotes were present in very low proportions, making this analysis difficult. This also limited insight into the ability of minor allele homozygotes to increase variance of DNAm, which was observed for *CHD5* and *VPS53* SNPs.

Again, much like the known mQTLs, p-values appeared to increase with oJIA cases present within the sample group (i.e. compared to the controls only analysis). This interesting link to oJIA suggests potential mQTLs identified here could also be oJIA susceptibility loci.

Table 5-2 Potentially new mQTLs identified in this study for CYP1A1, CHD5, TACSTD2 and VPS53.

Gene	Probe	Gene region	SNP	p-value	p-value (Fem)	p-value (Cont.)
CHD5	cg15205435	Enhancer	rs11121295	1.02×10^{-6}	2.15×10^{-5}	1.58×10^{-3}
			rs12067480	2.00×10^{-5}	1.39×10^{-4}	0.01
			rs12565328	1.63×10^{-14}	1.01×10^{-10}	1.30×10^{-7}
			rs17436816	**0.03**	0.19	0.14
			rs17489787	0.33	0.27	0.75
			rs1883604	1.25×10^{-3}	0.01	0.02
			rs1883767	0.40	0.50	0.11
			rs2485024	1.44×10^{-4}	9.89×10^{-4}	0.04
			rs4908529	0.90	0.68	0.60
			rs6495121	0.25	0.61	0.63
			rs747393	0.99	0.70	0.74
			rs7523555	**0.05**	0.29	0.14
			rs9434662	1.02×10^{-6}	2.15×10^{-5}	1.58×10^{-3}
			rs9435102	4.63×10^{-7}	1.36×10^{-5}	3.77×10^{-3}
		Regional	rs17029184	2.52×10^{-4}	7.65×10^{-4}	0.02
			rs2273033	0.13	0.48	0.96
			rs2746064	0.16	0.17	0.11
			rs2843494	0.92	0.92	0.53
			rs731975	0.95	0.95	0.62
			rs9434661	0.12	0.16	0.08
CYP1A1	cg12101586	Enhancer	rs2273032	0.93	0.67	0.78
			rs9434741	0.30	0.64	0.71
		Regional	rs1048943	0.88	0.77	0.67
			rs2229150	0.68	0.54	0.49
			rs2470893	0.91	0.51	0.58
			rs2472297	0.64	0.55	0.79
			rs4886605	0.83	0.39	0.73
			rs4986885	0.85	0.71	NA
			rs12441817	0.69	0.87	0.64

Table 5-2 Continued. Potentially new mQTLs identified in this study for CYP1A1, CHD5, TACSTD2 and VPS53.

TACSTD2	cg17210938	Regional	rs10493238	0.19	0.25	NA
			rs11581844	**2.56×10^{-5}**	**3.71×10^{-4}**	0.17
			rs12121124	0.12	0.15	0.23
			rs17117850	0.68	0.43	0.90
			rs17117853	0.35	0.06	0.08
			rs232828	**0.01**	**0.03**	0.66
			rs232840	0.33	0.12	0.17
			rs6666304	1.47×10^{-11}	3.47×10^{-9}	1.98×10^{-5}
			rs6683669	0.13	0.28	0.44
		Enhancer	rs6694699	0.01	9.51×10^{-3}	0.02
VPS53	cg04370829	Enhancer	**rs11247557**	**3.80×10^{-3}**	**0.02**	0.07
			rs2289621	0.11	0.37	0.10
			rs4968169	4.76×10^{-3}	2.75×10^{-5}	0.03
		Regional	**rs11649837**	**0.03**	0.19	0.40
			rs4286157	0.89	0.70	0.83
			rs759974	0.42	0.37	0.81
			rs9748016	0.24	0.05	0.80
			rs9895280	0.81	0.70	0.76
			rs9913674	0.06	0.34	0.37
		probe SNP	rs72477025	0.70	0.86	0.60

Each gene had SNPs investigated for potentially new mQTL from regional areas (+/-10kb from the probe listed), enhancer regions (+/- 100kb of the probe). Only VPS53 had a probe SNP was investigated, since probe SNPs for other probes had MAF's below 0.05 (thus unlikely to be observed in this preliminary study). Tests were performed using the Kruskal-Wallis method, with Chi-squared statistic used for p-value calculation. Grey highlighted SNPs represent significantly associated SNPs, which were also associated across sensitivity analyses. Bolded SNPs represent SNPs associated in the whole sample analysis but not replicated in both sensitivity analyses. Females from both cases and controls were used for the female only sensitivity analysis ("P-value females"). Controls only sensitivity analysis contained both males and females, and excluded cases ("P-value Controls"). NA: insufficient numbers in genotype groups to analyse.

Table 5-3 The most robustly associated SNPs were selected for post-hoc analysis.

Gene	SNP (genomic region)	Genotypes			Tests	
		Genotype	**N**	**Mean methylation**	**Genotypes tested**	**p-value**
CHD5	rs11121295 (enhancer)	AA	4	0.30	AA-AG	0.89
		AG	34	0.28	**AA-GG**	**0.03**
		GG	73	0.17	**AG-GG**	**5.06×10^{-7}**
	rs12565328 (enhancer)	CC	13	0.35	**CC-CT**	**0.045**
		CT	46	0.26	**CC-TT**	**2.72×10^{-10}**
		TT	52	0.13	**CT-TT**	**5.32×10^{-11}**
	rs9434662 (enhancer)	CC	73	0.17	**CC-TC**	**5.06×10^{-7}**
		TC	34	0.28	**CC-TT**	**0.03**
		TT	4	0.30	TC-TT	0.89
VPS53	rs4968169 (enhancer)	AA	12	0.60	AA-AG	0.32
		AG	46	0.69	**AA-GG**	**0.005**
		GG	51	0.75	**AG-GG**	**0.004**
TACSTD2	rs6666304 (regional SNP)	AA	2	0.21	AA-AG	0.29
		AG	37	0.35	**AA-GG**	**0.003**
		GG	72	0.48	**AG-GG**	**2.46×10^{-11}**
	rs6694699 (enhancer)	CC	1	0.20	CC-CG	0.30
		CG	17	0.37	CC-GG	0.08
		GG	93	0.44	CG-GG	0.01

This table includes SNPs that were associated in the whole sample analysis, as well as the sensitivity analyses (using females or controls only). Post-hoc analysis of individual genotypes was completed using Dunn's test. Genotype tests in bold are significant at $p<0.05$.

5.3.3 Variation in DNAm is potentially driven by cases

5.3.3.1 Introduction

DNAm variance, as opposed to changes in mean DNAm, has been hypothesised as mechanism for mQTL function. Indeed, two studies observed inter-individual differences were greater at mQTL-associated CpGs compared to other CpGs in immune cells [247, 435]. Feinberg *et al.* was arguably the first to suggest that genetic variants could explain the increased variability of DNAm, and produced data from cancer revealing genetic underpinnings for large variations in DNAm [236]. The same group published data from a large rheumatoid arthritis EWAS that provided further supporting evidence for this hypothesis, where 4 of the top 9 mQTLs associating with rheumatoid arthritis modulated DNAm variance [236, 245]. This also demonstrated the relevance of this hypothesis to autoimmune diseases.

Feinberg *et al.* identified genetic variation occurring in or near regulatory regions could be part of the overall mechanism behind DNA variation (and consequently, phenotypic effects) [236]. MQTLs located within regulatory regions such as enhancers are analogous to Feinberg's hypothesis and initial findings in cancer [236]. In this oJIA study, a number of potential mQTLs were enhancer linked, and there were already some hints of increased DNAm variation during post-hoc analysis of these potential mQTLs. This could be explained by Feinberg's hypothesis of genetically linked DNAm variation.

To understand if genotypes associate with changes in the variation of DNAm (vSNPs), SNPs were analysed for effects on DNAm variance. SNPs modulating DNAm variation, should any be identified in this study, could provide some explanation for earlier difficulties in replicating EWAS data as well as identify new potential mQTLs.

The same SNPs were utilised for this as for earlier analyses, which included known mQTLs, enhancer located SNPs and SNPs proximal to the probes of interest (+/-10kb).

5.3.3.2 Enhancer regions contain the most robust vSNPs

It was interesting to note that enhancers again featured as having significant associations with DNAm variance. Robust findings, where associations held in both sensitivity analyses, were only found in SNPs from enhancer regions (*CHD5*-rs12565328 p= 2.9 x10^{-3}, *TACSTD2*-rs6694699 p= 2.0 x10^{-3}) (Table 5-4).

Density plots for the robustly associated *CHD5* SNP rs12565328 suggest the addition of minor alleles increases the spread of DNAm value towards 50% (Figure 5-6B). In addition to the robustly associated *TACSTD2* SNP rs6694699, another *TACSTD2* SNP was observed with strong and suggestive evidence for female and control sensitivity analyses respectively (rs6683669, p= 0.04 in all samples analysis), see also Figure 5-6E. Density plots of DNAm as a function of rs6694699 genotypes suggest that the minor allele increases the spread of DNAm to both higher and lower values (Figure 5-6A).

Again, no SNPs from *CYP1A1* CpG site cg12101586 survived sensitivity analyses; however, two *VPS53* SNPs presented with strong and suggestive evidence for female and control sensitivity analyses respectively (rs759974 and rs9913674, p= 0.003 and p=0.002 respectively for all samples analysis). Minor alleles from both these SNPs appear to associate with distributing DNAm towards lower values (Figure 5-6C-D).

Table 5-4 DNAm variance analysis suggest enhancers and known mQTLs potentially contribute to DNAm variation.

Gene	Probe	Gene region	SNP	p-value	p-value (Females)	p-value (Controls)
CHD5	cg15205435*	Enhancer	rs11121295	0.84	0.79	0.77
			rs12067480	0.25	0.48	0.07
			rs12565328	0.003	0.02	0.047
			rs17436816	0.24	0.29	0.26
			rs17489787	0.34	0.41	0.63
			rs1883604	0.30	0.25	0.28
			rs1883767	0.51	0.82	0.10
			rs2485024	0.13	0.39	0.62
			rs4908529	0.94	0.92	0.87
			rs6495121	0.64	0.47	0.34
			rs747393	0.52	0.91	0.99
			rs7523555	0.25	0.20	0.32
			rs9434662	0.84	0.79	0.77
			rs9435102	**0.003**	**0.006**	0.55
		Regional	rs17029184	0.10	0.09	0.048
			rs2273033	0.85	0.78	0.58
			rs2746064	0.31	0.25	0.29
			rs2843494	0.37	0.24	0.62
			rs731975	0.99	0.86	0.83
			rs9434661	0.55	0.53	0.51
		Known mQTL	rs1883766	0.16	0.26	0.06
	cg17927613**		rs6697522	0.16	0.26	0.06
	cg12135344**		**rs1883764**	**3.96×10^{-4}**	**0.002**	0.40
	cg12135344**		**rs2250358**	**0.003**	**0.01**	0.26

Table 5-4 Continued. DNAm variance analysis suggest enhancers and known mQTLs potentially contribute to DNAm variation.

	cg12135344** cg12135344**	Known mQTL	**rs2273041** **rs9435080**	**3.94 x 10^{-4}** **0.002**	**0.001** **0.007**	0.46 0.26
CYP1A1	cg12101586	Enhancer	rs2273032	0.87	NA	0.96
			rs9434741	0.62	0.69	0.82
		Regional	rs1048943	0.23	0.71	0.33
			rs2229150	NA	NA	NA
			rs2470893	0.91	0.70	0.60
			rs2472297	0.14	NA	NA
			rs4886605	**0.02**	**0.02**	0.34
			rs4986885	NA	NA	NA
			rs12441817	0.17	0.43	0.07
TACSTD2	cg17210938	Regional	**rs10493238**	**0.03**	**0.03**	NA
			rs11581844	0.05	0.04	0.58
			rs12121124	0.27	0.13	0.20
			rs17117850	0.09	0.08	0.10
			rs17117853	0.11	0.01	0.08
			rs232828	0.10	0.07	0.21
			rs232840	0.06	0.13	0.05
			rs6666304	**0.03**	0.15	**0.03**
			rs6683669	**0.04**	0.06	**0.006**
		Enhancer	rs6694699	0.003	0.045	0.03
		Known mQTL	rs10889097	0.06	0.02	0.15
			rs11581395	**0.01**	**0.003**	0.15
			rs232847	**0.02**	**0.03**	0.09
			rs338228	0.21	0.18	0.40
			rs338230	**0.049**	**0.03**	0.46
			rs6587830	0.40	0.35	0.63

Table 5-4 Continued. DNAm variance analysis suggest enhancers and known mQTLs potentially contribute to DNAm variation.

	cg17210938	Known mQTL	**rs6667614**	**0.04**	0.14	0.25
VPS53	cg04370829	Enhancer	**rs11247557**	**0.004**	0.08	0.06
			rs9913674	**0.002**	**0.02**	0.09
	cg04370829 (cont.)	Enhancer (cont.)	**rs759974**	**0.004**	**0.01**	0.10
			rs4286157	0.65	0.61	0.48
			rs4968169	**0.003**	**0.045**	0.19
		Regional	rs11649837	0.10	0.16	0.42
			rs2289621	0.05	0.04	0.47
			rs9748016	0.04	0.14	0.48
			rs9895280	0.51	0.44	0.77
		Probe SNP	rs72477025	0.61	0.54	0.45

Published mQTLs, as well as SNPs from regional areas (+/-10kb from the probe listed), enhancer regions (+/- 100kb of the probe) were investigated for association with DNAm variance. P-values were calculated from Bartlett's test for normally distributed data (CYP1A1 only). For non-normally distributed data, Levene's test was used (CHD5, TACTSTD2, VPS53). Grey highlighted SNPs represent significantly associated SNPs, which were also associated across sensitivity analyses. Bolded SNPs represent SNPs associated in the whole sample analysis but not replicated in both sensitivity analyses. Females from both cases and controls were used for the female only sensitivity analysis ("P-value females"). Controls only sensitivity analysis contained both males and females, and excluded cases ("P-value Controls"). Only VPS53 had a probe SNP investigated, since probe SNPs for other probes had MAF's below 0.05 (thus unlikely to be observed in this preliminary study). NA: insufficient numbers in genotype groups to analyse. * denotes the lead probe identified and validated in the EWAS analysis. ** denotes secondary probes identified within the same region in the EWAS.

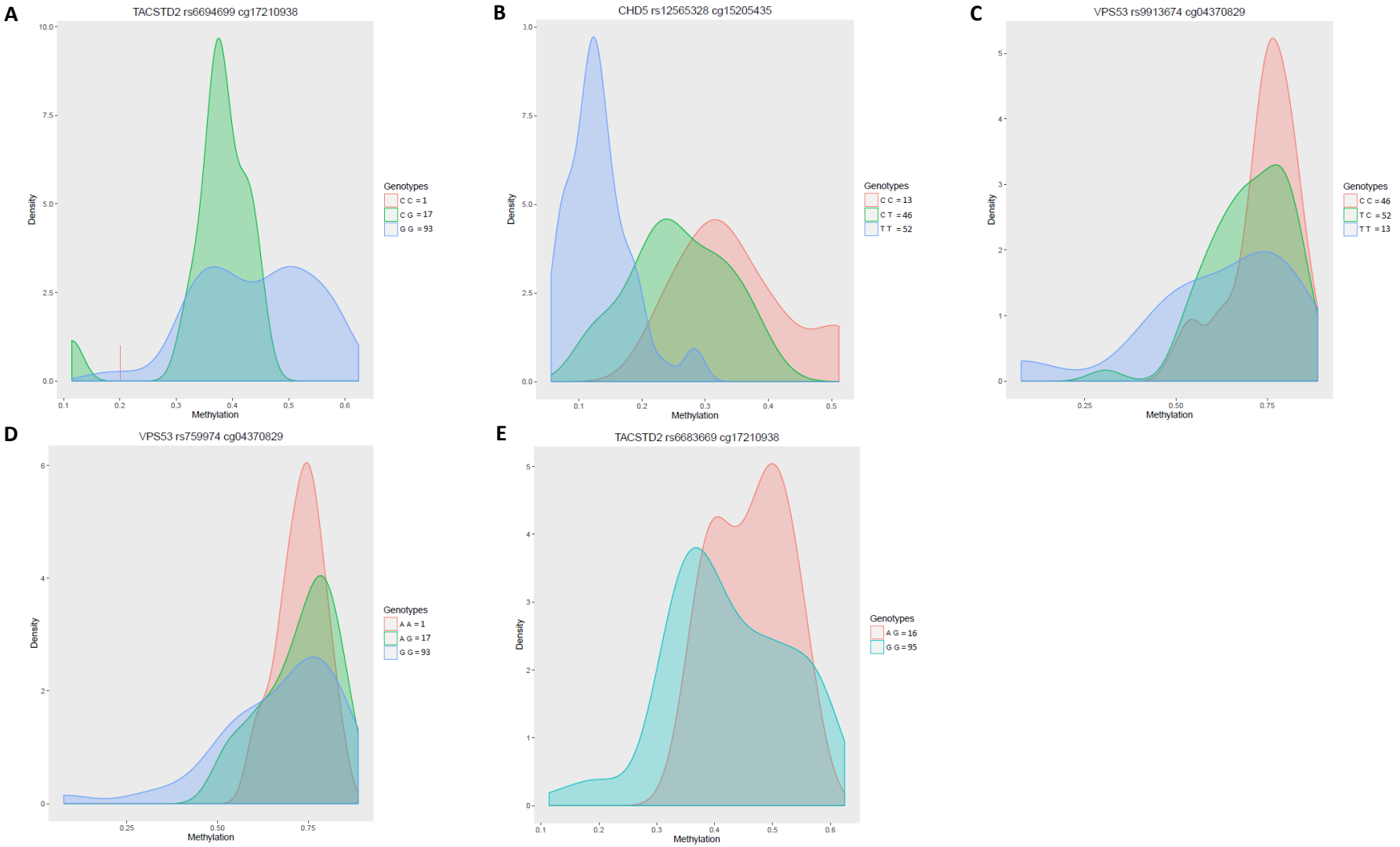

Figure 5-6 Density plots of SNPs potentially associating with DNAm variance.

Density plots in A) and B) are robustly associated SNPs, surviving both sensitivity analyses (i.e. females only and controls only analyses). Density plots in C) – E) are SNPs with significant associations in the full sample alongside 1 sensitivity analysis (and suggestive association in the other sensitivity analysis). SNPs in A) and B) are both located within enhancers.

5.3.3.3 Cases potentially mediate DNAm variation association

Interestingly, a number of other SNPs were associated with DNAm in the full sample and female only analyses, but did not replicate the association in the control only sensitivity analysis. That is, when cases were not present (e.g. the control only analysis), a number of SNPs did not associate with DNAm. This suggested that cases were mediating the association between genotype and DNAm. Thirteen SNPs were observed with this pattern of case driven DNAm variation. Again, enhancer SNPs appear to associate with DNAm in this context also. Seven of thirteen (7/13) SNPs (including known mQTLs which tag SNPs in enhancer region) followed this pattern, with another two enhancer SNPs narrowly missing significance.

This pattern of cases potentially driving associations with DNAm variance was observed more often in known mQTLs from *CHD5* and *TACSTD2* (e.g. rs2273041 $p=3.94 \times 10^{-4}$, rs11581395 p=0.01 being the strongest associations, respectively) (Figure 5-7A & D & Table 5-4). Indeed, 4/6 and 3/7 known mQTLs from *CHD5* and *TACSTD2* (with 1 other SNP narrowly missing significance at the p=0.05 level), respectively, were observed to have followed this pattern. The strongest associations for *CYP1A1* were rs4886605 (p=0.02) and p=0.002 for *VPS53* rs9913674 (Table 5-4).

Larger variation in *CHD5* DNAm generally appears to result from loss of minor alleles. The 3 most robustly associated SNPs are shown in Figure 5-7A-C. Whilst the pattern of variation in highly similar between all known *CHD5* mQTLs tested, this is incompletely explained by high LD amongst all SNPs. An LD of $r^2=0.95$ exists in European populations between rs2273041 and rs1883764; however, LD was less than $r^2=0.58$ between rs9435102 and the other SNPs [436].

One way to validate the potential for cases mediating the genotype x DNAm association is to undertake a case-control association analysis of alleles.

OJIA cases are over-represented by females in this study (~4:1), mirroring the female bias observed in the population [53]. If female cases do contain more DNAm variation, higher DNAm variation would likely be observed in the female only analysis. This is indeed what is observed in this study. Should cases truly influence DNAm, it may be possible to identify this association through an allelic association study focussed on cases vs. controls. This is the focus of the following section.

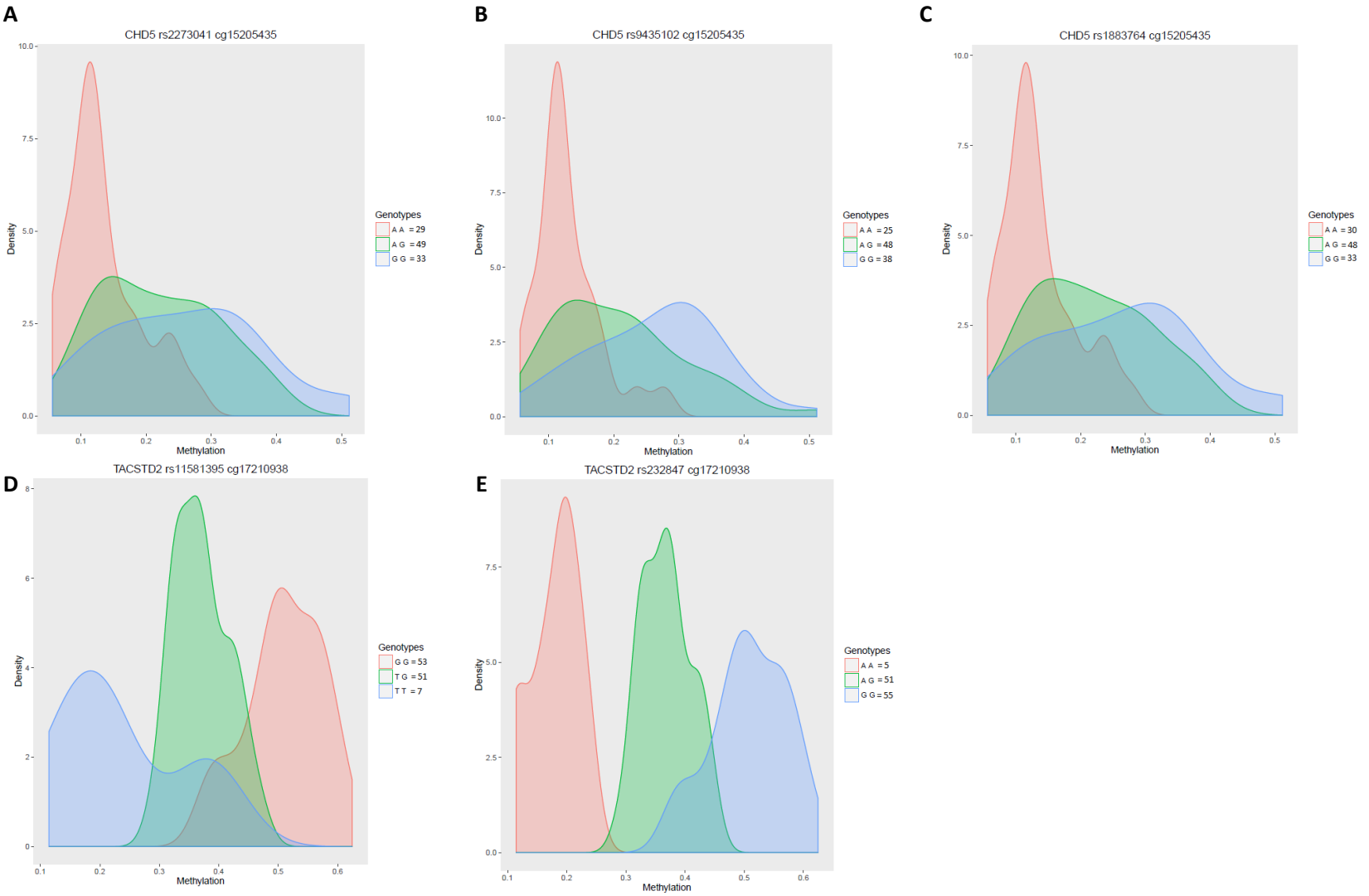

Figure 5-7 Density plots of SNPs with potentially case driven DNAm variance.

Density plots of SNPs associated with DNAm variance in all samples and females only, but not in the controls only analysis. Density plots in A) - C) are of CHD5 SNPs. Density plots in D) and E) are TACSTD2 SNPs. SNPs in A) and B) are both located within enhancers.

5.3.4 Top ranked oJIA EWAS probes point to new potential SNPs associating with oJIA

5.3.4.1 Introduction

The observation of 1) vSNP associations potentially driven by cases, as well as 2) enhancer located SNPs robustly associating with DNAm variation, tie in with previous hypotheses from Feinberg *et al.* [236]. He suggested that genetic variants increasing DNAm variation would potentially be located within regulatory regions, and would also increase susceptibility to disease (particularly complex diseases) as well as increase plasticity in developmental biology [236]. This hypothesis has been validated in a study of neonatal methylomes, a rheumatoid arthritis GWAS/EWAS and in populations at risk for chronic hepatitis B infection [231, 245, 246]. Therefore, potential vSNPs identified in this study were investigated for associations with disease to understand if these SNPs are important to disease susceptibility. This in turn may help explain the difficulty in replicating EWAS data i.e. oJIA sample heterogeneity introduced through vSNPs.

In addition, identified EWAS CpGs in this study may be a proxy for genetic variants of relevance to oJIA susceptibility. Indeed, analyses containing oJIA samples were observed to have stronger p-values compared to controls only analyses, a pattern observed across both known and potentially new mQTLs (sections 5.3.1-2), and in analyses of both difference in mean DNAm and DNAm variance. Therefore these SNPs may be particularly relevant to oJIA susceptibility. Again, oJIA sample heterogeneity leading to DNAm heterogeneity may explain some difficulties replicating EWAS data.

5.3.4.2 Suggestive evidence of genetically driven variable DNAm potentially predisposing to risk of oJIA.

Cases were potentially driving associations with DNAm variance (vSNPs) i.e. SNPs were associated with all samples and the female only sample group, but not the control only group). To understand if these vSNPs are relevant to oJIA, case-control allelic association tests were conducted with these SNPs identified in section 5.3.3.

Case-control analysis suggested that 9/13 potentially case driven vSNPs also associated with disease risk. Whilst the potential vSNP *CYP1A1* rs4886605 was not associated with case-control status, SNPs in *CHD5*, *TACSTD2* and *VPS53* that appeared to be driven by cases also associated with disease risk.

Within the 13 potential vSNPs, 7 were located within enhancer regions. In total, 4/7 enhancer associated potential vSNPs driven by cases were indeed associated with disease. These variants linked to *CHD5*, *TACSTD2* and *VPS53* Table 5-5.

All 5 SNPs identified in earlier *CHD5* analyses were associated with case-control status (Table 5-5). The p-values suggested strong evidence for this association, with values no higher than p=0.002. The strongest association belonged to rs1883764 (OR=2.49, 95% CI= 1.45-4.28, p= 0.0008). Three SNPs suggest quite a strong association with JIA (OR >2) (rs9435102, rs1883764, rs2273041), however 2 SNPs had the opposite direction of effect with OR's ~0.4 (rs2250358, rs9435080) (Table 5-5 and Figure 5-7A-C). Whilst LD (a measure of co-inheritance) between 2 SNPs with protective effect was very high, LD was also moderately high (0.58-0.75) between those and SNPs having opposite direction of effect. Therefore lack of LD likely did not explain differences in attributable risk between SNPs [436].

Meanwhile, 3 of 5 *TACSTD2* potential vSNPs were associated with case status. Whilst 1 SNP could not have an OR calculated due to the lack of minor alleles in the samples, 2 SNPs (rs11581395 and rs232847, Figure 5-7D-E) demonstrated consistent direction of effect, again with strong odds ratios of >2 and with low p-values <0.005 (Table 5-5). These 2 SNPs were known mQTLs, which are located within a number of enhancers. This is also a region of high LD, needing only a handful of SNPs to genotype the ~100kb region upstream of the EWAS probe within which these enhancers lie. It is likely these SNPs may indeed tag enhancer located SNPs (See methods section 5.2.1). However, due to small sample sizes, a larger study would be needed to validate these results.

One *VPS53* SNP suggested an association with cases, rs759974 (OR 2.23, 95% CI= 1.29 - 3.85, p= 0.003). This SNP was located within an enhancer. An odds ratio greater than 2 mirrored findings in *TACSTD2* and *CHD5*, suggesting these vSNPs may be strong susceptibility loci in these small genetic analyses.

Indeed, across 3 probes (excluding *CYP1A1*), SNPs were associated with DNAm variance and with case/control status. Further, intermediate levels of minor allele frequencies were a pattern across these SNPs, as were strong odd ratio's and p-values (OR's >2) particularly in enhancer located SNPs.

Table 5-5 Case driven potential vSNPs tested for associations with oJIA, a number of which are located in enhancers.

Gene	Probe	SNP	Gene region	Minor Allele	MAF (Cases/Controls)	P-value	OR	95% CI
CHD5	cg15205435*	**rs9435102**	Enhancer	**A**	**0.55/0.34**	**0.002**	**2.32**	**1.35 - 3.98**
	cg12135344**	**rs1883764**	Known mQTL	**A**	**0.60/0.38**	**0.001**	**2.50**	**1.45 - 4.28**
	cg12135344**	**rs2250358**		**G**	**0.37/0.57**	**0.003**	**0.44**	**0.26 - 0.76**
	cg12135344**	**rs2273041**		**A**	**0.59/0.38**	**0.001**	**2.40**	**1.40 - 4.12**
	cg12135344**	**rs9435080**		**G**	**0.36/0.57**	**0.002**	**0.43**	**0.25 - 0.73**
CYP1A1	cg12101586	rs4886605	Regional	T	0.20/0.19	0.84	1.07	0.55 - 2.07
TACSTD2	cg17210938	**rs10493238**	Regional	**T**	**0.05/0.00**	**0.02**	**NA**	**NA**
		rs11581395	Known mQTL	**T**	**0.39/0.20**	**0.002**	**2.52**	**1.38 - 4.58**
		rs232847		**A**	**0.36/0.19**	**0.005**	**2.36**	**1.29 - 4.35**
		rs338230		A	0.49/0.49	0.99	1.00	0.59 - 1.69
VPS53	cg04370829	rs4968169	Enhancer	A	0.27/0.37	0.14	0.65	0.37 - 1.16
		rs759974		**A**	**0.53/0.33**	**0.003**	**2.24**	**1.30 - 3.85**
		rs9913674		T	0.31/0.39	0.16	0.67	0.39 - 1.18

Case-control allelic association study was carried out, and asymptotic p-values are reported. TACSTD2 SNPs are located within a region with a number of enhancers, likely tagging enhancer located SNPs due to the high LD in the region. MAF: Minor allele frequency. NA: Insufficient alleles were present to analyse. SNPs in bold denote significant associations at p<0.05. N(cases) = 54, N(controls) = 57. CHD5* denotes the lead probe identified and validated in the EWAS analysis. CHD5** denotes another associated probe identified within the same region in the EWAS.

5.3.4.3 Potentially new mQTLs located within enhancers may also be oJIA susceptibility loci.

It was noted that in the new mQTL association analyses, control only associations were weaker when compared to analyses containing oJIA samples (section 5.3.1-2). These suggested cases were driving the mQTL relationship, and were potential oJIA risk loci. To investigate whether mQTLs suggest new oJIA susceptibility loci, cases vs. control allelic association analyses were carried out on robust newly identified mQTLs from section 5.3.2 (Table 5-2).

A number of new potential oJIA genetic associations were observed within all 4 DMPs interrogated ($p<0.05$). Indeed, 7/11 potentially new mQTLs were also associated with oJIA. As noted earlier, many of these SNPs are located within enhancer regions.

Analysis of the eight potentially new mQTLs in the *CHD5* region suggested six were associated with oJIA Table 5-6. A majority of these (5/6) were from enhancer located SNPs. Evidence for association was strong for all 6 identified SNPs, with the same direction of effect. The strongest evidence belonged to rs12565328 (OR=0.39, 95% CI= 0.22-0.71, p =0.001), rs9435102 (OR=2.32, 95% CI= 1.35-3.98, p= 0.002) and rs2485024 (OR=0.41, 95% CI= 0.23-0.73, p=0.002). Of note, rs12565328 was also associated with DNAm variance, and potentially also fits into Feinberg's hypothesis as a vSNP.

CYP1A1 contained no new mQTLs affecting mean DNAm, whilst the new mQTL identified with *VPS53* did not associate with oJIA. However, one *VPS53* SNP did associate with oJIA as noted in section 5.3.4.2 (rs759974), and evidence for an association for *CYP1A1* rs9434741 was potentially strong (p=0.002, OR=0.29 95%CI= 0.13-0.66). Both of these SNPs were located within an enhancer region.

One of two (1/2) potentially new mQTLs linked to *TACSTD2* was also associated with oJIA, located within the 10kb proximal region of *TACSTD2* (rs6666304, p=0.001, OR=3.60, 95% CI=1.70-7.63).

The abundance of enhancer located SNPs found to be proximal to DMPs and associated with oJIA suggests that DNAm may help guide the discovery of new oJIA susceptibility loci.

Table 5-6 Robust new mQTLs identified in this study were tested for associations with oJIA, a number of which are located in enhancers.

Gene	Probe	SNP	Gene region	Minor Allele	MAF (Cases/Controls)	P-value	OR	95% CI
CHD5	cg15205435	**rs11121295**	**Enhancer**	**A**	**0.12/0.25**	**0.01**	**0.40**	**0.20 - 0.82**
		rs12067480		T	0.13/0.16	0.55	0.79	0.37 - 1.69
		rs12565328		**C**	**0.22/0.42**	**0.002**	**0.39**	**0.22 - 0.71**
		rs1883604		**A**	**0.01/0.09**	**0.007**	**0.10**	**0.01 - 0.77**
		rs2485024		**G**	**0.25/0.45**	**0.002**	**0.41**	**0.23 - 0.73**
		rs9434662		**T**	**0.12/0.25**	**0.01**	**0.40**	**0.20 - 0.82**
		rs9435102		**A**	**0.55/0.34**	**0.002**	**2.32**	**1.35 - 3.98**
		rs17029184	Regional	T	0.06/0.08	0.68	0.81	0.29 - 2.25
TACSTD2	cg17210938	**rs6666304**	**Regional**	**A**	**0.28/0.10**	**0.001**	**3.60**	**1.70 - 7.63**
		rs6694699	Enhancer	C	0.07/0.10	0.55	0.75	0.29 - 1.94
VPS53	cg04370829	rs4968169	Enhancer	A	0.27/0.37	0.14	0.65	0.37 - 1.16

Case-control allelic association study was carried out, and asymptotic p-values are reported. MAF: Minor allele frequency. SNPs in bold denote significant associations at $p<0.05$. N(cases) = 54, N(controls) = 57.

5.4 Discussion

This study suggests the presence of genetic contributions to DNAm heterogeneity in oJIA. These genetic effects were on both mean DNAm and variation, through known and potentially newly identified mQTLs. Further, a number of these mQTLs were associated with oJIA also, suggesting the effect of genetics needs to be taken into account when designing DNAm studies in complex diseases.

5.4.1 Known mQTLs observed in oJIA enriched samples may contribute to DNAm complexity

The presence of published mQTLs, specifically associating with probes associated with oJIA in this study, provided an opportunity to investigate a potential source of variation that could explain the inability to replicate EWAS findings.

All published mQTLs identified were also found to associate with DNAm in this study, mirroring previous reports on these SNPs [245, 247]. Most of these SNPs remained associated through sensitivity analyses in control and female only samples, suggesting these findings are robust. The most robustly associated published mQTLs (for each probe) have very similar DNAm distributions between genotype groups, suggesting there is a lead causative SNP either amongst the list of known mQTLs or even outside of this region. Substantial differences in DNAm were observed between genotype groups, with mean DNAm differences up to 13% and 27% between genotype groups in the most robustly associated mQTLs for *CHD5* and *TACSTD2* respectively. Considering that minor alleles for these mQTLs are not rare, indeed they are typically present at MAF of >20%, there is a significant risk of introducing bias towards a particular genotype that can result from a small sample size. This may be the reason for the difficulty experienced in replicating discovery EWAS results.

These known mQTLs were identified from the literature, focussing on studies conducted in similar cell types (i.e. immune cells). Indeed one source used neonatal T-cells, further suggesting results in this study of children's immune cells are robust [247]. However, larger numbers are needed to validate these findings in oJIA, in children this age, and to find the causative SNP having this effect on DNAm.

It is unknown whether genetic effects on DNAm are working in isolation or in concert with environmental factors. For mQTLs originating from neonatal data, effects of the environment are likely to have limited impact. Alternatively, these childhood specific mQTLs may have been identified due to their sensitivity to the environment in neonatal and childhood years. Combining mQTL data with known environmental mediators of DNAm (e.g. exposure to smoking in pregnancy or early childhood) with larger sample sizes may help deciphering the links between genetics, DNAm and environment in oJIA. This link has been investigated in adult rheumatoid arthritis. A smoking and genotype interaction is reported to have DNAm as an intermediary, acting to integrate signals from biology and environment [437]. This was reported using 2 replication efforts, including ~1700 RA cases and a higher number of controls. Such efforts for a much rarer disease like oJIA will require large scale international collaboration.

Known mQTL associations identified in adult populations were also observed in this study of children. The observation of these mQTLs in younger people suggests the DNAm differences between genotypes are genetically driven, with environment likely to play a lesser role. It may also be a female sex-specific effect, since both this study and the source paper for adult mQTLs were enriched for females [245].

Nonetheless, observations mirroring published mQTL data suggests *CHD5* and *TACSTD2* DNAm data are likely to be modulated by genetics contributing to DNAm heterogeneity. This would likely affect replication efforts, and larger sample sizes would be needed to ensure robust discovery and replication data.

5.4.2 Enhancer mQTLs may contribute to DNAm complexity in oJIA

Enhancer-located SNPs have been known to associate with DNAm, and may better describe the link to gene expression and disease progression [245, 247, 249, 250]. Therefore, enhancer SNPs may have contributed to DNAm heterogeneity and the difficulty in replicating results in this study.

In three out of four probes investigated, SNPs located within tissue specific enhancer regions were associated with DNAm. A majority of robust findings (i.e. SNPs also associated in sensitivity analyses) for both DNAm mean and variance analyses were in enhancer regions and not in regions proximal to the CpG outside tissue specific enhancers. These findings mirror similar reports from Gutierrez-Arcelus *et al.*, who described enhancers as being

enriched for mQTLs in similar cell types and age group to this study [247]. A number of these SNPs represent new potential mQTLs, and a potential source of variation that may have affected EWAS replication.

Complex diseases like oJIA and other autoimmune diseases have also been linked to enhancer SNPs in immunochip studies [112, 149]. In particular, Farh *et al.* found oJIA susceptibility SNPs identified through immunochip were located within CD4+ T-cell enhancers including Th1 and Th2, and most strongly associating with Th17 specific enhancers [149]. Since immune cell enhancers (enriched for CD4+ T-cell specific enhancers) were selected for genotyping, it is likely that these incorporated Th17 specificity.

With enhancers being pivotal to both complex disease genetics and DNAm, this study identified a link between all enhancer SNPs and DNAm. In addition, analogous to Farh *et al.*, a number of enhancer mQTLs were also associated with oJIA. The precise relationship of DNAm to enhancer SNPs and oJIA is yet to be determined. A larger study would ideally to isolate cause from effect, independent effects, or bystander effects using methods such as causal inference testing [245, 251].

5.4.3 Enrichment of mQTLs in this study

It is worthy to note the contrast in mQTLs found between this study and those in the literature. A majority of mQTL studies report less than 10% of top ranked CpGs associated with an mQTL, yet this study suggests 3 out of 4 selected probes have an mQTL associated [219-229].

Perhaps the selection criteria for probes of interest may have biased the selection towards probes that are under genetic influence. Of studies noted above investigating mQTLs, none selected for CpGs with regional similarities in DNAm. A single study selected a p-value cut-off, whilst another used beta value limits to investigate mQTLs, as was done in this study. In attempting to select robust CpGs for validation and replication, this study may have enriched for mQTL associated probes.

Of note, most studies identified used samples from normal, non-diseased people. This study included samples from a complex disease with known strong genetic component, which may have also biased top ranking CpGs towards those influenced by genetics. In addition, this

study utilised half to a third the sample size of almost all studies and sampling bias may have inadvertently enriched for genetically driven CpGs.

5.4.4 Genetically driven DNAm variation potentially contributes to disease

The mQTLs published by Lui *et al.* and Gutierriez-Arcelus *et al.* modified mean DNAm [245, 247]. However, Feinberg *et al.* suggested genotypes could also affect DNAm variation (vSNPs) [236]. That is, at certain CpGs, particular genotypes would allow larger variations in DNAm levels between individuals [236]. In this study, SNPs that are able to affect DNAm variation could also affect the ability to replicate an EWAS with a small sample group.

All four probes in this study were associated with a vSNP, suggesting that a significant portion of DNAm heterogeneity may well be due to genetic sources. Of interest, the observation of potential genetic influence amongst CpGs with large inter-individual variation was also made by Islam *et al.* and provides, at least, a nominal type of validation of this thesis' finding [438]. Islam *et al.'s* top ranking variable probes of interest were visualised and produced patterns highly reminiscent of CpGs typically under genetic influence [438]. Whilst the variable CpG -mQTL association was not statistically tested, the list of CpGs tested substantially overlapped with known genetically influenced CpGs [438, 439].

In this study, a number of these vSNPs displayed two characteristics Feinberg *et al.* also postulated i.e. 1) association with a complex disease and 2) location within a regulatory region (enhancers). Eight of seventeen SNPs associated with DNAm variation had both these characteristics. These findings marry with Feinberg's hypothesis that vSNPs are a likely source of disease risk for complex diseases. According to this hypothesis, should a genotype allow a wider range of DNAm values than another genotype, this would allow a phenotype a greater ability to adapt. However, the potential for more DNAm variation also allows for extreme DNAm values, which could pre-dispose the cell/person to certain diseases. This hypothesis was validated in the large RA EWAS study, where 5/9 of the top ranking RA CpG sites were indeed vSNPs [245]. Initially, Feinberg's hypothesis can also be observed in this study.

However, unlike Feinberg *et al.*'s analysis, this analysis cannot differentiate between DNAm as a causal risk to oJIA, an effect of oJIA or as a mediator of genetic risk. Larger sample numbers, Mendelian randomisation studies or perhaps the causal inference test employed by Feinberg *et al.* would need to be utilised in order to better understand the relationship

between genetic, DNAm and oJIA risk, as well as the potential for these SNPs to complicate discovery EWAS findings in this study.

Indeed, this genetic study is a preliminary piece of work owing to the small sample sizes investigated. It attempted to explore a field yet to be understood in JIA, to identify potential avenues for further research. Due to the small sample sizes, multiple testing correction could not be completed. To attempt to identify some robust associations, sensitivity analyses were carried out utilising either controls only or females only (the 2 largest groups of samples outside of the overall analysis). Future studies will need to be larger in order to comprehensively control for multiple testing.

Other biological factors may also act as a source of variation in DNAm, such as case group heterogeneity. This will be investigated in the next chapter.

Chapter 6. **Investigating DNA methylation heterogeneity in oJIA by focussing on potential sources of clinical heterogeneity.**

6.1 Introduction

This thesis has suggested DNA methylation in oJIA is complex and not easily explained by case-control status. The DNAm variability observed potentially reflects genetic variation (as explored in Chapter 5). It may also mirror phenotypic variability in patients observed within the oJIA subtype, which may remain despite a homogeneous clinical classification.

Clinical and molecular evidence suggests age subgroups in oJIA are important (discussed in Chapter 4). Indeed, early seminal studies investigating classification criteria highlighted associations of age at onset with certain clinical markers like antibody status and eye disease, and even suggested this a part of a new classification criteria [135, 440]. Accumulating molecular data suggest that early efforts may have correctly identified associations with age of onset. This more recent data includes transcriptomic data grouping oJIA and RF-negative poly-JIA together by age of onset (≤6 and >6 years of age), and multiple publications from Ravelli *et al.*, Prakken *et al.* and Martini *et al.* on grouping the same subtypes according to age and anti-nuclear antibody status [12, 16, 17, 50]. These age subgroups put forward by Ravelli *et al.*, Barnes *et al.* and Martini *et al.* identify people who have a similar disease course, mirroring justifications for the use of age in early classification studies [440].

In addition to age subgroups, clinically recognised subgroups based on disease severity are also present in oJIA, namely, Persistent and Extended disease subgroups. A diagnosis of Extended oJIA is given 6 months post initial diagnosis, should a total of 5 or more joint be affected [53]. Otherwise, persistent oJIA is the final diagnosis [53]. Up to 50% of oJIA patients will progress to extended disease and these patients are more likely to have poorer clinical outcomes than persistent oJIA patients [66-68]. This includes being less likely to achieve remission off medication, having more time in an active disease state and being more likely to have disease flares even when in remission [67, 69, 70]. CD4+ T-cells are associated with these subgroups, e.g. altered CD4:CD8 ratios, Th17 and Tregs increased in joints as well as peripheral blood showing Treg subset differences [44, 45, 48].

Therefore, potential exists for clinical and biological heterogeneity even within a single subtype i.e. oJIA, that may involve CD4+ T-cells. This chapter explores EWAS data to identify sources of potential clinical heterogeneity, which may help explain the difficulty in replicating EWAS findings, and considers the potential for biological markers of disease.

6.2 Methods

6.2.1 Samples

An age subgroup analysis of the EWAS data was carried out utilising samples from the initial 'all-inclusive' analysis in chapter 4. Case subgroups were defined according to age of diagnosis, either ≤6 years of age or >6 years of age at the time of initial diagnosis. Samples were collected an average of 2.5 months from diagnosis (see Chapter 4.2.1). Controls were matched according to age (within 12 months) and sex. Extra details on the CLARITY project are provided in Chapter 2.1.

The summary characteristics of each age subgroup is shown below in Table 6-1.

Table 6-1 Summary characteristics of Discovery samples in the final analysis, as well as replication samples (for the younger diagnosed specific study)

	Discovery			
	Case		**Controls**	
Age at diagnosis[b]	≤6yrs	>6yrs	≤6yrs	>6yrs
No. samples	35	21	33	24
Mean age (SD)	3.47 (1.94)	10.47 (2.33)	3.21 (1.19)	10.79 (2.35)
Female | Male	30/6	15/6	25/8	18/6
	Replication			
	Case		**Control**	
Younger-diagnosed (≤6yrs) replication samples	10		11	
Mean age (SD)	3.2 (1.42)		3.9 (1.55)	
Female/Male	6/4		6/5	

.

Discordant numbers between cases and controls were due to loss of samples during data analysis QC. Mean age = age at blood collection. [a] Matched by age and sex. [b] Age at diagnosis is specific to Cases, whereas Controls in age-at-diagnosis category represent age at recruitment.

6.2.2 Statistical analyses – Genome scale methylation data

Genome scale DNA methylation data was generated using the Illumina Infinium HumanMethylation450 (HM450k) Beadchip kit (Illumina, San Diego USA). Details of the

generation of this data as well as the quality control and data processing are provided in Chapter 2.3.3-2.3.5.

6.2.2.1 Differential DNA methylation analysis: RUV

RUV was used for DMP analysis. For details on this method, see Chapter 2.3.5. Briefly, RUV is a method for removal of unwanted variation. It uses bottom ranked probes from an initial DMP analysis to identify covariates not related to the factor of interest (i.e. oJIA). These are then used as covariates in a subsequent DMP analyses.

6.2.2.2 Age specific analyses (≤6 and >6 years)

Specific to the younger age of diagnosis analysis (≤6 years of age at diagnosis), four iterations of the RUV DMP analysis were completed using the bottom 70% of ranked probes as empirical control probes (ECPs). For the older age of diagnosis (>6 years of age at diagnosis), three iterations of the DMP analysis were completed using the bottom 90% of ranked probes as empirical controls. P-values were adjusted for multiple testing using the Benjamini-Hochberg method [391]. MDS plots were used to identify case control clusters using top ranking probes, with as few of the top ranked probes. This clustering occurred using probes with unadjusted $p<0.0001$ and $p<0.0005$ from the younger and older diagnosed analysis respectively. Beta values and delta beta values were calculated as noted in Chapter 4.2.2.1.

6.2.2.3 Persistent vs Extended

This sub-group analysis focused on cases only, comparing persistent oJIA to extended oJIA cases. Differential methylation analysis was completed using four iterations of the RUV DMP analysis, where the bottom 70% of ranked probes were used as empirical control probes (ECPs). P-values were adjusted for multiple testing using the Benjamini-Hochberg method [391]. MDS plots were used to identify case control clusters using as few of the top ranking probes, in this case with unadjusted $p< 5 \times 10^{-5}$. Beta values and delta beta values were calculated as noted in Chapter 4.2.2.1. Investigation of age subgroups could not be done within this analysis due to the very small numbers available for analyses, particularly the older extended subgroup.

6.2.2.4 Classification analysis using ROC curve

The ability of DNAm to discriminate between persistent and Extended-to-be cases was assessed by utilising receiver operating characteristic (ROC) analysis and determining the

area under the curve (AUC), sensitivity (true positive call rate) and specificity (true negative call rate) of this data. The AUC is a probability of an event happening (disease extension) and is produced based on a scenario of 2 patients, in this case one who will progress to disease extension and one that won't. A higher probability of disease extension is produced for this first patient, allowing the discriminatory power of the data to be interpreted from a single value. Values greater than 0.5 are required for positive classification, with values of 1.0 indicative of a perfect ability to identify events (i.e. disease extension) [441, 442]. Average DNAm values for each sample were calculated based on selected number of probes, and used for plotting of ROC curves. These were plotted using GraphPad Prism 7.0.

6.2.3 EpiTYPER DNA Methylation Analysis

6.2.3.1 Assay design

Probes of interest were identified for technical validation and replication studies using a selection criteria outlined below (Chapter 6.3.2.1 and 3.4). Primers were designed for 4 genes, *ARSA*, *lncRNA*, *RGS12* and *POU2AF1* (listed in Table 6-3). Details of primer design are provided in Chapter 2.3.7. Details of the primers designed are provided in Table 6-3.

6.2.3.2 Polymerase Chain Reaction

Validation and replication of identified probes of interest was completed using the MassArray EpiTYPER system for loci specific DNA methylation analysis (Agena Bioscience, San Diego, USA), as described in Chapter 2.3.6. Up to 250ng of genomic DNA was bisulphite converted using the EZ DNA Gold kit (Zymo research, Irvine, USA). 10ng of converted DNA per sample was used in triplicates for each assay. Primers were designed for 4 genes, *ARSA*, lncRNA, *RGS12* and *POU2AF1*, as listed in Table 6-3. PCR was completed using FastStart mastermix (Roche, Basel, Switzerland), was cycled using the following temperatures in Table 6-2 and based on the protocol in Coolen *et al.* 2007 [341]. To check for amplification, samples were visualized on an agarose gel (Chapter 2.3.9).

SAP and T-Cleavage steps were completed as previously described in Chapter 2.3.10-11. Data quality was visualized and assessed using EpiTYPER V1.2 (Agena Bioscience, San Diego, USA). Replicates were used to calculate averages if methylation values were ±10% of the median value. Student's t-test was used to assess statistical significance in difference between group means.

Table 6-2 PCR thermocycling conditions for assays ARSA, lncRNA, RGS12 and POU2AF1.

Temperature (˚C)	Time	Cycles
95	10 min	5
95	10 sec	
56	30 sec	
72	1 min 30 sec	
95	10 sec	40
60	30 sec	
72	1 min 30 sec	
72	7 min	1
4	Hold	1

Table 6-3 List of primer sequences designed for EpiTYPER analysis of the JIA EWAS Age of diagnosis sub-study.

Gene/Primer	Sequence 5'-3'
*POU2AF1*_F	AGGAAGAGAGATTRGGAGTTAGTTAGTTGGGATAGAGT
*RGS12*_F	AGGAAGAGAGTTTTAGTTAAATGGTATGGGGTTTAAGGT
lncRNA_F	AGGAAGAGAGGGAGTATAGGATGAATTTGTAGGTAGGTAG
*ARSA*_F	AGGAAGAGAGGAGTGGTAGAGTTGGGGTGTTTAT
*POU2AF1*_R	CAGTAATACGACTCACTATAGGGAGAAGGCTAACTAAATCCTACCCATACCCTAATAACCTAA
*RGS12*_R	CAGTAATACGACTCACTATAGGGAGAAGGCTIGAIGTAACTAIGATCACACAAAAAATA
lncRNA_R	CAGTAATACGACTCACTATAGGGAGAAGGCTCTATTACTTCCTTTCCCTAAATCTCATCTAAAAAA
*ARSA*_R	CAGTAATACGACTCACTATAGGGAGAAGGCTGIAAAAAAACCTACTAAAACCAAATAACCC

Forward and reverse primers are indicated with a suffix (_F and _R, respectively).

6.3 Results

6.3.1 Age subgroups may be homogeneous at the DNA methylation level

6.3.1.1 Data structures within the whole cohort suggest large variation due to age subgroups

To explore the potential for age subgroups contributing to the variability in oJIA DNAm, the top hits from the EWAS analysis in Chapter 4.3.2.1 ($p<0.01$, 4166 probes) were visualised using MDS plots. Age subgroups ≤6 years and >6 years of age at diagnosis were used to label samples on the plot. MDS plots of the top hits at $p<0.005$ were also visualised using these age group labels. Figure 6-1 suggests that these 2 subgroups may indeed be present in cases and controls. Further, there is a difference in cluster patterns between age groups along dimension 1. This is observed as a generally tighter clustering of samples in the ≤6 years subgroup, compared to a wider spread of samples in the >6 years subgroup. The separation of the 2 age subgroups and the differences in subgroup clustering suggests that these age groups may have differing underlying biological characteristics. This sample group heterogeneity potentially makes an analysis incorporating both these age groups difficult. Therefore stratifying the analysis by age of diagnosis may yield a more sensitive analysis.

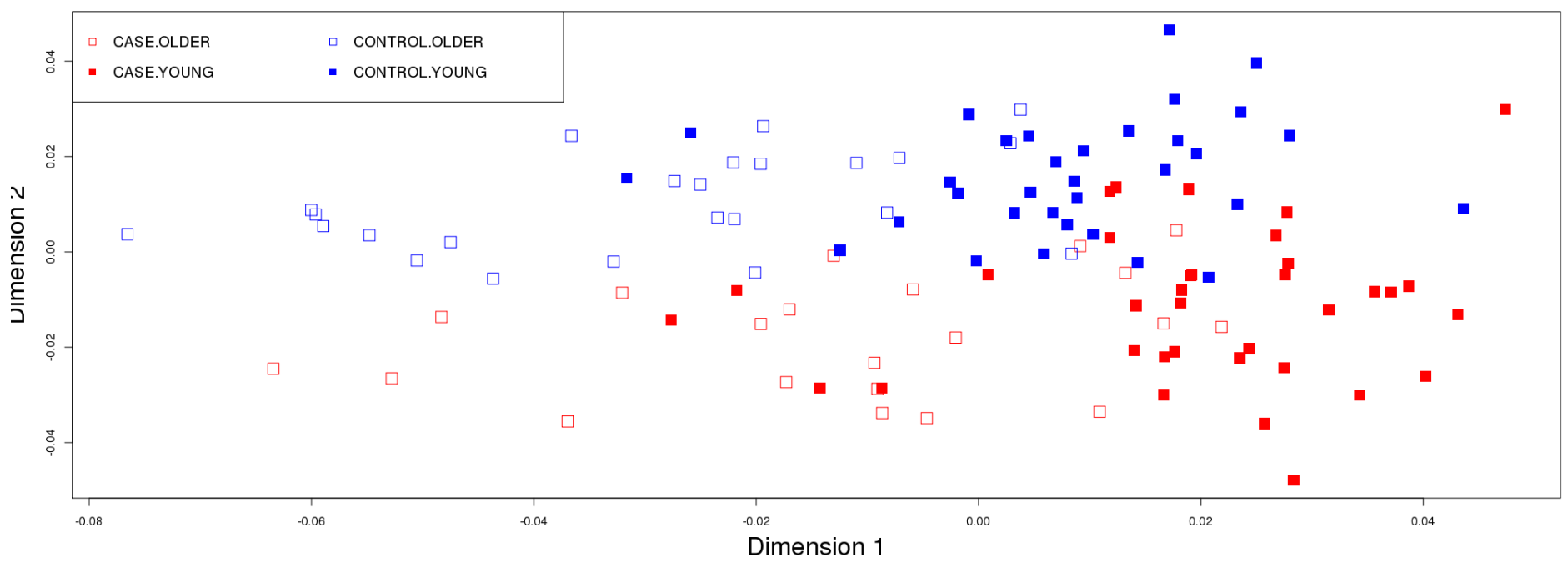

Figure 6-1 MDS plot of the all Inclusive analysis, using top hits from RUV analysis (unadj.p<0.01, 4166 probes)

Cases are in red, controls are in blue. In addition, age subgroups are shown: Closed squares represents younger (≤6 years) ages at diagnosis. Open squares are older (>6 years) samples at diagnosis.

6.3.1.2 Differential Methylation analysis of the older subgroup (>6yrs age) suggests a more homogeneous sample group.

To assess whether removing age subgroups as a variable results in a more sensitive analysis with robust results, samples were analysed using only the >6years of age subgroup. This included 21 cases and 24 controls for the analysis, see Table 6-1. Differential methylation analysis was undertaken using RUV, as it was able to remove unknown (technical or biological) sources of variation.

Implementing the RUV pipeline, analysis of older diagnosed samples against matched controls resulted in no significant hits. Despite this, MDS plots of top ranking probes demonstrated that far fewer probes were able to cluster cases from controls compared the all-inclusive analysis. Case and control clusters were easily identifiable using just 106 of the top ranking probes (unadjusted p <0.0005, Figure 6-2). This is compared to 4166 probes from the all-inclusive analysis required to identify case and control clusters. This substantial reduction in required probe numbers suggested that reducing the heterogeneity due to age has increased the sensitivity of the analysis. In the older diagnosed subgroup, it was still insufficient to obtain statistically significant hits. This may be due to a sample number issue, particularly since the sample group was almost halved in comparison to the all-inclusive analysis.

To assess if age remained a primary or secondary principal component of the top hits within this age subgroup, MDS plots were visualised using age labels. Figure 6-2 shows that age does not explain any spread in the data, suggesting the heterogeneity due to age in the all-inclusive analysis is not present in this older aged subgroup.

Interestingly, whilst age was not a source of sample heterogeneity in the top hits from this older diagnosed analysis, there appeared to be another factor that potentially add complexity to the oJIA subtype. This factor causes a noticeable clustering pattern within the top hits. It appears to affect both cases and controls, whilst not quite being a primary or secondary component explaining the data but somewhere in between. It is particularly noticeable as more probes are used to cluster samples (e.g. 248 probes at unadj.p<0.005. data not shown). This type of large visual separation resembled that seen in the LIMMA analysis in the all-inclusive analysis (Figure 4.9B). Similarly, both MDS plots show a large split in the data that cuts across both cases and controls. Chip effects were accounted for during

data processing and did not constitute one of the top 2 principal components (PC1 and PC2 p=0.5 and p=0.9 respectively) (Figure 4-5).

In order to attempt to identify what this source of variation was, a number of phenotype labels were added to samples in the MDS plots. No phenotype data available was able to categorically align itself with the split in the MDS plot samples (data not shown).

Therefore, whilst this age stratified analysis was perhaps more sensitive than the all-inclusive analysis, no significant hits were found and an unknown source of sample heterogeneity remained. The observed heterogeneity in this oJIA subgroup potentially adds further complexity to oJIA subtype and to the methylation analysis. Work therefore progressed to the younger diagnosed samples.

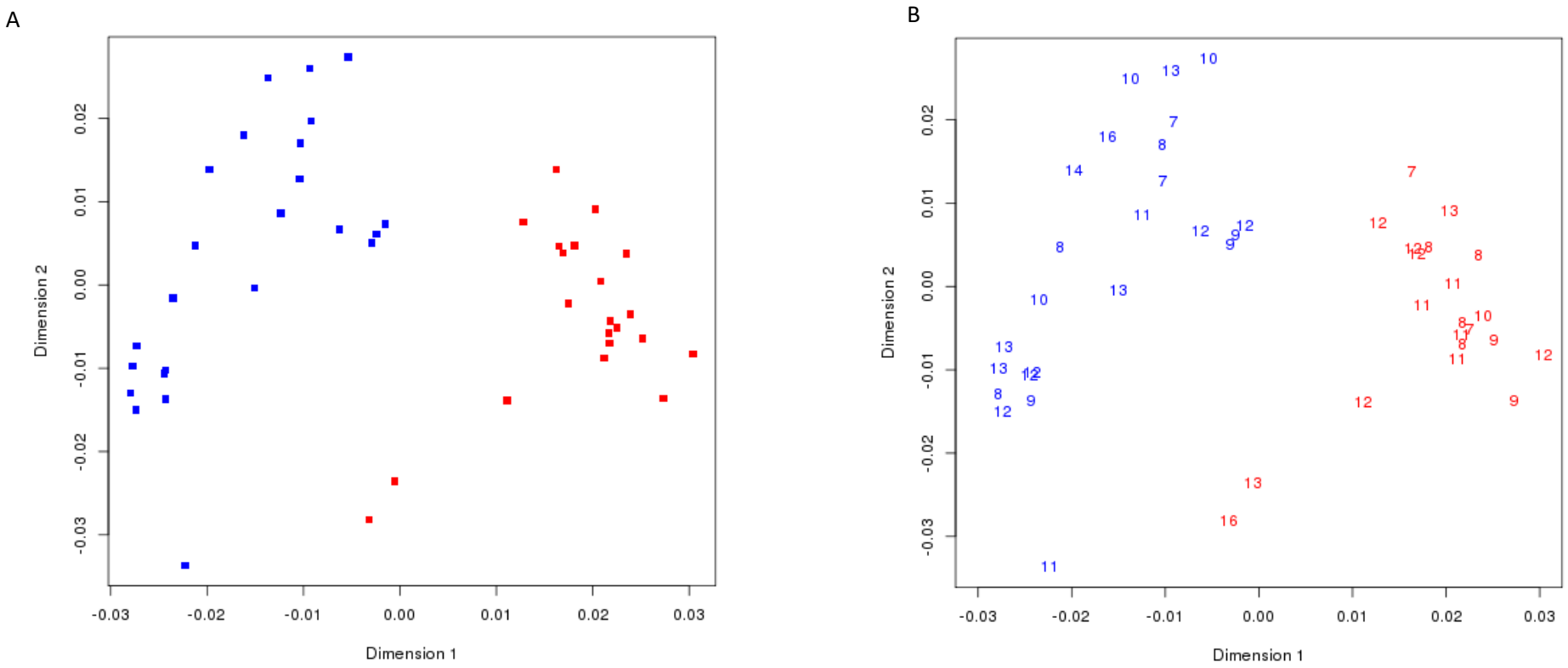

Figure 6-2 A-B. MDS plots of older diagnosed analysis, using the minimum number of probes able to cluster case-control groups (unadj.p<0.0005, 106 probes).

Cases are in red, controls are in blue. B. The same plot as in A), this time case samples are labelled with age at diagnosis (closely matching age at blood collection) and controls are labelled with age at blood collection.

6.3.1.3 Younger aged subgroup analysis shows greater sensitivity and reduced heterogeneity in top ranking probes

To further investigate the potential effect of age as a confounder in the all-inclusive analysis, the younger subgroup was also investigated in a sub analysis utilising 35 cases and 33 controls. The RUV pipeline was again implemented for differential methylation analysis. Despite a smaller sample group, this analysis resulted in 1 probe reaching significance (FDR 0.01). This probe, located on chromosome 6, is located within a recently annotated lncRNA and within an enhancer region.

Similar to the older diagnosed analysis, MDS plots of top hits suggested that sensitivity had also increased in the younger diagnosed analysis. Using just the 31 top ranking probes (unadj.p <0.0001, Figure 6-3) from the younger diagnosed analysis, cases and controls were able to be clustered together. These MDS plots demonstrated that far fewer probes were needed to cluster cases from controls compared to the all-inclusive analysis (4166 probes), and even compared to the older diagnosed samples (106 probes) (Figure 6-2). Despite smaller sample groups in age stratified analyses, the reduced heterogeneity appears to have increased the sensitivity of both analyses, particularly in the younger diagnosed analysis (aged ≤6 years).

To assess if age remained a primary or secondary principal component of the top hits within this age subgroup, MDS plots were visualised using age labels on samples. Figure 6-4 shows that age does not explain any spread in the data, even when using more probes to cluster samples. This suggests that the heterogeneity due to age and so strongly present in the all-inclusive analysis has been mitigated in this younger diagnosed sub-analysis.

In addition, whilst there appears to be a hint of clustering similar to the older diagnosed analysis (where 4 groups are forming), this disappears once more probes are used (e.g. unadj.p<0.005, 1939 probes Figure 6-4). This is in contrast to the older diagnosed analysis, where the addition of more probes led to a more intensified clustering of samples into 4 distinct groups. Therefore, this factor may be present in the younger diagnosed samples but is not a large source of variation inherent to DNAm in younger diagnosed samples, and is likely limited to the older diagnosed subgroup.

With both younger diagnosed and older diagnosed cases readily identifiable in MDS plots using far fewer probes, it appears as though age may be a source of biological heterogeneity

in oJIA reflected in DNAm data. The results from the age subgroup analysis contributed to a publication in January 2018 in the Journal of Autoimmunity, “The DNA methylation landscape of CD4+ T cells in oligoarticular juvenile idiopathic arthritis”, and is listed in Appendix III.

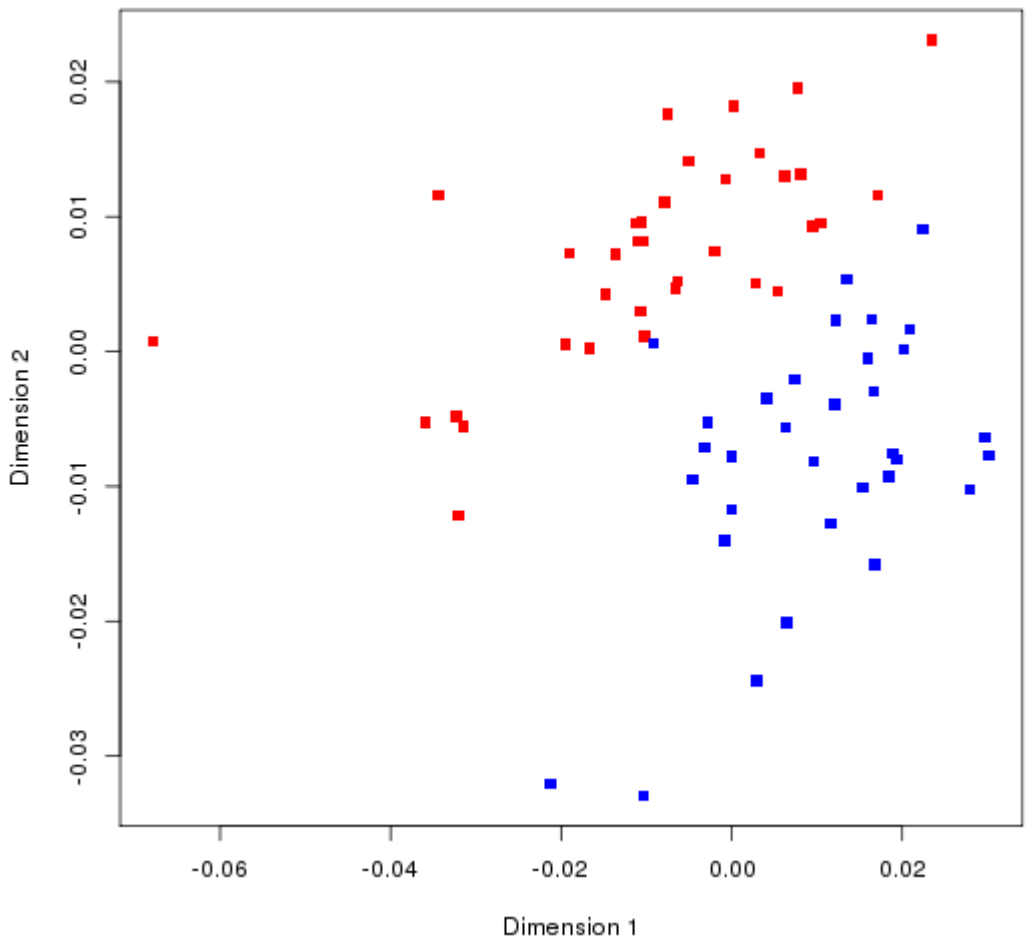

Figure 6-3 MDS plot of younger diagnosed analysis, using the top hits from the RUV analysis (unadj.p <0.0001, 31 probes).

Cases are in red and controls in blue.

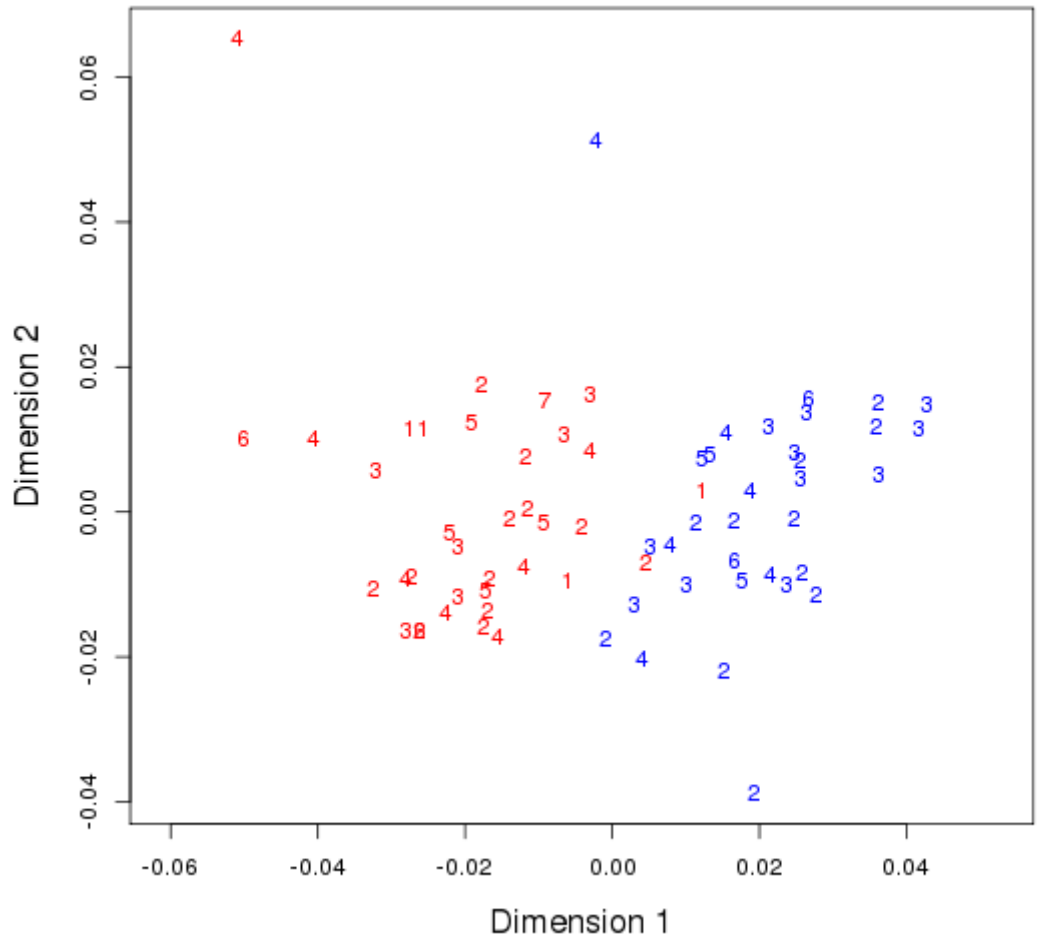

Figure 6-4 MDS plot of younger diagnosed only samples, using more of the top hits from RUV analysis (unadj.p<0.005, 1939 probes).

Age at blood collection (comparable to age at diagnosis) labels are added for each sample to demonstrate the spread of ages in this subgroup analysis.

6.3.1.4 Age subgroups are different at the DNA methylation level

To explore whether age subgroups have underlying biological differences, top hits from the stratified analyses were compared. From each subgroup analysis, top ranking probes able to clearly separate case-control clusters were used (≤6 years: 31 probes, >6 years: 106 probes). Comparing these probe sets, no probes were found to be overlapping. In order to get a better understanding for any overlaps present, a much larger set of top ranked probes was used (unadj. p-values <0.01: ≤6 years: 4040 probes, >6 years: 3094 probes). Just 30 probes were present across both lists of top ranking hits, constituting just a 0.74% and 0.96% overlap respectively.

This suggests that age subgroups, observed in MDS plots of the all-inclusive analysis, may be due to differences in underlying biology associated with DNAm.

6.3.1.5 Age subgroups have unique and non-interchangeable DNA methylation signatures

To further explore the age subgroups for differences in underlying biology, MDS analysis was used to cluster one subgroup based on the other group's top hits. In this instance, the top 31 probes that were able to cluster younger diagnosed cases from controls were selected. Samples from the older diagnosed analysis were then clustered using these 31 probes from the younger analysis, and the clustering pattern observed. The vice versa was also attempted, using the top 106 probes from the older analysis on sample from the younger analysis (see Figure 6-5).

Both subgroups could not be clustered in case control groups using the opposing subgroups top ranking hits (Figure 6-5). This suggested that top hits from age stratified analyses were specific to the age subgroups. Despite the clinical similarities between age groups, this data may point to these subgroups having a differing underlying disease biology.

Of interest, younger samples adopted the 4 group split when clustered using older diagnosed top hits. The separation in groups also appears as the first and second principal components of the data, similar to the older analysis clustering. This suggests that there is a strong source of variation inherent to the older samples, which is detected by DNAm data. This source of variation is present amongst the younger diagnosed samples, as indicated by Figure 6-5, but not present amongst the top hits from the younger diagnosed samples. This variation is unlikely to be due to technical error for two reasons.

1) Due to the randomisation of samples across the arrays. Technical effects would not be limited to a specific age group since samples were randomised.

2) Due to the randomisation of sample locations, the technical error is extremely unlikely to have a link to disease status in only the older diagnosed samples.

Therefore the most likely scenario is that the source of variation that is detected in the older samples is biological. It happens to be present in the younger samples but not strong enough to affect case-control status or be amongst the top hits in the younger analysis. It may potentially be an environmental or developmental exposure that takes time (more than 6 years) to affect one's risk of becoming affected by oJIA e.g. pathogenic cell type that increases in numbers and pathogenicity only after age 6.

This unknown source of sample heterogeneity may be the reason for the lack of significant findings in the older diagnosed analysis. DMP analysis in the older subgroup was likely confounded by such a variable. Whilst this variable is related to disease status in the older subgroup, it is unrelated to disease status in the younger age subgroup, evidenced by the lack of case-control grouping in the younger analysis samples.

Taken together 1) the age stratified analyses along with 2) the inability to cluster subgroups with the opposing groups top hits, this data suggests that oJIA age subgroups do differ at the DNAm level. This validates undertaking the stratification of the analysis according to age.

Since the younger diagnosed samples appeared to be the more homogenous sample group, as observed by producing a significant hit and clustering of samples using fewer of the top hits (and considering time, sample and resource limitations), this younger subgroup was the most prudent choice for attempting technical validation and replication studies.

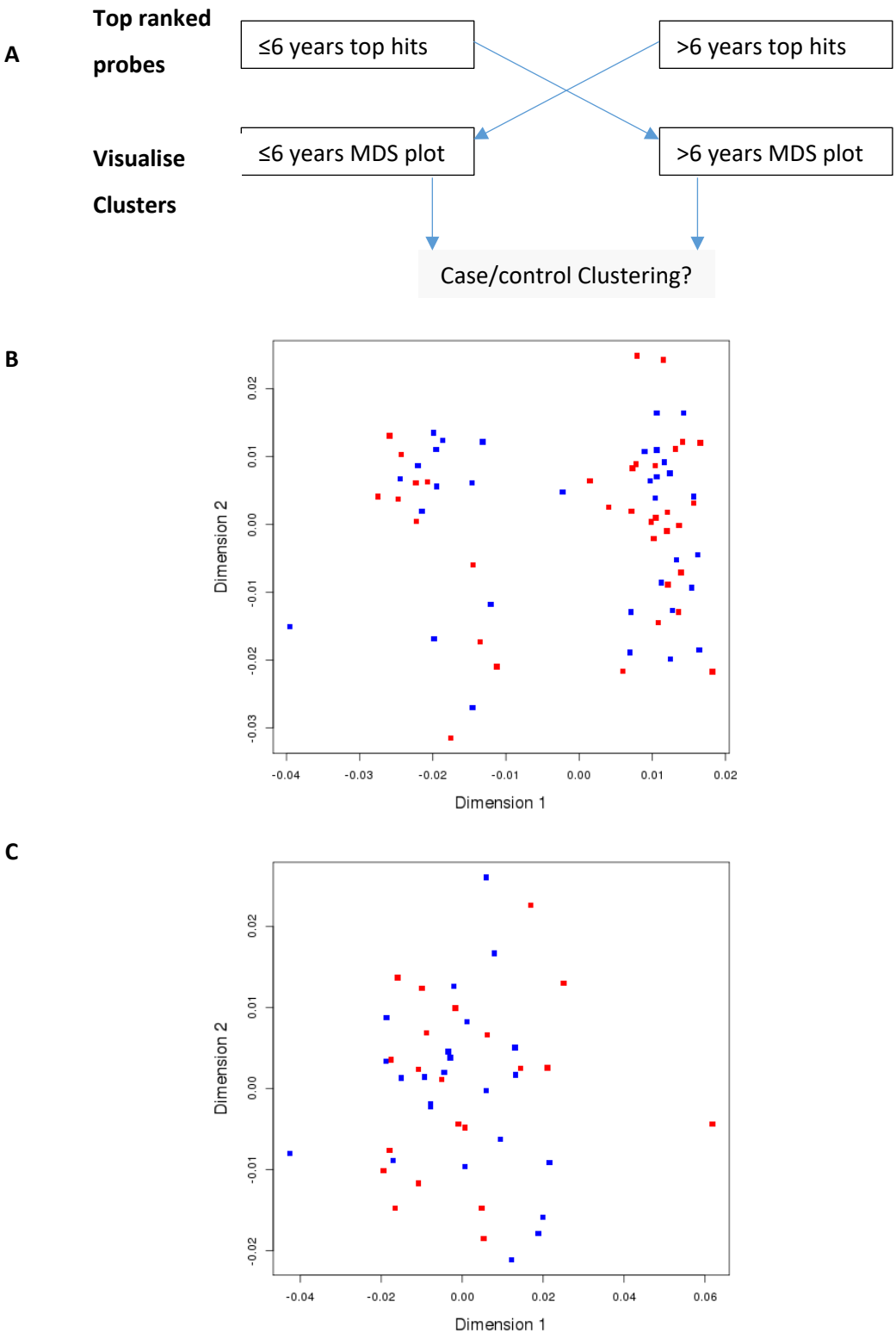

Figure 6-5 MDS plots exploring the age specificity of top ranked probes from age subgroup analyses.

Figure A: Schematic outlining the crossover design for investigating the age specific nature of top ranked probes. Probes originated from age specific sub-analyses. B-C: MDS plots using age subgroups. Cases are in red and Controls are in blue. B) Younger diagnosed samples plotted using the top 106 probes (unadj.p<0.0005) from the older diagnosed analysis. C) Older diagnosed samples plotted using the top 31 probes (unadj.p<0.0001) from the Younger diagnosed analysis.

6.3.2 Younger aged subgroup analysis

To explore the robustness of results from the younger subgroup analysis, technical validation and preliminary replication work was attempted. Technical validation was carried out using loci specific methylation analysis on a separate platform, MassArray EpiTYPER, utilising the same samples used in the younger diagnosed analysis.

6.3.2.1 Selection of probes to validate

Selection of top ranking probes for technical validation were selected based on the criteria developed in learnings from the vitamin D study (Chapter 3.4) and implemented in technical validation work from the all-inclusive analysis (Chapter 4.3.3-4). To reiterate, statistically significant probes do not always represent biologically relevant probes. Therefore, this criteria was developed in order to enrich the selected probes for biological relevance and to increase the likelihood of technical validation. Probes were selected for validation based on the following criteria:

- P-value cut-offs: using unadjusted p-values <0.01.
- Beta value cut-offs: delta beta values ~5% or greater.
- Regional correlation of delta beta values, and delta beta values ~5%: 2 or more CpGs <2kb apart, with delta beta values in same direction.
- Functional relevance: published data with associations with immune system or epigenetic mechanisms

In addition, the top ranking probe was also selected as it was the only hit to survive multiple testing adjustment. Based on this, 4 probes were selected for validation.

The top ranking probe (FDR 0.01) was selected, a probe sitting in the enhancer region of an lncRNA on chromosome 6 (interrogating CpG site cg09469870). Little is known about this particular lncRNA, yet a number of lncRNAs have been associated with Rheumatoid and osteoarthritis as well as JIA [443, 444]. In addition, published data suggests a number of pathways important in inflammation are associated with lncRNAs [444]. Further, correlations between lncRNA and DNAm are also recognised [445].

Also selected for technical validation was the CpG site cg25575065 in the *ARSA* gene. It has known links to the immune system, where stimulation with the cytokine IL-2 activates this

gene in Natural killer cells [446]. It is also known to inhibit NFKβ via TIFAB binding and is possibly involved in apoptosis via the CERK pathway [446-449].

RGS12 also contained a probe meeting the criteria (cg22655196). This gene is known to respond to treatment in RA synovial fluid macrophages, as well as also being involved in apoptosis (a pathway hypothesised to be important in JIA aetiology) [450, 451].

The CpG site cg01212326 in the gene *POU2AF1* was selected, and has a role in a number of immune mediated diseases including primary biliary cirrhosis, allergy and celiac disease [452-454].

6.3.2.2 Technical validation data supports EWAS data

The four selected probes were interrogated using the Agena MassArray EpiTYPER system. Assay data for lncRNA and *POU2AF1* did not demonstrate significant differences in the EWAS interrogated CpG site (Figure 6-6 B-C). The EWAS interrogated *POU2AF1* CpG site trended towards significance (p=0.15). One benefit of using EpiTYPER is that it enables the interrogation of DNAm at sites around the CpG of interest (through amplification of a region for assay –see chapter 2.3.6-7). Therefore I was able to ascertain whether the EWAS probe was a proxy for another CpG site nearby. Indeed, for both *POU2AF1* and lncRNA assays, neighbouring CpG sites were significantly different between cases and controls.

Regarding the *RGS12* assay, despite looking at a number of neighbouring CpG sites, none were associated with younger aged oJIA (Figure 6-6D). In contrast, the *ARSA* assay interrogated 6 CpG units, of which 5 were significantly associated with disease (Figure 6-6A). This included the same CpG site that was interrogated in the EWAS.

A number of sites across assays were significant different with delta beta values in the same direction as the EWAS data. However, EpiTYPER derived delta beta values were smaller than that seen in the EWAS. Nonetheless, the technical validation work was able to support the EWAS data either by showing significance at the EWAS interrogated CpG site or at neighbouring CpG sites. Of note, EpiTYPER derived delta beta values were smaller than that seen in the EWAS.

The successful technical validation again justified the use of the selection criteria developed in this thesis. The resulting data suggested that the increased sensitivity of the age stratified

analysis resulted in genuine, technically validated data. To further corroborate this data, replication work was carried out.

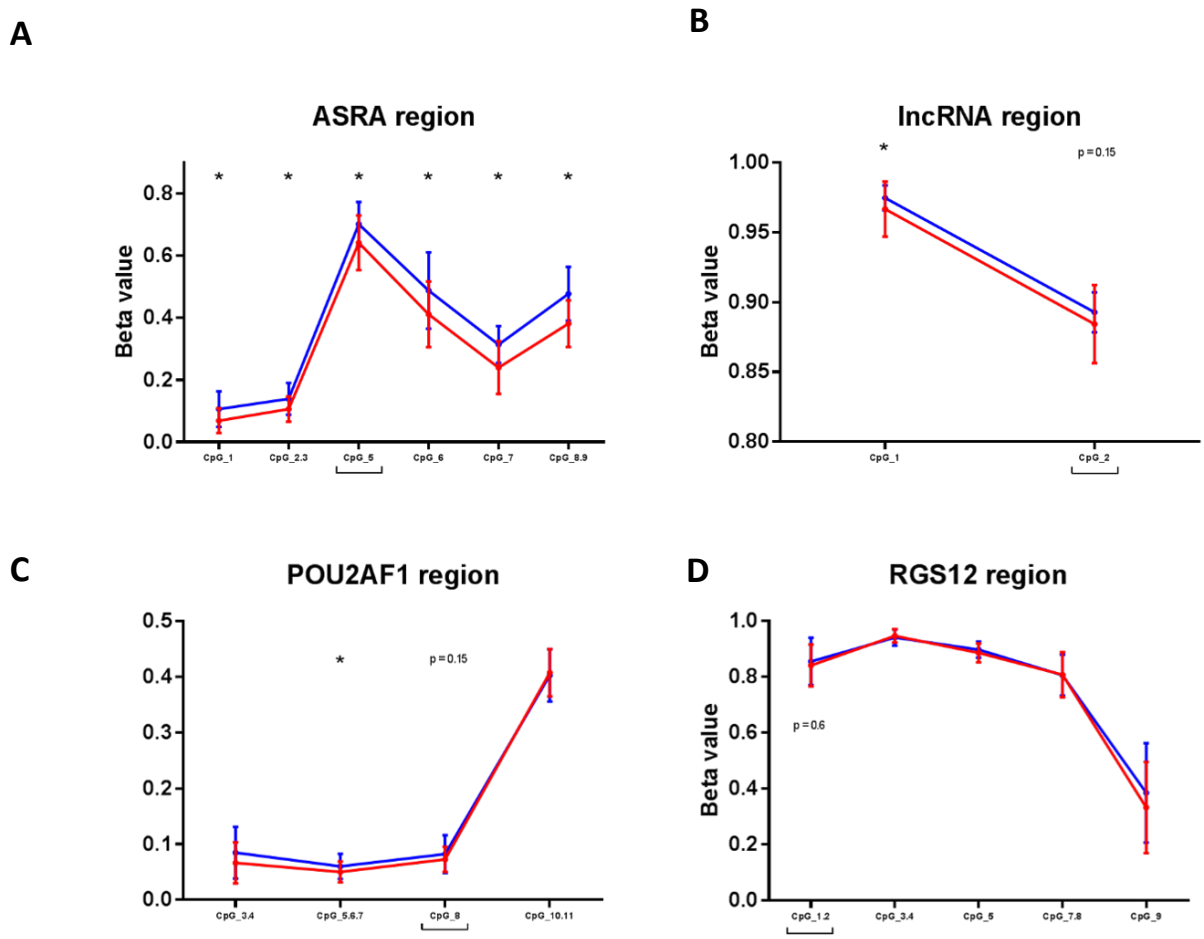

Figure 6-6 Plots of EpiTYPER methylation data from 4 genes selected for Technical validation.

Figures shown are A) ARSA B) lncRNA C) POU2AF1 D) RGS12. CpG site interrogated on the EWAS dataset is marked by brackets under the respective CpG sites, and p-values from the technical validation data are given for these CpG sites. Cases are in red and controls in blue. * indicates significance at $p<0.05$ (students t-test).

6.3.2.3 A preliminary replication study of younger diagnosed top hits again suggests a complex methylation landscape

A preliminary replication study was conducted, using a small replication sample group available from the CLARITY biobank. Again, as for previous CLARITY samples, JIA subtype classification was guided by the ILAR criteria [8]. Samples were selected if there was no prior exposure to methotrexate or bDMARDs, and must have had active disease at the time of blood collection. This followed previous selection criteria in order to avoid confounding due to medications known to affect the immune system or DNAm at CpGs (Chapter 2.1). Table 6-1 and Table 6-4 outlines the basic demographic data of each sample. Methylation analysis was completed using the MassArray EpiTYPER system, focussing on specific loci selected for technical validation, as noted above.

Figure 6-7 shows data from the selected loci, highlighting the significant CpG sites identified from EWAS data. The data presented is a meta-analysis including technical validation data (see section 6.3.2.2) and replication samples. No assay demonstrated significant associations with disease at any of the interrogated CpG sites. This included the single genome-wide significant site (lncRNA) and occurred despite observing significant differences at the technical validation stage in all but the *RGS12* assay (Figure 6-6).

For example, the *ARSA* assay displayed a number of CpG sites that were significantly different in the technical validation data (Table 6-5). Yet replication data displayed variability with some CpG sites having delta beta vales in the same direction as technical validation data, whilst delta beta values in other sites were in the opposite direction (Table 6-5). As mentioned, none of the CpG sites within the *ARSA* assay replicated in the small replication sample group available.

In addition to CpG loci present on the EWAS, additional CpG sites were assayed within ~200-300bp. Indeed, similar observations from the all-inclusive analysis were also observed here. Table 6-5 summarises the data and suggests that in addition to failing to replicate EWAS data, DNAm in younger diagnosed oJIA may still show some sample heterogeneity observed as variability in the data. For example, delta beta values across all 4 assays were either going in the same direction as the EWAS data, the opposite direction or was sometimes even variable within the one assay (*ARSA*). This complexity mirrors that seen in the all-inclusive analysis.

The meta-analysis including technical validation and replication data, again only *ARSA* and *POU2AF1* assays displayed any CpG sites with significant findings (Figure 6-7). These were noted in much the same CpG sites where technical validation was achieved. This suggests that the EWAS sample group (used for technical validation and larger in size vs the replication sample group) may be driving the observed significant differences in the meta-analysis. Therefore, results from this meta-analysis need to be interpreted with care.

Whilst there was a suggestion that the increased homogeneity of the younger diagnosed samples increased the sensitivity of the EWAS analysis, the sole significant finding could not be replicated. In addition, other sites that were selected did not replicate in the small sample group that was available.

The size of the replication sample available was perhaps the limitation in this replication study. Heterogeneity in the younger diagnosed subgroup would have prevented a small replication sample group from easily identifying the same association and was more likely to show variability in the data. This variability observed in DNAm suggests that additional case subgroups are worth exploring further, as these may help reduce DNAm variation and better identify the biological basis of oJIA.

The technical validation and replication results from the younger diagnosed age subgroup analysis contributed to a publication in January 2018 in the Journal of Autoimmunity, “The DNA methylation landscape of CD4+ T cells in oligoarticular juvenile idiopathic arthritis”, and is listed in Appendix III.

Table 6-4 Sample description table of replication samples for the younger diagnosed analysis.

Sample ID	Sample Group	Age	Sex	Matched Sample (age & sex) status (matched/unmatched)	Analysis	Persistent/Extended oJIA
J00037	CASE	3.2	F	Matched	Young only analysis	Extended
J00115	CASE	1.9	F	Matched	Young only analysis	Persistent
J00411	CASE	3.5	F	Matched	Young only analysis	Persistent
J00582	CASE	4.2	F	Matched	Young only analysis	Persistent
J00097	CASE	1.7	M	Matched	Young only analysis	Extended
J00584	CASE	2.9	M	Matched	Young only analysis	Persistent
J00530	CASE	3.7	M	Matched	Young only analysis	Extended
J00451	CASE	2.2	M	Matched	Young only analysis	Extended
J00354	CASE	6.8	M	Matched	Young only analysis	Extended
J00560	CASE	2.1	M	Unmatched	Young only analysis	Persistent
C00526	CONTROL	3.9	F	Matched	Young only analysis	~
C00397	CONTROL	1.7	F	Matched	Young only analysis	~
C00121	CONTROL	3.9	F	Matched	Young only analysis	~
C00593	CONTROL	4.1	F	Matched	Young only analysis	~
C00742	CONTROL	2.7	M	Matched	Young only analysis	~
C00015	CONTROL	2.5	M	Matched	Young only analysis	~
C00013	CONTROL	4.4	M	Matched	Young only analysis	~
C00059	CONTROL	2.8	M	Matched	Young only analysis	~
C00364	CONTROL	6.8	M	Matched	Young only analysis	~
C00421	CONTROL	3.1	F	Unmatched	Young only analysis	~
C00290	CONTROL	6.8	F	Unmatched	Young only analysis	~

Age is in years. Persistent/extended refers to the subgroup of oJIA that each sample belongs to.

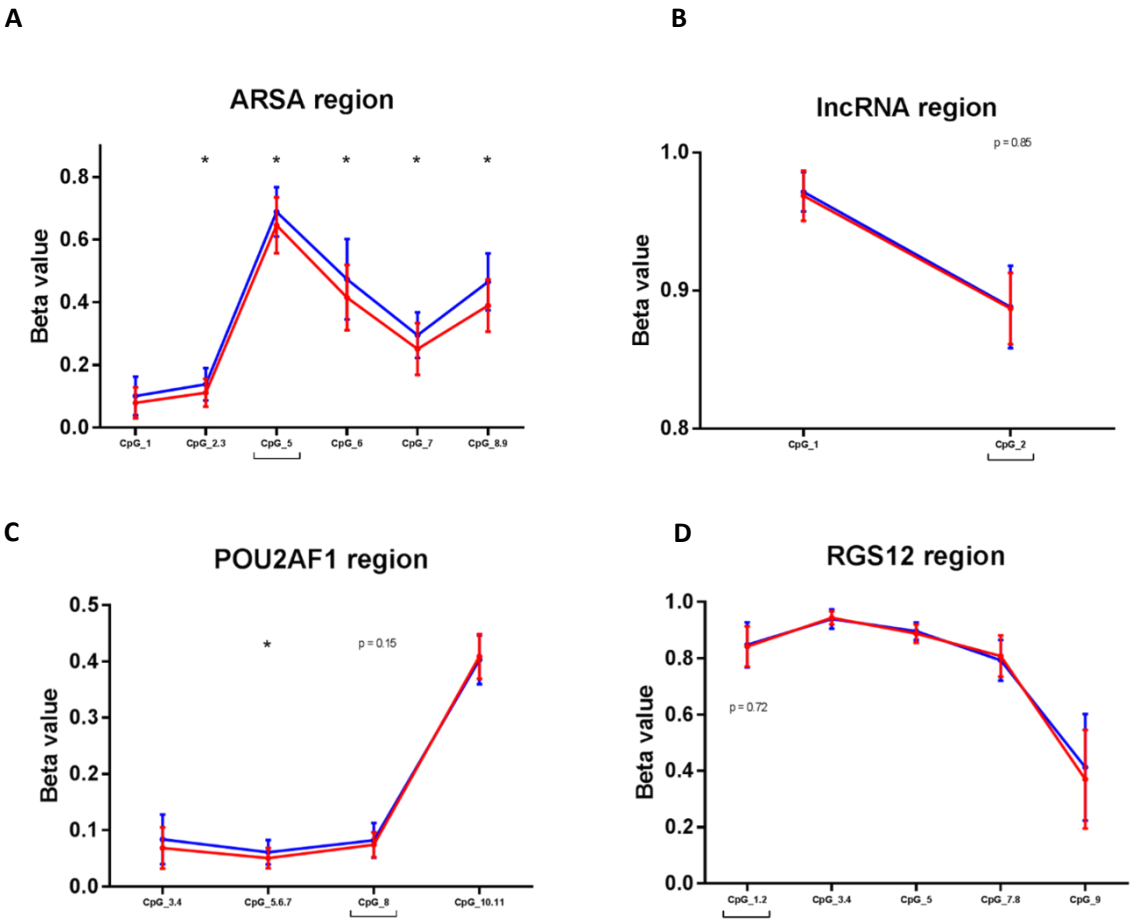

Figure 6-7 Plots of Meta-analysis of EpiTYPER methylation values from the 4 genes selected.

Data plotted is A) ARSA, B) lncRNA, C) POU2AF1, D) RGS12. CpG site interrogated on the EWAS dataset is marked by brackets under the respective CpG sites. Cases are in red and controls in blue. * indicates significance at p<0.05 (students t-test).

Table 6-5 Summary of the technical validation and replication work for each gene.

ID		Technical validation			Replication		
	EWAS p-value	**Same Δbeta Direction?**	**Sig. @ p<0.05**	**Notes**	**Same Δbeta Direction?**	**Sig. @ p<0.05**	**Notes**
lncRNA	0.01 (FDR adjusted)			CpG next to EWAS CpG is significant. ΔBeta is smaller			Opposite Δbeta direction
ARSA	0.0003 (unadj.)			All CpGs in same direction, all significant.			Variable from CpG to CpG. Some in same direction (very small Δbeta)
POU2AF1	0.0007 (unadj.)			CpG next to EWAS CpG is significant. ΔBeta is smaller			Same Δbeta direction, but much smaller Δbeta in surrounding probes.
RGS12	0.0015 (unadj.)			Smaller Δbeta			Opposite Δbeta direction, 1 CpG (non EWAS CpG) is significant

Green signifies validation/replication achieved. Red signifies validation/replication not achieved, orange signifies variable data. EWAS p-values were determined using the empirical Bayesian test. Technical validation and replication p-values were calculated using the student's t-test.

6.3.3 Exploring further clinical heterogeneity: Persistent vs Extended disease

To explore the potential for further sample group heterogeneity within oJIA affecting DNAm, another source of case group heterogeneity was analysed. That is, the subgroups of Persistent and Extended oJIA. The extended subgroup differs from persistent in being a more severe form of the disease. A patient can be sub classified into these groups from 6 months after initial diagnosis. A classification of 'Extended oJIA' is given if a child develops a total of more than 4 joints affected any time after the first 6 months of disease.

To explore these subgroups, only case samples were analysed. Extended disease cases were identified by clinicians guided by ILAR criteria. As this was an exploratory study, samples were not able to be matched according to age and sex Table 6-6. As was done for previous analyses, RUV was used to conduct a differential methylation analysis, comparing Persistent oJIA cases to extended oJIA cases. To allow for the potential to identify predictive markers of extended oJIA, only oJIA samples collected *prior* to an extended oJIA sub-classification were analysed. These sub-classifications can be made from 6 months post initial diagnosis with oJIA.

Therefore, 6 extended oJIA cases were excluded from the analysis. This was due to blood sampling occurring more than 6 months after an oJIA diagnosis.

Table 6-6 Summary of samples for the Extended vs Persistent subgroup analysis.

	Samples with blood <6 months post Diagnosis	
Subgroups	**Persistent**	**Extended**
No. Cases	39	11
Females/Males	31/8	8/3
Mean Age (SD)	5.51 (4.1)	6.09 (3.36)

Only samples with blood samples taken less than 6 months post diagnosis were used for this study. This enabled results to be potentially predictive of future disease extension.

Two questions could be asked using this data; 1) Are the persistent vs extended disease subgroups identifiable using DNAm (therefore potentially contribute to DNAm variation in oJIA) and 2) Can DNAm identify patients at risk of progressing to extended disease. A positive result to question 2 would provide proof of principle data for prediction of extended disease occurrence.

Using 11 extended vs 39 Persistent oJIA cases, 1 probe reached near significance at the genome wide level (FDR 0.19). This finding occurred despite having an even smaller sample group than the all-inclusive or the age subgroup analyses. This suggested that, again, there was an increased sensitivity in the analysis resulting from more homogenous sample grouping. Similar to previous analyses, MDS plots were used to explore the sample relationships using the top hits. Using just the top 27 probes (unadj. $p<5 \times 10^{-5}$), this was able to cluster persistent and extended disease subgroups (Figure 6-8). The Persistent-Extended disease variable was identified as the 1st principal component amongst the top hits. This suggests that this sub grouping is a strong component of DNAm adding heterogeneity to the case sample group. The heterogeneity resulting from these subgroups in oJIA may partly explain the lack of replication in the younger diagnosed and all-inclusive analyses.

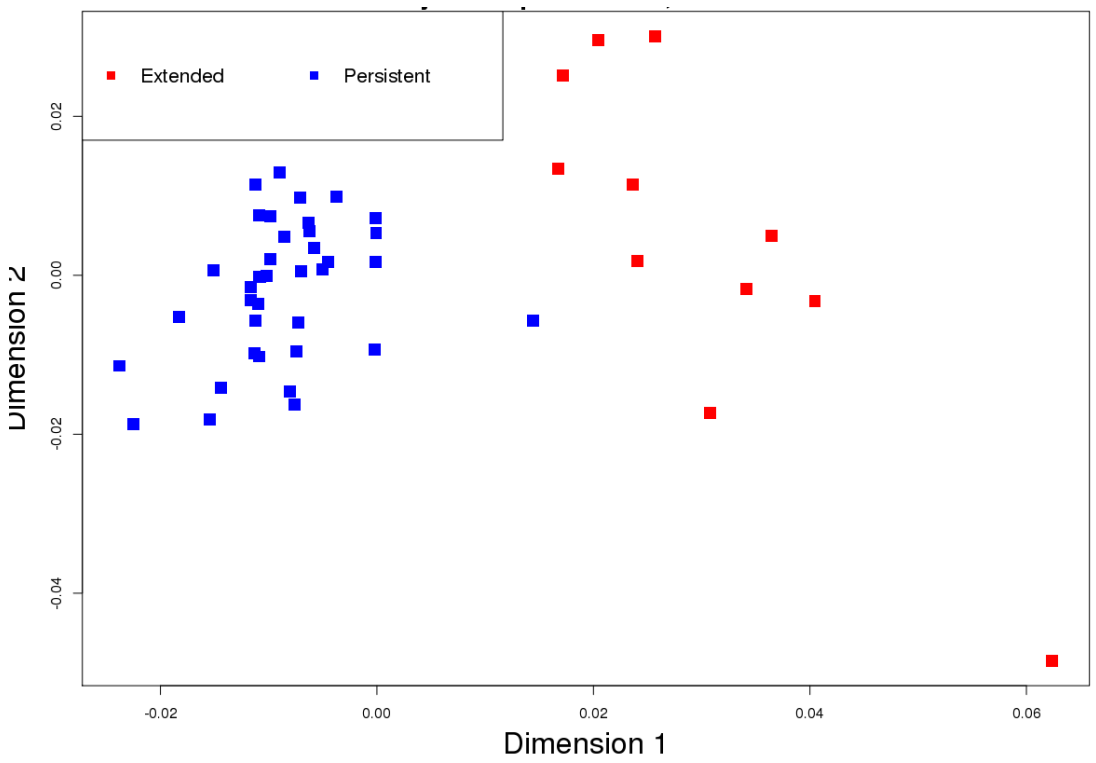

Figure 6-8 MDS plot of oJIA persistent vs extended cases. Top 27 probes (unadj. p< 5 x10^{-5}, 27 probes).

Probes from the RUV analysis were used to cluster samples. Extended cases are in red, Persistent cases in blue.

Further, since these extended oJIA cases were sampled prior to disease extension, this MDS plot is a potential proof of principle for prediction of disease outcome using DNAm. Even using non-genome-wide significant probes, DNAm data in this sample group contains a signature that identifies patients who will go on to get worse disease. This potential signature is detectable within 6 months of initial clinical presentation.

6.3.4 DNA methylation sites are sensitive and specific for predicting disease extension

To assess whether DNAm can be utilised as a clinically useful test, ROC curve analysis was carried out. Probes from 3 different p-value thresholds were used, utilised as these were all able to cluster Persistent from Extended-to-be cases. The best performing set of probes were those with $p<1 \times 10^{-4}$, with an area under the curve (AUC) of 0.93 (95% CI: 0.86- 1.0) Table 6-7. Indeed, the AUC did not fall below 0.75, with all p-values below $p<0.009$. The best performing set of probes also corresponded with a highly sensitive test, accurately identifying 89.7% of extended cases with a specificity of 72% Figure 6-9.

Table 6-7 Assessing the predictive value of DNAm status at disease diagnosis in identifying cases who will progress to disease extension.

Probes used (unadj. p-value)	Sensitivity %	Specificity %	AUC (95% CI)
27 probes, $p < 5 \times 10^{-5}$	71.7%	54.9%	0.66 (0.46-0.86)
58 probes, $p < 1 \times 10^{-4}$	89.7%	72.7%	0.93 (0.86-1.00)
241 probes, $p < 5 \times 10^{-4}$	97.4%	72.7%	0.88 (0.75-1.00)
490 probes, $p < 1 \times 10^{-3}$	89.7%	72.7%	0.76 (0.55-0.97)

P-values thresholds were selected due to their ability to cluster Persistent vs Extended-to-be cases in MDS plots. AUC: Area under the curve.

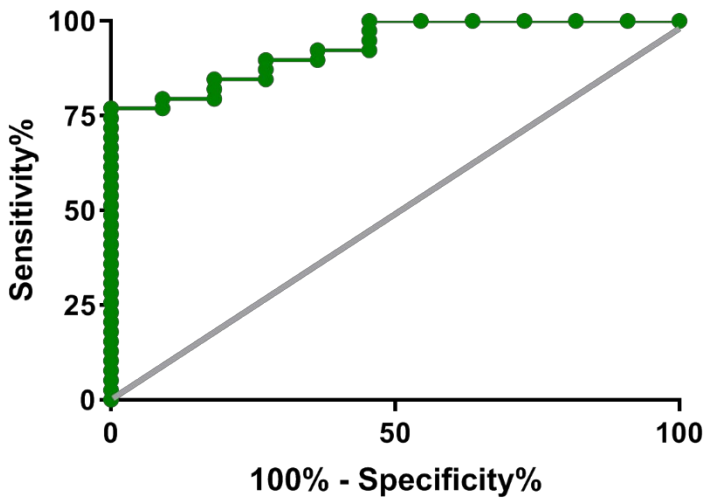

Figure 6-9 ROC curve of the top 58 probes ($p < 1 \times 10^{-4}$) in the Persistent vs Extended-to-be cases analysis.

Persistent (N= 39), Extended-to-be (N= 11). AUC=0.93 (Area under the curve). Straight line indicates values of a hypothetical random classifier.

6.4 Discussion

6.4.1 Heterogeneity in oJIA

JIA subtyping and the known heterogeneity in JIA as a whole is a point of contention within the literature. This debate is particularly robust in discussions on the oJIA subtype. Whilst oJIA appears as a clinically homogenous group using the current classification criteria, clinical and molecular evidence suggest that it may be representative of a number of different subgroups [12, 13, 17]. Indeed, some heterogeneity is already recognised in the oJIA subtype, namely the Persistent and Extended disease subgroups [8].

This study initially identified case-control clusters based on top hits from DNAm data (Chapter 4.3.2.1). Despite this clustering, age very clearly defined the spread in the data along the first principal component, suggesting that DNAm variation may likely be due to age differences. DNAm data appeared to reflect clinical and other molecular evidence pointing to age subgroups, therefore age stratified analyses were conducted.

Sub group analyses focussed only on younger or older diagnosed cases and controls (≤6 or >6 years age) resulted in far fewer top hits needed to cluster case-control groups compared to the all-inclusive analysis. This suggested that both subgroups consisted of more homogenous sample groups. This was highlighted with the finding of a probe reaching genome-wide significance (FDR 0.01) in the younger diagnosed analysis.

Comparing the analysis of older and younger diagnosed cases, older diagnosed cases appear to be a more heterogeneous group as suggested by 1) More probes from the top hits were needed to form case-control clusters than in the younger diagnosed analysis (248 vs 31 probes) and 2) MDS plots of older diagnosed cases demonstrated a clustering pattern defined by 2 potential variables i.e. In addition to case-control grouping, another factor appears to disperse these clusters. This additional factor could not be defined utilising the available sample data. Lastly, 3) Both points likely contributed to a sample group that was more heterogeneous, highlighted by an analysis that was unable to detect genome-wide significant hits.

The grouping of oJIA by age subgroups of ≤6 years and >6 years, both in this study and other studies related to oJIA, may reflect differences in a developing immune system. When looking at the CD4+ T-cell compartment specifically, T naïve cells gradually decrease after

the 1st year of life [429]. Inversely, T memory (Tmem) cells increase in that same time period, quickly increasing in numbers after 6 years of age [429].

The Tmem cells may also explain the reduced homogeneity in the older subgroup. Within Tmem cells there are further subsets (e.g. Th1, Th2, and Th17) and variability in the subset numbers will produce a more variable CD4+ T-cell compartment. Since Tmem numbers increase with age, this variability will be particularly evident when analysing the CD4 compartment in older age groups. In this study, Tmem variability in older diagnosed samples may be contributing to the relatively difficulty in clustering cases and control compared to younger diagnoses samples. The relative homogeneity of the younger diagnosed samples may reflect a smaller Tmem population where variability in cell subsets cannot have as marked an effect on DNAm.

Other studies appear to support this hypothesis. For example, DNAm in Tmem cells were found to be more variable compared to T naïve cells across a number of cytokine genes [455]. In addition, the known subsets of cells (Th1, Th2 etc.) within Tmem cells can be differentiated by DNAm, most notably in genes of subset specific cytokines [256-268, 456]. Therefore age subgroups in EWAS data may be partly explained by differences in numbers or function of T cell subsets.

Nonetheless, this study's EWAS data suggested DNAm may mirror or act as act as a proxy for the underlying biological differences between subgroups. This was highlighted by a highly specific DNAm signature defining age subgroups, demonstrated by a lack of concordance between subgroup DNAm signatures. The cluster analysis also suggests a highly specific signature since a subgroup's signature could not accurately identify cases from the other subgroup.

Regarding the technical validation efforts, a smaller Δbeta value was observed compared to EWAS data. This may be due to the technical limits of the EWAS platform which is less accurate within intermediate levels of DNAm (>0.2 and <0.8), with average Δbeta values of ~13% detectable [333, 365] [366]. However, differences less than 5% have been replicated, particularly in other AID EWAS' [245, 308, 313, 324]. However, smaller delta beta values were generally obtained from larger discovery samples, as well as larger replication sample groups, and this may be a shortcoming of this study.

Despite reducing sample group heterogeneity by focussing on age subgroups, difficulties were still observed in replicating the data. This may be due to the small replication cohort that was available. This also suggested that whilst age certainly appeared to be a major factor contributing to disease heterogeneity, there may be other factors contributing to oJIA heterogeneity and the complexity in DNAm observed.

To further understand the complexity in DNAm landscape in age subgroups, larger studies focussing on a specific JIA subtype are needed. This would assist in understanding the biological basis of the clinical and early molecular evidence which groups oJIA cases according to age, and may support efforts to adjust JIA classification criteria to take age into account.

This data supports further studies into the underlying differences in age specific immuno-biology that may contribute to identification of age specific targets for treatment and prognosis. However, future studies in DNAm must take into account sources of confounding, as was done in this study, in order to avoid type I errors and observe true differences in the biology of JIA in different age groups.

6.4.2 Potential of DNA methylation to predict disease outcome

Whilst a reduction in oJIA heterogeneity was observed when age stratified analyses were conducted, replication studies suggested that there was still some variability at CpG sites that may be due to other sources of disease heterogeneity. To explore this, case subgroups were analysed for differential methylation that may contribute to disease heterogeneity. Persistent and Extended oJIA known groups within the oJIA subtype, and differential methylation analysis revealed a CpG site that reached near genome wide significance (FDR 0.19). MDS plots of top hits revealed persistent vs extended clustering was achievable using just the top 27 probes, with the case status defining the primary reason for this grouping. These two pieces of evidence suggest the persistent vs extended subgroups have a strong signature present in DNAm data. Firstly, one CpG site was near genome wide significant, despite small sample numbers. Secondly, clustering of case groups needed fewer probes than the all-inclusive analysis to identify clusters, and fewer than the younger diagnosed samples.

Persistent and Extended oJIA has been associated with a number of different molecular and clinical markers. Initially, clinical markers such as symmetrical affected joints at initial

presentation, higher erythrocyte sedimentation rate as well as wrist and ankle involvement were associated with disease extension [66]. In the following years, a number of papers highlighting immunologically related differences between the sub groups were published. These included multiple antibody positivity (e.g. to CCP, type II collagen and rheumatoid factor), immune cell differences in the joint (lymphocyte expression of CD73, CD4:CD8 ratio, expression of inflammatory and macrophage differentiation genes) and genetic differences relating to T cell stimulation (*VTCN1*) [45, 457-459]. Not only do immune system perturbations in general appear to reflect persistent vs extended differences, CD4+ T cell compartment differences are also associated with these subgroups. Blood and joint CD4+CD25+(bright) T cells are associated with persistent disease, with IL-17+ T cells in the joint associated with extended oJIA [44, 48]. Therefore, it is likely that any differences in DNAm between Persistent and Extended disease reflects both specific CD4 T cell compartment differences and generalised immunological differences.

Persistent and Extended oJIA clusters in MDS plots of EWAS top ranking probes ($p< 5 \times 10^{-5}$) identified these subgroups as a potential source of disease heterogeneity within oJIA, potentially leading to DNAm heterogeneity and variation. As was observed for the age subgroups, it appears that DNAm partly reflects the clinical heterogeneity. Further, DNAm possibly reveals some of the underlying differences in biology between these two subgroups. Indeed, large variations in DNAm are observed in the only other JIA EWAS performed outside of our group. JIA cases who continued to have active disease after medication withdrawal have large interquartile ranges compared to those who will have inactive disease [51]. The authors suggest that this, too, may be driven by immune cell disturbances, particularly T-reg cell clonotypes within the CD4+ T cell compartment. Thus, another example exists of DNAm variation elicited by subgroups within JIA cases (i.e. response to treatment subgroups), supporting the potential for this phenomenon's occurrence within the current study's dataset.

This DNAm variation between oJIA subgroups such as age and persistent vs extended disease may help explain the difficulty in replicating top ranking probes in this study. These factors likely make up some of the major sources of variation that adds complexity to identifying robust DNAm associations in JIA. Future JIA EWAS' will need to take into account the variations due to these subgroups and include higher numbers of cases to accommodate them, or make use of more developed analysis techniques designed for identifying DNAm associations in complex diseases. An analysis segregating Persistent vs Extended disease

according to age could not be done in this study due to the very small numbers available for analysis (particularly in the older subgroup of extended disease patients). This shortcoming may be worthwhile following up, since both these factors do appear to be relevant to DNAm and may contain specific biological differences which could be targeted for intervention.

Besides potentially identifying subgroups, the persistent vs extended analysis also revealed that DNAm may potentially contain predictive markers of disease course. Extended samples were collected within 6 months of diagnosis i.e. prior to disease extension. Therefore, subgroup clustering using the top ranking hits from the EWAS data likely reflects biomarkers present before disease worsening. Indeed, ROC curve analysis suggests DNAm data may be an accurate predictor of disease extension at initial presentation.

Prediction of extended disease is clinically relevant and important for management of disease. Those with extended disease have a poly-articular course disease, thus are more likely to have the worse long term outcomes associated with that subtype. These include increased likelihoods for medication exposure and longer periods of active disease, as well as being less likely to achieve inactive disease compared to oligoarticular patients [39, 70, 460, 461]. Work to identify markers of prognostic value that are clinically useful, reliable and convenient is limited. Whilst specific cell types and their functions are potentially predictive of disease course, access to joint samples and sufficient numbers of cells can be difficult. Thus translating bench research to bedside test is hampered. Clinical markers of disease extension e.g. wrist and/or ankle involvement, despite being highly specific, are not yet reliable markers due to the potential for many false negatives [66]. Therefore a space exists for an objective test, one requiring an easily collected biospecimen and able to quickly return results to clinicians. DNAm based testing may be a solution.

Prior studies give credence to the use of DNAm for prognostic or diagnostic purposes. JIA patients who maintain inactive disease 8 months after medication withdrawal show potentially predictive signatures using DNAm [51]. Further, there are a number of commercially available DNAm based tests for cancer diagnosis including colorectal, lung and prostate cancer [462-471]. Whilst many milestones remain for development of an oJIA prognostic test [472], both the groundwork for acceptance of DNAm based tests and supporting evidence for the ability of DNAm to detect JIA subgroups suggest that such research can be successful, clinically relevant and potentially commercially viable.

This data therefore supports further work on proof of concept studies towards a diagnostic test for disease prediction at clinical presentation. Further work should include larger studies to identify robust markers of disease [472]. Samples should be collected at disease presentation in the clinic. In addition, a more convenient DNA source such as whole blood would reduce labour and costs of implementing such a test. Clinician driven indicators of development for such a test would allow alignment with current clinical practice, therefore increasing the value perceived by other clinicians, and foster uptake and dissemination of the test. Large, multicentre collaborations would be ideal for identification, replication and development of this work.

Chapter 7. **General Discussion**

7.1 Summary and discussion of key findings

7.1.1 Summary of the EWAS findings

This study was an exploratory study of oJIA DNAm in autoimmune relevant CD4+ T-cells. Prior to the study, publications in JIA had hinted at the potential for epigenetic aberrations to play a role in aetiology [44, 106, 269, 292]. In addition, differentiation of cells of the adaptive immune system and cytokine expression are known to be marked and sometimes regulated DNAm, adding further weight to the potential for DNAm to play a role in autoimmunity [252-255, 262, 264, 266-268, 376, 377, 419, 456]. Studies in other AID have indeed reported many associations with DNAm [51, 245, 281-284, 286-302, 304, 306-319, 423, 473-475]. However, many studies in AID have not accounted for many sources of variation such as medication-exposed samples, sex, age, highly variable disease duration and cell type heterogeneity [51, 287-291, 293, 295, 299, 300, 422]. This leads to heterogeneous sample groups and sources of confounding in reported associations. Therefore, this study explored the potential for an oJIA EWAS to identify aberrant DNAm free from many of these confounders.

The aim of this study was complete the largest JIA case-control analysis of genome scale DNAm to date. JIA cases that were MTX exposure naïve, of the same clinical subtype and all with active disease allowed for an analysis of clinically homogeneous samples. The main hypothesis tested was that DNAm differences exist in oJIA cases compared to age- and sex-matched controls. This study found that not only can cases and controls be clustered into discreet groups using DNAm data, but associations within biologically relevant pathways and functions can be detected, despite initial findings not being statistically significant. Relevant pathways identified were immune related, with links to receptor or intracellular effects of receptor binding such as *IL17*, *IL6* and *MHC* class II receptor activity. Cell death and survival, as well as cellular growth and proliferation, were higher level functions of some of these top ranking hits. In addition, gene expression data suggested similar pathways and functions correlated with those found with differential methylation.

Based on a selection criteria developed earlier in this study, five CpG sites were selected for technical validation: *CHD5, CYP1A1, SYNE1, TACSTD2* and *VPS53.*

The criteria used for selection of CpGs for further study was aimed at identifying top ranking CpGs that were most likely to be genuine associations and to technically validate. Functional

relevance of selected genes was secondary to this purpose, yet probes selected still had potentially relevant roles in immune function or within epigenetic regulation. *CHD5, CYP1A1, SYNE1* and *TACSTD2* all have reported or potential roles in immune system signalling or response, corresponding to enriched pathways identified [401, 405] [407] [476]. Further, both *CHD5* and *TACSTD2* expression has been reported to be modulated by DNAm. For example, *CHD5* reports strongly suggest DNAm is associated with a number of cancers, potentially acting as a master controller of a number of oncogenes [477-486]. Regarding DNAm associations in other selected genes, *CYP1A1* DNAm is often reported as modulated by smoking, an environmental risk factor in RA and currently a risk factor of debate in JIA [215-217, 437]. In addition, *TACSTD2* has been linked non-causally to childhood adiposity, and is known to be silenced via DNAm in some cancers [487-490]. *SYNE1* DNAm has also been associated with colorectal and lung cancers as well as leukaemia [491-494]. Thus, evidence suggests that there is potential for these genes to play a role in immune cells, and the reported associations with DNAm in other contexts suggests the DNAm associations in this study have precedence and may indeed be worth investigating further.

Technical validation was successful in 5/5 probes selected, with proximal CpGs also showing significant associations with oJIA. DNAm was also observed to be widely distributed at most probes validated, suggesting factors affecting oJIA DNAm may be complex. A small replication analysis was attempted in this study that suggested findings could not be replicated. The large variation in DNAm observed in most probes, likely due to biological sources, may have contributed to this.

Since technical validation was highly successful, biological sources of complexity needed to be identified e.g. sample/biological heterogeneity. Whilst the study design attempted to avoid biological confounding due to medication exposure, sex, age, disease subtype, disease activity and disease duration, it was clear some sample heterogeneity remained. Age and genetics were postulated as potential sources of sample heterogeneity potentially affecting DNAm that may also have affected replication efforts. Indeed, whilst samples were age matched and patients were clinically homogeneous according to current classification criteria, age has been postulated as a key source of subtype heterogeneity in JIA for over 15 years, and beyond, extending to revisions of classification criteria in the 1990s [17, 50, 135].

Indeed, should genetics be associated with DNAm by changing the mean or variability of DNAm, small replication analyses have the potential to be biased due to sampling error. As

such, genetics may also be a large component behind the complexity of DNAm and the difficulties in replicating this study. Therefore, age and genetics were the basis of further JIA DNAm investigation in this study.

7.1.2 Summary of MQTL analyses

To test the genetic contributions of sample heterogeneity, this study then explored genetic relationships to mean and stochastic variation in DNAm. This focussed on known mQTLs and identifying potentially new mQTLs in regions up to 100Kb from the CpGs of interest.

This chapter hypothesised that genetic variation may modulate DNAm, not only by affecting mean DNAm but also DNAm variability, irrespective of case or control status. Therefore, all cases and controls were combined for analyses of genotypic associations with DNAm.

Known mQTLs originating from studies in both young children and adults were replicated in this study, suggesting these may indeed be robust mQTLs [222, 251]. Potentially-new mQTLs were also identified in this study. Both known and potentially-new mQTLs were associated with mean DNAm and DNAm variance. Compared to SNPs generally closer in proximity, enhancer-located SNPs were the most robustly associated mQTLs in analyses of both mean DNAm and DNAm variability.

These analyses suggested differences in mean DNAm between genotype groups were substantially bigger than any differences between case and control groups. Mean differences in DNAm from known mQTLs were as large as 13% for *CHD5* and 27% for *TACSTD2*. This was many fold higher than differences observed between case and control groups (typically 5% in selected probes). This suggests genetics may indeed be a contributing factor to DNAm heterogeneity, by acting as a source of sample group heterogeneity in oJIA and in controls. Therefore, a strong potential arises for genetic factors as key contributors to difficulties replicating this study's EWAS.

The most robust mQTL associations were found within the known mQTLs (*CHD5* and *TACSTD2*). These mQTLs were identified in immune cell based EWAS' i.e. in a neonatal sample group and another in an older cohort of Scandinavian people enriched for females [222, 251]. *CHD5* is preferentially expressed in the nervous system; however, DNAm associations are reported in a number of cancers of various cell types including colon, breast as well as an immune cell cancer (leukaemia), suggesting expression of *CHD5* across a wide

range of cell types [477-486, 495, 496]. In addition, potential links to interferon gene expression suggest expression in immune cells may well occur [401]. CHD5 interacts with the NuRD (nucleosome remodelling and deacetylase) complex, a DNA remodelling complex involved in epigenetic regulation via histone deacetylase activity [496]. Through the NuRD complex, it is likely to modulate a number of oncogenes in renal cell carcinoma [477, 496]. Despite the number of DNAm associations between CHD5 and cancer, no mQTL analyses in cancer has been reported to our knowledge. Our study validates these known mQTLs in a disease setting in addition to the population data reported, and suggests that further mQTL studies may be worthwhile not only in JIA but also in cancer settings.

TACSTD2 is best recognised as the gene harbouring mutations causative of Gelatinous Drop-Like Corneal Dystrophy [497]. However, more recent reports identified it as having potential roles in cancer through regulation of cell growth and metastases [498]. Studies have also identified TACSTD2 as regulating important immune modulators such as STAT1, STAT3 and NF-KB [407]. It is also reported to identify a subset of dendritic cells in the epidermis [499, 500]. Much like *CHD5*, associations between DNAm and cancer are reported particularly in malignancies of liver, lung, and potentially in stomach and colorectal cancer [488-490, 501]. The links to both the immune system and DNAm modulation suggest TACSTD2 may play a role in JIA, potentially as a functional component in disease or as a biomarker relevant for identifying disease.

This is the first study of mQTLs in JIA, to our knowledge. Replication of known mQTLs in this study validated the robust nature of this EWAS, and also suggested probes of interest in this study may have genetics as a potential controlling component. Interestingly, 2-5% of EWAS CpGs are typically reported to have an mQTL association [219-229]. In AID EWAS', this number rises generally to about 12-16% of significant CpGs [284, 286, 307, 308]. A study in breast cancer identified ~25% of CpGs were likely to have genetic associations with short range SNP, and 1000 CpGs (at least 0.2%) with longer range SNPs [502]. CpGs with a genetic component may even be up to 48.5% [503]. The genetic contribution to DNAm variability has been estimated using heritability rates, particularly in twins, with genome wide heritability as low as 3% but typically ~16-20% reported [349, 504, 505]. These are similar to numbers from an extended family study [503]. However, heritability estimates contain high variability likely due to heritability differences in specific regions of the genome such as the MHC region [503, 506] or higher estimates at a number of individual CpGs (where heritability is >50%) [507], and also due to cell type [508, 509].

In accordance with studies reporting higher heritability, in this study 3 of 4 probes investigated were associated with an mQTL (and the 4th passing one of two sensitivity analyses). This is more in line with mQTLs reported in RA with 30% and 80-90% of DMPs identified being associated with an mQTL, and also aligns with reports of ~49% of CpGs with a genetic component [503] [310, 319]. The stronger presence of mQTLs perhaps reflects an arthritis-specific phenomenon. It may also reflect the potential for a very strong genetic component to JIA susceptibility, which may extend to genetic control of DNAm associations in JIA. Genetic risk loci acting through DNAm have been observed in another paediatric autoimmune disease, paediatric IBD [286]. Somineni *et al.* reported these methylation mediated risk variants despite using just half the sample size of very similar adult IBD study [286]. This study, together with Somineni *et al.*'s data, suggest that paediatric EWAS studies may have a stronger genetic component to take into account in future studies. Indeed, in this study, minor alleles associated with DNAm in this study are also moderately prevalent in the population (minor allele frequencies ranging from 0.25-0.37), lifting the potential for sampling biases to affect genotype sampling (thus, DNAm). These genetic sources of DNAm heterogeneity may partly explain the large variation in the JIA EWAS.

MQTLs associating with DNAm variation aligns with Feinberg's hypothesis on the genetic origins of stochastic variation [236]. Indeed, 2 other autoimmune EWAS' also identified genetically controlled DNAm variation, one in SLE and another in a disease highly related to JIA, rheumatoid arthritis [245]. In this study, many of the robust variation associated mQTLs were in enhancer regions. Of particular relevance, regulatory regions were also postulated by Feinberg *et al.* as potentially housing just these types of mQTLs [236]. Of particular interest, Feinberg *et al.* also hypothesised these mQTLs acting as risk loci in complex diseases [236]. This hypothesis found support in a large RA EWAS [245]. This study also suggested that mQTLs affecting DNAm variation are potential risk loci for oJIA. However, many of these mQTLs also affected mean DNAm. It is unknown what the interaction is between mean and variable DNAm. However, recent studies have begun exploring this and suggest closer investigation of these mQTLs be undertaken before DNAm variance mQTLs are robustly reported [510].

Enhancer-located mQTLs have been reported in a number of cancer types [248-250]. However, enhancer mQTLs in AID have not been reported widely, and may represent and avenue for investigation in the field. The finding of enhancer located mQTLs underscores the importance of enhancer regions to autoimmune genetics, and particularly to oJIA [149].

Indeed, many susceptibility loci for complex diseases have been associated with enhancer regions [112, 149]. With little understood regarding disease aetiology and mechanisms behind known risk loci in JIA, mQTL analysis (perhaps focussed on enhancer SNPs) may help unravel key biological processes behind disease, much like some had intended mQTL analyses had intended in the post-GWAS era [237].

Regarding mQTL links to disease, mQTLs identified in this study were found to be potential susceptibility loci for oJIA, mirroring rheumatoid arthritis data [245]. Whilst this study has not looked into the causal relationship between SNP, DNAm and disease, just five EWAS studies have done so [245, 284, 286, 308, 437]. Most find a very limited number of SNPs associated with disease acting through DNAm. More comprehensive studies investigating DNAm as a potential mechanism for susceptibility loci may well be the next step in AID EWAS.

This study suggested that even in clinically homogeneous samples, genotypes may contribute to DNAm heterogeneity in oJIA. This source of biological heterogeneity may contribute to or underlie differing disease processes, and may contribute to differences in disease progression and response. Further studies would be valuable in identifying gene-methylation interactions in oJIA, to identify robust associations and potential mechanisms behind genotypic associations.

7.1.3 Summary of oJIA subgroup heterogeneity studies

Next, this study looked at other potential sources of biological heterogeneity and hypothesised that within-subtype heterogeneity existed even within this clinically homogeneous group. Ravelli *et al.* and Martini *et al.* have advocated for new subtypes, suggesting current subtypes have heterogeneity based on age together with positive ANA status [17, 50]. Indeed, their studies reported patients identified by these factors go on to have poorer outcomes [17, 50]. Interestingly, these age subgroups were previously identified as a potential criteria for JIA classification in the 1990's [135]. Further, oJIA already has recognised subgroups within it i.e. persistent vs extended disease. These are likely to contribute to clinical heterogeneity, perhaps even at or near clinical diagnosis (when samples in this study were collected) [44, 45, 48, 458, 511]. This may be reflected in underlying biology such as DNAm on a genome scale. Therefore age and known oJIA subgroups were investigated for associations with DNAm.

Age subgroup analysis, particularly for the younger subgroup (≤6 years), exhibited better case-control clustering than the overall analysis, requiring far fewer CpGs. A genome wide significant hit was even detected in the younger analysis, despite an analysis half the size of the original EWAS analysis. Biological differences between age subgroups were potentially observed, since top ranking probes from the younger analysis could not cluster cases from the older subgroup (and vice versa).

Findings from the younger aged subgroup were technically validated in three out of four selected probes; however, these did not replicate in a small replication sample group. This suggested more biological heterogeneity may underlie age subgroups in oJIA.

This study suggests DNAm analysis can identify subgroups within a clinically homogeneous disease subtype. This ability to identify subgroups in oJIA with DNAm data mirrors the identification of subgroups within SLE, RA and IBD cases, reporting subgroups based on need for surgery, severity of disease and future need for treatment amongst others [282, 284, 308, 311]. The sensitivity of DNAm to detect these groups may actually increase difficulty of replicating findings from a sample group that may not be biologically homogeneous, despite the clinical homogeneity. This biological heterogeneity may explain the DNAm variation, as well as the difficulties in replicating findings.

The identification of age subgroup associations, which were more robust in the younger aged subgroup, is aligned with recent advocacy aimed at utilising age as a classification criteria in JIA [7]. It also aligns with gene expression data, suggesting biological differences do underlie age subgroups. However, disease mechanisms specific to age groups have not been fully elucidated. Whilst gene expression data suggests B-cells are important, this study in CD4+ T-cells suggests age subgroup effects may function through these cells [12]. Indeed, CD4+ T-cell subset differences do begin to emerge after 6 years of age, as do differences in immune responses (as observed for influenza) [429, 430]. These differences may be the mechanism behind age subgroup differences and partly contribute to oJIA heterogeneity, which may be reflected in DNAm variation.

This study contains an exploratory analysis which looked at another source of biological heterogeneity i.e. persistent vs extended-to-be disease. This suggested a DNAm signature is present in oJIA cases prior to disease extension and presentation of worse outcomes. This was proof of principle for a prognostic test for identifying patients at or near initial clinical

presentation who will go on to have worsening of disease. Some studies have explored markers predicting disease extension, but few have reported biological predictors of disease. Much like this study, CD4+ T-cells were important for predicting disease extension in Hunter *et al.*'s study [45]. The CD4:CD8 T-cell ratio from synovial fluid and interferon-γ expression (also produced by CD4+ T-cells) were associated with predicting disease extension [45]. In addition, Th17 CD4+ T-cells were enriched in extended oJIA joint samples, as are a subset of Tregs [44, 48]. In contrast to CD4+ T-cells, protein and gene expression data suggested macrophages and CD8+ T-cells were key cell types in predicting disease extension [44, 45]. This was highlighted by expression of markers of effector macrophages, as well as CCL5 or CD73 expression in CD8+ T-cells [44, 45]. Autoantibodies from B-cells may also hold prognostic power, as multiple autoantibodies were associated with extended oJIA [458]. In another study, vitamin D binding protein was also associated with extended-to-be oJIA [511]. However, many of these markers were identified in synovial fluid, which is not always accessible early in disease i.e. before the onset of poorer outcomes. Few have been identified in peripheral blood, an easier source of sample for such tests [458].

Nonetheless, even in such an exploratory study as this, sensitivity of the analysis appeared to increase as observed by near genome wide significant hits. Therefore, this subgroup of oJIA may potentially be a source of DNAm variation of relevance for this study and future DNAm EWAS' of oJIA. DNAm data may even be sensitive enough to identify extended-to-be oJIA patients, providing clinicians a tool to enable better clinical management of disease prior to joint damage and other co-morbidities.

7.2 Limitations

This study attempted to increase the sensitivity and clinical relevance of the analysis by focussing on a homogeneous clinical sample group at or near clinical diagnosis. This robust process may also have removed causes of significant DNAm associations. The lack of statistically significant findings in this study may indeed reflect a trend of similar reports in AID EWAS, where more powered and better designed analyses result in fewer associations [245, 284, 288, 295, 301, 308]. Much like the present study, these studies take into account potential confounding due to medication exposure, disease duration, cell type, and in many cases also genetics.

However, whilst numbers in this study were increased compared to prior efforts in JIA, numbers were still substantially smaller than other AID EWAS' by a factor of ~5-7 fold (compared to the highest powered studies in the field) [245, 284, 307, 308]. It is likely that this EWAS study was underpowered to detect small differences in DNAm values. To ensure sufficient power in future studies, larger numbers will be needed to ensure small differences in DNAm are detected. However, what such numbers should be is not entirely clear. One could use more recent studies in AID, or more specifically from paediatric AID EWAS' using immune cells, to estimate an appropriate sample size. IBD studies may help in this regard, suggesting approximately 36-164 cases are required in order to replicate findings [282, 286, 313].

However, it is clear that unlike GWAS studies where power calculations are robust and standardised, similar methodologies in EWAS have not been robustly developed and tested. This stands not only for power calculations but for also defining thresholds for significance and analysis methods. Indeed, recent discussions are not prescriptive and suggest a number of potential methods for design and analysis of DNAm studies using EWAS [512, 513]. However, this may be more a function of the dynamic nature of DNAm, as much as a shortcoming of analysis methods [512, 513]. Nonetheless, this remains a limitation for the EWAS field in general.

In this study, replication studies were performed for top ranking CpGs from the overall and age subgroup analyses using relatively small sample numbers compared to the EWAS. Inadvertent sampling bias may have skewed results. This may be particularly true in this study, where a wide distribution of DNAm exists for most CpGs of interest.

There is also the potential for medication e.g. non-steroidal anti-inflammatory drugs, to have affected DNAm. Prior to seeing rheumatologists for a diagnosis (thus, prior to blood sampling in this study), many parents are likely to have attempted to reduce inflammation using over the counter medication. Methods do exist to increase the number of JIA cases being completely medication naïve, including recruitment through community setting or GP clinics, or large prospective cohort studies. However, these require huge co-ordination efforts, personnel and funding to do effectively.

The mQTL analysis was an exploratory analysis and will require large numbers to validate. Correction for multiple testing was not used in this study due to the small sample sizes

available; however, sensitivity analysis was used to build evidence for the robustness of the associations. Whilst the small numbers limit the applicability of this data to the population, a strength of this study was the assessment of both mean DNAm and DNAm variance which opens up interesting directions for future research in JIA.

This study selected tissue specific enhancers in regions nearby CpGs of interest. However, pan tissue enhancer regions may also be relevant. Indeed, Feinberg's hypothesis on genetic regions housing mQTLs only noted regulatory regions as a general region, which could include DNAse hypersensitivity sites, transcription factor binding sites amongst many others [514]. Future studies could investigate enrichment of mQTLs within other regulatory regions.

In addition, enhancers no further than 100Kb from the CpG were used. Whilst a majority of mQTLs are located in that window [221-229], mQTLs associations have been reported to be as far as 300Kb from CpGs [222]. MQTLs of interest may yet exist in regions beyond 100kb from CpGs in this study.

This study was able to validate known mQTLs from the literature. This may have been due to the analogous nature of the source studies to this EWAS i.e. use of immune cells e.g. blood or immune cell subsets with HM450k data. MQTLs reported from non-immune cell studies may also be pertinent to DNAm in this study.

The JIA association with potential and known mQTLs could not be analysed further for causation vs consequence due to the limited sample sizes available. Therefore, it is unknown whether DNAm is mediating a relationship between genetics and disease, or coincidental to it, or a result of reverse causation i.e. the consequence of genetics x disease association. This aspect needs to be explored further using causal inference testing, or perhaps through Mendelian randomisation studies, and bigger sample numbers to comprehensively assess this would also be required.

The analysis of age subgroups within oJIA held the potential to increase the sensitivity of the analysis. However, this required sample numbers to reduce even further. Whilst a genome wide significant hit was identified and technically validated, larger analyses may reveal additional genuine associations. A small replication group in this study did not replicate findings from the age sub group analysis. Larger sample numbers would also be required to comprehensively replicate findings.

This study identified age as a large source of variation in the data. However, age may only be a proxy for another variable e.g. differences in T cell subset numbers, or responsiveness of subsets to stimuli [429, 430]. Analyses that take into account variables known to vary with age will be important in future JIA EWAS, and potentially with any future paediatric EWAS.

The persistent vs extended oJIA analysis provided intriguing insights into the potential for DNAm to identify patients with a poorer prognosis. Yet, this analysis was uncontrolled for age and sex. Whilst small differences between groups were apparent due to age, the proportion of females in the persistent-oJIA group was higher than in the extended group. It may be possible for sex differences to be driving the association. In addition, much like for the overall study, age subgroups may yet be a source of variation that needs to be accounted for in future studies. In addition, the sample groups in this analysis were small. Robust studies will require larger sample numbers.

7.3 Directions for future research

This was an exploratory study of DNAm in JIA, based on clinically homogeneous group. However, this study identified that that there are sources of sample group variation that may be reflected in DNAm data which potentially hamper identification of robust associations. Therefore, robust and replicable EWAS associations in future studies of JIA will require analysis of a more homogeneous sample group. However, this sample group needs to be clinically relevant, such that information from the research is clinically applicable. One example may be to focus on younger aged cases (≤6 years) with persistent oJIA, with follow up data on disease extension. Such patients would be easily identifiable in the clinic upon initial presentation, making associations found in that analysis more relevant for translation.

Studies with more specific hypotheses may be required in future JIA DNAm studies. There are two reasons for this: 1) The multiple testing burden high for genome scale studies and 2) The difficulty in accessing large enough sample numbers to carry out a sufficiently powered study. Therefore, focussing on a specific set of genes or regions based on specific hypotheses may help reduce this burden. Examples include a focus on immune related genes, or other hypotheses including premature ageing or cell death/apoptosis [515, 516]. This may increase the likelihood of identifying mechanisms in the aetiology of JIA.

EWAS and genetic interaction studies may help explain known JIA susceptibility loci. It has been hypothesised that mechanisms for susceptibility loci may be identified through mQTL analysis [237]. Indeed, a number of mQTL studies have already suggested an enrichment of susceptibility loci within mQTLs [222, 226, 242-244]. With large JIA GWAS data available, and with ability to draw on lessons from this JIA EWAS, new studies utilising GWAS and DNAm data may elucidate mechanisms behind JIA risk loci.

Prediction of disease outcomes is of high priority for clinicians, and has been the focus of recent collaborations with significant funding support, including the JIA focussed CLUSTER and CHART consortia (Childhood arthritis and associated uveitis: Stratification through Endotypes and mechanism and the CHildhood Arthritis Response to Treatment consortium, respectively). Whilst these focus on treatment response, identification of patients at diagnosis who will likely achieve poorer outcomes and require earlier treatment may be even more beneficial. As the old adage suggests, 'prevention is better than cure'. Preventing poor clinical outcomes before onset may help achieve better patient outcomes.

To this end, DNAm studies may hold promise. Larger studies utilising age and sex matched cases will be needed to elucidate this potential. Indeed, limiting analyses according to age subgroups may help ensure the analyses are sensitive enough to detect statistically significant changes whilst remaining clinically relevant. These analyses need not be limited to oJIA, as polyJIA RF- cases are also recognised as potentially similar to oJIA [17, 130]. Indeed, JIA cases under 6 years of age from a number of different subtypes may have similar biological processes underlying disease progression [7, 17]. Other clinical outcomes of relevance could include a future need for biological DMARDs, response to methotrexate or other treatments and progression to uveitis. As Spreafico *et al.* utilised, samples from clinical trials are likely to be valuable resources to leverage for these studies [51].

This study has identified DNAm as having the capability to identify oJIA cases from controls. However, biological heterogeneity was detected in this sample, that is, within one single JIA subtype, despite the clinical homogeneity at initial presentation and diagnosis. This heterogeneity is potentially reflected in DNAm variation, with suggestive evidence that this involves age subgroups, as well as known subgroups (persistent vs extended disease) and genetics (in this case, focussing on mQTLs). Whilst some JIA subtypes already have a biological basis for subtyping (which reflect or pre-empt clinical manifestations), a dire need exists for a biologically based classification system which reflects clinical outcomes or triage

according to treatments required. DNAm studies utilising at-diagnosis samples with follow up clinical data would allow analyses aimed at identifying such biological markers. Combining this with gene expression and genetics analysis would allow for a more comprehensive analysis.

References

1. Ravelli, A. and A. Martini, *Juvenile idiopathic arthritis.* Lancet, 2007. **369**(9563): p. 767-78.
2. Prakken, B., S. Albani, and A. Martini, *Juvenile idiopathic arthritis.* Lancet, 2011. **377**(9783): p. 2138-49.
3. Leden, I., *Doubts about Sandro Botticelli's depiction of juvenile rheumatoid arthritis.* Arthritis Rheum, 1984. **27**(10): p. 1197-8.
4. Espinel, C.H., *Caravaggio's "Il Amore Dormiente": a sleeping cupid with juvenile rheumatoid arthritis.* Lancet, 1994. **344**(8939-8940): p. 1750-2.
5. Frenk, S. and M.A. Faure-Fontenla, *Rachitis, not arthritis, in Caravaggio's sleeping child.* Lancet, 1995. **345**(8952): p. 801.
6. Alarcon-Segovia, D., *Botticelli and arthritis.* Arthritis Rheum, 1985. **28**(5): p. 600.
7. Martini, A., et al., *Toward New Classification Criteria for Juvenile Idiopathic Arthritis: First Steps, Pediatric Rheumatology International Trials Organization International Consensus.* J Rheumatol, 2019. **46**(2): p. 190-197.
8. Petty, R.E., et al., *International League of Associations for Rheumatology classification of juvenile idiopathic arthritis: second revision, Edmonton, 2001.* J Rheumatol, 2004. **31**(2): p. 390-2.
9. Swart, J.F., S. de Roock, and B.J. Prakken, *Understanding inflammation in juvenile idiopathic arthritis: How immune biomarkers guide clinical strategies in the systemic onset subtype.* European Journal of Immunology, 2016. **46**(9): p. 2068-2077.
10. Masters, S.L., et al., *Horror autoinflammaticus: the molecular pathophysiology of autoinflammatory disease (*).* Annu Rev Immunol, 2009. **27**: p. 621-68.
11. Petty, R.E. and J.T. Cassidy, *Chapter 13 - CHRONIC ARTHRITIS IN CHILDHOOD*, in *Textbook of Pediatric Rheumatology (Sixth Edition)*, J.T.C.M.L.E.P.B. Lindsley, Editor. 2011, W.B. Saunders: Philadelphia. p. 211-235.
12. Barnes, M.G., et al., *Biologic similarities based on age at onset in oligoarticular and polyarticular subtypes of juvenile idiopathic arthritis.* Arthritis Rheum, 2010. **62**(11): p. 3249-58.
13. Eng, S.W., et al., *The biologic basis of clinical heterogeneity in juvenile idiopathic arthritis.* Arthritis Rheumatol, 2014. **66**(12): p. 3463-75.
14. Hollenbach, J.A., et al., *Juvenile idiopathic arthritis and HLA class I and class II interactions and age-at-onset effects.* Arthritis Rheum, 2010. **62**(6): p. 1781-91.
15. Ma, X., et al., *Antinuclear antibody-positive cohort constitutes homogeneous entity in juvenile idiopathic arthritis.* Mod Rheumatol, 2015: p. 1-5.
16. Martini, A., *It is time to rethink juvenile idiopathic arthritis classification and nomenclature.* Ann Rheum Dis, 2012. **71**(9): p. 1437-9.
17. Ravelli, A., et al., *Antinuclear antibody-positive patients should be grouped as a separate category in the classification of juvenile idiopathic arthritis.* Arthritis Rheum, 2011. **63**(1): p. 267-75.
18. Vilca, I., et al., *Predictors of poor response to methotrexate in polyarticular-course juvenile idiopathic arthritis: analysis of the PRINTO methotrexate trial.* Ann Rheum Dis, 2010. **69**(8): p. 1479-83.
19. Beukelman, T., *Treatment advances in systemic juvenile idiopathic arthritis.* F1000Prime Rep, 2014. **6**: p. 21.
20. Webb, K. and L.R. Wedderburn, *Advances in the treatment of polyarticular juvenile idiopathic arthritis.* Curr Opin Rheumatol, 2015. **27**(5): p. 505-10.

21. De Benedetti, F. and R. Schneider, *Chapter 14 - SYSTEMIC JUVENILE IDIOPATHIC ARTHRITIS*, in *Textbook of Pediatric Rheumatology (Sixth Edition)*, J.T.C.M.L.E.P.B. Lindsley, Editor. 2011, W.B. Saunders: Philadelphia. p. 236-248.
22. Ravelli, A., et al., *2016 Classification Criteria for Macrophage Activation Syndrome Complicating Systemic Juvenile Idiopathic Arthritis: A European League Against Rheumatism/American College of Rheumatology/Paediatric Rheumatology International Trials Organisation Collaborative Initiative.* Arthritis Rheumatol, 2016. **68**(3): p. 566-76.
23. Ombrello, M.J., et al., *HLA-DRB1*11 and variants of the MHC class II locus are strong risk factors for systemic juvenile idiopathic arthritis.* Proc Natl Acad Sci U S A, 2015. **112**(52): p. 15970-5.
24. Pascual, V., et al., *Role of interleukin-1 (IL-1) in the pathogenesis of systemic onset juvenile idiopathic arthritis and clinical response to IL-1 blockade.* J Exp Med, 2005. **201**(9): p. 1479-86.
25. De Benedetti, F., et al., *Randomized trial of tocilizumab in systemic juvenile idiopathic arthritis.* N Engl J Med, 2012. **367**(25): p. 2385-95.
26. Gattorno, M., et al., *The pattern of response to anti-interleukin-1 treatment distinguishes two subsets of patients with systemic-onset juvenile idiopathic arthritis.* Arthritis Rheum, 2008. **58**(5): p. 1505-15.
27. Gerss, J., et al., *Phagocyte-specific S100 proteins and high-sensitivity C reactive protein as biomarkers for a risk-adapted treatment to maintain remission in juvenile idiopathic arthritis: a comparative study.* Ann Rheum Dis, 2012. **71**(12): p. 1991-7.
28. Holzinger, D., et al., *The Toll-like receptor 4 agonist MRP8/14 protein complex is a sensitive indicator for disease activity and predicts relapses in systemic-onset juvenile idiopathic arthritis.* Ann Rheum Dis, 2012. **71**(6): p. 974-80.
29. Nigrovic, P.A., R.P. Sundel, and R.E. Petty, *Chapter 18 - JUVENILE PSORIATIC ARTHRITIS*, in *Textbook of Pediatric Rheumatology (Sixth Edition)*, J.T.C.M.L.E.P.B. Lindsley, Editor. 2011, W.B. Saunders: Philadelphia. p. 287-297.
30. Ercan, A., et al., *Multiple juvenile idiopathic arthritis subtypes demonstrate proinflammatory IgG glycosylation.* Arthritis Rheum, 2012. **64**(9): p. 3025-33.
31. Stoll, M.L., et al., *Elevated IgG autoantibody production in oligoarticular juvenile idiopathic arthritis may predict a refractory course.* Clin Exp Rheumatol, 2011. **29**(4): p. 736-42.
32. Stoll, M.L., et al., *Patients with juvenile psoriatic arthritis comprise two distinct populations.* Arthritis Rheum, 2006. **54**(11): p. 3564-72.
33. Rosenberg, A.M. and K.G. Oen, *Chapter 15 - POLYARTHRITIS*, in *Textbook of Pediatric Rheumatology (Sixth Edition)*, J.T.C.M.L.E.P.B. Lindsley, Editor. 2011, W.B. Saunders: Philadelphia. p. 249-261.
34. Chipeta, J., et al., *Clinical patterns of juvenile idiopathic arthritis in Zambia.* Pediatr Rheumatol Online J, 2013. **11**(1): p. 33.
35. Oberle, E.J., J.G. Harris, and J.W. Verbsky, *Polyarticular juvenile idiopathic arthritis - epidemiology and management approaches.* Clin Epidemiol, 2014. **6**: p. 379-93.
36. Sailer, M., et al., *Rheumatoid factor positive, oligoarticular onset juvenile rheumatoid arthritis.* J Rheumatol, 1997. **24**(3): p. 586-8.
37. Klotsche, J., et al., *Long-term safety of etanercept and adalimumab compared to methotrexate in patients with juvenile idiopathic arthritis (JIA).* Ann Rheum Dis, 2016. **75**(5): p. 855-61.
38. Ungar, W.J., et al., *The use of biologic response modifiers in polyarticular-course juvenile idiopathic arthritis: a systematic review.* Semin Arthritis Rheum, 2013. **42**(6): p. 597-618.

39. Vidqvist, K.L., et al., *Disease activity of idiopathic juvenile arthritis continues through adolescence despite the use of biologic therapies.* Rheumatology (Oxford), 2013. **52**(11): p. 1999-2003.
40. Guzman, J., et al., *The outcomes of juvenile idiopathic arthritis in children managed with contemporary treatments: results from the ReACCh-Out cohort.* Ann Rheum Dis, 2015. **74**(10): p. 1854-60.
41. Petty, R.E. and J.T. Cassidy, *Chapter 17 - ENTHESITIS-RELATED ARTHRITIS (JUVENILE ANKYLOSING SPONDYLITIS)*, in *Textbook of Pediatric Rheumatology (Sixth Edition)*, J.T.C.M.L.E.P.B. Lindsley, Editor. 2011, W.B. Saunders: Philadelphia. p. 272-286.
42. Barnes, M.G., et al., *Subtype-specific peripheral blood gene expression profiles in recent-onset juvenile idiopathic arthritis.* Arthritis Rheum, 2009. **60**(7): p. 2102-12.
43. Consolaro, A., et al., *Advances in biomarkers for paediatric rheumatic diseases.* Nat Rev Rheumatol, 2014.
44. de Kleer, I.M., et al., *CD4+CD25bright regulatory T cells actively regulate inflammation in the joints of patients with the remitting form of juvenile idiopathic arthritis.* J Immunol, 2004. **172**(10): p. 6435-43.
45. Hunter, P.J., et al., *Biologic predictors of extension of oligoarticular juvenile idiopathic arthritis as determined from synovial fluid cellular composition and gene expression.* Arthritis Rheum, 2010. **62**(3): p. 896-907.
46. Barnes, M.G., et al., *Gene expression in juvenile arthritis and spondyloarthropathy: pro-angiogenic ELR+ chemokine genes relate to course of arthritis.* Rheumatology (Oxford), 2004. **43**(8): p. 973-9.
47. Campanilho-Marques, R., et al., *Prognostic value of antinuclear antibodies in juvenile idiopathic arthritis and anterior uveitis. Results from a systematic literature review.* Acta Reumatol Port, 2014. **39**(2): p. 116-22.
48. Nistala, K., et al., *Interleukin-17-producing T cells are enriched in the joints of children with arthritis, but have a reciprocal relationship to regulatory T cell numbers.* Arthritis Rheum, 2008. **58**(3): p. 875-87.
49. Nordal, E., et al., *Incidence and predictors of Uveitis in juvenile idiopathic arthritis in a Nordic long-term cohort study.* Pediatr Rheumatol Online J, 2017. **15**(1): p. 66.
50. Ravelli, A., et al., *Patients with antinuclear antibody-positive juvenile idiopathic arthritis constitute a homogeneous subgroup irrespective of the course of joint disease.* Arthritis Rheum, 2005. **52**(3): p. 826-32.
51. Spreafico, R., et al., *Epipolymorphisms associated with the clinical outcome of autoimmune arthritis affect CD4+ T cell activation pathways.* Proc Natl Acad Sci U S A, 2016. **113**(48): p. 13845-13850.
52. Saurenmann, R.K., et al., *Epidemiology of juvenile idiopathic arthritis in a multiethnic cohort: ethnicity as a risk factor.* Arthritis Rheum, 2007. **56**(6): p. 1974-84.
53. Petty, R.E. and J.T. Cassidy, *Chapter 16 - OLIGOARTHRITIS*, in *Textbook of Pediatric Rheumatology (Sixth Edition)*, J.T.C.M.L.E.P.B. Lindsley, Editor. 2011, W.B. Saunders: Philadelphia. p. 262-271.
54. Hinks, A., et al., *Fine-mapping the MHC locus in juvenile idiopathic arthritis (JIA) reveals genetic heterogeneity corresponding to distinct adult inflammatory arthritic diseases.* Ann Rheum Dis, 2017. **76**(4): p. 765-772.
55. Nordal, E., N.T. Songstad, and M. Rygg, *Difficulties in defining antinuclear antibody-positive patients as a separate category in the classification of juvenile idiopathic arthritis: comment on the article by Ravelli et al.* Arthritis Rheum, 2011. **63**(9): p. 2835; author reply 2835-6.
56. Sparchez, M., et al., *Antinuclear antibody screening by ELISA and IF techniques: discrepant results in juvenile idiopathic arthritis but consistency in childhood systemic lupus erythematous.* Clin Rheumatol, 2014. **33**(5): p. 643-7.

57. Rheumatology, A.C.o., *Position Statement: Methodology of Testing for Antinuclear Antibodies*. 2015.
58. Pisetsky, D.S., *Antinuclear antibody testing - misunderstood or misbegotten?* Nat Rev Rheumatol, 2017. **13**(8): p. 495-502.
59. Maguire, G.A., et al., *Clinical utility of ANA measured by ELISA compared with ANA measured by immunofluorescence.* Rheumatology (Oxford), 2009. **48**(8): p. 1013-4.
60. Nordal, E.B., et al., *Biomarkers of chronic uveitis in juvenile idiopathic arthritis: predictive value of antihistone antibodies and antinuclear antibodies.* J Rheumatol, 2009. **36**(8): p. 1737-43.
61. Kumar, Y., A. Bhatia, and R.W. Minz, *Antinuclear antibodies and their detection methods in diagnosis of connective tissue diseases: a journey revisited.* Diagn Pathol, 2009. **4**: p. 1.
62. Emlen, W. and L. O'Neill, *Clinical significance of antinuclear antibodies: comparison of detection with immunofluorescence and enzyme-linked immunosorbent assays.* Arthritis Rheum, 1997. **40**(9): p. 1612-8.
63. Homburger, H.A., et al., *Detection of antinuclear antibodies: comparative evaluation of enzyme immunoassay and indirect immunofluorescence methods.* Arch Pathol Lab Med, 1998. **122**(11): p. 993-9.
64. Jaskowski, T.D., et al., *Screening for antinuclear antibodies by enzyme immunoassay.* Am J Clin Pathol, 1996. **105**(4): p. 468-73.
65. Tonuttia, E., et al., *Diagnostic accuracy of ELISA methods as an alternative screening test to indirect immunofluorescence for the detection of antinuclear antibodies. Evaluation of five commercial kits.* Autoimmunity, 2004. **37**(2): p. 171-6.
66. Al-Matar, M.J., et al., *The early pattern of joint involvement predicts disease progression in children with oligoarticular (pauciarticular) juvenile rheumatoid arthritis.* Arthritis Rheum, 2002. **46**(10): p. 2708-15.
67. Guillaume, S., et al., *Long-term outcome and prognosis in oligoarticular-onset juvenile idiopathic arthritis.* Arthritis Rheum, 2000. **43**(8): p. 1858-65.
68. Firestein, G.S. and W.N. Kelley, *Kelley's textbook of rheumatology*. 2013, Philadelphia, PA: Elsevier/Saunders.
69. Foster, H.E., et al., *Outcome in adults with juvenile idiopathic arthritis: a quality of life study.* Arthritis Rheum, 2003. **48**(3): p. 767-75.
70. Wallace, C.A., et al., *Patterns of clinical remission in select categories of juvenile idiopathic arthritis.* Arthritis Rheum, 2005. **52**(11): p. 3554-62.
71. Navegantes, K.C., et al., *Immune modulation of some autoimmune diseases: the critical role of macrophages and neutrophils in the innate and adaptive immunity.* J Transl Med, 2017. **15**.
72. Alberts, B., A. Johnson, and J.e.a. Lewis, *Chapter 24, The Adaptive Immune System; Molecular Biology of the Cell*. 2002, Garland Science: New York.
73. Scherer, H.U. and G.R. Burmester, *Adaptive immunity in rheumatic diseases: bystander or pathogenic player?* Best Pract Res Clin Rheumatol, 2011. **25**(6): p. 785-800.
74. Komori, T., et al., *Lack of N regions in antigen receptor variable region genes of TdT-deficient lymphocytes.* Science, 1993. **261**(5125): p. 1171-5.
75. Pieper, K., B. Grimbacher, and H. Eibel, *B-cell biology and development.* Journal of Allergy and Clinical Immunology, 2013. **131**(4): p. 959-971.
76. Naradikian, M.S., et al., *Understanding B Cell Biology*, in *Drugs Targeting B-cells in Autoimmune diseases*, X. Bosch, M. Ramos-Casals, and M.A. Kamashta, Editors. 2014, Springer Basel. p. 11-35.

77. Cher, D.J. and T.R. Mosmann, *Two types of murine helper T cell clone. II. Delayed-type hypersensitivity is mediated by TH1 clones.* J Immunol, 1987. **138**(11): p. 3688-94.
78. Murphy, K.M. and S.L. Reiner, *The lineage decisions of helper T cells.* Nat Rev Immunol, 2002. **2**(12): p. 933-44.
79. Nurieva, R.I., et al., *Generation of T follicular helper cells is mediated by interleukin-21 but independent of T helper 1, 2, or 17 cell lineages.* Immunity, 2008. **29**(1): p. 138-49.
80. Swaidani, S., et al., *T cell-derived Act1 is necessary for IL-25-mediated Th2 responses and allergic airway inflammation.* J Immunol, 2011. **187**(6): p. 3155-64.
81. Veldhoen, M., et al., *Transforming growth factor-beta 'reprograms' the differentiation of T helper 2 cells and promotes an interleukin 9-producing subset.* Nat Immunol, 2008. **9**(12): p. 1341-6.
82. Weaver, C.T., et al., *Th17: an effector CD4 T cell lineage with regulatory T cell ties.* Immunity, 2006. **24**(6): p. 677-88.
83. Alberts, B., A. Johnson, and J.e.a. Lewis, *Molecular Biology of the Cell.*, in *Molecular Biology of the Cell*, J.A. Alberts B, Lewis J, et al., Editor. 2002, Garland Science: New York.
84. Josefowicz, S.Z. and A. Rudensky, *Control of regulatory T cell lineage commitment and maintenance.* Immunity, 2009. **30**(5): p. 616-25.
85. Mosmann, T.R., et al., *Two types of murine helper T cell clone. I. Definition according to profiles of lymphokine activities and secreted proteins.* J Immunol, 1986. **136**(7): p. 2348-57.
86. Raphael, I., et al., *T cell subsets and their signature cytokines in autoimmune and inflammatory diseases.* Cytokine, 2015. **74**(1): p. 5-17.
87. Nistala, K. and L.R. Wedderburn, *Th17 and regulatory T cells: rebalancing pro- and anti-inflammatory forces in autoimmune arthritis.* Rheumatology (Oxford), 2009. **48**(6): p. 602-6.
88. Licona-Limon, P., et al., *TH2, allergy and group 2 innate lymphoid cells.* Nat Immunol, 2013. **14**(6): p. 536-42.
89. Caza, T. and S. Landas, *Functional and Phenotypic Plasticity of CD4(+) T Cell Subsets.* Biomed Res Int, 2015. **2015**: p. 521957.
90. Skurkovich, S. and B. Skurkovich, *Anticytokine therapy, especially anti-interferon-gamma, as a pathogenetic treatment in TH-1 autoimmune diseases.* Ann N Y Acad Sci, 2005. **1051**: p. 684-700.
91. Chen, Z., A. Laurence, and J.J. O'Shea, *Signal transduction pathways and transcriptional regulation in the control of Th17 differentiation.* Semin Immunol, 2007. **19**(6): p. 400-8.
92. Lee, Y.K., et al., *Developmental plasticity of Th17 and Treg cells.* Curr Opin Immunol, 2009. **21**(3): p. 274-80.
93. Hori, S., T. Nomura, and S. Sakaguchi, *Control of regulatory T cell development by the transcription factor Foxp3.* Science, 2003. **299**(5609): p. 1057-61.
94. Fontenot, J.D., M.A. Gavin, and A.Y. Rudensky, *Foxp3 programs the development and function of CD4+CD25+ regulatory T cells.* Nat Immunol, 2003. **4**(4): p. 330-6.
95. Vignali, D.A., L.W. Collison, and C.J. Workman, *How regulatory T cells work.* Nat Rev Immunol, 2008. **8**(7): p. 523-32.
96. Miyara, M., et al., *Human FoxP3+ regulatory T cells in systemic autoimmune diseases.* Autoimmun Rev, 2011. **10**(12): p. 744-55.
97. Lundy, S.K., et al., *Cells of the synovium in rheumatoid arthritis. T lymphocytes.* Arthritis Res Ther, 2007. **9**(1): p. 202.

98. Klareskog, L., K. Amara, and V. Malmstrom, *Adaptive immunity in rheumatoid arthritis: anticitrulline and other antibodies in the pathogenesis of rheumatoid arthritis.* Curr Opin Rheumatol, 2014. **26**(1): p. 72-9.
99. Roberts, C.A., A.K. Dickinson, and L.S. Taams, *The Interplay Between Monocytes/Macrophages and CD4(+) T Cell Subsets in Rheumatoid Arthritis.* Front Immunol, 2015. **6**: p. 571.
100. Murray, K.J., et al., *Immunohistological characteristics of T cell infiltrates in different forms of childhood onset chronic arthritis.* J Rheumatol, 1996. **23**(12): p. 2116-24.
101. Gattorno, M., et al., *Phenotypic and functional characterisation of CCR7+ and CCR7- CD4+ memory T cells homing to the joints in juvenile idiopathic arthritis.* Arthritis Res Ther, 2005. **7**(2): p. R256-67.
102. Wedderburn, L.R., et al., *Selective recruitment of polarized T cells expressing CCR5 and CXCR3 to the inflamed joints of children with juvenile idiopathic arthritis.* Arthritis Rheum, 2000. **43**(4): p. 765-74.
103. Wedderburn, L.R., et al., *Divergence in the degree of clonal expansions in inflammatory T cell subpopulations mirrors HLA-associated risk alleles in genetically and clinically distinct subtypes of childhood arthritis.* Int Immunol, 2001. **13**(12): p. 1541-50.
104. Wehrens, E.J., et al., *Treating arthritis by immunomodulation: is there a role for regulatory T cells?* Rheumatology (Oxford), 2010. **49**(9): p. 1632-44.
105. Wu, S.A., et al., *Persistent improper upregulation of Th17 and T cells in patients with juvenile idiopathic arthritis.* J Microbiol Immunol Infect, 2014.
106. Bending, D., et al., *Hypomethylation at the Regulatory T Cell–Specific Demethylated Region in CD25(hi) T Cells Is Decoupled from FOXP3 Expression at the Inflamed Site in Childhood Arthritis.* J Immunol, 2014. **193**(6): p. 2699-708.
107. Cosmi, L., et al., *Evidence of the transient nature of the Th17 phenotype of CD4+CD161+ T cells in the synovial fluid of patients with juvenile idiopathic arthritis.* Arthritis Rheum, 2011. **63**(8): p. 2504-15.
108. Grose, R.H., et al., *Comparison of blood and synovial fluid th17 and novel peptidase inhibitor 16 Treg cell subsets in juvenile idiopathic arthritis.* J Rheumatol, 2012. **39**(10): p. 2021-31.
109. Olivito, B., et al., *Th17 transcription factor RORC2 is inversely correlated with FOXP3 expression in the joints of children with juvenile idiopathic arthritis.* J Rheumatol, 2009. **36**(9): p. 2017-24.
110. Szymanska-Kaluza, J., et al., *Imbalance of Th17 and T-regulatory cells in peripheral blood and synovial fluid in treatment naive children with juvenile idiopathic arthritis.* Cent Eur J Immunol, 2014. **39**(1): p. 71-6.
111. Venter, J.C., et al., *The sequence of the human genome.* Science, 2001. **291**(5507): p. 1304-51.
112. Consortium, T.E.P., *An integrated encyclopedia of DNA elements in the human genome.* Nature, 2012. **489**(7414): p. 57-74.
113. Sherry, S.T., et al., *dbSNP: the NCBI database of genetic variation.* Nucleic Acids Res, 2001. **29**(1): p. 308-11.
114. Risch, N. and K. Merikangas, *The future of genetic studies of complex human diseases.* Science, 1996. **273**(5281): p. 1516-7.
115. Clarke, A. and T.J. Vyse, *Genetics of rheumatic disease.* Arthritis Res Ther, 2009. **11**(5): p. 248.
116. Baum, J. and C. Fink, *Juvenile rheumatoid arthritis in monozygotic twins: a case report and review of the literature.* Arthritis Rheum, 1968. **11**(1): p. 33-6.
117. Clemens, L.E., E. Albert, and B.M. Ansell, *Sibling pairs affected by chronic arthritis of childhood: evidence for a genetic predisposition.* J Rheumatol, 1985. **12**(1): p. 108-13.

118. Glass, D.N. and E.H. Giannini, *Juvenile rheumatoid arthritis as a complex genetic trait.* Arthritis Rheum, 1999. **42**(11): p. 2261-8.
119. Rosenberg, A.M. and R.E. Petty, *Similar patterns of juvenile rheumatoid arthritis within families.* Arthritis Rheum, 1980. **23**(8): p. 951-3.
120. Rossen, R.D., et al., *Familial rheumatoid arthritis: linkage of HLA to disease susceptibility locus in four families where proband presented with juvenile rheumatoid arthritis.* J Clin Invest, 1980. **65**(3): p. 629-42.
121. Prahalad, S., et al., *Increased prevalence of familial autoimmunity in simplex and multiplex families with juvenile rheumatoid arthritis.* Arthritis Rheum, 2002. **46**(7): p. 1851-6.
122. Prahalad, S., et al., *Familial aggregation of juvenile idiopathic arthritis.* Arthritis Rheum, 2004. **50**(12): p. 4022-7.
123. Ansell, B.M., E.G. Bywaters, and J.S. Lawrence, *Familial aggregation and twin studies in Still's disease. Juvenile chronic polyarthritis.* Rheumatology, 1969. **2**: p. 37-61.
124. Savolainen, A., et al., *Magnitude of the genetic component in juvenile idiopathic arthritis.* Ann Rheum Dis, 2000. **59**(12): p. 1001.
125. Moroldo, M.B., et al., *Juvenile rheumatoid arthritis in affected sibpairs.* Arthritis Rheum, 1997. **40**(11): p. 1962-6.
126. Prahalad, S., et al., *Juvenile rheumatoid arthritis: linkage to HLA demonstrated by allele sharing in affected sibpairs.* Arthritis Rheum, 2000. **43**(10): p. 2335-8.
127. Fernandez-Vina, M.A., C.W. Fink, and P. Stastny, *HLA antigens in juvenile arthritis. Pauciarticular and polyarticular juvenile arthritis are immunogenetically distinct.* Arthritis Rheum, 1990. **33**(12): p. 1787-94.
128. Nepom, B.S., et al., *Specific HLA-DR4-associated histocompatibility molecules characterize patients with seropositive juvenile rheumatoid arthritis.* J Clin Invest, 1984. **74**(1): p. 287-91.
129. Rachelefsky, G.S., et al., *Increased prevalence of W27 in juvenile rheumatoid arthritis.* N Engl J Med, 1974. **290**(16): p. 892-3.
130. Hinks, A., et al., *Dense genotyping of immune-related disease regions identifies 14 new susceptibility loci for juvenile idiopathic arthritis.* Nat Genet, 2013. **45**(6): p. 664-9.
131. Li, Y.R., et al., *Genetic sharing and heritability of paediatric age of onset autoimmune diseases.* Nat Commun, 2015. **6**: p. 8442.
132. De Silvestri, A., et al., *HLA-DRB1 alleles and juvenile idiopathic arthritis: Diagnostic clues emerging from a meta-analysis.* Autoimmun Rev, 2017. **16**(12): p. 1230-1236.
133. Thomson, W., et al., *Juvenile idiopathic arthritis classified by the ILAR criteria: HLA associations in UK patients.* Rheumatology (Oxford), 2002. **41**(10): p. 1183-9.
134. Murray, K.J., et al., *Age-specific effects of juvenile rheumatoid arthritis-associated HLA alleles.* Arthritis Rheum, 1999. **42**(9): p. 1843-53.
135. Huppertz, H.I., *Viewpoint: time for change in pediatric rheumatology.* Rheumatol Int, 1993. **13**(1): p. 37-8.
136. Southwood, T.R. and P. Woo, *Juvenile chronic arthritis.* Baillieres Clin Rheumatol, 1995. **9**(2): p. 331-53.
137. Angeles-Han, S.T., et al., *HLA Associations in a Cohort of Children With Juvenile Idiopathic Arthritis With and Without Uveitis.* Invest Ophthalmol Vis Sci, 2015. **56**(10): p. 6043-8.
138. Davidsone, Z., et al., *HLA II class alleles in juvenile idiopathic arthritis patients with and without temporomandibular joint arthritis.* Pediatr Rheumatol Online J, 2016. **14**(1): p. 24.
139. Yanagimachi, M., et al., *Association of HLA-A*02:06 and HLA-DRB1*04:05 with clinical subtypes of juvenile idiopathic arthritis.* J Hum Genet, 2011. **56**(3): p. 196-9.

140. Finkel, T.H., et al., *Variants in CXCR4 associate with juvenile idiopathic arthritis susceptibility.* BMC Med Genet, 2016. **17**: p. 24.
141. Hindorff LA, M.J.E.B.I., Morales J (European Bioinformatics Institute), Junkins HA, Hall PN, Klemm AK, and Manolio TA, *A Catalog of Published Genome-Wide Association Studies.*
142. Hinks, A., et al., *Identification of a novel susceptibility locus for juvenile idiopathic arthritis by genome-wide association analysis.* Arthritis Rheum, 2009. **60**(1): p. 258-63.
143. Thompson, S.D., et al., *Genome-wide association analysis of juvenile idiopathic arthritis identifies a new susceptibility locus at chromosomal region 3q13.* Arthritis Rheum, 2012. **64**(8): p. 2781-91.
144. Behrens, E.M., et al., *Association of the TRAF1-C5 locus on chromosome 9 with juvenile idiopathic arthritis.* Arthritis Rheum, 2008. **58**(7): p. 2206-7.
145. McIntosh, L.A., et al., *Genome-Wide Association Meta-Analysis Reveals Novel Juvenile Idiopathic Arthritis Susceptibility Loci.* Arthritis Rheumatol, 2017. **69**(11): p. 2222-2232.
146. Hinks, A., et al., *Brief Report: The Genetic Profile of Rheumatoid Factor-Positive Polyarticular Juvenile Idiopathic Arthritis Resembles That of Adult Rheumatoid Arthritis.* Arthritis Rheumatol, 2018. **70**(6): p. 957-962.
147. Cortes, A. and M.A. Brown, *Promise and pitfalls of the Immunochip.* Arthritis Res Ther, 2011. **13**(1): p. 101.
148. Ellis, J.A., et al., *Independent replication analysis of genetic loci with previous evidence of association with juvenile idiopathic arthritis.* Pediatr Rheumatol Online J, 2013. **11**(1): p. 12.
149. Farh, K.K., et al., *Genetic and epigenetic fine mapping of causal autoimmune disease variants.* Nature, 2015. **518**(7539): p. 337-43.
150. Aydin-Son, Y., et al., *Systems-level analysis of genome wide association study results for a pilot juvenile idiopathic arthritis family study.* Turk J Pediatr, 2015. **57**(4): p. 324-33.
151. Haasnoot, A.J.W., et al., *Identification of an Amino Acid Motif in HLA-DRbeta1 That Distinguishes Uveitis in Patients With Juvenile Idiopathic Arthritis.* Arthritis Rheumatol, 2018. **70**(7): p. 1155-1165.
152. Buniello, A., et al., *The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019.* Nucleic Acids Res, 2019. **47**(D1): p. D1005-D1012.
153. Prahalad, S., *Genetic analysis of juvenile rheumatoid arthritis: approaches to complex traits.* Curr Probl Pediatr Adolesc Health Care, 2006. **36**(3): p. 83-90.
154. Ellis, J.A., J.E. Munro, and A.L. Ponsonby, *Possible environmental determinants of juvenile idiopathic arthritis.* Rheumatology (Oxford), 2010. **49**(3): p. 411-25.
155. Conradt, E., et al., *The roles of DNA methylation of NR3C1 and 11beta-HSD2 and exposure to maternal mood disorder in utero on newborn neurobehavior.* Epigenetics, 2013. **8**(12): p. 1321-9.
156. Massart, R., et al., *Hydroxymethylation and DNA methylation profiles in the prefrontal cortex of the non-human primate rhesus macaque and the impact of maternal deprivation on hydroxymethylation.* Neuroscience, 2014. **268**: p. 139-48.
157. Zhang, T.Y., et al., *Maternal care and DNA methylation of a glutamic acid decarboxylase 1 promoter in rat hippocampus.* J Neurosci, 2010. **30**(39): p. 13130-7.
158. Ellis, J.A., et al., *CLARITY - ChiLdhood Arthritis Risk factor Identification sTudY.* Pediatr Rheumatol Online J, 2012. **10**(1): p. 37.
159. Mason, T., et al., *Breast feeding and the development of juvenile rheumatoid arthritis.* J Rheumatol, 1995. **22**(6): p. 1166-70.

160. Rosenberg, A.M., *Evaluation of associations between breast feeding and subsequent development of juvenile rheumatoid arthritis.* J Rheumatol, 1996. **23**(6): p. 1080-2.
161. Miller, J., et al., *Sibling Exposure and Risk of Juvenile Idiopathic Arthritis.* Arthritis Rheumatol, 2015. **67**(7): p. 1951-8.
162. Nielsen, H.E., et al., *Epidemiology of juvenile chronic arthritis: risk dependent on sibship, parental income, and housing.* J Rheumatol, 1999. **26**(7): p. 1600-5.
163. Carlens, C., et al., *Perinatal characteristics, early life infections and later risk of rheumatoid arthritis and juvenile idiopathic arthritis.* Ann Rheum Dis, 2009. **68**(7): p. 1159-64.
164. Strachan, D.P., *Hay fever, hygiene, and household size.* BMJ, 1989. **299**(6710): p. 1259-60.
165. Al-Nakib, W. and H.A. Majeed, *Serologic studies on the association of rubella virus infection and juvenile rheumatoid arthritis.* Ann Trop Paediatr, 1981. **1**(2): p. 93-6.
166. Chantler, J.K., A.J. Tingle, and R.E. Petty, *Persistent rubella virus infection associated with chronic arthritis in children.* N Engl J Med, 1985. **313**(18): p. 1117-23.
167. Ogra, P.L., et al., *Rubella-virus infection in juvenile rheumatoid arthritis.* Lancet, 1975. **1**(7917): p. 1157-61.
168. Gonzalez, B., et al., *Parvovirus B19 may have a role in the pathogenesis of juvenile idiopathic arthritis.* J Rheumatol, 2007. **34**(6): p. 1336-40.
169. Nocton, J.J., et al., *Human parvovirus B19-associated arthritis in children.* J Pediatr, 1993. **122**(2): p. 186-90.
170. Oguz, F., et al., *Parvovirus B19 in the acute arthropathies and juvenile rheumatoid arthritis.* J Paediatr Child Health, 2002. **38**(4): p. 358-62.
171. Soderlund, M., et al., *Persistence of parvovirus B19 DNA in synovial membranes of young patients with and without chronic arthropathy.* Lancet, 1997. **349**(9058): p. 1063-5.
172. Weissbrich, B., Y. Suss-Frohlich, and H.J. Girschick, *Seroprevalence of parvovirus B19 IgG in children affected by juvenile idiopathic arthritis.* Arthritis Res Ther, 2007. **9**(4): p. R82.
173. Pritchard, M.H., N. Matthews, and J. Munro, *Antibodies to influenza A in a cluster of children with juvenile chronic arthritis.* Br J Rheumatol, 1988. **27**(3): p. 176-80.
174. Tsai, Y.T., et al., *Detection of Epstein-Barr virus and cytomegalovirus genome in white blood cells from patients with juvenile rheumatoid arthritis and childhood systemic lupus erythematosus.* Int Arch Allergy Immunol, 1995. **106**(3): p. 235-40.
175. Pritchard, M.H. and J. Munro, *Successful treatment of juvenile chronic arthritis with a specific antiviral agent.* Br J Rheumatol, 1989. **28**(6): p. 521-4.
176. Barash, J. and O. Goldzweig, *Possible role of streptococcal infection in flares of juvenile idiopathic arthritis.* Arthritis Rheum, 2007. **57**(5): p. 877-80.
177. Arvonen, M., et al., *Repeated exposure to antibiotics in infancy: a predisposing factor for juvenile idiopathic arthritis or a sign of this group's greater susceptibility to infections?* J Rheumatol, 2015. **42**(3): p. 521-6.
178. Horton, D.B., et al., *Antibiotic Exposure and Juvenile Idiopathic Arthritis: A Case-Control Study.* Pediatrics, 2015. **136**(2): p. e333-43.
179. Di Paola, M., et al., *Alteration of Fecal Microbiota Profiles in Juvenile Idiopathic Arthritis. Associations with HLA-B27 Allele and Disease Status.* Front Microbiol, 2016. **7**: p. 1703.
180. Picco, P., et al., *Increased gut permeability in juvenile chronic arthritides. A multivariate analysis of the diagnostic parameters.* Clin Exp Rheumatol, 2000. **18**(6): p. 773-8.
181. Tejesvi, M.V., et al., *Faecal microbiome in new-onset juvenile idiopathic arthritis.* Eur J Clin Microbiol Infect Dis, 2016. **35**(3): p. 363-70.

182. Jaakkola, J.J. and M. Gissler, *Maternal smoking in pregnancy as a determinant of rheumatoid arthritis and other inflammatory polyarthropathies during the first 7 years of life.* Int J Epidemiol, 2005. **34**(3): p. 664-71.
183. Shenoi, S., et al., *Juvenile idiopathic arthritis in relation to maternal prenatal smoking.* Arthritis Care Res (Hoboken), 2015. **67**(5): p. 725-30.
184. Mikeska, T. and J.M. Craig, *DNA methylation biomarkers: cancer and beyond.* Genes (Basel), 2014. **5**(3): p. 821-64.
185. Bird, A., *Perceptions of epigenetics.* Nature, 2007. **447**(7143): p. 396-8.
186. Russo, V.E.A.V.E.A.R., Arthur D & Martienssen, Robert A *Epigenetic mechanisms of gene regulation*. 1996, Cold Spring Harbour, NY: Cold Spring Harbour Laboratory Press.
187. Esteller, M., *Epigenetics in cancer.* N Engl J Med, 2008. **358**(11): p. 1148-59.
188. Gupta, B. and R.D. Hawkins, *Epigenomics of autoimmune diseases.* Immunol Cell Biol, 2015. **93**(3): p. 271-6.
189. Jaenisch, R. and A. Bird, *Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals.* Nat Genet, 2003. **33 Suppl**: p. 245-54.
190. Feng, S., S.E. Jacobsen, and W. Reik, *Epigenetic reprogramming in plant and animal development.* Science, 2010. **330**(6004): p. 622-7.
191. Ramsahoye, B.H., et al., *Non-CpG methylation is prevalent in embryonic stem cells and may be mediated by DNA methyltransferase 3a.* Proc Natl Acad Sci U S A, 2000. **97**(10): p. 5237-42.
192. Ziller, M.J., et al., *Genomic distribution and inter-sample variation of non-CpG methylation across human cell types.* PLoS Genet, 2011. **7**(12): p. e1002389.
193. Smith, Z.D. and A. Meissner, *DNA methylation: roles in mammalian development.* Nat Rev Genet, 2013. **14**(3): p. 204-220.
194. Bergman, Y. and H. Cedar, *DNA methylation dynamics in health and disease.* Nat Struct Mol Biol, 2013. **20**(3): p. 274-281.
195. Deaton, A.M. and A. Bird, *CpG islands and the regulation of transcription.* Genes Dev, 2011. **25**(10): p. 1010-22.
196. Li, E., T.H. Bestor, and R. Jaenisch, *Targeted mutation of the DNA methyltransferase gene results in embryonic lethality.* Cell, 1992. **69**(6): p. 915-26.
197. Okano, M., et al., *DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development.* Cell, 1999. **99**(3): p. 247-57.
198. Gruenbaum, Y., H. Cedar, and A. Razin, *Substrate and sequence specificity of a eukaryotic DNA methylase.* Nature, 1982. **295**(5850): p. 620-2.
199. Leonhardt, H., et al., *A targeting sequence directs DNA methyltransferase to sites of DNA replication in mammalian nuclei.* Cell, 1992. **71**(5): p. 865-73.
200. Laurent, L., et al., *Dynamic changes in the human methylome during differentiation.* Genome Res, 2010. **20**(3): p. 320-31.
201. Straussman, R., et al., *Developmental programming of CpG island methylation profiles in the human genome.* Nat Struct Mol Biol, 2009. **16**(5): p. 564-71.
202. Cortessis, V.K., et al., *Environmental epigenetics: prospects for studying epigenetic mediation of exposure-response relationships.* Hum Genet, 2012. **131**(10): p. 1565-89.
203. Martino, D.J., et al., *Evidence for age-related and individual-specific changes in DNA methylation profile of mononuclear cells during early immune development in humans.* Epigenetics, 2011. **6**(9): p. 1085-94.
204. Heijmans, B.T., et al., *Persistent epigenetic differences associated with prenatal exposure to famine in humans.* Proc Natl Acad Sci U S A, 2008. **105**(44): p. 17046-9.
205. Mazzio, E.A. and K.F. Soliman, *Basic concepts of epigenetics: impact of environmental signals on gene expression.* Epigenetics, 2012. **7**(2): p. 119-30.

206. Gordon, L., et al., *Expression discordance of monozygotic twins at birth: effect of intrauterine environment and a possible mechanism for fetal programming.* Epigenetics, 2011. **6**(5): p. 579-92.
207. Fraga, M.F., et al., *Epigenetic differences arise during the lifetime of monozygotic twins.* Proc Natl Acad Sci U S A, 2005. **102**(30): p. 10604-9.
208. Heyn, H., et al., *Distinct DNA methylomes of newborns and centenarians.* Proc Natl Acad Sci U S A, 2012. **109**(26): p. 10522-7.
209. Talens, R.P., et al., *Variation, patterns, and temporal stability of DNA methylation: considerations for epigenetic epidemiology.* FASEB J, 2010. **24**(9): p. 3135-44.
210. Hompes, T., et al., *Investigating the influence of maternal cortisol and emotional state during pregnancy on the DNA methylation status of the glucocorticoid receptor gene (NR3C1) promoter region in cord blood.* J Psychiatr Res, 2013. **47**(7): p. 880-91.
211. Majumdar, S., et al., *Arsenic exposure induces genomic hypermethylation.* Environ Toxicol, 2010. **25**(3): p. 315-8.
212. Oberlander, T.F., et al., *Prenatal exposure to maternal depression, neonatal methylation of human glucocorticoid receptor gene (NR3C1) and infant cortisol stress responses.* Epigenetics, 2008. **3**(2): p. 97-106.
213. Pilsner, J.R., et al., *Influence of prenatal lead exposure on genomic methylation of cord blood DNA.* Environ Health Perspect, 2009. **117**(9): p. 1466-71.
214. Pilsner, J.R., et al., *Genomic methylation of peripheral blood leukocyte DNA: influences of arsenic and folate in Bangladeshi adults.* Am J Clin Nutr, 2007. **86**(4): p. 1179-86.
215. Joubert, B.R., et al., *450K epigenome-wide scan identifies differential DNA methylation in newborns related to maternal smoking during pregnancy.* Environ Health Perspect, 2012. **120**(10): p. 1425-31.
216. Markunas, C.A., et al., *Identification of DNA methylation changes in newborns related to maternal smoking during pregnancy.* Environ Health Perspect, 2014. **122**(10): p. 1147-53.
217. Novakovic, B., et al., *Postnatal stability, tissue, and time specific effects of AHRR methylation change in response to maternal smoking in pregnancy.* Epigenetics, 2014. **9**(3): p. 377-86.
218. Zeilinger, S., et al., *Tobacco smoking leads to extensive genome-wide changes in DNA methylation.* PLoS One, 2013. **8**(5): p. e63812.
219. Gibbs, J.R., et al., *Abundant quantitative trait loci exist for DNA methylation and gene expression in human brain.* PLoS Genet, 2010. **6**(5): p. e1000952.
220. Zhang, D., et al., *Genetic control of individual differences in gene-specific methylation in human brain.* Am J Hum Genet, 2010. **86**(3): p. 411-9.
221. Grundberg, E., et al., *Global analysis of DNA methylation variation in adipose tissue from twins reveals links to disease-associated variants in distal regulatory elements.* Am J Hum Genet, 2013. **93**(5): p. 876-90.
222. Liu, Y., et al., *GeMes, clusters of DNA methylation under genetic control, can inform genetic and epigenetic analysis of disease.* Am J Hum Genet, 2014. **94**(4): p. 485-95.
223. Qiu, W., et al., *The impact of genetic variation and cigarette smoke on DNA methylation in current and former smokers from the COPDGene study.* Epigenetics, 2015. **10**(11): p. 1064-73.
224. Shi, J., et al., *Characterizing the genetic basis of methylome diversity in histologically normal human lung tissue.* Nat Commun, 2014. **5**: p. 3365.
225. Smith, A.K., et al., *Methylation quantitative trait loci (meQTLs) are consistently detected across ancestry, developmental stage, and tissue type.* BMC Genomics, 2014. **15**: p. 145.

226. Zhang, X., et al., *Linking the genetic architecture of cytosine modifications with human complex traits.* Hum Mol Genet, 2014. **23**(22): p. 5893-905.
227. Chen, L., et al., *Genetic Drivers of Epigenetic and Transcriptional Variation in Human Immune Cells.* Cell, 2016. **167**(5): p. 1398-1414 e24.
228. Day, K., et al., *Heritable DNA Methylation in CD4+ Cells among Complex Families Displays Genetic and Non-Genetic Effects.* PLoS One, 2016. **11**(10): p. e0165488.
229. Volkov, P., et al., *A Genome-Wide mQTL Analysis in Human Adipose Tissue Identifies Genetic Variants Associated with DNA Methylation, Gene Expression and Metabolic Traits.* PLoS One, 2016. **11**(6): p. e0157776.
230. Fraser, H.B., et al., *Population-specificity of human DNA methylation.* Genome biology, 2012. **13**(2): p. R8-R8.
231. Heyn, H., et al., *DNA methylation contributes to natural human variation.* Genome Res, 2013. **23**(9): p. 1363-72.
232. Moen, E.L., et al., *Genome-wide variation of cytosine modifications between European and African populations and the implications for complex traits.* Genetics, 2013. **194**(4): p. 987-96.
233. Zhang, F.F., et al., *Significant differences in global genomic DNA methylation by gender and race/ethnicity in peripheral blood.* Epigenetics, 2011. **6**(5): p. 623-629.
234. Czamara, D., et al., *Integrated analysis of environmental and genetic influences on cord blood DNA methylation in new-borns.* Nat Commun, 2019. **10**(1): p. 2548.
235. Teh, A.L., et al., *The effect of genotype and in utero environment on interindividual variation in neonate DNA methylomes.* Genome research, 2014. **24**(7): p. 1064-1074.
236. Feinberg, A.P. and R.A. Irizarry, *Evolution in health and medicine Sackler colloquium: Stochastic epigenetic variation as a driving force of development, evolutionary adaptation, and disease.* Proc Natl Acad Sci U S A, 2010. **107 Suppl 1**: p. 1757-64.
237. Feinberg, A.P., *Phenotypic plasticity and the epigenetics of human disease.* Nature, 2007. **447**(7143): p. 433-40.
238. Maurano, M.T., et al., *Systematic localization of common disease-associated variation in regulatory DNA.* Science, 2012. **337**(6099): p. 1190-5.
239. Schaub, M.A., et al., *Linking disease associations with regulatory information in the human genome.* Genome Res, 2012. **22**(9): p. 1748-59.
240. Do, C., et al., *Genetic-epigenetic interactions in cis: a major focus in the post-GWAS era.* Genome Biol, 2017. **18**(1): p. 120.
241. Gallagher, M.D. and A.S. Chen-Plotkin, *The Post-GWAS Era: From Association to Function.* Am J Hum Genet, 2018. **102**(5): p. 717-730.
242. Olsson, A.H., et al., *Genome-wide associations between genetic and epigenetic variation influence mRNA expression and insulin secretion in human pancreatic islets.* PLoS Genet, 2014. **10**(11): p. e1004735.
243. Hannon, E., et al., *An integrated genetic-epigenetic analysis of schizophrenia: evidence for co-localization of genetic associations and differential DNA methylation.* Genome Biol, 2016. **17**(1): p. 176.
244. Hannon, E., et al., *Methylation QTLs in the developing brain and their enrichment in schizophrenia risk loci.* Nat Neurosci, 2016. **19**(1): p. 48-54.
245. Liu, Y., et al., *Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis.* Nature biotechnology, 2013. **31**(2): p. 142-7.
246. Teh, A.L., et al., *The effect of genotype and in utero environment on interindividual variation in neonate DNA methylomes.* Genome Res, 2014. **24**(7): p. 1064-74.
247. Gutierrez-Arcelus, M., et al., *Passive and active DNA methylation and the interplay with genetic variation in gene regulation.* Elife, 2013. **2**: p. e00523.

248. Aran, D. and A. Hellman, *DNA methylation of transcriptional enhancers and cancer predisposition.* Cell, 2013. **154**(1): p. 11-3.
249. Kuroiwa-Trzmielina, J., et al., *SNP rs16906252C>T Is an Expression and Methylation Quantitative Trait Locus Associated with an Increased Risk of Developing MGMT-Methylated Colorectal Cancer.* Clin Cancer Res, 2016. **22**(24): p. 6266-6277.
250. McDonald, K.L., et al., *The T genotype of the MGMT C>T (rs16906252) enhancer single-nucleotide polymorphism (SNP) is associated with promoter methylation and longer survival in glioblastoma patients.* Eur J Cancer, 2013. **49**(2): p. 360-8.
251. Gutierrez-Arcelus, M., et al., *Tissue-specific effects of genetic and epigenetic variation on gene regulation and splicing.* PLoS Genet, 2015. **11**(1): p. e1004958.
252. Lee, P.P., et al., *A critical role for Dnmt1 and DNA methylation in T cell development, function, and survival.* Immunity, 2001. **15**(5): p. 763-74.
253. Makar, K.W. and C.B. Wilson, *DNA methylation is a nonredundant repressor of the Th2 effector program.* J Immunol, 2004. **173**(7): p. 4402-6.
254. Floess, S., et al., *Epigenetic control of the foxp3 locus in regulatory T cells.* PLoS Biol, 2007. **5**(2): p. e38.
255. Kim, H.P. and W.J. Leonard, *CREB/ATF-dependent T cell receptor-induced FoxP3 gene expression: a role for DNA methylation.* J Exp Med, 2007. **204**(7): p. 1543-51.
256. Kim, S.T., P.E. Fields, and R.A. Flavell, *Demethylation of a specific hypersensitive site in the Th2 locus control region.* Proc Natl Acad Sci U S A, 2007. **104**(43): p. 17052-7.
257. Gamper, C.J., et al., *Identification of DNA methyltransferase 3a as a T cell receptor-induced regulator of Th1 and Th2 differentiation.* J Immunol, 2009. **183**(4): p. 2267-76.
258. Thomas, R.M., et al., *De novo DNA methylation is required to restrict T helper lineage plasticity.* J Biol Chem, 2012. **287**(27): p. 22900-9.
259. Hashimoto, S., et al., *Coordinated changes in DNA methylation in antigen-specific memory CD4 T cells.* J Immunol, 2013. **190**(8): p. 4076-91.
260. Zhang, Y., et al., *Genome-wide DNA methylation analysis identifies hypomethylated genes regulated by FOXP3 in human regulatory T cells.* Blood, 2013. **122**(16): p. 2823-36.
261. Young, H.A., et al., *Differentiation of the T helper phenotypes by analysis of the methylation state of the IFN-gamma gene.* The Journal of Immunology, 1994. **153**(8): p. 3603-10.
262. Jones, B. and J. Chen, *Inhibition of IFN-gamma transcription by site-specific methylation during T helper cell development.* EMBO J, 2006. **25**(11): p. 2443-52.
263. Schoenborn, J.R., et al., *Comprehensive epigenetic profiling identifies multiple distal regulatory elements directing transcription of the gene encoding interferon-gamma.* Nat Immunol, 2007. **8**(7): p. 732-42.
264. Winders, B.R., R.H. Schwartz, and D. Bruniquel, *A distinct region of the murine IFN-gamma promoter is hypomethylated from early T cell development through mature naive and Th1 cell differentiation, but is hypermethylated in Th2 cells.* J Immunol, 2004. **173**(12): p. 7377-84.
265. Agarwal, S. and A. Rao, *Modulation of chromatin structure regulates cytokine gene expression during T cell differentiation.* Immunity, 1998. **9**(6): p. 765-75.
266. Bird, J.J., et al., *Helper T cell differentiation is controlled by the cell cycle.* Immunity, 1998. **9**(2): p. 229-37.
267. Lee, D.U., S. Agarwal, and A. Rao, *Th2 lineage commitment and efficient IL-4 production involves extended demethylation of the IL-4 gene.* Immunity, 2002. **16**(5): p. 649-60.

268. Thomas, R.M., L. Gao, and A.D. Wells, *Signals from CD28 Induce Stable Epigenetic Modification of the IL-2 Promoter.* The Journal of Immunology, 2005. **174**(8): p. 4639-4646.
269. Meyer, B., et al., *DNA methylation at IL32 in juvenile idiopathic arthritis.* Sci Rep, 2015. **5**: p. 11063.
270. Corvetta, A., et al., *5-Methylcytosine content of DNA in blood, synovial mononuclear cells and synovial tissue from patients affected by autoimmune rheumatic diseases.* J Chromatogr, 1991. **566**(2): p. 481-91.
271. Richardson, B., et al., *Evidence for impaired T cell DNA methylation in systemic lupus erythematosus and rheumatoid arthritis.* Arthritis Rheum, 1990. **33**(11): p. 1665-73.
272. Balasa, A., et al., *Assessment of DNA methylation at the interferon regulatory factor 5 (IRF5) promoter region in inflammatory bowel diseases.* Int J Colorectal Dis, 2010. **25**(5): p. 553-6.
273. Blomgren, S.E., J.J. Condemi, and J.H. Vaughan, *Procainamide-induced lupus erythematosus. Clinical and laboratory observations.* Am J Med, 1972. **52**(3): p. 338-48.
274. Cannat, A. and M. Seligmann, *Induction by isoniazid and hydrallazine of antinuclear factors in mice.* Clin Exp Immunol, 1968. **3**(1): p. 99-105.
275. Cornacchia, E., et al., *Hydralazine and procainamide inhibit T cell DNA methylation and induce autoreactivity.* J Immunol, 1988. **140**(7): p. 2197-200.
276. Neidhart, M., et al., *Retrotransposable L1 elements expressed in rheumatoid arthritis synovial tissue: association with genomic DNA hypomethylation and influence on gene expression.* Arthritis Rheum, 2000. **43**(12): p. 2634-47.
277. Quddus, J., et al., *Treating activated CD4+ T cells with either of two distinct DNA methyltransferase inhibitors, 5-azacytidine or procainamide, is sufficient to cause a lupus-like disease in syngeneic mice.* J Clin Invest, 1993. **92**(1): p. 38-53.
278. Richardson, B.C., et al., *Phenotypic and functional similarities between 5-azacytidine-treated T cells and a T cell subset in patients with active systemic lupus erythematosus.* Arthritis Rheum, 1992. **35**(6): p. 647-62.
279. Richardson, B., *Effect of an inhibitor of DNA methylation on T cells. II. 5-Azacytidine induces self-reactivity in antigen-specific T4+ cells.* Hum Immunol, 1986. **17**(4): p. 456-70.
280. Nile, C.J., et al., *Methylation status of a single CpG site in the IL6 promoter is related to IL6 messenger RNA levels and rheumatoid arthritis.* Arthritis Rheum, 2008. **58**(9): p. 2686-93.
281. Javierre, B.M., et al., *Changes in the pattern of DNA methylation associate with twin discordance in systemic lupus erythematosus.* Genome Res, 2010. **20**(2): p. 170-9.
282. Howell, K.J., et al., *DNA Methylation and Transcription Patterns in Intestinal Epithelial Cells From Pediatric Patients With Inflammatory Bowel Diseases Differentiate Disease Subtypes and Associate With Outcome.* Gastroenterology, 2018. **154**(3): p. 585-598.
283. McDermott, E., et al., *DNA Methylation Profiling in Inflammatory Bowel Disease Provides New Insights into Disease Pathogenesis.* J Crohns Colitis, 2016. **10**(1): p. 77-86.
284. Ventham, N.T., et al., *Integrative epigenome-wide analysis demonstrates that DNA methylation may mediate genetic risk in inflammatory bowel disease.* Nat Commun, 2016. **7**: p. 13507.
285. Coit, P., et al., *Renal involvement in lupus is characterized by unique DNA methylation changes in naive CD4+ T cells.* J Autoimmun, 2015. **61**: p. 29-35.

286. Somineni, H.K., et al., *Blood-Derived DNA Methylation Signatures of Crohn's Disease and Severity of Intestinal Inflammation.* Gastroenterology, 2019. **156**(8): p. 2254-2265 e3.
287. Absher, D.M., et al., *Genome-wide DNA methylation analysis of systemic lupus erythematosus reveals persistent hypomethylation of interferon genes and compositional changes to CD4+ T-cell populations.* PLoS genetics, 2013. **9**(8): p. e1003678.
288. Bell, C.G., et al., *Genome-wide DNA methylation analysis for diabetic nephropathy in type 1 diabetes mellitus.* BMC medical genomics, 2010. **3**: p. 33.
289. Coit, P., et al., *Genome-wide DNA methylation study suggests epigenetic accessibility andtranscriptional poising of interferon-regulated genes in naive CD4+ T cellsfrom lupus patients.* Journal of autoimmunity, 2013. **43**: p. 78-84.
290. Cooke, J., et al., *Mucosal genome-wide methylation changes in inflammatory bowel disease.* Inflammatory bowel diseases, 2012. **18**(11): p. 2128-37.
291. de la Rica, L., et al., *Identification of novel markers in rheumatoid arthritis through integrated analysis of DNA methylation and microRNA expression.* J Autoimmun, 2013. **41**: p. 6-16.
292. Ellis, J.A., et al., *Genome-scale case-control analysis of CD4+ T-cell DNA methylation in juvenile idiopathic arthritis reveals potential targets involved in disease.* Clinical epigenetics, 2012. **4**(1): p. 20.
293. Graves, M., et al., *Methylation differences at the HLA-DRB1 locus in CD4+ T-Cells are associated with multiple sclerosis.* Mult Scler, 2013.
294. Harris, R.A., et al., *Genome-wide peripheral blood leukocyte DNA methylation microarrays identified a single association with inflammatory bowel diseases.* Inflammatory bowel diseases, 2012. **18**(12): p. 2334-41.
295. Hasler, R., et al., *A functional methylome map of ulcerative colitis.* Genome research, 2012. **22**(11): p. 2130-7.
296. Jeffries, M.A., et al., *Genome-wide DNA methylation patterns in CD4+ T cells from patients with systemic lupus erythematosus.* Epigenetics : official journal of the DNA Methylation Society, 2011. **6**(5): p. 593-601.
297. Lin, S.Y., et al., *A whole genome methylation analysis of systemic lupus erythematosus: hypomethylation of the IL10 and IL1R2 promoters is associated with disease activity.* Genes and immunity, 2012. **13**(3): p. 214-20.
298. Lin, Z., et al., *Identification of disease-associated DNA methylation in intestinal tissues from patients with inflammatory bowel disease.* Clin Genet, 2011. **80**(1): p. 59-67.
299. Nakano, K., et al., *DNA methylome signature in rheumatoid arthritis.* Ann Rheum Dis, 2013. **72**(1): p. 110-7.
300. Nimmo, E.R., et al., *Genome-wide methylation profiling in Crohn's disease identifies altered epigenetic regulation of key host defense mechanisms including the Th17 pathway.* Inflammatory bowel diseases, 2012. **18**(5): p. 889-99.
301. Rakyan, V.K., et al., *Identification of type 1 diabetes-associated DNA methylation variable positions that precede disease diagnosis.* PLoS genetics, 2011. **7**(9): p. e1002300.
302. Whitaker, J.W., et al., *An imprinted rheumatoid arthritis methylome signature reflects pathogenic phenotype.* Genome Med, 2013. **5**(4): p. 40.
303. Selmi, C., et al., *Genome-wide analysis of DNA methylation, copy number variation, and gene expression in monozygotic twins discordant for primary biliary cirrhosis.* Front Immunol, 2014. **5**: p. 128.
304. Park, G.T., et al., *DNA methylation analysis of CD4+ T cells in patients with psoriasis.* Arch Dermatol Res, 2014. **306**(3): p. 259-68.

305. Zhang, P., et al., *Whole-genome DNA methylation in skin lesions from patients with psoriasis vulgaris.* J Autoimmun, 2013. **41**: p. 17-24.
306. Chung, S.A., et al., *Genome-Wide Assessment of Differential DNA Methylation Associated with Autoantibody Production in Systemic Lupus Erythematosus.* PLoS One, 2015. **10**(7): p. e0129813.
307. Imgenberg-Kreuz, J., et al., *DNA methylation mapping identifies gene regulatory effects in patients with systemic lupus erythematosus.* Ann Rheum Dis, 2018. **77**(5): p. 736-743.
308. Lanata, C.M., et al., *A phenotypic and genomics approach in a multi-ethnic cohort to subtype systemic lupus erythematosus.* Nat Commun, 2019. **10**(1): p. 3902.
309. Mok, A., et al., *Genome-wide profiling identifies associations between lupus nephritis and differential methylation of genes regulating tissue hypoxia and type 1 interferon responses.* Lupus Sci Med, 2016. **3**(1): p. e000183.
310. Julia, A., et al., *Epigenome-wide association study of rheumatoid arthritis identifies differentially methylated loci in B cells.* Hum Mol Genet, 2017. **26**(14): p. 2803-2811.
311. Mok, A., et al., *Hypomethylation of CYP2E1 and DUSP22 Promoters Associated With Disease Activity and Erosive Disease Among Rheumatoid Arthritis Patients.* Arthritis Rheumatol, 2018. **70**(4): p. 528-536.
312. Rodriguez-Ubreva, J., et al., *Inflammatory cytokines shape a changing DNA methylome in monocytes mirroring disease activity in rheumatoid arthritis.* Ann Rheum Dis, 2019. **78**(11): p. 1505-1516.
313. Adams, A.T., et al., *Two-stage genome-wide methylation profiling in childhood-onset Crohn's Disease implicates epigenetic alterations at the VMP1/MIR21 and HLA loci.* Inflamm Bowel Dis, 2014. **20**(10): p. 1784-93.
314. Paul, D.S., et al., *Increased DNA methylation variability in type 1 diabetes across three immune effector cell types.* Nat Commun, 2016. **7**: p. 13555.
315. Elboudwarej, E., et al., *Hypomethylation within gene promoter regions and type 1 diabetes in discordant monozygotic twins.* J Autoimmun, 2016. **68**: p. 23-9.
316. Rhead, B., et al., *Rheumatoid Arthritis Naive T Cells Share Hypermethylation Sites With Synoviocytes.* Arthritis Rheumatol, 2017. **69**(3): p. 550-559.
317. Glossop, J.R., et al., *DNA methylation at diagnosis is associated with response to disease-modifying drugs in early rheumatoid arthritis.* Epigenomics, 2017. **9**(4): p. 419-428.
318. Tseng, C.C., et al., *Next-Generation Sequencing Profiles of the Methylome and Transcriptome in Peripheral Blood Mononuclear Cells of Rheumatoid Arthritis.* J Clin Med, 2019. **8**(9).
319. Shao, X., et al., *Rheumatoid arthritis-relevant DNA methylation changes identified in ACPA-positive asymptomatic individuals using methylome capture sequencing.* Clin Epigenetics, 2019. **11**(1): p. 110.
320. Howell, K., et al., *Reply.* Gastroenterology, 2018. **155**(1): p. 230-231.
321. Houseman, E.A., J. Molitor, and C.J. Marsit, *Reference-free cell mixture adjustments in analysis of DNA methylation data.* Bioinformatics, 2014. **30**(10): p. 1431-9.
322. Nelis, M., et al., *Genetic structure of Europeans: a view from the North-East.* PLoS One, 2009. **4**(5): p. e5472.
323. Stover, P.J., *One-carbon metabolism-genome interactions in folate-associated pathologies.* J Nutr, 2009. **139**(12): p. 2402-5.
324. Imgenberg-Kreuz, J., et al., *Genome-wide DNA methylation analysis in multiple tissues in primary Sjogren's syndrome reveals regulatory effects at interferon-induced genes.* Ann Rheum Dis, 2016.

325. de Andres, M.C., et al., *Assessment of global DNA methylation in peripheral blood cell subpopulations of early rheumatoid arthritis before and after methotrexate.* Arthritis Res Ther, 2015. **17**: p. 233.
326. Cribbs, A.P., et al., *Methotrexate Restores Regulatory T Cell Function Through Demethylation of the FoxP3 Upstream Enhancer in Patients With Rheumatoid Arthritis.* Arthritis Rheumatol, 2015. **67**(5): p. 1182-92.
327. Ligthart, S., et al., *DNA methylation signatures of chronic low-grade inflammation are associated with complex diseases.* Genome Biol, 2016. **17**(1): p. 255.
328. Sokolove, J. and C.M. Lepus, *Role of inflammation in the pathogenesis of osteoarthritis: latest findings and interpretations.* Ther Adv Musculoskelet Dis, 2013. **5**(2): p. 77-94.
329. Boyum, A., *Isolation of mononuclear cells and granulocytes from human blood. Isolation of monuclear cells by one centrifugation, and of granulocytes by combining centrifugation and sedimentation at 1 g.* Scand J Clin Lab Invest Suppl, 1968. **97**: p. 77-89.
330. Frommer, M., et al., *A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands.* Proceedings of the National Academy of Sciences, 1992. **89**(5): p. 1827-1831.
331. Clark, S.J., et al., *High sensitivity mapping of methylated cytosines.* Nucleic Acids Res, 1994. **22**(15): p. 2990-7.
332. *Bisulphite treatment protocol.*
333. Dedeurwaerder, S., et al., *Evaluation of the Infinium Methylation 450K technology.* Epigenomics, 2011. **3**(6): p. 771-84.
334. Aryee, M.J., et al., *Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays.* Bioinformatics, 2014. **30**(10): p. 1363-9.
335. Maksimovic, J., L. Gordon, and A. Oshlack, *SWAN: Subset-quantile within array normalization for illumina infinium HumanMethylation450 BeadChips.* Genome Biol, 2012. **13**(6): p. R44.
336. Du, P., et al., *Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis.* BMC Bioinformatics, 2010. **11**: p. 587.
337. Langfelder, P. and S. Horvath, *WGCNA: an R package for weighted correlation network analysis.* BMC Bioinformatics, 2008. **9**(1): p. 559.
338. Phipson, B., J. Maksimovic, and A. Oshlack, *missMethyl: an R package for analyzing data from Illumina's HumanMethylation450 platform.* Bioinformatics, 2016. **32**(2): p. 286-8.
339. Maksimovic, J., et al., *Removing unwanted variation in a differential methylation analysis of Illumina HumanMethylation450 array data.* Nucleic Acids Res, 2015. **43**(16): p. e106.
340. Ehrich, M., et al., *Quantitative high-throughput analysis of DNA methylation patterns by base-specific cleavage and mass spectrometry.* Proc Natl Acad Sci U S A, 2005. **102**(44): p. 15785-90.
341. Coolen, M.W., et al., *Genomic profiling of CpG methylation and allelic specificity using quantitative high-throughput mass spectrometry: critical evaluation and improvements.* Nucleic Acids Res, 2007. **35**(18): p. e119.
342. Thompson, R.F., et al., *A pipeline for the quantitative analysis of CG dinucleotide methylation using mass spectrometry.* Bioinformatics, 2009. **25**(17): p. 2164-70.
343. Bioscience, A. *High-Resolution Quantitative Methylation Profiling with EpiTYPER® and the MassARRAY® System*. 2015; Available from: http://agenabio.com/wp-content/uploads/2015/06/51-20055R1.0-EpiTYPER-Brochure_WEB.pdf.

344. Ross, P., et al., *High level multiplex genotyping by MALDI-TOF mass spectrometry.* Nat Biotechnol, 1998. **16**(13): p. 1347-51.
345. Gabriel, S., L. Ziaugra, and D. Tabbaa, *SNP genotyping using the Sequenom MassARRAY iPLEX platform.* Curr Protoc Hum Genet, 2009. **Chapter 2**: p. Unit 2 12.
346. Bioscience, A., *Single Nucleotide Polymorphism Detection with the iPLEX® Assay and the MassARRAY® System*. 2015.
347. Stram, D.O., *Tag SNP selection for association studies.* Genetic Epidemiology, 2004. **27**(4): p. 365-374.
348. Barrett, J.C., et al., *Haploview: analysis and visualization of LD and haplotype maps.* Bioinformatics, 2005. **21**(2): p. 263-5.
349. Quon, G., et al., *Patterns of methylation heritability in a genome-wide analysis of four brain regions.* Nucleic Acids Res, 2013. **41**(4): p. 2095-104.
350. H., W., *ggplot2: Elegant Graphics for Data Analysis*. 2009, New York: Springer-Verlag.
351. Purcell, S., et al., *PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses.* Am J Hum Genet, 2007. **81**(3): p. 559-75.
352. Risso, D., et al., *Normalization of RNA-seq data using factor analysis of control genes or samples.* Nat Biotech, 2014. **32**(9): p. 896-902.
353. Risso, D., et al., *GC-content normalization for RNA-Seq data.* BMC Bioinformatics, 2011. **12**: p. 480.
354. McCarthy, D.J., Y. Chen, and G.K. Smyth, *Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation.* Nucleic Acids Res, 2012. **40**(10): p. 4288-97.
355. Wettenhall, J.M. and G.K. Smyth, *limmaGUI: a graphical user interface for linear modeling of microarray data.* Bioinformatics, 2004. **20**(18): p. 3705-6.
356. Vrba, L. and B.W. Futscher, *A suite of DNA methylation markers that can detect most common human cancers.* Epigenetics, 2018. **13**(1): p. 61-72.
357. Garcia-Pineres, A.J., et al., *DNAse treatment following thawing of cryopreserved PBMC is a procedure suitable for lymphocyte functional studies.* J Immunol Methods, 2006. **313**(1-2): p. 209-13.
358. Rawson, J.B., et al., *Vitamin D intake is negatively associated with promoter methylation of the Wnt antagonist gene DKK1 in a large group of colorectal cancer patients.* Nutr Cancer, 2012. **64**(7): p. 919-28.
359. Zhu, H., et al., *A genome-wide methylation study of severe vitamin D deficiency in African American adolescents.* J Pediatr, 2013. **162**(5): p. 1004-9 e1.
360. Chavez Valencia, R.A., et al., *In vitro exposure of human blood mononuclear cells to active vitamin D does not induce substantial change to DNA methylation on a genome-scale.* J Steroid Biochem Mol Biol, 2014. **141**: p. 144-9.
361. Wellcome Trust Case Control, C., *Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls.* Nature, 2007. **447**(7145): p. 661-78.
362. Hakonarson, H., et al., *A genome-wide association study identifies KIAA0350 as a type 1 diabetes gene.* Nature, 2007. **448**(7153): p. 591-4.
363. Blomberg Jensen, M., et al., *Vitamin D metabolism and effects on pluripotency genes and cell differentiation in testicular germ cell tumors in vitro and in vivo.* Neoplasia, 2012. **14**(10): p. 952-63.
364. Underwood, K.F., et al., *Regulation of RUNX2 transcription factor-DNA interactions and cell proliferation by vitamin D3 (cholecalciferol) prohormone activity.* J Bone Miner Res, 2012. **27**(4): p. 913-25.
365. Bibikova, M., et al., *High density DNA methylation array with single CpG site resolution.* Genomics, 2011. **98**(4): p. 288-95.

366. Bibikova, M., et al., *Genome-wide DNA methylation profiling using Infinium(R) assay.* Epigenomics, 2009. **1**(1): p. 177-200.
367. Eckhardt, F., et al., *DNA methylation profiling of human chromosomes 6, 20 and 22.* Nat Genet, 2006. **38**(12): p. 1378-85.
368. Hogg, K., et al., *Prenatal and perinatal environmental influences on the human fetal and placental epigenome.* Clin Pharmacol Ther, 2012. **92**(6): p. 716-26.
369. Mirbahai, L. and J.K. Chipman, *Epigenetic memory of environmental organisms: a reflection of lifetime stressor exposures.* Mutat Res Genet Toxicol Environ Mutagen, 2014. **764-765**: p. 10-7.
370. Barnes, S.K. and S.E. Ozanne, *Pathways linking the early environment to long-term health and lifespan.* Prog Biophys Mol Biol, 2011. **106**(1): p. 323-36.
371. Godfrey, K.M., et al., *Epigenetic gene promoter methylation at birth is associated with child's later adiposity.* Diabetes, 2011. **60**(5): p. 1528-34.
372. Stamp, L.K. and R.L. Roberts, *Effect of genetic polymorphisms in the folate pathway on methotrexate therapy in rheumatic diseases.* Pharmacogenomics, 2011. **12**(10): p. 1449-63.
373. Plant, D., et al., *Differential methylation as a biomarker of response to etanercept in patients with rheumatoid arthritis.* Arthritis Rheumatol, 2016.
374. Webster, A., et al., *OP0257 Differential DNA Methylation Related to Response to Adalimumab and Etanercept in Patients with Rheumatoid Arthritis.* Annals of the Rheumatic Diseases, 2014. **73**(Suppl 2): p. 158-159.
375. Mease, P.J., *Adalimumab in the treatment of arthritis.* Ther Clin Risk Manag, 2007. **3**(1): p. 133-48.
376. Ansel, K.M., et al., *Regulation of Th2 differentiation and Il4 locus accessibility.* Annu Rev Immunol, 2006. **24**: p. 607-56.
377. Hatton, R.D., et al., *A distal conserved sequence element controls Ifng gene expression by T cells and NK cells.* Immunity, 2006. **25**(5): p. 717-29.
378. Consolaro, A., et al., *Advances in biomarkers for paediatric rheumatic diseases.* Nat Rev Rheumatol, 2015. **11**(5): p. 265-275.
379. Mourao, A.F., et al., *Using the Juvenile Arthritis Disease Activity Score based on erythrocyte sedimentation rate or C-reactive protein level: results from the Portuguese register.* Arthritis Care Res (Hoboken), 2014. **66**(4): p. 585-91.
380. Dempster, E., et al., *Epigenetic Studies of Schizophrenia: Progress, Predicaments, and Promises for the Future.* Schizophr Bull, 2013. **39**(1): p. 11-6.
381. Ladd-Acosta, C. and M.D. Fallin, *The role of epigenetics in genetic and environmental epidemiology.* Epigenomics, 2016. **8**(2): p. 271-83.
382. Alisch, R.S., et al., *Age-associated DNA methylation in pediatric populations.* Genome Res, 2012. **22**(4): p. 623-32.
383. Horvath, S., et al., *Aging effects on DNA methylation modules in human brain and blood tissue.* Genome Biol, 2012. **13**(10): p. R97.
384. El-Maarri, O., et al., *Gender specific differences in levels of DNA methylation at selected loci from human total blood: a tendency toward higher methylation levels in males.* Hum Genet, 2007. **122**(5): p. 505-14.
385. Sarter, B., et al., *Sex differential in methylation patterns of selected genes in Singapore Chinese.* Human Genetics, 2005. **117**(4): p. 402-403.
386. McCarthy, N.S., et al., *Meta-analysis of human methylation data for evidence of sex-specific autosomal patterns.* BMC Genomics, 2014. **15**: p. 981.
387. Singmann, P., et al., *Characterization of whole-genome autosomal differences of DNA methylation between men and women.* Epigenetics Chromatin, 2015. **8**: p. 43.
388. Chiaroni-Clarke, R.C., et al., *The association of PTPN22 rs2476601 with juvenile idiopathic arthritis is specific to females.* Genes Immun, 2015. **16**(7): p. 495-8.

389. Chiaroni-Clarke, R.C., J.E. Munro, and J.A. Ellis, *Sex bias in paediatric autoimmune disease - Not just about sex hormones?* J Autoimmun, 2016. **69**: p. 12-23.
390. Mayne, B.T., et al., *Large Scale Gene Expression Meta-Analysis Reveals Tissue-Specific, Sex-Biased Gene Expression in Humans.* Front Genet, 2016. **7**: p. 183.
391. Benjamini, Y. and Y. Hochberg, *Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing.* Journal of the Royal Statistical Society. Series B (Methodological), 1995. **57**(1): p. 289-300.
392. Huang da, W., B.T. Sherman, and R.A. Lempicki, *Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists.* Nucleic Acids Res, 2009. **37**(1): p. 1-13.
393. Huang da, W., B.T. Sherman, and R.A. Lempicki, *Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources.* Nat Protoc, 2009. **4**(1): p. 44-57.
394. Supek, F., et al., *REVIGO summarizes and visualizes long lists of gene ontology terms.* PLoS One, 2011. **6**(7): p. e21800.
395. Evans, J.D., *Straightforward statistics for the behavioral sciences*. 1996, Pacific Grove: Brooks/Cole Pub. Co.
396. Olkhov-Mitsel, E. and B. Bapat, *Strategies for discovery and validation of methylated and hydroxymethylated DNA biomarkers.* Cancer Med, 2012. **1**(2): p. 237-60.
397. Michels, K.B., et al., *Recommendations for the design and analysis of epigenome-wide association studies.* Nat Meth, 2013. **10**(10): p. 949-955.
398. Rakyan, V.K., et al., *Epigenome-wide association studies for common human diseases.* Nat Rev Genet, 2011. **12**(8): p. 529-541.
399. Kolla, V., et al., *The tumour suppressor CHD5 forms a NuRD-type chromatin remodelling complex.* Biochem J, 2015. **468**(2): p. 345-52.
400. Thompson, P.M., et al., *CHD5, a new member of the chromodomain gene family, is preferentially expressed in the nervous system.* Oncogene, 2003. **22**(7): p. 1002-11.
401. Zaslavsky, E., et al., *Investigating immunological pathways and diseases with a comprehensive compendium of human data (HUM8P.347).* The Journal of Immunology, 2014. **192**(1 Supplement): p. 185.22.
402. Richmond, R.C., et al., *Prenatal exposure to maternal smoking and offspring DNA methylation across the lifecourse: findings from the Avon Longitudinal Study of Parents and Children (ALSPAC).* Hum Mol Genet, 2015. **24**(8): p. 2201-17.
403. Rotroff, D.M., et al., *Maternal smoking impacts key biological pathways in newborns through epigenetic modification in Utero.* BMC Genomics, 2016. **17**(1): p. 976.
404. Sugiyama, D., et al., *Impact of smoking as a risk factor for developing rheumatoid arthritis: a meta-analysis of observational studies.* Ann Rheum Dis, 2010. **69**(1): p. 70-81.
405. Rupasree, Y., et al., *Application of Various Statistical Models to Explore Gene-Gene Interactions in Folate, Xenobiotic, Toll-Like Receptor and STAT4 Pathways that Modulate Susceptibility to Systemic Lupus Erythematosus.* Mol Diagn Ther, 2016. **20**(1): p. 83-95.
406. Anderson, M.S. and M.A. Su, *Aire and T cell Development.* Curr Opin Immunol, 2011. **23**(2): p. 198-206.
407. Guerra, E., et al., *The Trop-2 signalling network in cancer growth.* Oncogene, 2013. **32**(12): p. 1594-600.
408. Zhang, Q., et al., *Nesprins: a novel family of spectrin-repeat-containing proteins that localize to the nuclear membrane in multiple tissues.* J Cell Sci, 2001. **114**(Pt 24): p. 4485-98.
409. Abazyan, B., et al., *Prenatal interaction of mutant DISC1 and immune activation produces adult psychopathology.* Biol Psychiatry, 2010. **68**(12): p. 1172-81.

410. Lipina, T.V., et al., *Maternal immune activation during gestation interacts with Disc1 point mutation to exacerbate schizophrenia-related behaviors in mice.* J Neurosci, 2013. **33**(18): p. 7654-66.
411. Ko, J.-K., et al., *The tail-anchoring domain of Bfl1 and HCCS1 targets mitochondrial membrane permeability to induce apoptosis.* Journal of Cell Science, 2007. **120**(16): p. 2912-2923.
412. Tuller, T., et al., *Common and specific signatures of gene expression and protein-protein interactions in autoimmune diseases.* Genes Immun, 2013. **14**(2): p. 67-82.
413. Strimbu, K. and J.A. Tavel, *What are biomarkers?* Curr Opin HIV AIDS, 2010. **5**(6): p. 463-6.
414. Staal, F.J., T.C. Luis, and M.M. Tiemessen, *WNT signalling in the immune system: WNT is spreading its wings.* Nat Rev Immunol, 2008. **8**(8): p. 581-93.
415. Rincon, M. and R.J. Davis, *Regulation of the immune response by stress-activated protein kinases.* Immunol Rev, 2009. **228**(1): p. 212-24.
416. Thurman, R.E., et al., *The accessible chromatin landscape of the human genome.* Nature, 2012. **489**(7414): p. 75-82.
417. Horvath, S., *DNA methylation age of human tissues and cell types.* Genome Biol, 2013. **14**(10): p. R115.
418. Donn, R., et al., *Network analysis and juvenile idiopathic arthritis (JIA): a new horizon for the understanding of disease pathogenesis and therapeutic target identification.* Pediatr Rheumatol Online J, 2016. **14**.
419. Akimzhanov, A.M., X.O. Yang, and C. Dong, *Chromatin remodeling of interleukin-17 (IL-17)-IL-17F cytokine gene locus during inflammatory helper T cell differentiation.* J Biol Chem, 2007. **282**(9): p. 5969-72.
420. Martino, D., et al., *Genome-scale profiling reveals a subset of genes regulated by DNA methylation that program somatic T-cell phenotypes in humans.* Genes Immun, 2012. **13**(5): p. 388-98.
421. Youngblood, B., J.S. Hale, and R. Ahmed, *T-cell memory differentiation: insights from transcriptional signatures and epigenetics.* Immunology, 2013. **139**(3): p. 277-84.
422. Glossop, J.R., et al., *Genome-wide profiling in treatment-naive early rheumatoid arthritis reveals DNA methylome changes in T- and B-lymphocytes.* Epigenomics, 2015.
423. Glossop, J.R., et al., *Genome-wide profiling in treatment-naive early rheumatoid arthritis reveals DNA methylome changes in T and B lymphocytes.* Epigenomics, 2016. **8**(2): p. 209-24.
424. Dugue, P.A., et al., *The repeatability of DNA methylation measures may also affect the power of epigenome-wide association studies.* Int J Epidemiol, 2015. **44**(4): p. 1460-1.
425. Rakyan, V.K., et al., *Epigenome-wide association studies for common human diseases.* Nat Rev Genet, 2011. **12**(8): p. 529-41.
426. Smolewska, E., et al., *Apoptosis of peripheral blood lymphocytes in patients with juvenile idiopathic arthritis.* Ann Rheum Dis, 2003. **62**(8): p. 761-3.
427. Smolewska, E., et al., *Regulation of peripheral blood and synovial fluid lymphocyte apoptosis in juvenile idiopathic arthritis.* Scand J Rheumatol, 2004. **33**(1): p. 7-12.
428. Manchia, M., et al., *The Impact of Phenotypic and Genetic Heterogeneity on Results of Genome Wide Association Studies of Complex Diseases.* PLoS ONE, 2013. **8**(10): p. e76295.
429. Shearer, W.T., et al., *Lymphocyte subsets in healthy children from birth through 18 years of age: the Pediatric AIDS Clinical Trials Group P1009 study.* J Allergy Clin Immunol, 2003. **112**(5): p. 973-80.

430. Simon, A.K., G.A. Hollander, and A. McMichael, *Evolution of the immune system in humans from infancy to old age.* Proc Biol Sci, 2015. **282**(1821): p. 20143085.
431. Kerkel, K., et al., *Genomic surveys by methylation-sensitive SNP analysis identify sequence-dependent allele-specific DNA methylation.* Nat Genet, 2008. **40**(7): p. 904-8.
432. Shoemaker, R., et al., *Allele-specific methylation is prevalent and is contributed by CpG-SNPs in the human genome.* Genome Research, 2010. **20**(7): p. 883-889.
433. Ernst, J., et al., *Mapping and analysis of chromatin state dynamics in nine human cell types.* Nature, 2011. **473**(7345): p. 43-9.
434. Kent, W.J., et al., *The human genome browser at UCSC.* Genome Res, 2002. **12**(6): p. 996-1006.
435. Garg, P., et al., *A survey of inter-individual variation in DNA methylation identifies environmentally responsive co-regulated networks of epigenetic variation in the human genome.* PLoS Genet, 2018. **14**(10): p. e1007707.
436. Hunt, S.E., et al., *Ensembl variation resources.* Database (Oxford), 2018. **2018**.
437. Meng, W., et al., *DNA methylation mediates genotype and smoking interaction in the development of anti-citrullinated peptide antibody-positive rheumatoid arthritis.* Arthritis Res Ther, 2017. **19**(1): p. 71.
438. Islam, S.A., et al., *Integration of DNA methylation patterns and genetic variation in human pediatric tissues help inform EWAS design and interpretation.* Epigenetics Chromatin, 2019. **12**(1): p. 1.
439. Edgar, R.D., et al., *BECon: a tool for interpreting DNA methylation findings from blood in the context of brain.* Transl Psychiatry, 2017. **7**(8): p. e1187.
440. Cassidy, J.T., et al., *A study of classification criteria for a diagnosis of juvenile rheumatoid arthritis.* Arthritis Rheum, 1986. **29**(2): p. 274-81.
441. Cook, N.R., *Statistical evaluation of prognostic versus diagnostic models: beyond the ROC curve.* Clin Chem, 2008. **54**(1): p. 17-23.
442. Pencina, M.J., et al., *Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond.* Stat Med, 2008. **27**(2): p. 157-72; discussion 207-12.
443. Jiang, K., et al., *RNA-Sequencing In Peripheral Blood Mononuclear Cells Identifies Novel Differentially Expressed Coding and Non-Coding Transcripts In Juvenile Idiopathic Arthritis. [abstract].* Arthritis Rheumatol, 2013. **65 Suppl 10**: p. 2194.
444. Pearson, M.J. and S.W. Jones, *Review: Long Noncoding RNAs in the Regulation of Inflammatory Pathways in Rheumatoid Arthritis and Osteoarthritis.* Arthritis Rheumatol, 2016. **68**(11): p. 2575-2583.
445. Li, Y., et al., *Genome-wide DNA methylome analysis reveals epigenetically dysregulated non-coding RNAs in human breast cancer.* Sci Rep, 2015. **5**: p. 8790.
446. Dybkaer, K., et al., *Genome wide transcriptional analysis of resting and IL2 activated human natural killer cells: gene expression signatures indicative of novel molecular signaling pathways.* BMC Genomics, 2007. **8**: p. 230.
447. Matsumura, T., et al., *TIFAB inhibits TIFA, TRAF-interacting protein with a forkhead-associated domain.* Biochem Biophys Res Commun, 2004. **317**(1): p. 230-4.
448. Pralhada Rao, R., et al., *Sphingolipid Metabolic Pathway: An Overview of Major Roles Played in Human Diseases.* J Lipids, 2013. **2013**.
449. Szklarczyk, D., et al., *STRING v10: protein-protein interaction networks, integrated over the tree of life.* Nucleic Acids Res, 2015. **43**(Database issue): p. D447-52.
450. Avila-Pedretti, G., et al., *Variation at FCGR2A and functionally related genes is associated with the response to anti-TNF therapy in rheumatoid arthritis.* PLoS One, 2015. **10**(4): p. e0122088.

451. Mao, J., et al., *Guanine nucleotide exchange factor GEF115 specifically mediates activation of Rho and serum response factor by the G protein alpha subunit Galpha13.* Proc Natl Acad Sci U S A, 1998. **95**(22): p. 12973-6.
452. Tomankova, T., et al., *MicroRNAs: Emerging Regulators of Immune-Mediated Diseases.* Scand J Immunol, 2012. **75**(2): p. 129-41.
453. Plaza-Izurieta, L., et al., *Expression analysis in intestinal mucosa reveals complex relations among genes under the association peaks in celiac disease.* Eur J Hum Genet, 2015. **23**(8): p. 1100-5.
454. Nakamura, M., et al., *Genome-wide association study identifies TNFSF15 and POU2AF1 as susceptibility loci for primary biliary cirrhosis in the Japanese population.* Am J Hum Genet, 2012. **91**(4): p. 721-8.
455. Komori, H.K., et al., *Defining CD4 T cell memory by the epigenetic landscape of CpG DNA methylation.* J Immunol, 2015. **194**(4): p. 1565-79.
456. Janson, P.C., et al., *FOXP3 promoter demethylation reveals the committed Treg population in humans.* PLoS One, 2008. **3**(2): p. e1612.
457. Albers, H.M., et al., *Genetic variation in VTCN1 (B7-H4) is associated with course of disease in juvenile idiopathic arthritis.* Ann Rheum Dis, 2014. **73**(6): p. 1198-201.
458. Berntson, L., et al., *Anti-type II collagen antibodies, anti-CCP, IgA RF and IgM RF are associated with joint damage, assessed eight years after onset of juvenile idiopathic arthritis (JIA).* Pediatr Rheumatol Online J, 2014. **12**: p. 22.
459. Botta Gordon-Smith, S., et al., *Correlation of low CD73 expression on synovial lymphocytes with reduced adenosine generation and higher disease severity in juvenile idiopathic arthritis.* Arthritis Rheumatol, 2015. **67**(2): p. 545-54.
460. Fantini, F., et al., *Remission in juvenile chronic arthritis: a cohort study of 683 consecutive cases with a mean 10 year followup.* J Rheumatol, 2003. **30**(3): p. 579-84.
461. Ringold, S., et al., *Inactive disease in polyarticular juvenile idiopathic arthritis: current patterns and associations.* Rheumatology (Oxford), 2009. **48**(8): p. 972-7.
462. Chen, Y., et al., *APC gene hypermethylation and prostate cancer: a systematic review and meta-analysis.* Eur J Hum Genet, 2013. **21**(9): p. 929-35.
463. Darwiche, K., et al., *Assessment of SHOX2 methylation in EBUS-TBNA specimen improves accuracy in lung cancer staging.* Ann Oncol, 2013. **24**(11): p. 2866-70.
464. Dietrich, D., et al., *DNA methylation of the homeobox genes PITX2 and SHOX2 predicts outcome in non-small-cell lung cancer patients.* Diagn Mol Pathol, 2012. **21**(2): p. 93-104.
465. Dietrich, D., et al., *Diagnostic and prognostic value of SHOX2 and SEPT9 DNA methylation and cytology in benign, paramalignant and malignant pleural effusions.* PLoS One, 2013. **8**(12): p. e84225.
466. Ilse, P., et al., *SHOX2 DNA methylation is a tumour marker in pleural effusions.* Cancer Genomics Proteomics, 2013. **10**(5): p. 217-23.
467. Kneip, C., et al., *SHOX2 DNA methylation is a biomarker for the diagnosis of lung cancer in plasma.* J Thorac Oncol, 2011. **6**(10): p. 1632-8.
468. Pan, J., et al., *Association between RASSF1A promoter methylation and prostate cancer: a systematic review and meta-analysis.* PLoS One, 2013. **8**(9): p. e75283.
469. Payne, S.R., *From discovery to the clinic: the novel DNA methylation biomarker (m)SEPT9 for the detection of colorectal cancer in blood.* Epigenomics, 2010. **2**(4): p. 575-85.
470. Schmidt, B., et al., *SHOX2 DNA methylation is a biomarker for the diagnosis of lung cancer based on bronchial aspirates.* BMC Cancer, 2010. **10**: p. 600.
471. Wu, T., et al., *Measurement of GSTP1 promoter methylation in body fluids may complement PSA screening: a meta-analysis.* Br J Cancer, 2011. **105**(1): p. 65-73.

472. Mikeska, T., et al., *DNA methylation biomarkers in cancer: progress towards clinical implementation.* Expert Rev Mol Diagn, 2012. **12**(5): p. 473-87.
473. Zhang, Y., et al., *Impaired DNA methylation and its mechanisms in CD4(+)T cells of systemic lupus erythematosus.* Journal of autoimmunity, 2013. **41**: p. 92-9.
474. Zhao, M., et al., *Increased 5-hydroxymethylcytosine in CD4 T cells in systemic lupus erythematosus.* J Autoimmun, 2016.
475. Zhao, M., et al., *IFI44L promoter methylation as a blood biomarker for systemic lupus erythematosus.* Ann Rheum Dis, 2016.
476. Oughtred, R., et al., *The BioGRID interaction database: 2019 update.* Nucleic Acids Res, 2019. **47**(D1): p. D529-D541.
477. Du, Z., et al., *The epigenetic modifier CHD5 functions as a novel tumor suppressor for renal cell carcinoma and is predominantly inactivated by promoter CpG methylation.* Oncotarget, 2016. **7**(16): p. 21618-30.
478. Fatemi, M., et al., *Epigenetic silencing of CHD5, a novel tumor-suppressor gene, occurs in early colorectal cancer stages.* Cancer, 2014. **120**(2): p. 172-80.
479. Liu, J.B., et al., *Influence of colorectal cancer tumor suppressor gene CHD5 methylation on its clinical and pathological characteristics.* J Biol Regul Homeost Agents, 2015. **29**(4): p. 889-93.
480. Ma, Z., et al., *Decreased expression of the CHD5 gene and its clinicopathological significance in breast cancer: Correlation with aberrant DNA methylation.* Oncol Lett, 2016. **12**(5): p. 4021-4026.
481. Mulero-Navarro, S. and M. Esteller, *Chromatin remodeling factor CHD5 is silenced by promoter CpG island hypermethylation in human cancer.* Epigenetics, 2008. **3**(4): p. 210-5.
482. Quan, J. and T. Yusufzai, *The tumor suppressor chromodomain helicase DNA-binding protein 5 (CHD5) remodels nucleosomes by unwrapping.* J Biol Chem, 2014. **289**(30): p. 20717-26.
483. Wang, J., et al., *The involvement of CHD5 hypermethylation in laryngeal squamous cell carcinoma.* Oral Oncol, 2011. **47**(7): p. 601-8.
484. Wang, X., et al., *CHD5 is down-regulated through promoter hypermethylation in gastric cancer.* J Biomed Sci, 2009. **16**: p. 95.
485. Zhao, R., et al., *Silencing of CHD5 gene by promoter methylation in leukemia.* PLoS One, 2014. **9**(1): p. e85172.
486. Zhao, R., et al., *CHD5 a tumour suppressor is epigenetically silenced in hepatocellular carcinoma.* Liver Int, 2014. **34**(6): p. e151-60.
487. Groom, A., et al., *Postnatal growth and DNA methylation are associated with differential gene expression of the TACSTD2 gene and childhood fat mass.* Diabetes, 2012. **61**(2): p. 391-400.
488. Kanai, Y., et al., *MRNA expression of genes altered by 5-azacytidine treatment in cancer cell lines is associated with clinicopathological parameters of human cancers.* J Cancer Res Clin Oncol, 2001. **127**(12): p. 697-706.
489. Lin, J.C., et al., *TROP2 is epigenetically inactivated and modulates IGF-1R signalling in lung adenocarcinoma.* EMBO Mol Med, 2012. **4**(6): p. 472-85.
490. Sawanyawisuth, K., et al., *Suppression of trophoblast cell surface antigen 2 enhances proliferation and migration in liver fluke-associated cholangiocarcinoma.* Ann Hepatol, 2016. **15**(1): p. 71-81.
491. Almamun, M., et al., *Integrated methylome and transcriptome analysis reveals novel regulatory elements in pediatric acute lymphoblastic leukemia.* Epigenetics, 2015. **10**(9): p. 882-90.

492. Melotte, V., et al., *Spectrin repeat containing nuclear envelope 1 and forkhead box protein E1 are promising markers for the detection of colorectal cancer in blood.* Cancer Prev Res (Phila), 2015. **8**(2): p. 157-64.
493. Papadia, C., et al., *FOXE1 and SYNE1 genes hypermethylation panel as promising biomarker in colitis-associated colorectal neoplasia.* Inflamm Bowel Dis, 2014. **20**(2): p. 271-7.
494. Tessema, M., et al., *Promoter methylation of genes in and around the candidate lung cancer susceptibility locus 6q23-25.* Cancer Res, 2008. **68**(6): p. 1707-14.
495. Kolla, V., et al., *Role of CHD5 in human cancers: 10 years later.* Cancer Res, 2014. **74**(3): p. 652-8.
496. Basta, J. and M. Rauchman, *The nucleosome remodeling and deacetylase complex in development and disease.* Transl Res, 2015. **165**(1): p. 36-47.
497. Tsujikawa, M., et al., *Identification of the gene responsible for gelatinous drop-like corneal dystrophy.* Nat Genet, 1999. **21**(4): p. 420-3.
498. Goldenberg, D.M., R. Stein, and R.M. Sharkey, *The emergence of trophoblast cell-surface antigen 2 (TROP-2) as a novel cancer target.* Oncotarget, 2018. **9**(48): p. 28989-29006.
499. Eisenwort, G., et al., *Identification of TROP2 (TACSTD2), an EpCAM-like molecule, as a specific marker for TGF-beta1-dependent human epidermal Langerhans cells.* J Invest Dermatol, 2011. **131**(10): p. 2049-57.
500. Picarda, G., et al., *Functional Langerinhigh-Expressing Langerhans-like Cells Can Arise from CD14highCD16- Human Blood Monocytes in Serum-Free Condition.* J Immunol, 2016. **196**(9): p. 3716-28.
501. Sin, S.T.K., et al., *Down-regulation of TROP-2 Predicts Poor Prognosis of Hepatocellular Carcinoma Patients.* Hepatol Commun, 2018. **2**(11): p. 1408-1414.
502. Joo, J.E., et al., *Heritable DNA methylation marks associated with susceptibility to breast cancer.* Nat Commun, 2018. **9**(1): p. 867.
503. McRae, A.F., et al., *Contribution of genetic variation to transgenerational inheritance of DNA methylation.* Genome Biol, 2014. **15**(5): p. R73.
504. Bell, J.T., et al., *Epigenome-wide scans identify differentially methylated regions for age and age-related phenotypes in a healthy ageing population.* PLoS Genet, 2012. **8**(4): p. e1002629.
505. Hannon, E., et al., *Characterizing genetic and environmental influences on variable DNA methylation using monozygotic and dizygotic twins.* PLoS Genet, 2018. **14**(8): p. e1007544.
506. Gervin, K., et al., *Extensive variation and low heritability of DNA methylation identified in a twin study.* Genome Res, 2011. **21**(11): p. 1813-21.
507. Boks, M.P., et al., *The relationship of DNA methylation with age, gender and genotype in twins and healthy controls.* PLoS One, 2009. **4**(8): p. e6767.
508. Gordon, L., et al., *Neonatal DNA methylation profile in human twins is specified by a complex interplay between intrauterine environmental and genetic factors, subject to tissue-specific influence.* Genome Res, 2012. **22**(8): p. 1395-406.
509. Kaminsky, Z.A., et al., *DNA methylation profiles in monozygotic and dizygotic twins.* Nat Genet, 2009. **41**(2): p. 240-5.
510. Ek, W.E., et al., *Genetic variants influencing phenotypic variance heterogeneity.* Hum Mol Genet, 2018. **27**(5): p. 799-810.
511. Gibson, D.S., et al., *Vitamin D binding protein isoforms as candidate predictors of disease extension in childhood arthritis.* J Proteomics, 2012. **75**(17): p. 5479-92.
512. Michels, K.B. and A.M. Binder, *Considerations for Design and Analysis of DNA Methylation Studies*, in *DNA Methylation Protocols*, J. Tost, Editor. 2018, Springer New York: New York, NY. p. 31-46.

513. Teschendorff, A.E. and C.L. Relton, *Statistical and integrative system-level analysis of DNA methylation data.* Nat Rev Genet, 2018. **19**(3): p. 129-147.
514. Feinberg, A.P., *Genome-scale approaches to the epigenetics of common human disease.* Virchows Arch, 2010. **456**(1): p. 13-21.
515. Mountz, J.D., et al., *Autoimmune disease. A problem of defective apoptosis.* Arthritis Rheum, 1994. **37**(10): p. 1415-20.
516. Lorenz, H.M., et al., *Role of apoptosis in autoimmunity.* Apoptosis, 2000. **5**(5): p. 443-9.

Appendices

Appendix I: Manuscript

DNA methylation at IL32 in juvenile idiopathic arthritis

SCIENTIFIC REPORTS

OPEN

DNA methylation at *IL32* in juvenile idiopathic arthritis

Braydon Meyer[1], Raul A Chavez[1,2], Jane E Munro[3,4], Rachel C Chiaroni-Clarke[1,2], Jonathan D Akikusa[3,4], Roger C Allen[3,4], Jeffrey M Craig[5,2], Anne-Louise Ponsonby[5,2], Richard Saffery[6,2] & Justine A Ellis[1,2]

Received: 08 February 2015
Accepted: 11 May 2015
Published: 09 June 2015

Juvenile idiopathic arthritis (JIA) is the most common autoimmune rheumatic disease of childhood. We recently showed that DNA methylation at the gene encoding the pro-inflammatory cytokine interleukin-32 (*IL32*) is reduced in JIA CD4+ T cells. To extend this finding, we measured *IL32* methylation in CD4+ T-cells from an additional sample of JIA cases and age- and sex-matched controls, and found a reduction in methylation associated with JIA consistent with the prior data (combined case-control dataset: 25.0% vs 37.7%, p = 0.0045). Further, JIA was associated with reduced *IL32* methylation in CD8+ T cells (15.2% vs 25.5%, p = 0.034), suggesting disease-associated changes to a T cell precursor. Additionally, we measured regional SNPs, along with CD4+ T cell expression of total *IL32*, and the γ and β isoforms. Several SNPs were associated with methylation. Two SNPs were also associated with JIA, and we found evidence of interaction such that methylation was only associated with JIA in minor allele carriers (e.g. rs10431961 $p_{interaction} = 0.011$). Methylation at one measured CpG was inversely correlated with total *IL32* expression (Spearman r = −0.73, p = 0.0009), but this was not a JIA-associated CpG. Overall, our data further confirms that reduced *IL32* methylation is associated with JIA, and that SNPs play an interactive role.

Juvenile idiopathic arthritis (JIA) is the most common autoimmune rheumatic disease in children, affecting approximately one per thousand of European descent, and causing significant disability[1,2]. JIA is considered to be a complex disease, and as such, risk is presumed to be determined by both genetic and environmental factors[3]. Recent work has identified a number of genetic risk loci, including HLA, *PTPN22*, and genes of the IL-2 pathway[4,5]. Specific environmental factors potentially contributing to risk remain largely uncharacterised; however, a role for environmental exposures is supported by numerous lines of evidence for other autoimmune diseases[6].

Epigenetic modification, including DNA methylation, is widely believed to provide a mechanism through which environmental and genetic risk factors interact to promote disease[7]. Indeed, there is a large and rapidly-growing body of evidence for altered DNA methylation in individuals with various autoimmune diseases, including rheumatoid arthritis (RA)[8–11], systemic lupus erythematosus[12–14], inflammatory bowel disease[15–17] and type 1 diabetes[18,19]. We recently performed a pilot genome-scale analysis of peripheral blood CD4+ T cell DNA methylation in (predominantly oligoarticular course) JIA[20]. Among the genomic regions showing evidence of differential methylation in JIA cases relative to age- and sex-matched controls, was the regulatory region of the cytokine interleukin-32 gene (*IL32*). IL-32 promotes the production of a number of pro-inflammatory molecules, including tumour necrosis factor alpha (TNFα) and interleukin 6[21–23], cytokines that are targeted in biologic therapies for JIA[24], and is therefore a biologically plausible candidate gene for JIA pathogenesis.

[1]Genes, Environment & Complex Disease, Murdoch Childrens Research Institute, Parkville, Victoria, Australia. [2]Department of Paediatrics, University of Melbourne, Victoria, Australia. [3]Arthritis & Rheumatology, Murdoch Childrens Research Institute, Parkville, Victoria, Australia. [4]Paediatric Rheumatology Unit, Royal Children's Hospital, Parkville, Victoria, Australia. [5]Environmental & Genetic Epidemiology Research, Murdoch Childrens Research Institute, Parkville, Victoria, Australia. [6]Cancer & Disease Epigenetics, Murdoch Childrens Research Institute, Parkville, Victoria, Australia. Correspondence and requests for materials should be addressed to J.A.E. (email: justine.ellis@mcri.edu.au)

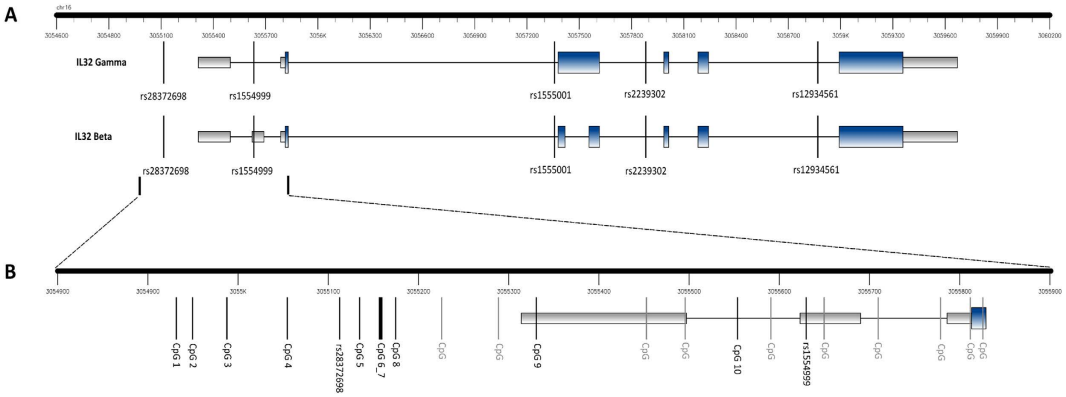

Figure 1. Graphical representation of the gene structure of *IL32*, and location of CpGs and SNPs measured in this study. A. *IL32* Gamma and Beta isoforms with regional SNPs annotated. B. Close-up of 5' gene region with measured CpGs annotated, other unmeasured CpGs in the region are shown in grey. Blue bars = exons. Grey bars = UTRs. Hg18 co-ordinates.

IL-32 has been implicated in a number of different biological pathways, including cell death, host defence, and immune function[25,26]. Several lines of evidence also suggest that IL-32 plays an important role in inflammatory arthritis (for review, see[27]). Shoda *et al.* showed that *IL32* expression is prominent in synovial-infiltrated lymphocytes of RA patients, and overexpression has the capacity to exacerbate inflammatory arthritis in mice[28]. Park *et al.* showed that injection of IL-32 into mouse knee joints led to increased pro-inflammatory cytokine production, synovial inflammation and bone destruction[29]. Promotion of joint inflammation by IL-32 appears to be TNFα dependent, since IL-32 did not induce inflammation in TNFα-deficient mice[30].

Six isoforms of IL-32 exist (γ, β, α, δ, ε, ζ), as the result of mRNA splicing[21,31,32]. Heinhuis *et al.* observed that IL-32γ expression was significantly greater in RA compared to osteoarthritis (OA) synovial biopsies, and that in RA synovial tissues, IL-32γ expression was correlated with expression of pro-inflammatory cytokines including TNFα[33]. Splicing of IL-32γ to IL-32β reduced pro-inflammatory potency, demonstrated by a lack of correlation of IL-32β with TNFα or IL-6, suggesting that splicing may serve as a mechanism to control inflammation[33].

Given our prior evidence of reduced DNA methylation at *IL32* in CD4+ T cells of individuals with JIA, and mounting evidence for a role for IL-32 in inflammatory arthritis, we sought to generate further evidence that *IL32* plays a role in JIA. We investigated *IL32* DNA methylation in circulating CD4+ T cells in a further JIA case-control sample, and in an additional immune cell subset, CD8+ T cells, considered the impact of single nucleotide polymorphisms (SNPs) on DNA methylation, and explored the relationship between CD4+ T cell *IL32* methylation and *IL32* expression.

Results

A 5′ region of *IL32* is differentially methylated in JIA CD4+ T cells. We previously found evidence for differential methylation of two CpG sites (designated CpG_10 and CpG_9; Fig. 1) in the 5' region of *IL32* in CD4+ T cells of children with JIA[20]. To confirm this association, we measured DNA methylation in the same genomic region (total of 10 CpG sites within nine assay units) in purified CD4+ T cells from a replication sample of 12 oligoarticular JIA cases and age- and sex-matched controls. The CD4+ *IL32* methylation beta values for each replication sample at each measured CpG unit are given in Supplementary Table 5. As with our previous pilot data, we identified a significant reduction in mean methylation in JIA cases relative to controls (CpG_10: 10.3% (95% CI 5.8–14.9%) vs 19.0% (95% CI 12.0–26.1%), t-test p = 0.032; CpG_9: 29.2% (95% CI 20.0–38.4%) vs 45.9% (95% CI 33.6–58.3%), t-test p = 0.028) (Fig. 2A,B). To maximise sample size for subsequent analyses, we combined all 'original' and 'replication' sample methylation data (Fig. 2C,D). Overall, mean DNA methylation in JIA cases at both CpG_10 and CpG_9 was significantly reduced relative to controls in the combined dataset (CpG_10: 10.7% (95% CI 7.9–13.7%) vs 17.1% (95% CI 12.3–21.9%), t test p = 0.025; CpG_9: 25.0% (95% CI 19.6–30.3%) vs 37.7% (95% CI 30.6–44.8%), p = 0.0045). Logistic regression demonstrated that, for CpG_9, a 1% increase in methylation was associated with a 5% decrease in the risk of JIA (OR = 0.95, 95% CI 0.92, 0.99, p = 0.008). CD4+ *IL32* methylation levels at seven other analytic units (comprising 8 CpG sites) were not associated with JIA (Supplementary Table 5).

To maximise homogeneity amongst the cases, we reanalysed the association of JIA with *IL32* methylation after removing the four polyarticular JIA cases from the 'original' sample set. Patterns of association were unchanged (data not shown).

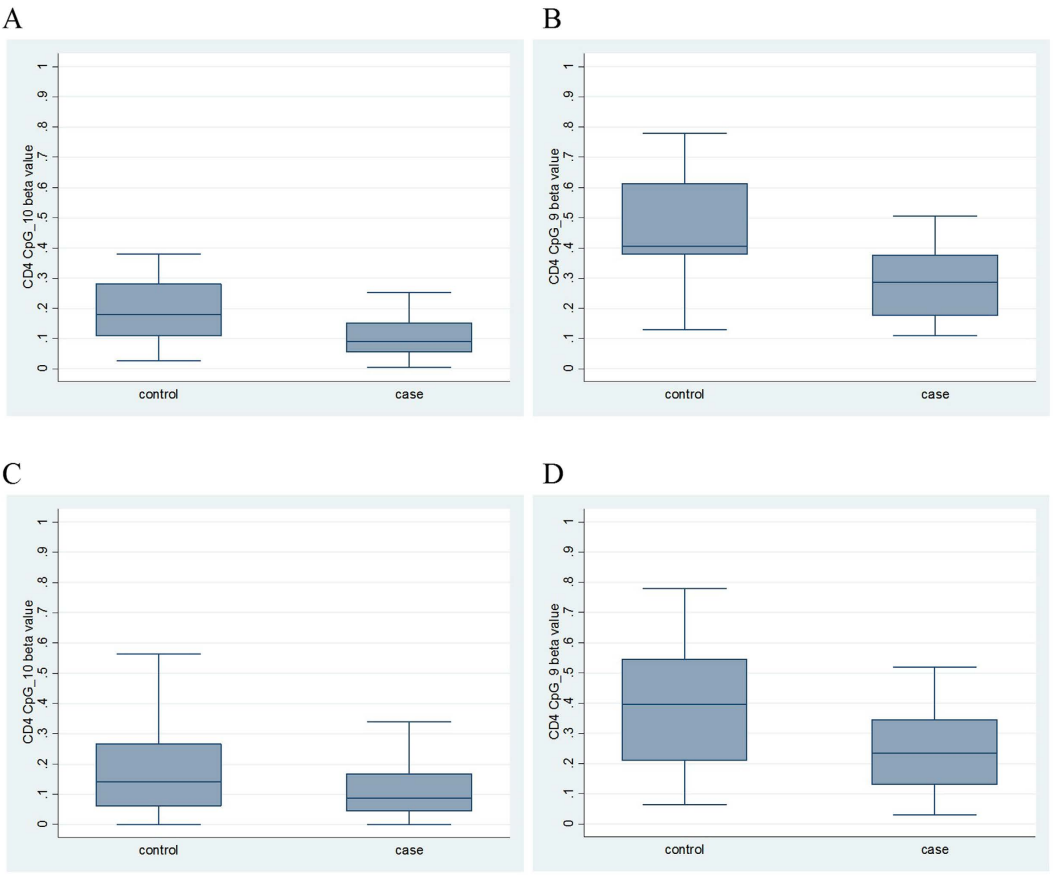

Figure 2. Methylation beta values in JIA cases and controls at the two significantly different *IL32* CpG sites. A: CpG_10 in replication samples. B: CpG_9 in replication samples. C: CpG_10 in original and replication samples combined. D: CpG_9 in original and replication samples combined.

A similar pattern of *IL32* differential methylation is seen in JIA CD8+ T cells. In order to explore the potential cell specificity of the observed CD4+ T cell differential methylation in JIA, we measured *IL32* methylation in CD8+ T cells isolated simultaneously with the CD4+ T cells in the 'replication' case-control pairs. Resulting methylation values are given in Supplementary Table 5. For all CpGs measured, mean CD8+ T cell *IL32* methylation was lower than in matched CD4+ T cells. As with CD4+ T cells, CD8+ T cell methylation at CpG_10 and CpG_9 was associated with JIA (CpG_10: 5.9% (95% CI 2.5–9.3%) vs 11.6% (95% CI 7.1–16.2%), t-test $p = 0.037$; CpG_9: 15.2% (95% CI 9.1–21.4%) vs 25.5% (95% CI 17.7–33.3%), t-test $p = 0.034$). Logistic regression demonstrated that, for CpG_9, a 1% increase in methylation was marginally associated with a 10% decrease in the risk of being a JIA case (OR = 0.91, 95% CI 0.83, 1.00, $p = 0.054$). CD8+ *IL32* methylation at other measured CpGs was not associated with JIA.

We measured the correlation of CpG methylation within and between CD4+ and CD8+ T cells. In general, there was a high level of correlation between CpG sites within each cell type (Supplementary Tables 6 and 7). CpG_9 and CpG_10 were highly correlated in both CD4+ T cells (Pearson $r = 0.86$, $p < 0.0001$) and CD8+ T cells (Pearson $r = 0.84$, $p < 0.0001$). When methylation levels were compared between cell types, CpG_1, CpG_3, CpG_8, CpG_9 and CpG_10 in CD4+ T cells were significantly correlated with their equivalent CpG in CD8+ T cells (Table 1).

SNPs in the *IL32* gene region are associated with methylation. We found the genotypes of a number of SNPs in the *IL32* region of chromosome 16 to be correlated with methylation at specific CpG sites. Table 2 shows the nine SNPs that were significantly associated with at least one CpG methylation level in at least one cell type. Methylation at CpG_10 and CpG_9 in CD4+ T cells appeared to be most commonly impacted by genotype, associated with six SNPs and four SNPs respectively. The SNP with the strongest overall evidence for association with *IL32* methylation (high number of CpGs and low p value for association) was rs1554999. Figure 3 shows the correlation plots of rs1554999 with methylation at the three CpGs with which it is correlated in CD4+ T cells.

		CD8+ T cells								
		CpG_1	CpG_2	CpG_3	CpG_4	CpG_5	CpG_6.7	CpG_8	CpG_9	CpG_10
CD4+ T cells	CpG_1	0.53 0.012								
	CpG_2		-0.0051 0.98							
	CpG_3			0.84 <0.0001						
	CpG_4				0.049 0.83					
	CpG_5					0.41 0.15				
	CpG_6.7						0.41 0.061			
	CpG_8							0.59 0.0080		
	CpG_9								0.56 0.011	
	CpG_10									0.57 0.0032

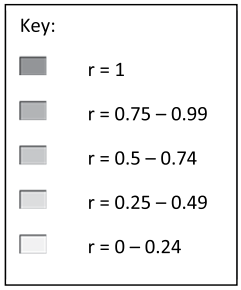

Table 1. Correlations (Pearson r followed by p-value) between CD4+ and CD8+ T cells for each measured CpG (cases and controls combined).

CD4+ T cell *IL32* methylation remains associated with JIA after taking into account the effect of SNPs. We noted that two SNPs found to be associated with *IL32* methylation, rs10431961 and rs7188573, were also associated with JIA by logistic regression (rs10431961: OR = 2.74; 95% CI 1.23, 6.12; p = 0.014; rs7188573: OR = 3.14; 95% CI 1.32, 7.49, p = 0.010). Given this, we considered them as confounders in the logistic regression model assessing the association between methylation and JIA. After adjusting for these two SNPs, CD4+ T cell CpG_10 methylation no longer remained associated with JIA (OR = 0.96, 95% CI 0.91, 1.01, p = 0.11). However, the association was still apparent for CpG_9 (OR = 0.96, 95% CI 0.92, 0.99, p = 0.034). Again, these patterns of association were not materially altered by the removal of polyarticular JIA cases.

Evidence of interaction between genotype and *IL32* methylation in JIA risk. We looked for evidence of interaction between SNPs rs10431961 and rs7188573 (main effect associations with JIA), and CD4+ DNA methylation at CpG_9. For rs10431961, we found that the association between JIA and CpG_9 methylation was present only in those with at least one minor allele (major allele homozygote group: AOR = 0.99, 95% CI 0.94–1.04, p = 0.71; heterozygote plus minor allele homozygote group: AOR = 0.85, 95% CI 0.74–0.98, p = 0.023). There was statistically significant evidence of multiplicative interaction between CpG_9 and rs10431961 (AOR_{int} = 0.86, 95% CI 0.76–0.97, p_{int} = 0.011). Similarly for rs7188573, we found that the association between JIA and CpG_9 was present only in those with at least one minor allele (major allele homozygote group: AOR = 0.98, 95% CI 0.93–1.03, p = 0.34; heterozygote plus minor allele homozygote group: AOR = 0.90, 95% CI 0.82–0.99, p = 0.028). Again, the test for interaction was statistically significant (AOR_{int} = 0.89, 95% CI 0.80–0.99, p_{int} = 0.039). SNPs rs10431961 and rs7188573 are in moderate linkage disequilibrium in our sample (r^2 = 0.76). However, conditioning the above interaction analyses on the opposite SNP did not materially alter the outcomes, suggesting that these two SNPs may be interacting independently with CD4+ T cell DNA methylation at CpG_9. The outcomes of these analyses were not altered by removal of polyarticular JIA cases.

Relationship between *IL32* methylation and expression in CD4+ T cells. In order to directly ascertain the functional relevance of the observed differential methylation of *IL32* in JIA, we measured expression of total *IL32* (the sum of all transcript variants), *IL32*γ, and *IL32*β in CD4+ T cells from JIA cases and controls. Expression values were normalised to the housekeeping gene *B2M*; normalisation to an alternative housekeeping gene, *RPLPO*, did not materially alter the values (data not shown). Figure 4 shows the distribution of relative expression values in JIA cases and controls. Supplementary Table 8 provides individual-level relative expression values. There was a large range of values for each measure of *IL32* expression in both cases and controls. Overall, we detected higher levels of expression of *IL32*β than *IL32*γ, consistent with prior evidence that *IL32*β is the more predominant *IL32* gene product in T cells[31]. Regression analysis adjusting for age, sex, and blood time to processing demonstrated no association between JIA and *IL32* expression (total *IL32*: mean difference = 0.59, 95% CI −1.91, 3.08; p = 0.63; *IL32*γ: mean difference = −0.00045; 95% CI −0.0051, 0.0042, p = 0.84; *IL32*β: mean difference = 0.060; 95% CI −0.093, 0.21; p = 0.43), although expression of total and β *IL32* was higher in the cases, as might be expected with reduced case methylation. The lack of association between JIA and *IL32* expression was unchanged on removal of the four polyarticular JIA cases from the dataset.

Pearson correlation				Linear regression*			
SNP	CpG	Cell type	Pearson r	p	coefficient	95% CI	p
rs3094471	10	CD4	0.31	0.014	5.57	1.37, 9.77	0.010
	3	CD4	0.30	0.019	9.18	1.15, 17.21	0.026
rs10431961	10	CD4	−0.29	0.020	−5.29	−9.49, −1.09	0.015
	9	CD4	−0.40	0.0018	−11.27	−17.62, −4.93	0.001
	5	CD4	−0.27	0.048	−6.32	−11.39, −1.26	0.015
	10	CD8	−0.42	0.041	−4.39	−8.15, −0.62	0.025
rs10438593	10	CD4	−0.32	0.010	−7.75	−13.44, −2.06	0.008
	9	CD4	−0.43	0.0008	−16.61	−24.73, −8.48	<0.001
rs7188573	10	CD4	−0.32	0.012	−6.17	−10.64, −1.71	0.008
	9	CD4	−0.44	0.0005	−12.77	−19.23, −6.32	<0.001
	6.7	CD4	−0.28	0.033	−4.17	−7.22, −1.13	0.008
	5	CD4	−0.33	0.016	−7.65	−12.74, −2.55	0.004
	10	CD8	−0.59	0.0025	−8.02	−11.83, −4.21	<0.001
	9	CD8	−0.40	0.059	−10.20	−18.87, −1.53	0.024
rs28372698	10	CD4	−0.27	0.030	−4.21	−8.05, −0.36	0.032
	3	CD4	−0.76	<0.0001	−21.6	−26.56, −16.64	<0.001
	3	CD8	−0.63	0.0011	−17.21	−26.20, −8.21	0.001
rs1554999	10	CD4	0.54	<0.0001	8.71	5.44, 11.99	<0.001
	9	CD4	0.60	<0.0001	14.09	9.20, 18.99	<0.001
	3	CD4	0.54	<0.0001	16.10	9.74, 22.46	<0.001
	9	CD8	0.43	0.040	8.52	1.71, 15.34	0.017
	3	CD8	0.66	0.0004	21.78	13.28, 30.29	<0.001
rs2239302	6.7	CD8	−0.45	0.027	−14.93	−29.04, −0.83	0.039
rs3789033	4	CD8	−0.44	0.033	−7.56	−13.96, −1.15	0.023
	2	CD8	−0.67	0.0008	−4.20	−7.17, −1.23	0.008
rs2239316	4	CD8	−0.44	0.034	−6.84	−12.65, −1.03	0.023
	2	CD8	−0.63	0.0020	−3.43	−6.22, −0.64	0.019

Table 2. Significant associations between genotyped SNPs and *IL32* CpG methylation in CD4+ and CD8+ T cells. *For linear regression, CpG methylation was converted to % methylation (β*100) for easier interpretation of the coefficient. The coefficient was adjusted for age and sex. The coefficient represents the change in % methylation for every additional minor allele of the SNP.

We next examined the correlations amongst total, γ and β isoforms in the CD4+ *IL32* combined case-control expression dataset. We observed correlation between total and γ IL32 relative expression (Spearman r = 0.45, p = 0.017) and total and β IL32 relative expression (Spearman r = 0.66, p = 0.0002). There was also a significant correlation between γ and β *IL32* (Spearman r = 0.70, p = < 0.0001). Next, we looked for correlation between CD4+ *IL32* methylation and expression amongst all available samples. There was evidence for a negative correlation between total *IL32* relative expression and CpG_3 (Spearman r = −0.73, p = 0.0009) (Fig. 5A) and *IL32*β relative expression and CpG_3 (Spearman r = −0.48, p = 0.049) (Fig. 5B). In both cases the correlation was in the biologically-predicted direction, that is, expression increased as CpG_3 methylation decreased. We used linear regression to adjust these associations for age, sex and time to blood sample processing and found that the association of total *IL32* with CpG_3 methylation remained significant (β coefficient = −0.071, 95% CI = −0.14, −0.0041, p = 0.039), as did the association between *IL32*β relative expression and CpG_3 methylation (β coefficient = −0.0041, 95%CI = −0.0081, −0.000045, p = 0.048. Methylation levels at no other CpGs were associated with *IL32* expression.

In totality, these analyses do not provide compelling evidence of a direct link between *IL32* expression level and 5' *IL32* methylation status in CD4+ T cells, or *IL32* gene expression levels and JIA. However, the identification of specific associations in a subset of comparisons warrants further investigation.

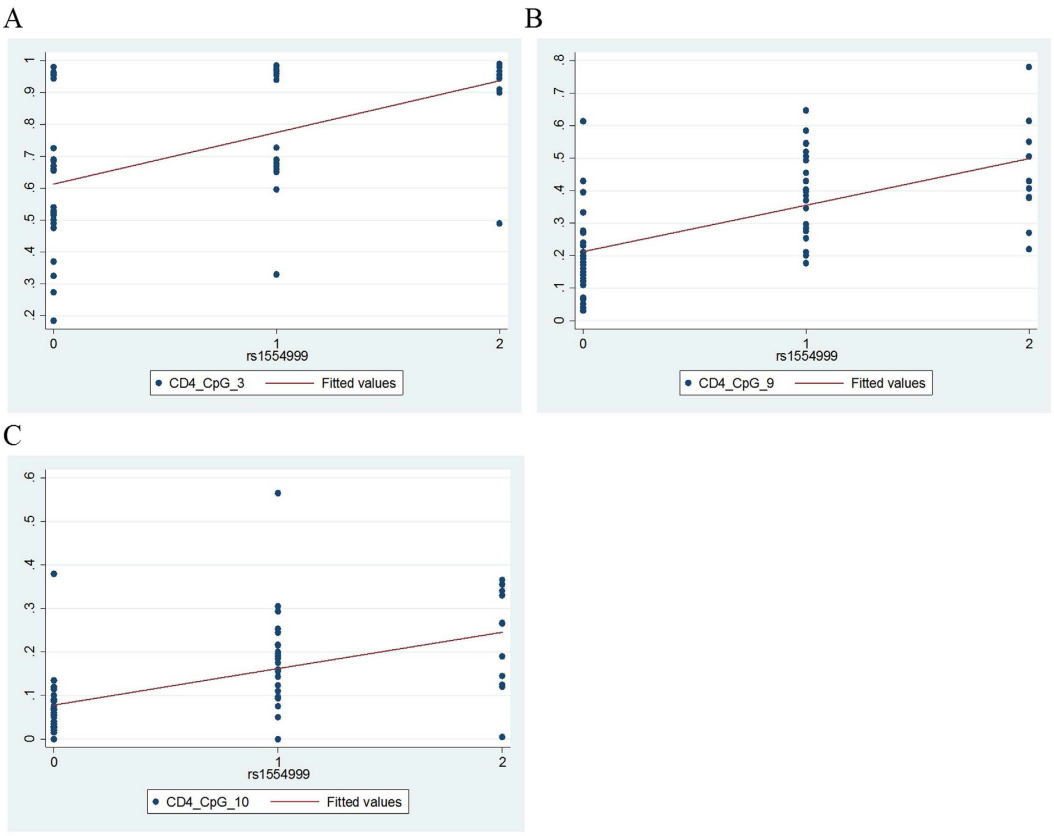

Figure 3. Correlations between genotype at rs1554999 and *IL32* β methylation values for CD4+ T cell CpGs (cases and controls combined). Only statistically significant correlations are shown. A. CD4 CpG_3, Pearson r = 0.54. B. CD4 CpG_9, Pearson r = 0.60. C. CD4 CpG_10, Pearson r = 0.54. SNP rs1554999 genotype 0 = CC, 1 = CA, 2 = AA (minor allele homozygote).

SNPs in the *IL32* gene region are not correlated with *IL32* expression. We did not find any evidence of correlation between *IL32* expression and genotyped SNPs (data not shown).

Overall relationships. The overall relationships of the *IL32* genomic measures amongst themselves, and with JIA, are shown in Fig. 6. Taken together our results show that the association between reduced *IL32* methylation and increased risk of JIA is robust, and that regional genotypes interact with *IL32* 5' methylation to determine disease risk.

Discussion

We have confirmed the presence of reduced DNA methylation at CpGs in the 5' region of *IL32* in circulating T cells in children with oligoarticular course JIA. Our previous data provided evidence, consistent in two case-control samples and across two assay platforms, of reduced *IL32* methylation in JIA CD4+ T cells[20]. Here, we have shown, in a replication case-control sample, a statistically significant association between JIA and *IL32* CpG methylation, with a direction of effect that is consistent with the prior work. When data from the original and replication samples (total 33 case-control pairs) was combined, the presence of association was further confirmed.

To complement these findings, we also measured DNA methylation at *IL32* in CD8+ T cells and found that, in general, this cell type was less methylated than CD4+ T cells. This is consistent with prior evidence that *IL32* is more highly expressed in CD8+ T cells[34], and suggests a role for *IL32* methylation in *IL32* expression. Given that CD4+ (helper) and CD8+ (cytotoxic) T cells have functionally distinct roles in the immune system, the function of *IL32* may differ between cell types. Indeed, there is evidence that both endogenous and exogenous agents stimulate *IL32* expression in a number of different cell types, suggesting that *IL32* might function in several aspects of immune defense via distinct pathways (for review see[26]). To our knowledge, however, no direct comparison of response of *IL32* to such stimuli between CD4+ and CD8+ T cells has been performed.

There was, however, a significant correlation between CD4+ and CD8+ T cells, and methylation at the key site CpG_9 in CD8+ T cells was also associated with JIA. These data suggest two things. First, the correlation between the two cell types, along with the presence of disease association with both cell types, suggests the establishment of the JIA-related differential methylation patterns in a T cell precursor.

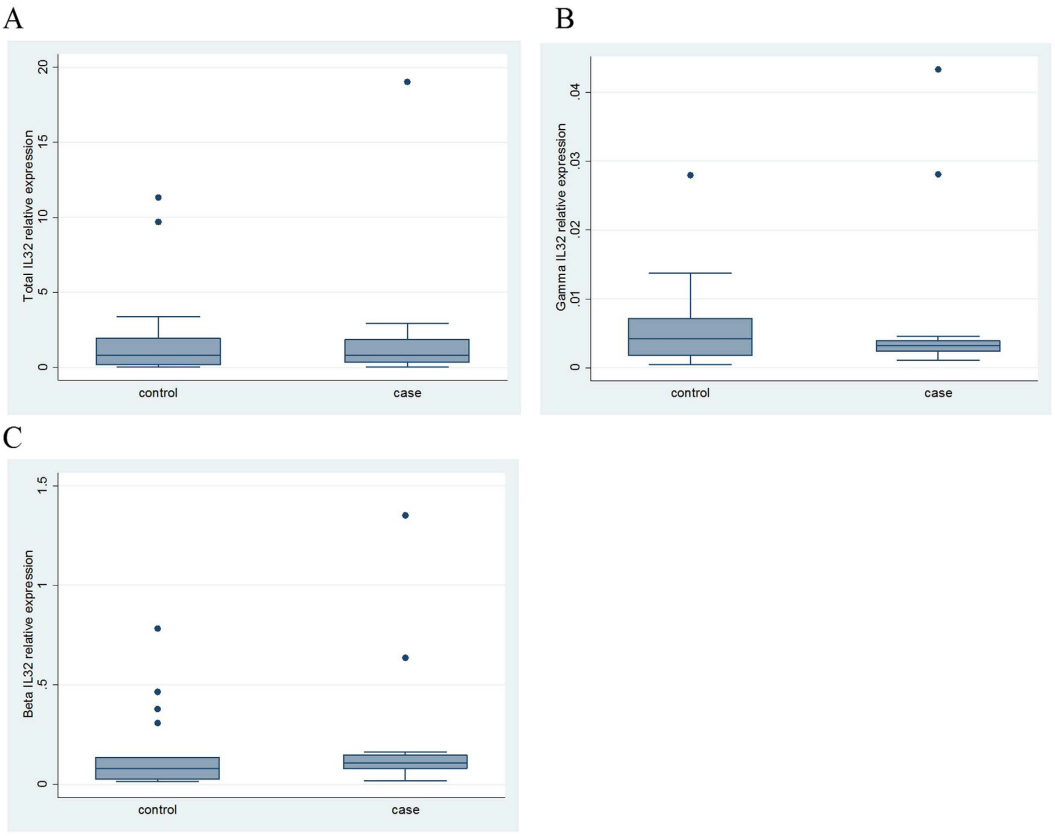

Figure 4. Distribution of relative expression of A. total *IL32*, B. *IL32*γ and C. *IL32*β in CD4+ T cells from cases and controls. *IL32* expression relative to *B2M* housekeeper gene expression.

To understand how far back in the immune cell lineage this differential methylation pattern extends, further work should examine *IL32* methylation in a wider range of immune cell types. Our data concerning CD4+ and CD8+ T cells suggests that the 'establishment' of differential methylation extends at least back prior to maturation to 'single positive' T cells in the thymus[35]. This supports a role for differential *IL32* methylation in disease causation, rather than disease consequence, which might more likely be reflected by changes in specific circulating immune cell subsets undertaking distinct functions in relation to disease processes. Correlation of disease-associated methylation changes across multiple immune cell types (T cells, B cells, monocytes) has also been reported in the autoimmune disease systemic lupus erythematosus (SLE)[12]. Such correlation amongst immune cells might provide the potential for DNA methylation to be used as a disease biomarker that is measureable in whole blood, simplifying the assay and thus increasing clinical utility.

By considering both regional SNP genotype and *IL32* gene expression, we have also significantly extended our understanding of the relationships amongst these genomic measures, and how they impact the association between *IL32* methylation and JIA. There is now considerable evidence that genetic sequence variants often impact methylation levels (methylation quantitative trait loci, or mQTLs), particularly early in life[36,37]. Data are also emerging implicating genotype-methylation interactions in phenotypic diversity apparent in conditions such as arthritis[9]. Recent data suggest the heritability of methylation profile per se to be around 20%[38]. A number of SNPs genotyped in this study were associated with *IL32* methylation, and associations differed between CpG sites and between T cell types. Of note, rs1554999 was strongly associated with methylation at CpG_10 in CD4+ T cells, and at CpG_9 and CpG_3 in both CD4+ and CD8+ T cells. This SNP is located in the 5' region of *IL32* (see Fig. 1), 642 bp, 291 bp and 76 bp downstream of CpG_3, CpG_9 and CpG_10 respectively. This SNP has also been strongly associated with *IL32* expression in peripheral blood (see Supplementary Table 3)[39]. This SNP in particular likely represents an *IL32 cis* methylation quantitative trait locus (mQTL). Two SNPs, rs10431961 and rs7188573, which lie 15 kb and 4 kb upstream of the *IL32* transcription start site respectively, were associated both with CD4+ T cell CpG_10 and CpG_9 methylation, and with JIA. Given that JIA is associated with reduced *IL32* methylation, that the minor alleles of the above two SNPs are associated with reduced *IL32* methylation, and that the minor alleles of the two SNPs are associated with an increased risk of JIA in our sample, these SNPs might explain the association between JIA and reduced *IL32* CpG methylation. When these SNPs were included as covariates in the logistic regression model assessing association between JIA and *IL32* methylation, evidence of association between JIA and

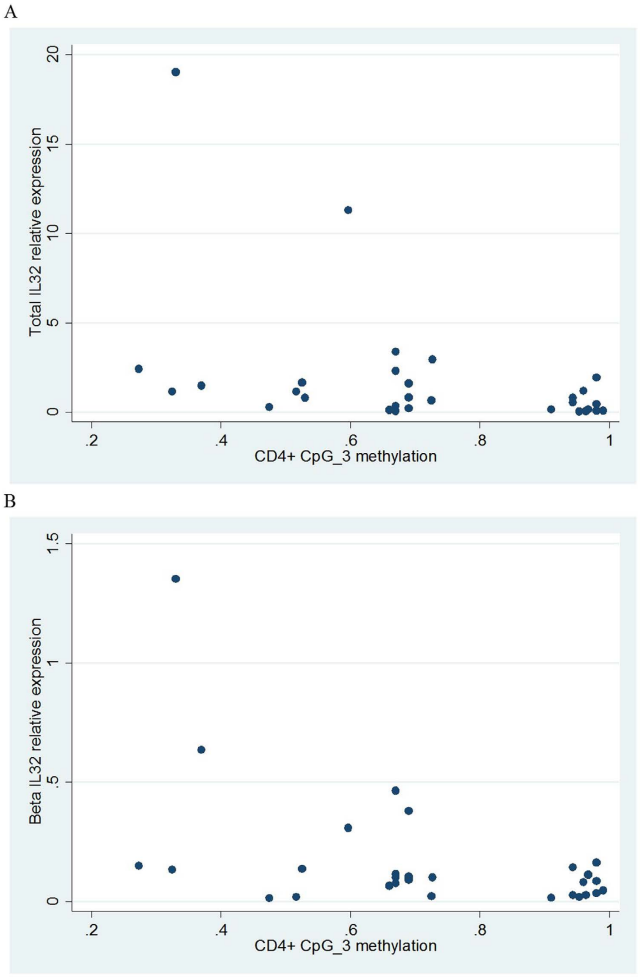

Figure 5. Correlation between *IL32* relative expression and methylation in CD4+ T cells (cases and controls combined). **A**: Total *IL32* and CpG_3, spearman r = −0.73, p = 0.0009. **B**: *IL32*β and CpG_3, spearman r = −0.48, p = 0.049.

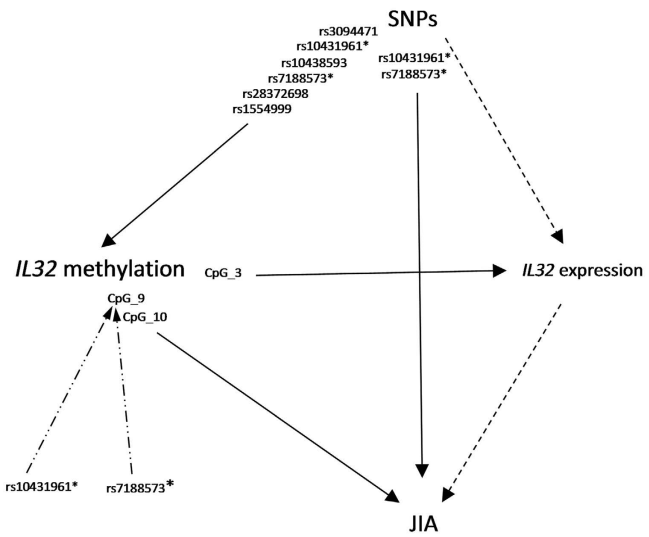

Figure 6. Diagram of the relationships between each of the genomic measures and JIA. Solid lines indicate a correlation/association. Dashed lines indicate no correlation/association. Dot-dash lines indicate interaction such that the minor allele of the SNP potentiates the association between *IL32* methylation and JIA. Asterisks indicate SNPs playing multiple roles in the causal pathway.

CD4+ T cell CpG_9 methylation remained. Therefore, genotype at these SNPs does not entirely account for the lower methylation-higher disease risk relationship. Further, we found evidence that these SNPs interact with CpG_9 methylation in determining disease risk, with reduced methylation increasing risk of JIA only in those carrying at least one minor allele. CpG_9 methylation was not associated with JIA in major allele homozygotes. This pattern of association would not be expected if methylation acted only as an intermediary between the SNP and disease risk.

Regulation of gene expression is considered to be one of the major functions of DNA methylation[40]. In general, reduced methylation, particularly in CpG rich gene promoters, is associated with elevated gene expression, although recent evidence suggests that the opposite may be true for some CpG rich sequences, especially those within gene bodies[40]. We therefore assessed the relationship between expression and methylation of *IL32* in CD4+ T cells. We found evidence of an inverse correlation between methylation at CpG_3 and expression of total IL32 and the β *IL32* isoform. However, neither CpG_3 methylation nor *IL32* expression were observed to be associated with JIA.

There are a few potential explanations for the lack of correlation we observed between the JIA-associated differential *IL32* methylation and gene expression. Firstly, other epigenetic mechanisms such as histone modification and chromatin remodelling may be important in regulating gene expression at this locus, obscuring correlation between methylation and expression[41]. Secondly, methylation at CpG sites for which methylation is correlated with expression can lie many kilobases from the relevant gene. It is therefore possible that methylation at the CpGs found to be associated with JIA are more relevant to expression of nearby genes. Genes from both the TNF receptor and matrix metalloproteinase superfamilies lie within 45 kb of CpG_9, and both gene families have been linked to inflammatory diseases[42,43]. Thirdly, If *IL32* expression is more dynamic and sensitive to environmental influence than methylation, the impact of sample processing procedures may have served to obscure the methylation – gene expression relationship.

Strengths of our study include careful attention to clinical phenotype to maximise case homogeneity, exclusion of cases treated with methotrexate (a folate inhibitor) and other disease-modifying anti-rheumatic drugs, and consideration of multiple genomic factors and their interactions. A limitation is the relatively small sample size, especially for association of genetic variants with outcomes. Additionally, treatment of cases with corticosteroids was not an exclusion criteria for this study. Although corticosteroids occur naturally in the body, it is possible that corticosteroid treatment (often injected directly to joints) may alter DNA methylation.

In summary, data presented here, in addition to data from our previous work[20], provides strong evidence that DNA methylation of CpGs in the 5' region of *IL32* is reduced in oligoarticular JIA. The association was observed in two different immune cell subsets suggestive of a role in determining disease risk. Methylation in this region is impacted by genotype, but the methylation-disease association persists after the effect of genotype is taken into account. In addition, genotype and methylation were seen to interact, with the effect of methylation on JIA risk only evident in those carrying certain alleles. A lack of strong correlation between disease-associated *IL32* methylation and *IL32* expression suggests that other unmeasured factors may need to be considered to elucidate the relationship. Given that *IL32* has an established role in promoting inflammatory arthritis in mouse models and adult rheumatoid arthritis, further work to understand the role of *IL32* in JIA is well justified.

Materials and methods

Participant Recruitment and Selection. Cases and controls for the current study were drawn from the ChiLdhood Arthritis Risk factor Identification sTudY (CLARITY). A detailed description of the study has been published previously[44,20]. Briefly, cases were recruited from the Royal Children's Hospital (RCH), Melbourne Australia. Cases were aged 18 years or under at recruitment, and were diagnosed with JIA and classified into clinical subtypes by a paediatric rheumatologist following International League of Associations for Rheumatology (ILAR) criteria[45]. Controls were healthy children aged 16 years or under attending the RCH day surgery unit for a minor surgical procedure. Peripheral blood mononuclear cells (PBMCs) were isolated using a ficoll procedure within 24 hours of blood collection, followed by storage in vapour-phase liquid nitrogen[20]. All study protocols were approved by the RCH Human Research Ethics Committee. Informed consent was obtained from all participants, and the research was carried out in accordance with the approved protocols.

Twenty-one CLARITY JIA cases (17 oligoarticular and four polyarticular) and 21 age and sex matched controls, for whom genome-scale and/or *IL32* locus specific DNA methylation data was previously generated[20], were included in the current study ('original' sample). The sample was augmented by the addition of twelve oligoarticular JIA cases (mean age 4.5 years, SD 3.5 years; 83.33% female) and twelve age and sex matched controls ('replication' sample). An additional unpaired case and two unpaired controls, for whom RNA was available, were included for the gene expression analyses (RNA-only sample). All selected cases were naive to methotrexate (MTX) and other biological disease modifying anti-rheumatic drugs at the time of blood collection. The characteristics of all original, replication, and RNA-only cases and controls are shown in Supplementary Table 1.

T Cell DNA/RNA Isolation. Total viable CD3+ CD4+ T cells, and CD3+ CD8+ T cells (replication samples only) were positively selected for from the PBMC population using flow cytometry (DAPI: Cat

ID D9542, Sigma Aldrich, St Louis, MO, USA; CD8-FITC: Cat ID 55536, CD3-APC Cat ID 340440, CD4-PE Cat ID 347327, Beckton Dickinson, San Jose, CA, USA). T cell purities were typically above 98% for CD3+CD4+ T cells and above 95% for CD3+CD8+ following cell sorting. DNA was extracted using the Flexigene DNA extraction kit (Qiagen). Where CD4+ T cell numbers were sufficient, an aliquot of cells was used for RNA extraction using a standard Trizol extraction adapted from Chomczynski, *et al.*[46], adjusting reagent volume to suit cell numbers. All RNA samples were DNase treated using the Ambion DNA Free kit (Life Technologies, Austin, TX USA).

Sequenom MassARRAY Epityper IL32 DNA Methylation analysis. DNA was bisulphite converted using the MethylEasy Xceed kit (Human Genetic Signatures, Randwick, NSW Australia) according to the manufacturer's instructions. Measurement of methylation at *IL32* was performed using two Sequenom MassARRAY Epityper assays as previously described[20]. Figure 1 shows the locations of CpGs in the 5' region of *IL32,* indicating those measured by the assays.

Single nucleotide polymorphism (SNP) selection and Sequenom MassARRAY iPlex genotyping. SNPs in the region of *IL32* were selected according to three criteria. First, we used CEU HapMap data[47] and the tagger function in Haploview[48] to identify tag-SNPs (defined as a proxy SNP representing a group of one or more SNPs amongst which $r^2 \geq 0.8$) in the region 3kb upstream to 10 kb downstream of the gene. A list of 7 SNPs representing variation at a total of 13 SNPs across the region was generated. Second, we identified any SNPs in the literature that had previously been examined for association with disease outcomes. Third, we used the Genevar eQTL database[49] to identify SNPs correlated with *IL32* gene expression in either T cells, lymphoblastoid cells or fibroblasts from the Gencord study[50]. We used the Sequenom Assay Design Tool to design a single iPlex multiplex assay that incorporated as many of the identified SNPs as possible. The resulting iPlex assay included a total of 17 SNPs. Rationale for the inclusion of each SNP in the final assay is provided in Supplementary Table 2. SNPs in close proximity to, or within, *IL32* are shown in Fig. 1. Correlations of SNPs with *IL32* expression in the Genevar eQTL database are depicted in Supplementary Figure 1. SNPs also associated with peripheral blood *IL32* expression in the Blood eQTL database[39] are shown in Supplementary Table 3. SNPs were genotyped using the iPlex chemistry on the Sequenom MassARRAY according to manufacturer's protocols. Primer sequences are provided in Supplementary Table 4. The Sequenom Typer program was used to call genotypes from raw data in a semi-automated fashion. Outliers on SNP cluster plots were visually inspected and calls rejected if unclear. DNA samples for which genotyping call rate was less than 90% were re-genotyped and subsequently discarded from analysis if call rate remained below 90%.

Quantitative Real Time PCR Expression Analysis. RNA samples were reverse transcribed to cDNA using the Tetro cDNA synthesis kit (Bioline, London, UK) as per the manufacturer's instructions. Quantitative real time PCR was then used to quantify expression levels of total *IL32* (all transcript variants), *IL32*γ and *IL32*β. Primers used to measure *IL32*γ and *IL32*β expression were as previously described[23]. Primers used to measure total *IL32* expression were as follows: sense 5'-GATGGATTACGGTGCCGAG-3'; antisense 5'-CACAAAAGCTCTCCCCAGG-3'. *IL32* expression was measured on the Applied Biosystems 7300 Real Time PCR System. Each sample was run in triplicate and a cut-off Cycle Threshold (CT) value of ± 1 from the median value of all 3 data points was used to remove outlying replicate values. Replicates that passed this quality control were then used to calculate the mean CT value. The ΔCT analysis method was employed to calculate the relative difference in expression of *IL32* and its isoforms to the endogenous control genes *B2M* (sense 5'-ATCATGGAGGTTTGAAGATGCC-3'; antisense 5'- ACATGGAGACAGCACTCAAAGTAGA) and *RPLP0* as per Dheda *et al*[51].

Statistical Analysis. We used t-tests (mean comparison tests) and logistic regression to compare methylation β values between cases and controls. We used logistic regression to generate odds ratios (ORs) to gauge the degree of change in risk of JIA with changes in methylation levels. For this purpose we converted methylation β values to % methylation (β x 100) to facilitate OR interpretation. We calculated Pearson correlations amongst methylation data, and between methylation and SNP data. We used linear regression, adjusting for the potential covariates age and sex, to further assess the relationships between methylation and SNPs. We looked for evidence of interaction between SNPs and methylation in relation to JIA risk by use of a product term in logistic regression, adjusting for age and sex. SNP genotype groups were dichotomised for this purpose into major allele homozygotes, and heterozygotes plus minor allele homozygotes (dominant genetic model). For analyses incorporating expression data, we employed non-parametric analyses (Spearman correlation) alongside parametric analyses (linear regression adjusting for the potential covariates age, sex and blood sample time to processing). All analyses were undertaken using Stata v13[52].

References

1. Prakken, B., Albani, S. & Martini, A. Juvenile idiopathic arthritis. *Lancet* **377,** 2138–2149, doi:10.1016/S0140-6736(11)60244-4 (2011).
2. Ravelli, A. & Martini, A. Juvenile idiopathic arthritis. *Lancet* **369,** 767–778, doi:10.1016/S0140-6736(07)60363-8 (2007).

3. Prahalad, S. & Glass, D. N. A comprehensive review of the genetics of juvenile idiopathic arthritis. *Pediatric rheumatology online journal* **6,** 11, doi:10.1186/1546-0096-6-11 (2008).
4. Cobb, J. E., Hinks, A. & Thomson, W. The genetics of juvenile idiopathic arthritis: current understanding and future prospects. *Rheumatology* **53,** 592–599, doi:10.1093/rheumatology/ket314 (2014).
5. Hinks, A. *et al.* Dense genotyping of immune-related disease regions identifies 14 new susceptibility loci for juvenile idiopathic arthritis. *Nature genetics* **45,** 664–669, doi:10.1038/ng.2614 (2013).
6. Ellis, J. A., Munro, J. E. & Ponsonby, A. L. Possible environmental determinants of juvenile idiopathic arthritis. *Rheumatology* **49,** 411–425, doi:10.1093/rheumatology/kep383 (2010).
7. Foley, D. L. *et al.* Prospects for epigenetic epidemiology. *American journal of epidemiology* **169,** 389–400, doi:10.1093/aje/kwn380 (2009).
8. de la Rica, L. *et al.* Identification of novel markers in rheumatoid arthritis through integrated analysis of DNA methylation and microRNA expression. *Journal of autoimmunity* **41,** 6–16, doi:10.1016/j.jaut.2012.12.005 (2013).
9. Liu, Y. *et al.* Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis. *Nature biotechnology* **31,** 142–147, doi:10.1038/nbt.2487 (2013).
10. Nakano, K., Whitaker, J. W., Boyle, D. L., Wang, W. & Firestein, G. S. DNA methylome signature in rheumatoid arthritis. *Annals of the rheumatic diseases* **72,** 110-117, doi:10.1136/annrheumdis-2012-201526 (2013).
11. Whitaker, J. W. *et al.* An imprinted rheumatoid arthritis methylome signature reflects pathogenic phenotype. *Genome medicine* **5,** 40, doi:10.1186/gm444 (2013).
12. Absher, D. M. *et al.* Genome-wide DNA methylation analysis of systemic lupus erythematosus reveals persistent hypomethylation of interferon genes and compositional changes to CD4+ T-cell populations. *PLoS genetics* **9,** e1003678, doi:10.1371/journal.pgen.1003678 (2013).
13. Coit, P. *et al.* Genome-wide DNA methylation study suggests epigenetic accessibility andtranscriptional poising of interferon-regulated genes in naive CD4+ T cellsfrom lupus patients. *Journal of autoimmunity* **43,** 78–84, doi:10.1016/j.jaut.2013.04.003 (2013).
14. Lin, S. Y. *et al.* A whole genome methylation analysis of systemic lupus erythematosus: hypomethylation of the IL10 and IL1R2 promoters is associated with disease activity. *Genes and immunity* **13,** 214–220, doi:10.1038/gene.2011.74 (2012).
15. Cooke, J. *et al.* Mucosal genome-wide methylation changes in inflammatory bowel disease. *Inflammatory bowel diseases* **18,** 2128–2137, doi:10.1002/ibd.22942 (2012).
16. Hasler, R. *et al.* A functional methylome map of ulcerative colitis. *Genome research* **22,** 2130–2137, doi:10.1101/gr.138347.112 (2012).
17. Nimmo, E. R. *et al.* Genome-wide methylation profiling in Crohn's disease identifies altered epigenetic regulation of key host defense mechanisms including the Th17 pathway. *Inflammatory bowel diseases* **18,** 889-899, doi:10.1002/ibd.21912 (2012).
18. Bell, C. G. *et al.* Genome-wide DNA methylation analysis for diabetic nephropathy in type 1 diabetes mellitus. *BMC medical genomics* **3,** 33, doi:10.1186/1755-8794-3-33 (2010).
19. Rakyan, V. K. *et al.* Identification of type 1 diabetes-associated DNA methylation variable positions that precede disease diagnosis. *PLoS genetics* **7,** e1002300, doi:10.1371/journal.pgen.1002300 (2011).
20. Ellis, J. A. *et al.* Genome-scale case-control analysis of CD4+ T-cell DNA methylation in juvenile idiopathic arthritis reveals potential targets involved in disease. *Clinical epigenetics* **4,** 20, doi:10.1186/1868-7083-4-20 (2012).
21. Kim, S. H., Han, S. Y., Azam, T., Yoon, D. Y. & Dinarello, C. A. Interleukin-32: a cytokine and inducer of TNFalpha. *Immunity* **22,** 131–142, doi:10.1016/j.immuni.2004.12.003 (2005).
22. Conti, P., Youinou, P. & Theoharides, T. C. Modulation of autoimmunity by the latest interleukins (with special emphasis on IL-32). *Autoimmunity reviews* **6,** 131–137, doi:10.1016/j.autrev.2006.08.015 (2007).
23. Heinhuis, B. *et al.* Tumour necrosis factor alpha-driven IL-32 expression in rheumatoid arthritis synovial tissue amplifies an inflammatory cascade. *Annals of the rheumatic diseases* **70,** 660–667, doi:10.1136/ard.2010.139196 (2011).
24. Hashkes, P. J., Uziel, Y. & Laxer, R. M. The safety profile of biologic therapies for juvenile idiopathic arthritis. *Nature reviews. Rheumatology* **6,** 561–571, doi:10.1038/nrrheum.2010.142 (2010).
25. Heinhuis, B., Netea, M. G., van den Berg, W. B., Dinarello, C. A. & Joosten, L. A. Interleukin-32: a predominantly intracellular proinflammatory mediator that controls cell activation and cell death. *Cytokine* **60,** 321–327, doi:10.1016/j.cyto.2012.07.010 (2012).
26. Joosten, L. A., Heinhuis, B., Netea, M. G. & Dinarello, C. A. Novel insights into the biology of interleukin-32. *Cellular and molecular life sciences : CMLS* **70,** 3883–3892, doi:10.1007/s00018-013-1301-9 (2013).
27. Xu, W. D. *et al.* IL-32 with potential insights into rheumatoid arthritis. *Clinical immunology* **147,** 89–94, doi:10.1016/j.clim.2013.02.021 (2013).
28. Shoda, H. *et al.* Interactions between IL-32 and tumor necrosis factor alpha contribute to the exacerbation of immune-inflammatory diseases. *Arthritis research & therapy* **8,** R166, doi:10.1186/ar2074 (2006).
29. Park, Y. E. *et al.* IL-32 aggravates synovial inflammation and bone destruction and increases synovial natural killer cells in experimental arthritis models. *Rheumatology international* **33,** 671–679, doi:10.1007/s00296-012-2385-5 (2013).
30. Joosten, L. A. *et al.* IL-32, a proinflammatory cytokine in rheumatoid arthritis. *Proceedings of the National Academy of Sciences of the United States of America* **103,** 3298–3303, doi:10.1073/pnas.0511233103 (2006).
31. Goda, C. *et al.* Involvement of IL-32 in activation-induced cell death in T cells. *International immunology* **18,** 233–240, doi:10.1093/intimm/dxh339 (2006).
32. Choi, J. D. *et al.* Identification of the most active interleukin-32 isoform. *Immunology* **126,** 535–542, doi:10.1111/j.1365-2567.2008.02917.x (2009).
33. Heinhuis, B. *et al.* Inflammation-dependent secretion and splicing of IL-32{gamma} in rheumatoid arthritis. *Proceedings of the National Academy of Sciences of the United States of America* **108,** 4962–4967, doi:10.1073/pnas.1016005108 (2011).
34. Tsitsiou, E. *et al.* Transcriptome analysis shows activation of circulating CD8+ T cells in patients with severe asthma. *The Journal of allergy and clinical immunology* **129,** 95–103, doi:10.1016/j.jaci.2011.08.011 (2012).
35. Xu, X. *et al.* Maturation and emigration of single-positive thymocytes. *Clinical & developmental immunology* **2013,** 282870, doi:10.1155/2013/282870 (2013).
36. Bell, J. T. *et al.* DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines. *Genome biology* **12,** R10, doi:10.1186/gb-2011-12-1-r10 (2011).
37. Teh, A. L. *et al.* The effect of genotype and in utero environment on interindividual variation in neonate DNA methylomes. *Genome research* **24,** 1064–1074, doi:10.1101/gr.171439.113 (2014).
38. McRae, A. F. *et al.* Contribution of genetic variation to transgenerational inheritance of DNA methylation. *Genome biology* **15,** R73, doi:10.1186/gb-2014-15-5-r73 (2014).
39. Westra, H. J. *et al.* Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nature genetics* **45,** 1238–1243, doi:10.1038/ng.2756 (2013).
40. Jones, P. A. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nature reviews. Genetics* **13,** 484–492, doi:10.1038/nrg3230 (2012).

41. Wagner, J. R. *et al.* The relationship between DNA methylation, genetic and expression inter-individual variation in untransformed human fibroblasts. *Genome biology* **15,** R37, doi:10.1186/gb-2014-15-2-r37 (2014).
42. Croft, M. *et al.* TNF superfamily in inflammatory disease: translating basic insights. *Trends in immunology* **33,** 144–152, doi:10.1016/j.it.2011.10.004 (2012).
43. Nissinen, L. & Kahari, V. M. Matrix metalloproteinases in inflammation. *Biochimica et biophysica acta* **1840,** 2571–2580, doi:10.1016/j.bbagen.2014.03.007 (2014).
44. Ellis, J. A. *et al.* CLARITY - ChiLdhood Arthritis Risk factor Identification sTudY. *Pediatric rheumatology online journal* **10,** 37, doi:10.1186/1546-0096-10-37 (2012).
45. Petty, R. E. *et al.* International League of Associations for Rheumatology classification of juvenile idiopathic arthritis: second revision, Edmonton, 2001. *The Journal of rheumatology* **31,** 390–392 (2004).
46. Chomczynski, P. & Sacchi, N. Single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction. *Anal. Biochem.* **162,** 156–159 (1987).
47. International HapMap, C. The International HapMap Project. *Nature* **426,** 789–796, doi:10.1038/nature02168 (2003).
48. Barrett, J. C., Fry, B., Maller, J. & Daly, M. J. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* **21,** 263–265, doi:10.1093/bioinformatics/bth457 (2005).
49. Yang, T. P. *et al.* Genevar: a database and Java application for the analysis and visualization of SNP-gene associations in eQTL studies. *Bioinformatics* **26,** 2474–2476, doi:10.1093/bioinformatics/btq452 (2010).
50. Dimas, A. S. *et al.* Common regulatory variation impacts gene expression in a cell type-dependent manner. *Science* **325,** 1246–1250, doi:10.1126/science.1174148 (2009).
51. Dheda, K. *et al.* Validation of housekeeping genes for normalizing RNA expression in real-time PCR. *BioTechniques* **37,** 112–119 (2004).
52. Stata Statistical Software: Release 13 (StataCorp LP, College Station, TX, 2013).

Acknowledgements

We thank the families who kindly participated in this study. We thank the research nurses and research assistants who undertook recruitment and data entry. We acknowledge support from the Australian National Health and Medical Research Council (NHMRC), the Australian Research Council (ARC), Arthritis Australia, Rebecca L Cooper Foundation, LEW Carty Charitable Fund, ANZ Medical Research and Technologies in Victoria Fund, Lynne Quayle Charitable Trust, The Victorian State Government Operational Infrastructure Support Program, and the Murdoch Childrens Research Institute. JAE was supported by an ARC Future Fellowship. RAC was supported by an Arthritis Australia Postgraduate Scholarship. RS and ALP were supported by NHMRC Senior Research Fellowships.

Author Contributions

JE, RS, JM, JC, ALP contributed to study concept and design. JE, JM, ALP, JA, RA contributed to participant acquisition. BM, RC performed laboratory experiments. BM, JE, RS, RC, RCC, ALP contributed to data analysis and interpretation. All authors contributed to drafting the manuscript to the final version.

Additional Information

Supplementary information accompanies this paper at http://www.nature.com/srep

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Meyer, B. *et al.* DNA methylation at *IL32* in juvenile idiopathic arthritis. *Sci. Rep.* **5**, 11063; doi: 10.1038/srep11063 (2015).

This work is licensed under a Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/

Appendix II: Manuscript

In vitro exposure of human blood mononuclear cells to active vitamin D does not induce substantial change to DNA methylation on a genome-scale

Journal of Steroid Biochemistry & Molecular Biology 141 (2014) 144–149

Contents lists available at ScienceDirect

Journal of Steroid Biochemistry and Molecular Biology

journal homepage: www.elsevier.com/locate/jsbmb

In vitro exposure of human blood mononuclear cells to active vitamin D does not induce substantial change to DNA methylation on a genome-scale

Raul A. Chavez Valencia[a,b,1], David J. Martino[b,c,1], Richard Saffery[b,c], Justine A. Ellis[a,b,*]

[a] Genes, Environment and Complex Disease, Murdoch Childrens Research Institute, 50 Flemington Road, Parkville, Victoria 3052, Australia
[b] Department of Paediatrics, The University of Melbourne, Victoria 3010, Australia
[c] Cancer and Disease Epigenetics, Murdoch Childrens Research Institute, 50 Flemington Road, Parkville, Victoria 3052, Australia

ARTICLE INFO

Article history:
Received 22 September 2013
Received in revised form
17 December 2013
Accepted 31 January 2014
Available online 10 February 2014

Keywords:
Vitamin D
Mononuclear cells
1,25-Dihydroxyvitamin D_3
DNA methylation
Epigenetics
Infinium

ABSTRACT

It is well-established that vitamin D impacts gene regulation via vitamin D response elements (VDREs) across the genome. Recent evidence, primarily at a locus-specific level, suggests that alterations to DNA methylation may also be a relevant mechanism through which vitamin D regulates gene expression. Given the intense interest in vitamin D, particularly as an immune modifier, we sought to examine the impact of vitamin D exposure on the immune cell methylome in vitro. We exposed primary human blood mononuclear cells with up to 100 nM calcitriol for up to 120 h, and measured genome-scale DNA methylation response using the Illumina Infinium HumanMethylation450 beadchip array. We observed that, while the expression of known vitamin D responsive genes was clearly altered by calcitriol exposure, substantial genome-scale changes to DNA methylation were not induced. Our data suggests that, over the exposure period measured, changes to DNA methylation may not be a predominant mechanism through which vitamin D impacts gene expression in human immune cells.

© 2014 Elsevier Ltd. All rights reserved.

1. Introduction

Vitamin D is a pleiotropic secosteroid hormone with important roles in regulating calcium homeostasis, cell proliferation, cell cycle control and immune function [1]. In addition to the well documented link between vitamin D deficiency and metabolic bone disorders, convincing data are emerging linking vitamin D deficiency to several complex disorders including cardiovascular disease, cancer, and immune disorders [1–3]. There has been much interest in determining whether vitamin D is linked to autoimmune disease risk, with considerable epidemiological and genetic data now emerging to suggest an association of suboptimal vitamin D with multiple sclerosis [4], rheumatoid arthritis [5] and type 1 diabetes [6]. Vitamin D is known to have multiple immuno-modulatory functions, including inhibition of TLR2 and TLR4 expression in monocytes, transcriptional repression of T helper cell type 1 (Th1) cytokines TNF-α, IL2 and IFN-γ, and suppression of IL17 production in CD4+ T cells (reviewed in [3,7]). These responses are considered to suppress autoimmunity by attenuating Th1 and Th17 inflammatory responses, thereby highlighting an apparent link between vitamin D, inflammation and autoimmunity [7].

Beyond the effects on immune-related gene regulation, vitamin D is also known to regulate gene expression on a genome-scale. For example, stimulation of lymphoblasts with the hormonally active form of vitamin D (1,25-dihydroxyvitamin D_3, also referred to as calcitriol) changes expression levels of over 200 genes [8]. Population-based cohort studies have demonstrated differing gene expression profiles in individuals with differing plasma 25-hydroxyvitamin D concentrations [9]. The prototypical mechanism through which vitamin D impacts expression is through binding of the vitamin D receptor (VDR) to widespread vitamin D response elements (VDRE), that facilitate transcriptional regulation via

Abbreviations: CDH1, cadherin-1; CpG, CG dinucleotides; CYP24A1, cytochrome P450, family 24, subfamily A, polypeptide 1; CYP27B1, cytochrome P450, family 27, subfamily B, polypeptide 1; DIO3, deiodinase, iodothyronine, Type III; DMSO, dimethyl sulfoxide; IL2, interleukin 2; IL17, interleukin 17; IFN-γ, interferon gamma; MAPRE2, microtubule-associated protein RP/EB family member 2; PBMC, peripheral blood mononuclear cell; PTEN, phosphatase and tensin homolog; RPMI, Roswell Park Memorial Institute; TLR, toll-like receptor; TNF-α, tumor necrosis factor-alpha.

* Corresponding author at: Genes, Environment and Complex Disease, Murdoch Childrens Research Institute, 50 Flemington Road, Parkville, Victoria 3052, Australia. Tel.: +61 3 8341 6311.
E-mail addresses: raul.chavez@mcri.edu.au (R.A. Chavez Valencia), david.martino@mcri.edu.au (D.J. Martino), richard.saffery@mcri.edu.au (R. Saffery), justine.ellis@mcri.edu.au (J.A. Ellis).
[1] These authors contributed equally to this work.

http://dx.doi.org/10.1016/j.jsbmb.2014.01.018
0960-0760/© 2014 Elsevier Ltd. All rights reserved.

histone modifications, chromatin remodeling and changes to RNA polymerase II binding [7,8]. Yet, only 26% of genes identified as differentially regulated by vitamin D are co-located with a VDRE in close proximity of the transcription start site [8]. This suggests that the full spectrum of mechanisms involved in the transcriptional control of vitamin D-responsive genes is yet to be elucidated.

Recent evidence implicates altered DNA methylation in mediating at least some of the effects of vitamin D on gene expression. Stefanska et al. reported *substantial* demethylation of 12–40% at the *PTEN* promoter along with a concurrent increase in gene expression in breast cancer cell lines cultured with (1.5 or 25) micromole concentrations of vitamin D_3 for 72 h [10]. Lopes et al. also demonstrated demethylation of CpG sites at the *CDH1* promoter along with increased gene expression in the same breast cancer cell line using far lower (100 nM) concentrations of calcitriol over a similar exposure period [11]. Rawson et al. demonstrated an association between vitamin D intake (measured by food frequency questionnaire) and reduced methylation at the promoter of *DKK1* [12]. Zhu et al. examined genome-scale methylation in leukocytes from vitamin D (plasma 25(OH)D) deficient adolescents. Compared to adolescents with sufficient circulating plasma 25(OH)D, several loci, including at *MAPRE2* and *DIO3*, were found to be differentially methylated by around 4–5% [13]. Additionally, the authors reported a trend toward differential methylation at three loci previously associated with circulating 25 (OH)D by genome-wide association analysis (*DHCR7*, *CYP2R1*, *CYP24A1*).

Given the intense interest in the role of vitamin D as a modifier of immune function, and the mounting evidence for a role of epigenetic modifications (including DNA methylation) in these processes, we sought for the first time to determine the extent to which genome-scale DNA methylation is altered in primary immune cells in response to prolonged calcitriol exposure.

2. Materials and methods

2.1. Mononuclear cell isolation

Samples used in this study (2 males and 2 females ages 26–33) were drawn from a collection of de-identified healthy adult blood samples gathered for pilot studies with informed consent. Experimental protocols were approved by the Human Research Ethics Committee of the Royal Children's Hospital, Melbourne Australia. Blood was collected in 2×9 mL K_3EDTA buffered vacutainers (Greiner Bio-one, Germany). Within 2 h of blood withdrawal, blood was centrifuged at $500 \times g$ for 10 min, and plasma removed. Blood was then diluted 1:1.5 with RPMI and layered over 12.5 mL of Ficoll Paque Plus (GE Healthcare Life Sciences, UK) to isolate peripheral blood mononuclear cells (PBMC). After centrifugation for 10 min at $350 \times g$ at ambient temperature, the PBMC cell layer was removed and diluted with 10 mL RPMI. Cells were then centrifuged at $350 \times g$ for 10 min and resuspended in 10 mL RPMI and the process repeated. The final cell pellet was resuspended in 1 mL of fetal bovine serum (FBS), and a cell count was performed. Cells were frozen viably at −80 °C in a freeze mix of FBS with 10% DMSO at 5×10^6 cells per mL, then transferred to liquid nitrogen.

2.2. Cell culture

Cryopreserved stocks of PBMCs were thawed in 37 °C water bath, and RPMI media added slowly to the cells over 2–3 min, mixing gently. Cells were pelleted at $250 \times g$ for 10 min. The cell pellets were resuspended to a concentration of 1×10^6 cells per mL and seeded at 2.5×10^5 cells per well in a 96-well round bottom polystyrene plate. Calcitriol (Sigma–Aldrich, Australia) was added to culture media (RPMI 1640 with 10% FBS, 10 mM L-glutamine, 10 μg/mL penicillin/streptomycin, 1 mM pyruvate) at a final concentration of 0, 0.1, 1, 10 or 100 nM in quadruplicate. This was replenished daily. Cells were harvested at 0, 24, 96 and 120 h. At each time point, three replicate wells were combined into a single tube for DNA and RNA extraction, and cell death was measured on the fourth replicate using the Annexin V apoptosis assay (see Section 2.4). A total of 16 wells were seeded per individual, per condition.

2.3. Extraction of nucleic acid

Triplicate wells were combined in a 1.5 mL Eppendorf tube and centrifuged at $450 \times g$ for 5 min. The cell supernatant was removed and cell pellets were processed for DNA and RNA extraction using Qiagen All Prep kits as per the manufacturer's instruction (Qiagen, Australia).

2.4. Annexin V assay

At each time point, cells from one replicate well were harvested and prepared for measurement of apoptosis using the Annexin V-fluorescein isothiocyanate (Annexin V-FITC) and propidium iodide (PI) Apoptosis/Necrosis Detection Kit (Abcam, UK). Briefly, cells were transferred to 10 mL centrifuge tubes and topped up with 5 mL of PBS and centrifuged for 5 min at $350 \times g$. The supernatant was discarded and the cell pellet was resuspended in 500 μL of binding buffer, then 2.5 μL of PI and Annexin V stain mixture was added and incubated in the dark for 5 min. Samples were then analyzed by flow cytometry (LSRII flow cytometer system, Becton-Dickinson, USA). Late and early apoptotic cells were identified using FCS Express4 software (De Novo Software, USA) following double positive staining (PI + Annexin V) and single positive staining (Annexin V only) respectively. Untreated cells were used as negative controls.

2.5. DNA pooling and bisulphite conversion

Equal amounts of genomic DNA from four donors were pooled prior to bisulphite conversion and duplicate pools (500 ng total DNA each) were constructed for each time point and condition which was bisulphite converted using MethylEasy Xceed kit according to manufacturer's instructions (Human Genetic Signatures, Australia). The conversion efficiency was assessed by bisulphite-specific PCR on all samples.

2.6. Illumina Infinium Human Methylation450 array analysis

Bisulphite DNA hybridization to Illumina Human Methlylation450 (HM450) arrays was carried out at the Australian Genome Research Facility. The HM450 platform interrogates the methylation status of >480,000 CpG dinucleotides at >95% of Refseq genes. Raw iDAT files were processed using the Minfi package [14] from the bioconductor project [15] in the R statistical environment. The Minfi package was used for array pre-processing using the Illumina method with normalization to internal control probes. The pre-processed data underwent subset quantile within-array normalization [16] to remove technical bias attributable to the different probe chemistries (Type I and Type II probes). Twenty-three of twenty-four samples produced methylation values of high confidence (average *P*-value detection of <0.05). One sample had an average *P*-value detection higher than 0.05 and was removed from the data set. Any poor-performing probes with a *P*-value detection call >0.01 for 1 or more samples were removed from the entire data set leaving a total of 454,563 common high confidence methylation values. Probes on the X and Y chromosomes were removed prior to analysis to eliminate gender bias. Probes previously described to cross-hybridize in the genome were removed [17]. Beta values

were derived from intensities as defined by the ratio of methylated probe intensity and the overall probe intensity [18]. Sample outlier detection was performed using multidimensional scaling analysis on normalized beta values. Sample correlations were calculated using Spearman's correlation coefficient. These correlation coefficients were visualized using heatmaps. To rank probes based on association with calcitriol dose, we employed a fixed effect linear model to regress all probes on calcitriol dose in the presence of covariates including time in culture and array batch.

2.7. Quantitative gene expression analysis

Total RNA was quantified by UV spectroscopy on the Nanodrop 1000 (Thermo Fisher, USA). RNA was converted into cDNA using the Qiagen Omniscript RT kit, according to manufacturer's instructions, and diluted 1/10 with RNase free water for use in gene expression assays. Amplification primers for *CYP24A1*, *DNMT1* and *DNMT3A* were designed to overlap splice sites. The following primers were used: 5′-CCCAAAGGAACAGTGCTCATG-3′ (forward) and 5′-TCTCCTGAAGCCAACGTTCAG-3′ (reverse) for *CYP24A1* [19]. 5′-CTTCTTCAGCACAACCGTCA-3′ (forward) and 5′-GAAGAGCCGGTAGGTGTCAG-3′ (reverse) for *DNMT1* [20] and 5′-GGGGACGTCCGCAGCGTCACAC-3′ (forward) and 5′-CAGGGTTGGACTCGAGAAATCGC-3′ (reverse) for *DNMT3A*. 5′-TGAAGAGAATCCACAAGGAATTGA-3′ (forward) and 5′-CAACAGGACCTGCTGAACACTG-3′ (reverse) for *UBC* (Sigma–Aldrich). A single reaction mixture containing 5 μL of SYBR green supermix (Qiagen), 1 μL each of 10 μM primer mix (*UBC* was used at 5 μM) and 2 μL of cDNA was run on the Roche Lightcycler 480 system. A total of 45 cycles were performed using the following conditions: 95 °C for 10 s, 60 °C for 30 s, 70 °C for 10 s, except for *UBC* which had conditions as follows: 95 °C for 10 s, 60 °C for 30 s, 70 °C for 5 s. Gene expression was quantitated relative to the housekeeping gene *UBC*.

2.8. Statistics

Gene expression data was analyzed using a two-way ANOVA repeated measures design with multiple testing corrections. Main effects for calcitriol dose, time in culture and interactions between these variables were analyzed using a significance test with Dunnett's multiple comparisons test. Statistics reported are significance at the adjusted alpha level 0.05 with 95% confidence interval of mean differences versus unstimulated sample. Analyses were performed in Prism 6.0 software.

3. Results

In this experiment, we cultured mononuclear cells from four healthy adult donors for up to 120 h with calcitriol over a four-log dose range. Media supplemented with calcitriol was replaced daily in order to maintain continual calcitriol stimulation over the required time course. At each time point, DNA and RNA were harvested from cells to assess biological activity and viability was assessed by flow cytometry as a marker of cellular toxicity. We found no evidence for calcitriol-induced cell death over the course of the experiment at the dose-range tested (Fig. 1A). To determine the effective dose of calcitriol at which clear regulation of transcription is apparent, we measured gene expression of *CYP24A1*, a well-characterized vitamin D responsive gene which is an excellent marker of biological activity [19]. We also measured gene expression of DNA methyltransferase enzymes *DNMT1* and *DNMT3A* as a proxy marker of epigenetic regulation. Analysis of gene expression regulation above un-stimulated levels showed a dose-dependent upregulation of *CYP24A1*, which was significant at 10 nM (adj. $P < 0.05$, CI 12.64–5.27) and 100 nM (adj. $P < 0.05$, CI 19.64–12.27) concentrations (Fig. 1B). A modest decline in relative gene expression of *DNMT1* (adj. $P < 0.05$, CI 0.3916–1.46) and *DNMT3A* (adj. $P < 0.05$, CI 0.1919–1.261) expression was also observed at the 10 nM and 100 nM concentrations (Fig. 1C and D). Based on these results we determined the 10 nM and 100 nM doses as effective regulators of the genome and these were used for further experiments.

Next, to determine whether prolonged exposure to calcitriol is associated with changes in DNA methylation, we performed genome-scale methylation analysis on pooled DNA using the HM450 BeadArray platform. We chose to study the 10 nM and 100 nM doses in the first instance as these were the most biologically active in the gene expression experiments. In order to determine whether DNA methylation profiles varied widely in response to treatment with calcitriol, we employed multidimensional scaling (MDS) as a robust means of visualizing the level of similarity of samples within the data set. We hypothesized that samples treated with calcitriol would cluster separately from untreated samples, based on variation in their methylation profiles attributable to the exposure. MDS analysis of sample relationships based on either genome-scale DNA methylation profiles or the top 5000 most variable HM450 probes did not reveal any evidence of an association between calcitriol exposure and changes to overall DNA methylation profile (Fig. 2A).

As we were interested in CpG sites that vary from background (untreated) levels in response to exposure to calcitriol, we performed a background adjustment by subtracting the unstimulated methylation values from each sample. This has the effect of adjusting all methylation values for baseline, retaining variation attributable to exposure or time in culture. A matrix of pairwise (Spearman) sample correlation coefficients was calculated for all background-adjusted samples and cluster analysis of the correlation coefficients was performed to test sample groupings. We only observed clustering by time in culture, with 120 h samples grouping together and exhibiting the highest correlations. No evidence for a dose-dependent effect on methylation status was evident from the cluster analysis (Fig. 2B).

To identify specific CpG sites associated with calcitriol exposure we regressed all probes on calcitriol dose using a fixed effect linear model to adjust for time in culture and chip number. Statistical significance was not reached for any loci tested after correction for multiple testing. We used the nominal *P*-values to rank probes on the array by evidence for an association with calcitriol dose and have provided the top 50 ranked probes in Supplementary Tables 1 and 2, for each of 10 nM and 100 nM doses. An examination of data points for all probes on the array annotated to genes described by others as either methylation-responsive on calcitriol or vitamin D_3 stimulation or nominally associated with circulating 25(OH)D deficient adolescents (*CYP24A1*, *CYP2R1*, *DHCR7*, *PTEN*, *CDH1*, etc.) revealed no differences between unstimulated and stimulated samples (Supplementary Figure 1) [10,11]. Collectively the data did not indicate substantial genome-scale changes in DNA methylation as mediating the biological effects of calcitriol exposure in human mononuclear cells over the timescale analysed.

4. Discussion

This is the first study to explore at the direct impact of calcitriol on DNA methylation on a genome scale. We found little evidence to suggest acute calcitriol exposure has a substantial or widespread effect on DNA methylation in peripheral blood mononuclear cells. This conclusion is based on two principal observations. Firstly, cluster analysis of beta methylation values, either genome-wide or restricted to the top 5000 most variable probes, did not

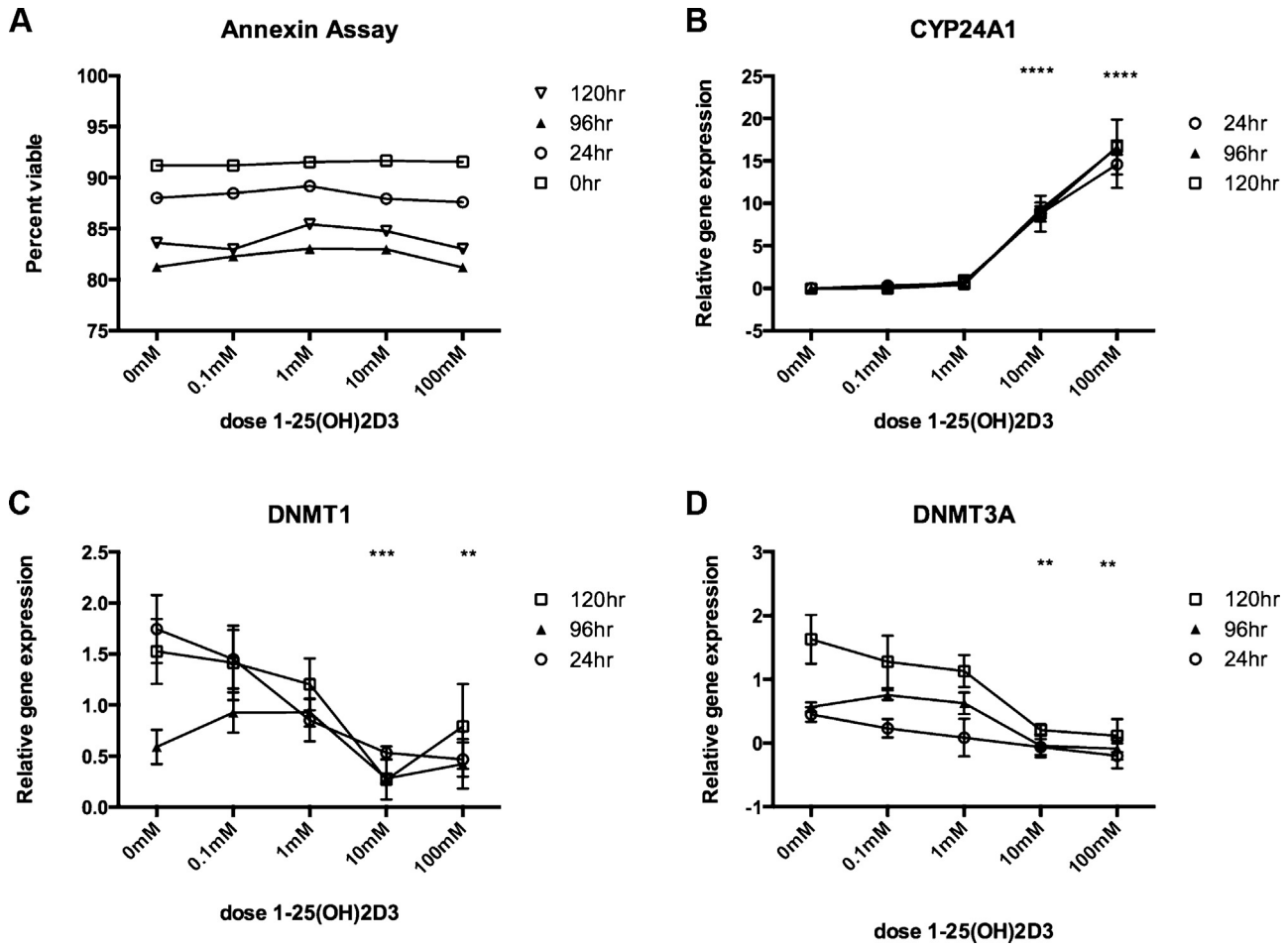

Fig. 1. Validation of experimental methodology. (A) Assessment of cell viability by flow cytometry. The graph shows the percentage of viable cells under each experimental condition derived from pooled data from 4 individuals. (B–D) Gene expression analysis of genes in the vitamin D and DNA methylation pathways by qRT-PCR. The graph shows mean gene expression with standard errors for the four replicates, normalized to the housekeeping gene. Statistical tests were performed by two-way ANOVA and significant changes at the alpha 0.05 level are denoted by **, at alpha 0.01 level by ***, and alpha 0.001 level by ****.

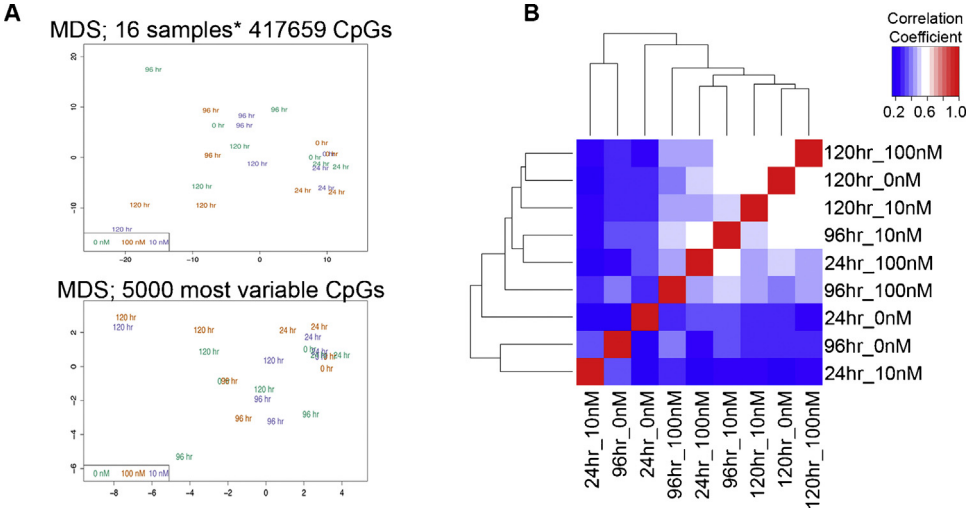

Fig. 2. Analysis of sample relationships based on whole genome methylation values reveals no correlation to dose. (A) Multidimensional scaling plots (MDS) of sample relationships based on whole-genome beta methylation values (16 samples × 417659 CpGs) (top panel) or the top 5000 most variable CpG sites (bottom panel). Samples are colourized by calcitriol dose and labeled by time in culture. (B) Cluster heatmap representation of sample correlation coefficients based on background-corrected genome-wide DNA methylation values. Cells are colorized according to strength of correlations.

suggest a dose-dependent relationship between samples. Secondly, following background correction for time zero methylation values and adjustment for time in culture, methylation at the top ranked CpG sites differed by less than 10% following 120 h of continual exposure, compared to unstimulated cells. This is within the range of normal measurement error expected with this platform [17]. Despite this, gene expression measurement for *CYP24A1* clearly indicates regulation of the vitamin D pathway in this experimental system. Interestingly, our data also suggests a dose-dependent down-regulation of DNA methyltransferase gene expression in response to calcitriol exposure. The biological significance of this is unknown, but it is consistent with a lack of an effect on genome-wide DNA methylation.

Genes identified as methylation-responsive on calcitriol stimulation by previous work, including *CDH1* [11] and *PTEN* [10], were not observed to be strikingly sensitive in our study, with changes of less than 5% observed. While Lopes et al. used the same concentrations as our study to identify expression and/or methylation differences at 72 h, Stefanska et al. used 15-fold greater concentrations of vitamin D_3. The inconsistencies between studies might be explained by differences in response to vitamin D of primary immune cells compared to cancer cell lines, and by compounds used.

Less substantial changes to the methylation profile of primary immune cells in response to calcitriol cannot be ruled out from our data, and may indeed be of relevance to immune function and thus immune disorders. Detection of *statistically* significant case–control methylation differences of the magnitude detected by our study (<10%) have been reported for a number of complex diseases, including autoimmune diseases [2,9,21–23]. However, the *biological* significance of such moderate change to DNA methylation has yet to be established. Zhu et al. associated differential methylation in leukocytes at *MAPRE2* and *DIO3* of less than 5% with circulating plasma 25(OH)D deficiency in adolescence [13]. Our data is essentially in accordance with that of Zhu et al., demonstrating methylation changes of small magnitude overall in response to calcitriol. Cumulatively, data generated so far on immune cells might suggest that if vitamin D is an environmental modulator of immune function and immune disorder risk, modulation via alterations to DNA methylation may be minimal. However, there may also be critical life stages yet to be examined, particularly during fetal development and early life when the immune system is developing. We (and others) have shown that the epigenome is potentially more susceptible to environmental perturbation during early development [24,25] and this may also be a time during which the effects of vitamin D on the immune system might be most relevant to later risk of disease [26,27]. Such life-course data linking the maturing DNA methylation landscape with vitamin D status remains to be generated.

Limitations of the current study are noteworthy. Firstly, given our dataset was generated from a mixed population of mononuclear cells, we cannot rule out a role for calcitriol in modulating changes in DNA methylation in specific leukocyte sub-populations, the effects of which may be masked in a mixed cell sample. Due to the potential confounding effects of variations in blood composition between individuals, we adopted a pooling strategy to circumvent this and enrich for relatively large and robust changes in methylation. This does not preclude a potentially significant number of small-scale changes in DNA methylation that might be detectable at an individual level. Further work should therefore focus on homogeneous cell populations. Lastly, our exposure and sampling methodology was both an advantage and disadvantage of this study. Our study extended to 120 h of calcitriol exposure, measuring methylation at 24, 96 and 120 h time-points. We analyzed the data collectively with the aim of detecting consistent changes across time, and to facilitate identification of sensitive CpGs. Our exposure times were generally longer than that of previous work, with sampling at various time-points. However, even longer term exposure may be required to elicit an immune cell methylome response to calcitriol.

Despite these limitations, our findings do suggest that the previously reported potent effects of vitamin D on gene expression in the immune cell compartment [7,8] occur largely independently of substantial genome scale changes to DNA methylation, and may instead be regulated by other epigenetic processes, such as a redistribution of histone modifications [7].

5. Conclusions

In summary, our study has provided no evidence to support substantial changes to DNA methylation in adult human mononuclear cells in response to calcitriol over the exposure period measured. Changes to DNA methylation may not be a predominant mechanism through which vitamin D impacts immune function. Further clarity may be provided by considering the response of specific immune cells to longer term calcitriol exposure across the life-course.

Acknowledgements

We thank the volunteers who made the kind donations of blood samples that facilitated this work, and Anna Czajko for assisting with blood processing. JAE is supported by an Australian Research Council Future Fellowship (#FT120100253). RS is supported by a National Health and Medical Research Council Senior Research Fellowship (#APP1045161). DJM is supported by a National Health and Medical Research Council Early Career Fellowship (APP#1072752). RACV is supported by an Arthritis Australia & State and Territory Affiliate grant (South Australia). This work was supported by the Victorian State Government Operational Infrastructure support program.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.jsbmb.2014.01.018.

References

[1] M.F. Holick, Vitamin D deficiency, N. Engl. J. Med. 357 (3) (2007) 266–281.
[2] S.S. Gunta, R.I. Thadhani, R.H. Mak, The effect of vitamin D status on risk factors for cardiovascular disease, Nat. Rev. Nephrol. 9 (6) (2013) 337–347.
[3] M. Hewison, Symposium 3: vitamin D and immune function: from pregnancy to adolescence vitamin D and immune function: an overview, Proc. Nutr. Soc. 71 (1) (2012) 50–61.
[4] K.C. Simon, K.L. Munger, A. Ascherio, Vitamin D and multiple sclerosis: epidemiology, immunology, and genetics, Curr. Opin. Neurol. 25 (3) (2012) 246–251.
[5] H. Wen, J.F. Baker, Vitamin D, immunoregulation, and rheumatoid arthritis, J. Clin. Rheumatol. 17 (2) (2011) 102–107.
[6] K. Badenhoop, H. Kahles, M. Penna-Martinez, Vitamin D: immune tolerance, and prevention of type 1 diabetes, Curr. Diab. Rep. 12 (6) (2012) 635–642.
[7] M. Di Rosa, et al., Vitamin D3: a helpful immuno-modulator, Immunology 134 (2) (2011) 123–139.
[8] S.V. Ramagopalan, A ChIP-seq defined genome wide map of vitamin D receptor binding: associations with disease and evolution, Genome Res. 20 (10) (2010) 1352–1360.
[9] K. Standahl Olsen, et al., Plasma 25 hydroxyvitamin D level and blood gene expression profiles: a cross-sectional study of the Norwegian Women and Cancer Post-genome Cohort, Eur. J. Clin. Nutr. 67 (7) (2013) 773–778.
[10] B. Stefanska, Comparative effects of retinoic acid: vitamin D and resveratrol alone and in combination with adenosine analogues on methylation and expression of phosphatase and tensin homologue tumour suppressor gene in breast cancer cells, Br. J. Nutr. 107 (6) (2012) 781–790.
[11] N. Lopes, 1Alpha,25-dihydroxyvitamin D3 induces de novo E-cadherin expression in triple negative breast cancer cells by CDH1-promoter demethylation, Anticancer Res. 32 (2012) 249–258.

[12] J.B. Rawson, et al., Vitamin D intake is negatively associated with promoter methylation of the Wnt antagonist gene DKK1 in a large group of colorectal cancer patients, Nutr. Cancer 64 (7) (2012) 919–928.
[13] H. Zhu, et al., A genome-wide methylation study of severe vitamin D deficiency in African American adolescents, J. Pediatr. 162 (5) (2013) 1004-9.e1.
[14] D. Kasper, H. Aryee, M. Aryee, Minfi: Analyse Illumina 450k methylation arrays. R package Version 1.2.0.
[15] R.C. Gentleman, et al., Bioconductor: open software development for computational biology and bioinformatics, Genome Biol. 5 (10) (2004) R80.
[16] J. Maksimovic, L. Gordon, A.S.W.A.N. Oshlack, Subset-quantile within array normalization for Illumina Infinium HumanMethylation450 BeadChips, Genome Biol. 13 (6) (2012) R44.
[17] Y.A. Chen, L.M. Choufani, S. Butcher, D.T. Grafodatskaya, D. Zanke, B.W. Gallinger, S. Hudson, T.J.R. Weksberg, Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray, Epigenetics 8 (2) (2013) 203–209.
[18] P. Du, et al., Comparison of beta-value and M-value methods for quantifying methylation levels by microarray analysis, BMC Bioinformatics 11 (2010) 587.
[19] B. Novakovic, et al., Placenta-specific methylation of the vitamin D 24-hydroxylase gene: implications for feedback autoregulation of active vitamin D levels at the fetomaternal interface, J. Biol. Chem. 284 (22) (2009) 14838–14848.
[20] R. Januchowski, et al., Trichostatin A down-regulate DNA methyltransferase 1 in Jurkat T cells, Cancer Lett. 246 (1–2) (2007) 313–317.
[21] B.M. Javierre, et al., Changes in the pattern of DNA methylation associate with twin discordance in systemic lupus erythematosus, Genome Res. 20 (2) (2010) 170–179.
[22] C.G. Bell, et al., Genome-wide DNA methylation analysis for diabetic nephropathy in type 1 diabetes mellitus, BMC Med. Genomics 3 (2010) 33.
[23] V.K. Rakyan, et al., Identification of type 1 diabetes-associated DNA methylation variable positions that precede disease diagnosis, PLoS Genet. 7 (9) (2011) e1002300.
[24] D. Martino, et al., Longitudinal, genome-scale analysis of DNA methylation in twins from birth to 18 months of age reveals rapid epigenetic change in early life and pair-specific effects of discordance, Genome Biol. 14 (5) (2013) R42.
[25] D.J. Martino, et al., Evidence for age-related and individual-specific changes in DNA methylation profile of mononuclear cells during early immune development in humans, Epigenetics 6 (9) (2011) 1085–1094.
[26] L. Harvey, et al., Developmental vitamin D3 deficiency induces alterations in immune organ morphology and function in adult offspring, J. Steroid Biochem. Mol. Biol. 121 (1–2) (2010) 239–242.
[27] R. Jacobsen, et al., The influence of early exposure to vitamin D for development of diseases later in life, BMC Public Health 13 (2013) 515.

Appendix III: Manuscript

The DNA methylation landscape of CD4+ T cells in oligoarticular juvenile idiopathic arthritis

Journal of Autoimmunity 86 (2018) 29–38

Contents lists available at ScienceDirect

Journal of Autoimmunity

journal homepage: www.elsevier.com/locate/jautimm

The DNA methylation landscape of CD4$^+$ T cells in oligoarticular juvenile idiopathic arthritis

Raul A. Chavez-Valencia [a, b, 1], Rachel C. Chiaroni-Clarke [a, b, 1], David J. Martino [a, b, c], Jane E. Munro [a, d], Roger C. Allen [a, d], Jonathan D. Akikusa [a, d], Anne-Louise Ponsonby [a, b], Jeffrey M. Craig [a, b], Richard Saffery [a, b], Justine A. Ellis [a, b, e, *]

[a] Murdoch Children's Research Institute, Parkville, Vic 3052, Australia
[b] Department of Paediatrics, The University of Melbourne, Vic 3010, Australia
[c] In-FLAME, the International Inflammation Network, World Universities Network (WUN), Australia
[d] Royal Children's Hospital, Parkville, Vic 3052, Australia
[e] Centre for Social and Early Emotional Development, Faculty of Health, Deakin University, Australia

ARTICLE INFO

Article history:
Received 28 August 2017
Received in revised form
25 September 2017
Accepted 25 September 2017
Available online 29 September 2017

Keywords:
Epigenetics
DNA methylation
Juvenile idiopathic arthritis
Autoimmunity
CD4$^+$ T cells

ABSTRACT

Juvenile idiopathic arthritis (JIA) is presumed to be driven by an adverse combination of genes and environment. Epigenetic processes, including DNA methylation, act as a conduit through which the environment can regulate gene activity. Altered DNA methylation has been associated with adult autoimmune rheumatic diseases such as rheumatoid arthritis, but studies are lacking for paediatric autoimmune rheumatic diseases including JIA. Here, we performed a genome-scale case-control analysis of CD4$^+$ T cell DNA methylation from 56 oligoarticular JIA (oJIA) cases and 57 age and sex matched controls using Illumina HumanMethylation450 arrays. DNA methylation at each array probe was tested for association with oJIA using RUV (Remove Unwanted Variation) together with a moderated *t*-test. Further to this 'all-inclusive' analysis, we stratified by age at diagnosis (≤6yrs, >6yrs) and by sex as potential sources of heterogeneity. Following False Discovery Rate (FDR) adjustment, no probes were associated with oJIA in the all-inclusive, >6yrs-diagnosed, or sex-stratified analyses, and only one probe was associated with oJIA in the ≤6yrs-diagnosed analysis. We attempted technical validation and replication of 14 probes ($p_{unadj}<0.01$) at genes of known/potential relevance to disease. At *VPS53*, we demonstrated a regional shift towards higher methylation in oJIA (all-inclusive) compared to controls. At *REEP3*, where polymorphism has been previously associated with JIA, we demonstrated higher DNA methylation in male oJIA compared to male controls. This is the most comprehensive JIA case-control analysis of DNA methylation to date. While we have generated some evidence of altered methylation in oJIA, substantial differences are not apparent in CD4$^+$ T cells. This may indicate a lesser relevance of DNA methylation levels in childhood, compared to adult, rheumatic disease.

© 2017 Elsevier Ltd. All rights reserved.

1. Introduction

Juvenile idiopathic arthritis (JIA) is a complex childhood autoimmune rheumatic disease that encompasses seven subtypes, as defined by the International League of Associations for Rheumatology (ILAR) [1,2]. These are based on clinical presentation, phenotypic features, and serologic markers, and are characteristically different in their age of onset and ratios of affected females to males [1]. It is generally accepted that the current subtype classification system requires refinement, because there remains significant within-subtype heterogeneity [3]. Disease heterogeneity represents a major challenge in uncovering the pathogenesis of JIA.

Numerous lines of evidence have established JIA as a complex disease [4] [5], arising from the interaction of genetic and environmental risk factors. Comprehensive twin studies to determine the relative impact of genes and environment on JIA disease susceptibility are lacking. However, there is some evidence that monozygotic twin discordance may be in the range of 25–40% [6]

* Corresponding author. Genes, Environment and Complex Disease, Murdoch Childrens Research Institute, Parkville, Vic 3052, Australia.
E-mail address: justine.ellis@mcri.edu.au (J.A. Ellis).
[1] Equal contribution.

https://doi.org/10.1016/j.jaut.2017.09.010
0896-8411/© 2017 Elsevier Ltd. All rights reserved.

Downloaded for Anonymous User (n/a) at Royal Children's Hospital - J W Grieve Lib from ClinicalKey.com.au by Elsevier on December 23, 2019.
For personal use only. No other uses without permission. Copyright ©2019. Elsevier Inc. All rights reserved.

indicating environmental contribution to disease risk. Recent work to quantify the contribution of common genetic variants to paediatric autoimmune diseases, including JIA, demonstrated that around three quarters of JIA susceptibility appears determined by such genetic variation, suggesting a moderate environmental component [7,8]. The presence of an environmental component is supported by recent data indicating sibling exposure, antibiotic exposure and sex as relevant environmental factors for JIA [9–12]. A key mediator of environmental influence on gene expression is epigenetic modification, defined as mitotically-stable alterations in gene expression in the absence of changes to DNA sequence [13]. The epigenome encompasses many layers including DNA methylation, histone modifications and non-coding RNAs, which interact to influence gene transcription, cell function, and disease risk [14].

There is a growing body of data associating epigenetic alterations with adult autoimmune rheumatic diseases, including rheumatoid arthritis (RA) and systemic lupus erythematosus (SLE) [15,16]. RA is an adult rheumatic disease with clinical and pathological similarities to JIA. At a genetic susceptibility level, RA and JIA overlap considerably [17]. However, evidence suggests that the environmental contribution to susceptibility may be greater for RA than for JIA, with RA monozygotic twin concordance rates around 15% [18], and several well-established environmental factors, such as smoking, that are known to impact risk [19]. DNA methylation variation has been associated with RA in disease-relevant cell types including whole blood [20], peripheral blood mononuclear cells (PBMCs) [21], synovial fibroblasts [22], T cells [23], and B cells [23]. In SLE, in which onset can occur at any age but is far more common in adulthood [24], there is also substantial evidence for the presence of DNA methylation alterations in various immune cell populations [25].

Determining the role of epigenetics in autoimmune rheumatic (and other) diseases provides opportunity to unveil previously-unrecognised disease pathways and new therapeutic molecular targets. Additionally, therapeutic approaches that correct epigenetic aberration by targeting epigenetic machinery are developing rapidly, and are of growing interest in rheumatic disease [26]. Given the evidence for a role for DNA methylation in adult autoimmune rheumatic diseases, but mindful of the fact that the environment may be relatively less relevant to JIA disease risk, we sought to determine whether substantial differences in DNA methylation are also associated with JIA.

Previously, we performed a small genome-scale analysis of $CD4^+$ T cell DNA methylation for JIA using medium density Illumina Infinium HumanMethylation27 (HM27) BeadChip arrays [27]. We chose $CD4^+$ T cells specifically since they play an established role in disease pathogenesis [28], and JIA-associated genetic variants are linked to $CD4^+$ T cell function [29]. We generated some evidence in this prior study to suggest that changes to DNA methylation may be associated with JIA. Therefore, in order to build on this past work, we have now carried out a larger and more comprehensive $CD4^+$ T cell JIA case-control analysis using the higher density HM450 array platform, with measurement of DNA methylation at >485,000 genomic sites [30,31]. Case heterogeneity was minimised by focusing on a single ILAR disease subtype – oligoarticular JIA (oJIA) [32], characterised by four or fewer joints affected at six months post-diagnosis, younger average age at onset, and a 2–3 times higher frequency in girls. We accounted for other sources of confounding by matching cases and controls by age and sex, by selecting only cases with active disease and no exposure to disease modifying anti-rheumatic drugs (DMARDs), and through the use of the RUV (Remove Unwanted Variation) algorithm [33] in our analyses. Further, we tested the hypothesis that accounting for age at onset and sex as sources of case heterogeneity [12,34] may more clearly reveal associations by re-performing our analyses following stratification by these factors.

2. Materials and methods

2.1. Participants

Subjects were drawn from the ChiLdhood Arthritis Risk factor Identification sTudY (CLARITY) [35]. Briefly, cases were diagnosed according to ILAR criteria by a paediatric rheumatologist at the Royal Children's Hospital (RCH) in Melbourne Australia. Controls were otherwise healthy children ≤16 years undergoing a minor surgical procedure in the RCH Day Surgery Unit. A peripheral blood sample was collected from each participant. The project was approved by the RCH Human Research Ethics Committee #27127.

For this study, 57 oJIA cases and 57 healthy controls were selected as the 'discovery' sample. Of these, 51 pairs were matched for both age (within 1 year) and sex. All cases had active disease and no prior exposure to methotrexate or biologic DMARDs. One case was subsequently excluded due to inactive disease. The average time from diagnosis to blood collection was 8.5 months. The female to male ratio was approximately 4:1. Using the same criteria, 19 oJIA age and sex matched case-control pairs were selected as a 'replication' sample. One additional unmatched male case and two additional unmatched female controls were also available for analysis and included to augment study power. Subsets of the replication sample were used for replication of different stratified analyses. The characteristics of the discovery and replication samples are shown in Supplementary Table 1.

2.2. $CD4^+$ T cell isolation and DNA extraction

The procedure for the isolation of total $CD4^+$ T cells from selected participants has been described previously [27]. Briefly, peripheral blood mononuclear cells (PBMCs) were isolated from blood samples using standard Ficoll gradient procedures within 24 h of blood collection and stored in vapour phase nitrogen until use. PBMCs were thawed at 37 °C. To isolate $CD3^+$ $CD4^+$ T cells, PBMCs were positively sorted using flow cytometry. Cells were gated according to viability, lymphocytes, then double positivity for CD3 and CD4. T-cell purity was typically ~98% and viability averaged 85%. $CD4^+$ T cell DNA was extracted using the Qiagen Flexigene kit (Qiagen, Hilden Germany) according to manufacturer's protocol.

2.3. Genome-scale DNA methylation raw data generation and preprocessing

Genomic DNA was submitted as a single batch to ServiceXS (Leiden, The Netherlands) and bisulphite converted using the EZ DNA Gold kit (Zymo, Irvine, USA). DNA methylation was measured using Illumina Infinium HumanMethylation450 BeadChip arrays (HM450) (San Diego, USA). To avoid potential confounding due to inter-array variability, cases were distributed randomly across arrays, with case-control pairs applied to the same array. DNA samples from two case-control pairs were used as technical replicates.

Raw data (idat files) were analysed in R using Bioconductor software packages [36]. The Minfi package [37] was used to preprocess raw signal intensities to methylation measurements, and also to detect any low-quality samples by inspecting the median of the methylated and unmethylated signal intensities. The SWAN (Subset-quantile Within Array Normalization) package [38] was used for probe type normalization by correcting for technical differences between type I and type II probes (which measure methylated and unmethylated signals using two paired probes, or use a single probe, respectively). X and Y chromosome probes were

Downloaded for Anonymous User (n/a) at Royal Children's Hospital - J W Grieve Lib from ClinicalKey.com.au by Elsevier on December 23, 2019.
For personal use only. No other uses without permission. Copyright ©2019. Elsevier Inc. All rights reserved.

removed to reduce confounding due to sex (except in the case of sex stratified analysis), as were probes with a detection p-value >0.01. Probes with SNPs at the single base extension site with minor allele frequency >0.05 were also removed, as well as probes known to bind non-specifically [39]. This left 430,191 probes common to all samples for subsequent analysis. Beta methylation ratios were derived for each probe and log transformed to M-values for statistical analysis as recommended for this platform [40].

2.4. All-inclusive oJIA differential DNA methylation analysis

We performed differential methylation analysis using the 'Remove Unwanted Variation' (RUV) algorithm together with the moderated *t*-test method of Smyth [41] implemented in the MissMethyl package to identify differentially methylated probes (DMPs) [33]. In this process, negative controls probes (i.e. those not associated with the phenotype of interest, in this case JIA) are used to estimate unwanted variation between samples, which is then adjusted for in the analysis. Three iterations of the bottom 90% of ranked probes were used to obtain a stable set of empirical negative control probes for estimation of the components of unwanted variation. The moderated *t*-test pools within-group variability between probes to prevent large t-statistics due to small per-probe within-group variance and results in a more accurate p value. P values were adjusted for multiple testing using the Benjamini-Hochberg method [42].

2.5. Stratified analyses

To explore potential disease heterogeneity associated with differing age at diagnosis in the all-inclusive analysis, cases were subsequently separated into two groups based on age at diagnosis ≤6 years (n = 35 cases, 29 with an age and sex matched control) or >6 years (n = 21 cases, each with an age and sex matched control). Data were reanalysed following the pipeline described above, except that the components of unwanted variation were estimated using a stable set of empirical negative control probes derived from four iterations of the bottom 70% of ranked probes.

To remove potential disease heterogeneity associated with sex, female (n = 45 cases, 41 with an age and sex matched control) and male (n = 11 cases, each with an age and sex matched control) samples were also analysed separately. In this instance sex chromosome probes were not removed prior to analysis.

2.6. Technical validation and replication of differentially methylated loci

Array probes that had an unadjusted p value of <0.01 and were annotated to a gene with prior evidence of association to immune system function or epigenetic mechanisms, were prioritised for validation and replication. Technical validation of specific sites in 'discovery' samples (where sufficient sample remained), and replication of associations in the 'replication' sample set, was performed using the MassARRAY EpiTYPER system (Agena Bioscience, San Diego, USA). In recognition of the technical limitations of this platform (accurate to ± 5% methylation) [43], larger case-control differences in methylation (Δβ) were also given priority. Amplification primers with appropriate 'tags' were designed manually and are provided in Supplementary Table 2. Where possible, EpiTYPER assays were designed to generate methylation data at the CpG corresponding to the probe of interest, as well as at other CpGs in the region (see Supplementary Table 2). Assays were run in triplicate and quality control carried out as previously described [44]. Student t-tests were used to compare group means. A consistent direction of effect and a $p < 0.05$ was considered evidence of validation/replication.

3. Results

3.1. All-inclusive oJIA analysis

Greater than 95% of the HM450 probes passed QC measures in all samples, with technical replicates showing a high level of reproducibility ($r = 0.997$, Supplementary Fig. 1). Case-control comparison of probe-wise average methylation levels did not identify any statistically significant (FDR < 0.05) probes after multiple testing correction (Table 1). Fig. 1A shows the Q-Q plot and Fig. 2A the Manhattan plot of the all-inclusive oJIA association analysis. Unsupervised clustering of samples using the top 4166 most variable probes between cases and controls (unadjusted $p < 0.01$) separated the two groups (Fig. 3A) suggesting a generalised difference in DNA methylation profile. Supplementary Table 3 lists the top 4166 DMPs included in the cluster analysis.

3.2. Stratified analyses: age at diagnosis

We next compared cases and controls after stratification according to age at diagnosis (≤6 years (younger-diagnosed) vs. >6 years (older-diagnosed)) in line with previously reported stratification in JIA PBMC gene expression patterns by these ages [34]. Fig. 1B shows the Q-Q plot and Fig. 2B the Manhattan plot of the younger-diagnosed oJIA association analysis. A single DMP with FDR <0.05 was identified in the younger-diagnosed analysis (Table 1). This probe (cg09469870, unadjusted $p = 3.5 \times 10^{-8}$, $\Delta\beta = -0.035$) is annotated to a long non-coding RNA (lncRNA) on chromosome 6. Unsupervised clustering of younger–diagnosed samples based on only the top 31 DMPs associated with oJIA (unadjusted $p < 0.0001$) showed clear separation of cases from controls (Fig. 3B). Supplementary Table 4 lists the top 31 DMPs included the younger-diagnosed cluster analysis. Fig. 1C shows the Q-Q plot and Fig. 2C the Manhattan plot of the older-diagnosed oJIA association analysis. No oJIA-associated DMPs were identified after FDR adjustment for the older-diagnosed analysis (Table 1). However, unsupervised clustering based on 248 case-control DMPs with unadjusted $p < 0.001$ largely separated cases from controls (Fig. 3C), again far fewer than for the all-inclusive analysis. Supplementary Table 5 lists the top 248 DMPs included in the older-diagnosed cluster analysis.

3.3. Stratified analyses: sex

Given the increased incidence [1] and different genetic architecture [12] of oJIA in females compared to males, we next considered whether there were any oJIA-associated methylation differences that were unique to females or males. Fig. 1D and E shows the Q-Q plots, and Fig. 2D and E shows the Manhattan plots for the female and male oJIA association analyses respectively. Case-control comparisons involving only females or males resulted in no DMPs at FDR<0.05 (Table 1). Unsupervised clustering based on 376 DMPs with unadjusted $p < 0.001$ separated female cases from female controls (Fig. 3D). Unsupervised clustering based on the top 16 DMPs with unadjusted $p < 0.0001$ was able to separate male cases from male controls with greater clarity than the females, despite the much smaller number of probes required (Fig. 3E). Supplementary Tables 6 and 7 list the top 376 and 16 DMPs included in the cluster analyses for females and males respectively.

The overlaps among probes identified (unadjusted $p < 0.01$) by the all-inclusive, age at diagnosis and sex stratified analyses are shown in Fig. 4. The vast majority of DMPs were unique to a single analysis.

Downloaded for Anonymous User (n/a) at Royal Children's Hospital - J W Grieve Lib from ClinicalKey.com.au by Elsevier on December 23, 2019.
For personal use only. No other uses without permission. Copyright ©2019. Elsevier Inc. All rights reserved.

Table 1
Differential methylation in case-control analysis for the all-inclusive sample, and for samples stratified by age at diagnosis and sex.

Analysis	Subgroup	N Case/Control	# probes analysed	DMPs FDR<0.05	DMPs p < 0.01*
All-inclusive		56/57	430,191	0	4166
Age at diagnosis	≤6yrs	35/33	430,191	1	4040
	>6yrs	21/24	430,191	0	3094
Sex	Female	45/43	440,360	0	4120
	Male	11/14	440,360	0	2858

FDR = False Discovery Rate. DMP = Differentially Methylated Probe.
*Bayesian p value.

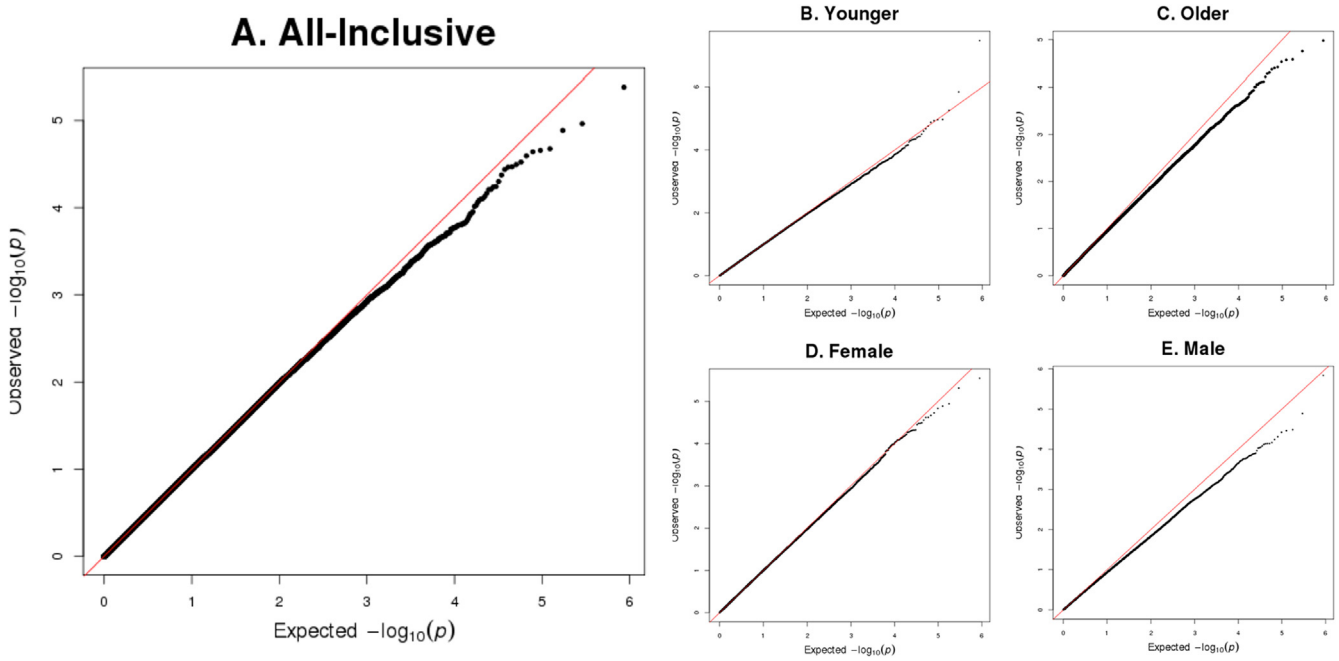

Fig. 1. Quantile-quantile (Q–Q) plots for the association between JIA and methylation at each analysed HM450 array probe. A. All-inclusive 'discovery' analysis. B. Younger-diagnosed analysis. C Older-diagnosed analysis. D. Female analysis. E. Male analysis.

3.4. Technical validation and replication

We chose a number of probes from the various analyses for technical validation and independent replication according to the criteria defined in Materials and Methods. In addition to DMPs from the all-inclusive analysis, we chose to focus on validating and replicating DMPs generated by the analysis of the younger-diagnosed and male only samples. These samples generated larger numbers of DMPs with case-control methylation differences ($\Delta\beta$) > 5%, compared to their counterparts. Selected probes and associated genes are shown in Table 2, along with the outcomes of validation and replication analyses. Overall, while most DMPs were technically validated in the discovery samples, no findings were replicated in the additional sample set using the criteria of same direction of effect and a case control difference of $p < 0.05$. Side-by-side dotplots of array, validation and replication data are provided in Supplementary Figs. 2–4.

We also looked for differential methylation at CpG sites other than those corresponding to the HM450 DMP captured by the EpiTYPER assays for each gene. For these analyses, to augment power, we combined both validation and replication data since these additional CpGs had not been analysed on any samples before. Plots showing methylation differences between cases and controls for CpGs surrounding the array probe as measured by the EpiTYPER assay (validation and replication data combined) are provided in Supplementary Figs. 5–7. For the 'all-inclusive' gene regions (Supplementary Fig. 5), two CpG units at *CYP1A1*, distal to the array probe site, were significantly (unadjusted $p < 0.05$) less methylated in cases compared to controls. Also, at *VPS53*, all but one CpG unit, including the array probe site, were significantly more highly methylated in cases than controls when the EpiTYPER validation and replication sample data were analysed together. This suggests a co-ordinated shift in methylation across the *VPS53* region in children with JIA, and is consistent with the original direction of differential methylation detected by the HM450 array. For the younger-diagnosed DMP gene regions (Supplementary Fig. 6), both *ARSA* and *POU2AF1* showed significant differential methylation at other CpG units not captured by the arrays, and in both instances the direction of effect was consistent with that detected at the array probe in that region. For the male-only DMP gene regions (Supplementary Fig. 7), only one CpG unit, at *REEP3*, was significantly differentially methylated between cases and controls. The direction of effect (higher methylation in cases) was consistent with that detected at the array probe for that region.

3.5. Analysis of previously reported JIA differentially methylated gene IL32

Previously, we reported association of oJIA with reduced $CD4^+$ T cell methylation at the pro-inflammatory gene *IL32* [45]. In the

Downloaded for Anonymous User (n/a) at Royal Children's Hospital - J W Grieve Lib from ClinicalKey.com.au by Elsevier on December 23, 2019.
For personal use only. No other uses without permission. Copyright ©2019. Elsevier Inc. All rights reserved.

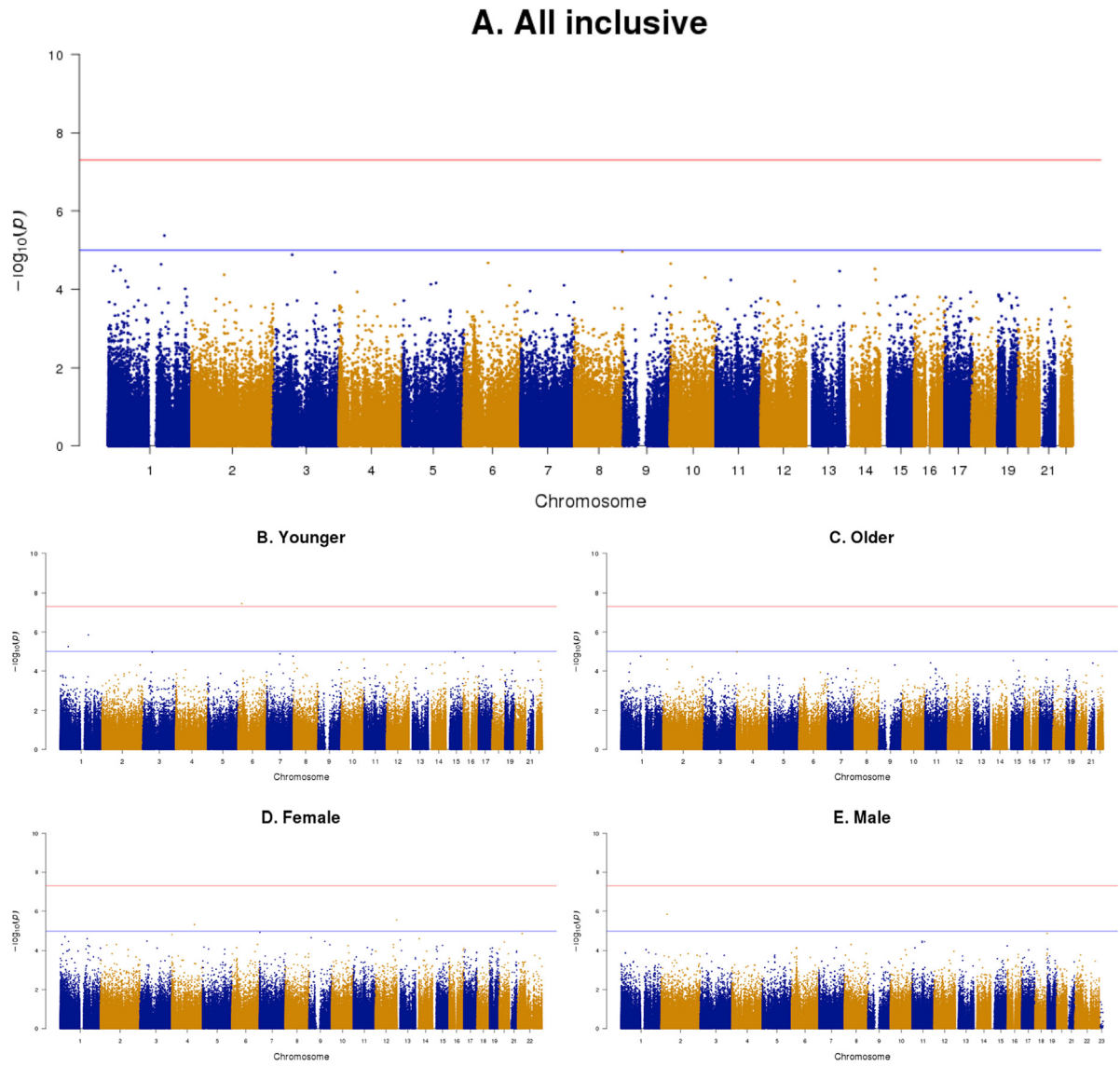

Fig. 2. Manhattan plots for the association between JIA and methylation at each analysed HM450 array probe. A. All-inclusive 'discovery' analysis. B. Younger-diagnosed analysis. C. Older-diagnosed analysis. D. Female analysis. E. Male analysis. Y axis shows the negative logarithm of the association p-value for each probe. Red and blue lines indicate the traditional genome-wide association study (GWAS) significant (10^{-8}) and suggestive (10^{-6}) p value cutoffs, respectively.

current study, we found no evidence of differential methylation at HM450 array probes annotated to *IL32*. Since the array did not measure methylation at the CpG site most strongly associated with JIA in our previous work, we used the MassARRAY platform to measure methylation at this additional site (CpG_9, see Ref. [45]). We did not find association of *IL32* methylation with oJIA using the current study samples alone, but when previous [45] and current samples were pooled (n = 65 cases, 64 controls), the association remained evident ($\Delta\beta = 0.067$, *t*-test p = 0.023).

4. Discussion

In this study we performed a carefully-designed analysis to identify DNA methylation variation associated with JIA in purified, disease-relevant $CD4^+$ T cells. This design was chosen to minimise potential epigenetic heterogeneity associated with mixed cell populations, minimising the need to adjust for cell type differences between samples. We opted to focus on paediatric-rheumatologist-confirmed oJIA to reduce case heterogeneity and we only selected cases with active disease and no prior exposure to DMARDs (including methotrexate) at the time of blood sampling. This is particularly important for DNA methylation analysis, since methotrexate is a folate analogue that inhibits one-carbon metabolism and methyl group production, needed for genomic DNA methylation. These stringent measures, along with case-control matching, and efforts to technically validate and replicate our findings, were taken to maximise our opportunity to identify true disease associations. Despite this careful study design, we were unable to detect any substantial differences in DNA methylation across the genome in children with oJIA compared to age- and sex-matched controls.

There is increasing recognition that JIA subtyping as defined according to current ILAR criteria results in biologically heterogeneous groupings of patients, prompting calls to reassess the way in which JIA is classified [3,34,46]. Indeed, molecular evidence supports a level of heterogeneity within the oligoarticular subtype, as gene expression data distinguishes cases whose age of onset was younger than 6 years from cases with an older age of onset [34]. Additionally, up to three times more girls than boys develop oJIA,

Downloaded for Anonymous User (n/a) at Royal Children's Hospital - J W Grieve Lib from ClinicalKey.com.au by Elsevier on December 23, 2019.
For personal use only. No other uses without permission. Copyright ©2019. Elsevier Inc. All rights reserved.

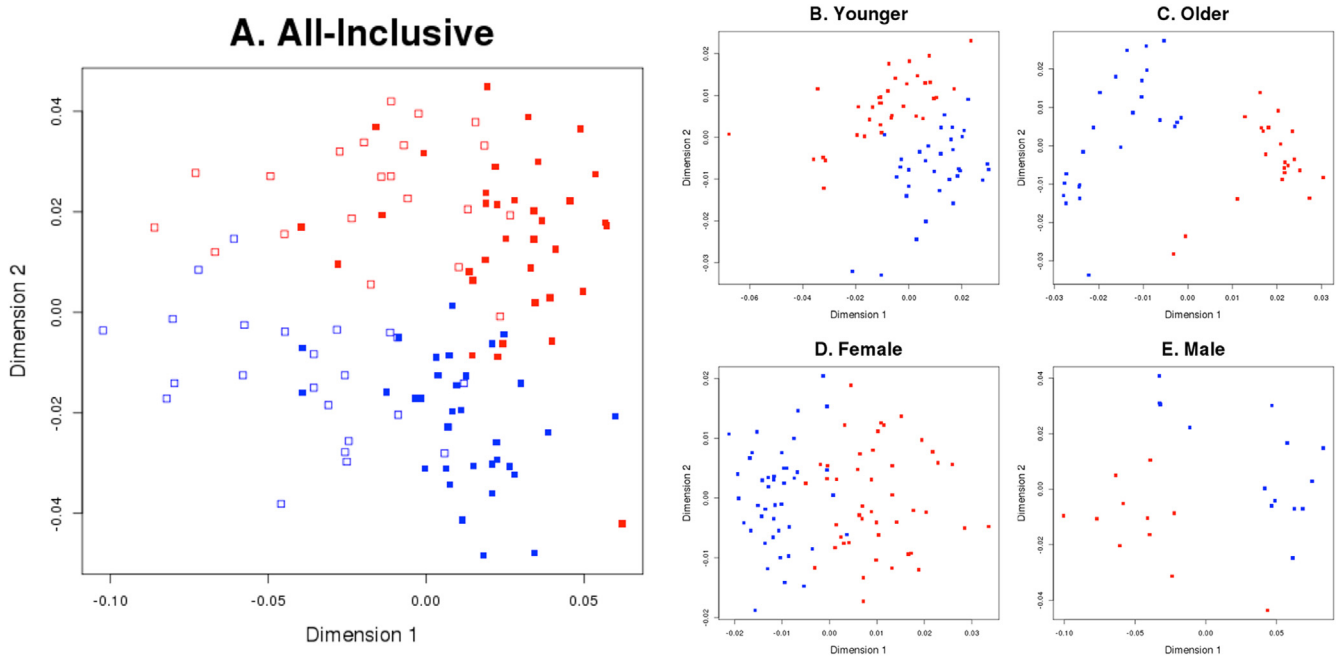

Fig. 3. Multidimensional scaling (MDS) plots showing clustering of cases (red squares) and controls (blue squares). A. All-inclusive 'discovery' cases and controls, clustered using 4166 DMPs with unadjusted $p < 0.01$. Open squares represent individuals >6 yrs and filled squares represent individuals ≤ 6 yrs at diagnosis (cases) or recruitment (controls). B. Younger-diagnosed cases and controls, clustered using 31 DMPs with unadjusted $p < 0.0001$. C Older-diagnosed cases and controls, clustered using 248 DMPs with unadjusted $p < 0.001$. D. Female cases and controls, clustered using 376 DMPs with $p < 0.001$. E. Male cases and controls, clustered using 16 DMPs with $p < 0.0001$.

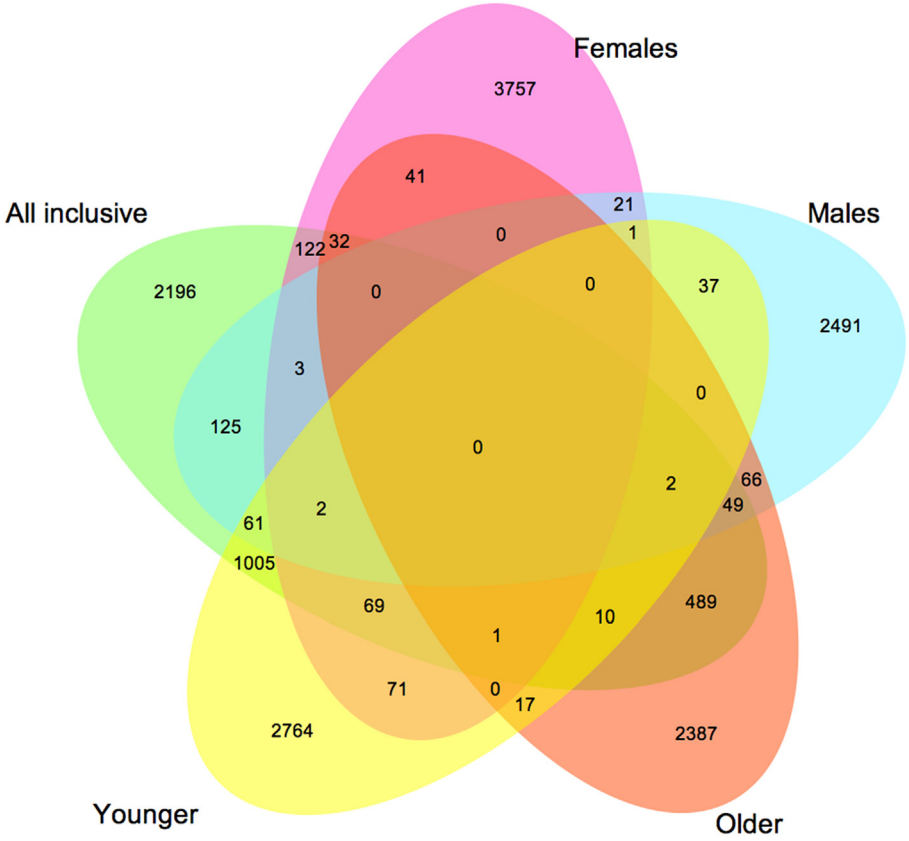

Fig. 4. Venn diagram demonstrating the overlaps amongst probes identified at unadjusted $p < 0.01$ by each of the all-inclusive, younger-diagnosed, older-diagnosed, female and male analyses.

Downloaded for Anonymous User (n/a) at Royal Children's Hospital - J W Grieve Lib from ClinicalKey.com.au by Elsevier on December 23, 2019.
For personal use only. No other uses without permission. Copyright ©2019. Elsevier Inc. All rights reserved.

Table 2
Validation and replication of selected DMPs for the all-inclusive and selected stratified analyses.

Analysis	DMP	Gene	Unadj P	Δβ	Validates?	Replicates?
Overall	cg12101586	*CYP1A1*	0.0050	−0.053	Y	N
	cg15205435	*CHD5*	0.0031	−0.062	Y	N
	cg17210938	*TACSTD2*	0.0072	−0.056	Y	N
	cg04370829	*VPS53*	0.0074	0.065	Y	N
≤6y diagnosed	cg01212326	*POU2AF1*	0.00075	−0.041	Y	N
	cg22655196	*RGS12*	0.0016	−0.065	S	N
	cg09469870	*lncRNA*	3.50×10^{-8}	−0.035	S	N
	cg13909895	*ARSA*	0.00054	−0.096	Y	N
Male	cg12424293	*NFATC1*	0.000013	0.057	N	N
	cg05592245	*REEP3*	0.00027	0.071	N	N
	cg26163153	*RHOH*	0.0016	−0.069	Y	N
	cg01498090	*TLR7*	0.0024	−0.11	Y	N
	cg06980173	*IL6R*	0.0067	−0.052	N	N
	cg12180028	*AR*	0.0067	−0.057	S	N

DMP = differentially methylated probe.
Gene = annotated UCSC RefGene
Unadj P = unadjusted p value.
Δβ = delta beta (difference in mean β methylation value, case − control).
Validates = Technical validation of array finding, Y = Yes - same direction of effect, $p < 0.05$, S = suggestive - same direction of effect, $p < 0.2$, N = no - different direction of effect, or $p > 0.2$.
Replicates = replication in a separate set of case control samples; Y, S, N as above.

raising the possibility of sex-specific pathways to disease. In support of this, we have previously reported that a variant of the major non-HLA JIA susceptibility gene, *PTPN22*, increases risk of JIA in females, but not in males [12]. Since case heterogeneity may obscure association of DNA methylation alterations with JIA, we endeavoured to create more homogenous case groupings by performing stratified analyses of cases and controls grouped by age at diagnosis, and by sex. While this division of the overall dataset reduced sample size, and thus our power to detect case-control differences, in each instance we observed a clearer clustering of cases from controls that was based on fewer DMPs (4166 DMPs for all inclusive clustering compared to 31 for younger-diagnosed, 248 for older-diagnosed, 16 for male and 376 for female clustering). This suggests that, as a group, the top-ranked DMPs may be useful in predicting a case from a control, especially when case heterogeneity is minimised. A future 'machine learning' study to explore the ability of genome-scale methylation data to predict oJIA would be ideal to explore this issue further.

There was very little overlap between the DMPs (unadjusted $p < 0.01$) detected by any of the various analyses performed. Thirty DMPs were common to the younger- and older-diagnosed analyses, and 27 DMPs common to the male and female analyses. These data support a high level of heterogeneity between the four age- and sex-based case sub-groups which may be consistent with a reduction in heterogeneity upon stratification. Case-control clustering using far fewer DMPs provides further support of this. However, it should be noted that the lack of overlap between analyses may also be a result of heterogeneity among the samples generally.

For younger-diagnosed cases we detected a single DMP that remained statistically significant following FDR adjustment. This probe (cg09469870) measures DNA methylation at a CpG site annotated to a long non-coding RNA (lncRNA) on chromosome 6. Interestingly, this probe has been reported to be differentially methylated in the inflammatory conditions giant cell arteritis [47] and primary Sjögren's syndrome [48]. However, we were unable to replicate the JIA case-control difference in methylation at this probe in a separate case-control sample. Although the replication sample was small, these data suggest that differential methylation at this locus is neither substantial nor common in oJIA, and thus its general relevance to oJIA disease processes is unclear.

The design of the EpiTYPER assays to cover probes of interest afforded the opportunity to consider case-control differences at CpG sites surrounding the array probes. Using this regional approach, we generated some further evidence of association of oJIA with altered methylation. For example, in the all-inclusive DMP regional analysis, we detected a regional shift to increased methylation in oJIA cases compared to controls at *VPS53*. This gene is involved in vesicular trafficking, and is related to the ability of HIV to successfully infect immune cells [49]. Also, at *REEP3* we identified a CpG site that was more highly methylated in oJIA male cases compared to male controls (Δβ 0.046, *t*-test $p = 0.048$). This gene formed part of a CpG 'module' (group of genes at which methylation is correlated) that was recently associated with clinical outcome following anti-TNF treatment in $CD4^+$ T cells from extended oligoarticular and polyarticular JIA cases [50]. Genetic variation in the *REEP3-JMJD1C* region has also previously been associated with oligoarticular and polyarticular rheumatoid factor negative JIA [51]. Genetic variation is known to contribute to the determination of DNA methylation across the genome (referred to as methylation quantitative trait loci, or meQTLs) [52], and alterations to DNA methylation may be the functional mechanism though which SNPs alter disease risk. Further work that brings together DNA methylation and SNP data at *REEP3* in large sample sets may shed further light on the role of this locus in oJIA.

The lack of a substantially different DNA methylation profile associated with oJIA is in contrast to studies of other autoimmune rheumatic diseases, including adult RA (for recent review, see Ref. [15]). In RA, a variety of tissue types have been analysed on a genomic scale, including PBMCs and synovial fibroblasts [15]. Of particular relevance to this study, recent published work compared DNA methylation in $CD3^+$ T lymphocytes from 23 cases of newly diagnosed RA with those from 11 healthy controls, using the HM450 array [53]. They identified 1951 differentially methylated probes (DMPs) between cases and controls at FDR<0.05. As with our study, they ensured that cases were naïve to DMARD, including methotrexate. However, in contrast to our study, they analysed a broader population of T cells (only around 60% of $CD3^+$ T cells are $CD4^+$ [54]), did not match cases and controls for age and sex, and did not describe specific statistical adjustments for these (or any

Downloaded for Anonymous User (n/a) at Royal Children's Hospital - J W Grieve Lib from ClinicalKey.com.au by Elsevier on December 23, 2019.
For personal use only. No other uses without permission. Copyright ©2019. Elsevier Inc. All rights reserved.

other) potential confounding factors. In particular, only 65% of cases were female, compared to 100% of controls. Additionally, they did not remove cross-reactive probes, or probes potentially associated with SNPs. Thus, while the difference in the number of DMPs identified by this RA study is in stark contrast to our findings for JIA despite the smaller RA study sample size, differences between the two studies in relation to management of confounding may account for at least some of the contrast.

A recent study sought to understand the molecular similarities between the adult autoimmune rheumatic diseases RA, SLE, and systemic sclerosis (SSc) [55]. Both DNA methylation and gene expression were measured in $CD4^+$ T cells from incident, mostly treatment-naive cases of each disease (RA $n = 13$, SLE $n = 12$, SSc $n = 19$) and from controls ($n = 8$). When cases of all three diseases were considered together and compared to controls, no statistically significant DMPs were identified following FDR adjustment. However, amongst the 130 identified DMPs with both an unadjusted all-case vs control $P < 0.0001$ and no significant differential methylation between case groups, the authors identified 33 genes at which there was also significant evidence of differential gene expression. Interestingly, 11 genes to which these probes are annotated also appear in our study's 'all-inclusive' oJIA list of DMPs at $p < 0.01$ (Supplementary Tables 3) and a further 11 appear in the lists of DMPs identified by our stratified analyses. In particular, one gene identified by the adult study, *PCCA*, is annotated to six probes in our oJIA 'all-inclusive' DMP list. *PCCA* encodes the propionyl CoA carboxylase alpha subunit, and mutations in this gene cause the metabolic disorder propionic acidemia [56]. How it might be related to autoimmune rheumatic disease is unclear. The top-ranked oJIA *PCCA* probe (cg24978725) has a $\Delta\beta$ −0.034 and unadjusted $p = 0.0009$, and appears also as the top ranked of 5 *PCCA*-annotated probes in the 'older-diagnosed' oJIA DMP list ($\Delta\beta = -0.054$, $p = 0.0008$). These data might indicate that oJIA shares some epigenetic overlap with adult autoimmune rheumatic diseases, possibly most strongly in those cases where oJIA has an older age at onset.

It has been hypothesised that genetics may play a larger role in juvenile-onset autoimmune diseases compared to their adult counterparts [57,58]. Juvenile onset autoimmune disease is often more severe and/or more strongly genetic (and sometimes monogenic), as shown for SLE [59] and inflammatory bowel disease (IBD) [60]. Given the reduced window of opportunity for accumulation of exposure to relevant environmental factors, it might make sense that there would be a stronger role for genetics in juvenile-onset disease. Indeed, in a recent variance components analysis of JIA susceptibility, Li et al. reported that greater than 70% of susceptibility to JIA appears determined by common SNPs [57]. These data, along with an apparent lack of substantial difference in the $CD4^+$ T cell DNA methylome in oJIA compared to controls, might suggest that the environment plays only a moderate role in JIA. Another juvenile-onset autoimmune disease shown by Li et al. to have a relatively small environmental susceptibility component was type 1 diabetes (T1D). Whilst some genome-scale studies in the past have reported T1D DMPs, these predominantly twin-based studies have often utilised mixed cell types [61] or cell lines [62], and did not always correct for multiple testing in their discovery analyses [61,63]. In contrast, a recent study of DNA methylation in 52 monozygotic twin pairs discordant for T1D using the HM450 array identified only one DMP across three purified primary immune cell types (T cells, B cells and monocytes) that remained significant after FDR adjustment [64]. The lack of DMPs identified by this study in this predominantly juvenile-onset autoimmune disease mirrors our findings in oJIA. Further study of the relative contributions of genes and environment to both adult and juvenile onset autoimmune disease susceptibility, and how these contributions correlate with the presence of disease-associated differential DNA methylation, are needed to better understand why some autoimmune diseases might be more strongly associated with differential DNA methylation than others.

Our careful approach to case phenotyping (confirmed single ILAR subtype, active disease, no exposure to DMARD) was a strength of the current study. However, this also limited the number of our biobanked cases available for discovery and replication datasets. Thus, our study was underpowered to identify more subtle disease associated differences in the DNA methylome, and the presence of these cannot be ruled out by our data. This issue was further exacerbated by stratification of the dataset by age at diagnosis and by sex for additional analyses. We also concentrated on a single immune cell type, $CD4^+$ T cells. While the selection of this cell type is supported by existing evidence for involvement of these cells in disease [5,28,29], it remains possible that the study of other immune cell types, or cells drawn from the site of inflammation such as synovial cells, may reveal more substantial DNA methylation differences relevant to disease. It is also important to note that alterations to methylation specifically found in $CD4^+$ T cell subsets, such as regulatory T cells that constitute around 8% of human $CD4^+$ T cells in healthy adults [65], may not have been detectable by our study. In addition, it remains possible that DNA methylation outside the regions interrogated here (around 5% of all CpG sites across the genome are measured by the HM450 array), or other forms of epigenetic modification not considered by this study such as histone modification, may be of more relevance.

5. Conclusion

In this carefully conducted study, we found no evidence of substantial and common alteration to the DNA methylome in peripheral $CD4^+$ T cells of children with oligoarticular JIA. This is in contrast to findings for adult-onset autoimmune rheumatic diseases such as RA. Our data do, however, support the hypothesis that heterogeneity in epigenetic variation might exist between cases, even within this single ILAR subtype. Consideration of sources of heterogeneity amongst JIA cases in future studies may assist to more clearly reveal disease-relevant epigenetic and other molecular associations.

Acknowledgements

We thank the families who kindly participated in this study. We thank the research nurses and research assistants who undertook recruitment and data entry. We acknowledge support from the Australian National Health and Medical Research Council (NHMRC) (GNT1026349), the Australian Research Council (ARC) (FT120100253), Arthritis Australia, Rebecca L Cooper Foundation, LEW Carty Charitable Fund, ANZ Medical Research and Technologies in Victoria Fund, Lynne Quayle Charitable Trust, The Victorian State Government Operational Infrastructure Support Program, and the Murdoch Childrens Research Institute. JAE was supported by an ARC Future Fellowship. RAC was supported by an Arthritis Australia Postgraduate Scholarship. RS and ALP were supported by NHMRC Senior Research Fellowships.

Appendix A. Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.jaut.2017.09.010.

Downloaded for Anonymous User (n/a) at Royal Children's Hospital - J W Grieve Lib from ClinicalKey.com.au by Elsevier on December 23, 2019.
For personal use only. No other uses without permission. Copyright ©2019. Elsevier Inc. All rights reserved.

References

[1] A. Ravelli, A. Martini, Juvenile idiopathic arthritis, Lancet 369 (2007) 767–778.
[2] R.E. Petty, T.R. Southwood, P. Manners, J. Baum, D.N. Glass, J. Goldenberg, et al., International League of Associations for Rheumatology classification of juvenile idiopathic arthritis: second revision, Edmonton, J. Rheumatol. 2004 (31) (2001) 390–392.
[3] A. Martini, It is time to rethink juvenile idiopathic arthritis classification and nomenclature, Ann. Rheum. Dis. 71 (2012) 1437–1439.
[4] S. Prahalad, Genetic analysis of juvenile rheumatoid arthritis: approaches to complex traits, Curr. Probl. Pediatr. Adolesc. Health Care 36 (2006) 83–90.
[5] A. Hinks, J. Cobb, M.C. Marion, S. Prahalad, M. Sudman, J. Bowes, et al., Dense genotyping of immune-related disease regions identifies 14 new susceptibility loci for juvenile idiopathic arthritis, Nat. Genet. 45 (2013) 664–669.
[6] A.O. Hersh, S. Prahalad, Immunogenetics of juvenile idiopathic arthritis: a comprehensive review, J. Autoimmun. 64 (2015) 113–124.
[7] Y.R. Li, J. Li, S.D. Zhao, J.P. Bradfield, F.D. Mentch, S.M. Maggadottir, et al., Meta-analysis of shared genetic architecture across ten pediatric autoimmune diseases, Nat. Med. 21 (2015) 1018–1027.
[8] J.A. Ellis, J.E. Munro, A.L. Ponsonby, Possible environmental determinants of juvenile idiopathic arthritis, Rheumatol. Oxf. 49 (2010) 411–425.
[9] J. Miller, A.L. Ponsonby, A. Pezic, A. Kemp, S.E. Piper, J.D. Akikusa, et al., Sibling exposure and risk of juvenile idiopathic arthritis in the ChiLdhood Arthritis Risk factor Identification sTudY (CLARITY), Arthritis Rheumatol. 67 (2015) 1951–1958.
[10] M. Arvonen, L.J. Virta, T. Pokka, L. Kroger, P. Vahasalo, Repeated exposure to antibiotics in infancy: a predisposing factor for juvenile idiopathic arthritis or a sign of this Group's greater susceptibility to infections? J. Rheumatol. 42 (2014) 521–526.
[11] D.B. Horton, F.I. Scott, K. Haynes, M.E. Putt, C.D. Rose, J.D. Lewis, et al., Antibiotic exposure and juvenile idiopathic arthritis: a case-control study, Pediatrics 136 (2015) e333–e343.
[12] R.C. Chiaroni-Clarke, Y.R. Li, J.E. Munro, R.A. Chavez, K.J. Scurrah, A. Pezic, et al., The association of PTPN22 rs2476601 with juvenile idiopathic arthritis is specific to females, Genes Immun. 16 (2015) 495–498.
[13] A. Bird, Perceptions of epigenetics, Nature 447 (2007) 396–398.
[14] B. Gupta, R.D. Hawkins, Epigenomics of autoimmune diseases, Immunol. Cell Biol. 93 (2015) 271–276.
[15] B. Sun, L. Hu, Z.Y. Luo, X.P. Chen, H.H. Zhou, W. Zhang, DNA methylation perspectives in the pathogenesis of autoimmune diseases, Clin. Immunol. 164 (2016) 21–27.
[16] Q. Lu, The critical importance of epigenetics in autoimmunity, J. Autoimmun. 41 (2013) 1–5.
[17] A. Hinks, S. Eyre, X.Y. Ke, A. Barton, P. Martin, E. Flynn, et al., Overlap of disease susceptibility loci for rheumatoid arthritis and juvenile idiopathic arthritis, Ann. Rheumat. Dis. 69 (2010) 1049–1053.
[18] E. Generali, A. Ceribelli, M.A. Stazi, C. Selmi, Lessons learned from twins in autoimmune and chronic inflammatory diseases, J. Autoimmun. 83 (2017) 51–61.
[19] D. Alpizar-Rodriguez, A. Finckh, Environmental factors and hormones in the development of rheumatoid arthritis, Semin. Immunopathol. 39 (2017) 461–468.
[20] Y. Liu, M.J. Aryee, L. Padyukov, M.D. Fallin, E. Hesselberg, A. Runarsson, et al., Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis, Nat. Biotechnol. 31 (2013) 142–147.
[21] C.J. Nile, R.C. Read, M. Akil, G.W. Duff, A.G. Wilson, Methylation status of a single CpG site in the IL6 promoter is related to IL6 messenger RNA levels and rheumatoid arthritis, Arthritis Rheum. 58 (2008) 2686–2693.
[22] E. Karouzakis, M. Trenkmann, R.E. Gay, B.A. Michel, S. Gay, M. Neidhart, Epigenome analysis reveals TBX5 as a novel transcription factor involved in the activation of rheumatoid arthritis synovial fibroblasts, J. Immunol. 193 (2014) 4945–4951.
[23] J.R. Glossop, R.D. Emes, N.B. Nixon, K.E. Haworth, J.C. Packham, P.T. Dawes, et al., Genome-wide DNA methylation profiling in rheumatoid arthritis identifies disease-associated methylation changes that are distinct to individual T- and B-lymphocyte populations, Epigenetics 9 (2014) 1228–1237.
[24] G.J. Pons-Estel, M.F. Ugarte-Gil, G.S. Alarcon, Epidemiology of systemic lupus erythematosus, Expert Rev. Clin. Immunol. 13 (2017) 799–814.
[25] H. Long, H. Yin, L. Wang, M.E. Gershwin, Q. Lu, The critical role of epigenetics in systemic lupus erythematosus and autoimmunity, J. Autoimmun. 74 (2016) 118–138.
[26] S.G. Gray, Epigenetic-based immune intervention for rheumatic diseases, Epigenomics 6 (2014) 253–271.
[27] J.A. Ellis, J.E. Munro, R.A. Chavez, L. Gordon, J.E. Joo, J.D. Akikusa, et al., Genome-scale case-control analysis of CD4+ T-cell DNA methylation in juvenile idiopathic arthritis reveals potential targets involved in disease, Clin. Epigene. 4 (2012) 20.
[28] B. Prakken, S. Albani, A. Martini, Juvenile idiopathic arthritis, Lancet 377 (2011) 2138–2149.
[29] K. Jiang, L. Zhu, M.J. Buck, Y. Chen, B. Carrier, T. Liu, et al., Disease-associated single-nucleotide polymorphisms from noncoding regions in juvenile idiopathic arthritis are located within or adjacent to functional genomic elements of human neutrophils and CD4+ T cells, Arthritis Rheumatol. 67 (2015) 1966–1977.
[30] M. Bibikova, B. Barnes, C. Tsan, V. Ho, B. Klotzle, J.M. Le, et al., High density DNA methylation array with single CpG site resolution, Genomics 98 (2011) 288–295.
[31] J. Sandoval, H. Heyn, S. Moran, J. Serra-Musach, M.A. Pujana, M. Bibikova, et al., Validation of a DNA methylation microarray for 450,000 CpG sites in the human genome, Epigenetics 6 (2011) 692–702.
[32] M. Manchia, J. Cullis, G. Turecki, G.A. Rouleau, R. Uher, M. Alda, The impact of phenotypic and genetic heterogeneity on results of genome wide association studies of complex diseases, PLoS One 8 (2013) e76295.
[33] J. Maksimovic, J.A. Gagnon-Bartsch, T.P. Speed, A. Oshlack, Removing unwanted variation in a differential methylation analysis of Illumina HumanMethylation450 array data, Nucleic Acids Res. 43 (2015) e106.
[34] M.G. Barnes, A.A. Grom, S.D. Thompson, T.A. Griffin, L.K. Luyrink, R.A. Colbert, et al., Biologic similarities based on age at onset in oligoarticular and polyarticular subtypes of juvenile idiopathic arthritis, Arthritis Rheum. 62 (2010) 3249–3258.
[35] J.A. Ellis, A.L. Ponsonby, A. Pezic, R.A. Chavez, R.C. Allen, J.D. Akikusa, et al., CLARITY - ChiLdhood arthritis risk factor identification sTudY, Pediatr. Rheumatol. Online J. 10 (2012) 37.
[36] W. Huber, V.J. Carey, R. Gentleman, S. Anders, M. Carlson, B.S. Carvalho, et al., Orchestrating high-throughput genomic analysis with bioconductor, Nat. Methods 12 (2015) 115–121.
[37] M.J. Aryee, A.E. Jaffe, H. Corrada-Bravo, C. Ladd-Acosta, A.P. Feinberg, K.D. Hansen, et al., Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays, Bioinformatics 30 (2014) 1363–1369.
[38] J. Maksimovic, L. Gordon, Oshlack A. SWAN, Subset-quantile within array normalization for illumina infinium HumanMethylation450 BeadChips, Genome Biol. 13 (2012) R44.
[39] Y.A. Chen, M. Lemire, S. Choufani, D.T. Butcher, D. Grafodatskaya, B.W. Zanke, et al., Discovery of cross-reactive probes and polymorphic CpGs in the illumina infinium HumanMethylation450 microarray, Epigenetics 8 (2013) 203–209.
[40] P. Du, X.A. Zhang, C.C. Huang, N. Jafari, W.A. Kibbe, L.F. Hou, et al., Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis, Bmc Bioinformat. (2010) 11.
[41] G.K. Smyth, Linear models and empirical bayes methods for assessing differential expression in microarray experiments, Stat. Appl. Genet. Mol. Biol. 3 (2004). Article3.
[42] Y. Benjamini, Y. Hochberg, Controlling the false discovery rate: a practical and powerful approach to multiple testing, J. R. Stat. Soc. Ser. B Methodol. 57 (1995) 289–300.
[43] M.W. Coolen, A.L. Statham, M. Gardiner-Garden, S.J. Clark, Genomic profiling of CpG methylation and allelic specificity using quantitative high-throughput mass spectrometry: critical evaluation and improvements, Nucleic Acids Res. 35 (2007) e119.
[44] D. Martino, Y.J. Loke, L. Gordon, M. Ollikainen, M.N. Cruickshank, R. Saffery, et al., Longitudinal, genome-scale analysis of DNA methylation in twins from birth to 18 months of age reveals rapid epigenetic change in early life and pair-specific effects of discordance, Genome Biol. 14 (2013) R42.
[45] B. Meyer, R.A. Chavez, J.E. Munro, R.C. Chiaroni-Clarke, J.D. Akikusa, R.C. Allen, et al., DNA methylation at IL32 in juvenile idiopathic arthritis, Sci. Rep. 5 (2015) 11063.
[46] A. Ravelli, G.C. Varnier, S. Oliveira, E. Castell, O. Arguedas, A. Magnani, et al., Antinuclear antibody-positive patients should be grouped as a separate category in the classification of juvenile idiopathic arthritis, Arthritis Rheum. 63 (2011) 267–275.
[47] P. Coit, L.B. De Lott, B. Nan, V.M. Elner, A.H. Sawalha, DNA methylation analysis of the temporal artery microenvironment in giant cell arteritis, Ann. Rheum. Dis. 75 (2016) 1196–1202.
[48] J. Imgenberg-Kreuz, J.K. Sandling, J.C. Almlof, J. Nordlund, L. Signer, K.B. Norheim, et al., Genome-wide DNA methylation analysis in multiple tissues in primary Sjogren's syndrome reveals regulatory effects at interferon-induced genes, Ann. Rheum. Dis. 75 (2016) 2029–2036.
[49] A.L. Brass, D.M. Dykxhoorn, Y. Benita, N. Yan, A. Engelman, R.J. Xavier, et al., Identification of host proteins required for HIV infection through a functional genomic screen, Science 319 (2008) 921–926.
[50] R. Spreafico, M. Rossetti, J.W. Whitaker, W. Wang, D.J. Lovell, S. Albani, Epipolymorphisms associated with the clinical outcome of autoimmune arthritis affect CD4+ T cell activation pathways, Proc. Natl. Acad. Sci. U. S. A. 113 (2016) 13845–13850.
[51] S.D. Thompson, M.C. Marion, M. Sudman, M. Ryan, M. Tsoras, T.D. Howard, et al., Genome-wide association analysis of juvenile idiopathic arthritis identifies a new susceptibility locus at chromosomal region 3q13, Arthritis Rheumat. 64 (2012) 2781–2791.
[52] J.T. Bell, A.A. Pai, J.K. Pickrell, D.J. Gaffney, R. Pique-Regi, J.F. Degner, et al., DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines, Genome Biol. 12 (2011) R10, 2011;12.
[53] J.R. Glossop, R.D. Emes, N.B. Nixon, J.C. Packham, A.A. Fryer, D.L. Mattey, et al., Genome-wide profiling in treatment-naive early rheumatoid arthritis reveals DNA methylome changes in T and B lymphocytes, Epigenomics 8 (2016) 209–224.
[54] R. Valiathan, K. Deeb, M. Diamante, M. Ashman, N. Sachdeva, D. Asthana, Reference ranges of lymphocyte subsets in healthy adults and adolescents

Downloaded for Anonymous User (n/a) at Royal Children's Hospital - J W Grieve Lib from ClinicalKey.com.au by Elsevier on December 23, 2019.
For personal use only. No other uses without permission. Copyright ©2019. Elsevier Inc. All rights reserved.

with special mention of T cell maturation subsets in adults of South Florida, Immunobiology 219 (2014) 487–496.
[55] M. Hudson, S. Bernatsky, I. Colmegna, M. Lora, T. Pastinen, K. Klein Oros, et al., Novel insights into systemic autoimmune rheumatic diseases using shared molecular signatures and an integrative analysis, Epigenetics 12 (2017) 433–440.
[56] L.R. Desviat, R. Sanchez-Alcudia, B. Perez, C. Perez-Cerda, R. Navarrete, R. Vijzelaar, et al., High frequency of large genomic deletions in the PCCA gene causing propionic acidemia, Mol. Genet. Metab. 96 (2009) 171–176.
[57] Y.R. Li, S.D. Zhao, J. Li, J.P. Bradfield, M. Mohebnasab, L. Steel, et al., Genetic sharing and heritability of paediatric age of onset autoimmune diseases, Nat. Commun. 6 (2015) 8442.
[58] L. de Ridder, R.K. Weersma, G. Dijkstra, G. van der Steege, M.A. Benninga, I.M. Nolte, et al., Genetic susceptibility has a more important role in pediatric-onset Crohn's disease than in adult-onset Crohn's disease, Inflamm. Bowel Dis. 13 (2007) 1083–1092.
[59] L.T. Hiraki, E.D. Silverman, Genomics of systemic lupus erythematosus: insights gained by studying monogenic young-onset systemic lupus erythematosus, Rheum. Dis. Clin. North Am. 43 (2017) 415–434.
[60] Q. Li, C.H. Lee, L.A. Peters, L.A. Mastropaolo, C. Thoeni, A. Elkadri, et al., Variants in TRIM22 that affect NOD2 signaling are associated with very-early-onset inflammatory bowel disease, Gastroenterology 150 (2016) 1196–1207.
[61] E. Elboudwarej, M. Cole, F.B. Briggs, A. Fouts, P.R. Fain, H. Quach, et al., Hypomethylation within gene promoter regions and type 1 diabetes in discordant monozygotic twins, J. Autoimmun. 68 (2016) 23–29.
[62] M. Stefan, W. Zhang, E. Concepcion, Z. Yi, Y. Tomer, DNA methylation profiles in type 1 diabetes twins point to strong epigenetic effects on etiology, J. Autoimmun. 50 (2014) 33–37.
[63] V.K. Rakyan, H. Beyan, T.A. Down, M.I. Hawa, S. Maslau, D. Aden, et al., Identification of type 1 diabetes-associated DNA methylation variable positions that precede disease diagnosis, PLoS Genet. 7 (2011) e1002300.
[64] D.S. Paul, A.E. Teschendorff, M.A. Dang, R. Lowe, M.I. Hawa, S. Ecker, et al., Increased DNA methylation variability in type 1 diabetes across three immune effector cell types, Nat. Commun. 7 (2016) 13555.
[65] G. Churlaud, F. Pitoiset, F. Jebbawi, R. Lorenzon, B. Bellier, M. Rosenzwajg, et al., Human and mouse CD8(+)CD25(+)FOXP3(+) regulatory T cells at steady state and during Interleukin-2 therapy, Front. Immunol. 6 (2015) 171.

Downloaded for Anonymous User (n/a) at Royal Children's Hospital - J W Grieve Lib from ClinicalKey.com.au by Elsevier on December 23, 2019.
For personal use only. No other uses without permission. Copyright ©2019. Elsevier Inc. All rights reserved.