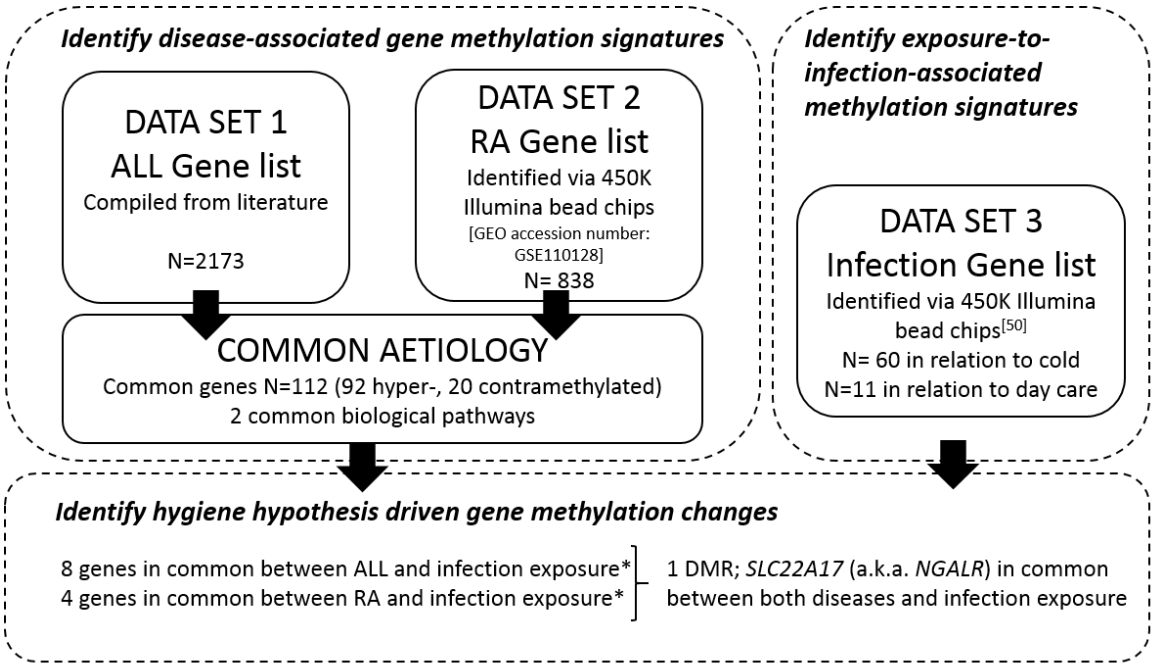
Structured Abstract

Aims; The hygiene hypothesis states lack of infection in early-life suppresses immune system development, and is linked to respiratory allergy (RA) and childhood acute lymphoblastic leukaemia (ALL) risk. Little is known about underlying mechanisms, but DNA methylation is altered in RA and ALL, and in response to infection. We investigated if aberrant methylation may be in common between these diseases and associated with infection. Methods; RA and ALL disease-associated methylation signatures were compared and related to exposure-to-infection-signatures. Results; A significant number of genes overlapped between RA and ALL signatures (p=0.0019). Significant overlaps were observed between exposure-to-infection signatures and disease-associated signatures. Conclusion; DNA methylation may be a mediating mechanism through which the hygiene hypothesis is associated with RA and ALL risk.

Graphical Abstract



Key words; epigenetics, allergy, DNA methylation, childhood acute lymphoblastic leukaemia, hygiene hypothesis, infection, exposure, proxy, day care, cold symptoms.

Abbreviations: Acute lymphoblastic leukaemia, ALL; Respiratory Allergy, RA.

Introduction

Rates of respiratory allergy and acute lymphoblastic leukaemia in childhood have risen over the last few decades [1-3]. Changes in lifestyle and environmental risk factors rather than genetic factors alone are thought to play an important role in disease aetiologies.

In 1989, Dr David Strachan suggested that the increased prevalence of allergic diseases in westernised populations in the 20th century may be explained by lower exposure to infections in early childhood [4, 5]. This proposal, coined the Hygiene Hypothesis, intimated that early childhood infection may protect against the development of allergies, and that increased allergy rates in western populations may be due to reduced infection rates via improvements in sanitation and reduced family size [4]. This statement is supported by studies in which proxy measures of early childhood infection such as birth order [6-11] or day care attendance [11-16] have been inversely associated with allergies. Furthermore, prolonged breastfeeding, which provides infants with a supply of IgA antibodies [17] and stems cells [18] as well as other properties [19-23], leading to a more mature immune system [24], have also been reported to be protective against allergies. But, the evidence for this association is fairly weak (reviewed in [25]).

At around the same time, Kinlen and Greaves were formulating related hypotheses with regard to the aetiology of childhood leukaemia. Kinlen proposed the Population Mixing Hypothesis that suggested an influx of people into rural areas introduces new infections to residing individuals who are more susceptible to infections, and subsequent leukaemia risk, due to a previous lack of exposure [26]. Meanwhile, Greaves proposed the Delayed Infection Hypothesis (later referred as Hygiene Hypothesis[27]), which suggests that in order to have an efficient immune response in childhood and later life, early life (i.e.<1 year) exposure to infections is required to build a proficient and adaptive immune reaction [28]. Without this early immune response, increased proliferation and errors are more likely to occur when an immune response is elicited later in life (i.e. delayed response). Greaves suggested this delayed response was likely to be the 'second hit' contributing to

59 the development of childhood leukaemia, specifically the more common acute lymphoblastic
60 leukaemias (ALL) [29, 30]. Epidemiological evidence in support of this latter disease model suggests
61 an association between lower rates of infection in infancy and an increased risk of leukaemia [31-
62 34]. Further inverse associations between proxy measures of infection or immune stimulation and
63 ALL risk have been reported (i.e. day care attendance [31-37], birth order [31, 34], exposure to
64 animals [31, 34] and breastfeeding [31, 33, 34, 37, 38]) supporting a role of the delayed infection, or
65 hygiene, hypothesis in ALL development.
66 Epidemiological evidence is suggestive of a role of the Hygiene Hypothesis in both RA and ALL.
67 However, more evidence in support of a role of early exposure to infection or immune stimulation
68 and risk of both RA and ALL is required. Furthermore, the underlying biological mechanisms possibly
69 involved in explaining these associations are yet to be identified. Since RA and ALL have this
70 suggested common aetiology, it is plausible that the mechanism involved in disease progression may
71 also be common between diseases. DNA methylation, an important epigenetic mechanism of gene
72 regulation, has been reported to be altered in both RA [39, 40] and ALL [41, 42], and has,
73 independently, been proposed as a disease mechanism [41, 43, 44]. DNA methylation is subject to
74 change in response to a range of environmental cues [45, 46] including infection [47-50], and may
75 act as a mediator between infection exposure and/or immune stimulation and disease risk.

76 If the Hygiene Hypothesis is aetiologically instrumental in the development of these diseases, we
77 hypothesized that a) there would be common disease-associated DNA methylation marks between
78 RA and ALL and b) methylation changes found to be associated with proxy measures for infection
79 and/or disease stimulus should also be observed (Figure 1). To test this, we determined and
80 compared disease-associated DNA methylation signatures from children with RA and ALL.
81 Additionally, using a meet-in-the-middle-approach, we compared DNA methylation changes
82 associated with proxy measures of infection (i.e. day care attendance and reported infection in early
83 life) with RA and ALL methylation signatures.

Materials and Methods

Disease-associated gene methylation signatures

We used data from studies that generated genome-wide DNA methylation data using the Illumina 450K Infinium array platform. Figure 2 gives a schematic overview of the design of the current study and overall outcomes.

Dataset 1:

For childhood ALL, the disease-associated methylation signature was determined through compiling data available from the literature, and has been previously described [51]. The criteria for study inclusion were that patient blood/bone marrow samples (across any, and all subtypes of childhood ALL) were compared with control samples for genome-wide methylation analysis that included use of the Illumina 450K Infinium array platform. Using the search term 'childhood acute lymphoblastic leukaemia' and 'methylation' in PubMed, data from two key studies [42, 52] were compiled to develop a list of genes reported to have aberrant DNA methylation in individuals with disease compared to controls. Whilst Chatteron *et al*., utilised the Illumina 27k platform in their study, only probes also present on the 450K were included in analysis, therefore in order to be inclusive this study was included [52]. Data were compiled at the gene, rather than CpG level, using the gene symbol as the identifier.

Dataset 2:

For RA, we used data from our previous study, in which methylation was assessed in blood samples using the Illumina 450K Infinium array platform (GEO accession number: GSE110128) in a sub-group (n=99) from the Flemish Environment and Health Study I (FLEHSI) birth cohort, followed-up at age 11 years. More information on the study design and details of the recruitment protocols have been previously reported (supplementary materials of [53]). Information on the allergy status of the children was collected based on the ISAAC questionnaire [54]. Plasma samples of the 11-year-old

children were used to determine specific IgE sensitization status for a mix of airborne allergens using an ImmunoCAP Phadiatop test (Thermo Fisher) as previously described [40]. Based on the questionnaire data and the IgE sensitization status 22 respiratory allergic cases vs 38 controls were identified. Cases were considered to have a respiratory allergy if they reported (either self-reported or doctor's diagnosed) at least one respiratory allergy symptom (occurrence of asthma, hay fever, other types of rhinitis, wheezing, or runny nose, in the past year and ever; as questioned in accordance with the ISAAC questionnaire[54]) and Phadiatop IgE ≥ 0.35 kU/ L. Control subjects were those who did not report any (doctor's diagnosed) allergy symptoms, and Phadiatop and FX5 IgE < 0.35 kU/L.

Peripheral blood (mononuclear cell fraction) DNA methylation profiles of respiratory allergy cases as a total group, as well as per respiratory allergy subtypes (i.e. allergic asthma, rhinitis, hay fever), were compared to controls in order to identify differentially methylated gene regions (DMRs). Results were compiled as a list of disease-associated differentially methylated genes. A detailed description of the 450K data processing, normalisation and deferential methylation analysis used here has been described [40, 53] previously. Briefly, DMRs were identified with comb-p analysis [40], using the list of uncorrected p values for all CpG sites as calculated from the differential methylation analysis together with their chromosomal location. This generates a list of gene regions that are differently methylated, and an aggregated, adjusted p value is assigned to each region. A region with an adjusted p value < 0.05 was deemed differently methylated. Gene symbols were used as the identifier. The proportion of various cell types was included in the analysis, using the statistical deconvolution method described by Houseman *et al*. and implemented in "minfi" as the "estimateCellCountsMset" function [55].

In an aim to generate a more comprehensive disease-associated gene methylation signature we also searched the literature for relevant data. Using the search terms 'respiratory allergy' and 'DNA methylation' in PubMed, no studies matching our criteria were found to have used the Illumina 450K

Infinium array platform to examine methylation differences in blood samples between children with RA vs controls.

Exposure-to infection-associated methylation signature

We have previously identified genes which have altered DNA methylation (measured at age 7) in response to early life exposures used as proxy for infection i.e. day care attendance at 8 months (maternally reported via questionnaire at 8 months) and reported cold symptoms in the first 6 months of life (maternally reported via questionnaire at 15 months) [56]. Briefly, variation in genome-wide DNA methylation patterns were assessed in a sub-population of the Avon Longitudinal Study of Parents and Children (ALSPAC) cohort using the Illumina Infinium® HumanMethylation450k BeadChip assay. Multiple linear regression analyses were then used to investigate associations between DNA methylation at individual CpG sites and day care attendance at 8 months and reported cold symptoms in the first 6 months of life. DNA methylation was modelled as a continuous variable (outcome), in a multivariate regression model accounting for relevant potential confounders (sex, parity, gestation, cell type, and batch)[57].

Cell Composition

Cell composition is likely to be different between the disease states, and that in particular the epigenetic profile of ALL patients do not represent normal blood cells due to expansion of leukaemic progenitor cells (indeed ALL patient samples used in Nordlund et al's study consist of >80% leukemic blasts)[42]. For ALL studies utilised to generate the disease-associated methylation signature here, appropriate cell types from control subjects (i.e. a range of B, T cells or hematopoietic progenitor cells from which leukemic blasts are derived) were used as comparators to allow the distinction of lineage- and cell type-specific methylation differences [42, 52]. For the RA methylation signature,

and for the exposure to infection signatures assessed in 'healthy' individuals, cell types were corrected for using the Houseman correction [40, 51, 53]. Due to the clonal nature of ALL being derived from an individual blast compared to the multi-cellular response in RA we argue that the cell correction methods are appropriate to individual diseases to accurately reflect specific disease-associated methylation, rather than cell-type specific methylation. Any further correction for cell type heterogeneity between diseases would have likely remove a large portion of the disease signals we sought to compare. The methylation profiles utilised here therefore reflect specific disease-associated methylation and thus encapsulated the similarities and differences in methylation observed between disease phenotypes, rather than between cell types.

Comparison of disease-associated and exposure-to-infection-associated methylation signatures

Compiled lists of genes with a given disease-associated and/or exposure-to-infection- associated methylation signature were compared using online list comparison tool (http://jura.wi.mit.edu/bioc/tools/compare.php) with gene symbols as the common identifiers. Hypergeometric tests were carried out using the GeneProf calculator (http://www.geneprof.org) to assess the probability that the observed overlapping changes in methylation were likely to be significant ($p \leq 0.05$) and not due to a chance observation, with an n of 21231 as the constant population size i.e. number of Refseq gene symbol identifiers on the array.

Gene Ontology Enrichment and Pathway Analysis

DAVID [58] was used to carry out Gene Ontology (GO) enrichment analysis and to investigate KEGG pathways affected by disease specific DNA methylation signatures. The threshold for significance for Gene Ontology enrichment analysis was set at $p<0.05$ (corrected for multiple testing), and at $p<0.05$ (uncorrected) for KEGG pathway enrichment analysis. Uncorrected p values were utilised for KEGG

pathway analysis due to the more limited number of genes associated with KEGG pathways compared to those associated with GO processes.

Results

Common gene-associated methylation changes in RA and ALL

Disease-associated methylation lists contained gene identifiers for which differential methylation was reported in each disease using probes present on the Illumina 450K Infinium array platform. The ALL list comprised of 2173 genes, whereas the RA list had 838 genes (see Supplementary File 1 for full lists of genes associated with ALL-related methyl signature and RA-related methyl signatures). Comparison of these disease associated signatures, found 112 genes in common (Table 1 and Figure 2). A hypergeometric test suggests that the probability of this overlap is not due to chance ($p=0.0019$). Of these disease associated gene methylation changes, 82% (92/112) had the same direction of change i.e. hypermethylated (Table 1). As expected, the magnitude of hypermethylation was much larger in ALL (ranging from 20-60% increase, indicated by beta values) compared to the more subtle change in methylation observed in RA (ranging from 1-4% increase, indicated by beta values) (for full range of values see Supplementary Table 1).

Common GO Processes and KEGG Pathways between Disease-Associated methylation signatures

Gene ontology and KEGG pathway analysis were carried out for each of the individual disease-associated methylation signatures. GO analysis revealed 801 biological processes significantly enriched in the ALL-specific methylation signature (See Supplementary Table 2 in Supplementary File 1), whereas 5 biological processes were enriched in the RA-specific methylation signature (see Supplementary Table 3 in Supplementary File 1). Interestingly, all 5 biological processes were also present in the list of processes based on the ALL methylation signature (Table 2). In addition, the

same 5 biological processes were identified when the analysis was run using the 92 genes in common with the same direction of methylation change (data not shown).

Sixteen KEGG pathways were identified in the ALL-specific methylation signature (Table 3), whereas 22 KEGG pathways were associated with the RA-specific methylation signature (Table 4). The 'Hippo Signalling Pathway' (see Supplementary Figure 1 in Supplementary File 2) and 'Signalling pathways regulating pluripotency of stem cells' (see Supplementary Figure 2 in Supplementary File 2) pathways were identified in common (Tables 3 & 4).

Common gene-associated methylation changes between exposure-to-infection and disease-associated methylation signatures

We have previously reported variation in methylation in 60 and 11 gene loci in response to cold symptoms and day care attendance, respectively [56]. We reported an overlap between variation in methylation observed in response to these exposures and gene-associated differential methylation reported in ALL [56]. Briefly, 5 genes (*GLB1L, PRKAA2, PTPRD, SCT, TMEM132c*) were found to have altered methylation in response to day care attendance at 8 months and in ALL (p=0.0003 for hypergeometric probability). Whilst 9 genes (*ARHGEF4, GLRB, JAKMIP1, KCNK1, KCNQ1DN, LOC647309, NRXN2, SLC22a17, SMOC2*) were found to have altered methylation in response to reported cold-like symptoms in the first year of life and in childhood ALL, with this overlap possibly due to chance (p=0.1562 for hypergeometric probability). Of these, 6/14 genes displayed consistent hypermethylation in response to a proxy exposure of infection and in ALL (Table 5).

When our previously defined exposure-to-infection-associated signatures were compared with the RA-associated methylation signature, we found an overlap of 8 genes (*AGER, APLP2, HDAC4, LMF1, NXN, SLC22a17, SMOC2, TPPP*) whose methylation status was altered in response to reported cold-like symptoms in the first year of life and also had altered methylation associated with RA (Table 5).

Hypergeometric probability suggests this overlap is not likely to be due to chance (p=0.0024). Only one gene (*TCP11*) was found to have an altered methylation status in response to day care attendance and in association to RA, with this finding likely to be a chance finding (p=0.3579 for hypergeometric test)(Table 5). Of these, 5/9 of these genes displayed consistent hypermethylation in response to a proxy exposure of infection and in RA (Table 5).

Two genes, *SLC22a17* and *SMOC2,* were found to have altered methylation in relation to exposure-to-infection (i.e. reported cold-like symptoms in first year of life) and in both ALL and RA-associated methylation signatures (Figure 3); *SMOC2* showing consistent hypermethylation and *SLC22a17* showing the opposite directional change of differential methylation. Hypergeometric probability suggests finding 2 genes with altered methylation from a pool of 60 with altered methylation associated with reported cold like symptoms in the first year of life and 112 with altered methylation in both RA and ALL is not likely to be a chance finding (p=0.040).

Comparison of RA and ALL disease-associated methylation signatures and exposure to infection signatures with additional disease-associated methylation signatures

A comparison between any unrelated DNA methylation disease study may uncover significant overlaps between disease-associated methylation signatures. To test this we carried out additional analysis, including hypergeometric tests, comparing disease-associated methylation signatures from a range of disease phenotypes; systemic lupus erythematosus[59], rheumatoid arthritis[60], multiple sclerosis[61], myocardial infarction[62], type 2 diabetes (T2D)[63] and obesity[64] (see Supplementary Table 4 results). Seven of twelve comparisons resulted in either no gene overlaps or overlaps which statistically are likely to be due to chance. Therefore 5/12 overlaps statistically are unlikely to be due to chance. Several of the diseases investigated are autoimmune in origin. Allergies (including RA) have been hypothesised to be precursors of autoimmune diseases, we therefore anticipated, and found, significant overlaps between RA and some autoimmune diseases (i.e. RA and

lupus; RA and rheumatoid arthritis). An overlap was also observed between RA and obesity, which may be due to comorbidity of these two phenotypes [65]. Overlaps also were significant between ALL and rheumatoid arthritis and myocardial infarction. Common treatments which can drive epigenetic change may explain the commonality between ALL and rheumatoid arthritis i.e. methotrexate is used in treatment of both. Additionally, since CVD is a common late effect observed in childhood cancer survivors, a degree of overlap may also be anticipated between ALL and myocardial infarction.

In comparing these additional disease-associated methylation signatures with exposure to infection signatures we found no significant overlaps (Supplementary Table 5).

Discussion

DNA methylation is suggested as molecular mechanism in the development of RA and ALL [39-41]. We postulated that the diseases may have DNA methylation marks in common due to their suggested shared aetiology i.e. the hygiene hypothesis. To the best of our knowledge, and for the first time, we report a significant number of 112 common gene loci with altered DNA methylation in both childhood ALL and childhood RA. The majority (82%) of these methylation marks were hypermethylation, further substantiating the plausibility that changes in DNA methylation may be a common underlying mechanism in the development of both diseases. Whilst significant overlaps were also identified between RA and ALL associated methylation signatures and signatures for additionally tested disease types, not all comparisons between the studies resulted in significant overlaps (7/12 either not showing any or non-significant overlaps), and for those that did there is plausible reasoning for the commonalities observed (see results section for details). We would therefore reason that the observed commonalities in gene-methylation between RA and ALL could plausibly be due to a common aetiology such as the hygiene hypothesis.

A large number of disease-associated methylation marks in our study remained specific to either ALL or RA i.e. 2061 and 726 genes respectively. Given that the development of RA and ALL is not necessarily concomitant, this is expected, and is likely to be due to other factors (i.e. genetic and/or environmental) which further contribute to the development of these individual disease outcomes. Certain genetic aberrations are the first hit in the pathway towards ALL, which could also plausibly influence the observed disease-associated methylation pattern. It is reasonable to assume highly disease-specific methylation patterns will be acquired as a result of the individual diseases themselves (potentially via changes in cell composition as discussed in methods section above), rather than being causal in the development of disease.

In examining the biological processes and pathways underlying the individual disease-associated DNA methylation signatures, all of the biological processes enriched in the RA-associated methylation signature were also identified in the ALL-associated methylation signature. These overlapping processes were associated with biological or cellular adhesion. Given that adhesion is integral to the inflammatory process, aberrations of these processes are likely to be key in the pathology of respiratory allergy, but also could be important aetiologically in the development of leukaemia since adhesion to the stromal niche is crucial for leukaemic stem cells to support proliferation and chemoresistance [66]. Furthermore, two biological pathways, Hippo signalling pathway and Signalling pathways regulating pluripotency of stem cells, are likely to be affected by both RA and ALL-associated methylation signatures. As Hippo signalling plays a critical role in stem cell and progenitor cell self-renewal and expansion [67], these pathways are likely to be linked. Indeed these pathways share 50 genes between them (see Supplementary File 1 for list of genes), and both involve Wnt signalling (see Supplementary File 2 for Supplementary Figures 1 and 2). Due to the role of the Hippo signalling pathway in controlling cell proliferation and apoptosis, aberrations in this pathway have been associated with a variety of health outcomes, including carcinogenesis [68] and asthma and allergy [69]. Furthermore, a number of tumour suppressor and oncogenes are involved in this pathway, including oncogene YAP1 which has been reported to be downregulated in

haematological malignancies [70] and hypermethylated in ALL [41, 52]. In addition, GLI2, which was recently reported to be associated with the development of RA in childhood [40], is also involved in this Hippo signalling pathway (Table 4).

In order to investigate if DNA methylation may be a mediating mechanism between early life infection exposure and ALL and RA, we used a meet-in-the-middle approach to determine if exposure-associated methylation overlapped with disease-associated methylation. We found overlapping gene loci with differential methylation status in response to a proxy measure of infection and disease. Furthermore, and perhaps more convincingly in the context of the hygiene hypothesis, comparing additional disease-associated methylation signatures with exposure to infection signatures no overlaps were observed that may not have been due to chance (Supplementary Table 5). Therefore proxy exposures to infection only show significant overlap with the two disease phenotypes (RA and ALL) we have investigated in the context of the hygiene hypothesis. Five genes were observed to have both altered methylation associated with day care attendance and ALL, and 8 genes were observed to have both altered methylation associated with reported cold (<6 months of age) and in RA. Whilst there were also 9 genes observed to have both altered methylation associated with reported cold and ALL and one gene observed to have altered methylation associated with day care attendance and RA, hypergeometric tests suggest these observations may be due to chance. Whilst the use of hypergeometric tests gives a statistically based rationale to suggest whether the overlap between disease and exposure associated genes is a chance finding or likely to be statistically significant, it could be argued that a biological rationale is equally likely, i.e. any one given key methylation change may have a significant functional effect which may be significant in the biological disease pathway. As such, for further discussion of their biological relevance, all identified genes were included, regardless if the hypergeometric test was statistically significant or not. Since the hygiene hypothesis suggests that early life infection is

329 protective against the development of allergies and childhood leukaemia, it would follow that any
330 changes in methylation as a result of 'protective' infection exposure could also be protective.
331 Therefore, we would anticipate that to be 'protective', exposure related methylation change would
332 be in the opposing direction to that observed in disease itself. Here we report 8 (*GLRB, JAKMIP1,*
333 *KCNK1, KCNQ1DN, NRXN2, GLB1L, SLC22A17, TMEM132c*) and 4 (*APLP2, SLC22A17, TPPP, TCP11*)
334 genes respectively to have the opposing direction of methylation change (i.e. hypomethylation)
335 compared to methylation change (i.e. hypermethylation) observed in ALL or RA (Tables 5 & 6).
336 Investigation of the function of these genes in the context of the associated diseases did not
337 highlight an obvious specific functional role of the 4 associated genes in the development of RA.
338 However, *JAKMIP1* may be functional in the development of ALL as it has previously overexpression
339 in tumour tissue has been associated with activation of the Wnt/beta-catenin pathway and
340 subsequent promotion of cancer cell proliferation [71]. Among the genes showing opposing
341 direction of methylation change in relation to early life infection exposure and ALL and RA, there was
342 1 gene in common between the diseases; *SLC22A17*. *SLC22A17*, alternatively known as *NGALR*,
343 codes for the neutrophil gelatinase-associated lipocalin receptor, which binds NGAL – a member of
344 lipocalin protein family [72]. NGAL has been reported to play a significant role in generating innate
345 immune response and safeguards against bacterial infections by sequestering iron, while on the
346 other hand elevated NGAL levels have been observed in various diseases ranging from infections to
347 chronic inflammation [72, 73]. Therefore an altered methylation status of this gene may be
348 mediating early life exposure to infection and subsequent risk of diseases linked to the hygiene
349 hypothesis. This may therefore be a candidate biomarker for investigation in future studies. Whilst
350 the functional significance of changes in DNA methylation of most of the genes identified as having a
351 potentially protective role in the development of ALL or RA may not be clear, it still remains plausible
352 that changes in methylation of these loci via exposure to infection may modulate disease risk
353 through yet unknown mechanisms/functions.

More changes in methylation in response to exposure to infection and related to ALL/RA are likely to exist, but could not be identified in this study. The primary reason for this is likely to be due to the use of proxy measures i.e. day care attendance and reported cold symptoms, to investigate methylation change in relation to infection rather than measures of specific infections or number of reported infections. Hence, changes in DNA methylation attributable to specific infections, high rates of infection or duration of infection, remain undetectable. Given the difficulty in obtaining data, particularly accurate measures relating to infection, the use of proxy measures is justifiable for proof of principle studies such as this. However, improved understanding of the impact of different types of infection on DNA methylation patterns is warranted in order to realise how exposure to infection is able to modulate the risk of developing childhood ALL and RA. To remove the complexities and confounding structures of human studies, controlled *in vitro* studies will be key to fully attribute and understand variation in DNA methylation in response to a given exposure to infection.

Conclusions

A meet-in-the-middle approach was used to investigate if DNA methylation may be a mediating mechanism in the hygiene hypothesis. We suggest a significant overlap in gene methylation marks exist between RA and childhood ALL, the two main diseases for which the hygiene hypothesis is suggested to be causal. Given that changes in methylation are thought to contribute to the development of these diseases, this significant overlap is suggestive of a biological mechanism through which exposure to infection may be aetiologically linked to both disease outcomes. Perhaps more importantly we observed significant overlaps between methylation changes associated with exposure to infection in early life and disease-associated methylation providing additional support for DNA methylation as a plausible mediating mechanism between exposure to infection and ALL and RA risk. Whilst these findings may be indicative our theory is plausible, further evidence will be required to provide proof of our hypothesis. In order to better understand the role of early life exposure to infection in the development of these diseases, further studies are needed to fully

elucidate which type of infections may be protective and which biological pathways and functions they influence in order to modulate disease risk, including more in depth investigation of the role of DNA methylation.

Future Perspective

It is plausible that diseases suggested to have a common aetiology will share underlying mechanisms and/or pathways towards development. DNA methylation, as an important regulator of gene expression and mechanism by which environment can orchestrate such regulation, is likely to be a key player. Where genome-wide association studies have provided genetic variants in common across seemingly unrelated diseases, epigenome-wide association studies may generate evidence for common methylation marks between diseases with shared environmental causes. We used a bioinformatics approach to provide evidence that a significant number of genes have altered methylation in common between respiratory allergy and acute lymphoblastic leukaemia for which the hygiene hypothesis has been implicated. In addition, we suggest that exposure to infections in early life may be linked to methylation at disease-associated loci. This early evidence needs to be further substantiated by follow-up studies.

Furthermore, for theories such as the hygiene hypothesis where no individual exposure is postulated to be causal, and the underlying mechanisms across multiple associated diseases not well understood, similar studies comparing other plausible mechanisms (i.e. other epigenetic factors, differences in microbiome etc.) across multiple diseases may aid understanding of these linked non-communicable diseases.

Summary Points

- We report a significant number (112) of common gene loci with altered DNA methylation in both childhood acute lymphoblastic leukaemia (ALL) and childhood respiratory allergy (RA).
- Most common gene loci displayed hypermethylation in both diseases (n=92).
- Investigating the biological processes underlying individual disease-associated DNA methylation signatures, uncovered that biological or cellular adhesion might be among the common key events in the pathways towards RA and ALL development.
- For RA, adhesion may be key to its pathology via inflammatory processes, but adhesion may also be important in ALL development in supporting proliferation and chemoresistance via adhesion to the stromal niche.
- Hippo signalling pathway and signalling pathways regulating pluripotency of stem cells, were identified in both RA and ALL-associated methylation signatures. Hippo signalling controls cell proliferation and apoptosis, with aberrations in this pathway being associated with carcinogenesis and asthma and allergy.
- Day care attendance was associated with variation of methylation in genes associated with ALL.
- Cold symptoms during early life were associated with methylation variation in genes associated with RA.
- Methylation of the *SLC22A17* gene, alternatively known as NGALR, was altered in both RA, ALL and in response to cold symptoms in early life. In binding NGAL, which plays a role in the innate immunity, NGALR may indirectly influence the innate immune response. Therefore methylation at of this gene may be a target locus mediating early life exposure to infection and subsequent risk of diseases linked to the hygiene hypothesis.

Figure 1. Proposed model depicting the mediating role of DNA methylation between early life exposures and development of RA and childhood ALL. A range of exposures to infection and/or immune stimuli, alter DNA methylation patterns, some of which are associated with either disease outcome. Where there is a common aetiological exposure between these diseases, (i.e. for the hygiene hypothesis and RA and ALL) some exposure-related methylation changes will be common to both diseases, whilst other disease specific DNA methylation changes may determine disease outcome.

Figure 2. Schematic overview of study design and outcome.

Figure 3. Venn diagram showing number of overlapping genes with differential methylation associated with combined reported proxy exposures of infection and in RA and ALL-associated methylation signatures. Hypergeometric probability suggests finding 2 genes with altered methylation from a pool of 71 with altered methylation associated with proxy exposure to infection and 112 with altered methylation in both RA and ALL is not due to chance (p=0.029).

Supplementary Figure 1. Modified KEGG Hippo Signaling Pathway showing genes with altered methylation in RA and ALL disease associated methylation signatures and exposure to infection (i.e. cold) methylation signatures.

Supplementary Figure 2. Modified KEGG Signaling Pathways Regulating Pluripotency of stem cells showing genes with altered methylation in RA and ALL disease associated methylation signatures and exposure to infection (i.e. cold) methylation signatures.

451 **Table 1.** Common genes with altered methylation in ALL and RA compared to healthy controls.

Gene ID	Hyper (↑)/ Hypomethylated (↓) in ALL	Hyper (↑)/ Hypomethylated (↓) in RA	Gene ID	Hyper (↑)/ Hypomethylated (↓) in ALL	Hyper (↑)/ Hypomethylated (↓) in RA
ABCA4	↑	↑	*PCDHA5*	↑	↑
ADAMTS15	↑	↑	*PCDHA6*	↑	↑
ADARB2	↓	↑	*PCDHA7*	↑	↑
ADRA2C	↑	↑	*PCDHGA1*	↑	↑
BLK	↓	↑	*PCDHGA10*	↑	↑
BMP7	↑	↑	*PCDHGA11*	↑	↑
CCDC85C	↑	↑	*PCDHGA12*	↑	↑
CELSR3	↑	↑	*PCDHGA2*	↑	↑
CNTN4	↑	↑	*PCDHGA3*	↑	↑
COL23A1	↑	↑	*PCDHGA4*	↑	↑
CPLX1	↑	↑	*PCDHGA5*	↑	↑
CTBP1	↑	↑	*PCDHGA6*	↑	↑
DDR1	↓	↑	*PCDHGA7*	↑	↑
DLK1	↑	↑	*PCDHGA8*	↑	↑
DOK6	↑	↑	*PCDHGA9*	↑	↑
DPP6	↑	↑	*PCDHGB1*	↑	↑
EGFLAM	↑	↑	*PCDHGB2*	↑	↑
EVI5L	↑	↑	*PCDHGB3*	↑	↑
F10	↑	↑	*PCDHGB4*	↑	↑
FAM110B	↑	↑	*PCDHGB5*	↑	↑
FBXO27	↑	↑	*PCDHGB6*	↑	↑
FSCN2	↑	↑	*PCDHGB7*	↑	↑
GABRB3	↑	↑	*PCSK1*	↑	↑
GABRG3	↑	↑	*PDE4C*	↑	↑
GALNT9	↑	↑	*PER3*	↑	↑
GCNT2	↑	↑	*PRDM16*	↑	↑
HHIPL1	↑	↑	*PTCHD3*	↑	↑
IGF2	↑	↑	*PTPRN2*	↑/↓	↑
IHH	↑	↑	*RAP1GAP*	↑	↑
INS-IGF2	↑	↑	*RGS12*	↓	↑
INTS1	↓	↑	*RNF126*	↓	↑
KCNE1	↑	↓	*SDK1*	↑	↑
KLB	↑	↑	*SGCD*	↑	↑
KRT18	↑	↑	*SHC2*	↑	↑
LCK	↓	↑	*SLC22A17*	↑	↑
LDHD	↑	↑	*SLC25A2*	↑	↑
LMO3	↑	↑	*SLC9A3*	↑	↑
LOC650226	↑	↑	*SMOC2*	↑	↑
LRP5	↓	↑	*SORCS2*	↑	↑
LRRC4C	↑	↑	*SOX8*	↑	↑
MAD1L1	↓	↑	*SPOCK1*	↑	↑
MGMT	↓	↑	*SPTBN2*	↓	↑
MLPH	↑	↑	*TACR3*	↑	↑
MSI2	↓	↑	*TBCD*	↑/↓	↑
MTUS2	↑	↑	*TBX4*	↑	↑

NADSYN1	↓	↑	***TCERG1L***	**↑**	**↑**
NFE2	**↑**	**↑**	***TDH***	**↑**	**↑**
NPFFR2	**↑**	**↑**	***TFAP2A***	**↑**	**↑**
NPTX2	**↑**	**↑**	***TMC2***	**↑**	**↑**
NRXN1	**↑**	**↑**	***TMEM161A***	**↑**	**↑**
NTM	**↑**	**↑**	***TNXB***	**↓**	↑
OSBPL5	↓	↑	***TRIM61***	↑	↓
PCDHA1	**↑**	**↑**	***TTLL7***	**↑**	**↑**
PCDHA2	**↑**	**↑**	***WNK4***	**↑**	**↑**
PCDHA3	**↑**	**↑**	***ZC3H3***	**↓**	↑
PCDHA4	**↑**	**↑**	***ZNF536***	**↑**	**↑**

↑ = hypermethylation; ↓ = hypomethylation; Where both ↑/↓ is stated, individual CpGs sites were found to have differential methylation within the same gene-associated loci between studies. Where the direction of change is common between diseases arrows are shown in bold.

Table 2. Biological Processes which may be affected through altered gene methylation present in ALL and RA.

GO ID	GO term	Number of genes included in process with altered methylation in ALL	Corrected P value for ALL (Benjamini)	Number of gene included in process with altered methylation in RA	Corrected P value for RA (Benjamini)	Number of overlapping genes between ALL and RA	Overlapping genes
GO:0007156	homophilic cell adhesion via plasma membrane adhesion molecules	76	1.06E-29	35	1.14E-13	25	*CELSR3, PCDHA1, PCDHA2, PCDHA3, PCDHA4, PCDHA5, PCDHA6, PCDHA7, PCDHGA1, PCDHGA2, PCDHGA3, PCDHGA4, PCDHGA5, PCDHGA6, PCDHGA7, PCDHGA8, PCDHGA9, PCDHGB1, PCDHGB2, PCDHGB3, PCDHGB4, PCDHGB5, PCDHGB6, PCDHGB7, SDK1*
GO:0098742	cell-cell adhesion via plasma-membrane adhesion molecules	90	8.32E-28	39	6.05E-12	29	*CELSR3, NRXN1, PCDHA1, PCDHA2, PCDHA3, PCDHA4, PCDHA5, PCDHA6, PCDHA7, PCDHGA1, PCDHGA10, PCDHGA11, PCDHGA12, PCDHGA2, PCDHGA3, PCDHGA4, PCDHGA5, PCDHGA6, PCDHGA7, PCDHGA8, PCDHGA9, PCDHGB1, PCDHGB2, PCDHGB3, PCDHGB4, PCDHGB5, PCDHGB6, PCDHGB7, SDK1*
GO:0007155	cell adhesion	260	8.74E-10	120	7.01E-10	44	*BMP7, CELSR3, CNTN4, DDR1, EGFLAM, GCNT2, IGF2, IHH, KRT18, LCK, MAD1L1, NRXN1, PCDHA1, PCDHA2, PCDHA3, PCDHA4, PCDHA5, PCDHA6, PCDHA7, PCDHGA1, PCDHGA10, PCDHGA11, PCDHGA12, PCDHGA2, PCDHGA3, PCDHGA4, PCDHGA5, PCDHGA6, PCDHGA7, PCDHGA8, PCDHGA9, PCDHGB1, PCDHGB2, PCDHGB3, PCDHGB4, PCDHGB5, PCDHGB6, PCDHGB7, SDK1, SMOC2, SPOCK1, SPTBN2, TBCD, TNXB*
							BMP7, CELSR3, CNTN4, DDR1, EGFLAM, GCNT2, IGF2, IHH, KRT18, LCK, MAD1L1, NRXN1, PCDHA1, PCDHA2, PCDHA3, PCDHA4, PCDHA5, PCDHA6, PCDHA7, PCDHGA1, PCDHGA10, PCDHGA11, PCDHGA12, PCDHGA2, PCDHGA3, PCDHGA4,

GO:0022610	biological adhesion	260	1.27E-09	120	6.84E-10	44	*PCDHGA5, PCDHGA6, PCDHGA7, PCDHGA8, PCDHGA9, PCDHGB1, PCDHGB2, PCDHGB3, PCDHGB4, PCDHGB5, PCDHGB6, PCDHGB7, SDK1, SMOC2, SPOCK1, SPTBN2, TBCD, TNXB*
GO:0098609	cell-cell adhesion	170	4.35E-05	91	2.84E-09	38	*BMP7, CELSR3, CNTN4, GCNT2, IGF2, IHH, KRT18, LCK, MAD1L1, NRXN1, PCDHA1, PCDHA2, PCDHA3, PCDHA4, PCDHA5, PCDHA6, PCDHA7, PCDHGA1, PCDHGA10, PCDHGA11, PCDHGA12, PCDHGA2, PCDHGA3, PCDHGA4, PCDHGA5, PCDHGA6, PCDHGA7, PCDHGA8, PCDHGA9, PCDHGB1, PCDHGB2, PCDHGB3, PCDHGB4, PCDHGB5, PCDHGB6, PCDHGB7, SDK1, SPTBN2*

Table 3. KEGG pathways potentially affected by the ALL-associated methylation signature.

KEGG Pathway Term	KEGG Pathway Name	Total number of genes on pathway	Number of genes affected by ALL-associated methylation signature	Genes affected	P value (uncorrected)	P value (FDR corrected)
hsa04080	Neuroactive ligand-receptor interaction	277	96	*GPR83, OPRM1, MCHR2, GRIK1, TACR3, THRB, GLRA1, GABRB3, GRIK2, GABRB2, GRIK3, TACR1, LEPR, GLRA3, GABRB1, GRIK4, LHCGR, GABBR2, GRIN3A, SCTR, EDNRA, EDNRB, AGTR1, HTR1B, HTR1A, GRIN2B, GALR1, GRIN2C, GRIN2D, GALR3, S1PR5, GRID2, CHRNA4, CALCRL, HTR5A, GRID1, HTR1E, GABRG2, GABRG3, CCKBR, RXFP3, GRIN2A, HTR4, GRM1, NTSR2, SSTR4, GRM5, GRM3, ADRB1, SSTR1, GRM8, CHRM2, GRM7, GRM6, GIPR, PTAFR, CTSG, CALCR, DRD1, PRLHR, DRD3, DRD2, NPY2R, DRD4, OXTR, FPR2, HCRTR2, ADRB3, P2RY2, PRSS3, NPFFR2, ADRA2A, ADRA2C, NPFFR1, GABRA2, GLRB, GABRA1, GABRA4, PTH2R, GRIN1, GABRA5, GRIA4, NPY1R, PTGFR, NPY5R, LEP, GRIA2, P2RX2, MTNR1B, AVPR1A, ADRA1A, CHRNB1, NMBR, GHSR, ADRA1D, MTNR1A*	3.34E-31	4.40E-28
hsa05033	Nicotine addiction	40	22	*GABRG2, GABRA2, GABRG3, GABRA1, GABRB3, GABRA4, GABRB2, GABRB1, GRIN1, GABRA5, GRIN2A, GRIN3A, GRIA4, SLC17A7, SLC17A6, GRIN2B, GRIA2, GRIN2C, GRIN2D, CHRNA4, CACNA1A, CACNA1B*	6.83E-12	9.00E-09
hsa04020	Calcium signaling pathway	179	47	*GNA14, ADCY1, DRD1, ADCY2, ERBB4, TACR3, ADCY8, TACR1, LHCGR, OXTR, EDNRA, AGTR1, EDNRB, ADRB3, GRIN2C, PDE1C, GRIN2D, PLCD3, HTR5A, EGFR, SLC8A2, NOS1, CCKBR, GRIN1, GRIN2A, HTR4, PRKCG, PTGFR, GRM1, VDAC3, GRM5, GNAL, PLCE1, ADRB1, CHRM2, P2RX2, RYR1, AVPR1A, RYR2, ADRA1A, CACNA1E, GNAS, CACNA1D, CACNA1A, PTAFR, ADRA1D, CACNA1B*	2.02E-10	2.66E-07

hsa05032	Morphine addiction	91	30	*OPRM1, ADCY1, DRD1, ADCY2, GABRB3, GABRB2, ADCY8, GABRB1, PDE11A, GABBR2, KCNJ3, PDE1C, PDE4A, PDE4B, GNG4, GABRG2, GABRG3, GABRA2, GNAO1, GABRA1, GABRA4, GABRA5, PDE10A, PRKCG, PDE4C, KCNJ6, PDE2A, GNAS, CACNA1A, CACNA1B*	2.47E-09	3.26E-06
hsa04724	Glutamatergic synapse	114	34	*ADCY1, ADCY2, GRIK1, ADCY8, GRIK2, GRIK3, GRIK4, GRIN3A, KCNJ3, GLS2, GRIN2B, GRIN2C, GRIN2D, GNG4, SLC1A1, TRPC1, GNAO1, GRIN1, GRIN2A, PRKCG, GRIA4, SHANK1, GRM1, SLC17A7, GRM5, GRM3, SLC17A6, GRIA2, GRM8, GRM7, GRM6, GNAS, CACNA1D, CACNA1A*	3.12E-09	4.11E-06
hsa04723	Retrograde endocannabinoid signaling	101	29	*ADCY1, ADCY2, PTGS2, GABRB3, ADCY8, GABRB2, GABRB1, RIMS1, KCNJ3, GNG4, GABRG2, GABRA2, GABRG3, GNAO1, GABRA1, GABRA4, GABRA5, PRKCG, GRIA4, GRM1, SLC17A7, GRM5, SLC17A6, NAPEPLD, KCNJ6, GRIA2, CACNA1D, CACNA1A, CACNA1B*	1.37E-07	1.81E-04
hsa04024	cAMP signaling pathway	198	41	*ADCY1, DRD1, ADCY2, ADCY8, DRD2, OXTR, GABBR2, CNGB1, GRIN3A, SOX9, EDNRA, BDNF, HTR1B, HTR1A, GRIN2B, GRIN2C, PDE4A, GRIN2D, PDE4B, HCN4, HTR1E, HCN2, GRIN1, GRIN2A, HTR4, CFTR, PDE4C, GRIA4, NPY1R, CNGA3, AMH, PLCE1, ADRB1, GRIA2, SSTR1, CHRM2, GIPR, RYR2, GNAS, GHSR, CACNA1D*	3.31E-06	0.004
hsa04713	Circadian entrainment	95	25	*ADCY1, ADCY2, ADCY8, PRKG2, KCNJ3, GRIN2B, GRIN2C, GRIN2D, GUCY1A2, PER3, GNG4, NOS1, GNAO1, GRIN1, GRIN2A, PRKCG, GRIA4, KCNJ6, GRIA2, MTNR1B, RYR1, RYR2, GNAS, CACNA1D, MTNR1A*	6.48E-06	0.009
hsa04727	GABAergic synapse	85	22	*GABRG2, GABRA2, ADCY1, GABRG3, GABRA1, ADCY2, GNAO1, GABRB3, SLC6A1, GABRA4, ADCY8, GABRB2, GABRB1, GABRA5, PRKCG, GABBR2, GLS2, KCNJ6, GNG4, CACNA1D, CACNA1A, CACNA1B*	3.51E-05	0.046
hsa04950	Maturity onset diabetes of the young	26	11	*FOXA2, ONECUT1, GCK, RFX6, SLC2A2, PKLR, MNX1, PAX6, NEUROD1, PDX1, NKX6-1*	7.79E-05	0.103
hsa05030	Cocaine addiction	49	15	*CDK5R1, DRD1, DRD2, SLC6A3, GRIN1, GRIN2A, GRIN3A, GRM3, BDNF, GRIN2B, GRIA2, GRIN2C, GRIN2D, SLC18A2, GNAS*	1.34E-04	0.177

hsa04726	Serotonergic synapse	111	24	*TRPC1, GNAO1, GABRB3, PTGS2, GABRB2, GABRB1, SLC6A4, HTR4, PRKCG, KCNJ3, CYP4X1, HTR1B, HTR1A, KCNJ6, KCNN2, SLC18A2, GNAS, GNG4, CACNA1D, CACNA1A, HTR5A, HTR3D, HTR1E, CACNA1B*	2.75E-04	0.362
hsa04728	Dopaminergic synapse	128	26	*DRD1, PPP2R3A, CALY, DRD3, DRD2, SLC6A3, DRD4, COMT, KCNJ3, GRIN2B, PPP2R2B, GNG4, PPP2R2C, GNAO1, KIF5C, GRIN2A, PRKCG, GRIA4, GNAL, KCNJ6, GRIA2, SLC18A2, GNAS, CACNA1D, CACNA1A, CACNA1B*	3.96E-04	0.521
hsa04550	**Signaling pathways regulating pluripotency of stem cells**	**140**	**25**	***FGFR2, WNT5A, WNT5B, ONECUT1, APC2, WNT3A, SOX2, PAX6, WNT2, HAND1, LHX5, WNT6, TCF3, FGF2, FZD9, TBX3, NODAL, MYF5, NEUROG1, FZD2, ISL1, FZD7, DLX5, ID4, WNT7A***	**0.003**	4.303
hsa04540	Gap junction	88	18	*GJD2, EGFR, DRD1, ADCY1, ADCY2, TUBB2B, ADCY8, DRD2, GJA1, PRKCG, PRKG2, GRM1, GRM5, TJP1, ADRB1, GUCY1A2, PDGFC, GNAS*	0.004	4.708
hsa04390	**Hippo signaling pathway**	**151**	**26**	***WNT5A, WNT5B, APC2, WNT3A, SOX2, WWC1, CDH1, WNT2, LIMD1, YAP1, PPP2R2B, WNT6, PPP2R2C, DLG2, FZD9, TEAD1, TEAD2, FZD2, WWTR1, FZD7, TP73, CTNNA2, AMH, BMP7, WNT7A, BMP8B***	**0.004**	5.707

Pathways highlighted using bold font are also potentially affected by the RA-associated methylation signature (underlined genes are common between data sets).

Table 4. KEGG pathways potentially affected by the RA-associated methylation signature.

KEGG Pathway Term	KEGG Pathway Name	Total number of genes on pathway	Number of genes affected by RA-associated methylation signature	Genes affected	P value (uncorrected)	P value (FDR corrected)
hsa05310	Asthma	30	7	*CCL11, IL4, HLA-DPA1, HLA-DPB1, HLA-DMB, HLA-DOA, HLA-DMA*	6.84E-04	0.879
hsa05416	Viral myocarditis	57	9	*CCND1, SGCD, HLA-DPA1, ITGB2, MYH6, HLA-DPB1, HLA-DMB, HLA-DOA, HLA-DMA*	0.001	1.395
hsa05152	Tuberculosis	177	16	*TCIRG1, MRC2, CREBBP, ITGB2, HLA-DMB, HLA-DMA, SRC, AKT1, ITGAX, JAK2, HLA-DPA1, CALML5, HLA-DPB1, HLA-DOA, KSR1, AKT3*	0.002	2.895
hsa04672	Intestinal immune network for IgA production	47	7	*IL4, LTBR, HLA-DPA1, HLA-DPB1, HLA-DMB, HLA-DOA, HLA-DMA*	0.007	9.042
hsa05321	Inflammatory bowel disease	64	8	*IL4, SMAD3, HLA-DPA1, HLA-DPB1, HLA-DMB, HLA-DOA, HLA-DMA, NFATC1*	0.009	11.063
hsa05330	Allograft rejection	37	6	*IL4, HLA-DPA1, HLA-DPB1, HLA-DMB, HLA-DOA, HLA-DMA*	0.011	13.404
hsa05164	Influenza A	174	14	*XPO1, CREBBP, DDX39B, HLA-DMB, HLA-DMA, FURIN, AKT1, IRF7, JAK2, HLA-DPA1, IRF3, HLA-DPB1, HLA-DOA, AKT3*	0.012	14.734
hsa05166	HTLV-I infection	256	18	*XPO1, LTBR, CREBBP, SMAD3, ITGB2, TRRAP, HLA-DMB, HLA-DMA, DVL1, AKT1, CCND1, LCK, HLA-DPA1, HLA-DPB1, WNT9A, HLA-DOA, AKT3, NFATC1*	0.014	16.629
hsa04550	**Signaling pathways regulating pluripotency of stem cells**	**140**	**12**	***AKT1, FGFR4, FGFR3, HNF1A, PCGF3, JARID2, SMAD3, JAK2, WNT9A, AKT3, AXIN1, DVL1***	**0.014**	17.202
hsa05140	Leishmaniasis	71	8	*IL4, HLA-DPA1, ITGB2, JAK2, HLA-DPB1, HLA-DMB, HLA-DOA, HLA-DMA*	0.016	18.336

hsa04917	Prolactin signaling pathway	71	8	*AKT1, CCND1, TNFRSF11A, ESR1, JAK2, SHC2, AKT3, SRC*	0.016	18.336
hsa05323	Rheumatoid arthritis	88	9	*TCIRG1, TNFRSF11A, ATP6V1G2, HLA-DPA1, ITGB2, HLA-DPB1, HLA-DMB, HLA-DOA, HLA-DMA*	0.016	18.640
hsa04340	Hedgehog signaling pathway	27	5	*CSNK1D, CSNK1G2, GLI2, GLI3, IHH*	0.016	19.326
hsa05168	Herpes simplex infection	183	14	*TRAF2, TAF4, CREBBP, HLA-DMB, HLA-DMA, TAP2, IRF7, SRSF8, JAK2, HLA-DPA1, IRF3, HLA-DPB1, PER3, HLA-DOA*	0.018	20.951
hsa04940	Type I diabetes mellitus	42	6	*PTPRN2, HLA-DPA1, HLA-DPB1, HLA-DMB, HLA-DOA, HLA-DMA*	0.019	21.563
hsa04390	**Hippo signaling pathway**	**151**	**12**	***PRKCZ, CCND1, CSNK1D, SMAD3, ITGB2, WNT9A, GLI2, SNAI2, BMP7, AXIN1, FRMD1, DVL1***	**0.024**	27.045
hsa04330	Notch signaling pathway	48	6	*NOTCH1, CTBP1, CTBP2, CREBBP, LFNG, DVL1*	0.031	33.759
hsa05332	Graft-versus-host disease	33	5	*HLA-DPA1, HLA-DPB1, HLA-DMB, HLA-DOA, HLA-DMA*	0.032	34.624
hsa04152	AMPK signaling pathway	122	10	*AKT1, SREBF1, CCND1, AKT1S1, PFKFB3, TSC2, FASN, EEF2, AKT3, RPTOR*	0.037	38.307
hsa04514	Cell adhesion molecules (CAMs)	142	11	*CLDN19, LRRC4B, ESAM, HLA-DPA1, ITGB2, NRXN1, HLA-DPB1, LRRC4C, HLA-DMB, HLA-DOA, HLA-DMA*	0.038	39.005
hsa05320	Autoimmune thyroid disease	52	6	*IL4, HLA-DPA1, HLA-DPB1, HLA-DMB, HLA-DOA, HLA-DMA*	0.042	42.806
hsa05150	Staphylococcus aureus infection	54	6	*HLA-DPA1, ITGB2, HLA-DPB1, HLA-DMB, HLA-DOA, HLA-DMA*	0.049	47.437

Pathways highlighted using bold font are also potentially affected by the ALL-associated methylation signature (underlined genes are common between data sets).

Table 5. Direction of methylation change of genes associated with infection exposure and in ALL or RA.

Gene Symbol	Gene Name	Associated Proxy Exposure of Infection	Direction of methylation changes associated with exposure	Direction of methylation changes associated with ALL	Direction of methylation changes associated with RA
ARHGEF4	Rho Guanine Nucleotide Exchange Factor 4	Cold	↑	↑	-
GLRB	Glycine Receptor Beta	Cold	↓	↑	-
JAKMIP1	Janus Kinase And Microtubule Interacting Protein 1	Cold	↓	↑	-
KCNK1	Potassium Two Pore Domain Channel Subfamily K Member 1	Cold	↓	↑	-
KCNQ1DN	KCNQ1 Downstream Neighbor	Cold	↓	↑	-
LOC647309	-	Cold	↑	↑	-
NRXN2	Neurexin 2	Cold	↓	↑	-
SLC22a17	Solute Carrier Family 22 Member 17	Cold	↓	↑	↑
SMOC2	SPARC Related Modular Calcium Binding 2	Cold	↑	↑	↑
GLB1L	Galactosidase Beta 1 Like	Day care	↓	↑	-
PRKAA2	Protein Kinase AMP-Activated Catalytic Subunit Alpha 2	Day care	↑	↑	-
PTPRD	Protein Tyrosine Phosphatase, Receptor Type D	Day care	↑	↑	-
SCT	Secretin	Day care	↑	↑	-
TMEM132c	Transmembrane Protein 132C	Day care	↓	↑	-
AGER	Advanced Glycosylation End-Product Specific Receptor	Cold	↑	-	↑
APLP2	Amyloid Beta Precursor Like Protein 2	Cold	↓	-	↑

HDAC4	Histone Deacetylase 4	Cold	↑	-	↑
LMF1	Lipase Maturation Factor 1	Cold	↑	-	↑
NXN	Nucleoredoxin	Cold	↑	-	↑
TPPP	Tubulin Polymerization Promoting Protein	Cold	**↓**	**-**	**↑**
TCP11	T-Complex 11	Day care	**↓**	**-**	**↑**

↑ = hypermethylation; ↓ = hypomethylation; Cold = reported cold symptoms (<age 6 month); Day care = Day care attendance at 8 month = Day Care.

Arrows are shown in bold when the direction of methylation is opposing between protective exposure and disease associated methylation change.

References

1. Asher MI, Montefort S, Bjorksten B *et al*. Worldwide time trends in the prevalence of symptoms of asthma, allergic rhinoconjunctivitis, and eczema in childhood: ISAAC Phases One and Three repeat multicountry cross-sectional surveys. *Lancet* 368(9537), 733-743 (2006).
2. R P, Gw C, St H, R L. *WAO White Book on Allergy 2011–2012: Executive Summary.* WAO, Milawaukee. (2011-2012).
3. Society. AC. Cancer Facts and Figures 2014. 2017(6.11.2017), (2014).
4. Strachan DP. Hay fever, hygiene, and household size. *Bmj* 299(6710), 1259-1260 (1989).
5. Strachan DP. Family size, infection and atopy: the first decade of the "hygiene hypothesis". *Thorax* 55 Suppl 1 S2-10 (2000). *Of interest. Describes the introduction of the hygiene hypothesis and its importance in atopic and allergic diseases.
6. Bernsen RM, De Jongste JC, Van Der Wouden JC. Birth order and sibship size as independent risk factors for asthma, allergy, and eczema. *Pediatr Allergy Immunol* 14(6), 464-469 (2003).
7. Lewis SA, Britton JR. Consistent effects of high socioeconomic status and low birth order, and the modifying effect of maternal smoking on the risk of allergic disease during childhood. *Respir Med* 92(10), 1237-1244 (1998).
8. Matricardi PM, Franzinelli F, Franco A *et al*. Sibship size, birth order, and atopy in 11,371 Italian young men. *J Allergy Clin Immunol* 101(4 Pt 1), 439-444 (1998).
9. Strachan DP, Taylor EM, Carpenter RG. Family structure, neonatal infection, and hay fever in adolescence. *Arch Dis Child* 74(5), 422-426 (1996).
10. Westergaard T, Rostgaard K, Wohlfahrt J, Andersen PK, Aaby P, Melbye M. Sibship characteristics and risk of allergic rhinitis and asthma. *Am J Epidemiol* 162(2), 125-132 (2005).
11. Kurosaka F, Terada T, Tanaka A *et al*. Risk factors for wheezing, eczema and rhinoconjunctivitis in the previous 12 months among six-year-old children in Himeji City, Japan: food allergy, older siblings, day-care attendance and parental allergy history. *Allergol Int* 60(3), 317-330 (2011).
12. Haby MM, Marks GB, Peat JK, Leeder SR. Daycare attendance before the age of two protects against atopy in preschool age children. *Pediatr Pulmonol* 30(5), 377-384 (2000).
13. Kramer U, Heinrich J, Wjst M, Wichmann HE. Age of entry to day nursery and allergy in later childhood. *Lancet* 353(9151), 450-454 (1999).
14. Nicolaou NC, Simpson A, Lowe LA, Murray CS, Woodcock A, Custovic A. Day-care attendance, position in sibship, and early childhood wheezing: a population-based birth cohort study. *J Allergy Clin Immunol* 122(3), 500-506.e505 (2008).
15. Cheng G, Smith AM, Levin L *et al*. Duration of day care attendance during infancy predicts asthma at the age of seven: the Cincinnati Childhood Allergy and Air Pollution Study. *Clin Exp Allergy* 44(10), 1274-1281 (2014).
16. Gabet S, Just J, Couderc R, Seta N, Momas I. Allergic sensitisation in early childhood: Patterns and related factors in PARIS birth cohort. *Int J Hyg Environ Health* 219(8), 792-800 (2016).
17. Coppa GV, Zampini L, Galeazzi T, Gabrielli O. Prebiotics in human milk: a review. *Dig Liver Dis* 38 Suppl 2 S291-294 (2006).
18. Svanborg C, Agerstam H, Aronson A *et al*. HAMLET kills tumor cells by an apoptosis-like mechanism--cellular, molecular, and therapeutic aspects. *Adv Cancer Res* 88 1-29 (2003).
19. Benno Y, Sawada K, Mitsuoka T. The intestinal microflora of infants: composition of fecal flora in breast-fed and bottle-fed infants. *Microbiol Immunol* 28(9), 975-986 (1984).
20. Bode L, Jantscher-Krenn E. Structure-function relationships of human milk oligosaccharides. *Adv Nutr* 3(3), 383s-391s (2012).
21. Brown P. Treatment of infant leukemias: challenge and promise. *Hematology Am Soc Hematol Educ Program* 2013 596-600 (2013).

22. Hawkes JS, Neumann MA, Gibson RA. The effect of breast feeding on lymphocyte subpopulations in healthy term infants at 6 months of age. *Pediatr Res* 45(5 Pt 1), 648-651 (1999).
23. Penders J, Thijs C, Vink C *et al*. Factors influencing the composition of the intestinal microbiota in early infancy. *Pediatrics* 118(2), 511-521 (2006).
24. Hanson LA. Session 1: Feeding and infant development breast-feeding and immune function. *Proc Nutr Soc* 66(3), 384-396 (2007).
25. Lodge CJ, Tan DJ, Lau MX *et al*. Breastfeeding and asthma and allergies: a systematic review and meta-analysis. *Acta Paediatr* 104(467), 38-53 (2015).
26. Kinlen L. Evidence for an infective cause of childhood leukaemia: comparison of a Scottish new town with nuclear reprocessing sites in Britain. *Lancet* 2(8624), 1323-1327 (1988).
27. Greaves M. The 'delayed infection' (aka 'hygiene') hypothesis for childhood leukaemia. In: *The Hygiene Hypothesis and Darwinian Medicine. Progress in Inflammation Research.*, G.A.W. R (Ed.^(Eds).Birkhäuser Basel (2009).
28. Greaves MF. Speculations on the cause of childhood acute lymphoblastic leukemia. *Leukemia* 2(2), 120-125 (1988).
29. Greaves M. Infection, immune responses and the aetiology of childhood leukaemia. *Nat Rev Cancer* 6(3), 193-203 (2006).
30. Greaves M. A causal mechanism for childhood acute lymphoblastic leukaemia. *Nat Rev Cancer.* 18(8), 471-484. doi: 410.1038/s41568-41018-40015-41566. (2018).
31. Ajrouche R, Rudant J, Orsi L *et al*. Childhood acute lymphoblastic leukaemia and indicators of early immune stimulation: the Estelle study (SFCE). *Br J Cancer* 112(6), 1017-1026 (2015).
32. Jourdan-Da Silva N, Perel Y, Mechinaud F *et al*. Infectious diseases in the first year of life, perinatal characteristics and childhood acute leukaemia. *Br J Cancer* 90(1), 139-145 (2004).
33. Perrillat F, Clavel J, Auclerc MF *et al*. Day-care, early common infections and childhood acute leukaemia: a multicentre French case-control study. *Br J Cancer* 86(7), 1064-1069 (2002).
34. Rudant J, Orsi L, Menegaux F *et al*. Childhood acute leukemia, early common infections, and allergy: The ESCALE Study. *Am J Epidemiol* 172(9), 1015-1027 (2010).
35. Gilham C, Peto J, Simpson J *et al*. Day care in infancy and risk of childhood acute lymphoblastic leukaemia: findings from UK case-control study. *Bmj* 330(7503), 1294 (2005).
36. Ma X, Urayama K, Chang J, Wiemels JL, Buffler PA. Infection and pediatric acute lymphoblastic leukemia. *Blood Cells Mol Dis* 42(2), 117-120 (2009).
37. Rudant J, Lightfoot T, Urayama KY *et al*. Childhood acute lymphoblastic leukemia and indicators of early immune stimulation: a Childhood Leukemia International Consortium study. *Am J Epidemiol* 181(8), 549-562 (2015).
38. Amitay EL, Keinan-Boker L. Breastfeeding and Childhood Leukemia Incidence: A Meta-analysis and Systematic Review. *JAMA Pediatr* 169(6), e151025 (2015).
39. Langie SA, Szarc Vel Szic K, Declerck K *et al*. Whole-Genome Saliva and Blood DNA Methylation Profiling in Individuals with a Respiratory Allergy. *PLoS One* 11(3), e0151109 (2016). **Of considerable interest. Provides evidence for DNA methylation in individuals with RA.
40. Langie SaS, Moisse M, Szarc Vel Szic K *et al*. GLI2 promoter hypermethylation in saliva of children with a respiratory allergy. *Clin Epigenetics.* 10:50.(doi), 10.1186/s13148-13018-10484-13141. eCollection 12018. (2018).
41. Nordlund J, Syvanen AC. Epigenetics in pediatric acute lymphoblastic leukemia. *Semin Cancer Biol* doi:10.1016/j.semcancer.2017.09.001 (2017).
42. Nordlund J, Backlin CL, Wahlberg P *et al*. Genome-wide signatures of differential DNA methylation in pediatric acute lymphoblastic leukemia. *Genome Biol.* 14(9), r105. doi: 110.1186/gb-2013-1114-1189-r1105. (2013). *Of interest. Epigenetic profiling of childhood ALL, suggesting epigenetic factors contribute to disease.

43. Devries A, Vercelli D. Early predictors of asthma and allergy in children: the role of epigenetics. *Curr Opin Allergy Clin Immunol* 15(5), 435-439 (2015). *Of interest. Review of the literature suggesting epigenetics may play a role in allergy development.
44. Timms JA, Relton CL, Rankin J, Strathdee G, Mckay JA. DNA methylation as a potential mediator of environmental risks in the development of childhood acute lymphoblastic leukemia. *Epigenomics* 8(4), 519-536 (2016).
45. Feil R, Fraga MF. Epigenetics and the environment: emerging patterns and implications. *Nat Rev Genet* 13(2), 97-109 (2012).
46. Mckay JA, Mathers JC. Diet induced epigenetic changes and their implications for health. *Acta Physiol (Oxf)* 202(2), 103-118 (2011).
47. Birdwell CE, Queen KJ, Kilgore PC *et al*. Genome-wide DNA methylation as an epigenetic consequence of Epstein-Barr virus infection of immortalized keratinocytes. *J Virol* 88(19), 11442-11458 (2014).
48. Mcerlean P, Favoreto S, Jr., Costa FF *et al*. Human rhinovirus infection causes different DNA methylation changes in nasal epithelial cells from healthy and asthmatic subjects. *BMC Med Genomics* 7 37 (2014).
49. Pacis A, Tailleux L, Morin AM *et al*. Bacterial infection remodels the DNA methylation landscape of human dendritic cells. *Genome Res* 25(12), 1801-1811 (2015).
50. Hattori N, Ushijima T. Epigenetic impact of infection on carcinogenesis: mechanisms and applications. *Genome Med* 8(1), 10 (2016).
51. Timms JA, Relton CL, Sharp GC, Rankin J, Strathdee G, Mckay JA. Exploring a potential mechanistic role of DNA methylation in the relationship between in utero and post-natal environmental exposures and risk of childhood acute lymphoblastic leukaemia. *Int J Cancer* 10(10), 32203 (2019). **Of considerable interest. Provides evidence for a role of DNA methylation as a mediating mechanism between environmental factors and ALL risk.
52. Chatterton Z, Morenos L, Mechinaud F *et al*. Epigenetic deregulation in pediatric acute lymphoblastic leukemia. *Epigenetics.* 9(3), 459-467. doi: 410.4161/epi.27585. Epub 22014 Jan 27586. (2014). *Of interest. Epigenetic profiling of childhood ALL, suggesting epigenetic factors contribute to disease.
53. Felix JF, Joubert BR, Baccarelli AA *et al*. Cohort Profile: Pregnancy And Childhood Epigenetics (PACE) Consortium. *Int J Epidemiol.* 47(1), 22-23u. doi: 10.1093/ije/dyx1190. (2018).
54. Asher MI, Keil U, Anderson HR *et al*. International Study of Asthma and Allergies in Childhood (ISAAC): rationale and methods. *Eur Respir J.* 8(3), 483-491. (1995).
55. Houseman EA, Accomando WP, Koestler DC *et al*. DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics.* 13:86.(doi), 10.1186/1471-2105-1113-1186. (2012).
56. Timms JA, Relton CL, Sharp GC, Rankin J, Strathdee G, Mckay JA. Exploring a potential mechanistic role of DNA methylation in the relationship between in utero and post-natal environmental exposures and risk of childhood acute lymphoblastic leukaemia *International Journal of Cancer* (In Press).
57. Sharp GC, Lawlor DA, Richmond RC *et al*. Maternal pre-pregnancy BMI and gestational weight gain, offspring DNA methylation and later offspring adiposity: findings from the Avon Longitudinal Study of Parents and Children. *International journal of epidemiology* 44(4), 1288-1304 (2015).
58. Sherman BT, Huang Da W, Tan Q *et al*. DAVID Knowledgebase: a gene-centered database integrating heterogeneous gene annotation resources to facilitate high-throughput gene functional analysis. *BMC bioinformatics* 8 426 (2007).
59. Imgenberg-Kreuz J, Carlsson Almlof J, Leonard D *et al*. DNA methylation mapping identifies gene regulatory effects in patients with systemic lupus erythematosus. *Ann Rheum Dis.* 77(5), 736-743. doi: 710.1136/annrheumdis-2017-212379. Epub 212018 Feb 212371. (2018).

60. Zhu H, Wu LF, Mo XB *et al*. Rheumatoid arthritis-associated DNA methylation sites in peripheral blood mononuclear cells*. Ann Rheum Dis.* 78(1), 36-42. doi: 10.1136/annrheumdis-2018-213970. Epub 212018 Oct 213978. (2019).
61. Souren NY, Gerdes LA, Lutsik P *et al*. DNA methylation signatures of monozygotic twins clinically discordant for multiple sclerosis*. Nat Commun.* 10(1), 2094. doi: 2010.1038/s41467-41019-09984-41463. (2019).
62. Rask-Andersen M, Martinsson D, Ahsan M *et al*. Epigenome-wide association study reveals differential DNA methylation in individuals with a history of myocardial infarction*. Hum Mol Genet.* 25(21), 4739-4748. doi: 4710.1093/hmg/ddw4302. (2016).
63. Meeks KaC, Henneman P, Venema A *et al*. Epigenome-wide association study in whole blood on type 2 diabetes among sub-Saharan African individuals: findings from the RODAM study*. Int J Epidemiol.* 48(1), 58-70. doi: 10.1093/ije/dyy1171. (2019).
64. Wahl S, Drong A, Lehne B *et al*. Epigenome-wide association study of body mass index, and the adverse outcomes of adiposity*. Nature.* 541(7635), 81-86. doi: 10.1038/nature20784. Epub 22016 Dec 20721. (2017).
65. Di Genova L, Penta L, Biscarini A, Di Cara G, Esposito S. Children with Obesity and Asthma: Which Are the Best Options for Their Management? *Nutrients.* 10(11).(pii), nu10111634. doi: 10111610.10113390/nu10111634. (2018).
66. Tabe Y, Konopleva M. Advances in understanding the leukaemia microenvironment*. Br J Haematol.* 164(6), 767-778. doi: 710.1111/bjh.12725. Epub 12014 Jan 12729. (2014).
67. Liu H, Jiang D, Chi F, Zhao B. The Hippo pathway regulates stem cell proliferation, self-renewal, and differentiation*. Protein Cell.* 3(4), 291-304. doi: 210.1007/s13238-13012-12919-13233. Epub 12012 May 13232. (2012).
68. Plouffe SW, Hong AW, Guan KL. Disease implications of the Hippo/YAP pathway*. Trends Mol Med.* 21(4), 212-222. doi: 210.1016/j.molmed.2015.1001.1003. Epub 2015 Feb 1018. (2015).
69. Fodor LE, Gezsi A, Ungvari L *et al*. Investigation of the Possible Role of the Hippo/YAP1 Pathway in Asthma and Allergy*. Allergy Asthma Immunol Res.* 9(3), 247-256. doi: 210.4168/aair.2017.4169.4163.4247. (2017).
70. Cottini F, Hideshima T, Xu C *et al*. Rescue of Hippo coactivator YAP1 triggers DNA damage-induced apoptosis in hematological cancers*. Nat Med.* 20(6), 599-606. doi: 510.1038/nm.3562. Epub 2014 May 1011. (2014).
71. Okai I, Wang L, Gong L *et al*. Overexpression of JAKMIP1 associates with Wnt/beta-catenin pathway activation and promotes cancer cell proliferation in vitro*. Biomed Pharmacother.* 67(3), 228-234. doi: 210.1016/j.biopha.2013.1001.1003. Epub 2013 Feb 1011. (2013).
72. Javadi M, Ganesan P, Bordoloi D, Roy NK, Kunnumakkara A. *Neutrophil Gelatinase-Associated Lipocalin (NGAL): A Promising Biomarker for Cancer Diagnosis and A Potential Target for Cancer Therapeutics*. 1, (2014).
73. Sm Z. NGAL Suppresses Immune Responses to Meningococcal Capsular Polysaccharide-Based Vaccine. . *International Journal of Vaccines and Vaccination* 2(6), 00051 (2016).