Identification of loci where DNA methylation potentially mediates genetic risk of type 1 diabetes

Jody Ye¹*, Tom G Richardson², Wendy L McArdle³, Caroline L Relton², Kathleen M Gillespie¹, Matthew Suderman², Gibran Hemani²

Author Affiliations:

¹Diabetes and Metabolism, Bristol Medical School (Translational Health Sciences), University of Bristol, Level 2 Learning and Research, Southmead Hospital, Bristol, UK, BS10 5NB

²MRC Integrative Epidemiology Unit, Bristol Medical School (Population Health Sciences), University of Bristol, Oakfield House, Oakfield Grove, Bristol, UK, BS8 2BN

³Population Health Sciences, Bristol Medical School, University of Bristol, Oakfield House, Oakfield Grove, Bristol, UK, BS8 2BN

Corresponding to: Dr Jody Ye, Diabetes and Metabolism, Bristol Medical School, Level 2 Learning and Research, University of Bristol, Southmead Hospital, Bristol, BS10 5NB, United Kingdom, jody.ye@bristol.ac.uk, Tel: +44 (0)117 414 8041

*Present address: Room 702, Division of Endocrinology, Department of Medicine, Albert Einstein College of Medicine, 1300 Morris Park Avenue, Bronx, New York, 10461, United States, yi.ye@einstein.yu.edu, Tel: +1 718-430-8503
## Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARIES</td>
<td>Accessible Resource for Integrated Epigenomic Studies</td>
</tr>
<tr>
<td>ALSPAC</td>
<td>Avon Longitudinal Study of Parents and Children</td>
</tr>
<tr>
<td>BOX</td>
<td>Bart’s Oxford family study of Type 1 Diabetes</td>
</tr>
<tr>
<td>CpG</td>
<td>Cytosine-phosphate-guanine dinucleotides</td>
</tr>
<tr>
<td>DNAm</td>
<td>DNA methylation</td>
</tr>
<tr>
<td>GWAS</td>
<td>Genome-wide association study</td>
</tr>
<tr>
<td>HLA</td>
<td>Human leukocyte antigen</td>
</tr>
<tr>
<td>JLIM</td>
<td>Joint likelihood mapping</td>
</tr>
<tr>
<td>LD</td>
<td>Linkage disequilibrium</td>
</tr>
<tr>
<td>SNP</td>
<td>Single nucleotide polymorphism</td>
</tr>
<tr>
<td>T1D</td>
<td>Type 1 diabetes</td>
</tr>
<tr>
<td>MAF</td>
<td>Minor allele frequency</td>
</tr>
<tr>
<td>MR</td>
<td>Mendelian Randomization</td>
</tr>
<tr>
<td>mQTL</td>
<td>methylation quantitative trait loci</td>
</tr>
<tr>
<td>2SMR</td>
<td>Two Sample Mendelian Randomization</td>
</tr>
</tbody>
</table>
Abstract

The risk of Type 1 Diabetes (T1D) comprises both genetic and environmental components. We investigated whether genetic susceptibility to T1D could be mediated by changes in DNA methylation, an epigenetic mechanism that potentially plays a role in autoimmune diabetes. From enrichment analysis, we found that there was a common genetic influence for both DNA methylation and T1D across the genome, implying that methylation could be either on the causal pathway to T1D or a non-causal biomarker of T1D genetic risk. Using data from a general population comprising blood samples taken at birth (n=844), childhood (n=846) and adolescence (n=907), we then evaluated the associations between 64 top GWAS single nucleotide polymorphisms (SNPs) and DNA methylation levels at 55 non-HLA loci. We identified 95 proximal SNP-cytosine phosphate guanine (CpG) pairs (cis) and 1 distal SNP-CpG association (trans) consistently at birth, childhood, and adolescence. Combining genetic co-localization and Mendelian Randomization analysis, we provided evidence that at 5 loci, ITGB3BP, AFF3, PTPN2, CTSH and CTLA4, DNA methylation is potentially mediating the genetic risk of T1D mainly by influencing local gene expression.

Word count: 177

Key words: DNA methylation; Epigenetics; Type 1 Diabetes; Mendelian Randomization
1. Introduction

Type 1 diabetes (T1D) is a polygenic disease with more than 50% of genetic susceptibility attributable to the human leukocyte antigen (HLA) class II region and the remaining attributable to the non-HLA region [1]. Over the last decade, genome-wide association studies (GWAS) identified 62 independent loci and over 100 single nucleotide polymorphisms (SNPs) associated with T1D risk [2, 3], but the biological pathways for most are unknown. DNA methylation (DNAm) is an epigenetic event that occurs at cytosine – phosphate – guanine (CpG) residues and can be modified by both genetic and environmental exposures. Genetic and epigenetic interactions have been postulated to contribute to susceptibility to a number of autoimmune disorders [5, 6].

For a multi-factorial disease like T1D, its liability is determined by polygenic risk plus environmental exposures; the varied genetic background plus environmental risk factors often result in varying rates of diabetes progression and degrees of beta cell destruction. Previous work focused on identifying methylation differences among T1D monozygotic twins [8], allowing changes of DNAm levels that are associated with environmental exposure to be captured [9-11], but no study has systematically investigated whether DNAm is causally influencing T1D due to common genetic risk.

Genetic variance has been reported to explain 24% of the methylation variance in childhood and 21% in middle age, when T1D can develop [12, 13]. We therefore asked whether genetic susceptibility of T1D could be mediated by DNAm, which potentially result in gene expression heterogeneities within a population. These intrinsic heterogeneities are increasingly recognised
to lead to immune and/or beta cell dysfunction in the presence of environmental stimuli [14, 15]. However, in true biological settings, the relationship between genetic risk, DNA methylation (DNAm) and type 1 diabetes (T1D) can be explained by four scenarios:

1. The SNP that influences DNA methylation is in linkage disequilibrium (LD) with a causal variant of T1D (Figure 1a);

2. The SNP has a causal effect on T1D mediated by the changes of DNA methylation levels (Figure 1b);

3. The SNP has a causal effect on T1D (i.e. via altering gene expression), which in turn alters DNA methylation levels (reverse causal, Figure 1c);

4. The SNP influences methylation and T1D via separate mechanisms, an effect known as horizontal pleiotropy (Figure 1d).

To disentangle the causal associations between DNA methylation and T1D risk, we employed the principles of Mendelian Randomization (MR), which is a powerful technique that has been used in both basic and clinical research [16-20]. Since genotypes are randomly determined at conception and not influenced by environmental confounders [21], they can be used as instruments to proxy exposures that are potentially influencing a trait, thereby mimicking a randomised controlled trial [22]. In the context of DNA methylation, SNPs that regulate methylation levels at nearby CpG sites (defined within 1Mb distance, known as cis-mQTLs) can be used to investigate the causal effect of DNA methylation on a trait [23]. Compared with traditional MR, where the effects of genetic instruments on exposure and on associated traits are measured in the same population (hence one-sample MR), Two-Sample MR (2SMR) has been developed to enable causal inference using summary statistics from GWAS alone, circumventing the requirement that DNA methylation levels and T1D status are measured in the same sample, enabling much larger
sample sizes [22]. Because biological heterogeneities could be present as long as an individual
inherited a risk allele, both the general population and T1D population can be used to partition
genetic risk on the liability scale, regardless of the individual’s disease affection status [24].

In this study, we first tested whether there was a genome-wide association between cis-mQTLs
and T1D by performing an enrichment analysis. Subsequently, at 55 non-HLA loci we
established associations between top T1D GWAS variants and DNAm in a well-characterised
general population. Next, we employed joint likelihood mapping (JLIM) to pinpoint loci where
DNAm and T1D share the same causal variants [25]. For these loci, we then used 2SMR to
assess the causal effect of DNAm on T1D risk [26]. We also tested whether the findings could
be seen in a subset of individuals from a T1D cohort. Finally, we assessed the relationship
between DNAm and local gene expression levels. A flow chart summarising the analysis
procedure is shown in Figure 2.

1. Materials and methods

2.1. Subjects

2.1.1. General population

The Avon Longitudinal Study of Parents and Children (ALSPAC) study is a large scale
prospective study based in Avon, UK. ALSPAC recruited 14,541 British pregnant women
with expected delivery dates between 1st April 1991 to 31st December 1992, clinical data and
biological samples were collected during pregnancy and at regular intervals postpartum from
both parents and offspring [27, 28]. The study website contains details of all the data that is
available through a fully searchable data dictionary

http://www.bris.ac.uk/alspac/researchers/data-access/data-dictionary/.
Accessible Resource for Integrated Epigenome Studies (ARIES) is a sub-study of ALSPAC. Mother-offspring pairs (n=1018) were selected to be included in ARIES based on the availability of DNA samples at two-time points of the mother (antenatal and at follow-up when the offspring were adolescents) and three-time points of the offspring (birth, childhood and adolescence) [29]. In the questionnaire completed by mothers at 10.5 years post pregnancy, mothers were asked if their child was currently taking any diabetes medication. Only two mothers answered ‘yes’. For childhood and adolescent samples, either whole blood or peripheral blood lymphocyte DNA was used; for samples taken at birth, cord blood DNA was used. Methylation data from the offspring at adolescence (mean age 17.1 years, n=907), childhood (mean age 7.5 years, n=846) and birth (n=844) were used in this study. The detailed sample characteristics were summarised in Supplemental Table 1.

2.1.2. A subset of individuals from a T1D cohort.

Sixteen families including 45 individuals were selected based on sample availability from the Bart’s Oxford family study of type 1 diabetes (BOX). The BOX study recruited probands diagnosed with T1D under the age of 21 and their family members in the Oxford region, UK, since 1985. Currently over 2000 families are under regular follow-up [30]. The selected families here include children with T1D (mean age 10.8 years, n=15), their parents (mean age 41.25 years, n=16) and grandparents (mean age 72.4 years, n= 14) with or without T1D. Whole blood DNA from this cohort was used in this study. The detailed clinical characteristics of BOX participants were described in Supplemental Table 2.

2.2. Methylation 450k array data and quality control
DNAm data of the ARIES [29] and BOX cohort were generated using the Illumina HumanMethylation450 BeadChip (‘450k array’) (Illumina, San Diego, CA, USA). The array quantifies DNAm levels of >485,000 CpG sites, which covers 99% RefSeq genes [31]. For quality control, probes with low signal to noise ratio (detection p-value > 0.01) or with methylated/unmethylated read counts of 0 were removed. Probes containing SNPs located at the CpG site at the single base extension site, as well as within 10 nucleotides within the query CpG site [32] were removed. Batch effect was normalized using the ‘wateRmelon’ R package [12]. Beta values (which approximate DNAm levels) were calculated using M/(M+U+100), where M is the methylated signal, U is the unmethylated signal. To remove outliers and heteroscedasticity of the data, bimodally distributed beta values were converted to M values using log2(beta/1-beta) [33] and then rank-transformed into normal distributions. In all subsequent analyses, rank-transformed M values were used; standard deviation units were also derived from rank-transformed M values. The final methylation data contained 459,734 probes per individual. To account for cellular heterogeneities, B cell, CD4+ T cell, CD8+ T cell, granulocyte, monocyte, and NK cell composition for each sample was estimated using a Reference-based algorithm developed by Houseman et al [34].

2.3. Individual level genotype data

Individual level genotype data on the ARIES cohort were generated using the Illumina HumanHap550-quad chips by Sample Logistics and Genotyping Facilities at the Wellcome Trust Sanger Institute and LabCorp (Laboratory Corporation of America). Details for the quality control have been described before [18]. T1D GWAS SNPs (n=113) were obtained from www.immunobase.org. Six additional SNPs associated with T1D were obtained from a recent meta-analysis [2]. Among them, genotypes of sixty-four SNPs spanning 55 regions (LD
r^2 < 0.1) were available in ARIES, where necessary, proxy SNPs (minimal r^2 = 0.6) were used to replace the original variants to obtain the required odds ratios for downstream MR analysis (Supplemental Table 3).

Genotype of the BOX participants were determined using Taqman® allele discrimination assays (Life Technologies, UK). Taqman probes for rs2269242, rs9653442, rs3087243, rs3825932, and rs1893217 were purchased from Life Technologies (Thermo Fisher, UK).

2.4. Summary statistics

T1D GWAS summary statistics were obtained from two meta-analyses: one for the initial analysis (Data 1) [2] and one for the replication (Data 2) [35]; details were summarized in Supplemental Table 4. For Data 1, beta coefficient, p-values, and standard errors were available. For Data 2, only p-values were available.

For SNP and DNAm associations, summary statistics were obtained from the ARIES participants from a previous study [12]. Beta coefficient for normalized effect sizes, p-values, and standard errors were available.

For SNP and gene expression associations, summary statistics for whole blood were downloaded from the Genotype-Tissue Expression (GTEx) consortium [36] https://www.gtexportal.org. Beta coefficient for normalized effect sizes, p-values, and standard errors were available.

2.5. Statistical analyses

2.5.1. Genome-wide cis-mQTL enrichment analyses
We investigated whether there is an overall association between cis-mQTLs and T1D more than expected by chance. Independent cis-mQTLs (LD $r^2 < 0.1$) that have shown strong associations ($p < 1 \times 10^{-14}$, Type I error rate 0.2%) with DNAm were retrieved from the ARIES cohort [12]; their $p$-values for T1D associations were retrieved from the two GWAS summary statistics (Data 1 and Data 2, respectively). The enrichment of cis-mQTLs could be either due to their (1) distinct SNP properties (allele frequency, LD, gene distance) or (2) due to distinct genomic locations (i.e. promoter, intron, exon, 5’ UTR, or 3’ UTR). To control for these two factors, null SNPs were selected to match cis-mQTLs in two ways. Firstly, null SNPs were chosen based on similarities in minor allele frequency (MAF) and LD structures [37]. For example, null SNPs must be at least 1000kb away from cis-mQTLs; the maximum MAF deviation of null SNPs from cis-mQTLs is 0.02; and LD scores [38] of null SNPs are in the same quintile bin of cis-mQTLs. Secondly, null SNPs were chosen based on similarities in genomic annotations, such as, intron, exon, 5’UTR, 3’UTR or promoter SNPs. For both methods, null SNPs were sampled without replacement. Fisher’s combined probability test was used to obtain an overall association with T1D:

$$X_{2k}^2 \sim -2 \sum_{i=1}^{k} \ln p_i$$

where $p$ is the $p$-value for SNP and T1D association. To generate a distribution for null SNPs, the same number of null SNPs was randomly drawn 10,000 times to obtain 10,000 combined $p$-values. The empirical $p$-value, reflecting the likelihood of observing a combined $p$-value at least as extreme as the combined $p$-value for cis-mQTLs in the null distribution, is calculated by ranking all the 10,000 $p$-values for null SNPs [12].

2.5.2. Epigenome-wide association analyses at T1D susceptible loci
We regressed 64 independent index SNPs with 459,734 CpG sites measured separately at adolescence, childhood and birth using a mixed effect linear model:

\[
y_j = \beta_0 + \beta_{\text{CpG}_j} \cdot \text{CpG}_j + \beta_{\text{age}} \cdot \text{age}_j + \beta_{\text{sex}} \cdot \text{Sex}_j + \beta_{\text{cell}} \cdot \text{cell count}_j + \epsilon_j
\]

where \(y_j\) is the rank-transformed M value for DNA methylation; \(\beta_0\) is the coefficient of the intercept; \(\beta_{\text{CpG}_j}\) is the coefficient for CpG site; \(\beta_{\text{age}}\) is the coefficient for age; \(\beta_{\text{sex}}\) is the coefficient for sex; \(\beta_{\text{cell}}\) is the coefficient for cell compositions estimated using the Houseman algorithm [34]; \(\epsilon_j\) is the random error term. A Bonferroni threshold of \(1.6 \times 10^{-9}\) \((0.05/64 \times 465,877)\) was used to correct for multiple-testing. Analyses were performed using the MatrixEQTL package in R 3.2.2 statistical software on the University of Bristol High Performance Computing (HPC) cluster.

2.5.3. Joint likelihood mapping

Given a CpG site, JLIM estimates the putative causal variant for this CpG site within a 1-Mb window centred on that mQTL. It also estimates the putative causal variant for T1D within the same region [25]. Concordance between top variants for the two traits would suggest that DNAm potentially reside in the causal pathway to T1D risk. The concordance rates were determined after accounting for chance, under 1000 permutations. This analysis was performed using the jlimR module in R 3.2.2 statistical software, available at https://github.com/cotsapaslab/jlim.

2.5.4. Causal effect between DNAm and T1D risk

Forward 2SMR was used to test whether DNAm (exposure) is causally influencing T1D (outcome, Figure 1b and Supplemental Figure 1). To reduce potential pleotropic effect, only
cis-mQTLs were used as instruments since trans-mQTLs may influence methylation via multiple biological pathways. To exclude potential instrument-confounder associations, we also ensured that instruments were not associated with fasting glucose concentration in a large GWAS meta-analysis involving 133,010 non-diabetic European individuals [39]. The associations between cis-mQTLs (genetic instruments) and DNAm (exposure) in the ARIES cohort (sample one) can be written as:

\[ X_i = \sum_{j=1}^{I} \gamma_j G_{ij} + U_i + \epsilon_i^X \] (1)

where \( X_i \) is the rank-transformed M value for DNA methylation; \( G_{ij} \) is the \( j \)th genetic variant for the \( i \)th participant; \( \gamma_j \) is the coefficient for each variant \( j \) represents the effect of the genetic variant on DNA methylation; \( U_i \) is the unknown confounding effect; \( \epsilon_i^X \) is the independent error term on methylation.

The associations between cis-mQTLs (genetic instruments) and T1D risk (outcome) obtained from T1D GWAS summary statistics (sample two) can be written as:

\[ Y_i = \Gamma_j G_{ij} + \epsilon_i^Y \] (2)

where \( Y_i \) is log odds for T1D, \( \Gamma_j \) is the coefficient of the genetic variant on T1D risk, \( \epsilon_i^Y \) is the error term.

The associations between DNAm (exposure) and T1D risk (outcome) can be written as:

\[ Y_i = \sum_{j=1}^{I} \alpha_j G_{ij} + \beta X_i + U_i + \epsilon_i^Y \] (3)

where \( \alpha_j \) is the coefficient for each variant \( j \) represents the pleiotropic effect that is not mediated by DNA methylation; \( \epsilon_i^Y \) is the independent error term; \( \beta \) (Wald ratio estimator) is the the causal effect of DNA methylation on T1D risk.
Combining equation (1) and (3), the relationship between DNA methylation and T1D variants are summarized as:

\[ Y_i = (\alpha_j + \beta \gamma_j)G_{ij} + \epsilon_{ij} \]

When \( \alpha_j = 0 \), meaning that genetic instruments have no pleiotropic effects \([40]\), the causal effect of DNA methylation on T1D risk can be estimated using the ratio method \([41]\):

\[ \beta = \frac{\Gamma_j}{\gamma_j} \]

\( \beta \) is interpreted as the change in log odds for T1D per standard deviation unit (rank transformed M value) increase in DNAm due to its associated SNP. A \( p \)-value is then reported to assess whether this causal effect is significantly different from zero.

To test the causal effect of T1D risk (exposure) on DNAm at each CpG site (outcome, Figure 1c and Supplemental Figure 2), multiple GWAS SNPs were used to proxy T1D risk, the causal effects for all genetic instruments were then averaged in a fixed-effect meta-analysis using MR – inverse variance weighting \([42]\):

\[ \frac{\sum_j \hat{\Gamma}_j \hat{\gamma}_j se(\hat{\Gamma}_j)^{-2}}{\sum_j \hat{\gamma}_j^2 se(\hat{\Gamma}_j)^{-2}} \]

Since the same exposure was tested against multiple CpG sites, adjustment for multiple testing was required. A Bonferroni threshold was therefore used to keep the Type I Error rate of the causal effect at 5%. Because multiple instruments were available in the reverse 2SMR, funnel plot and Cochran’s Q-test \([41]\) were used to assess directional pleiotropy and heterogeneities in the genetic instruments. Instruments that exhibited obvious pleiotropic effects were removed. Significant heterogeneity was determined if the \( p \)-value of the Cochran’s Q-test was less than 0.05.
2.5.5. **MR-Steiger directionally test**

In some situations, such as DNAm, it is difficult to ascertain whether genetic risk first causes changes in DNAm, which subsequently results in T1D risk or vice versa. As a verification of 2SMR results, MR-Steiger was used to assess whether DNAm was the likely exposure and T1D risk was the likely outcome. MR-Steiger estimated the proportion of variance in the exposure and in the outcome that was explained by genetic instruments. Causal direction was then determined based on whether exposure variance or outcome variance was subject to the primary effect of SNPs [43].

2.5.6. **Relationship between DNAm and gene expression**

We used 2SMR to assess whether DNAm potentially influenced gene expression in whole blood. Wald ratio was calculated using the beta coefficient for the mQTL and gene expression association (obtained from the GTEx consortium) divided by the beta coefficient for the mQTL and DNAm association (obtained from the ARIES). A p-value less than or equal to 0.05 was considered significant. 2SMR and MR-Steiger were performed using the ‘TwoSampleMR’ R package.

3. **Results**

3.1. **Genome-wide enrichment of cis-mQTLs and Type 1 diabetes associations**

As a primary analysis, we overlapped adolescent cis-mQTLs with an initial discovery dataset (Data 1). These cis-mQTLs were evenly distributed across the genome (data not shown) and
were devoid of HLA-SNPs (chr6: 28,477, 797-33, 448,354, hg19). Their overall associations with T1D were significantly enriched in SNPs with low GWAS p-values, matching null SNPs to cis-mQTLs either by SNP properties (Figure 3 a) or by genomic annotations (Figure 3 b). Secondary analyses using childhood and neonatal cis-mQTLs revealed the same findings (Supplemental Figure 3). To verify these observations, we performed a replication study using Data 2. Compared to Data 1, there was a stronger enrichment when cis-mQTLs were matched either by SNP properties or by genomic annotations (Supplemental Figure 3). These data suggested that there was a shared genetic influence of DNAm and T1D.

3.2. The associations between DNAm and index SNPs at T1D susceptible loci across three-time points of life

Of the 64 independent T1D GWAS variants, thirty-seven SNPs were consistently found to associate with a total of 96 CpG sites under the Bonferroni corrected threshold ($p < 1.6e-9$) at adolescence, childhood and birth in the ARIES cohort. One SNP-CpG pair was in trans (distance $>1$Mb) and the remainder were in cis (Supplemental Table 5). Figure 4 shows the genomic distribution of 96 total CpGs, excluding associations at the HLA locus. Methylation variance ($R^2$) explained by T1D SNPs varied largely, the strongest association was between rs7149271 and cg20045882 on chromosome 14, where rs7149271 explained greater than 78% methylation variance across all three-time points. The overall effect sizes of T1D variants on DNAm levels were consistent across three-time points, where the correlation between adolescence and childhood was 0.996 (95% CI: 0.995, 0.997, $p < 2.2e-16$); between adolescence and birth was 0.984 (95% CI: 0.979, 0.989, $p < 2.2e-16$); and between childhood and birth was 0.987 (95% CI: 0.982, 0.990, $p < 2.2e-16$).
3.3. Joint likelihood mapping pinpointed five loci where DNAm and T1D share the same causal variants

After removing SNPs with no beta coefficients for T1D, there were sixty-six CpG associations remaining, unique to 32 cis-mQTLs. JLIM pinpointed shared causal variants for DNAm and T1D at 5 loci, including ITGB3BP, AFF3, CTLA4, CTSH, and PTPN2, mediated by 10 CpG sites (Figure 5 and Table 1).

3.4. DNAm potentially mediates T1D genetic risk at five loci

Forward 2SMR showed that all the 10 CpGs potentially influenced T1D risk and the effects were consistent at adolescence, childhood and birth (Wald ratio $p < 0.05$, Table 1 and Supplemental Table 6). In the reverse direction, Bonferroni corrected threshold suggested that a $p$-value of 0.005 was required to keep the type I error rate at 5%. At this threshold, there was no evidence for reverse causation at all three-time points (Supplemental Table 7). Funnel plots (Supplemental Figure 4) and Cochran’s Q-tests (data not shown) did not demonstrate significant directional pleiotropy or heterogeneities in the genetic instruments. As a verification, MR-Steiger supported that DNAm was a mediator of T1D risk and this effect was also consistent at all three-time points (Supplemental Table 8).

3.5. Similar patterns were observed in individuals from a T1D cohort

The 10 SNP-CpG associations were further assessed in 45 individuals participating in an independent BOX study, where methylation 450k array data were available. In nine out of ten associations, DNAm patterns were consistent at the same direction as comparing to the ARIES cohort (Figure 6).
3.6. DNAm potentially alters local gene expression

Using DNAm as exposure and gene expression as outcome, 2SMR analysis suggested that in whole blood, apart from cg22572158, all the other 9 CpGs influenced the expression of their adjacent genes. For most genes we observed inverse relationship between DNAm and gene expression. For example, hypermethylation at cg05732488 decreased the expression of a pseudogene RN7SL130P; hypermethylation at cg18738367 decreased the expression of CTSH, whereas hypomethylation at cg06183267 and cg22572158 increased the expression of AFF3; hypomethylation at cg09945482, cg23544223, cg23598886 and cg24737193 increased the expression of two lincRNAs RP11-973H7.1 and RP11-973H7.4. We also observed positive associations between DNAm and gene expressions. For example, hypomethylation at cg25744700 decreased CTSH expression; hypomethylation at cg09945482, cg23544223, cg23598886 and cg24737193 decreased PTPN2 expression. These data are summarized in Table 2.

4. Discussion

One hypothesis of the mechanisms underlying T1D is that genetic variants alter DNAm levels, which in turn influence genes that are essential for immune tolerance as well as beta cell function. To the best of our knowledge, this is the first study that systematically evaluated the causal effect of DNAm on T1D genetic risk in a large-scale population. We showed genome-wide enrichment of cis-mQTLs and T1D associations in a young non-diabetic population at three-time points of life, implying that these intrinsic methylation heterogeneities could either be causally influencing the liability of T1D or facilitate as biomarkers for islet autoimmunity development and progression. A previous analysis based on SNP heritability however did not
show a significant association between cis-mQTLs and T1D [12]. Narrow sense heritability of T1D was estimated to be approximately 0.8 [44, 45]. The lack of enrichment of cis-mQTLs in SNP heritability found by the previous study, was probably due to limited number of cis-mQTLs used in the estimation. Since approximately 20,000 cis-mQTLs were identified in the ARIES study [12], more cis-mQTLs were perhaps required to capture enough genotypic variance to explain a highly heritable condition.

At the T1D susceptible loci, we identified widespread genetic and epigenetic interactions; more than half of the T1D-SNPs were associated with proximal methylation variation in the genome. However, JLIM suggested that majority of DNAm associations tested in the 450k array were non-causal due to LD-confounding and in 5 non-HLA loci, DNAm and T1D shared the same causal variants. A limitation of this test is that it does not report which SNPs are the likely causal variants. Additionally, the HLA loci could not be analysed due to the high LD structure and high false discovery rates for causal variant prediction [25].

We found that DNAm levels correlated more strongly between birth and childhood than between birth and adolescence, indicating that environment modulates DNAm overtime. Interestingly, consistent associations between T1D SNPs and DNAm were observed at the 5 non-HLA loci across three-time points of life, suggesting that genetic influences at these loci are important from birth to adolescence. More longitudinal data are required to test whether environment further modifies the risk of these loci later in life.

Mendelian randomization analyses indicated that DNAm is causally influencing T1D risk at the 5 non-HLA loci, DNAm – SNP relationships were observed in the same direction for 9 out
of 10 SNP-CpG pairs in both the ARIES and the BOX cohort. Although the sample size of the BOX cohort is small (n= 45) and there is a wide age range in the participants, the high replication rate in methylation patterns warrants our initial observations in ARIES [48].

Interrogating regulatory features from the Roadmap consortium revealed an enrichment of CpGs with chromosome regulatory elements. For example, cg05762488 overlaps with a DNase I hypersensitive region [49] in CD3+ primary cells and in CD20+ cells (Supplemental Figure 5). The SNP (rs2269242) associated with cg05762488 was identified as a novel T1D index variant [2] and ITGB3BP was proposed to be its candidate gene [2]. In our study cg05762488 potentially influenced the expression of a pseudogene RN7SL130P but not ITGB3BP, PGM1, or EFCAB7. The function of RN7SL130P however, is currently unknown.

The two CpGs (cg06183267 and cg07349094) associated with rs9653442 are located at the promoter of AFF3 (AF4/FMR2 Family Member 3), which overlaps with DNase I hypersensitive, H3K4me1, and H3K4me3 signals in a range of immune cells (Supplemental Figure 6). AFF3 is a risk gene for rheumatoid arthritis [50], juvenile idiopathic arthritis [51] and T1D [2, 52]; it encodes a transcription factor that may be involved in lymphoid development and plasma cell differentiation [53, 54]. Our data showed that hypomethylation at the two CpG sites potentially increase AFF3 expression, which subsequently increase T1D susceptibility.

cg22572158 is located at a previously defined enhancer upstream of the CTLA4 promoter overlapping a DNase I and H3K4me3 peak (Supplemental Figure 7) [55]. Although in this study cg22572158 did not alter CTLA4 gene expression in whole blood, the cg22572158
associated SNP (rs3087243) has been correlated with soluble \textit{CTLA4} expression in CD4+ T cells [56]. rs3087243 has also been associated with increased phosphorylation of T cell receptor downstream signalling molecules ZAP20 and SLP76 [57]. Therefore, it remains to be tested that whether cg22572158 is involved in regulating T cell function via alternative mechanisms other than gene expression.

The two rs3825932 associated CpGs (cg25744700 and cg18738367) are located in intron 1 and 5' upstream of the gene \textit{CTSH} (Cathepsin H), respectively. cg25744700 overlaps a DNase I cluster and H3K4me3 peak in a variety of immune cells (Supplemental Figure 8). Lowered gene expression of \textit{CTSH} in beta cells as well as in B-lymphoblastoid cell lines has been correlated with rs3825932 [14]. Interestingly, we found that DNAm at cg25744700 and cg18738367 exhibited opposite effect on \textit{CTSH} gene expression and this could be due to interactions with transcriptional factors with opposite functions.

The four CpG sites cg09945482, cg23544223, cg23598886, and cg24737193 were located within a CpG island 7kb downstream of \textit{PTPN2}. This region (chr18p11.21) has been associated with multiple autoimmune diseases and it contains three lincRNAs, a Y-RNA, as well as \textit{PTPN2}. We found negative effects of DNAm on \textit{RP11-973H7.1} and \textit{RP11-973H7.4} expression and a positive effect on \textit{PTPN2} expression. The four CpG sites overlap with a DNase I hypersensitivity signal and an H3K9me3 signal (Supplemental Figure 9). The detailed functional effects of these methylation variations require follow-up laboratory investigations.

One of the limitations in our study is that in the forward 2SMR, we were unable to robustly distinguish causal effect from horizontal pleiotropy (scenario 4, Figure 1d) using single genetic
instruments. More functional studies are needed to determine whether DNAm influence T1D aetiology in a causative manner by altering gene expression patterns. The second limitation is that all our analyses were performed using whole blood or peripheral blood lymphocyte samples. Despite correction for cell heterogeneity, attempts to interpret these findings in a different tissue must be conducted with caution [58]. Additionally, as supported by the mQTL enrichment analysis, DNAm could influence T1D risk in other loci via un-tested CpG sites, because 450k array typically covers 2% of the epigenome and approximately 20 CpGs per gene [31]. Large-scale bisulfite sequencing data are required to investigate this question. Finally, we reported mQTLs that regulate DNA methylation consistently at birth, childhood and adolescence, covering the spectrum of life when diagnosis of T1D peaks. It might be possible that some DNAm changes are only specific to a particular time point [12]. These changes are beyond the scope of the current analyses, future studies may focus on investigating epigenetic regulations during immune development.

In conclusion, we identified DNAm heterogeneities that are potentially causal to T1D due to common genetic risk. These heterogeneities were found to influence local gene expression, which may trigger islet autoimmunity under certain environmental stimuli. Our study provides important methylation candidates for downstream mechanistic studies as they may help us identifying novel pathways that could serve as therapeutic targets.

5. Declaration of interest
The authors declare no conflict of interests.

6. Acknowledgement
We are extremely grateful to all the families who took part in this study, the midwives for their help in recruiting them, and the whole ALSPAC team, which includes interviews, computer and laboratory technicians, clerical workers, research scientists, volunteers, managers, receptionists and nurses. We also sincerely appreciate the participants of the BOX study, nurses, administrators, and laboratory technicians for their contributions. The authors are grateful for Prof. John A Todd and Mr. Jamie Inshaw who kindly provided T1D GWAS summary statistics.

7. Web Resources

Matrix eQTL [http://www.bios.unc.edu/research/genomic_software/Matrix_eQTL/];

2SMR and MR-Steiger [https://github.com/MRCIEU/TwoSampleMR];

mQTL enrichment analysis [https://github.com/olegkagan/Ye-et-al.-2017];

JLIM [https://github.com/cotsapaslab/jlim];

Immunobase [www.immunobase.org];

UCSC genome browser [https://genome.ucsc.edu]

GTEx consortium [https://www.gtexportal.org/home]

8. Research funding

The UK medical Research Council and Wellcome (Grant ref: 102215/2/13/2) and the University of Bristol provide core support for ALSPAC. GWAS data were generated by Sample Logistics and Genotyping Facilities at the Wellcome Trust Sanger Institute and LabCorp (Laboratory Corporation of America) with support from 23andMe. Methylation data
in the ALSPAC cohort were generated as part of the UK BBSRC-funded (BB/I025751/1 and BB/I025263/1) Accessible Resource for Integrated Epigenomic Studies (ARIES). Methylation data used in the Bart’s Oxford Study (BOX) was generated using funding from Diabetes UK (14/0004869). This publication is the work of the authors; J.Y serves as guarantor for the contents of this paper. J.Y. was funded by a Diabetes Research & Wellness Foundation non-clinical fellowship N-C/2016/Ye. T.G.R. was supported by the Elizabeth Blackwell Institute Proximity to Discovery award EBI 424. M.S. was supported by the Economics and social research council ES/N000498/1. G.H. was supported by the Medical Research Council MC_UU_12013/1-9.
<table>
<thead>
<tr>
<th>SNP</th>
<th>Effect allele</th>
<th>Candidate gene</th>
<th>CpG ID</th>
<th>CpG position</th>
<th>CpG effect (SE)</th>
<th>T1D effect (SE)</th>
<th>2SMR effect (SE)</th>
<th>2SMR p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs2269242</td>
<td>A</td>
<td>ITGB3BP, PGM1</td>
<td>cg05762488</td>
<td>Chr1: 64140332</td>
<td>0.059 (0.006)</td>
<td>0.168 (0.030)</td>
<td>2.843 (0.508)</td>
<td>2.14E-08</td>
</tr>
<tr>
<td>rs9653442</td>
<td>C</td>
<td>AFF3</td>
<td>cg06183267</td>
<td>Chr2: 100759134</td>
<td>-0.100 (0.006)</td>
<td>0.124 (0.025)</td>
<td>-1.236 (0.250)</td>
<td>7.05E-07</td>
</tr>
<tr>
<td>rs9653442</td>
<td>C</td>
<td>AFF3</td>
<td>cg07349094</td>
<td>Chr2: 100759014</td>
<td>-0.157 (0.009)</td>
<td>0.124 (0.025)</td>
<td>-0.791 (0.160)</td>
<td>7.05E-07</td>
</tr>
<tr>
<td>rs3087243</td>
<td>A</td>
<td>CTLA4</td>
<td>cg22572158</td>
<td>Chr2: 204731068</td>
<td>-0.070 (0.008)</td>
<td>-0.178 (0.025)</td>
<td>2.534 (0.356)</td>
<td>1.08E-12</td>
</tr>
<tr>
<td>rs3825932</td>
<td>T</td>
<td>CTSN</td>
<td>cg25744700</td>
<td>Chr15: 79237217</td>
<td>-0.204 (0.012)</td>
<td>-0.141 (0.030)</td>
<td>0.692 (0.147)</td>
<td>2.60E-06</td>
</tr>
<tr>
<td>rs3825932</td>
<td>T</td>
<td>CTSN</td>
<td>cg18738367</td>
<td>Chr15: 79238723</td>
<td>0.093 (0.006)</td>
<td>-0.141 (0.030)</td>
<td>-1.523 (0.324)</td>
<td>2.60E-06</td>
</tr>
<tr>
<td>rs1893217</td>
<td>G</td>
<td>PTPN2</td>
<td>cg09945482</td>
<td>Chr18: 12777974</td>
<td>-0.188 (0.016)</td>
<td>0.250 (0.032)</td>
<td>-1.331 (0.170)</td>
<td>5.61E-15</td>
</tr>
<tr>
<td>rs1893217</td>
<td>G</td>
<td>PTPN2</td>
<td>cg23544223</td>
<td>Chr18: 12777786</td>
<td>-0.175 (0.011)</td>
<td>0.250 (0.032)</td>
<td>-1.428 (0.183)</td>
<td>5.61E-15</td>
</tr>
<tr>
<td>rs1893217</td>
<td>G</td>
<td>PTPN2</td>
<td>cg23598886</td>
<td>Chr18: 12777645</td>
<td>-0.414 (0.023)</td>
<td>0.250 (0.032)</td>
<td>-0.604 (0.077)</td>
<td>5.61E-15</td>
</tr>
<tr>
<td>rs1893217</td>
<td>G</td>
<td>PTPN2</td>
<td>cg24737193</td>
<td>Chr18: 12778029</td>
<td>-0.223 (0.159)</td>
<td>0.250 (0.032)</td>
<td>-1.123 (0.144)</td>
<td>5.61E-15</td>
</tr>
</tbody>
</table>

**Table 1: Representative 2SMR results from the adolescence dataset that survived the Joint Likelihood Mapping.** CpG effect denotes the addition of effect allele relative to other allele on CpG methylation changes (beta coefficient ±SE); T1D effect denotes the addition of effect allele relative to other allele on T1D risk (beta coefficient ±SE, beta coefficient equals to log OR); 2SMR effect denotes the change of log OR on T1D per SD unit (rank transformed M values) increase in DNAm due to its associated SNP. A p-value ≤ 0.05 was considered significant.
<table>
<thead>
<tr>
<th>SNP</th>
<th>Effect allele</th>
<th>Candidate gene</th>
<th>Cpg ID</th>
<th>Cpg position</th>
<th>Cpg effect (SE)</th>
<th>Expression effect (SE)</th>
<th>2SMR effect (SE)</th>
<th>2SMR p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs2269242</td>
<td>A</td>
<td>ITGB3BP</td>
<td>cg05762488</td>
<td>Chr1: 64140332</td>
<td>0.059 (0.006)</td>
<td>0.020 (0.043)</td>
<td>0.332 (0.745)</td>
<td>0.655</td>
</tr>
<tr>
<td>rs2269242</td>
<td>A</td>
<td>EFCA87</td>
<td>cg05762488</td>
<td>Chr1: 64140332</td>
<td>0.059 (0.006)</td>
<td>0.024 (0.047)</td>
<td>0.401 (0.796)</td>
<td>0.614</td>
</tr>
<tr>
<td>rs2269242</td>
<td>A</td>
<td>PGM1</td>
<td>cg05762488</td>
<td>Chr1: 64140332</td>
<td>0.059 (0.006)</td>
<td>0.038 (0.034)</td>
<td>0.640 (0.593)</td>
<td>0.280</td>
</tr>
<tr>
<td>rs2269242</td>
<td>A</td>
<td>RN7SL130P</td>
<td>cg05762488</td>
<td>Chr1: 64140332</td>
<td>0.059 (0.006)</td>
<td>-0.171 (0.078)</td>
<td>-2.900 (1.31)</td>
<td>0.027</td>
</tr>
<tr>
<td>rs9653442</td>
<td>C</td>
<td>AFF3</td>
<td>cg06183267</td>
<td>Chr2: 100759134</td>
<td>-0.100 (0.006)</td>
<td>0.086 (0.024)</td>
<td>-0.856 (0.244)</td>
<td>0.00046</td>
</tr>
<tr>
<td>rs9653442</td>
<td>C</td>
<td>AFF3</td>
<td>cg07349094</td>
<td>Chr2: 100759134</td>
<td>-0.157 (0.009)</td>
<td>0.086 (0.024)</td>
<td>-0.545 (0.156)</td>
<td>0.00046</td>
</tr>
<tr>
<td>rs3087243</td>
<td>A</td>
<td>CTLA4</td>
<td>cg22572158</td>
<td>Chr2: 204731068</td>
<td>-0.070 (0.008)</td>
<td>0.035 (0.036)</td>
<td>-0.498 (0.513)</td>
<td>0.332</td>
</tr>
<tr>
<td>rs3825932</td>
<td>T</td>
<td>CTSH</td>
<td>cg25744700</td>
<td>Chr15: 79237217</td>
<td>-0.204 (0.012)</td>
<td>-0.278 (0.036)</td>
<td>1.362 (0.177)</td>
<td>1.37e-14</td>
</tr>
<tr>
<td>rs3825932</td>
<td>T</td>
<td>CTSH</td>
<td>cg18738367</td>
<td>Chr15: 79238723</td>
<td>0.093 (0.006)</td>
<td>-0.278 (0.036)</td>
<td>-2.987 (0.388)</td>
<td>1.37e-14</td>
</tr>
<tr>
<td>rs1893217</td>
<td>G</td>
<td>PTPN2</td>
<td>cg09945482</td>
<td>Chr18: 12777974</td>
<td>-0.188 (0.016)</td>
<td>-0.090 (0.036)</td>
<td>0.477 (0.1917)</td>
<td>0.0128</td>
</tr>
<tr>
<td>rs1893217</td>
<td>G</td>
<td>RP11-973H7.1</td>
<td>cg09945482</td>
<td>Chr18: 12777974</td>
<td>-0.188 (0.016)</td>
<td>0.242 (0.0580)</td>
<td>-1.29 (0.308)</td>
<td>2.90e-5</td>
</tr>
<tr>
<td>rs1893217</td>
<td>G</td>
<td>RP11-973H7.4</td>
<td>cg09945482</td>
<td>Chr18: 12777974</td>
<td>-0.188 (0.016)</td>
<td>0.287 (0.0781)</td>
<td>-1.527 (0.416)</td>
<td>0.000237</td>
</tr>
<tr>
<td>rs1893217</td>
<td>G</td>
<td>PTPN2</td>
<td>cg23544223</td>
<td>Chr18: 12777786</td>
<td>-0.175 (0.011)</td>
<td>-0.090 (0.036)</td>
<td>0.513 (0.206)</td>
<td>0.0128</td>
</tr>
<tr>
<td>rs1893217</td>
<td>G</td>
<td>RP11-973H7.1</td>
<td>cg23544223</td>
<td>Chr18: 12777786</td>
<td>-0.175 (0.011)</td>
<td>0.242 (0.0580)</td>
<td>-1.385 (0.331)</td>
<td>2.90e-5</td>
</tr>
<tr>
<td>rs1893217</td>
<td>G</td>
<td>RP11-973H7.4</td>
<td>cg23544223</td>
<td>Chr18: 12777786</td>
<td>-0.175 (0.011)</td>
<td>0.287 (0.0781)</td>
<td>-1.641 (0.446)</td>
<td>0.000237</td>
</tr>
<tr>
<td>rs1893217</td>
<td>G</td>
<td>PTPN2</td>
<td>cg23598886</td>
<td>Chr18: 12777645</td>
<td>-0.414 (0.023)</td>
<td>-0.090 (0.0360)</td>
<td>0.217 (0.0870)</td>
<td>0.0128</td>
</tr>
<tr>
<td>rs1893217</td>
<td>G</td>
<td>RP11-973H7.1</td>
<td>cg23598886</td>
<td>Chr18: 12777645</td>
<td>-0.414 (0.023)</td>
<td>0.242 (0.0580)</td>
<td>-0.586 (0.140)</td>
<td>2.90e-5</td>
</tr>
<tr>
<td>rs1893217</td>
<td>G</td>
<td>RP11-973H7.4</td>
<td>cg23598886</td>
<td>Chr18: 12777645</td>
<td>-0.414 (0.023)</td>
<td>0.287 (0.0781)</td>
<td>-0.694 (0.189)</td>
<td>0.000237</td>
</tr>
<tr>
<td>-----------</td>
<td>---------</td>
<td>--------------</td>
<td>------------</td>
<td>-----------------</td>
<td>----------------</td>
<td>----------------</td>
<td>----------------</td>
<td>----------</td>
</tr>
<tr>
<td>rs1893217</td>
<td>G</td>
<td>PTPN2</td>
<td>cg24737193</td>
<td>Chr18: 12778029</td>
<td>-0.223 (0.159)</td>
<td>-0.090 (0.0360)</td>
<td>0.402 (0.161)</td>
<td>0.0128</td>
</tr>
<tr>
<td>rs1893217</td>
<td>G</td>
<td>RP11-973H7.1</td>
<td>cg24737193</td>
<td>Chr18: 12778029</td>
<td>-0.223 (0.159)</td>
<td>0.242 (0.0580)</td>
<td>-1.087 (0.260)</td>
<td>2.90e-5</td>
</tr>
<tr>
<td>rs1893217</td>
<td>G</td>
<td>RP11-973H7.4</td>
<td>cg24737193</td>
<td>Chr18: 12778029</td>
<td>-0.223 (0.159)</td>
<td>0.287 (0.0781)</td>
<td>-1.288 (0.350)</td>
<td>0.000237</td>
</tr>
</tbody>
</table>

**Table 2:** Casual relationship between DNA methylation (DNAm) and expression levels of genes that are located adjacent to the 10 CpG sites. 2SMR was used to assess the potential causal effect of DNAm on whole blood gene expression. 2SMR effect denotes the change in gene expression (normalized effect size) per standard deviation unit of rank transformed M values) increase in DNAm due to its associated SNP. SNP-gene expression summary data were downloaded from the GTEx consortium. SNP-CpG summary data were generated from the EWAS of the present study. A p-value ≤ 0.05 was considered significant.
Figure legend

Figure 1: Four possible scenarios that could explain the associations between SNP, DNAm and T1D.

a, a SNP that regulates DNAm is simply in LD with another causal variant that influences T1D;
b, DNAm mediates the genetic risk of T1D; c, SNPs first increase T1D risk (i.e. via changes in gene expression), which in turn changes DNAm; d, a SNP is associated with DNAm and T1D via independent biological pathways (horizontal pleiotropy).

Figure 2: Flow chart summarising the overall analysis procedure in this study.

EWAS: epigenome wide association analysis; DNAm: DNA methylation; mQTL: methylation quantitative trait loci

Figure 3: cis-mQTLs are enriched in SNPs with low GWAS p-values associated with T1D.

a, A representative plot showing the enrichment analysis conducted using the adolescent data, when null SNPs were matched to cis-mQTLs via SNP properties; b, when null SNPs were matched to cis-mQTLs via genomic annotations. T1D GWAS p-values were extracted from meta-analysis Data 1.

Figure 4: Genomic distribution of CpG sites that are associated with T1D GWAS variants.
Manhattan plot showing CpG sites (n=96) significantly associated with 37 non-HLA T1D GWAS variants below the Bonferroni threshold 1.6e-9 (redline); these associations ) were consistently detected at adolescence, childhood and birth.

**Figure 5:** DNAm levels of CpG sites and their associations with T1D GWAS variants that survived JLIM analyses, obtained from the ARIES adolescent participants.

Values on the Y - axis represent beta values for CpG sites. The inner most genotype on the X - axis is the reference genotype.

**Figure 6:** Results of SNP and DNA methylation associations in a subset of individuals participated the Bart’s oxford T1D cohort.

45 individuals were analysed in the BOX cohort. Nine out of ten SNP-CpG pairs showed associations in the same direction as compared to the ARIES participants. The SNP – CpG pair that did not follow the same pattern as compared to the ARIES was highlighted in red.
References


Ma C, Staudt LM. LAF-4 encodes a lymphoid nuclear protein with transactivation potential that is homologous to AF-4, the gene fused to MLL in t(4;11) leukemias. Blood, 1996;87:734-45.


