bioRxiv preprint doi: https://doi.org/10.1101/248260; this version posted January 15, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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

Jody Ye[1], Tom G Richardson[2], Wendy McArdle[2], Caroline L. Relton[2], Kathleen M. Gillespie[1], Matthew Suderman[2], Gibran Hemani[2]

Author Affiliations:

[1]Diabetes and Metabolism, Bristol Medical School (Translational Health Sciences), University of Bristol, UK

[2]MRC Integrative Epidemiology Unit, Bristol Medical School (Population Health Sciences), University of Bristol, UK

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.yi.ye@bristol.ac.uk, Tel: +44 (0)117 414 8041

Key words: DNA methylation, Type 1 diabetes, Mendelian Randomization, ALSPAC, BOX

bioRxiv preprint doi: https://doi.org/10.1101/248260; this version posted January 15, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Abbreviations

ARIES	Accessible Resource for Integrated Epigenomic Studies
ALSPAC	Avon Longitudinal Study of Parents and Children
BOX	Bart's Oxford family study of Type 1 Diabetes
CpG	Cytosine-phosphate-guanine dinucleotides
GWAS	Genome-wide association study
HLA	Human leukocyte antigen
JLIM	Joint likelihood mapping
LD	Linkage disequilibrium
SNP	Single nucleotide polymorphism
T1D	Type 1 diabetes
MAF	Minor allele frequency
MR	Mendelian Randomization
mQTL	methylation quantitative trait loci
RA	Rheumatoid arthritis
2SMR	Two Sample Mendelian Randomization

bioRxiv preprint doi: https://doi.org/10.1101/248260; this version posted January 15, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Abstract

The risk of Type 1 diabetes comprises both genetic and environmental components. We investigated whether genetic susceptibility could be mediated by changes in DNA methylation, an epigenetic mechanism that potentially plays a role in autoimmune diabetes. Using data from a non-diabetic population comprising blood samples taken at birth (n=844), childhood (n=911) and adolescence (n=907), we evaluated the association between 65 top GWAS single nucleotide polymorphisms (SNPs) and genome-wide DNA methylation levels interrogating 99% RefSeq genes. We identified 159 proximal SNP-cytosine phosphate guanine (CpG) pairs (*cis*), and 7 distal SNP-CpG associations (*trans*) at birth, childhood, and adolescence. We also found systematic enrichment for DNA methylation related SNPs to be associated with T1D across the genome, after controlling for the SNPs' genomic characteristics. For each of the proximal CpG site identified, we used the principles of Mendelian Randomization to infer the putative causal relationship between DNA methylation levels and T1D. With genetic colocalization analysis, we discovered 10 CpGs at 5 loci, including *ITGB3BP*, *AFF3*, *PTPN2*, *CTSH* and *CTLA4*, where DNA methylation is potentially on the causal pathway to T1D. Nine out of ten SNP – CpG associations showed similar patterns in an independent T1D cohort (n=45). Our data imply that DNA methylation mediate the polygenic risk of T1D and dissecting their molecular mechanisms may uncover novel disease aetiologies.

Word count: 215

bioRxiv preprint doi: https://doi.org/10.1101/248260; this version posted January 15, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Significance statement

So far genome wide association studies identified 62 loci contributing to the genetic risk of Type 1 diabetes. However, the underlying mechanisms mediating genetic susceptibilities are largely unknown. DNA methylation is an epigenetic mechanism, which can be influenced by genetic polymorphisms and is potentially causal to Type 1 diabetes by altering chromatin conformation and gene expression. We investigated the causal relationships between type 1 diabetes-associated loci and DNA methylation. Our data suggest that methylation potentially mediates the genetic risk at 5 loci. These effects were consistently detected in a non-diabetic population at birth, childhood and adolescence, and the majority of which were replicated in a type 1 diabetes cohort. Dissecting their molecular mechanisms may uncover novel mechanisms of disease aetiology.

Word count:120

bioRxiv preprint doi: https://doi.org/10.1101/248260; this version posted January 15, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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 (1). Beta cell autoimmunity is thought to result from impaired tolerance to islet autoantigens, which were presented on the HLA complex to autoreactive T cells. Many non-HLA genes also contribute to the dysregulation of the immune system with relatively small effects. Over the last decade, genome-wide association studies have identified 62 independent loci and over 100 GWAS SNPs associated with T1D risk (2), but the most associated variants are not necessarily causal due to linkage disequilibrium (LD) with many other SNPs. Fine mapping has pinpointed a number of credible variants, many of which are localized to enhancers, implying that they may influence disease through gene regulation (3).

DNA methylation and histone modification are epigenetic events that can regulate gene expression. DNA methylation occurs at cytosine – phosphate – guanine (CpG) residues and can be modified by genetic and / or environmental exposure. Genetic and epigenetic interactions have been postulated to contribute to susceptibility to a number of autoimmune disorders (4, 5).

Previous work was focused on identifying methylation differences among T1D monozygotic twins where genetic differences, age, gender, and *in utero* environmental effects are normalized (6), allowing changes of DNA methylation levels that are purely introduced by non-shared environment to be captured (7-9) . It is however unknown whether the genetic susceptibility of T1D is mediated by changes of DNA methylation, which subsequently lead to altered gene expression and immune cell function.

bioRxiv preprint doi: https://doi.org/10.1101/248260; this version posted January 15, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Mendelian Randomization (MR) is a statistical framework to infer causal relationships in hypothesis testing, using genetic variants to create a pseudo-experimental design. Since genetic polymorphisms are randomly assigned at conception and not influenced by environmental confounders, they can be used as instruments to proxy exposures that are potentially influencing a trait, thereby mimicking a randomised controlled trial (10). 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 (11). If a CpG site mediates genetic risk of T1D, the casual effect of this CpG site can be interpreted as the change in log odds for T1D per unit increase in the CpG methylation level due to its associated SNP. 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 (10). When a SNP is associated with both DNA methylation and T1D, four potential scenarios can occur:

(1) the genetic variant has a causal effect on T1D mediated by the changes of DNA methylation levels (illustrated in Figure 1a);

(2) the genetic variant has a causal effect on T1D (i.e. via altering gene expression), which in turn alters DNA methylation levels (Figure 1b);

(3) the genetic variant that causes changes in DNA methylation levels is in LD with the causal variance of T1D (Figure 1c);

(4) the genetic variant causes changes in DNA methylation levels and T1D via separate mechanisms, an effect known as horizontal pleiotropy (Figure 1d).

bioRxiv preprint doi: https://doi.org/10.1101/248260; this version posted January 15, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

In this study, we aimed to determine whether the genetic risk of T1D can be mediated by DNA methylation (scenario 1, Figure 1a) and to distinguish scenarios 2 & 3 from scenario 1. Firstly, in a non-diabetic population (Accessible Resource For Integrated Epigenomic Studies, also known as the ARIES cohort) we performed an epigenome-wide association analysis identifying CpGs that are related to top T1D GWAS variants. Secondly, we tested whether there is an overall association between cis-mQTLs and T1D. Subsequently, following the framework outlined in Richardson et al 2017 (12), we combined the principles of 2SMR with genetic colocalization methods to assess whether these GWAS variants related CpG sites mediate T1D genetic risk. Finally, we tested whether the findings can be replicated in an independent cohort with T1D patients and their relatives (the BOX cohort). A flow chart summarising the analysis procedure is shown in Figure 2.

Results

The association between DNA methylation and T1D susceptible SNPs across three-time points

To establish associations between top T1D GWAS variants and DNA methylation, we regressed 65 independent SNPs with 459,734 CpG sites using a mixed effect linear model. Whole blood DNA methylation levels of these CpG sites were measured in the ARIES cohort at three-time points during life (adolescence, childhood and birth). Of the 65 independent T1D GWAS variants, thirty-eight SNPs were consistently found to associate with a total of 166 CpG sites under the Bonferroni corrected threshold (0.05/65*459,734, $p < 1.6$ e-9) at three-time points. Seven T1D variant-CpG pairs were in trans (distance >1Mb) and the remaining were in cis; these data are summarised in SI Table 1. Figure 3 shows the genomic

bioRxiv preprint doi: https://doi.org/10.1101/248260; this version posted January 15, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

distribution of 166 total CpGs, including extensive associations at the HLA locus. According to 450k array annotation, approximately 45% CpGs were located in the gene body / introns; 24% were located in promoters / distal promoters. Methylation variance (R^2) explained by T1D SNPs varied largely, the strongest association lies between rs7149271 and cg20045882 on chromosome 14, where rs7149271 explained greater than 78% methylation variance across all three-time points. At the HLA locus, rs3104163 is in high LD with rs9272346 (r^2=0.84), a variant most strongly associated with T1D (T1D OR=18.5). rs3104363 regulates 70 CpGs within the HLA locus and the most strongly associated CpG site was cg01889448, for which rs3104363 explained at least 53% methylation variance across three-time points. However, it is important to note that the estimates of methylation variances explained by T1D SNPs might be inflated by the "winner's curse". The overall effect sizes of T1D variants on DNA methylation levels are consistent at three-time points, where the correlation between adolescence and childhood is 0.996 (95% CI: 0.995, 0.997, p < 2.2e-16); between adolescence and birth is 0.984 (95% CI: 0.979, 0.989, p < 2.2e-16); and between childhood and birth is 0.987 (95% CI: 0.982, 0.990, p < 2.2e-16).

cis-mQTLs are enriched in SNPs with low GWAS *p*-values associated with Type 1 diabetes

We next investigated whether cis-mQTLs are associated with T1D more than expected by chance. We hypothesized that cis-mQTLs would have more extreme *p*-values associated with T1D than non-mQTLs that are matched by SNP properties (allele frequency, LD, gene distance), or by genomic annotations (i.e. promoter, intron, exon, 5' UTR, or 3' UTR). As a primary analysis, we overlapped cis-mQTLs detected at adolescence that are known to pose strong effects on DNA methylation (p < 1e-14) (13) with an initial discovery GWAS

bioRxiv preprint doi: https://doi.org/10.1101/248260; this version posted January 15, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

summary statistic dataset (Data 1) and obtained 4,562 cis-mQTLs together with their T1D GWAS *p*-values. 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 4 a) or by genomic annotations (Figure 4 b). Secondary analyses using cis-mQTLs detected at childhood and birth revealed the same findings (SI Figure 1). To verify these observations, we performed the same analyses using the replication summary statistics (Data 2). Compared to enrichment analysis using Data 1, there was a stronger enrichment when cis-mQTLs were matched by SNP properties (SI Figure 1) and when cis-mQTLs were matched by genomic annotations (SI Figure 1). These data suggest that there is a shared genetic influence of DNA methylation levels and T1D. Furthermore, from each of the enrichment analyses we identified a number of cis-mQTLs that have smaller observed GWAS *p*-values than theoretical *p*-values (the probability of T1D association by chance). Most of these cis-mQTLs lie within known T1D susceptible loci and are in strong to moderate LD with index SNPs. However, rs605093 resides at Chr11q24.3 that was not reported by previous GWA studies or in LD with any index variants ($p_{\text{observed. meta}} = 4.22\text{x}10^{-6}$, $p_{\text{theoretical}} = 1.01\text{x}10^{-5}$).

DNA methylation potentially mediate T1D genetic risk

To assess the causal effect of DNA methylation levels for each of the detected proximal CpG sites on T1D risk, we performed forward 2SMR using CpGs as individual exposures and T1D as outcome (Figure 1a). After removing SNPs in high LD, SNPs on X chromosomes, and those without odds ratios as well as palindromic SNPs with harmonizing issues (14). One hundred and twenty-eight CpG associations unique to 33 cis-mQTLs remained. CpGs with

bioRxiv preprint doi: https://doi.org/10.1101/248260; this version posted January 15, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

trans associations were removed from the analysis in order to minimize the possibility of horizontal pleiotropy because trans-mQTLs may regulate CpGs and T1D via different pathways. We ensured that none of the genetic instruments was associated with confounder - fasting glucose concentrations as identified in the European GWAS meta-analysis (p < 5e-8) (15). Results showed that the Wald ratios for all the 128 CpGs were significant and the effects were consistent at adolescence, childhood and birth (p < 0.05, SI Table 2).

Forward 2SMR using a single genetic instrument could suffer from weak instrumental bias. In addition, reverse causation and horizontal pleiotropy are not easily distinguishable (as shown in Figure 1 b & d, respectively). To test the causal effect of T1D liability on DNA methylation, we performed reverse MR using T1D as exposure and CpG sites as outcomes. 128 CpGs were tested as outcomes individually and multiple mQTLs (instruments) were used for exposure (Figure 1b). Since multiple testing needs to be accounted for the outcomes, but some of the tests are not independent due to correlations of the tested CpG sites (i.e. CpGs are co-methylated if they are located close to each other on a chromosome), we used matSpDlite to estimate the number of independent CpGs. It suggested that there were 124 independent CpGs in the outcome and a p-value of 4.12e-4 is 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. However, it is important to note that the statistical power to detect an effect in this direction is low because the outcome sample size was small. These data are summarised in SI Table 2.

MR-Steiger test to verify the direction of causality

bioRxiv preprint doi: https://doi.org/10.1101/248260; this version posted January 15, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

We also used the MR-Steiger test to verify the findings in 2SMR. MR-Steiger estimates whether DNA methylation or T1D is more likely to be the exposure by testing whether mQTLs primarily explain the variance of methylation or T1D (16). Results showed that within the HLA region, SNPs explained methylation variance more than T1D variance for most CpG sites (36 out of the 64 CpGs showed methylation as the exposure consistent at three-time points, $p < 0.01$, sensitivity ratio >2.35, SI Table 3). Outside the HLA, the same causal direction was inferred for all the CpGs and this effect was consistent at all three-time points (SI Table 3). We did not perform MR-Steiger test at the reverse direction due to insufficient statistical power to detect an effect at this direction.

Bivariate fine mapping pinpointed overlapping methylation and T1D causal variants

To exclude the possibility that mQTLs are simply in LD with T1D causal variants (illustrated in Figure 1c), we searched for evidence of the shared causal variants for the 128 CpG sites and for T1D within the 1Mb window centred around their top associated mQTLs. For the 32 non-HLA loci, although most causal variants for DNA methylation were simply in LD with the causal variants for T1D, JLIM analysis found colocalization of shared causal variants in 5 loci, mediated by 10 CpG sites in total ($p < 1e\text{-}3$, Table 1, Figure 5). CpGs in the HLA region were however excluded from the analysis owing to the extensive LD structure and high false discovery rate (17).

Replication in T1D cohort

To check whether genetic and CpG relationships discovered in the ARIES general population can be replicated in a T1D population, we assessed the 10 SNP - CpG associations that survived JLIM analysis in 45 individuals participating the Bart's Oxford family study of type

bioRxiv preprint doi: https://doi.org/10.1101/248260; this version posted January 15, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

1 diabetes (BOX), where genome-wide methylation data were available. In this cohort T1D probands together with their parents and grandparents with or without diabetes were analysed (detailed clinical characteristics are summarised in SI Table 4). Nine out of ten associations showed similar patterns comparing to the ARIES cohort, after fitting the DNA methylation levels and SNP genotype into linear regression models (Figure 6). However, given the small sample size of the replication cohort, there is insufficient power to obtain significant *p* values for cg05762488 and for cg025744700 (Figure 6).

Discussion

One hypothesis of the mechanisms underlying T1D is that genetic variants alter DNA methylation levels, which in turn influence genes that are essential to immune tolerance as well as beta cell function, increasing the risk of T1D. To the best of our knowledge, this is the first study that systematically evaluated this hypothesis. We showed that of the 65 T1D top GWAS variants, 38 influence DNA methylation levels of 166 CpG sites consistently at birth, childhood and adolescence; 40% of CpGs were located at the HLA region and strongly associated with rs3104363; cis-mQTLs were found to influence T1D more than expected by chance. Using the principles of Mendelian randomization, we showed that DNA methylation potentially mediate the polygenic risk of T1D in many loci. However, subsequent joint likelihood mapping restricted these to 5 non-HLA loci where T1D susceptibility is putatively mediated by the differential methylation levels at 10 CpGs. In an independent T1D cohort containing 45 individuals, we observed similar patterns in nine of ten SNP-CpG pairs.

Our EWAS analysis identified widespread genetic and epigenetic associations in the known T1D susceptible loci. The only previous study by Fradin et al., took a candidate gene

bioRxiv preprint doi: https://doi.org/10.1101/248260; this version posted January 15, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

approach and investigated 7 CpG sites at the insulin promoter in a total of 802 individuals using pyrosequencing (18). Our data generally agree with theirs, but we showed rs689 dependent associations with cg21574853 and cg24338752. The differences observed may be because we adjusted for cellular composition in whole blood data and removed potential confounders, whereas Fradin et al., could not perform such adjustment when methylation levels were measured using pyrosequencing. Unfortunately, there were no beta coefficients available for rs689 and its proxy SNPs in our GWAS summary statistics, we therefore could not proceed to test the causality of methylation at the *INS* locus.

We observed strong associations between cis-mQTLs and T1D in the enrichment analyses based on GWAS meta-analysis *p*-values. However, a previous enrichment analysis based on SNP heritability did not show significant association with T1D (13). Narrow sense heritability of T1D was estimated to be approximately 0.8 (19, 20), T1D was thus considered highly heritable. 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 (13), more cis-mQTLs are perhaps required to capture enough genotypic variance to explain a highly heritable condition. The number of cis-mQTLs is less of a concern for *p*-value based enrichment analysis because rather than testing a null hypothesis of SNP heritability, it tests the null hypothesis of uniformly distributed *p*-values. However, one caveat of our analysis is that there was an overlapping control population (WTCCC control samples) between the data source of Data 1 and Data 2. Data 2 was therefore not completely independent of Data 1. We found that rs605093 (chr11: 128604232, hg19) was associated with T1D more than expected by chance. In a previous T1D meta-analysis, rs605093 was detected as one of the most associated T1D SNPs but failed to reach genome-wide significance, although no follow-up

bioRxiv preprint doi: https://doi.org/10.1101/248260; this version posted January 15, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

study was performed to confirm its true association (21). Since GWAS tends to omit variants with small effect sizes due to the burden of multiple testing, our data may suggest a weak effect of rs605093 on T1D. This SNP is located at the intron 1 of *FLI-1* (Friend leukemia integration 1 transcription factor), which overlaps with a known susceptible region for rheumatoid arthritis (RA) and systemic lupus erythematosus (SLE). It is possible that this locus contributes to autoimmune risk via a shared mechanism.

In the HLA locus, the majority of the rs3104363 associated CpGs were tested as potentially causal to T1D at the forward direction. Previously, a similar conclusion at the HLA region was reached by a study of Rheumatoid arthritis (RA) using causal inference test (CIT), a regression based approach rather than mendelian randomization, which is an instrumental variable approach (4, 16). Since T1D and Rheumatoid arthritis (RA) share *HLA-DRB1* susceptibilities (22), our findings generally support a mediatory role of DNA methylation. Due to the extensive LD structure within the HLA region, however, we cannot use JLIM to reliably estimate whether there are shared causal variants for methylation and T1D, and thus cannot rule out the possibility that causal variants for DNA methylation and causal variants for T1D are simply in LD. As high-risk variants at the HLA regions are classically thought to influence autoantigen presentation by modulating the affinity and conformation between the peptide and the HLA binding pocket (23), whether DNA methylation is truly involved in mediating HLA risk requires further functional confirmation.

In non-HLA loci our data suggest a potential functional role of DNA methylation in T1D risk. The 5 loci included genes that are known to alter immune / islet cell function, such as *CTLA4*, *CTSH* and *PTPN2*. Interrogating regulatory features from the ENCODE/Roadmap

bioRxiv preprint doi: https://doi.org/10.1101/248260; this version posted January 15, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

consortium revealed that CpGs within 3 of 5 loci were overlapping with apparent chromosome regulatory elements (SI Figure 2). For example, the rs2269242 (tagging *ITGB3BP*) associated CpG site cg05762488 is located within a dense DNase I hypersensitive region and transcription factor binding site adjacent to the gene *PGM1* (phosphoglucomutase -1) and upstream of the gene *ITGB3BP* (integrin beta 3 binding protein beta3-endonexin). *ITGB3BP* is a new candidate gene to T1D identified in a recent study (2), it encodes a transcriptional coregulator that is involved in signalling pathways of apoptosis (24). The two rs9653442 (tagging *AFF3*) related CpG sites cg06183267 and cg07349094 are situated in the exon1 of *AFF3* (AF4/FMR2 Family Member 3), which is a region enriched with DNase I hypersensitivity and H3K4me3 (associates with enhancers) signals in a range of immune cells. *AFF3* is a risk gene for rheumatoid arthritis (RA) (25), juvenile idiopathic arthritis (26) and T1D (2, 3); it encodes a nuclear transcriptional activator that is preferentially expressed in lymphoid tissue, which may be involved in lymphoid development and plasma cell differentiation (27, 28). It has also been shown to contribute to anti-TNF treatment responses in RA patients (29). The two rs3825932 (tagging *CTSH*) associated CpG sites cg25744700 and cg18738367 are located in intron 1 and 5' upstream of the gene *CTSH* (Cathepsin H), respectively. cg25744700 overlaps with a H3K27Ac peak (marks active chromatin), DNase I cluster, as well as a transcription factor binding region; cg18738367 co-localizes with a DNase I cluster. Lowered gene expression of *CTSH* in beta cells has been correlated with increased beta cell apoptosis upon cytokine exposure as well as faster diabetes progression (30). These data imply potential gene regulatory functions of the identified CpG sites, for which laboratory investigations are worth to follow up.

There are several limitations of our study. Firstly, given single genetic instruments used in the forward 2SMR, we were unable to robustly distinguish whether genetic risk influences DNA

bioRxiv preprint doi: https://doi.org/10.1101/248260; this version posted January 15, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

methylation and T1D separately via horizontal pleiotropy. For example, if a causal variant really influences gene expression level first and that subsequently influences DNA methylation and T1D together, then this will give a false positive result when testing the mediatory effect of DNA methylation using forward 2SMR. It is therefore recommended to verify the 2SMR findings using laboratory approaches. Secondly, regardless of non-HLA and HLA loci, 2SMR analyses, as well as the MR-Steiger directionality test, suffer from insufficient statistical power to detect an effect in the reverse direction. This is because the sample size available for SNP effects on CpG levels was small (approximately 1000 individuals in the ARIES participants per time point). Thirdly, although JLIM analysis suggested the potential mediatory roles of 10 CpG sites, JLIM does not report which is the potential shared causal variants between DNA methylation and T1D, leaving the true causal variants to be pinpointed by other statistical and/or biological methods. Fourthly, all our analyses were performed using whole blood or peripheral blood lymphocyte samples. As the mQTL effects are likely to be tissue and cell type specific (31), attempts to interpret our findings in a different tissue must be conducted with caution. Another limitation is that the methylation 450k array typically interrogates on average 17 CpG sites per gene (32) and we were unable to investigate causal effects of uncovered CpG sites, it is thus possible that DNA methylation may mediate T1D susceptibility via other regions. Finally, in this study we only 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 mediatory effects of DNA methylation are only specific to a particular time point, especially knowing that environmental impact on DNA methylation increases later in life and genetic determinants of methylation heritability decreases (13). These changes are beyond the scope of the current analyses.

bioRxiv preprint doi: https://doi.org/10.1101/248260; this version posted January 15, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Within these limitations, we observed similar association patterns in 9 out of 10 SNP-CpG pairs in an independent cohort containing T1D probands and their relatives. Given a small sample size in this cohort (n= 45), we were unable to obtain significant associations for all the SNP-CpG pairs. Because all the T1D probands were children and their non-diabetic relatives were older, stratification of methylation levels by disease status was not possible due to confounding by age. Therefore, we could only plot our data against genotype. In a previous T1D monozygotic twin study, Paul et al., suggested that variations in DNA methylation found in T1D affected twins were not associated with any post-zygotic genetic mutations (8) and therefore not genetically driven. Post-zygotic mutations occur at a frequency of 1.2e-7 per base pair per twin pair (33), these rare variants are unlikely to provide sufficient power to detect genetic and epigenetic associations. Our analyses using common genetic variants thus have an advantage of increased power to detect such an effect.

In conclusion, the identification of putative genetically driven DNA methylation changes provides a rich source for follow-up functional verifications, as dissecting genetic and epigenetic interactions may help to uncover novel mechanism that contribute to the risk of T1D development.

Materials and methods

DNA methylation data

Non-diabetic general population. DNA methylation data was obtained from the Avon Longitudinal Study of Parents and Children (ALSPAC) study, a large scale prospective study based in Avon, UK. ALSPAC recruited 14,541 pregnant women with expected delivery dates

bioRxiv preprint doi: https://doi.org/10.1101/248260; this version posted January 15, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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 (34, 35). Please note that 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/. DNA methylation data were generated either from whole blood or from peripheral blood leukocyte samples, derived from 1,018 mother-offspring pairs using the Illumina HumanMethylation450 BeadChip ('450k array') (Illumina, San Diego, CA, USA) as part of the Accessible Resource for Integrated Epigenomic Studies (ARIES) project (36). The array quantifies DNA methylation levels of >485,000 CpG sites, which covers 99% RefSeq genes (32). The ARIES participants were selected based on availability of DNA samples at two-time points for the mother (antenatal and at follow-up when the offspring was adolescents) and three-time points for the offspring (cord blood, childhood and adolescence). Methylation data from the offspring at adolescence (mean age 17.1 years, n=907), childhood (mean age 7.5 years, n=911) and birth (mean range, n=844) were used in this study. For quality control, probes with low signal to noise ratio (detection *p*-value > 0.01) or with methylated or unmethylated read counts of 0 were removed. Additionally, SNPs on 450k array were compared with individual SNP-chip data on same individuals, samples with mismatched genotypes were removed. Probes with SNPs at the CpG site as well as probes with SNPs located at the single base extension site at any minor allele frequency were removed using annotations in the Minfi package, batch effect was normalized using the wateRmelon package (13). To retain the maximum number of CpG candidates, we did not remove probes that contain SNPs greater than 10 nucleotides from the query CpG site, as they were shown to have negligible influences on the beta values of the query CpG site (37). To account for cellular heterogeneity, B cell, CD4+ T cell, CD8+ T cell, granulocyte, monocyte, and NK cell composition in each sample was estimated using

bioRxiv preprint doi: https://doi.org/10.1101/248260; this version posted January 15, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

a Reference-based algorithm developed by Houseman et al (38). To remove outliers, bimodally distributed beta values were converted to M values (39) and then rank-transformed into normal distributions. The final methylation data contained 459,734 probes per individual.

T1D population. As part of the methylation study, 16 families including 45 individuals were selected from the Bart's Oxford (BOX) family study of type 1 diabetes (40). Probands (mean age ± SD: 11.3 ± 3.6 years), their parents (mean age ± SD: 41.8 ± 5.1 years) and grandparents (mean age ± SD: 72.7 ± 7.5 years) were analysed. Parents and grandparents may or may not have T1D, these individuals were summarized in SI Table 2. Briefly, DNA was extracted from whole blood, approximately 350ng DNA per sample were bisulfite converted and loaded onto the Infinium Methylation 450k BeadChip (Illumina). Samples were processed in Population Health Sciences, University of Bristol. DNA methylation data were normalized using SWAN normalization and further corrected for batch effect, age, gender and cell heterogeneity using the Minfi and SVA package under the R programming environment (version 3.2.2). methylation levels were then plotted against SNP genotypes.

Individual level genotype data

General population. Individual level genotype data on the ARIES cohort were generated using Illumina HumanHap550-quad chips by Sample Logistics and Genotyping Facilities at the Wellcome Trust Sanger Institute and LabCorp (Laboratory Corporation of America). For quality control, individuals with gender mismatches, minimal or excessive heterozygosity or >3% missingness on genotype data were removed. SNPs with minor allele frequencies (MAF) of >1%, a call rate of <95% or violations of Hardy-Weinberg equilibrium were removed. Imputation of unmeasured genotypes was performed using IMPUTE2 based on the

bioRxiv preprint doi: https://doi.org/10.1101/248260; this version posted January 15, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

1000 genomes phase 1 version 3 as a reference panel (36, 41, 42). 113 SNPs spanning 57 genomic regions that are associated with T1D at genome-wide significant level were obtained from immunobase.org, including six additional SNPs associated with T1D from a recent GWAS study (2). Among them, sixty-seven independent SNPs (LD $r^2 < 0.1$) were selected and where necessary, proxy SNPs (minimal $r^2 = 0.6$) were used to replace the original variants in order to obtain the required odds ratios for downstream MR analysis. Genotype data of sixty-five SNPs were available and extracted from the ARIES participants, which were summarised in SI Table 5.

T1D population. Taqman probes for rs2269242, rs9653442, rs3087243, rs3825932, and rs1893217 were purchased from Life Technologies (Thermo Fisher, UK). SNP genotypes for T1D probands as well as their parents and grandparents were determined using Taqman® allele discrimination assays (Life Technologies, UK).

mQTL-CpG association analyses

A mixed effect linear regression model was used. Typically, rank transformed M values of the methylation data were regressed against each of the 65 T1D GWAS variants, with age, gender, cellular compositions also included as covariates in the model. A Bonferroni threshold of 1.6×10^{-9} (0.05/65*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.

T1D GWAS summary statistics

bioRxiv preprint doi: https://doi.org/10.1101/248260; this version posted January 15, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Summary statistics were obtained from meta-analyses: one for the initial analysis (Data 1) and one for replication (Data 2). Data 1 combined Affymetrix and Illumina genotyping data derived from the UKBS, 1958 Birth cohort, WTCCC Bipolar disorder samples, and UK GRID cohort(2). The final summary statistics were obtained by comparing 5913 cases and 8828 controls. Results contained *p*-values, odds ratios, regression coefficients (log odds ratios) and their standard errors; 9,037,957 SNPs were available in this dataset. Data 2 combined Affymetrix and Illumina genotyping data derived from the McGill University cohort, Children's hospital Philadelphia (CHOP) cohort, DCCi-EDIC cohort, T1DGC, GoKinD, and WTCCC (43). The summary statistics was obtained by comparing 9,934 cases and 16,956 controls (43). Data were retrieved from Immunobase.org; only *p*-values from this meta-analysis were available; 2,060,920 SNPs were available in this dataset. For the raw data where both summary statistics were derived from, there were some population overlap in the control samples from WTCCC, which were summarised in SI Table 6.

mQTL summary data

mQTL summary data was obtained from ARIES participants from a previous study (13), in which 450k DNA methylation data derived from blood samples of children collected at birth, childhood, adolescence as well as blood samples collected from their mothers at pregnancy and middle age were regressed against individual level genotype at genome-wide scale. CpGs that showed significant associations (p < 1e-14, Type I error ate 0.2%) with SNPs were retrieved using the TwoSampleMR R package.

mQTL enrichment analyses

bioRxiv preprint doi: https://doi.org/10.1101/248260; this version posted January 15, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

To test the hypothesis that mQTLs are enriched in SNPs with low T1D GWAS *p*-values more than expected by chance, independent mQTLs (LD $r^2<0.1$) that have shown strong effect ($p<1x10^{-14}$) on DNA methylation during adolescence in ARIES cohort (13) were retrieved using the TwoSampleMR R package. The likelihood of mQTL enrichment could either be due to their 1) distinct SNP properties or 2) due to distinct genomic locations. To control for these two factors, null SNPs were selected to match mQTLs in two ways. Firstly, null SNPs were chosen based on similarities in MAF and LD structures (44). Briefly, null SNPs must be at least 1000kb away from mQTLs; the maximum MAF deviation of null SNPs from mQTLs is 0.02; and LD scores of null SNPs are in the same quintile bin of 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 for all the mQTLs:

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

To generate a distribution for null SNPs, the same number of null SNPs (4,562) was randomly drawn and 10,000 iterations were generated. Fisher's combined probability test was used to estimate 10,000 combined *p*-values for null SNPs. The empirical *p*-value, reflecting the likelihood of observing a combined *p*-value at least as extreme as the combined *p*-value for mQTLs in the null distribution, is calculated by ranking all the 10,000 null *p*-values.

Two-sample bi-directional Mendelian Randomization

To test the causal effect of DNA methylation on T1D genetic susceptibility, forward 2SMR was used (Figure 1a). To be considered as valid instruments to proxy DNA methylation,

bioRxiv preprint doi: https://doi.org/10.1101/248260; this version posted January 15, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

SNPs must meet three key assumptions (45). First, SNPs must be strongly associated with DNA methylation; second, SNPs must only influence T1D via DNA methylation; third, SNPs must be independent of confounders of the methylation-T1D associations (i.e. hyperglycemia and medications). The associations between mQTLs and CpG sites (beta coefficient) were established using the ARIES cohort (sample 1). The associations between mQTLs and T1D (log odds) were obtained from the GWAS summary statistics Data 1 (sample 2). To reduce potential pleotropic effect (SNPs that influence multiple CpG sites via independent pathways), only cis-mQTLs were chosen as instruments. This was necessary as the majority of CpG sites in this analysis can only be instrumented using a single cis-acting mQTL, which means we were unable to robustly investigate pleiotropy. To exclude potential instrument-confounder associations, we also examined whether instruments were associated with fasting glucose concentration in a large GWAS meta-analysis involving 133,010 non-diabetic European individuals (15). The causal effect of CpG to T1D was then determined using a Wald ratio estimator, calculated by dividing the log odds of cis-mQTLs on T1D association by the beta coefficient of cis-mQTLs on CpG association (46).

To test the causal effect of T1D on DNA methylation, T1D GWAS SNPs were used as multiple instruments for each CpG as outcome. For each CpG site, mQTL that was used as an instrument in the forward 2SMR was excluded. The causal effects of multiple SNPs were combined in a fixed-effect meta-analysis using MR – inverse variance weighting (IVW). All the above analyses were performed using the TwoSampleMR R package.

We used matrix spectral decomposition (matSpDlite) (47) to determine the number of independent tests in the outcome. “matSpDlite” calculates the number of independent

bioRxiv preprint doi: https://doi.org/10.1101/248260; this version posted January 15, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

variables in the correlation matrix generated from all the CpG sites, by examining the ratio of observed eigenvalue variance to its theoretical maximum (47). It also reports an alpha level that is required to keep the Type I Error rate at 5%.

MR-Steiger directionally test

2SMR estimates the causal effect under an important assumption that the exposure is known. This however, in some situations particularly in the case of DNA methylation, is difficult to ascertain as it is unclear whether genetic risk first causes changes in DNA methylation which subsequently results in T1D risk or vice versa. To evaluate this, we used MR-Steiger to access whether DNA methylation is likely the exposure and T1D risk is likely the outcome. MR-Steiger estimates the proportion of variance in the exposure and in the outcome that is explained by genetic instruments. Causal direction is then determined based on whether exposure variance or outcome variance is subject to the primary effect of SNPs (16). This was performed in TwoSampleMR R package.

Bivariate fine mapping

Another MR assumption is that cis-mQTLs only influence T1D via DNA methylation. This is however not always true because some cis-mQTLs may simply be in LD with a causal variant influencing T1D (Figure 1 c). To this end, we implemented joint likelihood mapping (JLIM) (17) to investigate whether causal variants for DNA methylation (GWAS SNPs that cause DNA methylation changes) are likely to be the same causal variants for T1D. Given a mQTL - CpG pair, JLIM estimates the putative causal SNP for this CpG site within a 1-Mb window centred around that mQTL. It also estimates the putative causal SNP for T1D within the same region. Concordance between top SNPs for the two sets of traits would suggest that

bioRxiv preprint doi: https://doi.org/10.1101/248260; this version posted January 15, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

DNA methylation potentially reside in the causal pathway to T1D risk. The concordance rates were determined after accounting for chance, under 1000 permutations. However, JLIM cannot rule out the possibility that cis-mQTLs influence methylation and T1D via two independent mechanisms (horizontal pleiotropy, as shown in Figure 1 d); in addition, it does not specify which SNP is the putative causal variant in a particular region.

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. The authors are grateful for Prof. John A Todd and Mr. Jamie Inshaw who kindly provided T1D GWAS summary statistics, as well as Prof. Yaron Tomer for his valuable advice on methylation data interpretation.

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; and J.Y will serve as guarantor for the contents of this paper. J.Y. was funded by a Diabetes Wellness & Research 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.

bioRxiv preprint doi: https://doi.org/10.1101/248260; this version posted January 15, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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.

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;

matSpDlite

https://github.com/snewhouse/BRC_MH_Bioinformatics/blob/master/misc_sh/matSpDlite.R;

Immunobase www.immunobase.org;

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

bioRxiv preprint doi: https://doi.org/10.1101/248260; this version posted January 15, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

References

1. Mehers KL & Gillespie KM (2008) The genetic basis for type 1 diabetes. *Br Med Bull* 88(1):115-129.
2. Cooper NJ*, et al.* (2017) Type 1 diabetes genome-wide association analysis with imputation identifies five new risk regions. *bioRxiv*.
3. Onengut-Gumuscu S*, et al.* (2015) Fine mapping of type 1 diabetes susceptibility loci and evidence for colocalization of causal variants with lymphoid gene enhancers. *Nat Genet* 47(4):381-386.
4. Liu Y*, et al.* (2013) Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis. *Nat Biotechnol* 31(2):142-147.
5. Farh KK*, et al.* (2015) Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature* 518(7539):337-343.
6. Castillo-Fernandez JE, Spector TD, & Bell JT (2014) Epigenetics of discordant monozygotic twins: implications for disease. *Genome Med* 6(7):60.
7. Stefan M, Zhang W, Concepcion E, Yi Z, & Tomer Y (2014) DNA methylation profiles in type 1 diabetes twins point to strong epigenetic effects on etiology. *J Autoimmun* 50:33-37.
8. Paul DS*, et al.* (2016) Increased DNA methylation variability in type 1 diabetes across three immune effector cell types. *Nat Commun* 7:13555.
9. Elboudwarej E*, et al.* (2016) Hypomethylation within gene promoter regions and type 1 diabetes in discordant monozygotic twins. *J Autoimmun* 68:23-29.
10. Davey Smith G & Hemani G (2014) Mendelian randomization: genetic anchors for causal inference in epidemiological studies. *Hum Mol Genet* 23(R1):R89-98.
11. Relton CL & Davey Smith G (2012) Two-step epigenetic Mendelian randomization: a strategy for establishing the causal role of epigenetic processes in pathways to disease. *Int J Epidemiol* 41(1):161-176.
12. Richardson TG*, et al.* (2017) Causal epigenome-wide association study identifies CpG sites that influence cardiovascular disease risk. *bioRxiv*.
13. Gaunt TR*, et al.* (2016) Systematic identification of genetic influences on methylation across the human life course. *Genome Biol* 17:61.
14. Hartwig FP, Davies NM, Hemani G, & Davey Smith G (2016) Two-sample Mendelian randomization: avoiding the downsides of a powerful, widely applicable but potentially fallible technique. *Int J Epidemiol* 45(6):1717-1726.
15. Scott RA*, et al.* (2012) Large-scale association analyses identify new loci influencing glycemic traits and provide insight into the underlying biological pathways. *Nat Genet* 44(9):991-1005.
16. Hemani G, Tilling K, & Davey Smith G (2017) Orienting the causal relationship between imprecisely measured traits using GWAS summary data. *PLoS Genet* 13(11):e1007081.
17. Chun S*, et al.* (2017) Limited statistical evidence for shared genetic effects of eQTLs and autoimmune-disease-associated loci in three major immune-cell types. *Nat Genet* 49(4):600-605.
18. Fradin D*, et al.* (2012) Association of the CpG methylation pattern of the proximal insulin gene promoter with type 1 diabetes. *PLoS One* 7(5):e36278.
19. Hyttinen V, Kaprio J, Kinnunen L, Koskenvuo M, & Tuomilehto J (2003) Genetic liability of type 1 diabetes and the onset age among 22,650 young Finnish twin pairs: a nationwide follow-up study. *Diabetes* 52(4):1052-1055.
20. Li YR*, et al.* (2015) Genetic sharing and heritability of paediatric age of onset autoimmune diseases. *Nat Commun* 6:8442.
21. Cooper JD*, et al.* (2008) Meta-analysis of genome-wide association study data identifies additional type 1 diabetes risk loci. *Nat Genet* 40(12):1399-1401.

bioRxiv preprint doi: https://doi.org/10.1101/248260; this version posted January 15, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

22. Gough SC & Simmonds MJ (2007) The HLA Region and Autoimmune Disease: Associations and Mechanisms of Action. *Curr Genomics* 8(7):453-465.
23. Noble JA & Valdes AM (2011) Genetics of the HLA region in the prediction of type 1 diabetes. *Curr Diab Rep* 11(6):533-542.
24. Li D, Das S, Yamada T, & Samuels HH (2004) The NRIF3 family of transcriptional coregulators induces rapid and profound apoptosis in breast cancer cells. *Mol Cell Biol* 24(9):3838-3848.
25. Barton A, *et al.* (2009) Identification of AF4/FMR2 family, member 3 (AFF3) as a novel rheumatoid arthritis susceptibility locus and confirmation of two further pan-autoimmune susceptibility genes. *Hum Mol Genet* 18(13):2518-2522.
26. Hinks A, *et al.* (2010) Association of the AFF3 gene and IL2/IL21 gene region with juvenile idiopathic arthritis. *Genes Immun* 11(2):194-198.
27. Ma C & Staudt LM (1996) 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* 87(2):734-745.
28. Minnich M, *et al.* (2016) Multifunctional role of the transcription factor Blimp-1 in coordinating plasma cell differentiation. *Nat Immunol* 17(3):331-343.
29. Tan RJ, *et al.* (2010) Investigation of rheumatoid arthritis susceptibility genes identifies association of AFF3 and CD226 variants with response to anti-tumour necrosis factor treatment. *Ann Rheum Dis* 69(6):1029-1035.
30. Floyel T, *et al.* (2014) CTSH regulates beta-cell function and disease progression in newly diagnosed type 1 diabetes patients. *Proc Natl Acad Sci U S A* 111(28):10305-10310.
31. Gutierrez-Arcelus M, *et al.* (2015) Tissue-specific effects of genetic and epigenetic variation on gene regulation and splicing. *PLoS Genet* 11(1):e1004958.
32. Bibikova M, *et al.* (2011) High density DNA methylation array with single CpG site resolution. *Genomics* 98(4):288-295.
33. Li R, *et al.* (2014) Somatic point mutations occurring early in development: a monozygotic twin study. *J Med Genet* 51(1):28-34.
34. Boyd A, *et al.* (2013) Cohort Profile: The 'Children of the 90s'-the index offspring of the Avon Longitudinal Study of Parents and Children. *International Journal of Epidemiology* 42(1):111-127.
35. Fraser A, *et al.* (2013) Cohort Profile: The Avon Longitudinal Study of Parents and Children: ALSPAC mothers cohort. *International Journal of Epidemiology* 42(1):97-110.
36. Relton CL, *et al.* (2015) Data Resource Profile: Accessible Resource for Integrated Epigenomic Studies (ARIES). *Int J Epidemiol* 44(4):1181-1190.
37. Zhou W, Laird PW, & Shen H (2017) Comprehensive characterization, annotation and innovative use of Infinium DNA methylation BeadChip probes. *Nucleic Acids Res* 45(4):e22.
38. Houseman EA, *et al.* (2012) DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics* 13:86.
39. Stahl EA, *et al.* (2010) Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci. *Nat Genet* 42(6):508-514.
40. Gardner SG, Bingley PJ, Sawtell PA, Weeks S, & Gale EA (1997) Rising incidence of insulin dependent diabetes in children aged under 5 years in the Oxford region: time trend analysis. The Bart's-Oxford Study Group. *BMJ* 315(7110):713-717.
41. Howie BN, Donnelly P, & Marchini J (2009) A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet* 5(6):e1000529.
42. Howie B, Marchini J, & Stephens M (2011) Genotype imputation with thousands of genomes. *G3 (Bethesda)* 1(6):457-470.
43. Bradfield JP, *et al.* (2011) A genome-wide meta-analysis of six type 1 diabetes cohorts identifies multiple associated loci. *PLoS Genet* 7(9):e1002293.

bioRxiv preprint doi: https://doi.org/10.1101/248260; this version posted January 15, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

44. Bulik-Sullivan B, *et al.* (2015) An atlas of genetic correlations across human diseases and traits. *Nat Genet* 47(11):1236-1241.
45. Burgess S, *et al.* (2015) Using published data in Mendelian randomization: a blueprint for efficient identification of causal risk factors. *Eur J Epidemiol* 30(7):543-552.
46. Haycock PC, *et al.* (2016) Best (but oft-forgotten) practices: the design, analysis, and interpretation of Mendelian randomization studies. *Am J Clin Nutr* 103(4):965-978.
47. Nyholt DR (2004) A simple correction for multiple testing for single-nucleotide polymorphisms in linkage disequilibrium with each other. *Am J Hum Genet* 74(4):765-769.

SNP	Effect allele	Candidate gene	CpG ID	CpG position	CpG effect (beta±SE)	T1D effect (beta±SE)	2SMR effect	2SMR p value
rs2269242	A	*ITGB3BP,PGM1*	cg05762488	Chr1: 64140332	0.059 (0.006)	0.168 (0.030)	2.843 (0.508)	2.14E-08
rs9653442	C	*AFF3*	cg06183267	Chr2: 100759134	-0.100 (0.006)	0.124 (0.025)	-1.236 (0.250)	7.05E-07
rs9653442	C	*AFF3*	cg07349094	Chr2: 100759014	-0.157 (0.009)	0.124 (0.025)	-0.791 (0.160)	7.05E-07
rs3087243	A	*CTLA4*	cg22572158	Chr2: 204731068	-0.070 (0.008)	-0.178 (0.025)	2.534 (0.356)	1.08E-12
rs3825932	T	*CTSH*	cg25744700	Chr15: 79237217	-0.204 (0.012)	-0.141 (0.030)	0.692 (0.147)	2.60E-06
rs3825932	T	*CTSH*	cg18738367	Chr15: 79238723	0.093 (0.006)	-0.141 (0.030)	-1.523 (0.324)	2.60E-06
rs1893217	G	*PTPN2*	cg09945482	Chr18: 12777974	-0.188 (0.016)	0.250 (0.032)	-1.331 (0.170)	5.61E-15
rs1893217	G	*PTPN2*	cg23544223	Chr18: 12777786	-0.175 (0.011)	0.250 (0.032)	-1.428 (0.183)	5.61E-15
rs1893217	G	*PTPN2*	cg23598886	Chr18: 12777645	-0.414 (0.023)	0.250 (0.032)	-0.604 (0.077)	5.61E-15
rs1893217	G	*PTPN2*	cg24737193	Chr18: 12778029	-0.223 (0.159)	0.250 (0.032)	-1.123 (0.144)	5.61E-15

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 log odds ratio); 2SMR effect denotes the change of log odds on T1D per unit increase in DNA methylation due to its associated SNP.

bioRxiv preprint doi: https://doi.org/10.1101/248260; this version posted January 15, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

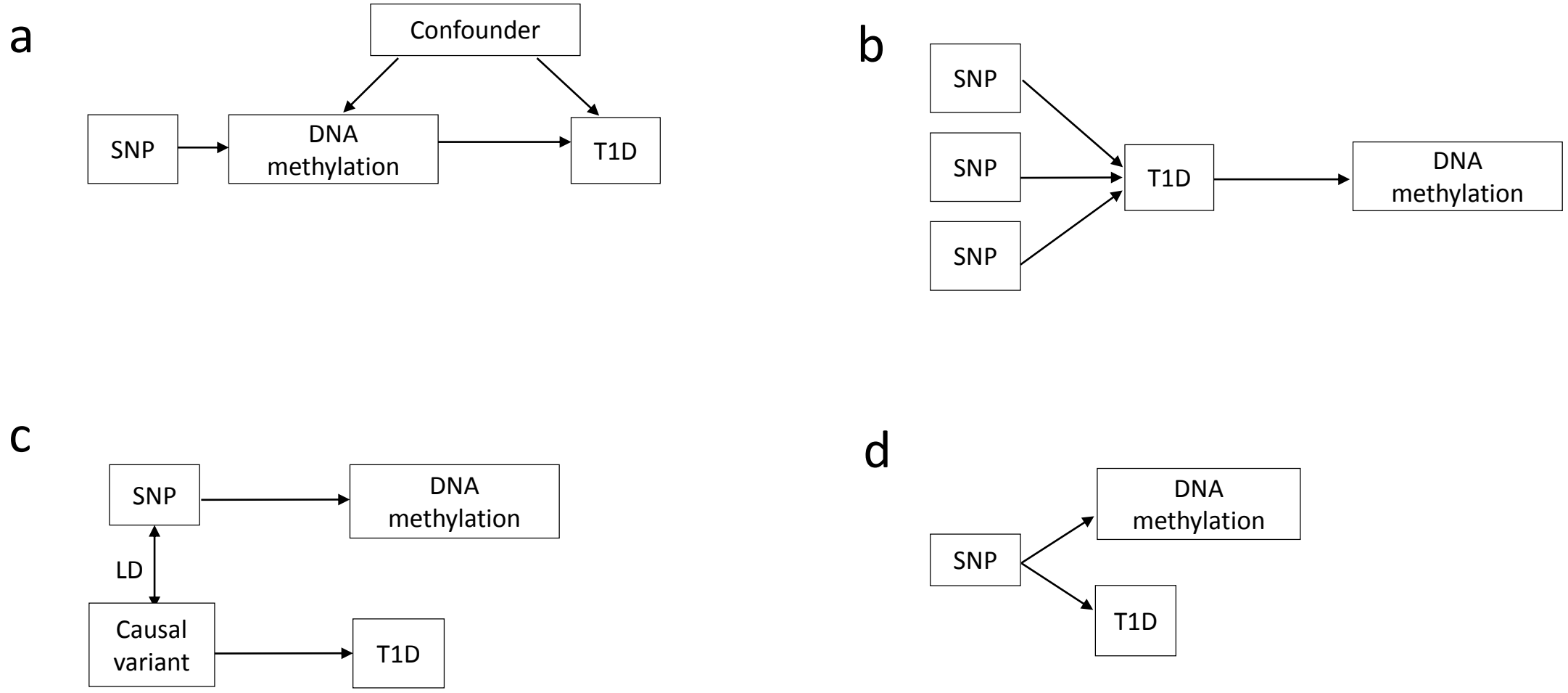

Figure 1: Four possible scenarios that could explain the associations between SNP, DNA methylation and T1D.

a, DNA methylation mediates the genetic risk of T1D; b, SNPs first increases T1D liability, which in turn changes DNA methylation levels; c, a SNP that regulates DNA methylation could simply be in LD with another causal variant that influences T1D; d, a SNP is associated with DNA methylation and T1D via independent biological pathways (horizontal pleiotropy).

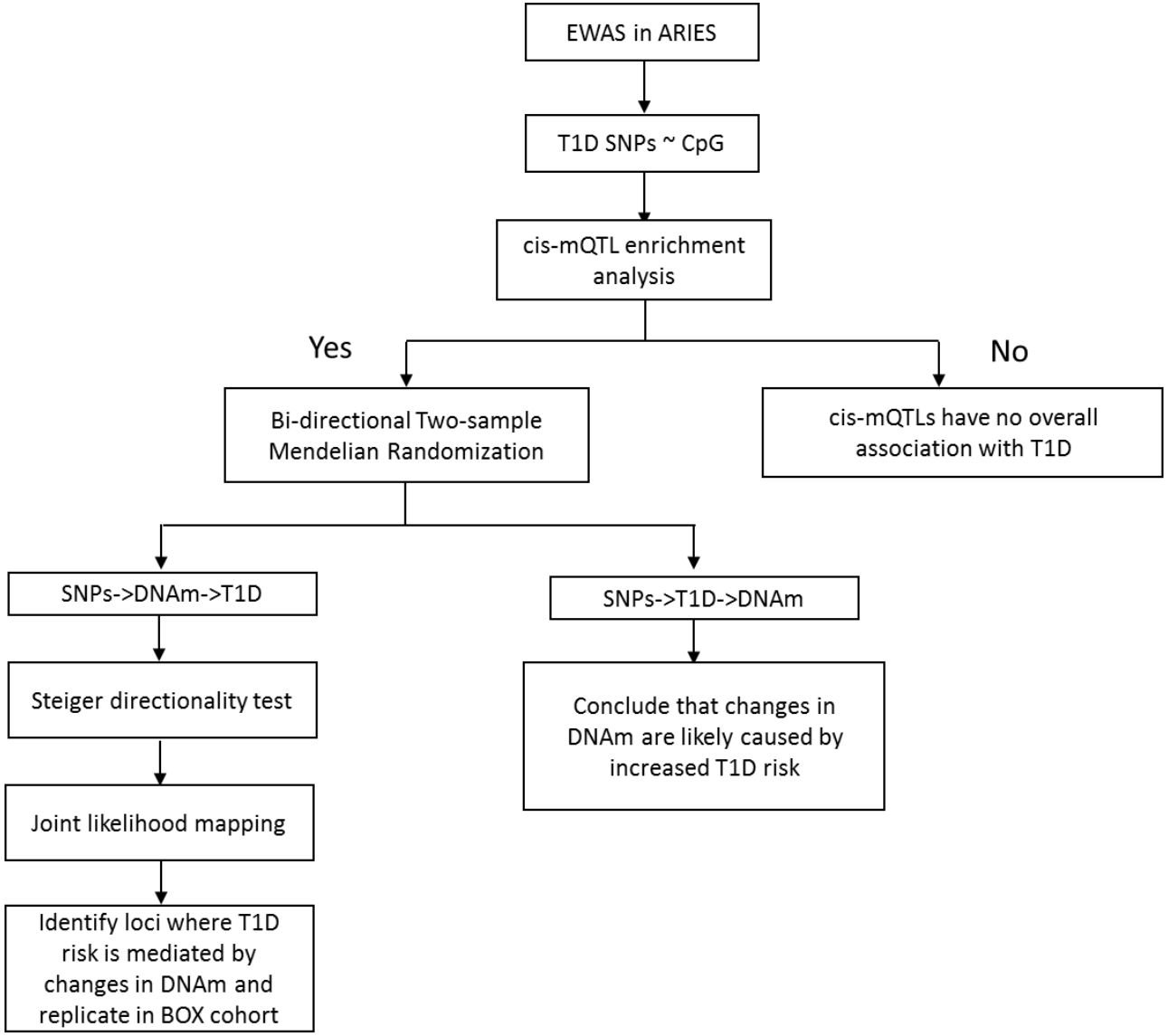

Figure 2: Flow chart summarising the overall analysis procedure in this study.

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

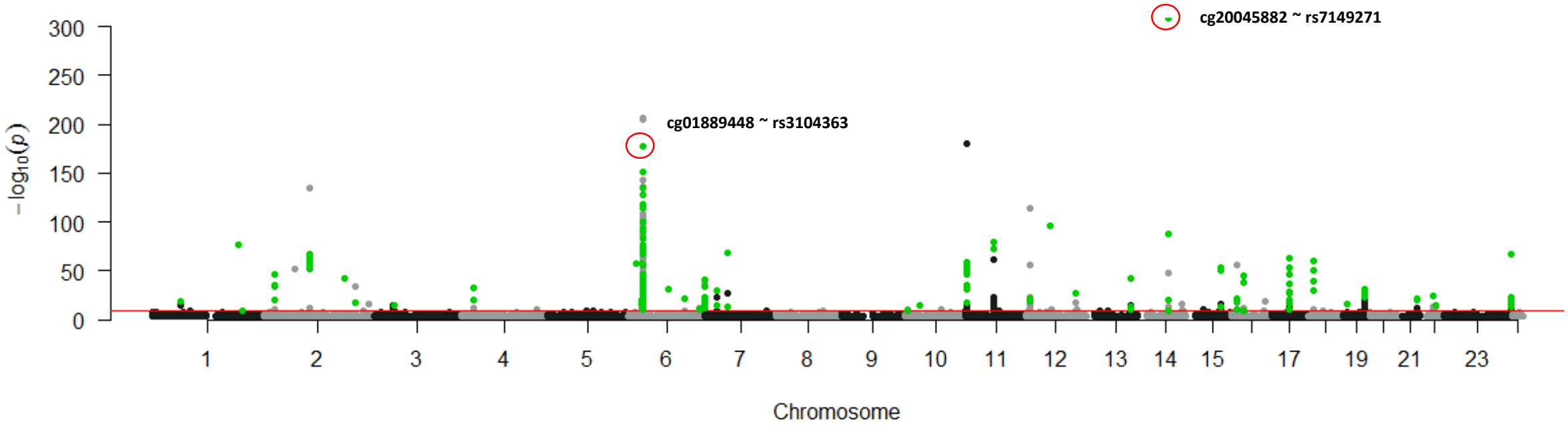

Figure 3: Genomic distribution of CpG sites that are associated with T1D GWAS variants.

Manhattan plot showing the CpG sites associated with 38 T1D GWAS variants above the Bonferroni threshold 1.6e-9 (redline); green highlighted dots were those (n=166) that were consistently detected at adolescence, childhood and birth; there is a peak reflecting mQTL-CpG association at the HLA locus (chromosome 6).

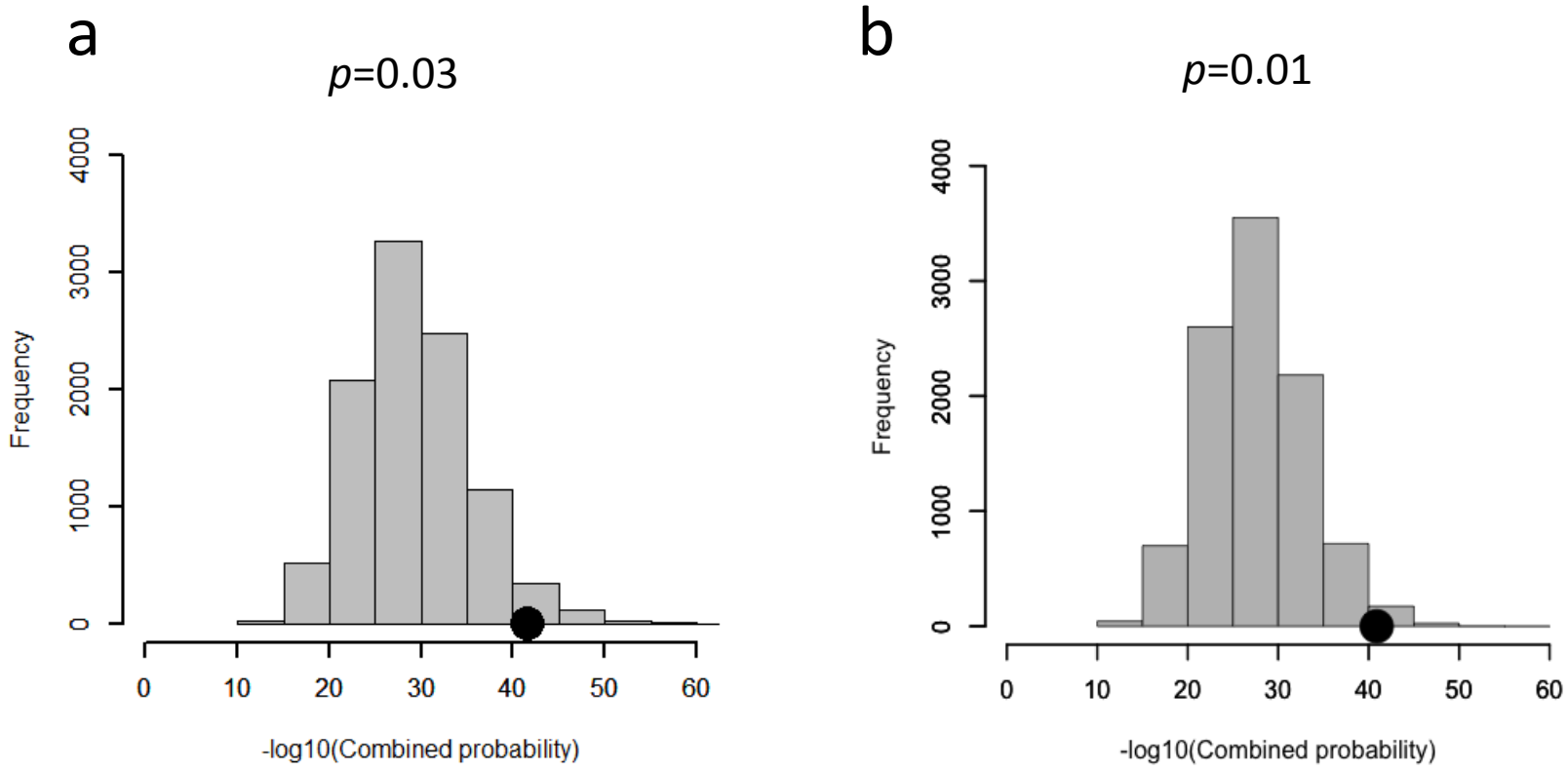

Figure 4: cis-mQTLs are enriched in SNPs with low GWAS *p*-values associated with T1D.

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.

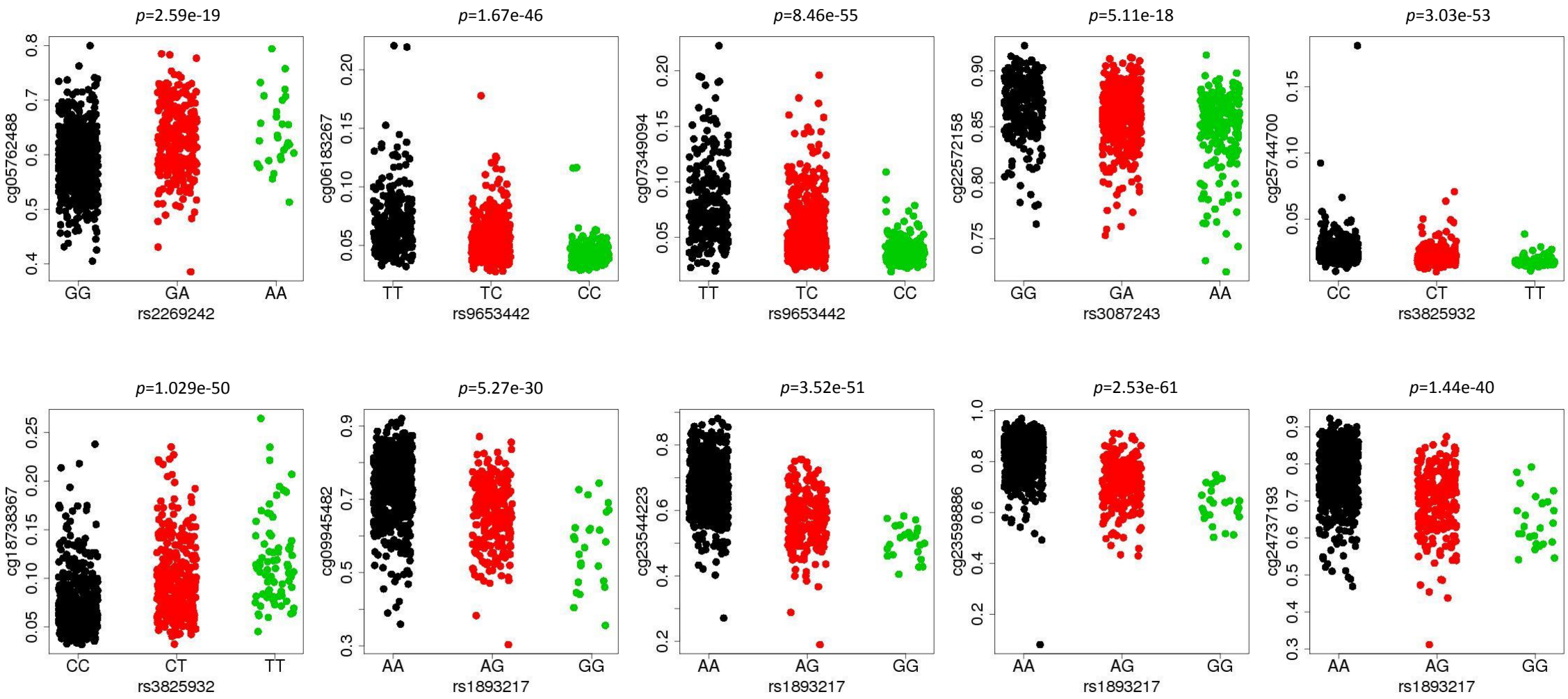

Figure 5: DNA methylation levels of CpG sites and their associations with T1D GWAS variants that survived JLIM analyses, obtained from the ARIES adolescent participants.

Y - axis represents beta values for each CpG site. The inner most genotype in X - axis is comprised of two other alleles, the outer most genotype is comprised of two effect alleles.

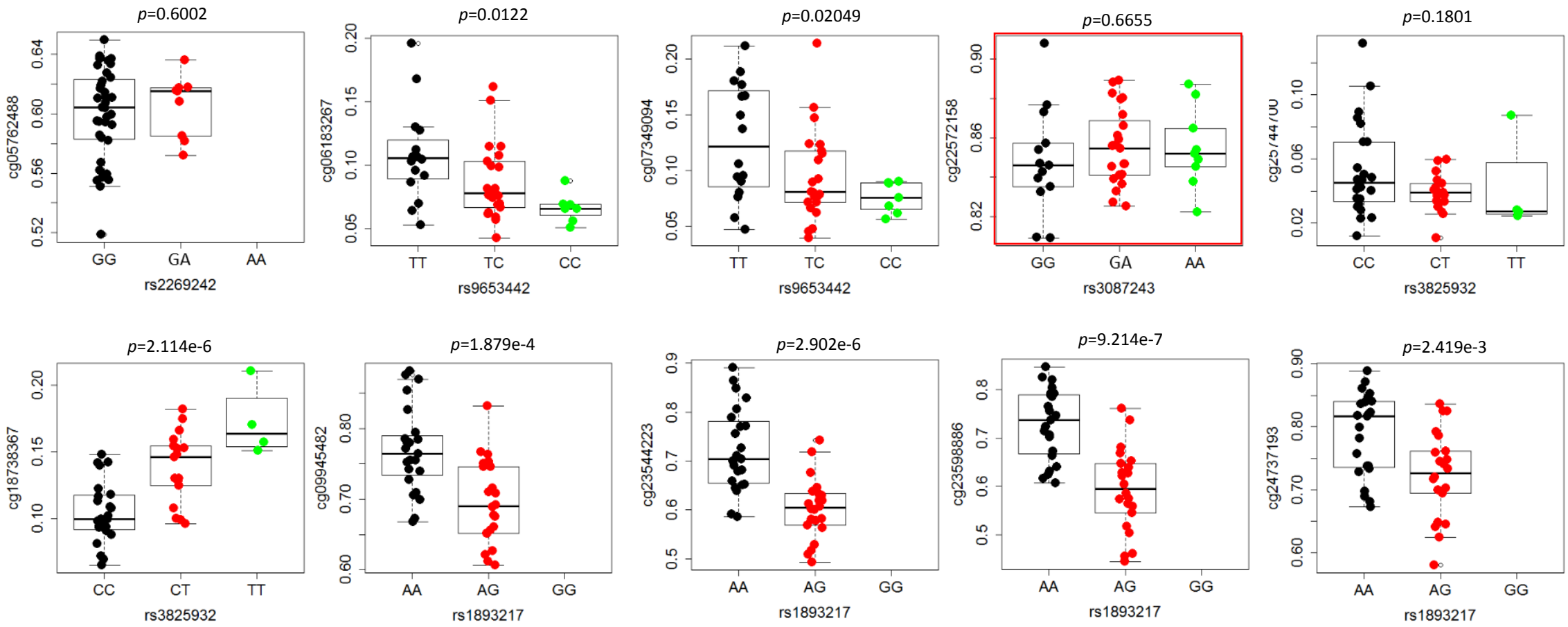

Figure 6: Replication of SNP and DNA methylation associations in the Bart's oxford T1D cohort.

45 individuals were analysed in the BOX cohort. Nine out of ten SNP-CpG pairs showed similar associations compared to the ARIES participants. The SNP – CpG pair that did not replicate the ARIES result was highlighted in red.