DNA Methylome profiling in identical twin pairs discordant for body mass index

Weilong Li¹, Dongfeng Zhang², Weijing Wang², Yili Wu², Jan Baumbach³,⁴, Lene Christiansen¹,⁵, Qihua Tan¹,⁶, *

1. Epidemiology and Biostatistics, Department of Public Health, University of Southern Denmark, Odense, Denmark
2. Division of Epidemiology and Health Statistics, Qingdao University Medical College, Qingdao, China
3. Chair of Experimental Bioinformatics, TUM School of Life Sciences, Technical University of Munich, Munich, Germany
4. Department of Mathematics and Computer Science, University of Southern Denmark, Odense, Denmark
5. Department of Clinical Immunology, Copenhagen University Hospital, Rigshospitalet, Copenhagen, Denmark
6. Unit of Human Genetics, Department of Clinical Research, University of Southern Denmark, Odense, Denmark

E-mails:
WL: wli@health.sdu.dk
DZ: zhangdf1961@126.com
WW: wangwj793@126.com
YW: ywu@health.sdu.dk
JB: jan.baumbach@wzw.tum.de
LC: lchristiansen@health.sdu.dk
QT: qtan@health.sdu.dk

*Addresses for correspondence:
Professor Qihua Tan, MD, PhD
Epidemiology and Biostatistics, Department of Public Health
University of Southern Denmark
J.B. Winsløws Vej 9 B, DK-5000, Odense C, Denmark
Tel. 0045 65503536 e-mail: qtan@health.sdu.dk
ABSTRACT

Body mass index (BMI) serves as an important measurement of obesity and adiposity which are highly correlated with cardiometabolic diseases. Although high heritability has been estimated, the identified genetic variants by genetic association studies only explain a small proportion of BMI variation. As an active effort for further exploring the molecular basis of BMI variation, large scale epigenome-wide association studies have been conducted but with limited number of loci reported, perhaps due to poorly controlled confounding factors including genetic factors. Being genetically identical, monozygotic twins discordant for BMI are ideal subjects for analysing the epigenetic association between DNA methylation and BMI, providing perfect control on their genetic makeups largely responsible for BMI variation. We performed an epigenome-wide association study on BMI using 30 identical twin pairs (15 male and 15 female pairs) with age ranging from 39 to 72 years and degree of BMI discordance ranging from 3-7.5 kg/m². Methylation data from whole blood samples were collected using the reduced representation bisulfite sequencing technique. After adjusting for blood cell composition and clinical variables, we identified 136 CpGs with p-value < 1e-4, 30 CpGs with p<1e-05 but no CpGs reached genome-wide significance. Genomic region based analysis found 11 differentially methylated regions harbouring coding and non-coding genes some of which were validated by gene expression analysis on independent samples. Our DNA methylation sequencing analysis on identical twins provides new references for the epigenetic regulation on BMI and obesity.

Key words: BMI, epigenetics, identical twins, bisulfite sequencing
INTRODUCTION

Diabetes and cardiovascular diseases are common conditions in developed and developing countries, which have big impact on life quality and mortality [1–3]. These conditions are often related to obesity for which body mass index (BMI) is an important measurement. BMI calculated as weight in kilograms divided by the square of height in meters \((kg/m^2)\) is a simple and inexpensive surrogate measurement for body mass and obesity [4]. BMI has been widely studied using traditional epidemiology revealing its association with environmental factors including occupation [5], incomes [6], and many other factors. Importantly, twin studies have estimated high heritability of up to 90% indicating large contribution by genetic factors to the variation of BMI [7].

In the past decade, the increasing popularity of genome-wide association studies (GWAS) has led to the discovery of both common genetic polymorphisms and rare genetic variants associated with complex diseases and health traits [8] including BMI [9]. Unfortunately, the reported static genetic polymorphisms explain only a limited proportion of BMI variation in comparison with the estimated high heritability [7]. Efforts have been taken to look for alternative mechanisms underlying BMI variation such as the epigenetics [10] defined as the study of heritable changes in gene expression that do not constitute changes in the DNA sequence. Large scale epigenome-wide association studies (EWAS) have been conducted using DNA methylation profiling but only a handful of significant genomic sites have been reported [11–13]. The limited discovery could perhaps be partly due to poorly controlled confounding factors, both genetic and environmental. Considering the high heritability estimate of BMI, the uncontrolled sequential genetic variations could serve as confounding factors that reduce the
Because identical twins or monozygotic (MZ) twins share their genetic makeups, MZ twin pairs discordant for a disease or trait serve as ideal samples for EWAS on highly heritable diseases or phenotypes advantaged by a perfect control over the genetic background [14]. Our recent computer simulation study has shown the highly enriched statistical power of the discordant twin design in EWASs especially on diseases with strong genetic basis [15]. We have recently conducted an EWAS on 30 pairs of identical twin pairs discordant for BMI using the reduced representation bisulfite sequencing (RRBS) technique for analyzing the DNA methylome with a broader coverage compared with the DNA methylation chips [16]. This paper reports our findings and discuss potential implications in the etiology of obesity.

MATERIALS AND METHODS

Samples
The discordant twin pairs were selected from the databases of Qingdao Twin Registry [17] and the large Qingdao Diabetes Prevention Program [18]. Based on our recent statistical power estimates for the discordant twin design in EWAS [15], thirty monozygotic twin pairs with 15 male pairs and 15 female pairs were included in this study. All samples included were free from medication and from diagnosed diseases. The age ranges from 39 to 72 and BMI ranges from 18.9 to 37.7 across the samples. Absolute intra-pair BMI discordance (ΔBMI) ranges from 3-7.5 kg/m² and percentage discordance calculated as |ΔBMI|/max(BMI) from 8-22%. For each participant, whole blood was taken for DNA methylation analysis. Written consents were obtained from all the participants. The study was approved by Regional Ethics Committee at
Qingdao CDC Institutional Review Boards (QDCDC-IRB) and conducted according to the principles of the Helsinki Declaration.

**DNA methylation analysis**

DNA was extracted from whole blood using CTAB (cetyltrimethyl ammonium bromide) following standard procedure. DNA methylation library was constructed by Biomarker Technologies Corporation, Beijing, China (http://www.biomarker.com.cn/) using RRBS. Genomic DNA was first digested using the MspI enzyme, which restrictively cut the DNA at sites CCGG, and then the fragments were end-repaired and dA-tailing to blunt end products, followed by adaptor-ligation with T overhang. The ligation products were purified by 2% agarose gel electrophoresis and size-selected of DNA fragments 150–400 bp long (including 100 bp adaptor). Size-selected DNA was bisulfite-converted with the NEXTflex Bisulfite-Seq Kit (Bioo Scientific, Austin, TX, USA). The final library was generated by PCR-amplification, enriching for fragments with adapters on both ends. The RRBS was performed by Illumina HiSeq X Ten (Illumina Inc., San Diego, CA, USA).

**Data preprocessing**

Raw data were preprocessed using the pipeline recommended by Bismark [19]. Sequencing data were first trimmed using Trim Galore [20] and then aligned to Genome Reference Consortium Human Build 37 (hg19) using Bowtie2 [21]. The coverage outputs from Bismark were imported to R package BiSeq [22] to smooth the methylation level, and were used for further analysis. As a result of quality control, one twin pair were dropped due to unexpected extremely high coverage. To reduce bias, we limited the coverage to the 90% quantile which led to 1368869 remaining sites. Next, we further removed CpGs that had an average methylation beta value <
0.01 or more than 4 missing observations. This resulted in 952848 CpG sites for subsequent analysis. The methylation beta values were transformed into M-value using logit (base 2) transformation for statistical modeling.

**Cell-type composition**

Since the DNA methylation data were measured on whole blood consisting of different cell types, different methylation profiles of distinct cell types may lead to false discoveries. In order to control for cell-type composition effect on DNA methylation, we introduced ReFACTor, a reference-free adjustment for cell-type composition based on principal component analysis (PCA) [23]. The algorithm calculates components that are correlated with the cell-type composition of the samples by applying an unsupervised feature selection step followed by PCA. Instead of estimating absolute cell count values, ReFACTor calculates the linear transformations of the cell-type composition as PCA components. We selected the top 5 components to add as covariates in subsequent statistical analysis to account for cell-type heterogeneity.

**Statistical analysis**

For single CpG analysis we applied linear mixed effect model regressing methylation level on BMI adjusting for cell-type composition and clinical variables, with twin pairing modeled as a random effect variable. The mixed effect model here is equivalent to the regression model for discordant twin design as proposed by Tan et al. [24]. To take multiple testing into account, we calculated the false discovery rate (FDR) [25] and defined genome-wide significance as FDR<0.05.

**Functional annotations**

1. Over-representation analysis
The identified genomic sites reaching nominal significance of p<0.05 were first linked to their nearest genes. The resulting list of genes was submitted to the Gene-set Enrichment Analysis (GSEA) website at http://software.broadinstitute.org/gsea/index.jsp for over-representation analysis of the input gene list in the Molecular Signatures Database (MSigDB). The analysis was performed using a hypergeometric test to calculate the probability of having the number of overlapping genes between the input list and a functional cluster (pathway) of genes when randomly taking the same number of genes as in the input list from all genomic sites analyzed. Multiple testing is corrected by estimating the FDR [25].

2. Genomic regions enrichment analysis

For the identified methylation sites with nominal significance of p<0.05, we examined their enrichment in the regulatory domain of different functional pathways assuming cis-regulation using the Genomic Regions Enrichment of Annotations Tool (GREAT) [26]. Instead of the nearest-gene approach based on proximal binding events, GREAT applies a binomial test on genomic regions using the total fraction of the genome associated with a given ontology term as the expected fraction of input regions associated with the term by chance. A binomial fold enrichment is calculated as the ratio between the actual number of genomic regions in the test set and the expected number of genomic regions in the test set. Annotation of GREAT is based on Genome Reference Consortium Human Build 37 (hg19).

Detecting differentially methylated regions (DMRs)

Based on the bisulfite sequencing data and corresponding EWAS result, we detected significant DMRs for BMI using the comb-p approach proposed by Petersen et al. [27]. The method combines p values of adjacent CpGs using the Stouffer–Liptak method to find regions of
significant enrichment after performing false discovery adjustment. The method is implemented in a python library \textit{comb-p} (https://github.com/brentp/combined-pvalues).

\textbf{RESULTS}

In Table 1, we show the description for anthropometric and biochemical variables for samples. The last column of Table 1 is the intra-pair correlation of each variable. Most variables had considerably high and/or significant correlation, a situation that our discordant twin design can benefit. Some showed low and insignificant intra-pair correlation (e.g. GLU, TG) which we included in our association analysis for adjustment. Figure 1 displays the BMI measurements for each twin pair. All twin pairs distance themselves from the diagonal line indicating intra-pair discordance in BMI measurements. Meanwhile, the overall pattern in Figure 1 also indicate a weak intra-pair correlation on BMI as reported in Table 1 (r=0.307, p=0.099). No sex-specific pattern for BMI discordance was observed in the male (solid dots) and the female (empty dots) twin pairs.

\textit{Single-site based association analysis}

We first performed a single site based EWAS as described in the Methods section. A total of 136 CpGs were identified with p<1e-04, among them 30 CpGs with p<1e-05, 2 CpGs with p<1e-06. No site reached genome-wide significance defined as FDR<0.05 with the lowest FDR=0.195 (supplementary Table S1). In Figure 2, we show the Manhattan plot for the p-values of each CpG site against its chromosomal location (Figure 2a) and a Q-Q plot for the estimated p-values plotted against the randomly expected p value distribution (Figure 2b). The p values of the top
significant CpGs show deviation from the random distribution and its shaded area of 95% confidence intervals (Figure 2b) suggesting non-random association of these CpGs with BMI.

_Biological pathway analysis_

Using biomaRt [28,29], a total of 23863 CpGs out of 38487 CpGs with p-value < 0.05 (supplementary Table S1) were successfully annotated to 2542 functional genes and submitted to GSEA [30] for over-representation analysis of canonical pathways in the BioCarta gene sets, KEGG gene sets and Reactome gene sets. Table 2 shows the 10 pathways with high statistical significance even after correcting for multiple testing (FDR<4.32e-12). Among them, there are pathways related to ensemble of genes encoding extracellular matrix (ECM) and extracellular matrix-associated proteins, including ECM glycoproteins, collagens and proteoglycans, genes involved in neuronal system, genes involved in developmental biology. There are also pathways of neuron functions and signaling include genes involved in transmission across chemical synapses, in axon guidance, and in signaling by NGF (nerve growth factor).

The 38487 CpGs with p-value < 0.05 were also used as input to GREAT for analyzing their enrichment in the regulatory domain. The analysis found 15 functional clusters of biological process with very high statistical significance (binomial p value<1.072e-10) (Table 3) involving positive regulations of gliogenesis (Fold enrichment or FE:2.49), negative regulations of gliogenesis (Fold enrichment or FE:2.03), negative regulation of glial cell differentiation (FE:2.19), negative regulation of lipid biosynthesis (FE:2.61), thyroid gland development (FE:2.83), retinal rod cell development (FE: 3.54), etc.

_Genomic-region based analysis_
By applying comb-p, we identified 11 DMRs with FDR<0.1, among them 9 DMRs were found with FDR<0.05 on chromosomes 1, 2, 3, 5, 6, 9, 10 and 18 (Table 4). Interestingly, 4 significant DMRs (from 1103393 to 1103564 bp (basepair) and from 990300 to 990641 bp on chromosome 1; from 1594287 to 1594438 bp on chromosome 6; from 102026134 to 102026689 bp on chromosome 10) cover the top significant CpGs in supplementary Table S1. Four DMRs in Table 4 are located in the intergenic regions and the other 7 DMRs are annotated to LINC01503 on chromosome 9 (an RNA gene), DPYSL3 on chromosome 5, CHD5 and AGRN on chromosome 1, C1QL2 on chromosome 2, CWF19L1 on chromosome 10, and AKAP12 on chromosome 6. Figure 3 displays the methylation patterns for the 11 DMRs in Table 4. Except the DMRs in Figures 3f and 3j which are hypomethylated with increasing BMI, methylation levels at the other 9 DMRs are all positively correlated with BMI.

**DISCUSSION**

Based on bisulfite sequencing analysis of the DNA methylome, we have performed an EWAS on BMI using the discordant twin design. As shown in Figure 2, our study did not identify DNA methylation sites of genome-wide significance predefined as FDR<0.05. However, the tail of the observed p values deviates clearly from random, indicating epigenetic association with BMI. Among the top CpGs in supplementary Table S1, 18 CpGs are from the significant DMR on chromosome 1 from 1103393-1103564 bp (Figure 3h, Table 4), with p value <5.47e-05. Most importantly, although no protein coding gene is linked to the DMR, it is in the vicinity of miR429 and miR200a both have been reported to regulate adipogenesis [31,32]. Crepin et al. observed differential expression of 11 miRNAs in the hypothalamus (a crucial central nervous system area controlling appetite, body weight and metabolism) from leptin-deficient ob/ob mice.
among which three miRNAs, miR-200a, miR-200b and miR-429, were confirmed in independent samplers [33], suggesting their implications in the hypothalamic regulation of energy homeostasis.

Although our single CpG based EWAS did not identify sites of genome-wide significance, significant functional pathways have been found enriched by their annotated genes or regulatory profiles. Our results show that DNA methylation changes due to BMI discordance could be linked to development (Table 2, REACTOME: genes involved in developmental biology) involving perhaps early-life events such as low birthweight. Among the 10 significant pathways in Table 2, 3 are related to the extracellular matrix (ECM) which excess deposition in the adipose tissue deteriorates insulin sensitivity [34]. In both Tables 2 and 3, several neural pathways are enriched by the DNA methylation patterns associated with BMI, including axon guidance (Table 2), glial cell differentiation, gliogenesis (Table 3). The neural system directly controls the production of many hormones, which are involved in regulating metabolism and behavior [35]. The importance of glial cells in gastrointestinal functions is notably highlighted by a multitude of digestive and even extradigestive disorders (such as obesity) that are associated with altered enteric glia [36]. In another development, an animal study showed that a partial loss of neural stem cells in the hypothalamus of the forebrain caused weight gain and glucose intolerance [37] which are in line with the significant functional cluster of forebrain regionalization in Table 3. Neuron cell fate commitment or differentiation is frequently observed in Table 3, perhaps due to the fact that modulation of energy metabolism is a key aspect associated with cell fate transition [38]. In fact, neurons in the hindbrain, forebrain and midbrain are critical for normal energy homeostasis involving obesity pathogenesis through controlling food intake and energy balance [39, 40]. Retinal degeneration and several other degenerative
diseases have been associated with metabolic disorders [41]. Finding of the biological process in retinal rod cell development in Table 3 could suggest that relevant epigenetic modification could have already started in BMI discordant twins.

The power of genomic region based analysis is clearly shown by the results in Table 4 where 9 DMRs have been found with FDR<0.05 and 2 with 0.05<FDR<0.1. Among the genes linked to the DMRs, expression of DPYSL3 has been associated with nonalcoholic fatty liver disease [42]. The AGRN gene displayed strong evidence for a causal intrauterine effect of maternal BMI on newborn blood DNA methylation [43]. Our result directly links the gene with BMI discordance in identical twin pairs. The gene CWF19L1 has been shown to influence liver fat deposition together with ERLIN1 and CHUK [44]. Recently, the same gene was discovered as potential epigenetic biomarkers of obesity-related insulin resistance in human whole blood [45]. LINC01503 is over expressed in squamous cell carcinoma [46]. Its role in regulating BMI and obesity requires further investigation.

Using gene expression data on Danish twins, we analyzed the 7 genes linked to the 11 in DMRs in Table 4 by testing the correlation between their expression levels and BMI. Two genes, LINC01503 and AKAP12 have been validated with very high significance. In the gene expression data, 11 probes were matched to LINC01503 among which 8 had p value<0.05 with the lowest p value of 9.23e-06. There were 3 probes in AKAP12, 1 probe with p value 0.067 and the other with p value 1.89e-04.

Although this study used the powerful discordant twin design, the single site based EWAS discovered no CpG of genome-wide significance. This can be explained partly by the small sample size but also by the high heritability estimate for BMI [7] indicating that individual
variation in BMI can be mainly explained by DNA sequence variations. The limited impact of nongenetic factors requires relatively large sample sizes to obtain sufficient statistical power. Different from single site analysis, biological pathway analyses using GSEA based on functional gene annotation and GREAT based on the regulatory domain were able to reveal significant functional clusters and pathways potentially implicated in metabolic disorders. The fact that both GSEA and GREAT revealed high importance of developmental biology (cell fate commitment, cell and gland development, forebrain regionalization, hindbrain maturation, layer formation in cerebral cortex) could emphasize the impacts of early-life factors (such as low birthweight) on BMI at adult ages. Meanwhile, the results also remind us that studying the epigenetics of BMI using discordant twins design might put high weight to the epigenetic influence of early-life events on BMI at adult ages. Generalization of the findings to the non-twin population should be made with caution.
Conflicts of Interest

None declared

Acknowledgements

This project was supported by the EFSD/CDS/Lilly Collaborative Diabetes Research Programme (2013), the Lundbeck Foundation [grant number R170-2014-1353]; the DFF research project 1 from the Danish Council for Independent Research, Medical Sciences (DFF-FSS): DFF – 6110-00114; the Novo Nordisk Foundation Medical and Natural Sciences Research Grant [grant number NNF13OC0007493, and by the National Natural Science Foundation of China grant # 81773506.
REFERENCES


10. Allis CD, Jenuwein T: The molecular hallmarks of epigenetic control. Nat Rev Genet 2016; DOI: 10.1038/nrg.2016.59


2013;1:46–51.


Figure captions

**Figure 1.** Scatter plot displaying intra pair BMI discordance by plotting BMI measurements for twin 1 on x-axis, for twin 2 on y-axis. All twin pairs distance themselves from the diagonal line of equal BMI.

**Figure 2.** Circular Manhattan (a) and Q-Q (b) plots for single CpG based EWAS. No CpGs was found as genome-wide significant. The tail of observed p values shows a clear deviation from the random p value distribution and its shaded area of 95% confidence interval.

**Figure 3.** Differential methylation patterns from the 11 DMRs. Except the 2 hypomethylated DMRs (h, j), 9 are hypermethylated with increasing BMI.