LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT
TECHNISCHE UNIVERSITÄT MÜNCHEN TUM

Helmholtz Zentrum München
Institute of Computational Biology

Bachelor's Thesis
in Bioinformatics

Epigenetic signatures of environmental factors in children with familial risk for type 1 diabetes

Maria Wörheide

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT
TECHNISCHE UNIVERSITÄT MÜNCHEN TUM

Helmholtz Zentrum München
Institute of Computational Biology

Bachelor's Thesis
in Bioinformatics

Epigenetic signatures of environmental factors in children with familial risk for type 1 diabetes

~

Epigenetische Muster von Umweltfaktoren in Kindern mit einem familiären Risiko für Typ 1 Diabetes

Maria Wörheide

Aufgabensteller:	Prof. Dr. Dr. Fabian Theis
Betreuer:	Dr. Alida Kindt, Dr. Jan Krumsiek
Abgabedatum:	15.03.2016

I confirm that this bachelor's thesis is my own work and I have documented all sources and material used.

15.03.2016 _____

Maria Wörheide

Acknowledgments

I would like to thank the Institute of Computational Biology at the Helmholtz Zentrum München and Prof. Dr. Anette Ziegler for making this thesis possible. Especially Dr. Alida Kindt for her constant assistance, supervision and input.

Abstract

Type 1 Diabetes (T1D) is an autoimmune disease caused by the destruction of insulin-producing β-cells in the islets of Langerhans by autoantibodies. Its prevalence in the population is rapidly increasing and the precise triggers and underlying mechanisms of disease onset are still not fully understood. The methylation of cytosines in cytosine – guanine dinucleotides is an important epigenetic mechanism to control gene expression. The aim of this thesis was to establish a link between DNA methylation patterns in umbilical cord blood of children with familial risk of type 1 diabetes, environmental factors and information on future events (seroconversion, T1D onset). Using multivariate regression analysis, epigenome wide association studies were conducted with children from the BABYDIET cohort. Findings support the assumption of heritable disease-associated methylation patterns and give evidence for environmentally induced epigenetic programming *in utero*. Further examination of the found results, enabled novel insight into epigenetic mechanisms and disease susceptibility.

Zusammenfassung

Typ 1 Diabetes ist eine Autoimmunkrankheit bei der die Insulin produzierenden β-Zellen der Bauchspeicheldrüse durch das körpereigene Immunsystem zerstört werden. Das Auftreten dieser Krankheit in der Bevölkerung häuft sich zunehmend und die genauen Auslöser und zugrundeliegenden Mechanismen sind noch weitestgehend unbekannt. Die Methylierung von Cytosin-Guanosin-Dinukleotiden is ein wichtiger epigenetischer Mechanismus der Genregulierung. Das Ziel dieser Thesis war es, einen Zusammenhang zwischen DNA-Methylierungsmustern in Nabelschnurblut von Kindern mit familiärem T1D Risiko, Umweltfaktoren und Informationen über den zukünftigen Gesundheitsverlauf der Kinder (Seroconversion, T1D Diagnose) zu finden. Durch multivariate lineare Regression wurden Epigenomweite Assoziationsstudien mit Kindern der BABYDIET Kohorte durchgeführt. Die Ergebnisse stützen die Annahme, dass es vererbbare, krankheits-assoziierte Methylierungsmuster gibt. Des weitern wurden Hinweise auf epigenetische Programmierung im uterus gefunden die durch Umweltfaktoren beeinflusst werden. Anschließende Überprüfung der gefundenen Ergebnisse ermöglichten neue Einblicke in epigenetische Mechanismen und Krankheitsempfänglichkeit.

Contents

1 Introduction

In the following section the biological background of type 1 diabetes will be explained and a short introduction to epigenetics with focus on DNA methylation given as well as an outline of the thesis.

1.1 Type 1 Diabetes

Type 1 diabetes (T1D) is a chronic autoimmune disease caused by the destruction of insulin-producing β-cells in the islets of Langerhans by autoantibodies [1]. It is presumed to develop as the result of genetic predisposition, environmental factors and stochastic events [1]. T1D only accounts for 5-10 % of all cases of diabetes but once diagnosed, patients require strict glucose level monitoring and lifelong insulin treatment [1, 2]. Although often associated with children and adolescents, T1D can occur at any age [3]. The predominant form of diabetes, accounting for 90 - 95 % of patients is type 2 diabetes [3]. It is characterized by chronic insulin resistance and declining β-cell function (relative insulin deficiency or insulin secretory defect) [3, 4]. Opposed to T1D patients, patients of type 2 diabetes often do not require a life long insulin treatment to survive[3]. This form of diabetes is associated with obesity and its risk increases with age [3].

β-cells produce insulin and play an important role in maintaining delicate physiological glucose levels [1]. Their destruction by autoantibodies in the course of T1D leads to a loss of blood glucose control and usually absolute insulin deficiency [3]. This can result in multiple consequences ranging from ketoacidosis to severe hypoglycaemia or even amputations, blindness and kidney failure [1].

Patients with a genetic predisposition do not necessarily develop T1D in later life. Additionally it requires exposure to one or more environmental factors that initiate β-cell autoimmunity [2, 5]. Suggested triggers include viral infections or early exposure to certain foods such as cow's milk or gluten [2]. Prior to T1D onset, patients develop islet tissue-specific autoantibodies (AABs). The 4 AABs associated with T1D are: glutamic acid decarboxylase (GADA), islet cell anti-

bodies (ICA), insulin autoantibodies (IAA) and ZnT8 antibodies (ZnTA) [5, 6]. The process of developing one of these AABs is referred to as "seroconversion" and studies have shown that the risk of progression to T1D later in life increases, reaching near certainty, depending on the number of circulating AABs and at which age seroconversion took place [5]. Children who develop multiple islet autoantibodies at an age younger than 3 years are especially at risk [7]. Furthermore it is known that *a priori* family history of T1D in first degree relatives, particularly if more than one relative is affected, also represents a major risk factor [5].

Many T1D susceptibility loci are found within the human leukocyte antigen (HLA) region, along with strong resistance alleles, making the HLA complex on chromosome 6 very useful in the context of determining high risk patients [1]. Furthermore, inherited susceptibility also resides predominantly in HLA genotypes (DR and DQ) [2]. Evidence from animal models and humans have further indicated that auto-reactive T cells play an important role in disease initiation and progression [1].

The incidence of T1D is increasing rapidly worldwide and simultaneously a shift of disease onset to an earlier age (more incidences in children ages <5 years) is observable [8]. The time period in which this development has taken place is too short in order to explain it with genetic effects [8]. An increase in frequency of T1D risk loci has also not been observed [8]. Additionally, there is a large geographical variation in the incidence rates around the world. China is reported to have one of the lowest T1D occurrences, with about 0.57 cases per 100 000 population (<18 years of age/year) [2]. The UK is reporting an incidence rate that is around 30 times higher [2]. The countries with the highest rates (reaching nearly 100-fold) are Finland and Sardinia with about 48–49 per 100 000/year [2]. It is notable that migrating populations, for example south Asian children in the UK, fairly quickly take on incidence rates of their new population which can be very different to their native one [2]. These observations strongly support that the causes of T1D are environmentally linked, making the effect of non-genetic factors a crucial key to understanding etiology and pathogenesis of T1D. [8].

1.2 Epigenetics

The epigenome refers to the entity of mitotically or meiotically heritable changes across the genome which also effects gene expression in a cell at any given point in time [9]. It is highly dynamic, influenced by the interplay of multiple factors such as genetic determinants, lineage-specific cues and envi-

ronmental factors [10]. The term "epigenetics" means "above the genetics" and indicates that these changes do not affect the DNA sequence itself [9]. Initially it referred to chemical modifications to DNA molecules and histone proteins but presently the term also includes other molecules that can transmit epigenetic information (such as non-coding RNAs) [9].

The most extensively investigated mechanisms are post-translational modifications of histone proteins and methylation of DNA, the latter of which we will focus in this thesis. By regulating chromatin structure and DNA accessibility, these changes influence the control of gene expression and gene silencing across different developmental stages, tissues and diseases [11].

1.2.1 Cytosine Methylation

DNA methylation (DNAm) is an epigenetic modification, in which cytosines in cytosine – guanine dinucleotides (CpGs) are methylated by adding a methyl group to the 5′ position on the cytosine pyrimidine ring (see fig. 1.2) [12]. Although the methylation of cytosines in CpGs is presumed to be the predominant form of DNAm, recent studies suggest that CpH methylation (with $H = C|A|T$) may also play an important role. In this thesis DNAm will be referring to CpG methylation if not stated otherwise. [10].

Regions of the genome with higher G+C and CpG frequency than expected are called 'CpG Islands' (CGIs). Although no objective standard exits, generally regions with a minimum length of 200bp, an observed-to-expected CpG ratio > 60 % and a GC content of > 50 % are considered as CGIs [13, 14]. Around half of the mammalian genes are associated with one or more CGIs, which are often located in the promotor region [15].

DNAm plays an important role in a diverse range of cellular pathways including tissue-specific gene expression, cell differentiation and X chromosome inactivation [11]. For example, most promoter-associated CGIs are unmethylated, but in silenced areas such as the inactivated X chromosome in females, they generally show methylation [17]. Transcription can thereby be repressed by two distinct mechanisms; either directly by inhibiting the binding of transcription factors (TFs) or indirectly by recruitment of methyl-CpG-binding proteins and their associated histone-modifying enzymes which can establish a silenced chromatin state (see figure 1.1) [9].

In early embryogenesis the DNA is largely free of methylation. *De novo* methylation is initiated by DNA (cytosine-5-)-methyltransferase-3α (DNMT3A) and -3β (DNMT3B) [17]. In CpG islands this can trigger a silencing cascade and consequently repress transcription [17]. In order to preserve methyla-

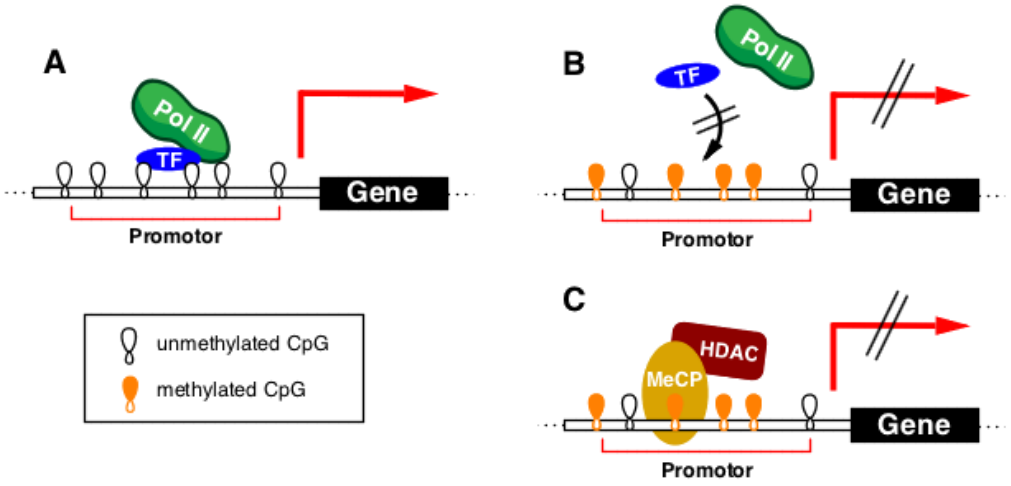

Figure 1.1: DNAm and transcriptional repression. **A** No methylation in promotor sequence. Transcriptional factors (TFs) and RNA polymerase II (pol II) can bind, resulting in transcription of gene. **B** Direct gene silencing by methylation in the DNA-binding sequence of some TFs. This results in inhibition of TF binding and consequently repression of transcription. **C** Indirect gene silencing by methyl-CpG-binding proteins (MBPs). Here the targeted binding of MeCP2, a dynamic repressor of neuronal genes, is shown. After binding, MeCP2 recruits histone deacetylase 1 (HDAC) to create a region of silenced chromatin [16].

tion patterns and maintain silencing during replication, the newly synthesized DNA must be methylated accordingly. This is accomplished by methyltransferase DNMT1 which has specificity for hemi-methylated CpG dinucleotides and can methylate CpGs based on the presence of methylation on the complementary template strand [11]. In early embryogenesis, epigenetic reprogramming largely erases adult pattern of methylation [17].

It is important that epigenetic marks can be accumulated (crucial role in cell differentiation) and stably maintained as well as erased in the germ line to allow gender-specific methylation (see figure 1.2) [18]. However, not all epigenetic marks seem to be erased between generations (for example some mutagenic retrotransposons) leading to multi-generational influences of unknown extent [18, 19]. The underlying mechanisms of epigenetic reprogramming and how certain marks resist global demethylation are not fully understood [19]. In addition to these dynamic DNAm patterns in normal development, DNAm variation can also be a consequence or cause of disease as seen in extensive studies in

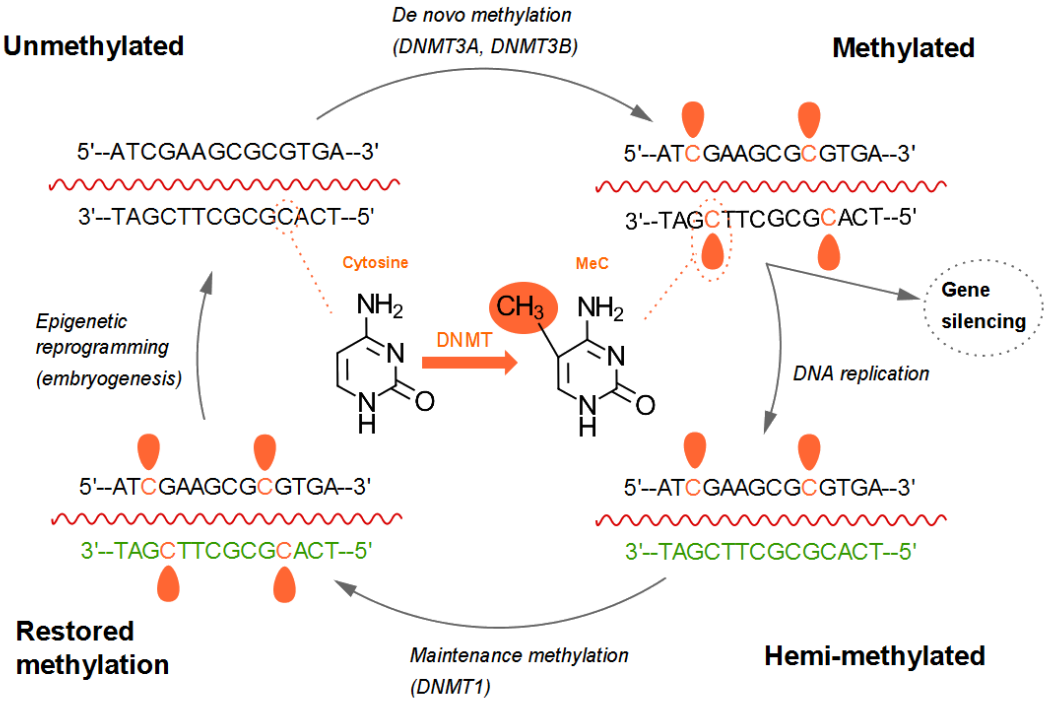

Figure 1.2: Cytosine Methylation. Early embryogenesis with methylation free DNA (top left). *De novo* methylation leads to methylated DNA sequence and gene silencing (top right). DNA replication resulting in hemi-methylated DNA (bottom right). Maintenance reaction through DNMT1 restores methylation (bottom left). Demethylation (epigenetic reprogramming) restores initial state (left). The methylation reaction *cytosine + DNMT → MeC (methylated cytosine)* is depicted in the center. *Figure adapted from* [17].

the context of cancer [20]. This may provide promising and novel opportunities for identifying loci associated with common diseases.

1.2.2 Environmental Factors and Disease Susceptibility

Evidence from animal studies suggest that prenatal and early postnatal environmental factors can alter epigenetic programming (especially DNAm patterns) and play a role in susceptibility to disease in later life [21, 22]. Furthermore Wolff *at al.* found evidence of maternal inheritability of epigenetic phenotypes

using an Avy/a mice model, suggesting that epigenetic marks may potentially also affect the health of future generations [23].

Epigenome wide association studies (EWASs) and advances in genomic technologies have recently made it possible to conduct large-scale studies of epigenetic variation in human diseases [10]. The results of these studies in combination with findings in animal models strongly support the assumption that 'developmental programming' plays an important role in the development and outcome of disease in adult life [10]. By encountering a specific environment during critical windows in fetal development and infancy, adaptive responses can lead to long-term changes in tissue structure or function [24]. If the environment leading to such programming is sub optimal or does not reflect the one encountered in adulthood, these changes can be unfavorable or even lead to disease susceptibility [25].

Exposure to nutritional, chemical and physical factors can lead to long-term effects on gene expression [9]. Genomic regions that are likely to be affected by environmentally induced epigenetic marks are often CpG enriched, for example promoter regions of housekeeping genes or regulatory elements of imprinted genes [9]. Furthermore, the characterization of expression profiles of genes prone to such marks will hopefully identify epigenetic biomarkers [9]. This resembles an important step towards early diagnosis of individuals with a high risk of adult-onset disease and could also assist in prevention and treatment.

1.2.3 Epigenome-wide Association Studies

Epigenome-wide association studies (EWASs) are large-scale studies of disease-associated variation, designed to help elucidate non-genetic determinants of human diseases [10]. They represent an epigenomic equivalent to genome-wide association studies (GWASs) which have uncovered a vast range of single nucleotide polymorphism (SNP) associations for diseases and other traits [26]. In order to distinguish between inter-individual and disease associated variation, it is essential that such studies are performed with adequate genome coverage and sample size [10]. Due to these requirements, large-scale EWASs have only recently become practical. As a result of advances in epigenomic profiling technologies, DNAm (specifically CpG methylation) profiling of large sample populations is now feasible with high throughput and at an affordable price [10].

Though similar to GWASs, the design of EWASs requires more specific considerations concerning sample selection, as DNAm patterns are known to vary across different developmental stages and tissues [10].

1.3 Umbilical Cord Blood

Umbilical cord blood is often used in studies investigating the impact of environmental factors on fetal development. It contains the newborn's epigenome and can therefore help understand and analyze the effect of *in utero* exposures on DNAm and long-term health effects [27]. It has been suggested by previous studies that prenatal exposure to chemical compounds [28], nutritional supplements [29] or maternal smoking [30] can influence methylation patterns in fetal cord blood and furthermore play a critical role in the development of disease later in life [31]. While analyzing and interpreting whole cord blood it is important to regard tissue specific DNAm and control for different cell types and cell concentrations [29]. Cord blood is often collected in the context of birth cohort studies [27].

1.4 Thesis Outline

Epigenetic variation can be causal or a consequence of disease [10]. Furthermore variation can be inherited to some extent and influenced by non-genomic factors. This makes the task of characterizing causative patterns arising prior to any signs of disease very difficult. [10].

The aim of this thesis was therefore to investigate the influence of environmental factors on methylation patterns in children with familial risk of T1D. Utilizing data provided by the German-wide interventional trial BABYDIET, epigenome-wide association studies were conducted in order to find methylation sites associated with maternal exposures such as smoking or dietary supplements and child characteristics such as sex, birth weight, seroconversion and T1D onset. Methylation data from umbilical cord blood was used with the fundamental assumption that the children have not been exposed to any environmental factors except those of the maternal uterine environment. By establishing a link between DNAm patterns in umbilical cord blood, environmental factors and information on future events (seroconversion, T1D onset), I hope to gain insight into developmental programming *in utero* and investigate potential influence of environmental factors on susceptibility to T1D.

2 Materials and Methods

This chapter gives an overview over the cohort and data used. Furthermore, it introduces methods that were used for quality control, data analysis and evaluation.

2.1 Dataset

The following section describes and provides background information on the study cohort and used data.

2.1.1 BABYDIET

BABYDIET is a German-wide study, initially conducted to explore the effects of dietary gluten delay in infants with a strong genetic predisposition for T1D [32]. Early introduction of gluten was presumed to increase the risk of islet autoimmunity in childhood [32]. Newborn children that were younger than 2 months and without prior exposure to gluten or cereals were eligible to participate in the study [32]. Furthermore, they had to have two or more first-degree relatives with T1D or one first-degree relative and additionally, one of the known HLA genotypes that confer a high T1D risk [33]. The children were divided into two groups; control group (exposure to gluten at 6 months of age) and late-exposure group (at 12 months) and were monitored by monthly and later yearly checkups [33]. The study could not observe a significant decrease in the risk for islet autoimmunity for the late exposed children and also found no evidence of an effect on growth [33]. Despite the non-effective dietary intervention, the study participants were further followed and are since used as a birth cohort study, observing the development of children with a high familial risk for T1D and potential pathogenesis of disease.

In this thesis, data from 126 children from the BABYDIET cohort was used. For each sample, identified with a unique identifier consisting of numbers and letters, information on birth weight, sex, current state of health or disease (seroconversion/progression to T1D) and various maternal factors was provided

Table 2.1: Descriptive characteristics of the BABYDIET cohort (n = 126).

Characteristic	n (%)	male	female
Child specific			
Gender	-	54	72
Seroconversion	21 (16.7)	10	11
First degree relatives (≥2)	22 (17.5)	6	16
C-section	45 (35.7)	18	27
Birth weight (average in g)	3461.2	3492.453	3438.194
Health State (current)			
Healthy	105 (83.3)	44	61
Seroconversion	21 (16.7)	10	11
T1D	13 (10.3)	6	7
Mother specific			
Dietary			
Salt-water fish	79 (62.7)	35	44
Fish oil	9 (7.1)	0	9
Folic acid	92 (73.0)	39	53
Iron	64 (50.8)	28	36
Other supplements	86 (68.3)	37	49
Smoking			
Non - smoker	97 (77.0)	40	57
During conception	22 (17.5)	11	11
During pregnancy	10 (7.9)	4	6

Table 2.2: Samples that were excluded from the dataset (see section 2.2).

ID	Sex	Fdr[a]	Health[b]	Csec[c]	Bw[d]	Gw[e]	Sm_c[f]	Sm_p[g]
91C3	f	0	h	0	3550	38	0	0
75J5	m	0	h	0	4570	39	0	NA
73C1	m	0	h	1	4200	38	0	0

[a] definition see section 2.1.1 [b] h=healthy [c] C-section [d] birth weight (g)
[e] gestation week [f] smoking (conception) [g] smoking (pregnancy)

as listed in table 2.1. Additionally, for all samples information concerning first-degree relatives with T1D (mother, father, sibling) was available. The dataset consisted of 72 (57%) females and 54 (43 %) males, of which 22 % and 11 % had more than one first-degree relative with T1D, respectively (see figure 2.1B). After quality control (as described in section 2.2) the sample size was reduced to n=123 with 71 (58 %) females and 52 (42 %) males (see figure 2.1A) .

Environmental factors were given either as a continuous (e.g. birth weight) or a categorical variable. Categorical variables such as intake of supplements, which were reported with the levels "none", "low dosage", "medium dosage" and "high dosage" , were transformed into binary variables with 0 = "no intake" and 1 = "intake" to ensure appropriate numbers for category size. For example folic acid was taken as a supplement in high dosage by 80 women. In contrast to this only 8 women reported medium intake and low only 2. Smoking was also re-coded as a binary variable for the analysis. The variable referred to as "fdr" confers information about the familial history of T1D. Children with more than one first degree relative (fdr) are at a markedly higher risk for islet autoantibodies [5] and were assigned "1", children with only one first degree relative "0".

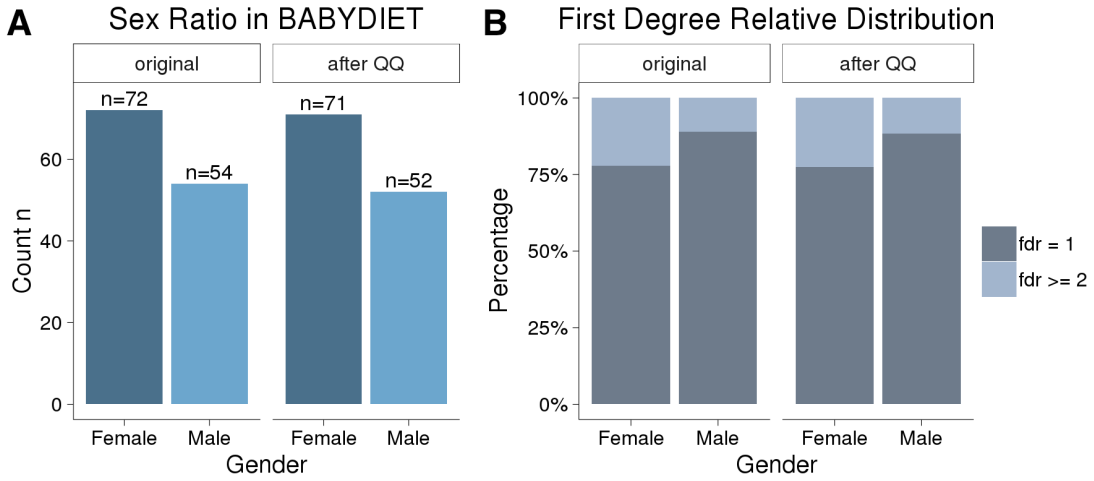

Figure 2.1: (A) Sex ratio in BABYDIET cohort before (n=126) and after quality control (QQ) (n=123). (B) Proportion of children at higher risk of developing islet autoantibodies due to more than two first-degree relatives with T1D. Proportion is shown for female and male samples, before and after QQ.

Methylation Data

Genome-wide DNAm data was generated using the Illumina Infinium Human Methylation450k BeadChip assay. The methylation level for a specific CpG site i is represented as a β-value, which is defined by a combination of the methylated (M_i) and unmethylated (U_i) signal intensities [34]:

$$\beta - value_i = \frac{max(M_i, 0)}{max(M_i, 0) + max(U_i, 0)} \tag{2.1}$$

This value ranges between 0 and 1, corresponding to 0 % and 100 % methylation at this site, respectively. A β-value of 0.5 indicates "hemi-methylation" as seen in monoallelically expressed genes (see figure 2.2B) [35]. The distribution of these raw beta values is often skewed and displays heteroscedasticity, making them inappropriate for statistical models that assume a normal distribution [34]. Therefore, the data was further processed using methods described by Simone Wahl [34], including quantile normalization and PCA on control groups, in order to transform the values and eliminate technical variance. The resulting values (referred to as β-values in this thesis) are beta residuals, indicating if a CpG site is hypo- or hypermethylated (negative or positive sign) compared to the control probes.

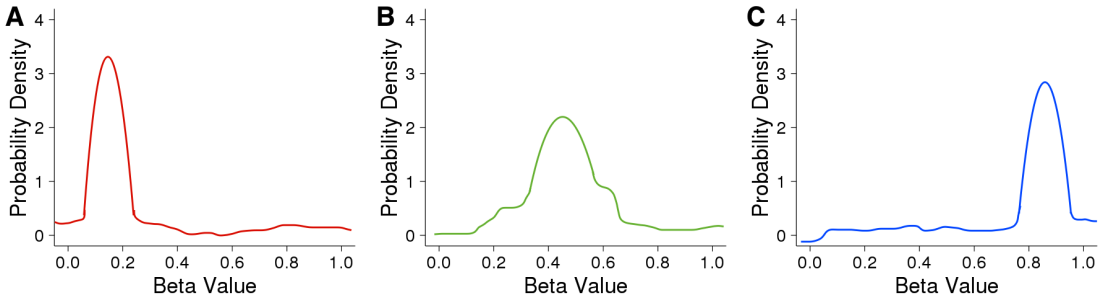

Figure 2.2: Probability density plot showing different frequency distributions of β-values. (A) An abundance of hypomethylated CpGs lead to a spike in very low β-values. (B) The methylation of only one allele (hemi-methylation) shows as an accumulation of β-values around 0.5. (C) When both alleles are methylated (hypermethylated CpGs), the distribution of β-values peaks close to 1.

Cell Estimates

Whole blood is composed of a variety of different cell types, each with its own unique DNAm pattern [36, 37]. In order to evaluate if differences in methylation are potentially disease-related or attributable to some other exposure (such as smoking), it is important to rule out variation due to different distributions of cell type proportions [38]. Houseman *et al.* [37] designed an algorithm that can estimate the proportions of immune cells in whole blood. This is important because changes in DNAm often reflect the underlying immune response rather than changes induced directly by disease. Furthermore the complete assessment of the immune profile requires extensive and expensive measurements [37]. The method proposed by Houseman *et al.* does not require these steps. It uses the the principal immune components of whole blood (consisting of CD8+ and CD4+ T cells, natural killer (NK) cells, B cells, monocytes and granulocytes) and 500 CpGs that show differential composition regarding those cell types, to fit regression models for these sites. The thereby estimated coefficients of cellular composition are then used to predict the relative proportion of cell type components in whole blood [37].

Cell estimates obtained by the procedure described above were included as covariates in the regression model (see equation 2.5) in order to adjust for possible confounding. It is important to note that this method was developed for an adult population and may not be accurate for cord blood analysis.

2.2 Quality Control

Outliers are defined as measures or observations that are suspiciously bigger or smaller than the majority [39]. These data points can influence statistics by altering the mean, increasing variability and in the case of linear regression, distorting the fitted model. Causes can reach from technical failures to spontaneous biological events or simply non representative samples.

This sections gives detail on the quality control methods applied in order to detect and remove such outliers. Before quality control the dataset consisted of 126 samples and 485512 methylation sites (also referred to as probes).

2.2.1 Filtering

In an initial step, samples and probes with poor quality were filtered and excluded. If methylation data regarding a sample was missing for more than 20 %

of the probes, the sample was excluded. Fortunately no sample met this criteria. Probes were filtered by the same procedure, if data concerning a methylation site was missing for over 20 % of the samples it was precluded. On this basis (and after removal of 3 samples, see section 2.2.2 and section 2.2.3) 1326 probes were excluded. In this step all CpG sites located on the Y chromosome were already automatically excluded (see sex ratio of children, table 2.1). To further minimize sex-specific methylation bias [28], all remaining probes on the X chromosome (n=11218) were removed, resulting in a total of 472968 methylation sites used for further analysis.

2.2.2 Z-Score Cutoff

In order to detect global sample outliers and isolated outliers that do not occur systematically, a z-score based cutoff was applied.

Sample Outlier Detection

In the first step, the amount of potential outliers per sample for a given cutoff was determined. A β-value was regarded as an outlier, if its z-score was over a specific threshold. This is a common method to detect outliers and can be applied if the data is assumed to follow a normal distribution [39]. The z-score is defined as follows

$$z = \frac{x - \mu}{\sigma} \tag{2.2}$$

with x being the raw value, μ the mean of the population and σ the standard deviation of the population [40]. It indicates how many standard deviations an observation is away from the mean. Positive or negative values imply above or below the mean, respectively.

Z-scores regarding a methylation site were calculated for the β-values measured per sample (maximal n=126). This was done separately for all 472993 sites. Subsequently the amount of probes with a z-score higher than a defined cutoff was determined for every sample and plotted (as seen in figure 2.3). As proposed by Cousineau *et al.* , the effects of a cutoff between 3σ and 4σ were evaluated in order to find the most suitable for the dataset. The results of this analysis can be seen in figure 2.3. It is notable that regardless of which cutoff was used, both result in the same four samples with the highest count of "outliers", while the ranking of the rest changes. Furthermore, a steep decline in

the overall number of CpG sites over the threshold was observable between the cutoffs (see figure 2.3, A and B).

Taking into account the results of the principle component analysis (section 2.2.3), the samples "91C3", "75J5" and "73C1" were excluded from the dataset, as they appear to be sample outliers.

Individual Cell Filtering

Initial analysis showed that the three excluded samples did not account for all observed outliers, making an additional, more flexible "local" outlier filtering indispensable. In this context all measured values (regardless of sample or methylation site) were set to missing if they exceeded 4 standard deviations from the mean.

A z-score of 4 was chosen to prevent too harsh quality control and limit the extent of data exclusion. A total of 65613 data points were set to missing, which is only a fifth of what would have been excluded if the stricter cutoff was chosen. Furthermore, EWAS performed on the factor "first degree relative" with a cutoff of 3 showed decreased association with HLA genes in contrast to data filtered by a cutoff of 4 (results not shown). As these HLA genes are known to carry heritable T1D-associated information [2], this association can be regarded as true. Thus a z-score cutoff of 4 was chosen, resulting in a dataset with a total of 296975 missing values (representing a 28% increase of missing values).

2.2.3 Principal Component Analysis

After removal of the X and Y chromosomal probes, principal component analysis (PCA) was performed using probes with methylation data present for all 126 samples. 345522 probes met this criteria. PCA is a technique that helps reduce dimensionality of a dataset, transforming it into a more interpretable form, while preserving as much of the initial variation and information as possible [41]. The transformed data set consists of so called "principal components" (PCs). These uncorrelated variables are ordered so that the first principle component accounts for the largest variance in the data, followed by the component with the second largest and so on [41]. This allows a 2-dimensional representation of the data as seen in figure 2.4[42].

Examination of the association between the first three PCs and sample specific factors (sex, birth weight, maternal factors) did not show any distinguishable clustering. However, the choice to exclude the sample outliers identified in section 2.2.2 can be further supported by the PCA plots depicted in figure 2.4.

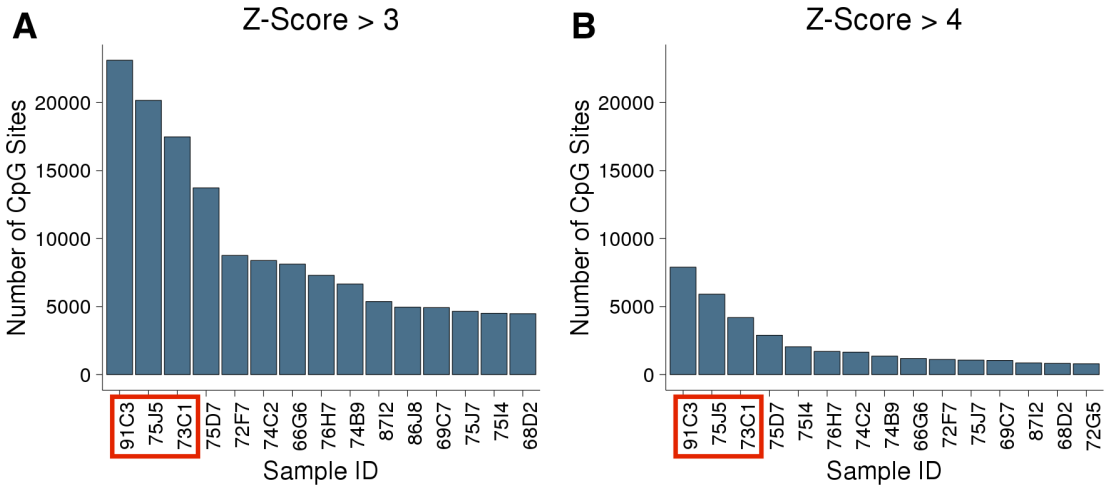

Figure 2.3: Bar plot showing the number of CpG sites per sample that have a z-score over 3 (A) or 4 (B). The 15 samples with the most deviating data points are shown in descending order. A cutoff of 3σ results in 340300 outliers in total (A), whereas the less strict threshold of 4σ only declares 65613 data points as such (B). The samples which were ultimately excluded are shown in the red box.

The distribution of samples in the first and second PC, accounting for 17.6% and 7.9% of variation, shows a dichotomous distribution. No factor could be identified separating the samples. However, the sample "91C3" (figure 2.4A) is clearly in an outlying position, supporting the findings of section 2.2.2. Although "68J5" may also be considered as outlying, no additional information (PC2-3, z-scores) suggested the exclusion of this sample.

Figure 2.4B depicts the second and third PC and even though this plain reflects less variance (7,9% and 4,2%), it is interesting to see that the other two identified sample outliers ("75J5" and "73C1") show the highest deviation from the distribution.

All in all these findings support the descicion to exclude the samples "91C3", "75J5" and "73C1" while finding no evidence of other severe outliers.

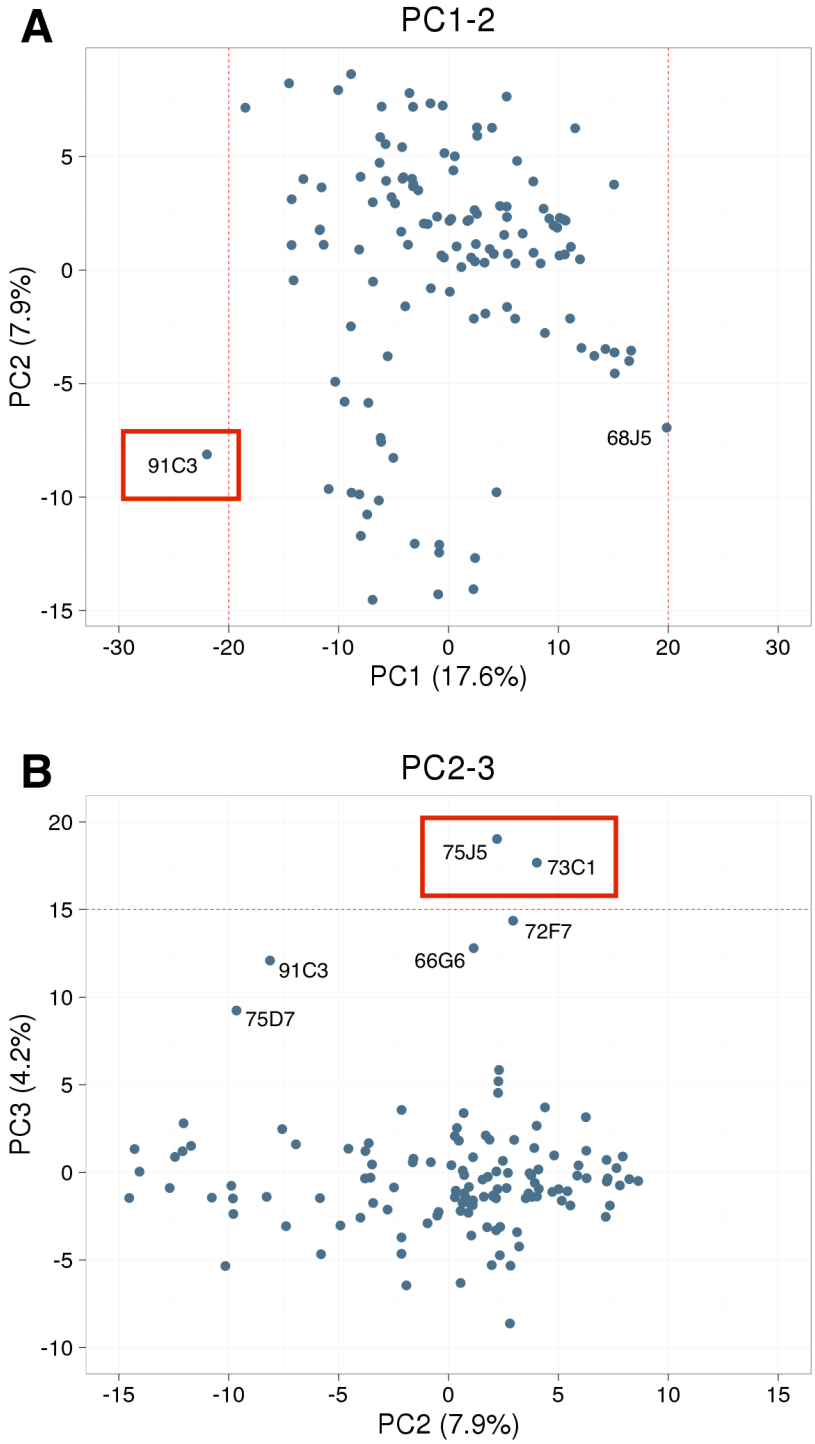

Figure 2.4: Principle Component Analysis (PCA) of complete probes (n=345522) for 126 samples. (A) First and second components (PC1 and PC2). (B) Second and third component (PC2 and PC3). Red lines depict values chosen for cutoff. Samples 91C3 (A), 75J5 and 73C1 (B) were excluded by this criteria.

2.3 Multivariate Linear Regression

2.3.1 Theoretical Background

Multivariate regression is used to describe the relationship between a response variable Y with $Y = y_1, y_2, ..., y_n$ and a set of p predictor variables X with $X = x_1, x_2, ..., x_n$ in terms of a linear function [43]. The general regression model for $i = 1, ..., n$ is given by

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip} + \epsilon_i \tag{2.3}$$

where ϵ_i denotes the independent random error and the unknown parameters are the overall mean (β_0) and regression coefficients ($\beta_k, withk = 1, ..., p$) [43].

The objective is to find a linear function with $b_0, b_1, ..., b_p$ such that the fitted values of y_i, given by

$$\hat{y}_i = b_0 + b_1 x_{i1} + ... + b_p x_{ip} \tag{2.4}$$

are as close to the observed values y_i as possible [43]. The difference between y_i and $\hat{y}_i$ is called a residual.

2.3.2 Multivariate Linear Regression Model

The multivariate linear regression model used for the conducted EWASs was:

$$site_i \sim term_a + fdr + gender + cellestimates \tag{2.5}$$

where $site_i$ is the methylation site i, $term_a$ resembles the environmental factor of interest, fdr is the first degree relative (coded 0/1), $gender$ is also coded binary (0=male, 1=female) and $cellestimates$ gives information regarding the concentration of CD8T-, CD4T-, NK-, Bcell-, Mono- and Gran-cells.

Regression analysis of an environmental factor with methylation levels at each site always included sex and fdr as covariates. This was done to minimize confounding by these variables and to account for their influence on methylation patterns [5, 38] . Regression on both terms separately, only including cell estimates (see chapter 3), further supported this decision. Additionally, it was

corrected for cell concentrations in whole blood (as discussed in section 2.1.1) to account for variability of DNAm patterns in different cell types [36]).

Regression was performed for all methylation sites and the estimate and *p*-value of each association was extracted for further analysis. To evaluate the excess of significant associations, expected *p*-values were plotted against the observed with a log quantile-quantile (Q-Q) plot. The diagonal line (as for example seen in figure 3.1) corresponds to the null hypothesis and the deviation of small *p*-values from that line can give a first impression of potential significance [44]. Close adherence to the line for most *p*-values implies that there is no evidence of systematic sources of specious association or genomic inflation [44]. This was further assessed by calculating the genomic inflation factor λ from *p*-values which is typically used in genome-wide association studies to address the extent of substructure leading to inflation [45]. Epigenome-wide association studies have also used this measure [46]. The inflation factor is defined as the median of the resulting chi-squared test statistics divided by the expected median [47].

For further evaluation, highly associated CpG sites were visualized as scatterplots (for continuous variables) or boxplots (binary variables). Here, the residual of methylation is plotted for each sample against the environmental term of interest. By extracting the model residuals from equation 2.5 without $term_a$ for the respective methylation $site_i$, the direction of association can be shown. Thus the residual of methylation indicates if a CpG site is hypo- or hypermethylated compared to the average.

2.3.3 Interaction Terms

In order to evaluate to which extent the association between one environmental $term_a$ and methylation depends on a second $term_b$, linear regression was performed with interaction terms. An example question may be: Does the association of DNAm patterns with a child's birth weight depend on the condition that their mothers smoked during pregnancy?

For this purpose, besides including the main effects ($term_a + term_b$), an additional interaction term is introduced into the regression model (denoted as $term_a * term_b$):

$$site_i \sim term_a + term_b + \underline{(term_a * term_b)} + fdr + gender + cellestimates \quad (2.6)$$

Methylation residuals for plotting were obtained by extracting the model residuals from equation 2.6 without the interaction term ($term_a * term_b$).

2.4 Evaluation Methods

This section explains the applied methods for significance evaluation and lists programs and packages used throughout the thesis.

2.4.1 Multiple Testing Correction

In order to assess the genome-wide significance of associations between methylation level and environmental factors, regression analyses compute a statistical confidence measure [48]. This *p*-value resembles the probability that an observed association with the same strength or larger would occur under the null hypothesis [48]. To determine if an association is statistically significant a confidence threshold α must be chosen. While performing many tests, as done in EWASs with regression analysis of over 480000 CpG sites, the chance of finding associations with very small *p*-values increases just by the mere number of analysis performed [48]. Therefore, adjusting *p*-values is essential to obtain meaningful results.

The most commonly used methods are the Bonferroni correction and the Benjamini-Hochberg false discovery rate, which both result in multiplicity adjusted *p*-values that can be compared to the desired confidence threshold α .

Bonferroni Method

The Bonferroni correction is used to control the "family-wise error rate" which is the rejection of at least one true null hypothesis (type I error) [44]. This means it ensures that for a confidence threshold of $\alpha = 0.05$ and a set of n scores for association, it is 95 % certain that none of the scores would have been observed by chance if the null hypothesis (no association) is true [48]. This is done by designating a *p*-value statistically significant if it satisfies following equation:

$$p - value \leq \frac{\alpha}{n} \tag{2.7}$$

with confidence threshold α and n scores/tests. The multiple testing adjusted

p-values are easily calculated by multiplying the uncorrected p-value with n [48].

Benjamini-Hochberg False Discovery Rate

The false discovery rate (FDR) correction aims at controlling the expected proportion of falsely rejected null hypotheses (false positives) according to a desired FDR level q [49]. It is less strict than the bonferroni and more powerful at the cost of increased rates of type I errors [50]. For exploratory studies, which are not compromised by a certain amount of false positives this method of correction seams most promising [50].

The FDR is defined by Benjamini and Hochberg (1995) [49] as

$$FDR = Q_e = E[Q] = E[\frac{V}{V+S}] \tag{2.8}$$

with Q as the proportion of false positives (V) among the false positives and true positves (S). P-values can be adjusted by the FDR controlling procedure which firstly sorts the values in ascending order and then divides them by their respective percentile rank [48]. The resulting multiple testing adjusted p-values represent the lowest level of FDR for which this observation would be considered significant (null hypothesis rejected) [50].

In this thesis, FDR calculated with the Benjamini–Hochberg controlling procedure (as explained above) was used to determine the significance of CpG sites. The confidence threshold was set to $\alpha = 0.05$.

2.4.2 Programs and R Packages used

Figures in chapter 1 were made with ChemDoodle (v8.0.1) [51]. All analyses and visualizations were performed with *R* (v3.2.0) [52] and the below listed *R packages* if not stated otherwise.

Linear regression was performed using the *lm()* function of the *stat* package. Annotations regarding methylation sites were retrieved by probe name with *get450k()* and *getNearest()* from *FDb.InfiniumMethylation.hg19* (v2.2.0). QQplots and manhattan plots were generated with the package *qqman* (v0.1.2).

> sessionInfo():

- R version 3.2.0 (2015-04-16), `x86_64-redhat-linux-gnu`
- Locale: `LC_CTYPE=en_US.UTF-8`, `LC_NUMERIC=C`, `LC_TIME=en_US.UTF-8`, `LC_COLLATE=en_US.UTF-8`, `LC_MONETARY=en_US.UTF-8`, `LC_MESSAGES=en_US.UTF-8`, `LC_PAPER=en_US.UTF-8`, `LC_NAME=C`, `LC_ADDRESS=C`, `LC_TELEPHONE=C`, `LC_MEASUREMENT=en_US.UTF-8`, `LC_IDENTIFICATION=C`
- Base packages: base, datasets, graphics, grDevices, grid, methods, parallel, stats, stats4, utils
- Other packages: AnnotationDbi 1.32.3, Biobase 2.30.0, BiocGenerics 0.16.1, BiocInstaller 1.20.1, DBI 0.3.1, FDb.InfiniumMethylation.hg19 2.2.0, GenomeInfoDb 1.6.3, GenomicFeatures 1.22.13, GenomicRanges 1.22.4, ggplot2 2.0.0, gridExtra 2.0.0, IRanges 2.4.7, org.Hs.eg.db 3.2.3, qqman 0.1.2, reshape2 1.4.1, RSQLite 1.0.0, S4Vectors 0.8.11, scales 0.3.0, TxDb.Hsapiens.UCSC.hg19.knownGene 3.2.2
- Loaded via a namespace (and not attached): BiocParallel 1.4.3, biomaRt 2.26.1, Biostrings 2.38.4, bitops 1.0-6, colorspace 1.2-6, futile.logger 1.4.1, futile.options 1.0.0, GenomicAlignments 1.6.3, gtable 0.1.2, lambda.r 1.1.7, munsell 0.4.3, plyr 1.8.3, Rcpp 0.12.3, RCurl 1.95-4.7, Rsamtools 1.22.0, rtracklayer 1.30.2, stringr 0.6.2, SummarizedExperiment 1.0.2, tools 3.2.0, XML 3.98-1.3, XVector 0.10.0, zlibbioc 1.16.0

3 Results

This chapter presents the results of the performed epigenome-wide association studies. The -$\log_{10}p$-values of the linear regression for 472968 methylation sites were plotted as quantile-quantile and manhattan plots. In addition to this, box plots and scatterplots helped illustrate direction and strength of differential methylation for key cytosine – guanine dinucleotide (CpG) sites. Information on the genomic context of CpGs was inferred with the R package *FDb.InfiniumMethylation.hg19* and the UCSC Human Genome Browser [53] on *Human Feb. 2009 (GRCh37/hg19) Assembly*.

3.1 Environmental Factors

In order to adjust for gender-specific methylation [28] and take into account potential differences due to heritable methylation marks [2], both first degree relative (fdr) and gender were included in the regression model as covariates. In the following subchapter the findings for gender, fdr, birth weight and maternal smoking are shown. Analysis of the association between dietary supplements and methylation levels did not result in significant findings and are not depicted. Following analyses were also not informative: C-section, seroconversion, birth weight percentile, maternal type 1 diabetes (T1D) and birth weight, maternal T1D and seroconversion, first degree relative (fdr) and seroconversion. Additionally, evidence of differential methylation in cord blood associated with T1D onset later in life is presented.

3.1.1 Gender of Child

Gender-specific methylation is an important factor that must be considered as a potential confounding variable in EWAS. This has been demonstrated in previous studies [30, 31, 54] and recommended by Michels *et al.* [38].

In order to assess the differences in DNAm patterns across both genders, linear regression was performed with the model described in section 2.3 excluding the covariate fdr. As seen in figure 3.1, the log quantile-quantile (QQ) plot

Table 3.1: CpG site with significant association between gender and methylation level. Only one site (cg03769704) is significant ($P_{FDR} < 0.05$). It is located in the promotor region of SLFN5, a member of the Schlafen family.

	Est.[a]	*P*-value	FDR[b]	Chr.[c]	Gene	TSS[d]
cg03769704	0.038	3.43e-08	0.016	17	SLFN5	SLFN5

[a]estimate of linear regression [b]FDR adjusted *p*-value [c]chromosome [d]transcription start site

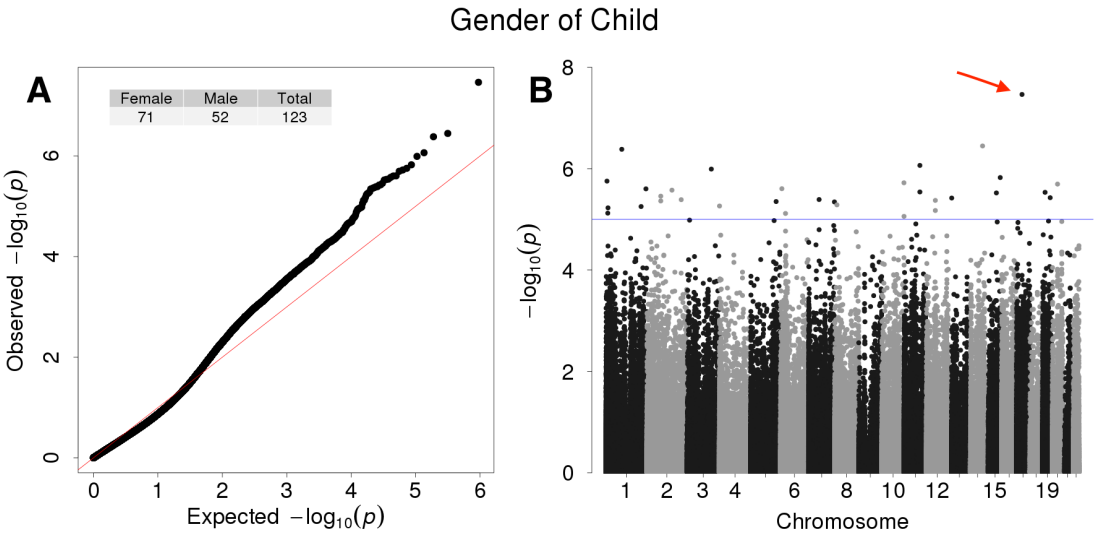

Figure 3.1: Quantile-quantile plot (A) and manhattan plot (B) of epigenome-wide association between gender of child and methylation of 472968 CpGs. Cord blood was analyzed for 71 females and 52 males resulting in one statistical significant association (indicated by red arrow). The blue line is given as orientation, it represents P_{unadj}=0.05.

shows a distinct deviation from the null hypothesis (diagonal line). Calculation of the genomic inflation factor λ showed no evidence of systematic inflation ($\lambda < 1$) thus we can assume that the *p*-value distribution can be entirely accounted for by sex-specific methylation. Although only one CpG dinucleotide fell bellow the significance threshold (table 3.1, $P_{FDR} < 0.05$), the observed *p*-values were overall mostly lower than expected (figure 3.1A). CpG cg03769704 ($P_{FDR} \approx 0.016$) is located in the promoter region of SLFN5 (Schlafen family) and shows a higher methylation level in females than males (figure 3.2). The variability of methylation seems evenly distributed (as seen in figure 3.2) for both genders and outliers are only observed for one site (figure 3.2C). Boxplot

outliers are defined as points with methylation levels lower or higher than 1.5 interquartile range (IQR).

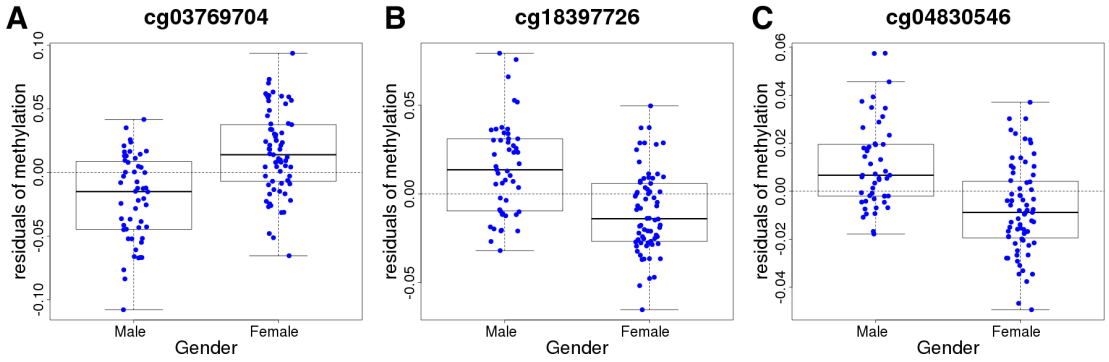

Figure 3.2: Boxplot showing methylation level (residual) by gender for the three highest associated CpGs. cg03769704 ($P_{FDR} \approx 0.016$) shows an increased methylation in females whilst cg18397726 and cg18397726 (both $P_{FDR} \approx 0.065$) are hypermethylated in males compared to females. The horizontal thick line denotes the median, the box itself the upper and lower quantiles of the data. Dots represent samples.

3.1.2 First Degree Relative

Regression analysis on the association between DNAm and number of first degree relatives with T1D (coded 0/1) did not result in statistical significance. However, the QQ plot shows slight deviation for lower *p*-values and the manhattan plot clearly shows an elevation of high *p*-values in a region of the chromosome 6 (colored green in figure 3.3B). As seen in table 3.2 the 10 highest associations comprised of 50 % of CpGs that are located within the human leukocyte antigen (HLA) region.

When isolating key sites (P_{FDR} <0.5) on chromosome 6 (n=9), the proportion of CpG sites located in HLA reaches 100 %, including methylation sites near HLA-DQB1/B2, HLA-DQA1, HLA-DRB1, HLA-DRB1/6 and the GABA-B Receptor 1. The directional changes in methylation levels between children with more than one first degree relative ("1") and children with only one ("0") can be seen in figure 3.4. The five highest associated sites all display decreased methylation in fdr "1" children, though the boxplot also shows that there is great variability in methylation levels for samples with more than one first degree relative (figure 3.4A,D-F).

Table 3.2: Top 10 CpG sites with the highest fdr-association. No methylation site reaches statistical significance ($P_{FDR} < 0.05$) but a notable enrichment in sites located within the major histocompatibility (HLA) complex can be seen.

	Est.[a]	*P*-value	FDR[b]	Chr.[c]	Gene	TSS[d]
cg22984282	-0.173	5.22e-07	0.244	6	HLA-DQB1	HLA-DQB1
cg20720056	-0.046	1.83e-06	0.244	10	ERLIN1	ERLIN1
cg08148418	-0.016	2.35e-06	0.244	21	PTTG1IP	PTTG1IP
cg23785275	-0.088	2.40e-06	0.244	6	HLA-DQB2	HLA-DQB2
cg25306444	-0.023	2.58e-06	0.244	17	LINC00511	LINC00511
cg02919082	0.118	3.49e-06	0.255	6	HLA-DQA1	HLA-DQA1
cg21663668	-0.016	3.81e-06	0.255	2	ANTXR1	MIR3126
cg19301366	0.251	4.42e-06	0.255	6	HLA-DQB1	HLA-DQB1
cg07984380	0.145	5.68e-06	0.255	6	HLA-DRB5	HLA-DRB1
cg05608716	0.006	5.98e-06	0.255	16	MMP25	MMP25

[a] estimate of linear regression [b] FDR adjusted *p*-value [c] chromosome [d] transcription start site

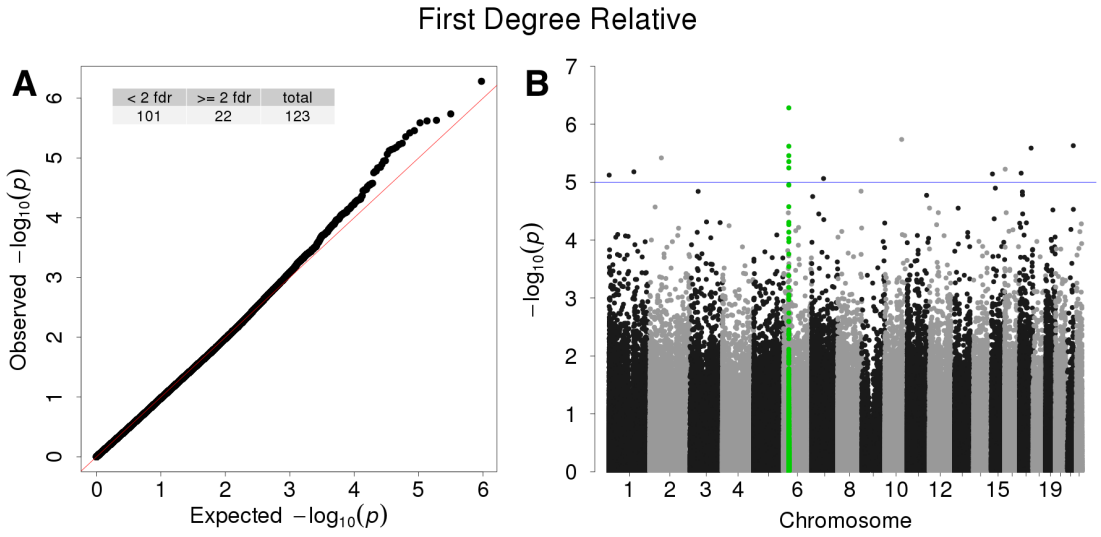

Figure 3.3: Quantile-quantile plot (A) and manhattan plot (B) of epigenome-wide association between the degree of T1D history in first degree relatives and methylation of 472968 CpGs. Cord blood analysis of 123 samples resulted in no statistical significant association but elevated association in a region of the chromosome 6 (colored green). The blue line is given as orientation, it represents P_{unadj}=0.05.

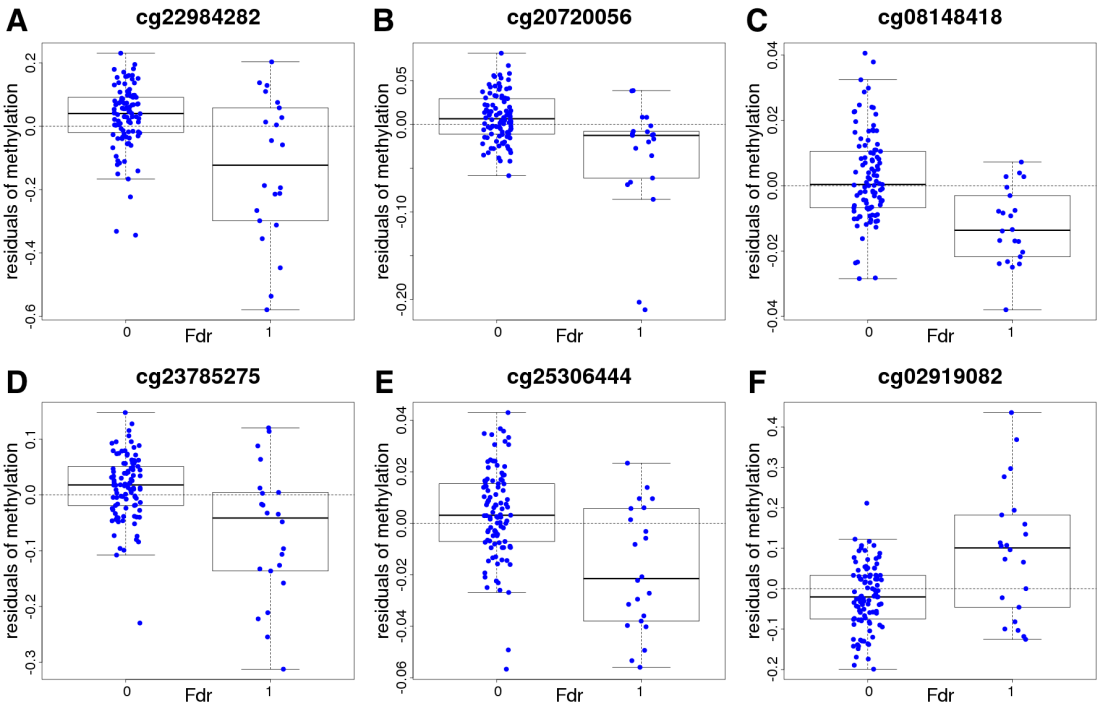

Figure 3.4: Boxplot showing methylation level (residual) by fdr category (0 = only one fdr and 1= two or more fdr) for the six CpGs with the highest association. Nearly all sites (exept F) are hypomethylated in children with more than 1 case of T1D in first degree relatives in comparison to those with only one. Methylation levels for category "1" samples show a larger variation than those of "0" and outliers are present in all sites. The horizontal thick line denotes the median, the box itself the upper and lower quantiles of the data. Dots represent samples.

3.1.3 Birth Weight of Child

Methylation differences associated with the birth weight of children did not reach statistical significance ($P_{FDR} < 0.05$). Regression analysis using birth weight percentiles (taking into account gestation week) resulted in even less significance (results not shown) and is not further discussed.

Interestingly two CpG sites display a higher divergence from the expected *p*-value than the rest (figure 3.6A). cg15681239 and cg05409131 both have a P_{FDR} of 0.07 and are located near the DLEC1 gene and in an intron of INHBA-AS1 as part of a small CpG island (<300bp), respectively (table 3.3). Methylation of cg15681239 shows a negative correlation with birth weight (r=-0.46, figure 3.5A) while methylation at cg05409131 is positively correlated (r=0.45, figure 3.5B). The manhattan plot (figure 3.6B) shows a similar pattern in high *p*-value aggregation in a region of chromosome 6 as seen in figure 3.3B. In contrast to genes in the HLA region, 4 of the top 10 key sites on chromosome 6 are located in a CpG island in the α subunit of the nuclear transcription factor NF-Y (NFYA). They are in close proximity to the transcription start site of the adenylate cyclase 10 (Soluble) pseudogene 1 ADCY10P1 and all show an increase in methylation for children with higher birth weight (table 3.3).

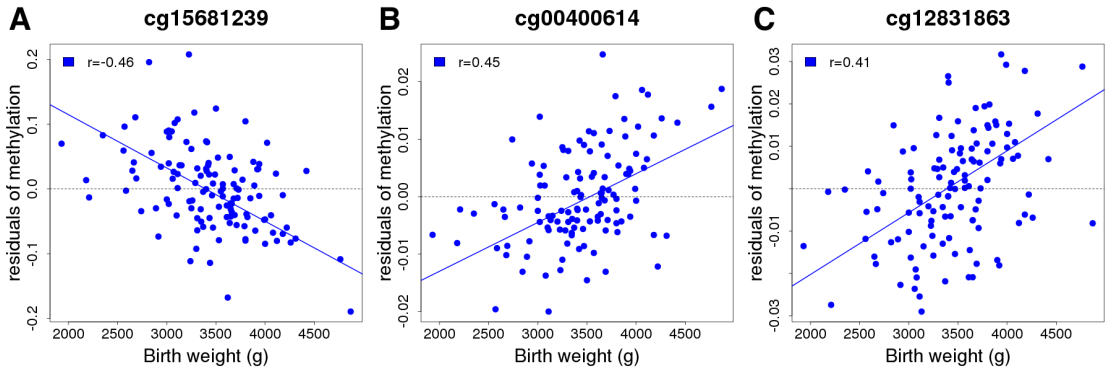

Figure 3.5: Scatterplot showing the directional association between methylation level (residual) and birth weight for the three CpG sites with the lowest *p*-value. The less associated site cg12831863 (r=0.41, $P_{FDR} \approx 0.46$) shows a larger dispersion around the line of best fit than cg15681239 and cg00400614 (with correlation coefficient r = -0.46 and 0.45 and $P_{FDR} < 0.08$). The blue line denotes the linear regression line. Dots represent samples.

Table 3.3: CpG sites with the highest genome-wide association ($P_{FDR} < 0.45$) between methylation level and birth weight (top). No methylation site reaches statistical significance ($P_{FDR} < 0.05$). Chromosome 6 shows an enrichment in CpGs located in the gene NFYA on Chromosome 6. The top 10 sites are shown (bottom).

	Est.[a]	*P*-value	FDR[b]	Chr.[c]	Gene	TSS[d]
Genome						
cg15681239	-6.06e-05	1.63e-07	0.071	3	DLEC1	DLEC1
cg00400614	7.72e-06	3.02e-07	0.071	7	INHBA-AS1	INHBA
On chr.[c] 6						
cg12831863	1.12e-05	2.95e-06	0.46	6	GPX6	GPX6
cg03644281	9.77e-05	1.26e-05	0.479	6	NFYA	ADCY10P1
cg09118053	1.69e-05	2.37e-05	0.561	6	LINC01016	LINC01016
cg27643910	-7.85e-06	3.27e-05	0.596	6	TNXB	TNXB
cg04346459	1.06e-04	3.60e-05	0.607	6	NFYA	ADCY10P1
cg25110423	8.74e-05	4.36e-05	0.607	6	NFYA	ADCY10P1
cg18949415	2.12e-05	4.70e-05	0.607	6	C6orf223	C6orf223
cg05155704	-1.69e-05	4.97e-05	0.607	6	FAM83B	FAM83B
cg02167203	6.22e-05	5.86e-05	0.613	6	NFYA	ADCY10P1
cg26797676	-1.19e-05	6.56e-05	0.613	6	FLOT1	FLOT1

[a] estimate of linear regression [b] FDR adjusted *p*-value [c] chromosome [d] transcription start site

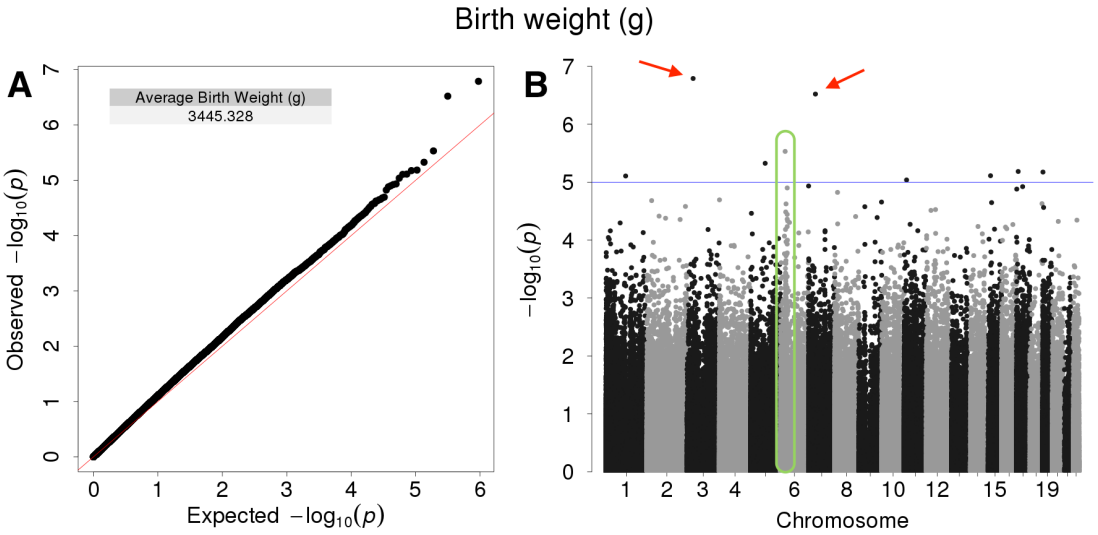

Figure 3.6: Quantile-quantile plot (A) and manhattan plot (B) of epigenome-wide association between the birth weight of the child (in grams) and methylation of 472968 CpGs. Cord blood analysis of 123 samples resulted in no statistical significant association. (A) The QQ plot shows minor inflation (λ=1.19) and notable lower p-values than expected for 2 sites (indicated by red arrows in B). (B) Systematic aggregation of low p-values indicating an association between methylation and birth weight on chromosome 6 (green). The blue line is given as orientation, it represents P_{unadj}=0.05.

3.1.4 Maternal Smoking

Influence of maternal smoking on *in utero* developmental programming and methylation patterns has been reported by various studies [30, 31, 55]. Regression analysis of the BABYDIET cohort found no CpG site that fell below the significance threshold (table 3.4, $P_{FDR} < 0.05$). Smoking was coded as a binary variable, with "0" no smoking during the regarded period of time and "1" smoking reported. Information on the extent of smoking (cigarettes per day) was given but not included due to the very small fraction of smokers paired with high discrepancies in their smoking habits (for example during pregnancy: minimum 3 to maximum 15 cigarettes/day).

Table 3.4: CpG sites with the highest association between methylation level and maternal smoking. The three highest associated sites are shown for the separately performed analysis regarding smoking at conception and smoking during pregnancy. No methylation site reaches statistical significance ($P_{FDR} < 0.05$).

	Est.[a]	*P*-value	FDR[b]	Chr.[c]	Gene	TSS[d]
Conception						
cg06864895	0.015	1.62e-07	0.077	12	SLC38A2	SLC38A2
cg05409131	-0.041	6.47e-07	0.153	3	ACPP	ACPP
cg02762752	0.031	3.13e-06	0.295	16	ZCCHC14	ZCCHC14
Pregnancy						
cg06864895	0.019	1.92e-06	0.425	12	SLC38A2	SLC38A2
cg01574787	0.009	2.59e-06	0.425	7	SLC4A2	SLC4A2
cg21207665	0.028	2.70e-06	0.425	14	PAX9	PAX9

[a] estimate of linear regression [b] FDR adjusted *p*-value [c] chromosome [d] transcription start site

Two analyses were performed separately to determine effects on methylation in children whose mothers reported smoking at conception (n=22) and whose mothers reported to have continued smoking throughout the pregnancy (n=10). Although no site was significant after FDR correction (table 3.4), it is notable that in both analysis the most significant difference in methylation was reported for the CpG cg06864895 with $P_{FDR} \approx 0.076$ (smoking at conception) and $P_{FDR} \approx 0.43$ (smoking during pregnancy). This site is located upstream of the amino acid transporter SLC38A2 and displays increased methylation in both analysis (figure 3.8 and figure 3.9). The distribution of -$\log_{10}$*p*-values indicated higher significance for lower *p*-values for smoking at conception than during pregnancy but also showed signs of a slightly higher inflation with a

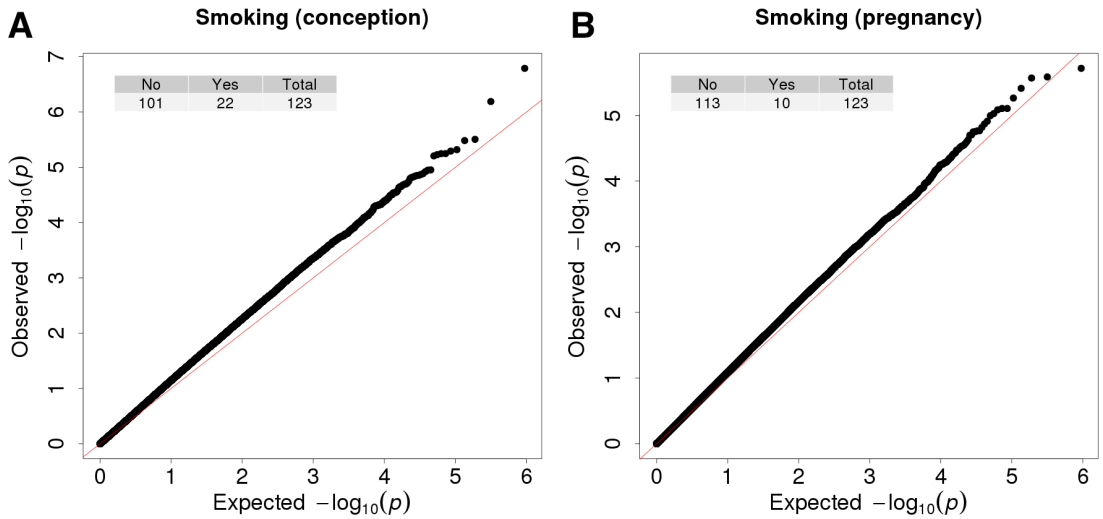

Figure 3.7: Quantile-quantile plot for the association between methylation level and maternal smoking at conception (A) and during pregnancy (B). Cord blood analysis of 123 samples resulted in no statistical significant association. Slight genomic inflation occures in both models, stronger in A with $\lambda \approx 1.22$ than in B with $\lambda \approx 1.13$.

genomic inflation factor of $\lambda \approx 1.22$ (in contrast to smoking during pregnancy with $\lambda \approx 1.13$).

Smoking during conception is further associated with less methylation at cg05409131 ($P_{FDR} \approx 0.15$), located in an intron of the prostatic acid phosphatase ACPP and an increased methylation of cg02762752 ($P_{FDR} \approx 0.29$), located in an intron of the zink finger domain ZCCHC14 (figure 3.8). Smoking during pregnancy was additionally associated with higher methylation of cg01574787 ($P_{FDR} \approx 0.43$) which is located in a short CpG island (<300bp) in the intron of the anion carrier SLC4A2 (cg01574787) and higher methylation of the PAX9 CpG cg21207665 (both $P_{FDR} \approx 0.43$). The boxplots of methylation residuals for both analysis show a higher variability in children born to mothers that smoked during pregnancy (larger boxes as seen in figure 3.9A and B) and the outliers seen in figure 3.9C may be underlying a potentially significant association.

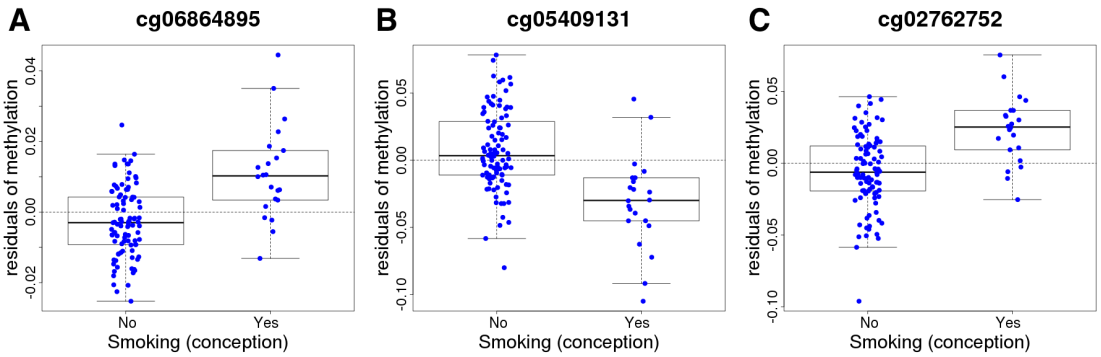

Figure 3.8: Boxplot showing methylation level (residual) for maternal smoking at conception (no = "0" and yes = "1") for the most significant CpGs. A and C show an increase in methylation while B is less methylated. All plots show outliers of which none is extrem. The horizontal thick line denotes the median, the box itself the upper and lower quantiles of the data. Dots represent samples.

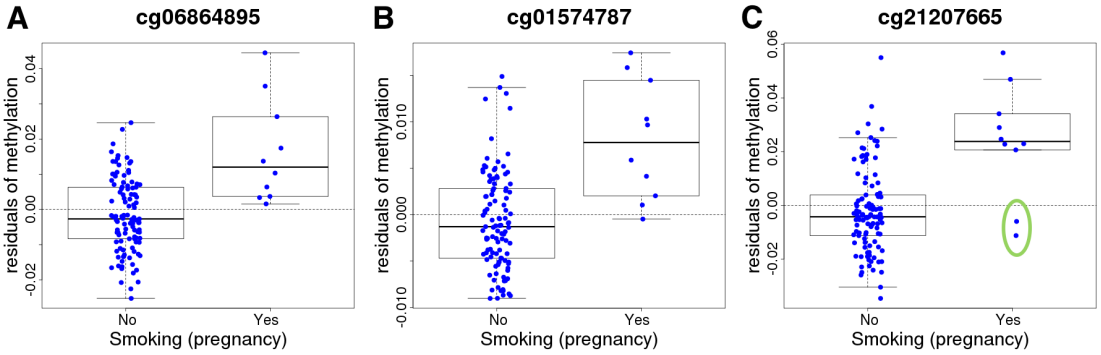

Figure 3.9: Boxplot showing methylation level (residual) for maternal smoking during pregnancy (no = "0" and yes = "1") for the highest associated CpGs. All show an increase in methylation for maternal smoking. The outliers seen in C may have influenced regression, masking an even higher association at this site (green circle). The interquartile region is very large for samples of category "yes" in both A and B. The horizontal thick line denotes the median, the box itself the upper and lower quantiles of the data. Dots represent samples.

3.1.5 Future Event: Type 1 Diabetes

Regression analysis incorporating information on "future" events was also conducted. Seroconversion and the onset of T1D are factors that cannot have directly impacted DNAm *in utero* because this event had not yet occurred at the time when cord blood samples were extracted. However, with this information it is possible to investigate if differences in methylation are observable in children who will develop islet autoantibodies or progress to T1D compared to healthy children. Such CpG sites would have great value for early diagnosis and treatment of T1D.

The analysis of association between seroconversion and methylation levels did not find any significant methylation patterns and is therefore not shown. The results of the regression conducted with samples of 13 children diagnosed with T1D and 110 healthy children can be seen in figure 3.10. After FDR correction cg15293181 reaches statistical significance ($P_{FDR} \approx 0.025$). In total three CpG sites can be seen with differing observed to expected low *p*-values figure 3.10. The aforementioned significant site is located in the SNTG2 gene with close proximity to the transcription start site of thyroid peroxidase (TPO). SNTG2 encodes a protein that belongs to the syntrophin family which is comprised of cytoplasmic peripheral membrane proteins.

cg03153658 is part of a CpG island upstream of the zinc finger protein 470 (ZNF470) and cg10563643 is located near the parathyroid secretory protein CHGA. Boxplots for all three sites show a high variability in the distribution of residuals of methylation for children who develop T1D later in life in contrast to those who do not (figure 3.11). This may partially be attributed to the very small fraction of reported T1D cases.

Table 3.5: CpG sites with the highest association ($P_{FDR} < 0.65$) between methylation level and future onset of T1D. The hypomethylation of cg15293181 in children with diagnosed T1D reaches statistical significance ($P_{FDR} < 0.05$).

	Est.[a]	*P*-value	FDR[b]	Chr.[c]	Gene	TSS[d]
cg15293181	-0.074	5.36e-08	0.025	2	SNTG2	TPO
cg03153658	0.009	6.97e-07	0.165	19	ZNF470	ZNF470
cg10563643	-0.020	1.67e-06	0.263	14	CHGA	CHGA

[a] estimate of linear regression [b] FDR adjusted *p*-value [c] chromosome [d] transcription start site

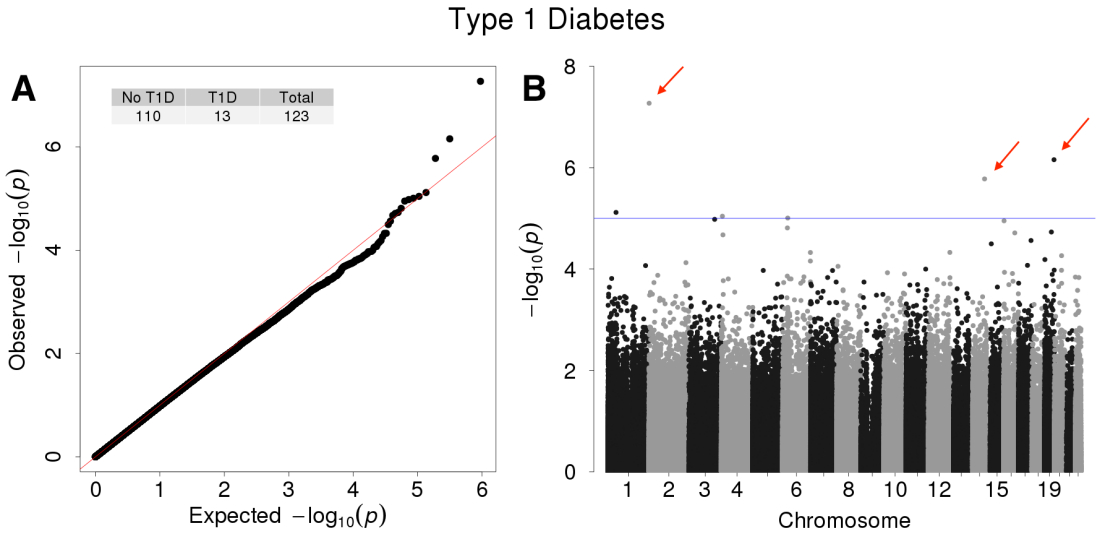

Figure 3.10: Quantile-quantile plot (A) and manhattan plot (B) of epigenome-wide association between methylation level and future T1D onset. Cord blood analysis of 110 children that are healthy (up to last doctor's visit) and 13 children that developed T1D resulted in one statistical significant association. This site and two other CpGs that did not fall below the threshold, but also have low p-values are indicated by red arrows (B).

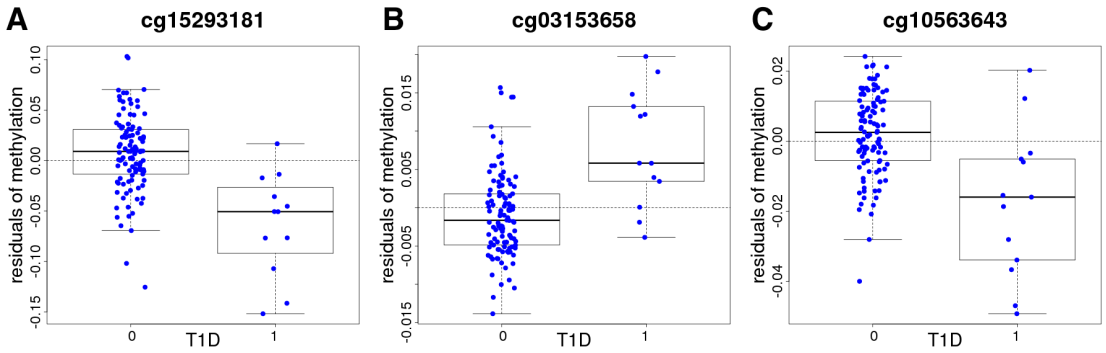

Figure 3.11: Boxplot showing methylation level (residual) by health state (0 = healthy and 1 = has developed T1D) for the three CpG sites with the highest association between methylation and future T1D diagnosis. Methylation sites depicted in A and C show less methylation in children with future T1D onset and the methylation of site B marks an increase. The horizontal thick line denotes the median, the box itself the upper and lower quantiles of the data. Dots represent samples.

3.2 Interaction between Environmental Factors

To investigate co-dependent influences of environmental factors on methylation marks, regression analysis was performed incorporating interaction terms. Unfortunately many did not produce significant results, for example the effect of maternal iron intake during pregnancy and later T1D onset or first degree relatives with T1D and seroconversion. In other cases there were not enough or no children to represent all scenarios. No child that developed T1D later in life was exposed to maternal smoking during pregnancy, making the analysis of interaction between T1D onset and smoking not applicable.

3.2.1 Maternal Smoking and Birth Weight of Child

Effects of maternal smoking on DNAm have been reported by various studies [30, 55] and it is also associated with a reduction in birth weight [56]. To investigate the interaction between maternal smoking and birth weight in the BABYDIET cohort, regression analysis with an interaction term incorporating both variables (equation 2.6) was performed.

The mean birth weight for children exposed to maternal smoking during pregnancy was slightly lower (3413.5 g) than the mean for children of non smoking mothers (3454.954 g), coinciding with the a fore mentioned association (figure 3.13)[56]. The analysis of association between the interaction of smoking at

Table 3.6: CpG sites with statistically significant association between maternal smoking during pregnancy, birth weight and methylation level ($P_{FDR} < 0.05$).

	Est.[a]	*P*-value	FDR[b]	Chr.[c]	Gene	TSS[d]
cg06724462	-4.06e-05	4.93e-08	0.023	1	KCNQ4	KCNQ4
cg09856467	8.24e-05	4.42e-07	0.046	22	TPST2	MIR548J
cg19042497	-5.58e-05	4.91e-07	0.046	18	MYO5B	MYO5B
cg23514537	-7.54e-05	5.98e-07	0.046	5	F12	F12
cg13158344	-9.17e-06	6.34e-07	0.046	6	ULBP1	ULBP1
cg19629818	-3.65e-05	7.91e-07	0.046	7	GPC2	GPC2
cg02331198	1.25e-04	8.11e-07	0.046	6	AIM1	AIM1
cg17208467	-4.38e-05	8.19e-07	0.046	8	TNFRSF10D	TNFRSF10D
cg11217193	-1.07e-04	9.51e-07	0.046	1	VPS13D	SNORA59B
cg05057515	5.36e-05	9.66e-07	0.046	16	CBFA2T3	CBFA2T3

[a] estimate of linear regression [b] FDR adjusted *p*-value [c] chromosome [d] transcription start site

conception and birth weight on DNAm did not produce significant results but when regarding smoking during pregnancy ten statistically significant methylation sites were found (table 3.6, $P_{FDR} < 0.05$).

In figure 3.13 scatter plots of the residuals of methylation can be seen for the significantly associated CpG sites. Correlation between birth weight and level of methylation varies but is overall stronger for samples with exposure to maternal smoking (colored red in figure 3.13). Samples without exposure are only loosely distributed around the line of best fit and with little to no clear direction of association (colored blue in figure 3.13). The depicted CpG sites are predominantly hypomethylated in children with higher birth weight and exposure to smoking than lighter children (A,C-F). The plots indicate that the methylation level at these sites is normally not or only slightly associated with birth weight. Maternal smoking establishes or reverses and intensifies this association, leading to birth weight dependent hypo- or hypermethylation. Especially cg06724462, located in the KCNQ3 gene ($P_{FDR} \approx 0.023$) and cg19629818, located in the glypican 2 (GPC2) gene ($P_{FDR} \approx 0.046$) show very high correlation (r=-0.93 and r=-0.94). Other sites that show a high correlation are cg17208467, situated in proximity of the tumor necrosis factor receptor superfamily member 10D (TNFRSF10D) and cg11217193, located within the vacuolar protein sorting 13 homolog D (VPS13D). On the contrary, for example cg23514537 and cg02331198 seem to be strongly influenced by outlying samples (figure 3.13D,G) challenging their meaningfulness.

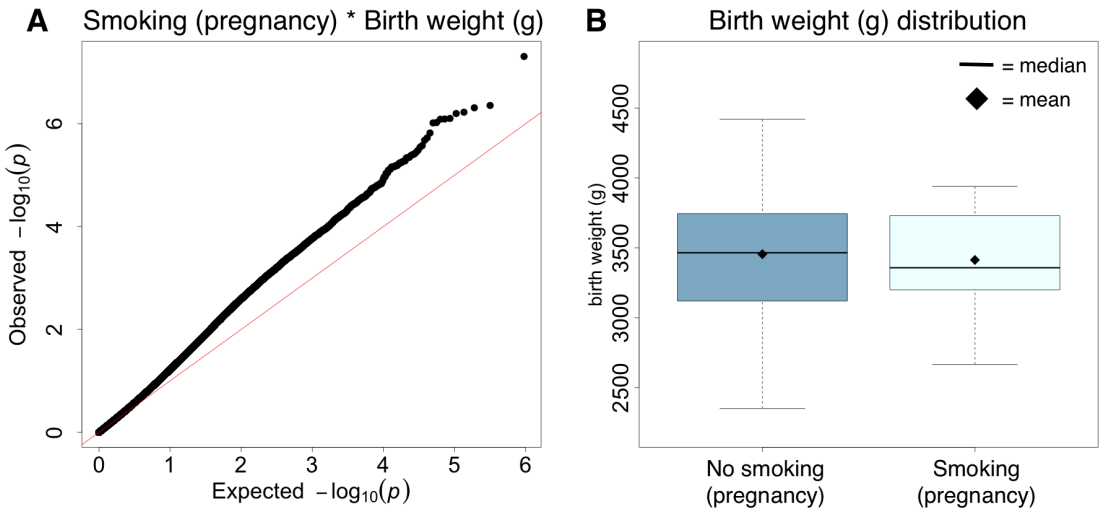

Figure 3.12: (A) QQ plot of epigenome-wide association between methylation level and interaction between maternal smoking during pregnancy and birth weight. A clear upward trend for lower p-values is observable, partly due to genomic inflation ($\lambda \approx 1.99$). (B) Boxplot of the birth weight distribution in children without exposure to smoking (mean 3454.954g) and maternal smoking during pregnancy (mean 3413.5g).

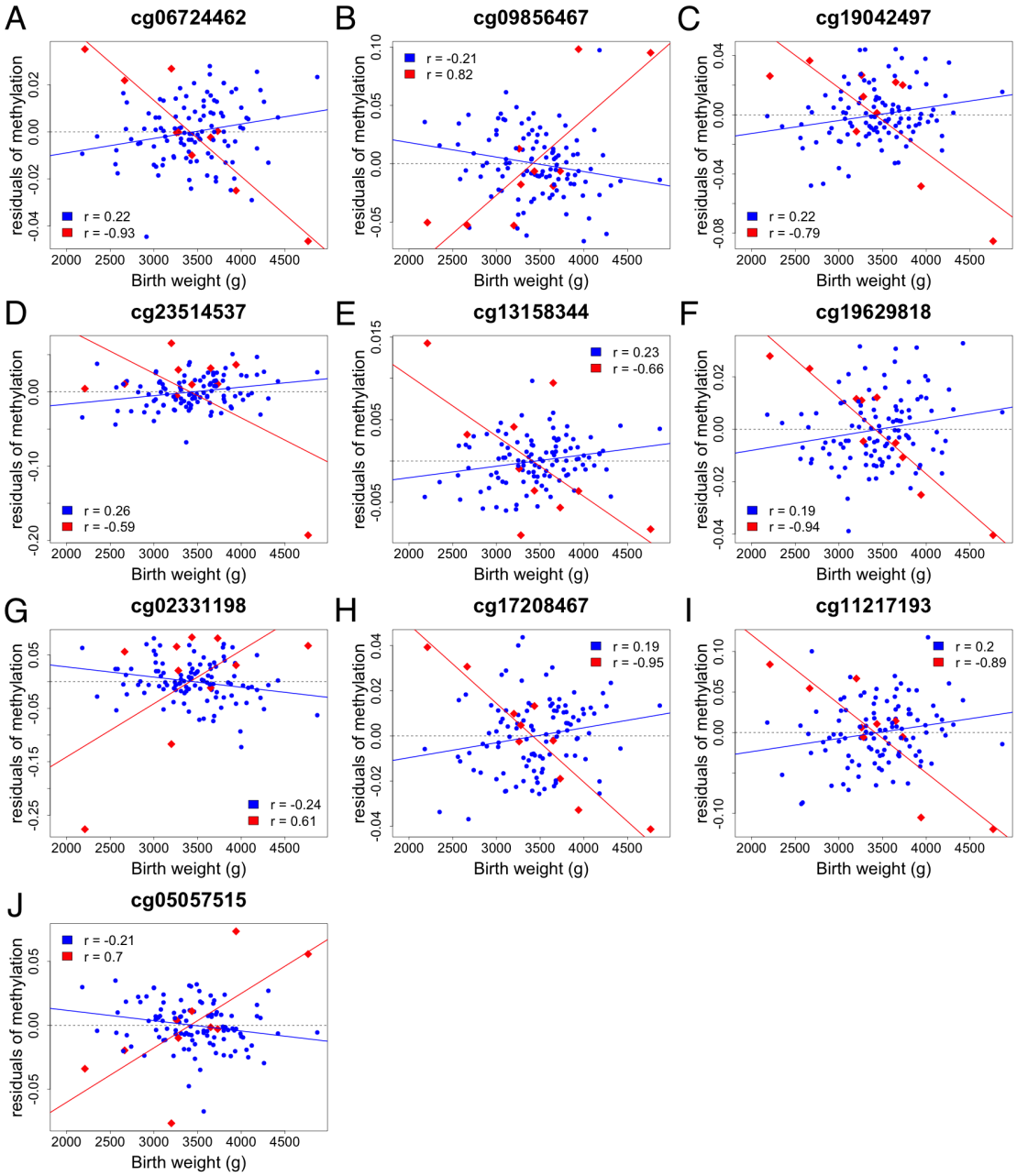

Figure 3.13: Scatterplot showing the directional association between methylation level (residual) and birth weight for the ten statistically significant CpG sites. Blue dots represent children that were not exposed to smoking during pregnancy and red dots are samples that were exposed to maternal smoking. In general the correlation of blue dots is very low and the regression line shallow. In contrast to this, the correlation and regression line of red samples is much higher and steeper. Direction of associations are always opposite. The blue and red lines denote the linear regression lines for the respective sample group.

3.2.2 Birth Weight and Progression to T1D

The epigenome-wide association analysis regarding the interaction of birth weight and progression to T1D in later life did not result in any statistically significant findings, although CpG cg09488090 is very close to the threshold with $P_{FDR} \approx 0.053$. This site is located within a non-protein coding RNA and is in proximity of the polyamine modulated factor 1 binding protein 1 (PMFBP1). Further inspection of the association between methylation residuals and birth weight (figure 3.14) shows that the significance of this methylation site may be falsely induced by outlying samples (circled green, figure 3.14).

Table 3.7: CpG site with the highest association between diagnosed T1D, birth weight and methylation level. cg09488090 is very close to the significance threshold but not significant($P_{FDR} < 0.05$). It is located in proximity of the polyamine modulated factor 1 binding protein 1 (PMFBP1).

	Est.[a]	*P*-value	FDR[b]	Chr.[c]	Gene	TSS[d]
cg09488090	0.0001	1.13e-07	0.05344949	16	PMFBP1	PMFBP1

[a]estimate of linear regression [b]FDR adjusted *p*-value [c]chromosome [d]transcription start site

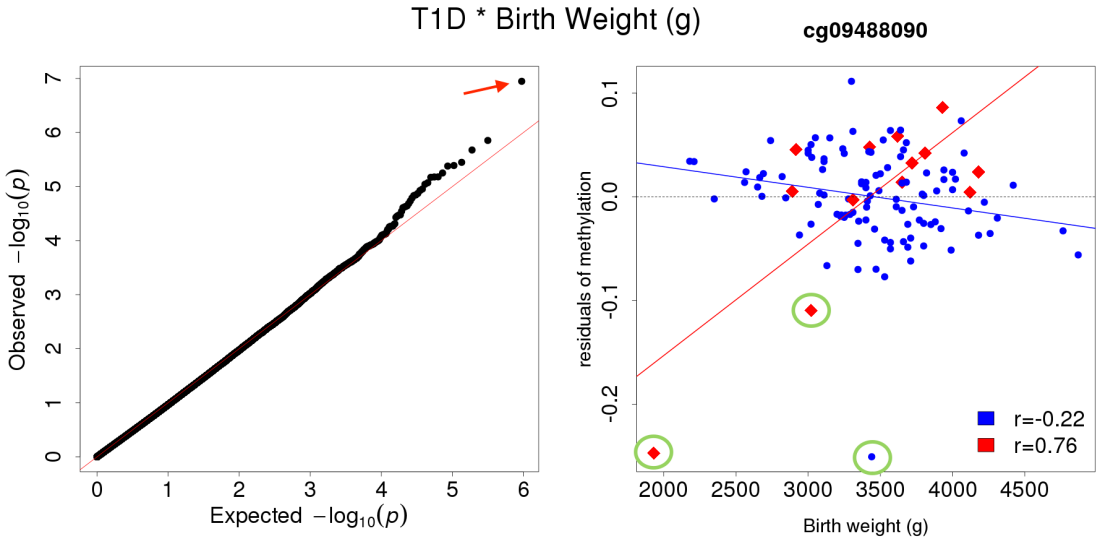

Figure 3.14: (A) Quantile-quantile plot of epigenome-wide association between methylation level and birth weight in dependence of T1D onset in later life. (B) Scatterplot showing the directional association between methylation level (residual) and birth weight for cg09488090. Blue dotes represent children that are currently healthy and the red dots children that have been diagnosed with T1D. Correlation for the blue dots is very low (r=22) and shows a slightly negative association. Red dots in contrast are higher correlated and display a strong positive association which is most likely influenced by outliers (circled green). The blue and red lines denote linear regression lines for the respective sample group.

4 Discussion

The following chapter will discuss the results presented in section 3 and elucidate potential limitations and drawbacks of this study. The UCSC Human Genome Browser [53] on *Human Feb. 2009 (GRCh37/hg19) Assembly* was used to infer further information on the identified cytosine – guanine dinucleotide (CpG) sites, utilizing information on CpG islands, transcription factor binding sites and measured methylation profiles in different cell types. In this context umbilical vein endothelial cells (HUVEC), embryonic stem cells (H1hESC) and GM12878, a lymphoblastoid cell line provided further insight into cell specific methylation patterns [53].

4.1 Effects of Environmental Factor on DNAm

The conducted epigenome wide association study of DNA methylation (DNAm) patterns in newborns of the BABYDIET cohort observed statistically significant associations between methylation levels and environmental factors. Although not all findings met the strict FDR significance threshold, further investigation often revealed meaningful biological context. This is for example illustrated by the multiple cytosine – guanine dinucleotides (CpGs) highly associated with fdr within the human leukocyte antigen (HLA) region on chromosome 6 (figure 3.3B). Various studies have shown that in this gene complex many type 1 diabetes (T1D) susceptibility and resistance loci are located [1]. Genetic inheritable susceptibility resides predominantly in HLA-DR and HLA-DQ genotypes which correspond to the identified CpG sites presented in table 3.2. These findings indicate that children with more than one first degree relative indeed have an increased genetic susceptibility for T1D. This is in union with previous findings [5] and supports the decision to control further regression analyses for familial history of T1D. Additionally, an overall higher distribution of lower p-values than expected for the association between methylation levels and gender of the child gave evidence of sex-specific patterns. Consequently further analyses were also adjusted for gender. The CpG site cg03769704 ($P_{FDR} \approx 0.016$) , which is located in the in the promoter region of SLFN5 (Schlafen family) showed a statistically significant associa-

tion. However within the scope of this study, designated to find links between environmental factors and DNAm, this finding will not further be discussed.

Various studies have investigated the influence of different cell proportions in whole blood on epigenome-wide association studies [36, 37, 38]. Due to the fact that blood is a heterogeneous mixture of different cell types which each have unique DNAm patterns, it is essential to control for their concentrations [36]. These studies have shown that without consideration of the heterogeneity of cell populations, strong confounding effects can be observed [36]. Therefore cell concentrations were estimated with the method of Houseman *et al.* [8], which has already found application in epigenome-wide association studies of cord blood samples [28].

An increased birth weight is known to increase the risk of developing type 2 diabetes [57] and a weak link between birth weight and T1D susceptibility has also been suggested [58]. Analysis regarding associations between birth weight of children (in grams) and methylation levels in the BABYDIET cohort showed increased methylation at cg00400614 in INHBA-ASI. The nearest transcription start site is INHBA, which encodes the inhibin beta A subunit that has been identified as a critical modulator of somatic growth and survival in mice [59]. Epigenetic modifications may influence expression levels of this gene and have effects on fetal growth in humans. McMinn *et al.* studied the differential expression of genes in intrauterine growth restriction (IUGR)[60], which is a condition characterized by insufficient fetal growth and low birth weights [61]. The study found evidence of differential expression of INHBA between IUGR and non-IUGR placentae [60], supporting the assumption that altered gene expression through methylation of INHBA may play an important role in fetal growth. Additionally, linear regression found associations of higher birth weight and hypermethylation of CpG sites located in the NFYA gene on chromosome 6. These sites are part of a CpG cluster overlapping the transcription start site of the adenylate cyclase 10 pseudogene 1 (ADCY10P1). Furthermore, the most significant of the CpG sites is located within a transcription factor binding motif for YY1 (Yin and Yang 1). This multifunctional transcription factor YYI is involved in many biological processes such as embryogenesis, differentiation and replication [62] and can activate or inhibit transcription [62]. Animal studies in rats have identified SNPs in YY1 as potential candidates for increased diabetes susceptibility and suggested a possible role of YY1 in human type 1 diabetes [63]. Even though the findings of Klöting *et al.* do not correspond to the chromosomal region of cg03644281, a general role of YY1 in birth weight and T1D susceptibility may be plausible but will need further investigation and validation.

Maternal smoking at conception showed higher associations in DNAm than smoking throughout pregnancy. This is interesting because it would have been expected to be directly opposite due to the fact that the time window for methylation changes to occur during pregnancy is much bigger than the short window of conception. Although in this context it is important to regard the fact that the definition of conception is very unspecific as it will often not refer to the exact time point of conception (pregnancy may have been discovered a few weeks after conception). Tobi *et al.* found evidence suggesting that the early gestation period (week 1-10) is a critical time window during which the maternal uterine environment can influence methylation marks of the fetus [54]. This would provide a potential explanation for the observed differences, as 22 women were smoking before they knew they were pregnant and 12 of these stopped smoking during pregnancy. These women reduced their cigarette consumption resulting in an even lower exposure to nicotine and it has been suggested that not only the duration, but also the intensity of smoking has an effect on methylation of the epigenome *in utero* [55].

Despite different significance values, both analysis reported an increase of methylation in cg06864895 in children exposed to maternal smoking. CpG cg06864895 lies within a strong enhancer region in umbilical vein endothelial cells (HUVEC) [64] which maps to the SLC38A2 gene. Increased methylation at this site may inactivate or alter the enhancing abilities, leading to a decreased expression of the gene. As the SLC38A2 gene, together with IGF, is involved in regulation of placental nutrient transfer and fetal nutrient acquisition [65], an altered expression may have effects on fetal development and growth. This may provide an epigenetic mechanism linking reduced birth weight in children with maternal smoking during pregnancy, which was observable in the BABYDIET cohort (figure 3.12) [56]. Smoking at conception was further associated with a reduction in methylation level in ACPP, a prostatic acid phosphatase at cg05409131 ($P_{FDR} \approx 0.15$) and an increase in methylation in the zink finger domain ZCCHC14 at cg02762752 ($P_{FDR} \approx 0.29$). Single nucleotide polymorphisms (SNPs) in ZCCHC14 have previously been associated with nicotine dependence [66]. Smoking during pregnancy was overall associated with less significance, however the higher methylation of the PAX9 CpG cg21207665 ($P_{FDR} \approx 0.43$) in children that were exposed to smoking during pregnancy, shows influence of two outliers (figure 3.9) that might be masking significance. PAX9 encodes a gene of the paired box (PAX) family of transcription factors, which play critical roles in fetal development.

The CpG cg15293181 is significantly associated ($P_{FDR} \approx 0.023$) with lower methylation levels in children with future T1D onset and located in an intron of the syntrophin gamma 2 (SNTG) gene. It is methylated in multiple cells

lines (HUVEC, H1heSC, GM12878) and the decreased amount of methylation may induce transcriptional changes if regulatory elements are affected. These changes may control the transcription of TPO which flanks SNTG. TPO encodes thyroid peroxidase , an enzyme that plays a major role in the biosynthesis of thyroid hormones [67]. Studies have shown that the risk of thyroid dysfunction is two to threefold higher in patients with T1D than in the general population, especially in those with prevalence of positive TPO antibodies [68]. This observation could confirm an association between autoimmune thyroid dysfunction and type 1 diabetes. The hypermethylation of cg15293181 in children who develop T1D later in life may mirror this relationship and if true, present a*priori* indication for increased susceptibility to thyroid diseases through high genetic predisposition for T1D. But this is very speculative and will need further analysis in order to allow a profound assumption.

Two other CpGs showed increased but not significant association with T1D. Hypomethylation of CpG site cg10563643, located in proximity of chromogranin A (CHGA) and hypermethylation of a CpG island upstream of the zinc finger protein 470 (ZNF470) at cg03153658 was observed for children who developed T1D compared to healthy children. ZNF470 is associated with transcriptional regulation and transcription factor activity. It is normally unmethylated in cell lines such as HUVEC, H1hESC and GM12878 and has been identified as an active promoter region in HUVEC cells [53]. Methylation of promoter regions can lead to inhibition or alterations of gene expression providing a plausible mechanism for an epigenetic effect on T1D susceptibility. Furthermore CHGA has been identified as a novel autoantigene in T1D [**G**, 69], which may be influenced in reactivity and association to disease by post-translational modifications [70]. It is notable that CHGA also showed hypomethylation in females compared to males (in the analysis of the association between gender and methylation), though not at the same CpG site (cg18397726).

Analysis of the association between maternal smoking during pregnancy and birthweight on methylation levels produced the most significant results.The effects on methylation levels were predominantly a decrease in methylation (seen in figure 3.13) in children with higher birth weight and exposure to maternal smoking. Four CpG sites showed the opposite effect but seemed largely influenced by outliers (figure 3.13B,D,G,J). cg06724462 and cg19629818 showed a very high correlation of methylation residuals, indicating a strong linear association of methylation levels and birth weight in children that were exposed to maternal smoking table 3.4. Both sites show methylation in multiple cells lines (HUVEC, H1hESC, GM12878) [53]. In children with exposure to smoking, methylation is increased for infants with low birth weight and strongly reduced for higher birth weights. Methylation levels in children without exposure show

no or very weak association with birth weight figure 3.13. A study conducted by Haworth *et al.* investigating the effects of maternal smoking on gene-specific cord blood methylation and birth weight percentile, found that higher methylation levels in APOB is associated with a higher risk of lower birth weight percentiles. The CpG cg19629818 (P_{FDR} ≈0.046) is located within the glypican 2 (GPC2) encoding gene which is reported to interact with APOB, thus potentially linking both to the same pathway that alters fetal methylation patterns through maternal smoking [71]. However, in this study we found no evidence of differential methylation in APOB, contradicting the a fore mentioned relationship. The highest association was reported for the cg06724462 (P_{FDR} ≈0.023) which is located in an intron region of a gene that codes for a potassium channel forming protein (KCNQ4).

Type 1 diabetes in interaction with birth weight was found to be highly (but not significantly) associated with the CpG site cg09488090 that is located in proximity of PMFBP1 which encodes for a polyamine modulated factor 1 binding protein 1 (P_{FDR} ≈0.053). However, the scatterplot of the residuals of methylation depicted in figure 3.14 shows highly influencing outliers which is why this site is not further discussed.

4.2 Limitations and Drawbacks

Even though this study presents insights into DNAm *in utero*, it was not able to reproduce findings of previous studies. EWAS have been conducted finding significant associations between DNAm *in utero* and birth weight [31] as well as maternal smoking [30, 55]. Reasons for discrepancies in results can have multiple causes. Firstly, and perhaps the greatest limitation of this study, the BABYDIET cohort is comparatively small with only 123. For example Joubert *et al.* [30] studied the effects of maternal smoking during pregnancy in 1062 cord blood samples. It is also important to consider that this cohort was assembled regarding familial T1D risk and therefore automatically represents a biased genetic background. As the participating mothers were not chosen in respect of for example smoking habits, some environmental factors are not equally distributed or do not have sufficient occurrences and therefore cannot ensure significant and meaningful results (analysis of dose-dependent effects of smoking). Secondly, many studies adjusted their regression model differently or used information that was not available for the children of this cohort (such as BMI, maternal age). In addition to these limitations in comparability, the question arises if a significance threshold of P_{FDR}< 0,05 may be to strict, as seen in the context of first degree relative or maternal smoking (section 3). Another very impor-

tant aspect is that cell-specific methylation was accounted for by estimating cell component proportions using the method proposed by Houseman *et al.* which was developed and validated for adult whole blood [37]. The correction for cell proportions is very important for epigenome wide analyses of whole blood as differences in cell proportions can devoid the outcome of regression analysis or infer association between a factor and methylation that is solely attributable to differential cell composition [36]. However if the estimated and observed proportions do not coincide the results may also be falsified. This is why careful interpretation of the presented findings is crucial.

5 Summary and Outlook

Type 1 Diabetes (T1D) is an autoimmune disease caused by the destruction of insulin-producing β-cells in the islets of Langerhans by autoantibodies. Its prevalence in the population is rapidly increasing and the precise triggers and underlying mechanisms of disease onset are still not fully understood. Patients with a genetic predisposition do not necessarily develop T1D in later life as environmental factors play a crucial role in the initiation of β-cell autoimmunity. Epigenetic modifications are heritable changes that do not effect the DNA sequence. One of the most extensively investigated epigenetic marks is the methylation of cytosines in cytosine – guanine dinucleotides (CpGs). DNA methylation (DNAm) is an important mechanism in the control of gene expression and can be determining for disease susceptibility. Due to technological advances it is now possible to conduct large scale studies, investigating disease-associated DNAm patterns.

The aim of this thesis was to identify DNAm patterns in umbilical cord blood of children with a high familial risk of T1D in order to gain insight into developmental programming *in utero* and potential influence of environmental factors on susceptibility to T1D. This was achieved by incorporating different environmental factors and information on future events (seroconversion, T1D onset) into a multivariate regression model and analyzing umbilical cord blood samples of 123 children of the BABYDIET cohort.

Effects of gender and familial T1D history were assessed and subsequently included in the regression model to prevent confounding effects. The model was further adjusted with estimates of cell composition to account for differential DNAm patterns across different cell types. Initial analyses showed that methylation levels show great variability, making appropriate and flexible outlier detection a crucial part of data preprocessing. A z-score based cutoff procedure paired with principle component analysis was used to eliminate spurious samples and data points. Despite these efforts the potential influence of outlying methylation data could not be excluded entirely.

The results of the regression analysis showed statistical significant associations between methylation levels and gender of child, future T1D progression and the interaction between maternal smoking and birth weight. Although

many findings did not meet the strict FDR significance threshold, meaningful biological context could be inferred for the effect of fdr, birth weight and maternal smoking on methylation. The findings discussed in this thesis support the assumption, that methylation marks are heritable and can transfer disease susceptibility. For example differential methylation was detected in T1D associated HLA genotypes for children with more than one affected first degree relative. Further more, evidence for *in utero* fetal programming is presented and potential mechanisms of epigenetic influence on disease susceptibility proposed.

Although the results of this thesis present novel insight on the effect of environmental factors on methylation patterns *in utero*, there are many aspects that will need further exploration. For example interactions between environmental factors that have not been analyzed in this study may play an important, yet unidentified role in disease onset. This is especially problematic due to the fact that we are exposed to a large variety of perceived and unperceived environmental factors and to identify and model every combination is an impossible task. Therefore the study of umbilical cord blood paired with close observations of maternal exposures seems to be a very promising approach for determining fundamental epigenetic mechanisms. Furthermore, more extensive research and stricter studies exploring the dose-dependent association between intake of dietary supplements and methylation patterns of the fetal epigenome may help elucidate the process and extent of developmental programming *in utero*.

The identified methylation patterns of key CpG sites throughout the genome are yet to be verified by supporting evidence from animal models or independent reproduction with a different cohort. However despite the limitations, this thesis presents a comprehensive analysis of the fetal epigenome in children with a familial risk of T1D. Findings were presented giving evidence of inheritable methylation patterns as well as novel methylation sites that may help elucidate the effects of environmental factors on epigenetic marks and disease susceptibility.

Bibliography

[1] Jeffrey A Bluestone, Kevan Herold, and George Eisenbarth. "Genetics, pathogenesis and clinical interventions in type 1 diabetes". In: *Nature* 464.7293 (Apr. 2010), pp. 1293–1300. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/nature08933.

[2] Denis Daneman. "Type 1 diabetes". In: *Lancet* 367.9513 (Mar. 2006), pp. 847–858. ISSN: 0140-6736. DOI: 10.1016/S0140-6736(06)68341-4.

[3] American Diabetes Association. "Diagnosis and classification of diabetes mellitus". In: *Diabetes Care* 29 Suppl 1 (Jan. 2006), S43–8. ISSN: 0149-5992.

[4] Alexandra E Butler et al. "Beta-cell deficit and increased beta-cell apoptosis in humans with type 2 diabetes". In: *Diabetes* 52.1 (Jan. 2003), pp. 102–110. ISSN: 0012-1797. DOI: 10.2337/diabetes.52.1.102.

[5] Peter Achenbach et al. "Natural history of type 1 diabetes". In: *Diabetes* 54 Suppl 2 (Dec. 2005), S25–31. ISSN: 0012-1797.

[6] Janet M Wenzlau et al. "The cation efflux transporter ZnT8 (Slc30A8) is a major autoantigen in human type 1 diabetes". In: *Proc. Natl. Acad. Sci. U. S. A.* 104.43 (Oct. 2007), pp. 17040–17045. ISSN: 0027-8424. DOI: 10.1073/pnas.0705894104.

[7] Anette G Ziegler et al. "Seroconversion to multiple islet autoantibodies and risk of progression to diabetes in children". In: *JAMA* 309.23 (June 2013), pp. 2473–2479. ISSN: 0098-7484, 1538-3598. DOI: 10.1001/jama.2013.6285.

[8] Alia Hasham and Yaron Tomer. "The recent rise in the frequency of type 1 diabetes: who pulled the trigger?" In: *J. Autoimmun.* 37.1 (Aug. 2011), pp. 1–2. ISSN: 0896-8411, 1095-9157. DOI: 10.1016/j.jaut.2011.04.001.

[9] Randy L Jirtle and Michael K Skinner. "Environmental epigenomics and disease susceptibility". In: *Nat. Rev. Genet.* 8.4 (Apr. 2007), pp. 253–262. ISSN: 1471-0056. DOI: 10.1038/nrg2045.

[10] Vardhman K Rakyan et al. "Epigenome-wide association studies for common human diseases". In: *Nat. Rev. Genet.* 12.8 (Aug. 2011), pp. 529–541. ISSN: 1471-0056, 1471-0064. DOI: 10.1038/nrg3000.

[11] Bradley E Bernstein, Alexander Meissner, and Eric S Lander. "The mammalian epigenome". In: *Cell* 128.4 (Feb. 2007), pp. 669–681. ISSN: 0092-8674. DOI: 10.1016/j.cell.2007.01.033.

[12] Jeremy J Day and J David Sweatt. "DNA methylation and memory formation". In: *Nat. Neurosci.* 13.11 (Nov. 2010), pp. 1319–1323. ISSN: 1097-6256, 1546-1726. DOI: 10.1038/nn.2666.

[13] Serge Saxonov, Paul Berg, and Douglas L Brutlag. "A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters". In: *Proc. Natl. Acad. Sci. U. S. A.* 103.5 (Jan. 2006), pp. 1412–1417. ISSN: 0027-8424. DOI: 10.1073/pnas.0510310103.

[14] M Gardiner-Garden and M Frommer. "CpG islands in vertebrate genomes". In: *J. Mol. Biol.* 196.2 (July 1987), pp. 261–282. ISSN: 0022-2836.

[15] I P Ioshikhes and M Q Zhang. "Large-scale human promoter mapping using CpG islands". In: *Nat. Genet.* 26.1 (Sept. 2000), pp. 61–63. ISSN: 1061-4036. DOI: 10.1038/79189.

[16] Robert J Klose and Adrian P Bird. "Genomic DNA methylation: the mark and its mediators". In: *Trends Biochem. Sci.* 31.2 (Feb. 2006), pp. 89–97. ISSN: 0968-0004. DOI: 10.1016/j.tibs.2005.12.008.

[17] Jean-Pierre Issa. "CpG island methylator phenotype in cancer". In: *Nat. Rev. Cancer* 4.12 (Dec. 2004), pp. 988–993. ISSN: 1474-175X. DOI: 10.1038/nrc1507.

[18] Wolf Reik. "Stability and flexibility of epigenetic gene regulation in mammalian development". In: *Nature* 447.7143 (May 2007), pp. 425–432. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/nature05918.

[19] Stefanie Seisenberger et al. "Reprogramming DNA methylation in the mammalian life cycle: building and breaking epigenetic barriers". In: *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 368.1609 (Jan. 2013), p. 20110330. ISSN: 0962-8436, 1471-2970. DOI: 10.1098/rstb.2011.0330.

[20] Marta Kulis and Manel Esteller. "DNA methylation and cancer". In: *Adv. Genet.* 70 (2010), pp. 27–56. ISSN: 0065-2660. DOI: 10.1016/B978-0-12-380866-0.60002-2.

[21] Robert A Waterland and Randy L Jirtle. "Transposable elements: targets for early nutritional effects on epigenetic gene regulation". In: *Mol. Cell. Biol.* 23.15 (Aug. 2003), pp. 5293–5300. ISSN: 0270-7306.

[22] Igor Koturbash et al. "Epigenetic dysregulation underlies radiation-induced transgenerational genome instability in vivo". In: *Int. J. Radiat. Oncol. Biol. Phys.* 66.2 (Oct. 2006), pp. 327–330. ISSN: 0360-3016. DOI: 10.1016/j.ijrobp.2006.06.012.

[23] G L Wolff et al. "Maternal epigenetics and methyl supplements affect agouti gene expression in Avy/a mice". In: *FASEB J.* 12.11 (Aug. 1998), pp. 949–957. ISSN: 0892-6638.

[24] Simon C Langley-Evans. "Developmental programming of health and disease". In: *Proc. Nutr. Soc.* 65.1 (Feb. 2006), pp. 97–105. ISSN: 0029-6651.

[25] Peter D Gluckman and Mark A Hanson. "The developmental origins of the metabolic syndrome". In: *Trends Endocrinol. Metab.* 15.4 (May 2004), pp. 183–187. ISSN: 1043-2760. DOI: 10.1016/j.tem.2004.03.002.

[26] Lucia A Hindorff et al. "Potential etiologic and functional implications of genome-wide association loci for human diseases and traits". In: *Proc. Natl. Acad. Sci. U. S. A.* 106.23 (June 2009), pp. 9362–9367. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.0903103106.

[27] David A Armstrong et al. "Global and gene-specific DNA methylation across multiple tissues in early infancy: implications for children's health research". In: *FASEB J.* 28.5 (May 2014), pp. 2088–2097. ISSN: 0892-6638, 1530-6860. DOI: 10.1096/fj.13-238402.

[28] Devin C Koestler et al. "Differential DNA methylation in umbilical cord blood of infants exposed to low levels of arsenic in utero". In: *Environ. Health Perspect.* 121.8 (Aug. 2013), pp. 971–977. ISSN: 0091-6765, 1552-9924. DOI: 10.1289/ehp.1205925.

[29] Y Ba et al. "Relationship of folate, vitamin B12 and methylation of insulin-like growth factor-II in maternal and cord blood". In: *Eur. J. Clin. Nutr.* 65.4 (Apr. 2011), pp. 480–485. ISSN: 0954-3007, 1476-5640. DOI: 10.1038/ejcn.2010.294.

[30] Bonnie R Joubert et al. "450K epigenome-wide scan identifies differential DNA methylation in newborns related to maternal smoking during pregnancy". In: *Environ. Health Perspect.* 120.10 (Oct. 2012), pp. 1425–1431. ISSN: 0091-6765, 1552-9924. DOI: 10.1289/ehp.1205412.

[31] Kim E Haworth et al. "Combined influence of gene-specific cord blood methylation and maternal smoking habit on birth weight". In: *Epigenomics* 5.1 (Feb. 2013), pp. 37–49. ISSN: 1750-1911, 1750-192X. DOI: 10.2217/epi.12.72.

[32] S Schmid et al. "BABYDIET, a feasibility study to prevent the appearance of islet autoantibodies in relatives of patients with Type 1 diabetes by delaying exposure to gluten". In: *Diabetologia* 47.6 (June 2004), pp. 1130–1131. ISSN: 0012-186X. DOI: 10.1007/s00125-004-1420-9.

[33] Sandra Hummel et al. "Primary dietary intervention study to reduce the risk of islet autoimmunity in children at increased risk for type 1 diabetes: the BABYDIET study". In: *Diabetes Care* 34.6 (June 2011), pp. 1301–1305. ISSN: 0149-5992, 1935-5548. DOI: 10.2337/dc10-2456.

[34] Simone Wahl et al. "On the potential of models for location and scale for genome-wide DNA methylation data". In: *BMC Bioinformatics* 15 (July 2014), p. 232. ISSN: 1471-2105. DOI: 10.1186/1471-2105-15-232.

[35] Ruth Pidsley et al. "A data-driven approach to preprocessing Illumina 450K methylation array data". In: *BMC Genomics* 14 (May 2013), p. 293. ISSN: 1471-2164. DOI: 10.1186/1471-2164-14-293.

[36] Andrew E Jaffe and Rafael A Irizarry. "Accounting for cellular heterogeneity is critical in epigenome-wide association studies". In: *Genome Biol.* 15.2 (Feb. 2014), R31. ISSN: 1465-6906. DOI: 10.1186/gb-2014-15-2-r31.

[37] Eugene Andres Houseman et al. "DNA methylation arrays as surrogate measures of cell mixture distribution". In: *BMC Bioinformatics* 13 (May 2012), p. 86. ISSN: 1471-2105. DOI: 10.1186/1471-2105-13-86.

[38] Karin B Michels et al. "Recommendations for the design and analysis of epigenome-wide association studies". In: *Nat. Methods* 10.10 (Oct. 2013), pp. 949–955. ISSN: 1548-7091, 1548-7105. DOI: 10.1038/nmeth.2632.

[39] Denis Cousineau and Sylvain Chartier. "Outliers detection and treatment: a review". In: *International Journal of Psychological Research* 3.1 (June 2015), pp. 58–67. ISSN: 2011-7922, 2011-7922.

[40] Erwin Kreyszig. "Applied mathematics". In: *Hoboken, NJ: John Wiley & Sons* (1979), p. 880.

[41] Ian Jolliffe. *Principal component analysis*. Wiley Online Library, p. 1.

[42] Howard H Yang et al. "Influence of genetic background and tissue types on global DNA methylation patterns". In: *PLoS One* 5.2 (Feb. 2010), e9355. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0009355.

[43] Daniel Wollschläger. *Grundlagen der Datenanalyse mit R: eine anwendungsorientierte Einführung*. Springer-Verlag, 2015.

[44] David J Balding. "A tutorial on statistical methods for population association studies". In: *Nat. Rev. Genet.* 7.10 (Oct. 2006), pp. 781–791. ISSN: 1471-0056. DOI: 10.1038/nrg1916.

[45] Arend Voorman et al. "Behavior of QQ-plots and genomic control in studies of gene-environment interaction". In: *PLoS One* 6.5 (2011), e19416. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0019416.

[46] Marco P Boks et al. "The relationship of DNA methylation with age, gender and genotype in twins and healthy controls". In: *PLoS One* 4.8 (2009), e6767. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0006767.

[47] Paul I W de Bakker et al. "Practical aspects of imputation-driven meta-analysis of genome-wide association studies". In: *Hum. Mol. Genet.* 17.R2 (2008), R122–8. ISSN: 0964-6906, 1460-2083. DOI: 10.1093/hmg/ddn288.

[48] William S Noble. "How does multiple testing correction work?" In: *Nat. Biotechnol.* 27.12 (Dec. 2009), pp. 1135–1137. ISSN: 1087-0156, 1546-1696. DOI: 10.1038/nbt1209-1135.

[49] Yoav Benjamini and Yosef Hochberg. "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing". In: *J. R. Stat. Soc. Series B Stat. Methodol.* 57.1 (1995), pp. 289–300. ISSN: 1369-7412, 0035-9246.

[50] Anat Reiner, Daniel Yekutieli, and Yoav Benjamini. "Identifying differentially expressed genes using false discovery rate controlling procedures". In: *Bioinformatics* 19.3 (Feb. 2003), pp. 368–375. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btf877.

[51] LLC iChemLabs. *ChemDoodle: Chemical and Biological Publishing Software*. 2015.

[52] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2015.

[53] W James Kent et al. "The human genome browser at UCSC". In: *Genome Res.* 12.6 (June 2002), pp. 996–1006. ISSN: 1088-9051. DOI: 10.1101/gr.229102.Articlepublishedonlinebeforeprintin May 2002.

[54] Elmar W Tobi et al. "Early gestation as the critical time-window for changes in the prenatal environment to affect the adult human blood methylome". In: *Int. J. Epidemiol.* 44.4 (Aug. 2015), pp. 1211–1223. ISSN: 0300-5771, 1464-3685. DOI: 10.1093/ije/dyv043.

[55] Rebecca C Richmond et al. "Prenatal exposure to maternal smoking and offspring DNA methylation across the lifecourse: findings from the Avon Longitudinal Study of Parents and Children (ALSPAC)". In: *Hum. Mol. Genet.* 24.8 (2015), pp. 2201–2217. ISSN: 0964-6906, 1460-2083. DOI: 10.1093/hmg/ddu739.

[56] Xiaobin Wang et al. "Maternal cigarette smoking, metabolic gene polymorphism, and infant birth weight". In: *JAMA* 287.2 (2002), pp. 195–202. ISSN: 0098-7484. DOI: 10.1001/jama.287.2.195.

[57] Charlotte M Boney et al. "Metabolic syndrome in childhood: association with birth weight, maternal obesity, and gestational diabetes mellitus". In: *Pediatrics* 115.3 (Mar. 2005), e290–6. ISSN: 0031-4005, 1098-4275. DOI: 10.1542/peds.2004-1808.

[58] L C Stene et al. "Birth weight and childhood onset type 1 diabetes: population based cohort study". In: *BMJ* 322.7291 (2001), pp. 889–892. ISSN: 0959-8138, 0959-535X.

[59] Chester W Brown et al. "Activins are critical modulators of growth and survival". In: *Mol. Endocrinol.* 17.12 (Dec. 2003), pp. 2404–2417. ISSN: 0888-8809. DOI: 10.1210/me.2003-0051.

[60] J McMinn et al. "Unbalanced placental expression of imprinted genes in human intrauterine growth restriction". In: *Placenta* 27.6-7 (June 2006), pp. 540–549. ISSN: 0143-4004. DOI: 10.1016/j.placenta.2005.07.004.

[61] Ira M Bernstein et al. "Morbidity and mortality among very-low-birth-weight neonates with intrauterine growth restriction". In: *Am. J. Obstet. Gynecol.* 182.1 (2000), pp. 198–206. ISSN: 0002-9378. DOI: 10.1016/S0002-9378(00)70513-8.

[62] S Gordon et al. "Transcription factor YY1: structure, function, and therapeutic implications in cancer biology". In: *Oncogene* 25.8 (2006), pp. 1125–1142. ISSN: 0950-9232. DOI: 10.1038/sj.onc.1209080.

[63] Nora Klöting and Ingrid Klöting. "Genetic variation in the multifunctional transcription factor Yy1 and type 1 diabetes mellitus in the BB rat". In: *Mol. Genet. Metab.* 82.3 (July 2004), pp. 255–259. ISSN: 1096-7192. DOI: 10.1016/j.ymgme.2004.04.007.

[64] Jason Ernst et al. "Mapping and analysis of chromatin state dynamics in nine human cell types". In: *Nature* 473.7345 (May 2011), pp. 43–49. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/nature09906.

[65] Emily Angiolini et al. "Developmental adaptations to increased fetal nutrient demand in mouse genetic models of Igf2-mediated overgrowth". In: *FASEB J.* 25.5 (May 2011), pp. 1737–1745. ISSN: 0892-6638, 1530-6860. DOI: 10.1096/fj.10-175273.

[66] Ke-Sheng Wang et al. "ANAPC1 and SLCO3A1 are associated with nicotine dependence: meta-analysis of genome-wide association studies". In: *Drug Alcohol Depend.* 124.3 (2012), pp. 325–332. ISSN: 0376-8716, 1879-0046. DOI: 10.1016/j.drugalcdep.2012.02.003.

[67] B Czarnocka et al. "Purification of the human thyroid peroxidase and its identification as the microsomal antigen involved in autoimmune thyroid diseases". In: *FEBS Lett.* 190.1 (1985), pp. 147–152. ISSN: 0014-5793. DOI: 10.1016/0014-5793(85)80446-4.

[68] Guillermo E Umpierrez et al. "Thyroid dysfunction in patients with type 1 diabetes: a longitudinal study". In: *Diabetes Care* 26.4 (Apr. 2003), pp. 1181–1185. ISSN: 0149-5992. DOI: 10.2337/diacare.26.4.1181.

[69] Shuhong Han et al. "Novel autoantigens in type 1 diabetes". In: *Am. J. Transl. Res.* 5.4 (2013), pp. 379–392. ISSN: 1943-8141.

[70] Peter A Gottlieb et al. "Chromogranin A is a T cell antigen in human type 1 diabetes". In: *J. Autoimmun.* 50 (May 2014), pp. 38–41. ISSN: 0896-8411, 1095-9157. DOI: 10.1016/j.jaut.2013.10.003.

[71] Lars J Jensen et al. "STRING 8–a global view on proteins and their functional interactions in 630 organisms". In: *Nucleic Acids Res.* 37.Database issue (Jan. 2009), pp. D412–6. ISSN: 0305-1048, 1362-4962. DOI: 10.1093/nar/gkn760.