Investigation of DNA methylation in type 2 diabetes genetic risk loci

Björn Þór Aðalsteinsson

Thesis for the degree of Master of Science
University of Iceland
School of Health Sciences
Faculty of Medicine

HÁSKÓLI ÍSLANDS

Investigation of DNA methylation in type 2 diabetes genetic risk loci

Björn Þór Aðalsteinsson

Thesis for the degree of Master of Science

Master's degree committee:

Vilmundur Guðnason (supervisor)

Albert Vernon Smith (instructor)

Haukur Guðnason

Faculty of Medicine

School of Health Sciences

University of Iceland

April 2012

Rannsókn á DNA metýlun í erfðasætum tengdum áhættu á sykursýki af týpu 2

Björn Þór Aðalsteinsson

Ritgerð til meistaragráðu

Meistaranámsnefnd:

Vilmundur Guðnason (umsjónarkennari)

Albert Vernon Smith (leiðbeinandi)

Haukur Guðnason

Læknadeild

Heilbrigðisvísindasvið

Háskóli Íslands

Apríl 2012

Printing office: Háskólaprent

Reykjavík, Iceland, 2012

Abstract

Epigenetic studies are commonly conducted on DNA from tissue samples. However, tissues are ensembles of cells that may each have their own epigenetic profile and therefore inter-individual difference in cellular heterogeneity may compromise these studies. In work presented here, the potential for such confounding on DNA methylation measurement outcomes when using DNA from whole blood was explored. DNA methylation was measured using pyrosequencing based methodology in two white blood cell fractions, isolated using density gradient centrifugation. In three out of four regions tested, significant differential DNA methylation between the two fractions was detected. The difference was very moderate in all but one region where the average absolute methylation difference per CpG site ranged between 3.4-15.7 percentage points. In this same region, inter-individual variation in cellular heterogeneity explained up to 36% ($p<0.0001$) of the variation in measured whole blood DNA methylation levels. In the examined regions, methylation levels were highly correlated between cell fractions. In summary, the analysis detects region-specific differential DNA methylation between white blood cell sub-types, which can confound the outcome of whole blood DNA methylation measurements. Finally, by demonstrating the high correlation between methylation levels in cell fractions, the results suggest a possibility to use a proportional number of a single white blood cell type to correct for this confounding effect in analyses.

Type 2 diabetes mellitus (T2DM) is a complex disease (i.e., multifactorial and polygenic) characterized by high blood glucose levels due to reduced insulin sensitivity and β-cell function. Heritable as well as lifestyle and environmental factors contribute to risk of development of the disease. Despite recent advances in identifying T2DM genetic risk variants, a large proportion of the disease's heritable component remains unidentified. One potential explanation is the existence of inherited epigenetic aberration(s) that contribute to the disease. Additionally, it is possible that the effects of environmental and lifestyle factors are mediated through induced epigenetic aberration(s). On this basis, it was hypothesized that aberrant DNAm could be associated with T2DM, and to address this hypothesis, the second study presented here aimed to identify such aberrations. DNA methylation levels were measured using pyrosequencing based methodology in whole blood DNA. DNA methylation levels were compared between individuals with and without T2DM in seven regions, located in three loci previously associated with T2DM through genetic studies. A single region located in an intragenic CpG island in the *HHEX* gene was selected for further study, and comparing DNAm levels between 214 cases and 164 controls, lower average DNAm levels were observed in individuals with T2DM. Further, the difference was significant after correction for cellular heterogeneity, age and gender, and was not carried by an association with obesity. These results support the hypothesis that DNAm aberrations may be associated with T2DM.

Ágrip

Utangenaerfðarannsóknir eru oft á tíðum framkvæmdar með erfðaefni úr vefjasýnum. Vefir eru hins vegar samsettir úr fjölda frumutegunda sem hver um sig hefur mögulega einkennandi utangenaerðamerki og þar af leiðandi getur breytileyki í hlutfallslegum frumufjölda milli einstaklinga gruggað (e.confound) þessar rannsóknir. Í einni af tveimur rannsóknum sem hér verða kynntar verður möguleikinn á slíkri gruggun rannsakaður m.t.t. rannsókna á DNA metýlun sem framkvæmdar eru með DNA úr heilblóði. DNA metýlun var mæld með aðferð sem byggir á raðgreiningu á bísúlfíð umbreyttu DNA í tveimur heilblóðs frumuhlutum sem einangraðir voru með aðferð sem byggir á skilvindun blóðsins. Marktækur munur á DNA metýlun milli frumuhlutanna greindist í þremur af þeim fjórum svæðum sem rannsökuð voru. Munurinn var mjög lítill í tveimur svæðanna, en í einu þeirra mældist munur í meðaltals DNA metýlun hlutanna 3.4-15.7 prósentustig per CpG set. Í þessu sama svæði skýrði breytileiki í hlutfallslegum frumufjölda milli einstaklinga allt að 36% ($p<0.0001$) af breytileika í mældri DNA metýlun í heilblóði. Í öllum svæðanna var fylgni milli mældrar DNA metýlunar mjög há milli frumuhlutanna tveggja. Niðurstöðurnar benda til svæðis-sértæks munar í DNA metýlun mismunandi hvítra frumutegunda sem gruggað geta niðurstöður mælinga á DNA metýlun í heilblóði. Með því að sýna fram á fylgni milli DNA metýlunar í hinum mismunandi frumuhlutum gefa niðurstöðurnar til kynna að leiðrétta megi fyrir þessari gruggun með notkun hlutfallslegs fjölda einnar frumutegundar hvítra blóðkorna við greiningu gagna á DNA metýlun í heilblóði.

Sykursýki af týpu 2 (SST2) er sjúkdómur sem einkennist af háum blóðsykurstyrk vegna skerts insúlínnæmis og skertrar virkni β frumna. Lífstíls, umhverfis og erfðaþættir valda áhættu á SST2. Þrátt fyrir miklar framfarir á síðustu árum í að finna erfðabreytileika tengda SST2, þá er stór hluti arfbundna þáttar áhættu á SST2 enn óþekktur. Ein möguleg skýring á því felst í því að afbrigðileg arfgeng utangenaerfðamörk tengist sjúkdómnum. Að auki er mögulegt að umhverfis og lífstílsþættir miðli áhrifum sínum gegnum utangenaerfðaþætti. Á þessum rökum var sú tilgáta sett fram að afbrigðileg utangenaerfðamörk gætu tengst sjúkdómnum og markmið seinni rannsóknarinnar sem hér verður kynnt var að prófa tilgátuna með því að leita slíkra afbrigðileika. DNA metýlun var mæld með sömu aðferð og áður í erfðaefni úr heilblóði. DNA metýlun í einstaklingum með og án SST2 var borin saman í sjö svæðum sem staðsett eru í þremur erfðasætum sem áður hafa verið tengd SST2 með erfðafræðirannsóknum. Eitt svæðanna, sem staðsett er í CpG eyju innan *HHEX* gensins, var valið til frekari skoðunar. Lægri meðaltals DNA metýlun greindist í einstaklingum með SST2 í þessu svæði þegar 214 sykursjúkir og 164 einstaklingar án sjúkdómsins voru bornir saman. Munurinn var marktækur eftir leiðréttingu fyrir kyni, aldri, líkamsþyngdarstuðli og breytileyka í hlutfallslegum frumufjölda. Niðurstöðurnar styðja tilgátuna sem sett var fram, að afbrigðileg DNA metýlun tengist sykursýki af týpu 2.

Acknowledgements

First, I would like to thank Vilmundur Guðnason (supervisor) and Albert Vernon Smith (instructor) for their excellent supervision. Secondly, I thank Haukur Guðnason (masters committee member) for his assistance and discussions, especially about laboratory matters.

The work presented here was conducted at the Icelandic Heart Association. The Association's entire staff, past and present, has contributed to the project directly or indirectly and therefore deserves my gratitude. I would like to thank a few of them specifically: Thor Aspelund for his guidance on the statistical aspect of the project, Guðný Eiríksdóttir for discussions and assistance on a variety of matters, Alda Hauksdóttir for assistance on laboratory matters, and finally, Sigrún Halldórsdóttir and Friðrik Þórðarson for help with database matters. Finally, I thank my friend and fellow student, Valborg Guðmundsdóttir for her companionship and discussions.

The participants of cohort studies conducted at the Icelandic Heart Association made this study possible and I am indebted to them for their contribution. Finally, I would like to thank those who funded the present study and The Age, Gene/Environment Susceptibility Reykjavik and the Risk Evaluation For Infarct Estimates Reykjavik studies: Rannsóknarmiðstöð Íslands (Markáætlun um erfðafræði og örtækni), the Icelandic Heart Association, the Icelandic Parliament and the National Institute on Ageing, the National Institutes of Health.

Table of contents

List of figures

List of tables

List of frequently used abbreviations

AGES Reykjavik study – Age Gene/Environment Susceptibility Reykjavik study

AUC – Area under receiver-operating characteristic curve

BMI – Body mass index

cat.nr – catalog number

CD – Cluster of differentiation

CGI – CpG island

CpG – C phosphate G (i.e., DNA sequence 5'-CpG-3')

DNA – Deoxyribonucleic acid

DNAm – DNA methylation

EWAS – Epigenome wide association study

GWAS – Genome wide association study

HHEX – Homeobox, hematopoietically expressed

HOMA – Homeostasis model assessment (-IR: Insulin resistance, and -β: Beta cell function)

HPLC – High performance liquid chromatography

IFG – Impaired fasting glucose

IHA – Icelandic Heart Association

kb – Kilobase (1.000 bases)

KCNJ11 – Potassium inwardly rectifying channel, subfamily J, member 11

KCNQ1 – Potassium voltage-gated channel, KCT-like subfamily, member 1

MNCs – Mononuclear cells (the white blood cells lymphocytes and monocytes)

NFG – Normal fasting glucose

PCR – Polymerase chain reaction

PMNCs – Polymorphonuclear cells (the white blood cell neutrophils, basophils and eosinophils)

PM20D1 – Peptidase M20 domain containing 1

pp – percentage points

REFINE Reykjavik study – Risk Evaluation For Infarct Estimates Reykjavik study

Rev. in ref. – reviewed in reference

RNA – Ribonucleic acid

TSS – Transcription start site

T2DM – Type 2 diabetes mellitus

%DNAm – Percent DNA methylation

1 Introduction

1.1 Epigenetics

Epigenetics refers to the heritable, but reversible, regulation of various genetic functions, including gene expression, mediated through modifications of DNA or chromatin (1,2). The most extensively studied epigenetic mark, described in detail below, is DNA methylation (DNAm). Histone modifications are a second well established epigenetic mark and will be discussed briefly.

1.1.1 DNA methylation, CpG sites, CpG islands and distribution

DNAm is a covalent addition of a methyl molecule on a cytosine base (specifically to position 5 of cytosine's pyrimidine ring) in DNA. It can be maintained through cell division, possibly between generations and can affect gene expression. Such marking is sequence dependent, in humans cytosines followed by guanine bases, i.e., 5'-CpG-3' (C phosphate G) sequences, termed CpG sites are the predominant target for methylation (3). Extensive methylation in the sequences CHG and CHH (H = A, C or T) has also been observed in embryonic stem cells (3). In all following text DNA methylation only refers to methylation in the CpG sequence context.

In the human genome, the number of CpG dinucleotides is underrepresented with respect to expected numbers. A frequently suggested cause is that spontaneous deamination of methylated cytosines results in formation of a thymine base, while the same event occurring on unmethylated cytosines results in the formation of a uracil. While the resulting uracil-guanine mismatch is easily detectable and correctly repaired, the thymine-guanine mismatch is frequently erroneously repaired to a thymine-adenine basepair which may have led to gradual loss of CpG sites over evolutionary time (reviewed in reference (rev. in ref.) (4)). A lack of correlation between TpG excess and CpG shortage (5) does however suggest that other mechanisms contribute to the CpG shortage. In certain regions of DNA, called CpG islands (CGI), the CpG dinucleotide occurs in close to its expected frequency (6).

In the human genome, approximately 70-80% of CpG sites are methylated (7,8). The pattern of DNAm in mammals with respect to genomic context is global, i.e., genes, intergenic regions, repetitive elements such as satellite DNA and transposons are generally methylated (6). The predominant exception is that CpG islands are generally not methylated; Illingworth *et al*. investigated DNAm in 14.318 CpG islands in four somatic human tissues and found that only 11.6% are methylated (9). As 56-72% of gene promoters are estimated to contain CpG islands (depending e.g., on definition of a CGI) (4,10), it follows that the 5' extremities of genes are an exception of the global methylation pattern found in the genome.

1.1.2 Establishment and erasure of DNA methylation and reprogramming

Three enzymes that catalyze DNAm have been identified, the DNA methyltransferases (DNMT) DNMT1, DNMT3A and DNMT3B. These enzymes catalyze *de novo* and/or maintenance DNAm, which is necessary to maintain normal methylation patterns through cell divisions and after DNA repair. DNMT1 has a stronger affinity for hemi-methylated DNA than completely unmethylated DNA (11), and is therefore commonly referred to as a maintenance methyltransferase, while DNMT3A and DNMT3B

are referred to as *de novo* methyltransferases (rev. in ref. (12,13)). Demethylation can occur passively or actively. Passive demethylation occurs during cell division when the DNAm marks are not maintained. The cellular processes involved in active demethylation are not well established, but results from a recent study suggest that they may involve oxidation of methylated cytosines by Tet proteins, resulting in formation of hydroxymethylated cytosine (14).

DNAm is globally erased and re-established during development, a phenomenon termed reprogramming. These events have been studied in considerable detail in mice, but sampling difficulties have prevented similar studies for humans. Following fertilization, a global demethylation occurs in the zygote, the paternally derived chromosomes are actively demethylated and the maternally derived chromosomes passively. After implantation, *de novo* methylation occurs and is maintained in most tissues of the embryo. An exception is primordial germ cells, which after embryonic day 7.25 undergo a global demethylation event. In male embryos, the germ-cell precursor undergoes a *de novo* methylation event which is completed before birth, while in females, this event occurs after birth (rev. in ref. (15)).

1.1.3 Tissue and cell specificity of DNA methylation

Tissue and cell specific methylation are well established in human DNA. In 2006 Eckhardt *et al*. presented data from the Human Epigenome Project (HEP, a project which aims to identify, catalog and interpret genome-wide DNAm profiles of all human genes in all major tissues) that suggest that tissue-specific differential methylation is very common in the genome (16). The dataset describes DNAm of CpG sites in 2524 sequenced amplicons on chromosomes 6, 20 and 22 in 12 different tissues. Differential DNAm between tissues was observed in approximately 22% of the investigated amplicons and the average absolute methylation levels differed by up to 20% (or up to 15% if only somatic tissues are compared). Recently, Fan and Zhang analyzed DNAm in selected (CpG site coverage > 30%) CGIs using the HEP dataset (17). Similarly, their results indicate that a substantial proportion of CGIs (~18%) are differentially methylated between tissues. Three recent independent studies using microarray based methods also identify differences in DNAm between tissues after interrogating CpG sites across the whole genome (18), in CGIs across the genome (9), and in non-CGI regions on chromosome 1 (19).

Relatively few studies have addressed the question whether different white blood cell types have specific DNAm levels. Common types of white blood cells include the granulocytes neutrophils basophils and eosinophils, collectively called polymorphonuclear cells (PMNCs) here, and the mononuclear cells (MNCs) lymphocytes and monocytes. All white blood cells originate from multipotent hemopoietic stem cells in the bone marrow that differentiate into a lymphoid progenitor which gives rise to lymphocytes and a myeloid progenitor which gives rise to the other four cell types. Basophils mediate inflammatory reactions by releasing histamine and lymphocytes produce antibodies, kill virus infected cells and regulate activities of other white blood cells. Monocytes and neutrophils both phagocytose invading bacteria while eosinophils participate in removing larger parasites (rev. in ref. (20)). In two papers from 1990 and 1991 Kochanek *et al*. reported results from a study investigating DNAm of the *TNFα* and *TNFβ* (tumor necrosis factor α and β respectively) genes in

multiple cancerous and non-cancerous white blood cells and cell lines (21,22). Their results revealed gross differences in *TNFβ* methylation in lymphocytes versus granulo- and monocytes as well as minor distinctions in the *TNFα* gene between cell types in the control samples. A comparison of DNAm levels in CD4+ and CD8+ (CD; cluster of differentiation) lymphocytes was included in the HEP report which showed that these highly developmentally related cell types exhibit on average ~5% absolute difference in DNAm (16). Finally, Wu *et al*. compared different methods and sources of DNA for measuring global DNAm (i.e., methylation of the genome as a whole) in whole blood (23). DNA derived from whole blood and two blood fractions, MNCs and PMNCs was measured using five global methylation assays that interrogate methylation at CpG sites located in different genome contexts, e.g., in different repetitive elements. In four of the five assays, global methylation levels in MNCs and PMNCs were not correlated, suggesting a widespread difference in DNAm between the two cell groups.

1.1.4 Histone modifications

Histone modification is a covalent addition of a chemical group to a core histone (the proteins which DNA is wrapped around in the eukaryotic cell nuclei forming nucleosomes). The chemical group may be of several types, e.g., histones can be methylated, acetylated or phosphorylated, some may be added singly or in multiple copies e.g., monomethylation, dimethylation and trimethylation, and they can be added on various locations on the histone protein tails, i.e., on different amino acid residues on the polypeptide tails that extend from the core of the nucleosome. The full histone code, i.e., the full repertoire of histone modifications and their effects on cell function, is therefore very complicated and the vast majority of them is poorly understood. Some histone modifications have nevertheless been relatively well characterized, especially lysine acetylation and methylation (i.e., a lysine residue on the histone tail is methylated or acetylated). Lysine acetylation almost always correlates with chromatin accessibility and transcriptional activity, but lysine methylation correlates with both transcriptional activity and repression, depending on which histone residue is modified. It is worth noting that not all histone modifications have been shown to be maintained through cell divisions, and therefore they may not all represent an epigenetic mark (rev. in ref. (24,25)).

1.1.5 Correlations between epigenetic marks

In the present study, DNAm was the only epigenetic mark under investigation. Although other epigenetic marks may be just as important for studies on normal and abnormal phenotypes their investigation was beyond the scope of this study. Patterns of different epigenetic marks have however been observed to be correlated both in mouse and human cells (3,26). Although not the purpose here, it is worth noting that profiling DNAm may therefore provide indirect information on other epigenetic marks, i.e., DNAm may potentially be used as surrogate of an epigenetic state (collective epigenetic marks in a particular region under investigation) (27).

1.1.6 Measuring DNA methylation

During early DNAm research, DNAm levels could only be investigated in a global manner, using HPLC (high-performance liquid chromatography) or other chromatography. The first method for detecting

DNAm at specific loci was to digest DNA using methylation sensitive restriction endonucleases followed by Southern blotting (28). The drawback of the method is that only CpG sites that occur in a restriction enzyme recognized sequence can be investigated, and in addition it requires large amounts of DNA. Analysis of DNAm was revolutionized in 1992 after Frommer *et al*. demonstrated the usefulness of bisulfite conversion of DNA for such analyses. Treating DNA with bisulfite converts unmethylated cytosines to uracil while methylated cytosines are unaffected (29). Following such conversion, a region of interest can be PCR (polymerase chain reaction) amplified and DNAm levels analyzed with multiple different techniques. Multiple methods have been developed on basis of this principle (rev. in ref. (28)), and either involve amplification with primers that bind sequences that contain no CpG sites and reveal the methylation status of CpGs in the amplicons in downstream analyses or involve amplification using primers that anneal to the CpG sites whose methylation status is to be investigated (methylation specific polymerase chain reaction, MSP). Downstream analyses for the former methodology include COBRA (combined bisulfite restriction analysis) where the products are digested and separated on gels to reveal the extent of methylation (30), and sequencing, e.g., using pyrosequencing.

Pyrosequencing is a sequencing by synthesis method which relies on detecting nucleotide incorporation by light emission (31). In short, its principle is to add dNTPs to a solution containing the DNA strand to be analyzed and a team of enzymes. If the dNTP is complementary to the DNA sequence, which is made single stranded and a sequencing primer annealed adjacent to a region of interest, it is incorporated and a pyrophosphate is released. In a series of enzymatic steps (involving sulfurylase and luciferase enzymes and luciferin substrate) the pyrophosphate causes light emission in amounts that are proportional to the numbers of dNTPs incorporated and thus the number of complementary bases in the analyzed DNA molecules. Pyrosequencing of amplified bisulfite treated DNA can be used to analyze DNAm levels in a single or a series of CpG sites in a highly reproducible and accurate quantitative manner (32,33). Its principle relies on that after amplification of bisulfite treated DNA, potential methylated positions can be treated as polymorphisms, and the allele frequencies determined to reveal their methylation status (e.g., a C/T polymorphism for C in a CpG sequence).

Studying DNAm epigenome wide (complete collection of all epigenetic marks present in a cell) has become a feasible option in the last few years, thus allowing for epigenome wide associations studies (EWAS or EWA study). Multiple technologies have emerged which enable such profiling, based on DNA microarray chip or sequencing based methods. In a recent article by Rakyan *et al*. (ref. (27)) where study designs for EWAS are discussed, an insight into which of these technologies are preferred by researchers conducting EWAS is provided. Although whole genome bisulfite sequencing is ideal because it provides the highest level of genome coverage and a single base resolution, it is currently considered too expensive. Due to rapidly falling costs, this method may however become feasible and prevail in the future. Enrichment methods, such as MeDIP (Methylated DNA Immunoprecipitation) can be employed followed by sequencing (or analysis on chip arrays) to cut costs, but although they may be suitable for some studies, the lack of single base resolution reduces their usefulness. Rakyan *et al*. state that in their view, the Illumina 450K Infinium Methylation

BeadChip array is currently the best option for EWA studies as it provides a single-base resolution, high genome coverage (over 450 thousand CpG sites interrogated) and is suitable for high throughput, analyzing up to 96 samples per run.

1.2 DNA methylation and gene expression

1.2.1 Examples from normal cellular processes; imprinting and X inactivation

Normal cellular processes, such as imprinting and X-chromosome inactivation provide examples of the role of DNAm in controlling gene expression.

Specific genes, often clustered together, are non-randomly expressed in a parental-origin specific manner in all somatic cells or in specific cells or tissues of the body, a phenomenon termed imprinting. Control of parental-origin specific expression requires differential DNAm of the parental alleles in regions in close proximity to the genes, so called imprinting specific differentially methylated regions (iDMRs) (rev. in ref. (34)). The allele specific methylation is inherited from parent to offspring and escapes the genome-wide demethylation event in the zygote. Methylation in iDMRs is erased along with other methylation marks in primordial germ cells and subsequently reestablished at specific iDMRs depending on the embryo's gender (rev. in ref. (15,34)). In section 1.3.3 (pg. 8), several examples are discussed of aberrations in iDMRs that lead to loss of parental-origin specific expression of imprinted genes, highlighting the causal role of DNAm in control of gene expression.

In placental mammals, the different dosage of X-chromosomes in males and females is compensated by inactivation of one of the X-chromosomes in females. Multiple mechanisms take part in achieving X-inactivation, including coating of the inactive chromosome with the Xist non-coding RNA, histone modifications and DNAm of gene promoters (rev. in ref. (35)). Evidence for a causal role of DNAm in controlling X-inactivation is discussed in the next section.

1.2.2 Evidence for a causal role

Experimental evidence for a role of DNAm in regulating gene expression date as far back as to the 1980s. Vardimon *et al*. injected *in vitro* methylated and unmethylated DNA containing a reporter gene into frog oocytes, and observed expression of the gene in oocytes that were injected with unmethylated DNA, but not in those that were injected with methylated DNA (36). In a similar experiment, Stein *et al*. found that when transfected into cultured mouse cells, the *Aprt* (adenine phosphoribosiltransferase) gene was silenced when methylated, but not when it was unmethylated (37). Shortly after it was discovered that silencing of genes on the inactive X-chromosome was correlated with promoter DNAm levels, a study was conducted to test their causal relationship. Treating cells with a methyltransferase inhibitor, 5-azacytidine, caused expression of genes from a previously inactive X-chromosome (38). In a subsequent study, transfecting cells with DNA obtained from cells that had received inhibitor treatment showed the same effect on a specific gene on the X-chromosome, indicating that the observation was not caused by secondary mechanisms of the inhibitor, but rather changed DNAm state (39).

Results from more recent studies, which have investigated the correlation between gene expression and DNAm at a large number of loci (whole chromosomes or across the whole genome)

provide evidence that these findings are not confined to a few genes. In three recent reports, where DNAm was measured in normal human cell lines (i.e., not from diseased tissues) in 66.000 CpG sites on chromosomes 12 and 20 and at 7.000 CpG sites in ENCODE (Encyclopedia of DNA elements) pilot project regions (44 genomic regions comprising 1% of the human genome (40)) at a single-base resolution and in the entire genome at 100 bp resolution, all found that DNAm in proximity to transcription start sites (TSS) correlated negatively with gene expression. In addition, all studies found that DNAm in gene-bodies correlated positively with gene expression (41–43). This has also been observed in other experimental settings, e.g., when comparing DNAm levels between the inactive and active X-chromosomes, researchers found that the inactive chromosome is more heavily methylated in the region flanking TSS, but less methylated in gene bodies compared to the active chromosome (44).

1.2.3 Mechanisms

How DNAm affects gene expression may be mediated through multiple mechanisms, but two are described here. First, it is possible that the addition of methyl groups to DNA interferes with transcription factor binding. This is supported by experimental evidence, e.g., showing that a particular transcription factor, MLTF, is unable to bind to DNA when CpG sites in its recognition sequence are methylated (45). Second, methylated DNA may attract proteins that mediate gene silencing. Multiple proteins with affinity for methylated DNA sequences have been identified, and evidence suggests that some may recruit other proteins or enzymes that cause compacting of chromatin, e.g., through histone modifications resulting in repression of gene expression. These and other potential mechanisms are reviewed in reference (46).

1.3 DNA methylation aberrations in disease

Investigation of the role of DNAm in human disease has largely been limited to cancer, where aberrant DNAm of multiple genes has been linked to multiple types of the disease. Additionally, DNAm aberrations have been well established in so called imprinting disorders. Their association with other diseases has however not been studied in detail, but recently EWA studies for a few non-malignant complex diseases have been reported. It is worth noting that at least in some cases, differences in DNAm between cases and controls discussed below may represent normal DNAm variation, rather than an aberration per se, but for simplification they are referred to as DNAm aberrations. Before examples of DNAm aberrations are discussed, the next section considers where inter-individual variation in DNAm may originate.

1.3.1 When and how do variations of DNAm arise

Inter-individual variation of DNAm may arise by three means (rev. in ref. (27)) or by a combination of these means; it may be 1. inherited, 2. due to stochastic events or 3. environmentally or lifestyle induced. If DNAm of a particular region is inherited, it may be observable in all tissues of the body. This is an important concept for studies linking DNAm with disease, because in this situation DNAm may potentially be measured in easily acquired surrogate tissues such as whole blood rather than target tissues which may be hard to obtain. In addition, if stochastic or environmentally induced events occur early in development, they may also be observable in many tissue types.

Inherited inter-individual variation has to be divided into two classes; one that depends on the underlying genotype, and a second that does not. Due to the epigenetic reprogramming that follows oocyte fertilization, transgenerational epigenetic inheritance that is independent of genotype may be limited. However, evidence for such inheritance comes from studies of the agouti locus in mice. The agouti gene, *A,* is responsible for yellow coat color of mice, *a/a* mice are black. The A^{vy} allele carries an intra-cisternal A particle (IAP) retrotransposon upstream from the gene, and in A^{vy}/a mice, coat color depends on methylation status of the IAP insertion; heavy methylation results in black (termed pseudoagouti) coat color while lower methylation causes yellow color. Interestingly, the coat color phenotype of A^{vy}/a offspring that inherit the A^{vy} allele maternally depends on the maternal phenotype, suggesting that DNAm of the A^{vy} allele is transgenerationally inherited (47). Whether this type of inheritance extends to other genomic loci or is observable in humans is unclear. Methylation of IAP transposons has been observed to escape reprogramming (48,49), and it is therefore possible that this is an isolated occurrence. It nevertheless presents an example of the potential for non-sequence dependent intra-individual variation of DNAm due to transgenerational inheritance. Several studies have demonstrated associations between DNAm and the underlying genotype that extend over both short and long distances, even across chromosomes (50–52). Most of these associations are not absolute, i.e., methylation is not unequivocally varied by genetic variants. A specific genotype rather generates an increased probability of a particular methylation level (27).

Evidence for both stochastic and environmentally induced epigenetic changes come from studies on twins. Fraga *et al*. showed that monozygotic twins that had shared more of their life together had more similar epigenetic patterns than those that had shared less of their lives together (53). This may suggest an environmental contribution to epigenetic patterns. In addition, all twin pairs, including those that had spent more of their life together and very young twin pairs (youngest pair was 3 years old), had some dissimilarity in terms of epigenetic states. Although this does not rule out environmental effects, it may suggest that stochastic events occur. Multiple environmental and lifestyle factors have been associated with DNAm variation, including diet, smoking, environmental toxins and etc. (rev. in ref. (54)) and stochastic events may occur e.g., due to errors in maintaining DNAm through cell division.

1.3.2 DNAm aberrations in cancer

DNAm aberrations in cancer were first reported in 1983 by Feinberg and Vogelstein. Using methylation sensitive restriction enzymes on human normal and cancerous cell derived DNA followed by southern blotting, they found several CpG sites to be unmethylated in cancers, where they were methylated in normal cells from the same tissue (55). In the same year, Gama-Sosa *et al*. reported that global DNAm levels were reduced in human tumor samples compared to normal tissue using HPLC methodology (56). Since these studies were reported, DNAm aberrations have been associated with multiple forms of the disease (rev. in ref. (57,58)). An example of the scale of these aberrations is provided in a report by Costello *et al*. where it was estimated that in the 98 primary tumor types investigated, an average of 600 (range 0-4.500) of 45.000 CpG islands were aberrantly methylated (59).

A DNAm aberration is often referred to as a hypermethylation or a hypomethylation, depending on whether the DNA sequence under investigation is methylated in a normally unmethylated region, or unmethylated in a normally methylated region, respectively. Hypermethylation of multiple tumor suppressor genes has been observed and hypomethylation of repeat sequences and a few oncogenes (58). In a review published in 2004, Feinberg and Tycko mention that although hypomethylation was the first DNAm aberration identified in cancer, it was subsequently mostly overlooked and hypermethylations are better defined (57). They note that this is due to a bias in experimental design, researchers focused on aberrations in normally unmethylated sites, and thus only identified hypermethylations. Examples of aberrantly methylated genes in cancer include hypermethylation of the *CDKN2A* (INK4A cyclin dependent kinase inhibitor) tumor suppressor in bladder tumors and hypomethylation of the *SNCG* (gamma-synuclein) oncogene in breast and ovarian tumors, which correlate with decreased and increased gene expression respectively (60,61).

1.3.3 DNAm aberrations in imprinting disorders

Imprinting disorders are diseases that arise due to defects in imprinted regions leading to loss of parental-origin specific expression. The defects can be of genetic or epigenetic nature. Genetic defects such as uniparental disomy (a genetic region is inherited in two copies from one parent and not from the second parent), deletions, duplications, translocations and etc, can lead to imprinting disorders when they include an imprinted region. As these defects are not epigenetic in nature, they will not be discussed here, but they account for a large proportion of the cases affected by the diseases discussed. Epigenetic defects in the context of imprinting disorders are termed loss of imprinting (LOI). LOI is a loss or gain of DNAm in iDMRs leading to loss of expression in the normally expressed allele or gain of expression in the normally silent allele.

Examples of imprinting disorders are Beckwith-Wiedemann syndrome (BWS), characterized by overgrowth and predisposition to embryonic tumors, Prader-Willi syndrome (PWS) and Angelman syndrome (AS), characterized by mental retardation and behavioral abnormalities and Transient neonatal diabetes mellitus (TNDM), characterized by diabetes which presents in the first weeks after birth, followed by remission in following months, and frequently a recurrence in adult life. All disorders can be caused by LOI and provide examples of how DNAm aberrations can affect gene expression and phenotypes.

Epigenetic defects leading to BWS include LOI of *KCNQ10T1* (KCNQ1 overlapping transcript 1) due to loss of DNAm on the maternal allele and LOI of *IGF2* (insulin-like growth factor 2) due to *de novo* DNAm on the maternal allele, both resulting in biallelic expression of the genes. Upregulation of *IGF2* is thought to contribute to the predisposition to embryonic tumors, while upregulation of *KCNQ10T1,* which causes a downregulation of proximal genes, is thought to contribute to the other phenotypic characteristics of BWS. Epigenetic defects leading to PWS and AS occur in the same genomic region, 15q11-q13. In PWS, LOI of multiple genes in the region through *de novo* DNAm of the paternal allele results in loss of their expression. In AS, LOI only affects a single gene in the region, *UBE3A* (ubiquitin protein ligase E3A), through loss of DNAm on the maternal allele*,* resulting in

loss of its expression. Finally, in TNDM, LOI of *PLAG1* (pleomorphic adenoma gene-like 1) through loss of DNAm of the maternal allele results in biallelic expression of the gene (rev. in ref. (58)).

1.3.4 DNA methylation aberrations in non-malignant complex diseases

Whether aberrant DNAm is associated with non-malignant complex diseases has not been investigated in detail. With the advent of technologies that allow determining the methylation levels of thousands of CpG sites across the genome, a few EWA studies for non-malignant complex diseases have been reported in the last few years. The diseases investigated in these studies include auto-immune related diseases; type 1 diabetes (T1D), systemic lupus erythematosus (SLC), rheumatoid arthritis (RA), dermatomyositis (DM) and multiple schlerosis (MS) and mental disorders such as autism, schizophrenia and bipolar disorder, and other disease related traits such as body mass index (BMI) and obesity (1,62–67). In addition, two studies, published after the present study was initiated, have investigated the association between DNAm and type 2 diabetes. These will be discussed in the discussion chapter of the thesis in relation to the results of the present study.

Rakyan *et al*. investigated DNAm in 27.458 CpG sites in 14.475 promoters (i.e., about 2 CpG sites were investigated per promoter) in CD14+ monocytes from 15 monozygotic (MZ) twin pairs discordant for T1D (62). In using discordant MZ twins, confounding due to underlying genotype that may affect DNAm levels is prevented and thus serves as a very convenient model for studying DNAm aberrations. Their analysis identified 132 CpG sites in 132 different loci where the direction of the DNAm difference in the intra-twin pairs was significantly associated with T1D. Several of these CpG sites were located in genes involved in immune function or in genes previously associated with T1D, including *HLA-DQB1* which contains a genetic variant conferring the highest known genetic predisposition for the disease. In addition, Rakyan *et al*. investigated the temporal origin of the differences observed and found that for 71% of the differentially methylated CpG sites, the same directionality of the aberration was also found in pre-disease manifestation samples from individuals that later developed the disease, compared to controls. Javierre *et al*. investigated DNAm in 1505 CpG sites in 807 gene promoters in whole blood DNA from 15 MZ twin pairs discordant for three diseases SLC, RA and DM (i.e., five pairs discordant for each disease) (63). No significant differences were observed in DNAm levels of intra-pairs discordant for RA or DM. Significant intra-pair DNAm differences were observed in 49 genes in twins discordant for SLC. Many of these genes have disease relevant functions, such as in immune response and cytokine production. In addition, Javierre *et al*. investigated correlations between DNAm and expression in seven differentially methylated genes, and found that five showed a significant association. Baranzini *et al*. investigated DNAm in 1.7 million CpG sites in CD4+ lymphocytes in 3 MZ twin pairs discordant for MS (64). Their analysis revealed no DNAm differences that were shared between all twin pairs. However, they only considered marked differences in DNAm of an on/off fashion; changes to/from less than 20% to/from more than 80%. Nguyen *et al*. investigated DNAm in 8109 CpG islands in lymphoblastoid cell lines derived from 3 MZ twin pairs discordant for autism (65). Their analysis revealed significant intra-pair differential methylation in 73 islands, including in multiple genes with biologically relevant functions such as in nervous system development and in several genes involved in neurological disorders.

Additionally, Nguyen *et al*. investigated expression of two genes that were differentially methylated, *BLC-2* (B cell CLL/lymphoma 2) and *RORA* (retinoic acid-related orphan receptor alpha), and found that they were downregulated in autistic samples. Further they found that the downregulation was alleviated with treatment with methyltransferase inhibitor, suggesting a causal role for DNAm in their regulation. Finally, Feinberg *et al*. investigated DNAm in 227 highly variably methylated regions in 74 DNA samples from whole blood, selected randomly from individuals recruited by the Icelandic Heart Association and tested for associations between DNAm and BMI (66). Their analysis revealed that methylation levels in or near four genes, including *PM20D1* (peptidase M20 domain containing 1) was significantly associated with BMI. These genes have previously been implicated in body weight regulation or diabetes.

These studies have revealed evidence that aberrant DNAm is associated with many non-malignant complex diseases. Yet these studies only interrogated a small fraction of the genomes total number of CpG sites. In addition, they were generally performed using arrays that have a biased selection for promoter regions and specific genes, such as oncogenes, differentially expressed genes and etc. These results may therefore only represent the tip of the iceberg. With one exception, they have been conducted in a retrospective setting (comparing cases and controls) and thus the observed differences in DNAm potentially do not represent predisposing aberrations. Until this has been established, their clinical relevance is uncertain.

1.3.5 Applications for DNAm aberrations

The identification of DNAm aberrations in relation to disease has at least three potential applications: Identified aberrations may 1. contribute information on disease pathogenesis, 2. serve as drug targets and 3. be used for disease detection or prediction.

Both correlational and more direct experimental evidence suggests that DNAm aberrations of specific genes associated with cancer, imprinting disorders and non-malignant complex diseases affect gene expression. These findings provide evidence for a potential role for DNAm aberrations in disease pathogenesis. The first step towards elucidating possible etiological role will be to establish their temporal origins, i.e., whether these aberrations are present prior to disease onset and are thus potentially causal in disease development, or whether they arise after disease onset.

In principle, epigenetic marks are reversible and are therefore attractive targets for therapeutic drug treatment (28). The methyltransferase inhibitor drug 5-azaCdR has been used in chemotherapy treatments for leukemias (57). In line with its inhibitory effects on methyltransferases, it is possible that its therapeutic influence stems from hypomethylating effects that may induce expression of tumor suppressor genes (57). However, evidence suggests that its effect, or at least part of it, may be mediated through a different mechanism. The drug incorporates into DNA, where it inhibits methyltransferases by covalently binding to them, and experimental evidence suggests that the methyltransferase-DNA adduct is rather the causal factor in the drugs cytotoxicity than its effects on DNAm (68).

DNAm aberrations may have an application as diagnostic markers of disease or markers for disease risk assessment and perhaps in addition, monitoring of recurrence and disease stratification

(28). The concept of using DNAm for disease detection has been studied in some detail for cancer. Early detection of cancer is very important for disease prognosis but for some types of the disease detectable symptoms do not arise until after it has metastasized. In such cases, biomarkers present one of the best options for early detection (28). Measurements of DNAm levels in DNA obtained from bodily fluids, such as plasma, serum, urine, saliva and etc. have been tested for use as such biomarkers because tumor cells or tumor cell DNA can often be found in such fluids. In a review from 2003, it was noted that such biomarkers can be highly specific (i.e., do not detect aberrations in controls), and that their sensitivity (i.e., detection of the aberration in cases) is generally around 50% (28). Using DNAm aberrations for detection for many non-malignant complex diseases may not be of much relevance because they often present with obvious symptoms or are detectable with other robust means (e.g., blood glucose measurements for diabetes). For these diseases, DNAm aberrations may however be useful for disease risk assessment, particularly if they are detected in easily collectable tissues such as blood. Their ability for such application is however unclear due to lack of studies on associations between DNAm and diseases performed in a longitudinal setting (69).

1.4 Diabetes

1.4.1 Characteristics, prevalence and risk factors

Type 2 diabetes mellitus (T2DM) is a chronic disease characterized by high blood glucose levels (hyperglycemia) due to insulin resistance (the biological effects of insulin on liver and skeletal muscle that normally result in decreased glucose production and increased glucose uptake respectively are reduced) and reduced β cell function (insulin production in the pancreas is decreased) (70). T2DM risk factors include aging, physical inactivity, overweight and hypertension (71). Disease symptoms include excessive urine production accompanied by thirst and increased fluid intake, weight loss, lethargy, blurred vision and changes to energy metabolism (72). The disease is accompanied by serious complications, mostly micro- and macro-vascular in nature, including renal failure and cardiovascular disease. These complications present the greatest medical and socioeconomic burden of the disease (73). Other forms of diabetes, e.g., type 1 diabetes and maturity onset diabetes of the young share some or all of the characteristics, symptoms and complications of type 2 diabetes. In the following discussion diabetes will refer to all types of the disease combined, and T2DM explicitly to type 2 diabetes. In other chapters of the thesis, diabetes only refers to T2DM unless specified otherwise.

The International Diabetes Federation estimated that in 2010, 285 million individuals aged between 20-79 years were diabetic globally, of which 90-95% had T2DM (rev. in ref. (71)). Global prevalence of the disease has risen rapidly over the last decades; Danaei *et al*. estimated that it was 8.3% in men and 7.5% in women in 1980 and increased to 9.8% in men and 9.2% in women in 2008 for individuals 25 years and older (74), and it has been predicted to continue to increase (75). The prevalence in the Icelandic population specifically, has also grown, although it is still lower than global estimates. In a report from 1953, Albertsson describes studies and observations made by medical doctors in Iceland between ca. 1850-1940, which all concluded that diabetes was a very rare disease, and that it was an “extraordinary occurrence for doctors to come across it” (76). Using data from the Icelandic Heart Association, Bergsveinsson *et al*. reported that that the prevalence of T2DM in 45-64 year old

Icelanders rose by 50% between 1967-72 and 1997-2002 and was 3.8% in both genders in the latter period (77).

A number of demographic and health care related factors are thought to contribute to the increased global prevalence estimates; elderly continually comprise a larger proportion of the population with increasing life spans and reduced birth rates, mortality among diabetes patients has decreased and diabetes is more readily diagnosed. However, lifestyle and environmental changes in the past decades concerning e.g., diet, exercise and stress are also though to contribute to the increased prevalence due to their effects on disease risk factors (i.e., higher weight, higher blood pressure and etc.) (71). As noted above, multiple examples of associations between DNAm and environmental or lifestyle factors have been reported. It is possible that these factors mediate their effects on T2DM risk through epigenetic factors.

The increased prevalence of T2DM cannot be explained by emerging genetic factors, because the increase is taking place at such a fast rate, over one or two generations (71). Nevertheless, even before genetic factors were found to associate with the disease, they were a known contributing factor. The evidence supporting this presumption included; high concordance rates for T2DM in monozygotic twins, 35-58% compared to 17-20% in dizygotic twins, increased risk of T2DM development in offspring with affected parent(s) and differences in prevalence between ethnic groups (71,73). Although estimates vary, the heritability component of T2DM risk may exceed 50% (71). Perhaps interaction with the aforementioned lifestyle and environmental factors, which were not present until very recently are necessary to reveal the genetic predisposition (71). Again, it is possible that these interactions are mediated through epigenetic factors.

1.4.2 Genetic contribution to T2DM

Prior to the advent of genome wide association studies (GWAS) in 2005, genetic studies on complex diseases and other phenotypes depended on candidate gene and family based linkage study approaches. In a report from 2005, these studies were described as disappointingly unsuccessful in identifying genetic variants associated with T2DM, only revealing a handful of associated genes or genetic regions, of which most associations were not robust (i.e., results were conflicting between studies) and only conferred a very small risk for development of the disease (73). From 2007, when the first GWAS for T2DM was reported by Sladek *et al*. (78) and to date, about 50 genetic risk variants have been associated with the disease (79). These include variants in loci containing the *HHEX* (homeobox, hematopoietically expressed) (78), *KCNJ11* (potassium inwardly rectifying channel, subfamily J, member 11) (80) and *KCNQ1* (potassium voltage-gated channel, KCT-like subfamily, member 1) (81) genes investigated in the present study. *HHEX* is expressed in the pancreas and codes for a transcription factor implicated in pancreatic development, *KCNJ11* and *KCNQ1* are expressed in several tissues, including the pancreas and code for proteins that form a part of potassium channels involved regulation of cell membrane potential (discussed in references (78,80,81)). The variants in all three loci are thought to associate with T2DM through effects on pancreatic β cell function (71) due to the aforementioned functions, and/or due to results from studies investigating the variants' association with indices of β cell function and insulin sensitivity. One such

index is homeostasis model assessment (HOMA, -IR for insulin resistance and –β for β cell function) used in the present study. Most other indices of insulin sensitivity and β cell function require data from complex measurements, such as time series data from a glucose tolerance test. The HOMA model is based on experimental data, i.e., it is constructed from physiological dose responses of glucose uptake and insulin production, and can be used to estimate IR and β cell function relative to a normal reference population using routinely measured clinical parameters, fasting glucose and insulin concentrations (82,83).

Similar to previous studies, the T2DM risk variants identified through GWAS are common, generally with allele frequencies of > 25% and only confer a very small increase in risk for T2DM. Odds ratios for having the disease if carrying a risk allele is generally less than 1.2 (79).

1.4.3 Missing heritability

Despite the advances in identifying genetic T2DM risk variants in the last years, 90% of the total heritable component of T2DM risk remains unidentified (71). This scenario is common for other complex diseases and phenotypes, e.g., Chron's disease, systemic lupus erythematosus and height, and has raised the question where the missing heritability is hiding (rev in ref. (84,85)). A few possible explanations are discussed here: First, it is possible that the proportion of the identified heritable component is underestimated because the genetic variants identified through GWAS are usually not the actual causal variants, but rather are identified as associated with the trait in question because they are in linkage disequilibrium (LD) with the causal variants. The causal variants may be more strongly associated with the phenotype of interest, and thus explain more of the risk. Second, much larger number of risk variants conferring smaller risk may exist that require much larger sample sizes for detection than have been used previously. Third, causal variants that are independent from those identified through GWAS may exist in the loci identified. Fourth, structural variants, such as copy number variants may contribute to these phenotypes, but have not been investigated in detail. Fifth, rare and low frequency variants may contribute to the disease, and perhaps confer greater risk than common variants but GWA studies have limited potential to capture them. Finally, epigenetic variation may contribute to the heritability estimates. Epigenetic variants that depend on genotype would presumably be readily detectable by genetic studies. Evidence for association between GWAS identified disease risk variants and DNAm levels have been reported; the *FTO* (fat mass and obesity associated) variants rs8050136 and rs1121980, associated with T2DM and obesity risk, have been shown to correlate with proximal DNAm (86,87). In this scenario, it is possible that the identified genetic variant is tagging variations in DNAm that it or other variants it is in LD with cause and through which the genetic variant affects cell function. Epigenetic variation that is inherited from parent to offspring but does not depend on the underlying genetic variation would however not be captured by genetic studies.

1.4.4 T2DM prevention and prediction

Impaired fasting glucose (IFG, hyperglycemia in fasting state, although not as severe as in T2DM) and impaired glucose tolerance (IGT, hyperglycemia after glucose load, i.e., consumption of a large dose of glucose, although not as severe as in T2DM) precede overt T2DM. If untreated, 7% of individuals

with IFG and IGT develop T2DM every year (73). A few studies have investigated whether simple interventions involving lifestyle changes of diet and exercise could delay or prevent these conditions progressing to T2DM, with promising results.

The first study of this kind was performed in China on 577 individuals with IGT (88). The individuals were randomized into four groups, of which three received active treatment, and the fourth served as a control group. The treatments consisted of lifestyle change with regard to diet in one group, exercise in a second group and both diet and exercise in the third group with active counseling over a six year period. The diet prescribed for individuals with BMI < 25 kg/m^2 contained 25-30 kcal/kg body weight and these participants were encouraged to reduce consumption of alcohol and simple sugars and increase vegetable intake. Individuals with BMI ≥ 25 kg/m^2 were encouraged to reduce caloric intake to reduce body weight by 0.5-1.0 kg per month until they reached a BMI of 23 kg/m^2. The exercise recommendation consisted of 5-60 minutes of exercise per day depending on intensity and physical form of the individuals. After six years the cumulative incidence of diabetes was 67.7% in the control group compared with 43.8% in the diet group, 41.1% in the exercise group and 46.0% in the combined diet and exercise group. The difference corresponds to a 31-46% reduced risk of disease development. A follow up study was conducted 14 years later, a period of no intervention, to investigate whether any long term benefits of such interventions could be observed (89). The study showed that in 20 years after the intervention was initiated, the individuals receiving any form of intervention had a 43% lower incidence of diabetes compared to controls. Other studies have revealed similar results. For example, in the largest study (i.e., with the largest number of participants) on effects of lifestyle intervention on T2DM risk conducted to date, significantly reduced risk for T2DM progression with simple lifestyle interventions was also observed (90). The study followed 3234 individuals with IFG and IGT for an average of 2.8 years, randomly assigned to three groups; one receiving a lifestyle modification program with the goal of 150 minutes of physical activity per week and a 7% weight loss, a group receiving metformin (a drug used for managing T2DM that increases insulin sensitivity of muscle cells and reduces hepatic glucose production), and a control group receiving a placebo. The lifestyle intervention reduced the incidence of T2DM by 58% and metformin by 31% compared to the placebo group, suggesting that physical activity is a more effective prevention for T2DM development than metformin. Due to the effectiveness of these interventions, identifying individuals at increased risk for T2DM is highly beneficial, both for the individual and society.

Current risk prediction models cannot accurately predict individual risk of T2DM development. It was expected that identified genetic T2DM risk variants would improve such models due to the high genetic contribution to the disease (71). Recently, several studies have been conducted comparing the accuracy of T2DM risk prediction models relying on conventional risk factors with those relying on genetic risk scores and combination of the two. Eleven such studies were reviewed by Herder and Roden which revealed that genetic risk scores were less accurate predictors of type 2 diabetes development than scores using conventional risk factors (71). The area under the receiver-operating characteristic curve (AUC) was about 10-35% lower when using genetic risk scores. Additionally, although combining the two resulted in a significant increase in AUC compared to models only relying

on conventional factors, the increase was too subtle to be of any clinical relevance, AUC changed by ≤ 0.02.

2 Aims

2.1 Investigation of DNA methylation in type 2 diabetes genetic risk loci

2.1.1 Hypothesis

In the past few years, genetic studies have been very successful in identifying multiple genetic T2DM risk variants, but a large proportion of the disease's heritable component nevertheless remains unidentified. Several explanations may account for missing heritability, but one is the existence of inherited epigenetic aberration(s). Additionally, multiple environmental and lifestyle factors are associated with T2DM risk and it is possible that their effects are mediated through induced epigenetic aberration(s). On this basis, it was hypothesized that aberrant DNAm could be associated with T2DM, and to address this hypothesis, the present study aimed to identify such aberrations.

2.1.1.1 Specific aims to address hypothesis:

1. Measure and compare methylation levels in DNA extracted from whole blood in selected regions between a small number of diabetics and controls in order to identify any aberrantly methylated region(s). Loci that have been associated with type 2 diabetes through genetic studies were targeted because it was hypothesized that such loci may be good candidates for identifying epigenetic aberrations that may be associated with the disease, due to the assumption that altered expression that they may cause may have similar consequences as alterations in gene products (at least if such alterations change the product's activity). The *HHEX*, *KCNJ11* and *KCNQ1* loci were selected as they were considered representative of loci exhibiting a genetic association with the disease.
2. Select a promising candidate from the above experiment (if one is identified), and repeat the comparison in a larger set of samples to establish the robustness of the finding, and to enable adjusting for potential confounders.

2.2 Investigation of whether cellular heterogeneity may confound analyses of DNA methylation data

2.2.1 Hypothesis

As peripheral blood cell DNA is relatively easily accessible it has been an essential source for genetic experiments for the past decades. However whether it is appropriate material for studies on epigenetics has been debated (91) because inter-individual variation in the number of specific white blood cells in combination with cell specific methylation could compromise measurement outcomes for DNAm carried out on cells from whole blood. This concern has largely been theoretical due to lack of experimental data. The hypothesized confounding on whole blood DNAm measurement outcomes due to cellular heterogeneity was therefore investigated.

Analyses were conducted in a non-disease specific context to understand the potential for confounding in general. A confounding effect may be region-specific, depending on two factors; first, the size of the difference in methylation level between cell types, and second due to the relative size of

the difference compared to the variation in methylation levels caused by other factors. DNAm was therefore analyzed in four regions, in genes *HHEX*, *KCNJ11*, *KCNQ1* and *PM20D1*, which represented a range of inter-individual variation in DNAm from very low to very high.

2.2.1.1 Specific aims to address hypothesis:

1. Measure DNAm in whole blood DNA samples in selected regions and test for an association between the measured levels and cellular heterogeneity.
2. Fractionate whole blood and measure and compare DNAm levels in the fractions to test whether methylation levels differ in the fractions and may thus underlie the observed association(s), if any.

3 Materials and methods

A part of the results presented in this thesis have been submitted for publication in PLoS ONE. The manuscript, which is provided as supplement at the end of the thesis, was used as a basis for section 1.1.3 in the introduction, section 4.1 of the results and section 5.1 of the discussion.

3.1 Sample acquisition and processing

3.1.1 Participants

Samples were obtained from two cohort studies conducted at the Icelandic Heart Association (IHA); the Age, Gene/Environment Susceptibility (AGES)-Reykjavik (92) and the Risk Evaluation For Infarct Estimates (REFINE)-Reykjavik studies. DNA samples extracted from whole blood were obtained from both studies and blood collected for fractionation from individuals taking part in the REFINE-Reykjavik study. In the latter case, the REFINE-Reykjavik study was not selected per se, individuals were visiting the clinic for participation in the REFINE-Reykjavik study but not the AGES-Reykjavik study at the time when these samples were needed.

Briefly, the AGES-Reykjavik study was the seventh visit of the Reykjavik Study, a population-based cohort study initiated in 1967, inviting all Reykjavik inhabitants born between 1907 and 1935 to participate. In this visit, 5764 of the surviving members were recruited. REFINE-Reykjavik is a prospective study on risk factors and cause of atherosclerotic disease in a population of Icelandic people. The main goal of the study is to improve the predictability of cardiovascular disease risk estimates. The study was initiated in 2005 and recruitment of the first phase was completed in spring 2011 recruiting 6942 men and women born in the years 1936-1980 living in the Reykjavik city area. Both studies are approved by the Icelandic National Bioethics Committee (VSN: 05-112, VSN: 00-063) and the Data Protection Authority and all participants gave written informed consent on entry into the studies.

3.1.2 Fractionation of blood samples

Blood samples collected for fractionation were processed as soon as possible, never later than 4 hours after being drawn. About a third of the volume collected from each individual was taken aside for extraction of whole blood DNA, and the rest used for the fractionation. The blood was fractionated by density gradient centrifugation using Histopaque-1077 Ficoll medium in Accuspin^{TM} Tubes (Sigma-Aldrich, St. Louis, MO, USA, cat.nr.: 10771 and A1930 respectively). In short (the protocol is provided in full in the appendix, pg. 54), blood was poured on top of Ficoll medium in an Accuspin Tube, where a porous barrier separates the two liquids and prevents them from mixing prior to centrifugation. On centrifugation, polymorphonuclear cells (PMNCs) descended through the membrane and to the bottom of the tubes while the mononuclear cells (MNCs) were retained in a thin layer between the plasma and the Ficoll medium. The MNC fraction was extracted from the plasma/medium boundary and the PMNC fraction from the bottom of the tubes using pipettes. DNA was subsequently extracted from each fraction and thus three DNA samples were obtained from each blood sample collected; from whole blood (MNCs and PMNCs combined), MNCs and PMNCs.

3.1.3 DNA extraction

A simple salting out method was used for DNA extraction, based on an extraction method developed by Scotlab Bioscience (Coatbridge, Scotland, UK, protocol is provided in the appendix, pg. 54). After extraction, the DNA was dissolved in TE buffer (see aforementioned protocol) and its concentration measured using ultra-violet absorbance quantification (260 nm) on a Spectramax M2 microplate reader (Molecular Devices, Sunnyvale, CA, USA).

3.2 Selection of loci and regions and design of assays

3.2.1 Selection of loci for the study of DNAm in T2DM

DNAm levels in four loci had been investigated at the IHA by sequencing of amplified, bisulfite converted DNA using Sanger methodology (referred to as the "Sanger study" hereafter). The four loci, *TCF7L2* (transcription factor 7-like2), *CDKAL1* (CDK5 regulatory subunit associated protein 1-like1), *HHEX* and *KCNJ11*, had been chosen on the basis that, according to the literature and a GWAS of a large proportion of the AGES-Reykjavik cohort, they were representative of loci exhibiting a strong genetic association with type 2 diabetes. The results from this analysis (data not shown) revealed little or no DNAm in the regions investigated in the *TCF7L2* and *CDKAL1* loci, and in one of two regions investigated in the *KCNJ11* locus. Evidence for low methylation levels were observed in the *HHEX* locus and high methylation levels in the second region investigated in the *KCNJ11* locus. More importantly, considerable inter-individual variation of DNAm levels was observed in this second region in the *KCNJ11* locus and in the *HHEX* locus. Therefore, the *HHEX* and *KCNJ11* loci were selected for further investigation. In addition, a third locus, *KCNQ1*, was selected for the present study because of its genetic association with T2DM and because DNAm levels in the locus had been shown to associate with the genotype of the identified T2DM risk variant (SNP (single nucleotide polymorphism) rs2334499) (93).

3.2.2 Selection of regions per locus for the study of DNAm in T2DM

In the present study, selected regions in one gene per locus were investigated. In each case the gene most proximal to the genetic variant exhibiting the strongest association with the disease was selected: in the *HHEX* locus, the *HHEX* gene, in the *KCNJ11* locus the *KCNJ11* gene and in the *KCNQ1* locus the *KCNQ1* gene. Genetic regions classified as CpG islands (CGI) or CGI shores (2 kb DNA stretch up- and downstream of CpG islands (18)) were targeted, principally those proximal to predicted transcription start sites (TSS). The regions were located with the University of California, Santa Cruz genome browser (Human March 2006 NCBI36/hg18 assembly) (94) using default track settings for identifying CGIs (the track is termed "CpG islands" in the browser) and TSS (both the "SwitchGear TSS" and "Epionine TSS" tracks were used, which predict TSS based on cDNA (complementary DNA) alignment (95) and computational analysis of sequence motifs (96), respectively). *KCNJ11* contains two CGIs (**figure 1A**), and both had been investigated in the Sanger study. The CGI located at the 3' end of the gene contains a predicted TSS. Results from the Sanger study revealed no methylation in this CGI and it was therefore not investigated in the present study. The other CGI in the *KCNJ11* gene resides in the gene body and is where inter-individual variation

was detected in the Sanger study. This region was therefore investigated further in the present study and is termed “the KCNJ11 region” in the following text. The *HHEX* gene contains four CGIs, and two were selected for investigation in the present study (**figure 1B**); the largest CGI, located in the 5’ region, which contains a predicted TSS (termed “the HHEXI region”) and an island residing in the gene body (termed “the HHEXII region”). In *HHEX*, two CGI shores were additionally investigated and are termed “the HHEXIII region“ and “the HHEXIV region”. Finally, in *KCNQ1* (**figure 1C**), two CGIs were selected, both containing a predicted TSS, located in the 5’ end (termed “the KCNQ1I region”) and the intragenically (“the KCNQ1II region”). The specific sequences analyzed per region are discussed in section 3.2.4 (pg. 21). Note that all region names are non-italicized to distinguish them from gene and loci names, e.g., the KCNJ11 region vs. the *KCNJ11* gene and the *KCNJ11* locus.

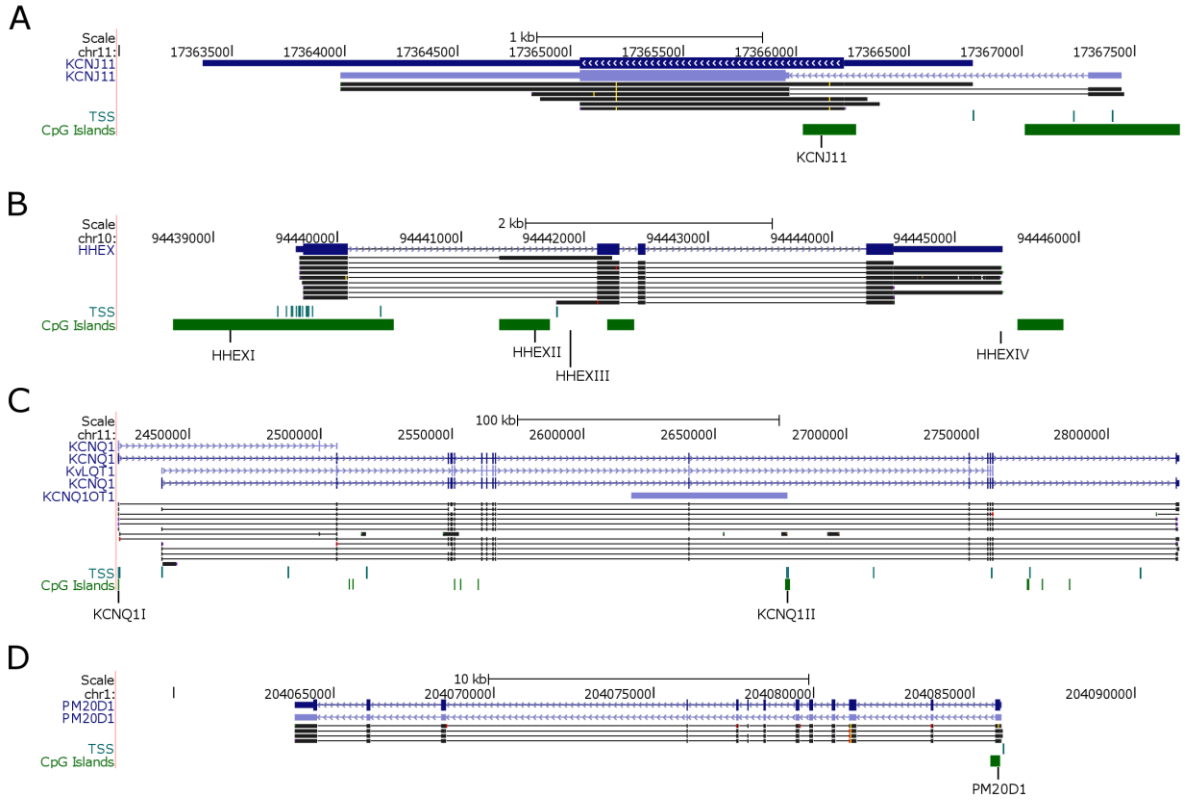


Figure 1. Gene-maps of the four genes investigated in the study.
The gene maps (adapted from the UCSC genome browser (94)) depict A. *KCNJ11*, B. *HHEX*, C. *KCNQ1* and D. *PM20D1*. At the top of each map, a ruler indicates the scale and genomic position of the regions shown. Genes are depicted in blue and mRNAs in black, the exons as blocks, the introns as thin lines connecting the blocks and arrows indicating the direction of transcription. Transcription start sites (TSS, the “Epionine TSS” and “SwitchGear TSS” tracks were combined) are shown as light blue ticks and CpG islands are shown as green blocks and ticks (depending on their size relative to the gene). At the bottom of each map the positions of the regions investigated are indicated (thin line points approximately to the middle of each region, with the name of the region below; KCNJ11, HHEXI-IV, KCNQ1I-II and PM20D1).

3.2.3 Selection of regions for the study of confounding

For the study of whether measured DNAm levels in whole blood may be confounded due to cellular heterogeneity, four regions were chosen to represent a range of inter-individual variation of DNAm from very low to very high. HHEXII, KCNJ11 and KCNQ1II were chosen to represent low to intermediate variability regions. A fourth region located in the only CGI in the *PM20D1* gene (termed "the PM20D1 region", **figure 1D**), was selected from previous, published work (66) to represent a highly variable DNAm region. More specifically, these regions were selected from a larger set of candidates (e.g., the other regions selected for the analysis of DNAm in T2DM) based on two criteria: First, on basis of the size of the inter-individual variability present in each region to represent a spectrum of variability from very low to very high and second, on basis of which region in each variability category had available data on DNAm in the largest number of whole blood DNA.

3.2.4 Design of assays for measuring DNAm in the selected regions

Assays were designed to measure DNAm levels using pyrosequencing methodology in a series of adjacent CpG sites in the selected regions. The assays are entitled with the name of the region they were designed for, e.g., "the HHEXI assay" for the HHEXI region. Primer sets (forward and reverse PCR primers, one tagged with biotin, and a sequencing primer) were designed using PyroMark Assay Design software (version 2.0.1.15, QIAGEN). The same software provided each assay's dispensation sequence for the pyrosequencer. The primer sets chosen, and thus the specific CpG sites interrogated per region, depended on their quality as determined by the software. Primers that would bind to a sequence containing a CpG site were not considered. Primer sequences for the selected sets are provided in **table 5** (appendix, pg. 57) and the genomic positions of the CpG sites analyzed in each region are listed in **table 1.** Assays were tested for PCR bias and their robustness assessed, and the results suggest that their precision is high but that their accuracy is low (sections 7.1.2.2 and 7.1.2.3, pg. 51 and 52). PCR bias, i.e., preferential amplification of particular alleles in a heterogeneous pool of alleles, has been demonstrated for bisulfite treated DNA (97) and most of the assays used here were biased towards unmethylated DNA to some degree. Consequently measured DNAm levels presented here may be lower than actual DNAm levels. Because this bias does not affect analysis outcomes, i.e., whether DNAm is or is not associated with diabetes status or cellular heterogeneity, but only the effect sizes, the assays are suitable to address the aims of the present study.

Table 1. Genomic positions* of the CpG sites analyzed per region.

		CpG site #, chromosomal position**									
Region	**Chromosome**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
KCNJ11	11	17366204	187	178	168	135	129	123	114		
HHEXI	10	94439119	130	137	141						
HHEXII	10	94441605	607	619	627	633	637	644	646	662	676
HHEXIII	10	94441831	855	875							
HHEXIV	10	94445518	522	529	539						
KCNQ1I	11	2423364	376	379	387	393					
KCNQ1II	11	2677095	111	115	117						
PM20D1	1	204085711	713	716	733	740	749	760			

* Human March 2006 NCBI36/hg18 assembly

**The genomic position is given in full for CpG site 1 per region. For the other CpG sites, only the last three digits of the position are provided.

3.3 DNA methylation measurements

DNAm was measured using a pyrosequencing based methodology. It involved three steps; 1. bisulfite conversion of DNA, 2. amplification of the converted DNA using PCR, and 3. sequencing of the products on a pyrosequencer.

3.3.1 Bisulfite conversion of DNA samples

Bisulfite conversion of DNA samples was carried out using the EZ DNA Methylation™ kit (Zymo Research, Irvine, CA, USA, cat.nr.: D5004) following the manufacturer's instructions (the protocol is provided in the appendix, pg. 55). In short, the DNA was diluted in a buffer and heated to separate the strands before bisulfite was added to the solution and incubated overnight. The DNA was subsequently washed in a series of steps and finally eluted using 15 µl of the provided elution buffer. When the DNA was not analyzed immediately following the conversion process it was stored at -20°C for later use. As a general rule, when certain groups of individuals were being compared, e.g., individuals with and without type 2 diabetes, each “conversion batch” contained the same number of samples from each group so as to minimize potential confounding due to batch effects. For the same reason, DNA from blood fractions and the corresponding whole blood DNA for each individual was also converted in the same batch. Signal strength in the pyrosequencing reactions and data quality obtained from the reactions was observed to be positively associated with DNA concentration used in preceding PCRs (for details see sections 7.1.1.1 and 7.1.2.1, pg. 48 and 50). On basis of these results, 400 ng of DNA were converted for each sample used for the experiments presented in the results chapter.

3.3.2 Amplification of converted DNA

For nested-PCR, two reactions were performed; the first in a total volume of 10 µl containing 2 µl of the bisulfite converted DNA, and the second in a total volume of 30 µl containing 3 µl of the reaction mixture from the previous PCR. For one-step PCR, the reaction was performed in a total volume of 30 µl containing 3 µl of the bisulfite converted DNA. Signal strength in pyrosequencing reactions on PCR products from a nested and a one-step PCR was not observed to differ (for details see section 7.1.1.2, pg. 49). A one-step PCR was therefore generally used for the experiments presented in the results chapter.

PCRs were carried out on a 2720 Thermal cycler (Applied Biosystems, Foster City, CA, USA) with 1X Standard Taq Reaction Buffer (New England Biolabs (NEB), Ipswich, MA, USA, cat.nr.: B90145), 0.2 mM dNTP (NEB, cat.nr.: N04465), 0.25 µM of each primer (Sigma-Aldrich) and the amount of polymerase according to the manufacturer’s instructions in each case (Taq polymerase from NEB (“Standard-Taq”), OneTaq from NEB, AmpliTaq from Life Technologies, Carlsbad, CA, USA and TITANIUM-Taq from Clontech, Mountain View, CA, USA (cat.nrs.: M0273, M0481, N8080160 and 639208 respectively) were tested). The signal strength in pyrosequencing reactions on PCR products from amplification using the four different polymerase brands was observed to differ considerably (for details see section 7.1.1.3, pg. 49). On basis of these results, TITANIUM-taq polymerase was generally used for the experiments presented in the results section.

For the nested PCR, the cycling conditions for both reactions were as follows; 5 cycles of 30s at 96°C, 90s at 55°C and 120s at 72°C followed by 35 cycles of 30s at 96°C, 90s at an assay specific annealing temperature and 90s at 72°C. For the one-step PCR, the cycling conditions were as follows; 40 cycles of 30s at 96°C, 90s at an assay specific annealing temperature and 90s at 72°C. Annealing temperatures were 62°C for all assays except for KCNQ1I (64°C) and HHEXIV (58°C). Pre-cycling conditions varied depending on the manufacturer's instructions for the polymerase used in each reaction and post-cycling conditions were; 240s at 72°C followed by a hold at 4°C.

3.3.3 Analysis of amplicons on the pyrosequencer

Preparation of PCR products for analysis on the pyrosequencer was performed according to the manufacturer's instructions (protocol is provided in the appendix, pg. 56). In short, the biotinilated sequencing template was extracted from the PCR product mixture by annealing with streptavidin coated sepharose beads (Streptavidin Sepharose™ High Performance, GE Healthcare, cat.nr.: 17-5113-01). The template was subsequently washed and made single stranded in a series of steps using a Vacuum prep workstation (QIAGEN cat.nr.: 9001518) and released onto a sequencing plate (QIAGEN, cat.nr.: 979201) containing annealing buffer (QIAGEN, cat.nr.: 979309) with the appropriate sequencing primer (**table 5** in appendix, pg. 57). The amplicons were sequenced using a PyroMark Q24 pyrosequencer (QIAGEN) and PyroMark™ Gold Q24 reagents (QIAGEN, cat.nr.: 97082).

3.3.4 Standard DNA

DNA samples of known methylation state, both 100% methylated (catalog number (cat.nr.): 59655) and 0% methylated (cat.nr.: 59665) were obtained from QIAGEN (Hilden, Germany, bisulfite converted by the manufacturer). To acquire DNA of intermediate methylation states, these samples were mixed in the desired proportions.

3.4 Data processing and statistical analyses

3.4.1 Analysis of pyrograms

Pyrograms from the pyrosequencing reactions were analyzed with the "PyroMark Q24 Software" (v1.0.10, QIAGEN). Two types of data were extracted, DNAm data and signal strength data.

DNAm data: Methylation levels were calculated from the pyrograms as the ratio between peak heights for methylated C's and the sum of methylated and unmethylated C's for each CpG site. Default software settings were used for quality assessment of the pyrograms per CpG site and measurements that failed the assessment were discarded when appropriate. Consequently, some individuals had missing values for one or more CpG site. A bisulfite conversion control was included in all assays where the conversion of a cytosine base in a non-CpG sequence was tested, which should theoretically be 100% if the conversion is complete. DNAm data from samples that failed this control were discarded as a whole (i.e., not on a per CpG basis) when appropriate. The DNAm data was deposited into the IHA's database as a percentage value per CpG site per region for each individual.

Signal strength data: Signal strength during the reactions was assessed by extracting data on light emission per nucleotide dispensation from the pyrograms. The average signal strength from all

dispensations was calculated and compared. The comparison was made in a relative manner, where the average signal strength under specific conditions was divided by the average signal strength from reactions that gave the lowest signal strength per analysis. The condition causing the lowest average signal strength thus had relative average signal strength of 1.

3.4.2 General data processing

For each statistical analysis, DNAm data was retrieved from the IHA's database with other information for each individual, e.g., age, gender, BMI and et cetera. Mixed models were used to compare DNAm levels between groups of individuals, which assume normal distribution of the data. Data on a percentage scale has a skewed distribution when the bulk of the data is close to either 0% or 100%. DNAm data for the HHEXI, HHEXII, HHEXIII and KCNQ1I regions was transformed prior to mixed model analysis due to such skewing by taking the arcsine of the square root of each percentage value. When multiple DNAm measurements had been conducted on a particular sample using the same assay, average DNAm values from the measurements were used in subsequent analyses. For each analysis, outliers in the DNAm data were identified per CpG site. Outliers were defined as values outside mean ± 2.698s, where s is standard deviation. For a standard Gaussian distribution, this criterion defines 0.35% of the data farthest from the mean in both directions as outliers. Data was not pooled from measurements conducted under varying conditions (e.g., nested or one-step PCR, or with different polymerases) unless it had been established that no measurement bias was detectable due to the differing conditions (see section 7.1.2.4, pg. 52).

3.4.3 Definitions of variables

The characteristics used in statistical analyses of the DNAm data were obtained from the IHA's database. This information was originally obtained from the participants in the AGES-Reykjavik and REFINE-Reykjavik studies through questionnaires, direct measurements, or from measurements of biosamples. Definitions, laboratory acquisition and derivations are explained in short:

Body mass index, **BMI** (kg/m^2), was calculated from measurements of weight (kg) and height (m), using the formula BMI = weight / $height^2$.

Homeostasis model assessment, HOMA indices, were derived from laboratory results of fasting glucose and insulin levels. Glucose (mmol/l) and insulin levels (µIU/ml) were measured in serum using automated clinical chemistry analyzers (Cobas c311 and e411 respectively, from Roche, Basel, Switzerland). HOMA-IR was calculated using the formula HOMA-IR = [glucose]*[insulin]/22.5 and HOMA-β with the formula HOMA-β = 20*[insulin]/([glucose]-3.5), where [glucose] and [insulin] are glucose and insulin concentrations in mmol/l and µIU/ml respectively. The output of the model is calibrated to give normal β-cell function of 100% and normal insulin resistance of 1 (82,83).

Type 2 diabetes mellitus, T2DM, status was derived from answers to the questionnaires and laboratory results (fasting serum glucose level). If an individual had a history of T2DM and/or was taking T2DM medication and/or his blood glucose level was ≥ 7 mmol/l (WHO 2006 diagnostic criteria (98)), he/she was defined as having T2DM. If an individual was not diabetic according to this criteria, but had a fasting plasma glucose level of 6.1-6.9 mmol/l (WHO 2006), he/she was defined as having **impaired fasting glucose, IFG**. Finally, individuals were defined as having **normal fasting glucose, NFG,** when they were non-diabetic according to the criteria above and had a fasting plasma glucose level of < 6.1 mmol/l (WHO 2006).

Blood cell ratios (%) were obtained from laboratory results. White blood cells (monocytes, lymphocytes, eosinophils, basophils and neutrophils) were counted in whole blood by an automated cell counter (Coulter HmX AL Hematology Analyzer, Beckman Coulter, High Wycombe, England, UK). The proportion of each cell type was calculated as the ratio of the count for the respective cell type of total white blood cell counts.

3.4.4 Statistical analysis for assessing correlation between DNAm and cellular heterogeneity

The coefficient of determination, R^2, was used to estimate the proportion of inter-individual variation in measured whole blood DNAm levels explained by differential white blood cell counts. The analysis was conducted using unadjusted mixed models with DNAm as the dependent variable and a random intercept term to account for the correlation of DNAm levels between CpG sites within a person (performed in SAS Enterprise Guide version 4.2). Since R^2 cannot be obtained directly from such analyses, two models were applied, an intercept only model containing only CpG sites as fixed effects and a full model where additionally, the proportional number of a specific white blood cell type was added to the intercept model as fixed effect. R^2 was then calculated from the residual variance (v_r) and variance of the random intercept (v_s) terms using the formula $R^2=(V_i-V_f)/V_i$ where $V_i=v_r+v_s$ for the intercept only model and $V_f=v_r+v_s$ for the full model. Individuals that had missing DNAm data for one or more CpG site (due to failed quality check, see section 3.4.1) or outliers in one or more CpG sites (for criteria, see 3.4.2) were analyzed separately.

3.4.5 Statistical analysis for comparing DNAm between cell fractions

For comparison of DNAm in two blood cell fractions, PMNC and MNCs, non-parametric statistical tests were used to avoid making a generalized assumption about the distribution of the data, which may differ between regions. Paired Wilcoxon signed rank test was used to assess statistical differences in methylation levels between the two cell populations and their correlation assessed with Spearman's ρ correlation coefficient using R version 2.13.2.

3.4.6 Statistical analysis for assessing correlation between DNAm and T2DM or HOMA indices

Mixed models were used to estimate the association between DNAm levels and diabetes status or HOMA indices. For these analyses, both unadjusted and adjusted models with a random intercept term were used where DNAm was the dependent variable, CpG sites fixed effects, and the independent variable and the covariables were added to the model as fixed effects. Individuals that had missing DNAm data for one or more CpG site or outliers in one or more CpG sites were analyzed separately. Such exclusion was not conducted when sample sizes were small ($n < 25$) to avoid losing statistical power and because identifying outliers from measurements on such small numbers of individuals may be misleading.

3.4.7 Other statistical analyses

Statistical tests used in analyses that are not described in detail above are specified in the results chapter when applied. These are Student's t-test (Welch, i.e. assuming unequal variance in the groups being compared) and Spearman's ρ correlation coefficient, performed in R version 2.13.2.

3.4.8 Plotting

All plots presented in the results and appendix chapters were prepared using R version 2.13.2. Figures 2-5, 10 and 12 were prepared using built-in functions while plotting figures 6-9, 11, 13 and 14 required additional packages called VennDiagram and gglplot2. A script was created, and is provided in the appendix (section 7.4, pg. 59), for preparing the plots presented in figures 6, 7 and 14. As multiple variations of boxplots exist, it is worth noting that the default settings for the boxplot() function in R were used here. In short the box's hinges correspond the first and third quartiles, and a horizontal line is drawn through the box marking the median. The whiskers extend to the most extreme data point that is no more than 1.5 IQR (inter quartile range, i.e., length of the box) from the box. Finally, circles or dots are drawn for data points that fall outside the range of the whiskers. In some cases, figures were edited, e.g., to indicate statistical significance between groups of data. Such editing was performed using GIMP graphics software, version 2.6

4 Results

4.1 Investigation of whether white blood cell heterogeneity can confound analyses of DNA methylation data

DNAm levels measured in DNA isolated from whole blood are dependent on the methylation levels in each white blood cell type and on the ratio of each cell type of the total cell count. A study was conducted to estimate whether cellular heterogeneity had potential to confound analyses of whole blood DNAm data by first testing for an association between whole blood DNAm levels and cellular heterogeneity, and second to test whether differential methylation in two cell fractions (MNCs and PMNCs) might underlie the association, if any. The analysis was not conducted in a disease specific context but rather its aim was to investigate the potential for such confounding in general.

4.1.1 Analysis testing for association between DNAm levels and cellular heterogeneity

DNAm was measured using pyrosequencing based methodology in four regions; HHEXII (10 CpG sites), KCNJ11 (8 CpG sites), KCNQ1II (4 CpG sites) and PM20D1 (7 CpG sites), i.e., in CGIs in four genes; *HHEX*, *KCNJ11*, *KCNQ1* and *PM20D1* respectively, in DNA isolated from whole blood to test whether the measured levels were associated with cellular heterogeneity. Samples from apparently healthy adults (~ 45% males), aged between 22-96 years were selected for this study from the AGES-Reykjavik and REFINE-Reykjavik cohorts (n=211 in total). Samples were analyzed independently for each region, and therefore there was only partial sample overlap between regions (**figure 13**, pg. 57). After exclusion of individuals due to missing values and outliers, whole blood DNAm data was successfully obtained for 169 individuals for HHEXII, 54 for KCNJ11, 49 for KCNQ1II and 59 for PM20D1.

Whole blood DNAm levels differed between CpG sites within each region (**figure 2**), but were generally low for HHEXII (range ~0-20%), intermediate for KCNQ1II (range ~40-60%), intermediate to very high for KCNJ11 (range ~60-100%) and ranging from very low to very high for PM20D1 (range ~0-100%). The results also indicated that intra-individual variability in DNAm differed between regions; it was high for KCNJ11 and KCNQ1II, but low for HHEXII and PM20D1. In general, inter-individual variability was very low for KCNQ1II, intermediate for HHEXII and KCNJ11 and very high for PM20D1; the standard deviation per CpG site ranged between 1.4-1.9 percentage points (pp) for KCNQ1II, 1.5-3.0 pp for HHEXII, 1.3-3.4 pp for KCNJ11 and 22.8-25.3 pp for PM20D1.

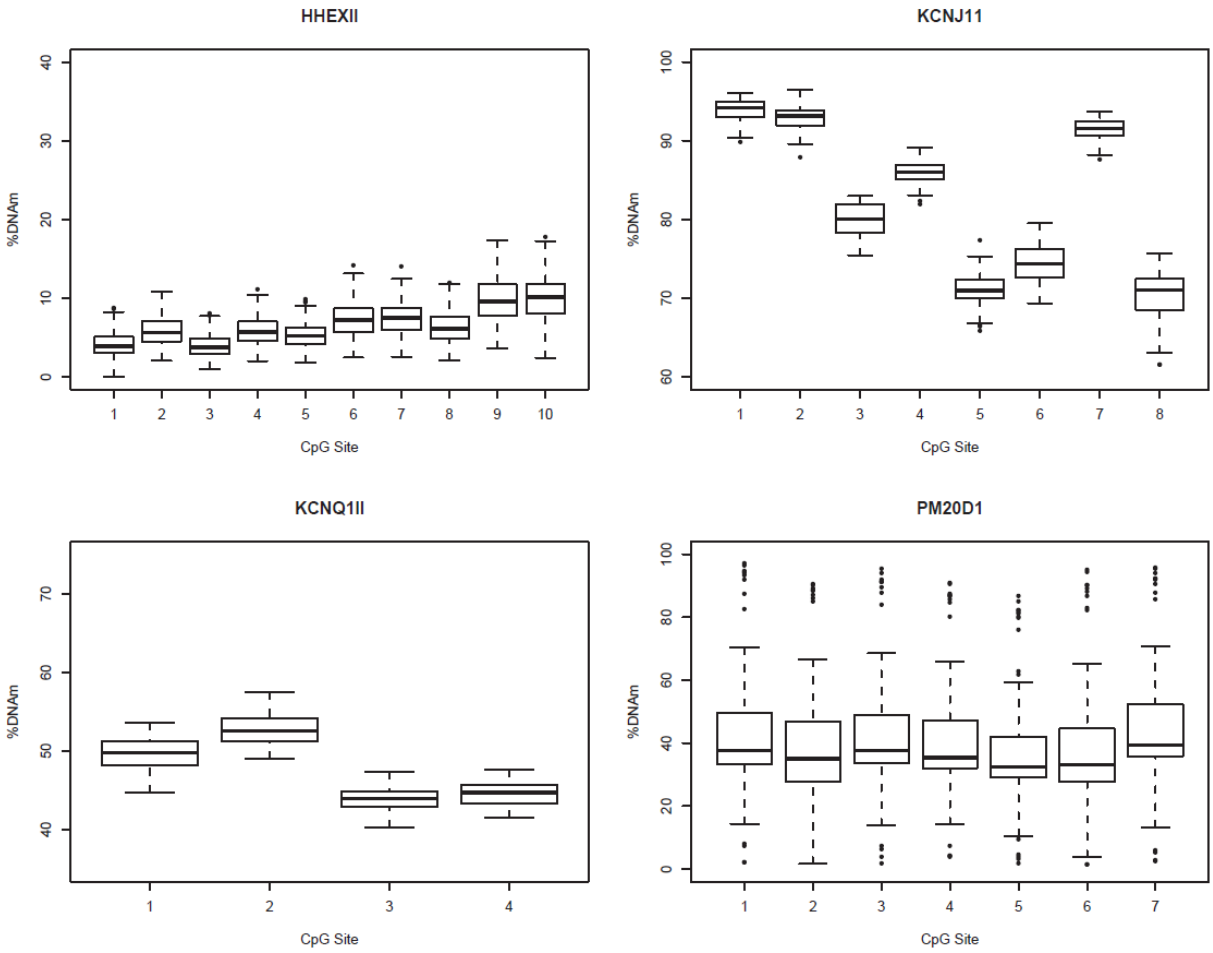


Figure 2. Percent DNA methylation in whole blood samples.
Percent DNAm (y-axis) in whole blood DNA per CpG site (x-axis) in the HHEXII (n=169), KCNJ11 (n=54), KCNQ1II (n=49) and PM20D1 (n=59) regions. Data for each region is depicted in a separate boxplot (the boxplot is defined in detail in the materials and methods, pg. 26). Note the varying scale on the y-axis per plot.

The inter-individual variability in whole blood DNAm level could in theory, at least partly, be explained in terms of differential white blood cell composition between the studied individuals. The numbers of white blood cells, neutrophils, lymphocytes, monocytes, eosinophils and basophils, were counted using an automated cell counter. In the samples used for this study (n=211), the cell counts varied considerably between individuals. The relative standard deviation for the neutrophil proportion was 14.7% (56.8% ± 8.3%), lymphocytes 25.8% (29.7% ± 7.7%), monocytes 30.8% (9.4% ± 2.9%), eosinophils 61.1% (3.6% ± 2.2%) and basophils 96.9% (0.5% ± 0.5%). An analysis was conducted to test whether the variation in proportional numbers of specific white blood cell types were associated with variation in measured DNAm levels. Statistical analyses indicated that a significant proportion of the DNAm level variability in the HHEXII region could be explained by this factor, or up to 36% ($p<0.0001$, **table 2**). Additionally, DNAm levels in the KCNJ11 region were suggestively associated with the basophil proportion, explaining 3% of the variation ($p=0.04$, **table 2**), perhaps only owing to multiple testing. None of the five white blood cell ratios were significantly associated with measurement outcomes in the KCNQ1II and PM20D1 regions (**table 2**). These results were minimally affected by outliers and missing values (in total, n=10 for the HHEXII data, n=10 for KCNJ11, n=1 for

KCNQ1II and n=0 for PM20D1), except for the association between DNAm in KCNJ11 and the basophil proportion, which was not significant when this data was included.

Table 2. **Proportion of variation in measured DNA methylation level accounted for by cellular heterogeneity.**

	Variance explained by cell proportion (%)				
Region	**Lymphocytes**	**Monocytes**	**Neutrophils**	**Eosinophils**	**Basophils**
HHEXII	36**	0	27**	0	0
KCNJ11	0	0	0	0	3*
KCNQ1II	1	0	0	0	0
PM20D1	0	0	0	0	0

*p<0.05, **p<0.0001

Note: the small discrepancy between the results for HHEXII in the table presented here, and in the article manuscript (see end of the thesis) stems from transformation of the data, which was not conducted prior to submission of the manuscript.

4.1.2 Comparison of DNAm levels in cell fractions

To examine if the variability in measured methylation level at different CGIs in whole blood was attributable to differential methylation in the white blood cell types comprising whole blood, whole blood samples from 20 individuals were fractionated into mononuclear cells (MNCs, containing lymphocytes and monocytes) and polymorphonuclear cells (PMNCs, containing neutrophils, basophils and eosinophils), DNA isolated and methylation levels measured at the four regions in each fraction (DNA was also isolated from whole blood for these individuals, methylation levels measured in each of the four regions and the data included in the analysis described in section 4.1.1 above).

A comparison of the methylation levels measured in MNCs and PMNCs indicated higher average methylation in MNCs in 21 of the 29 CpG sites investigated in total. Paired Wilcoxon signed rank test revealed that 18 of these CpGs were significantly differentially methylated in the two different cell fractions, located in the HHEXII, KCNJ11 and KCNQ1II regions (**figure 3**). The average absolute difference between the two cell fractions was highest in the HHEXII region. All ten CpGs studied in this region showed significantly higher methylation in MNCs, the average DNAm levels differed by 3.4-15.7 pp per CpG site (corresponding to ~2.3-4.0 fold higher methylation levels in MNCs). Methylation in the KCNJ11 region was also significantly higher in MNCs. The average difference was more moderate than in the HHEXII region, but nonetheless significant in 7 out of 8 CpGs, ranging between 0.4-6.1 pp per site (corresponding up to ~1.1 fold higher methylation levels). In the case of the KCNQ1II region, only one CpG site was significantly differentially methylated between the two fractions, where average levels differed by 1.2 pp between cell fractions, corresponding to ~2% higher methylation levels in MNCs. DNAm levels did not differ significantly between cell fractions in the PM20D1 region.

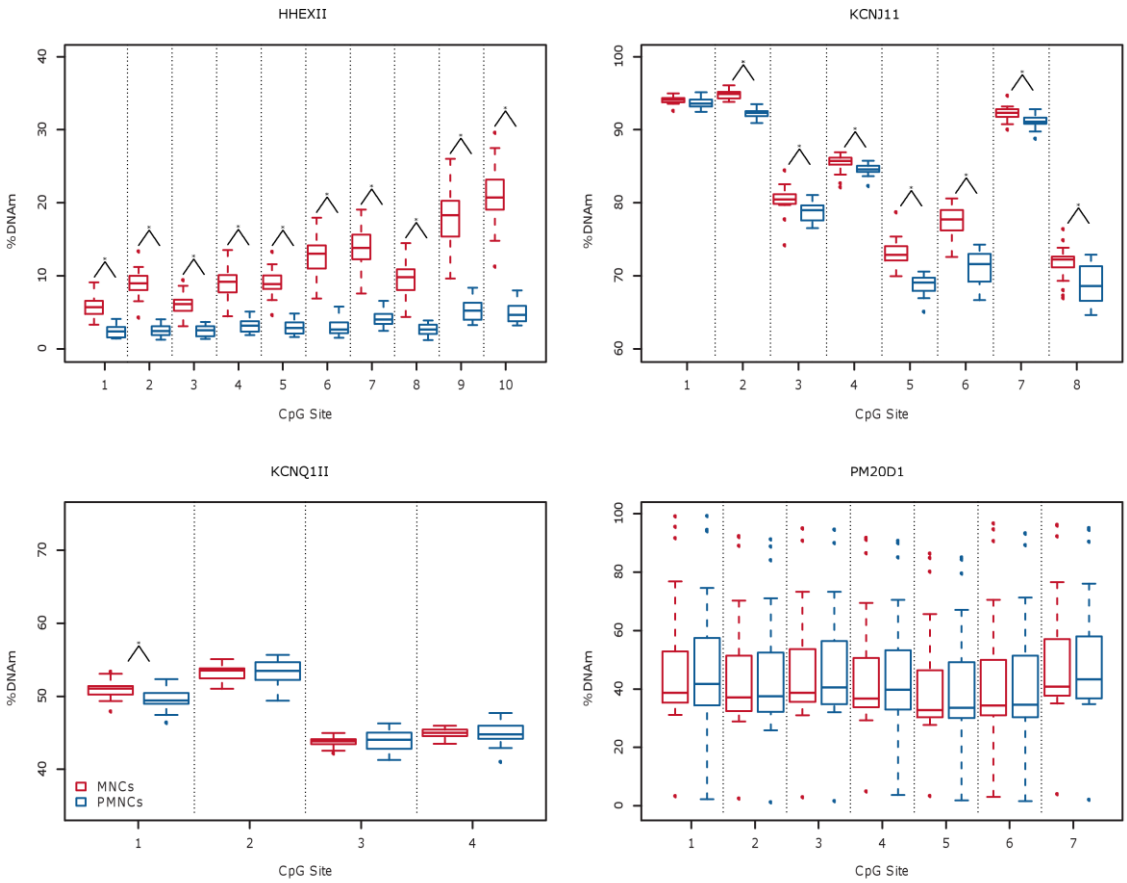


Figure 3. Percent DNA methylation in mononuclear and polymorphonuclear cells.
Percent DNA methylation (y-axis) in mononuclear and polymorphonuclear cells (MNCs and PMNCs, n=20) per CpG site (x-axis) in the HHEXII, KCNJ11, KCNQ1II and PM20D1 regions. Data for each region is depicted in a separate boxplot where measurements for MNCs are shown in red and for PMNCs in blue. The dotted lines separating the boxes indicate that at each CpG site a pair of data is being compared (i.e., for MNCs and PMNCs). Significantly ($p<0.05$) differentially methylated CpG sites (MNCs versus PMNCs DNAm) are indicated with an asterisk. Note the varying y-axis scale.

4.1.3 Analysis testing for correlation between DNAm in blood cell fractions

The results presented in figure 3 suggest that the methylation patterns between cell fractions are highly similar overall. To quantify this observation the correlation between methylation levels for the two different fractions was analyzed.

The correlation was very high in all regions, irrespective of whether methylation levels differed between cell fractions or not (**figure 4**); Spearman's ρ was 0.72 for the HHEXII region, 0.93 for KCNJ11, 0.80 for KCNQ1II, and 0.95 for PM20D1.

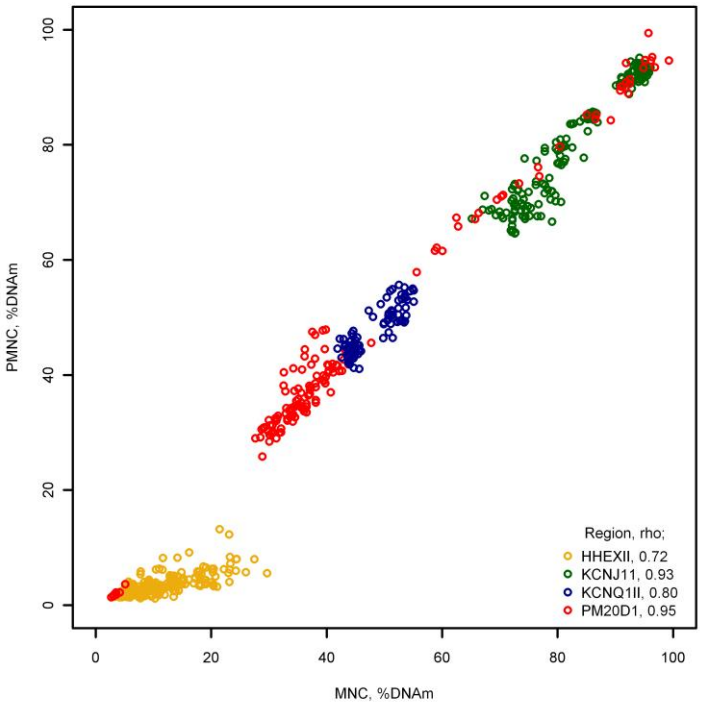


Figure 4. Correlation between DNA methylation in mononuclear and polymorphonuclear cells. Comparison of DNA methylation levels measured in two cell fractions (n=20), mononuclear cells (MNCs) and polymorphonuclear cells (PMNCs). Percent methylation in PMNCs (y-axis) is plotted against percent methylation in MNCs (x-axis). Each dot represents the two measurements for a single CpG per individual. The Spearman's ρ for correlation between measurements in MNCs and PMNCs for each assay is shown in the legend.

4.2 Comparison of blood cell counts in diabetics and controls

The results presented in section 4.1 demonstrate a potential for confounding in the analysis of DNAm data due to cellular heterogeneity in whole blood. For this reason, an analysis was conducted to test whether the proportional number of specific cell types differed in individuals with type 2 diabetes compared to individuals without the disease.

4.2.1 Comparison of proportional cell counts in diabetics and controls

The proportional numbers of five cell types (monocytes, neutrophils, lymphocytes, basophils and eosinophils) were compared between individuals with and without type 2 diabetes in the AGES-Reykjavik cohort (i.e., all individuals with T2DM were defined as cases and all individuals without the disease as controls, without regard to any other factors). Of the 5764 individuals recruited, data on both diabetes status and the five white cell counts was available for 5688, 781 of whom were diabetic. The statistical comparison was unadjusted, conducted using Student's t-tests.

The average proportional number of neutrophils was higher in individuals with type 2 diabetes compared to controls (it was 59.5% in cases and 57.6% in controls, $p = 5 \cdot 10^{-10}$), the average proportion of lymphocytes was lower (it was 27.6% in cases and 29.3% in controls, $p = 3 \cdot 10^{-9}$) and similarly the average monocyte proportion was lower in diabetic individuals (it was 8.8% in cases and 9.2% in controls, $p = 2 \cdot 10^{-5}$, **figure 5**). A subtle difference in average basophil and eosinophil proportions was also detected between cases and controls, but it was not statistically significant (p=0.4 and p=0.2 respectively).

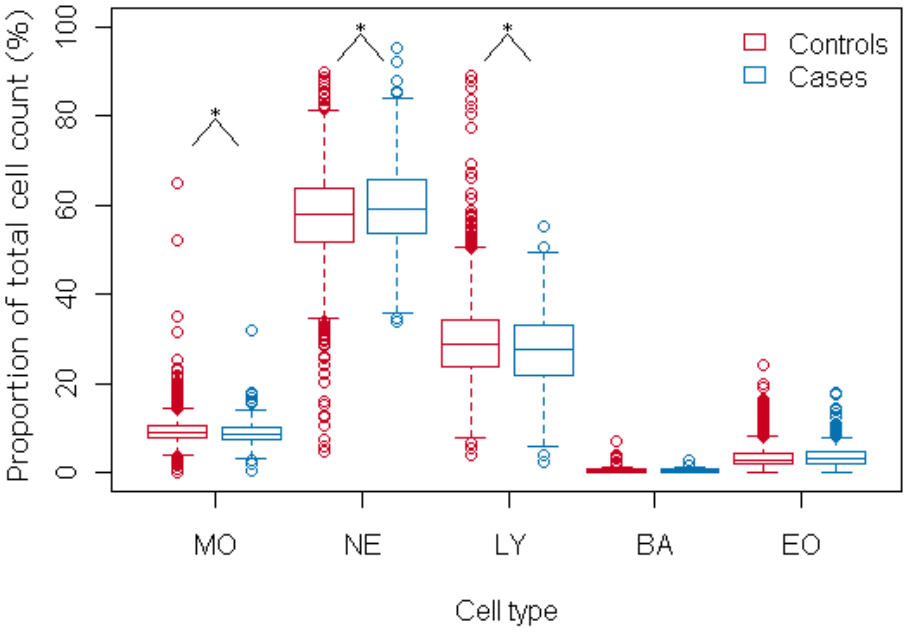


Figure 5. Comparison of proportional numbers of five white blood cell types in diabetic and non-diabetic individuals.
The proportional number of total cell counts (y-axis) for five white blood cell types (x-axis) were compared in individuals with (cases, n=781) and without (controls, n=4907) type 2 diabetes. On the x-axis, names are abbreviated, where MO is monocytes, NE is neutrophils, LY is lymphocytes, BA is basophils and EO is eosinophils. An asterisk indicates where significant (p<0.05) differences in proportional numbers were detected between the cases and controls.

4.3 Investigation of DNA methylation in type 2 diabetes genetic risk loci

4.3.1 A comparison of DNAm levels in six regions between diabetics and controls

As a first stage of this study (referred to hereafter as the "discovery stage") DNAm levels in whole blood from type 2 diabetic individuals (referred to as diabetics or cases) were compared to controls (individuals with normal fasting glucose (NFG), also referred to as non-diabetics) to attempt to identify aberrantly methylated regions in individuals with diabetes. Six regions in three genes were investigated; the HHEXI (4 CpG sites), HHEXII (10 CpGs) and HHEXIV (4 CpGs) regions in the *HHEX* gene, the KCNJ11 region (8 CpGs) in the *KCNJ11* gene and the KCNQ1I (5 CpGs) and KCNQ1II (4 CpGs) regions in the *KCNQ1* gene. Each region was investigated in an independent set of DNA samples (there was only a single sample overlap between KCNQ1I and KCNQ1II, otherwise no sample overlapped between any regions) from 13-23 individuals each, roughly half diabetic and half controls, selected randomly from the AGES-Reykjavik cohort.

DNAm levels differed between CpG sites within each region but were generally close to 0% in the HHEXI and KCNQ1I regions (ranging between ~0-5%), low in the HHEXII region (ranging between ~0-20%), intermediate in the KCNQ1II region (ranging between ~50-70%) and intermediate to high in the KCNJ11 and HHEXIV regions (ranging between ~70-100%, **figure 6**). Inter-individual variation was close to none in the HHEXI and KCNQ1I regions (standard deviation per CpG site ranged between 0.4-0.7 pp and 0.1-0.5 pp respectively), very low in the KCNQ1II region (ranging between 1.0-1.8 pp) but higher in the other three regions (ranging between 1.5-3.8 pp in HHEXII, 3.7-8.3 pp in HHEXIV and 0.8-3.0 pp in KCNJ11, **figure 6**). Comparison of average DNAm levels in diabetics and controls indicated that for four of the regions tested, HHEXI, HHEXIV, KCNJ11 and KCNQ1II the inter-individual variation was not explained in terms of T2DM phenotype as the average difference per CpG site was either close to none and/or was alternately positive and negative across the respective CpG sites (the difference between the two groups is better displayed on **figure 14** in the appendix, pg. 58, than on figure 6). However, in the HHEXII and KCNQ1I regions, average DNAm levels were lower in diabetics compared to controls in all respective CpG sites (**figure 14**). The difference ranged between 0.9-3.0 pp in HHEXII and between 0.1-0.5 pp in KCNQ1I per CpG site. Statistical analysis of the data indicated that average DNAm levels across the respective CpG sites differed between cases and controls in HHEXII and KCNQ1I (p=0.0445 and p=0.0393 respectively), but not in the other four regions (**table 3**).

Table 3. Association between DNAm and T2DM.

Region	n (T2DM / NFG)	β*	p*
HHEXI**	23 (9/14)	-0.26	0.5515
HHEXII**	19 (11/8)	-2.85	0.0445
HHEXIV	16 (9/7)	-1.41	0.5627
KCNJ11	20 (6/14)	0.05	0.9411
KCNQ1I**	13 (8/5)	-1.03	0.0393
KCNQ1II	14 (8/6)	-0.36	0.6328

*From an unadjusted model
**DNAm values were transformed, see materials and methods for details.

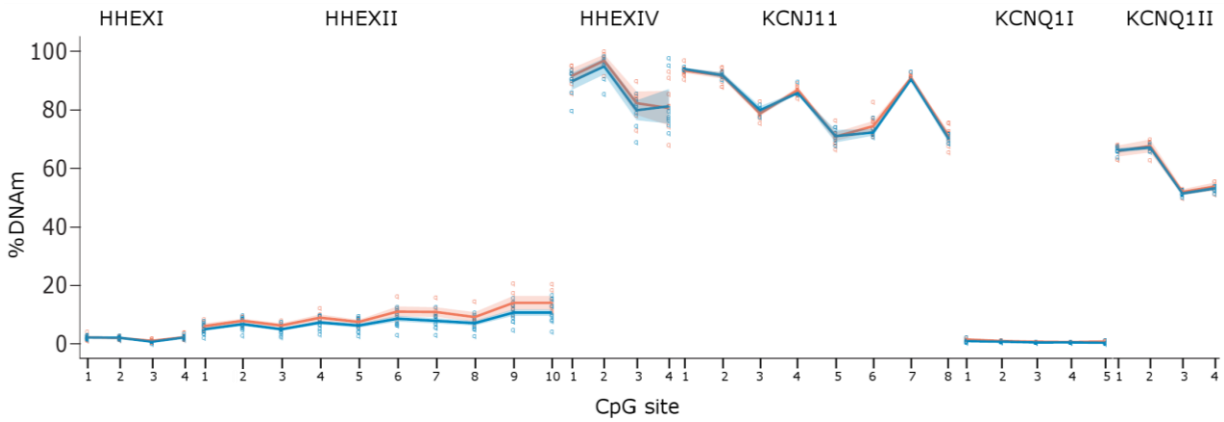


Figure 6. Comparison of DNAm in diabetics and controls in six regions in three genes. DNAm levels (y-axis) per CpG site (x-axis) in six regions in three genes (indicated at the top of the figure) are compared between diabetics (blue) and controls (red). The dots indicate individual measurements for each sample, the lines the mean methylation, and the ribbons the 95% confidence intervals per CpG site. The lines and ribbons extend between the CpG sites for better visualization of the data and are not intended to imply any interrelation between measurements. The same plots are shown in an enlarged form in the appendix (**figure 14**).

An assay was prepared to investigate DNAm in CpG sites located directly adjacent to the HHEXII region (~150 base pairs downstream, region was termed HHEXIII). DNAm levels were compared between diabetics (n=11) and controls (n=8), as before, in an independent set of samples.

Methylation levels in the HHEXIII region ranged between ~0-15% across the three CpG sites investigated, and the standard deviation between 1.4-2.3 pp. Average DNAm levels were lower in individuals with type 2 diabetes in the region; the difference ranged between 0.9-1.8 pp per CpG site. However, unadjusted mixed model statistical analysis of DNAm levels across the three CpG sites indicated that the difference was not significant (β=-1.72, p=0.1119).

4.3.2 Comparison of DNAm levels in the HHEXII region in a larger set of samples

To test whether the result that average DNAm levels in the HHEXII region differ between cases and controls in the HHEXII region was robust, and to be able to account for covariates and confounding factors that are associated with type 2 diabetes (either the disease state itself, or with the development of T2DM) and DNAm levels (age (99,100), gender (99,101), BMI (66,102) and cellular heterogeneity), the analysis was repeated in a larger set of samples. Samples were again selected from the AGES-Reykjavik and REFINE-Reykjavik cohorts from male and female (~55% male ratio) adults (age range was 56-96 years). After exclusion of individuals with missing values and outliers, DNAm data was analyzed for 378 individuals, of which 164 controls and 214 cases. This analysis is referred to hereafter as “the replication stage”.

The results indicated that DNAm levels ranged between ~0-20% in the HHEXII region in both diabetics and controls (**figure 7**) and that the overlap of measured levels in the two groups was very high. Nonetheless, in agreement with the previous results, lower average DNAm levels were observed in diabetic individuals compared to controls (**figure 7**), although more moderate than in the discovery stage the difference per CpG site ranged between 0.5-1.1 pp. The difference was statistically significant both in an unadjusted model and after adjusting for the potential confounders (**table 4)**. In

the full model (model 3), cellular heterogeneity (% lymphocytes of total white blood cell count) was significantly associated with DNAm levels but age, BMI and gender were not. The results from the unadjusted model (model 1) and the model only adjusting for cellular heterogeneity (model 2) allow for estimation of the confounding effect caused by this covariate; the unadjusted beta estimate for DNAm difference in diabetics versus controls was 67% higher than the beta estimate from the model adjusting for cellular heterogeneity. These results (applies to all results discussed in the paragraph) were minimally affected when individuals were included that had been identified as outliers (n=3) or had missing DNAm values (n=22)

Table 4. Association between DNAm in HHEXII and T2DM.

	β	p
model 1*	-0.90	0.0010
model 2**	-0.54	0.0090
model 3***	-0.62	0.0047

*Unadjusted.

**Adjusting for cellular heterogeneity (using the percentage of lymphocytes of total white blood cell counts).

***Adjusting for cellular heterogeneity, age, gender, and BMI.

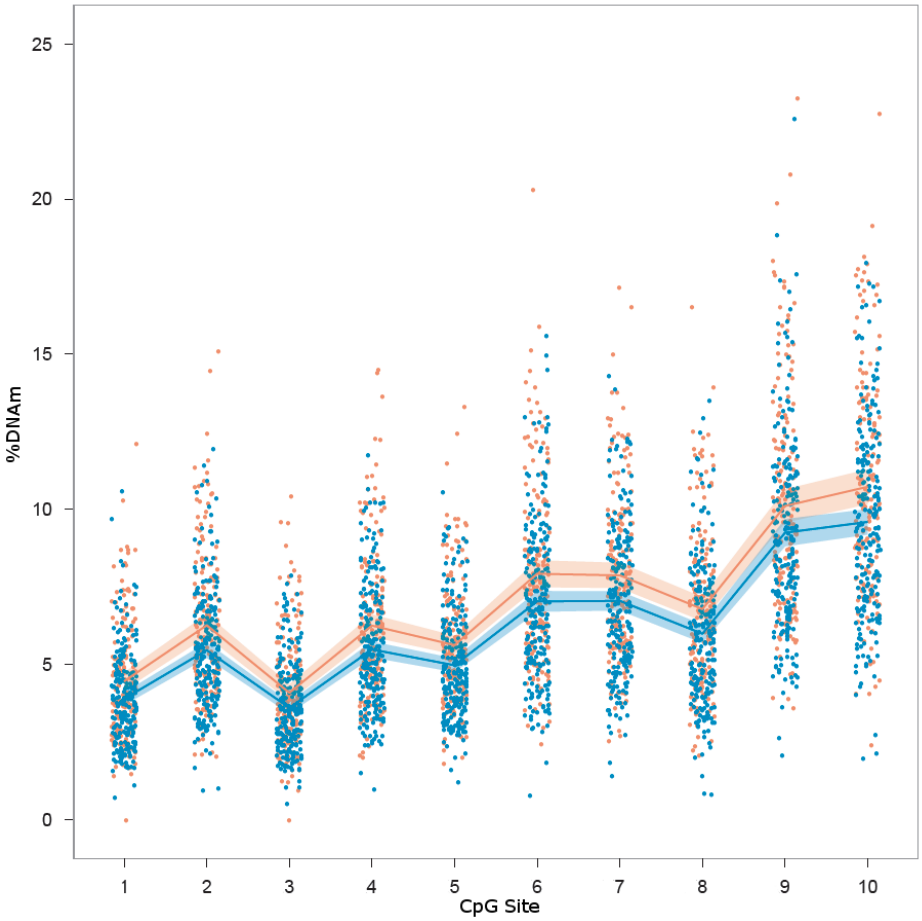


Figure 7. Comparison of DNAm levels in HHEXII in diabetics and controls. DNAm levels (y-axis) per CpG site (x-axis) in the HHEXII region is compared between diabetics (blue) and controls (red). The dots indicate individual measurements for each sample and the lines and "ribbons" the mean methylation and the 95% confidence intervals respectively, per CpG site. The dots are "jittered" for better visualization.

4.3.3 Analysis testing for association between HHEXII DNAm and HOMA indices

The genetic association between the *HHEX* locus and T2DM is thought to be mediated through effects on β cell function (71,103–105) rather than insulin sensitivity. An analysis was conducted to test whether DNAm in this locus was associated with β cell function or insulin sensitivity. A subset of the

samples that were used in the analysis presented in section 4.3.2 had available data on fasting serum glucose and insulin levels (n=244), enabling an estimation of the β cell function and insulin sensitivity of these individuals through the HOMA-β and HOMA-IR indices. The association between these indices and DNAm in HHEXII was therefore tested.

The variation in both HOMA-β and HOMA-IR values was very high in the 244 samples analyzed; the relative standard deviation was 87.3% for HOMA-β (average and standard deviation were 104% ± 90.8%) and 151% for HOMA-IR (4.88 ± 7.37). Both indices were log transformed prior to statistical analysis of the data due to their highly skewed distribution. The statistical analysis revealed a significant positive association between DNAm in the HHEXII region and HOMA-β ($p=0.0242$, $\beta=0.47$) after adjusting for cellular heterogeneity. No association was however detected between DNAm in the region and the HOMA-IR index ($p=0.9385$, $\beta=-0.01$, in the same model).

5 Discussion

5.1 Heterogeneity in white blood cells has potential to confound analyses of DNA methylation data

5.1.1 Summary of aim and results

Studies on DNAm using whole blood DNA frequently do not control for inter-individual variation in the cellular population from which the DNA is derived, the white blood cells; lymphocytes, neutrophils, eosinophils, basophils and monocytes. This has been criticized due to hypothesized potential for confounding effect when cellular heterogeneity is present in conjunction with cell type specific DNAm (91). Here, the aim was to study this hypothesis in a more comprehensive manner than has been done previously by first testing for an association between whole blood DNAm levels and cellular heterogeneity, and second to test whether differential methylation in two cell fractions might underlie the observed association, if any.

The results indicated that indeed a locus specific association between measured DNAm levels and cellular heterogeneity in whole blood can be observed. Further, significant differences in locus specific DNAm levels were observed between two blood fractions, MNCs and PMNCs, suggesting that it could be the underlying cause of the observed association between DNAm levels and white blood cell counts. Finally, in all loci tested DNAm in MNCs and PMNCs was highly correlated independent of differential methylation levels in these fractions.

5.1.2 Interpretation of the results, their comparability with other studies and the hypothesis under question

The results indicated that up to 36% of the inter-individual variation in whole blood DNAm in the HHEXII region was attributed to cellular heterogeneity, suggesting that a considerable confounding can affect measured levels of whole blood DNAm due to differences in the cellular population. A weak association between DNAm in the KCNJ11 region and the basophil ratio of small effect size (3%) was also detected. However, given a very small proportion of basophils, as well as a suggestive p-value of 0.04, this result does not convincingly suggest this as an additional example of cellular heterogeneity confounding methylation measurements. Additionally, given the number of tests performed, a correction for multiple testing may be appropriate. Any such correction would presumably deem the association between the basophil ratio and methylation in the KCNJ11 region not significant, while even a conservative correction (e.g., Bonferroni) would not affect the significance of the association between cell fractions and DNAm in the HHEXII region. No effect on measurements for the KCNQ1II and PM20D1 regions was observed, suggesting that this type of confounding does not affect DNAm measurement outcomes universally throughout the genome, but may be locus-specific. These results are in concordance with a previous study (106) where out of a total of 16 loci assayed, only a single locus was affected in similar magnitude as the HHEXII region. Together, these studies indicate that while measured DNAm levels in a substantial proportion of loci may not be affected by cellular heterogeneity, measurement outcomes in some loci may be.

DNAm levels in PMNCs and MNCs differed significantly in individual CpG sites in three out of four regions examined; i.e., in all CpG sites analyzed in the HHEXII region, 7 of 8 CpG sites analyzed in the KCNJ11 region and 1 of 4 CpG sites analyzed in the KCNQ1II region but not in the PM20D1 region. The gross difference observed in the HHEXII region may reflect the fact that the *HHEX* gene is differentially expressed in the various blood cells (107–109). Just as in whole blood DNAm measurements, this analysis may have been confounded by cellular heterogeneity because PMNCs and MNCs both consist of groups of cells. However, the fractionation split up the two white blood cell groups that affected whole blood DNAm measurements and their numbers are so dominant relative to the other groups that the analysis is likely to be minimally affected. Kerkel *et al.* have previously studied methylation in these fractions, and identified multiple differentially methylated loci (110). Their analysis was however not described in detail. Nonetheless, together these studies indicate that differential methylation between white blood cell types may be relatively common.

Analysis of DNAm both in whole blood and blood fractions has allowed evaluation of the hypothesis that measured DNAm levels in whole blood can be confounded by cellular heterogeneity due to differential methylation levels in the various white blood cell types. Differential methylation was observed between cell fractions in the HHEXII, KCNJ11 and KCNQ1II regions and not in the PM20D1 region, but a significant effect due to cellular heterogeneity on whole blood DNAm measurement outcomes was only detected (convincingly) for the HHEXII region. However, the difference in DNAm between fractions was very moderate in the KCNJ11 and KCNQ1II regions and in the KCNQ1II region only one of four CpG sites was differentially methylated. It is therefore possible that the effect of cellular heterogeneity on measurement outcomes for the KCNJ11 and KCNQ1II regions, if any, is subtle, and thus undetectable by the methods employed in the study. It can therefore be argued that the results presented here support the hypothesis, and that they suggest a need to control for cellular heterogeneity in the analysis of methylation in blood cells.

5.1.3 Discussion of ideas to address the confounding effect of cellular heterogeneity

Since the confounding effect would only be observed when both the genomic region of interest is differentially methylated amongst white blood cell types, and when there is blood cell count heterogeneity in the individuals being compared, controlling for this problem may be addressed in different ways depending on available data. Differences in white blood cell composition may be assessed, and controlled for when applicable, if blood cell counts for the individuals under investigation are available. In that scenario, subjects can alternatively be paired with controls that are concordant in terms of cellular composition prior to the analysis. Furthermore, whole blood can be fractionated to assess possible differential methylation in the area of interest. This may be done with the Ficoll medium method used here which is relatively easy to perform, but due to heterogeneity in the fractions, as noted previously, this approach may not be sufficient to address the problem. Finally, referring to the literature may be advisable to assess the risk of altered blood cell counts in the groups of individuals under study. For example white blood cell counts have been shown to be associated with the development of cancers (111) and coronary heart disease (112). This raises the issue that

whenever there is a difference in cell fractions associated with disease, an adjustment for blood cell proportions could be essential for better controlled analyses.

The different approaches may cause inconsistent results, and therefore it may be important to standardize methods for this correction. As has been discussed previously (106), adjusting for white blood cell counts can be achieved with standard statistical approaches. Such an approach may be well suited for that purpose since such data is presumably readily available at many laboratories conducting experiments on whole blood DNA. This could be achieved in two ways: One is to use multiple variables accounting for the absolute number of each cell type (commonly five; neutrophils, lymphocytes, monocytes, basophils and eosinophils) or alternatively use a single variable accounting for the proportion of one cell type. Using a single variable is more appealing because the other option would reduce the number of degrees of freedom. However, to be able to correct for the confounding effect of cellular heterogeneity in statistical models by using a variable accounting for the proportional number of one cell type, there needs to be a correlation between methylation levels in the different types of white blood cells. Results from this study indicate that in the analyzed regions, methylation patterns across the corresponding CpG sites within a region are very similar between the different cell types irrespective of demonstrable differences in the cell specific absolute methylation levels. The study therefore suggests that use of a single variable to account for the proportional number of a single cell type (e.g., neutrophils or lymphocytes) in statistical analyses might be sufficient to correct for the confounding effect of cellular heterogeneity on DNAm measurements conducted using whole blood DNA.

In line with the suggestions above, proportional cell counts were compared between diabetics and non-diabetics in the present study to assess the need to control for cellular heterogeneity in the subsequent analyses testing for association between DNAm and T2DM. The results indicate that average proportional numbers of neutrophils and lymphocytes (together comprising about 85% of the total white blood cell count) differ significantly between cases and controls (section 4.2). Although a more detailed analysis needs to be conducted to verify these results (e.g., adjusting for potential confounders), they, in conjunction with the data indicating an association between proportional cell counts and DNAm levels, suggest a need to control for cellular heterogeneity in the analysis of DNAm data in association with type 2 diabetes. Finally, the study provides an example of how this confounding factor can affect analysis of DNAm data: When comparing DNAm differences between diabetics and non-diabetics in HHEXII in the replication stage, the unadjusted β-value was highly overestimated, specifically 67% higher than the estimate from the analysis controlling for cellular heterogeneity (section 4.3.2).

5.1.4 Other considerations

The findings from the present study may not only be relevant for methylation measurements using whole blood DNA. Other tissues are samples of different types of cells as well, so a similar problem could affect measurements in these tissues. The data presented here indicates that although methylation levels may differ between blood cell types in some loci, the methylation pattern may at the same time be very similar (as indicated by the high correlation between methylation levels). This is in

agreement with previous studies which have shown that different cells and tissues, even from separate germ layers, generally have similar DNAm patterns (17,106,113). If blood cell DNAm measurements could be used as surrogates for methylation in other tissues based on this feature, it might be preferable to use blood.

DNAm levels are sometimes assessed in a global manner, assaying CpG sites across the entire genome. Since the study presented here was conducted in a gene-specific manner the results may not apply to global DNAm measurements. Indeed, in a previous study using LUMA (Luminometric methylation assay) to estimate global methylation, it was reported that no association was detected between methylation levels and white blood cell counts (100). However, as mentioned in the introduction, Wu *et al*. report that global methylation levels in PMNCs, as measured by LUMA, are significantly higher than in MNCs and are not correlated (23). In the same study, results from three other assays for global DNAm showed no association between PMNCs and MNCs methylation levels. It is therefore possible that global methylation measurements are also confounded by cellular heterogeneity. A more detailed analysis, comparing both the association between global methylation levels in whole blood and cellular composition and global methylation levels in cell fractions, such as in the present study, should be conducted in order to extend these observations.

5.1.5 Future directions

The results from the present study call for an analysis of larger number of regions to reveal the full extent of how confounding effects may influence analyses on DNAm conducted using whole blood DNA. It is important to assess whether measured methylation levels at a considerable amount of loci are affected by this effect or whether only a small fraction of loci are affected. Second, it would be of value to study whether methylation of CpGs positioned in certain genes is more prone to be affected by this factor than others (e.g., in genes that are differentially expressed in the different cell subtypes such as HHEX). Finally, it would be interesting to investigate whether certain sequences (e.g., introns, exons, CGIs, CGI shores, transcription start sites or promoter regions) are more likely to be affected by this confounding effect.

5.2 Type 2 diabetes associated DNA methylation identified in genetic diabetes risk locus

5.2.1 Summary of hypothesis, aim and results

It was hypothesized here that aberrant DNAm could be associated with T2DM due to possible inherited or environmentally or lifestyle induced epigenetic aberrations. To address this hypothesis the study aimed to identify aberrant DNAm in individuals with T2DM. Genetic studies, most recently GWA studies, have been very successful in identifying loci containing genetic variants associated with type 2 diabetes. It was hypothesized that these loci may present good candidates for studying epigenetic aberrations that may be associated with the disease, due to the assumption that altered expression caused by epigenetic mechanisms may have similar consequences as alterations in gene products caused by genetic variants (at least if such alterations change the product's activity). DNAm in three such loci were investigated in the present study.

The results indicated that average DNAm levels were significantly lower in individuals with type 2 diabetes compared to controls in at least one of the seven regions investigated, located in the *HHEX* locus (HHEXII). Importantly, the data indicates that the observed difference in DNAm in the HHEXII region is not carried by an association with obesity and that it is not an artifact due to differences in the white blood cell composition of diabetics and controls. Finally, the results indicate that methylation in the region may be associated with T2DM through effects on pancreatic β cell function, perhaps an indication that the aberration affects the same molecular mechanism in the cell as does the genetic T2DM risk variant in the locus.

5.2.2 Interpretation of the results and their comparability with other studies

Of the seven regions investigated in total in the discovery stage, average DNAm levels were observed to differ significantly between cases and controls in two; in the HHEXII and KCNQ1I regions. However, DNAm levels in the KCNQ1I region were close to 0% and the difference was therefore not considered a convincing example of DNAm aberration in T2DM. Using close to 400 samples, average DNAm levels were again observed to be significantly lower in diabetic individuals in the HHEXII region and the result robust after controlling for cellular heterogeneity, age, and gender and not carried by an association with BMI. These findings support the hypothesis that DNAm aberrations may be associated with T2DM. Whether they are environmentally or lifestyle induced, inherited or stochastic cannot be inferred from the results presented here.

Two recent studies by Bell *et al*. and Toperoff *et al*. (86,87) have included an investigation of whole blood DNAm levels in the loci investigated here, *HHEX*, *KCNJ11* and *KCNQ1*, in attempts to identify DNAm aberrations in individuals with type 2 diabetes, but neither study detected the aberration in *HHEX*. Bell *et al*. investigated DNAm in 20 loci previously associated with type 2 diabetes through genetic studies and about a hundred other loci associated with monogenic forms of diabetes and obesity, in imprinted genes an others, using MeDIP-chip (methylated DNA immunoprecipitation followed by analysis on microarray chips) methodology. Their analysis did not reveal any significant methylation differences between diabetic and non-diabetic individuals in any of the loci. Toperoff *et al*.

conducted an epigenome wide association study (EWAS) using a methodology involving methylation sensitive restriction digest followed by microarray chip analysis. Their analysis identified several differentially methylated loci between diabetics and non-diabetics, enriched in genes previously associated with T2DM through GWAS. The epigenome wide analysis was followed by sequencing of 93 selected single CpG sites embedded in the microarray probed fragments, including 2 CpG sites in the *HHEX* locus, 1 in the *KCNJ11* locus and 13 in the *KCNQ1* locus. This CpG-specific analysis revealed that 13 CpG sites in 6 loci were significantly differentially methylated between diabetics and non-diabetics, including three sites in the *KCNQ1* locus, but none in the *HHEX* and *KCNJ11* loci. The discordance between the results presented here and these two studies may be explained both by differences in the specific regions investigated per locus and by the different experimental approaches used. While a small number of consecutive CpG sites were interrogated here at single base resolution, Bell *et al.* measured DNAm at a 100 bp resolution and in the EWAS stage, Toperoff *et al.* measured methylation at an average of about 1000 bp resolution. The subtle methylation differences in a few CpG sites, as observed in the present study, may have been lost by averaging the levels across multiple CpG sites, many of which with no DNAm differences. The differentially methylated CpG sites identified in the *KCNQ1* locus by Toperoff *et al*. in the CpG-specific analysis are located over 100 kb downstream from the regions investigated here and the CpG sites they interrogated in the *HHEX* locus, over 10 kb downstream from the HHEXII region.

The difference in average DNAm levels between cases and controls in the HHEXII region observed in the present study was moderate. This is in agreement both with the results from the study by Toperoff *et al*., where average DNAm levels were reported to differ by ~0.5-4 pp between diabetics and non-diabetics in the CpG sites where significant differences were observed and with other studies using a candidate gene approach on pancreatic islet DNA, revealing differences in average DNAm between diabetics and non-diabetics in the *PPARGC1A* (peroxisome proliferator activated receptor gamma coactivator-1 alpha) (114) and the insulin genes (115) of about 5-10 pp in the CpG sites investigated. In contrast, studies on cancer have revealed gross DNAm differences between cancerous and normal cells in an on/off fashion, i.e., methylation completely abrogated in normally methylated regions or introduced in normally unmethylated regions (see e.g., ref. (60)). Such gross differences are however perhaps nonexistent in association with non-malignant complex diseases. Recent EWA studies on non-malignant complex diseases discussed in the introduction have generally revealed subtle differences between cases and controls (1,62–67). Although such subtle differences may appear to be inconsequential, it is however important to realize that any difference in average DNAm levels between groups, irrespective of its size, corresponds to an absolute change in methylation status (i.e., cytosine is methylated or not methylated) on the allele level for a respective proportion of the alleles investigated.

The results indicate that average DNAm levels in the HHEXIII region, which is located adjacent to the HHEXII region, separated by a 155 bp sequence (i.e., distance between 10th CpG site in HHEXII and 1st CpG site in HHEXIII) containing 8 CpG sites, do not differ significantly between diabetics and non-diabetics. This may suggest that the difference observed in the HHEXII region does not extend downstream (at least not 155 bp), but although not significant, average DNAm levels were lower in

diabetics in the HHEXIII region similar to what was observed in HHEXII and it is possible that the lack of significance of the result stems from insufficient statistical power. To allow assessment of the region specificity of the difference observed in HHEXII, the HHEXIII region needs to be investigated further, and additionally, a region directly upstream from HHEXII has to be examined.

The results presented here indicate a positive association between DNAm in the *HHEX* locus, specifically in an intragenic CGI in the *HHEX* gene (HHEXII region), and HOMA-β but no association with HOMA-IR. Identical to the genetic T2DM risk variants in the HHEX locus, the association between DNAm in the region and T2DM may therefore be mediated through effects on β cell function. It can therefore be speculated that the aberration in DNAm levels in the HHEXII region may affect the same molecular mechanism as does the genetic variant in the locus through which the association with type 2 diabetes is caused.

This result may strengthen the hypothesis that was used as a basis for selecting loci for this study. The study was not designed to assess whether selecting loci for the purpose of identifying epigenetic aberrations on the basis of this hypothesis is more effective than using any other criteria and the results presented here cannot be used for such an assessment because they are limited to a small number of loci and no comparison with other approaches was included. Results from the EWA study by Toperoff *et al*. do however suggest that this approach may be more effective than others: In an analysis assessing whether the distribution of regions exhibiting differential methylation between diabetics and non-diabetics across the genome was concentrated at specific genomic locations, e.g., in T2DM risk loci, in loci containing genes involved in metabolic pathways that have not been associated with T2DM and various other gene-ontology terms (not defined further in the paper), a significant enrichment of such regions was observed in T2DM risk loci relative to the genome while they were not enriched in any other location tested. In addition, in some of the EWA studies on non-malignant complex diseases discussed in the introduction an enrichment of differentially methylated regions was often observed in loci with previously identified genetic association with the respective disease (1,62–67).

5.2.3 Consideration about temporal origins and speculations about applied relevance

A cross-sectional comparison of DNAm levels cannot distinguish between differences that are present prior to the disease onset, and thus possibly causal in the disease progression, and those that arise after disease onset, possibly due to effects of the disease state itself or the disease treatment, such as drugs. This is a well defined issue in epidemiological studies, but is in stark contrast to genetic studies, where such comparison reveals truly predisposing associations. To determine the temporal origin of the DNAm difference observed in the HHEXII region, and thus whether it can potentially be predisposing to diabetes, a longitudinal comparison of non-diabetic individuals that later develop T2DM to those who do not has to be conducted. An alternative approach has recently been suggested by Relton and Davey Smith (54), to adjust the principle of Mendelian Randomization for epigenetic studies (the method was named “Genetical Epigenomics”). The approach requires establishing an association between genetic variants that are associated with the disease in question and DNAm in

the region where methylation differences have been observed, and therefore cannot be used here. In the EWA study report by Toperoff *et al*., an analysis was conducted of the temporal origin of the DNAm difference observed between cases and controls in a single region, specifically a single CpG site in the *FTO* locus. Non-diabetic individuals that later developed IFG or T2DM were reported to have lower average methylation levels in the CpG site compared to those that did not develop the disease (in 13.1 year follow-up on average), similar to diabetics compared to controls. Such pre-disease manifestation differences have also been reported prior to the onset of type 1 diabetes, but have not been assessed in other non-malignant complex disease epigenome studies (62). Until the temporal origin of the DNAm difference observed between diabetics and non-diabetics in the HHEXII region presented here has been established, its possible functional implications and use as a biomarker for disease prediction can only be speculated under the condition that the results can be replicated in such a setting.

While average DNAm levels were observed to differ significantly between cases and controls in the HHEXII region, the range of DNAm levels was observed to overlap considerably between the two groups. If it is assumed that a similar scenario is present at baseline between individuals that will develop the disease and those that will not, it can be speculated that the power of DNAm levels in HHEXII as a biomarker for T2DM risk prediction cannot be absolute, i.e., will develop, will not develop disease. Whether it will exceed the prediction power of genetic risk variants, which similar to the above scenario non-diabetic individuals commonly carry but are present in a higher frequency in diabetic individuals than controls, cannot be speculated on grounds of the results presented here. In their EWAS report, Toperoff *et al*. included a receiver-operating characteristic analysis of cross sectional DNAm data for a single CpG site in *FTO.* On basis of the results, they suggested that DNAm level in this single CpG site in the *FTO* locus was more closely related to T2DM than 18 most established genetic T2DM risk variants combined (area under curve was 0.638 versus 0.6). This observation needs to be extended in a longitudinal setting.

Although the data from the present study do not allow for claims of definitive functional implications, it is interesting to speculate about their possible biological relevance. The CpG sites interrogated in the HHEXII region are positioned in the gene-body and previous studies have revealed a positive association between gene-body methylation and gene-expression (41–43). A study on Hhex knockout mice suggests that the gene product is essential for normal embryonic development of the pancreas (116) and it has therefore been suggested that the association between the genetic variation in the *HHEX* locus and diabetes arises from alterations in pancreatic development caused by the causal allele (105). The data presented here may present an additional mechanism leading to the same outcome, whereby aberrantly reduced DNAm would cause a reduction in *HHEX* gene expression which may cause aberrant development of the pancreas and thus reduced pancreatic β cell function in the adult individual. This above scenario is dependent on that the DNAm difference observed in whole blood is present in the pre-pancreatic tissue during development.

5.2.4 Other considerations

The present study was conducted using DNA from whole blood, rather than the target tissue, pancreatic β cells. Since epigenetic marks can be tissue specific (9,16–19), this may limit the biological inferences to be made based on the data. However, it is possible that epigenetic aberrations in a particular tissue are observable in surrogate tissues such as blood. For example, LOI in *IGF2* is found in both white blood cells and in the colon (117).

Using blood derived DNA rather than pancreatic DNA has at least three important advantages. First, the subtle differences in DNAm observed in this and other studies on non-malignant complex diseases require large numbers of samples to be detected, which is much easier to acquire through the use of whole blood DNA than pancreatic β cells. Second, pancreatic β cells are only available post-mortem or in pancreatic resections and could therefore only be used in a case-control setting, which limits the clinical importance of any findings while DNA from whole blood can be taken on multiple time points to establish the temporal origins of observed differences in DNAm. It may therefore be sensible to use the experimental approach applied in the present study for the investigation of epigenetic marks in non-malignant complex diseases as a locus discovery step, and subsequently try to replicate the results in the target tissue to assess their potential mechanistic impact. Finally, for studies where the aim is to identify DNAm aberrations for use as biomarkers for the disease, pancreatic DNAm is not useful because it cannot be obtained for such purposes, while using whole blood DNA is ideal.

5.2.5 Future directions

The experiment needs to be replicated in an independent cohort, to provide independent biological confirmation of the results. This was beyond the scope of the present study, a weakness that limits its confidence. As has been discussed above, it is of crucial importance to establish the temporal origin of the DNAm difference observed between diabetics and non-diabetics in the HHEXII region. The same data could be used to investigate the potential use of DNAm in the region as a biomarker for assessing risk for development of the disease. From a functional perspective, a comparison of DNAm levels in HHEXII in pancreatic tissue of diabetic and non-diabetic individuals needs to be conducted. If the DNAm difference is observed in the target tissue, a study on the association between DNAm levels in the region and *HHEX* gene expression in the pancreas is presumably the first step towards elucidating the functional importance of the finding. A second step might involve studying effects on HHEX protein levels, and/or transcription of genes regulated by the transcription factor.

The present study was limited to investigating small regions within much larger loci. Each of the three loci contains a number of genes and all could be as valid targets as the ones investigated here. For example, in addition to containing the *HHEX* gene, the *HHEX* locus contains the *IDE* (insulin degrading enzyme) gene, whose product degrades amylin that may cause β cell dysfunction through amyloid deposition (118). In addition, at present, about 50 genetic variants have been associated with T2DM risk (79) and only three loci were considered here. Although the studies by Bell *et al*. and Toperoff *et al*. investigated these loci, their analysis failed to detect the difference in HHEXII observed

here. The results presented here may therefore warrant a more detailed analysis, focused on identifying DNAm aberrations in T2DM genetic risk loci.

6 Conclusions

6.1 Heterogeneity in white blood cells has potential to confound DNA methylation measurements

The results indicate region-specific differential DNA methylation between white blood cell sub-types and region-specific association between DNAm levels measured in whole blood and cellular heterogeneity. Together these results allowed evaluation of the hypothesis that measured DNA methylation levels in whole blood can be confounded by cellular heterogeneity due to differential methylation levels in the various white blood cell types. In the region where gross white blood cell type specific DNA methylation differences were detected (HHEXII), an association between whole blood DNA methylation levels and cellular heterogeneity was observed, but where only subtle or no DNAm differences were observed between the cell types (KCNJ11, KCNQ1II and PM20D1), no convincing association was observed. This suggests that the results from the study support the hypothesis, and a need to control for cellular heterogeneity in the analysis of whole blood DNAm data. Finally, a high correlation between DNA methylation levels in cell fractions was observed, which suggest a possibility to use a proportional number of a single white blood cell type to correct for this confounding effect in analyses.

6.2 Type 2 diabetes associated DNA methylation identified in genetic diabetes risk locus

The results indicate that average DNAm levels differ between diabetic and non-diabetic individuals in at least one of the seven regions investigated in total. The region is located in an intragenic CpG island in the *HHEX* gene, which resides in a locus previously associated with type 2 diabetes through genetic studies. The results indicate that the observed difference is not carried by an association with obesity and that it is not an artifact due to differences in the white blood cell composition of diabetics and controls. These findings support the hypothesis that DNAm aberrations may be associated with T2DM.

7 Appendix

7.1 Technical aspects of the DNA methylation assays

Multiple tests were conducted on technical aspects of the DNAm assays used in the study in order to optimize the assays and to gain information on potential sources of measurement error. Samples used for the analyses presented in this section are from diabetic and non-diabetic individuals, both whole blood and blood fraction DNA, and unless otherwise specified did not overlap between the different conditions under study in each case.

7.1.1 Assay optimization

Prior to analysis of DNA samples on the pyrosequencer, PCRs are conducted in order to amplify the sequence which is to be analyzed. It is important to optimize this step in the process because the number of DNA strands present in the pyrosequencing reaction is directly related to the signal strength during sequencing, and thus (presumably) the quality of the DNAm data. In order to establish optimum PCR conditions for subsequent analysis on the pyrosequencer, the effects of three factors on pyrosequencing signal strength were evaluated; the amount of input DNA (section 7.1.1.1), number of PCR steps (section 7.1.1.2) and polymerase type (section 7.1.1.3) used in preceding PCRs.

7.1.1.1 Effect of the amount of input DNA

An experiment was conducted to test whether the DNA concentration in preceding PCRs affected pyrosequencing signal strength. DNA samples from three individuals were bisulfite converted, each in eight reactions with varying amounts of input DNA, ranging from ~6 ng to 800 ng. PCR was subsequently performed using the HHEXII assay with 3 µl of DNA eluted from each of the 24 conversions. All PCRs were conducted under the same conditions in terms of all other aspects than DNA concentration. The average signal strength from pyrosequencing reactions on the PCR products were compared and two factors were evaluated; whether the signal strength was dependent on the amount of input DNA, and if so, whether a “plateau” would be reached where increasing the amount had no additional effect on the signal strength.

The results indicated that average signal strength in the pyrosequencing reaction increased with increasing DNA concentration in the preceding PCRs (**figure 8A**). In addition, the results indicated that the increase in average signal strength reached a plateau when products from PCRs conducted with DNA eluted from bisulfite conversion of 400 ng of DNA were sequenced, i.e., the signal strength did not increase further.

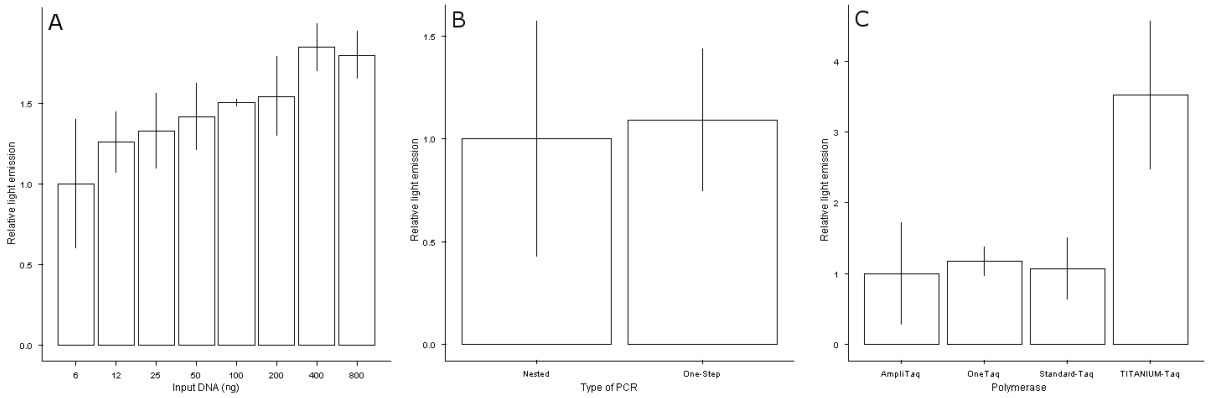


Figure 8. The effect of DNA concentration (A), PCR steps (B) and polymerase types (C) in preceding PCRs on pyrosequencing signal strength.
In each of the three plots, the relative average light emission, an indicator of signal strength (y-axis, note the varying scale) is compared by the different factors in question (x-axis). The HHEXII assay was utilized in all cases. The height of the bars indicate the relative average signal strength observed, and the whiskers extend a single standard deviation above and below the average.

7.1.1.2 Nested versus one-step PCR

An analysis was conducted to test whether employing a one-step PCR rather than a nested (two-step) PCR preceding pyrosequencing affected signal strength in the pyrosequencing reactions. A total of 96 one-step and 96 nested PCRs using the HHEXII assay were conducted and analyzed on the pyrosequencer using the same conditions in all other aspects than varying cycling conditions. Average signal strength in the pyrosequencing reactions on the PCR products from one step and nested PCRs were compared.

The results indicated that the average signal strength in the pyrosequencing reactions conducted on PCR products from a nested and a one-step PCR was very similar (**figure 8B**). They also indicated that variation of the signal strength was much higher when analyzing products from a nested PCR (the standard deviation was ~1.6 times higher), resulting in reads of very poor and very good signal strength, while the reads after a one-step PCR were more uniform.

7.1.1.3 Comparison of DNA polymerases

Amplification of bisulfite converted DNA was tested with four different DNA polymerases (Taq polymerase from NEB (“Standard-Taq”), AmpliTaq from Life Technologies, OneTaq from NEB and TITANIUM-Taq from Clontech) to assess whether use of different types of polymerases affected downstream analyses on the pyrosequencer in terms of signal strength. For each polymerase, 24 DNA samples were amplified under the same conditions using the HHEXII assay and the products analyzed on the pyrosequencer. Average signal strength produced in pyrosequencing reactions on products from each polymerase was compared.

Average signal strengths in the pyrosequencing reactions were observed to differ depending on which polymerase was used for the preceding PCR (**figure 8C**). The average signal strength was highest after PCR using the TITANIUM-Taq. It was ≥ 3.0 times stronger than the signal strength observed after PCR with the other three polymerases, which were all very similar (**figure 8C**).

7.1.2 Tests for biases and robustness of measures

Several analyses were conducted in order to identify factors that cause biases in the DNAm measurements and to establish the robustness of the measurements.

7.1.2.1 Amount of input DNA and data quality

The data presented in section 7.1.1.1 suggests that signal strength in the pyrosequencing reactions is dependent on the amount of input DNA in the preceding PCRs, but it remains to evaluate whether it affects DNAm measurement quality, and if so, whether at a certain level, no further gain in data quality is obtained by the stronger signal. DNAm data from the experiments described in section 7.1.1.1 were therefore used to assess these questions. An average (n=3) intra-individual comparison was made of measured DNAm levels from pyrosequencing of PCR products from amplification of DNA from the eight conversions (which were conducted on varying amounts of input DNA, ranging between ~6 to 800 ng) to reveal whether at a certain level of input DNA, measurements become inconsistent with the other data.

The comparison indicated that measurements on products from PCRs conducted on DNA from conversion of 100-800 ng generally gave consistent results, i.e., similar DNAm pattern across the 10 CpG sites and similar methylation levels at each site (**figures 9A** and **9B**). The measurements obtained from the analysis of amplicons from PCRs conducted on DNA from conversion of less than 50 ng did however appear to be less consistent (**figures 9A** and **9C**).

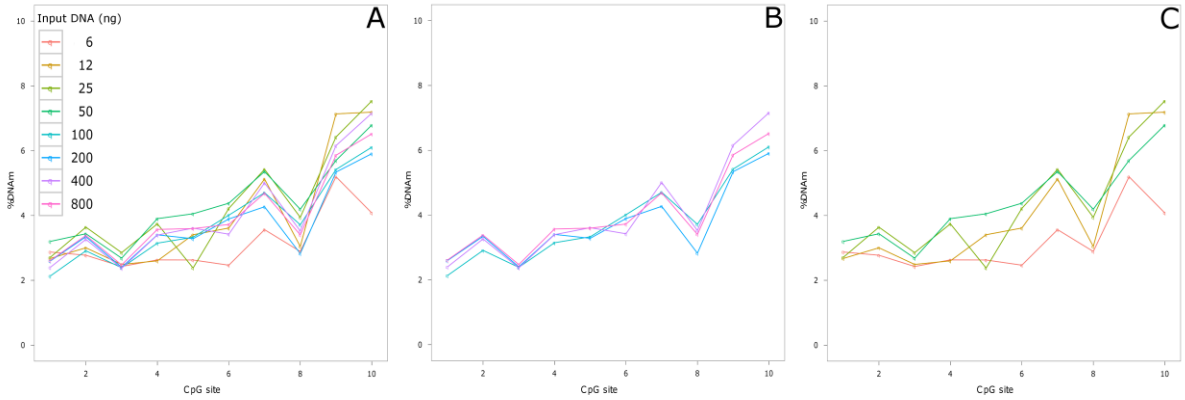


Figure 9. Amount of input DNA and data quality.
The amount of input DNA for bisulfite conversion was varied from ~6 to 800 ng for three individuals, PCR performed on the eluted DNA and the products analyzed on the pyrosequencer using the HHEXII assay. An average intra-individual comparison of measured DNAm levels (y-axis) per CpG site (x-axis) is plotted. Data from analysis of each concentration of DNA is presented with points and lines of a specific color (see legend – the two blue lines are hard to distinguish when printed, but can be viewed in the electronic version). First, the total data is presented (A), second, only the data obtained from measurements on DNA from conversion of 100-800 ng is compared (B), and third only the data obtained from measurements on DNA from conversion of ~6-50 ng is compared (C), i.e., data from A is re-plotted in B and C for better visualization of the data. A line is drawn through the data points for each CpG site solely for the purpose of better visualization of the data and is not intended to imply any interdependency between measurements per CpG site

7.1.2.2 Robustness of measures

An experiment was conducted to assess the robustness of the DNAm measurements. The measurement process, i.e., conversion, amplification and analysis on the pyrosequencer, was performed in duplicate under the same conditions on a number of samples for the HHEXII (n=9), KCNJ11 (n=16), KCNQ1II (n=14) and PM20D1 (n=13) assays. Measured DNAm levels in all CpG sites per assay in the first measurement were compared with the corresponding level in the second measurement and the correlation between measures calculated.

The analysis indicated a high correlation between measurements for all assays, the Spearman's ρ correlation coefficient was ~≥ 0.9 in all cases (**figure 10**). The average absolute difference between measurements was 0.9 percentage points (pp) for HHEXII, 2.5 pp for KCNJ11, 2.5 pp for KCNQ1II and 4.1 pp for PM20D1. Additionally, the data indicated that the measurement error that was detected was not systematic, which would result in a clustering of data on either side of the diagonal on **figure 10**. Rather, the measurement error is sporadic, and the data points fall on either side of the diagonal.

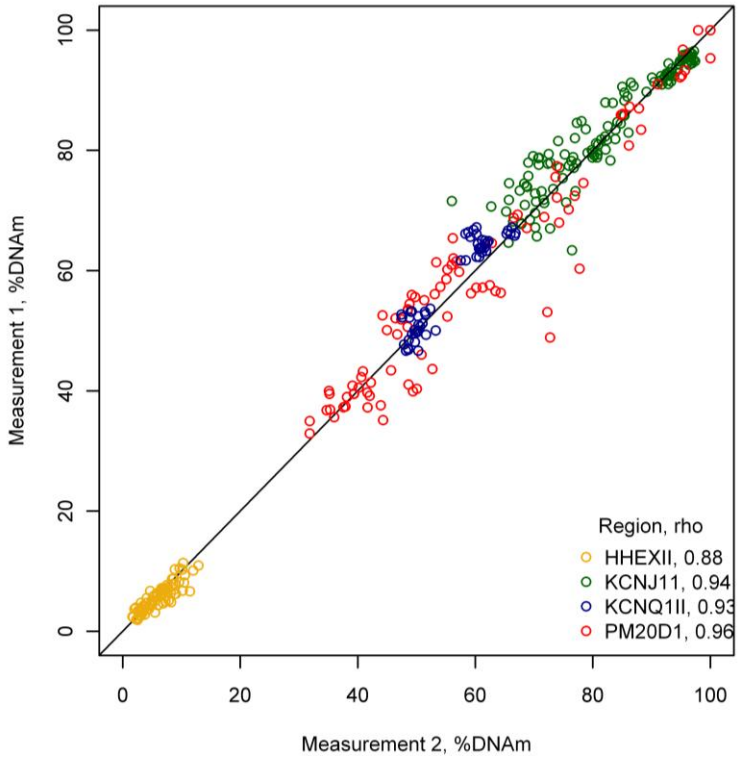


Figure 10. Robustness of DNA methylation measurements. Two separate DNAm measurements were conducted on DNA samples after separate conversion and amplification using four assays: HHEXII (n=9), KCNJ11 (n=16), KCNQ1II (n=14) and PM20D1 (n=13). Results from the two measurements were compared per CpG site, i.e., measured DNA methylation at each CpG site in the first measurement (y-axis) was plotted against the measured DNA methylation at the corresponding CpG site in the second measurement (x-axis). Each dot constitutes a comparison of the DNA methylation level measured per CpG site per assay.

In order to assess whether the measurement error detected above was attributed to the conversion process specifically, rather than the PCR or pyrosequencing steps, a second test was performed. A single bisulfite conversion of 23 DNA samples was performed, two separate PCRs conducted for each sample and the products analyzed on the pyrosequencer. This analysis was only conducted for the HHEXII assay. As in the previous analysis, measured DNAm levels in all CpG sites in the first measurement were compared with the corresponding level in the second measurement and the correlation between them calculated.

The results indicated that average absolute DNAm levels differed by 0.5 pp between measurements and that the Spearman's ρ for correlation between the two measurements was 0.93.

7.1.2.3 PCR bias

PCR bias, i.e., preferential amplification of particular alleles in a heterogeneous pool of alleles, has been demonstrated for amplification of bisulfite treated DNA (97). All assays designed for the present study were tested for such bias, except the PM20D1 assay because methylation levels measured with the assay spanned the whole scale between 0-100%. Control DNA of known methylation state, 0% and 100%, was mixed to obtain DNA of methylation states ranging between 0-100% (specifically, 0%, 25%, 50%, 75% and 100% methylated). The DNA was PCR amplified and the methylation level measured by pyrosequencing of the PCR products. Average measured DNAm levels across the corresponding CpG sites per assay were compared to the expected levels.

The quality of the DNAm data obtained for the KCNQ1II assay was very poor (due to sporadic measurement failure rather than a problem with the assay) and should therefore be interpreted with caution. Additionally, it should be noted that for the KCNQ1I assay, data was only obtained from the measurement of DNA of three methylation states, 0%, 25% and 100%. The comparison indicated that measured methylation levels were in general lower than the expected level, i.e., a bias was observed in all assays, although of very varying degree, towards the unmethylated alleles (**figure 11**). The data for KCNQ1II appears to be in stark contrast with the other results, but given the low quality of the data as noted above, this result should be considered with caution.

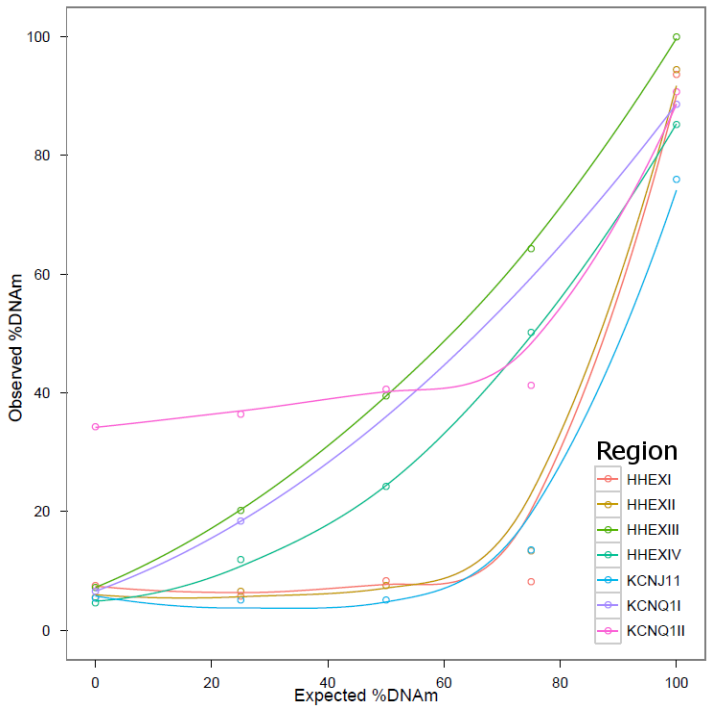


Figure 11. Test for PCR bias. DNA of known methylation state (0%, 25%, 50%, 75% and 100%) was converted, amplified and its methylation level measured on the pyrosequencer using seven different assays (see legend – it is hard to distinguish between the two green colors in the legend; HHEXIII is depicted in a lighter green color than HHEXIV). The expected DNA methylation level (x-axis) was compared to the average measured level across the respective CpG sites per assay (y-axis). Points indicate a single comparison of an expected and measured level per assay, and an approximate trend is indicated by drawing a line through the points of a given assay.

7.1.2.4 Bias due to other factors

An analysis was performed to assess whether a systematic measurement bias could occur when measuring DNAm on the pyrosequencer if different polymerases were used in the preceding PCRs. DNA samples were converted, PCR amplified and analyzed on the pyrosequencer in duplicate under the same conditions in each case, except varying the DNA polymerase used in the preceding PCR,

using TITANIUM-Taq in one reaction and OneTaq in the other reaction. The analysis was conducted for four assays, HHEXII (n=21), KCNJ11 (n=14), KCNQ1II (n=16) and PM20D1 (n=16), and in each case measured DNAm levels from analysis of PCR products using the two polymerases were compared per CpG site per assay.

The analysis indicated that correlation between measurement outcomes was very high for all assays, the Spearman's ρ was ~ ≥ 0.8 (**figure 12**). However, DNAm measurement outcomes obtained from analysis of PCR products from the OneTaq polymerase were consistently lower than those obtained from TITANIUM-Taq for three of the four assays tested; HHEXII, KCNQ1II and PM20D1, but not KCNJ11 (**figure 12**, exemplified by the unidirectional deviation of the data points from the diagonal). For HHEXII, the average absolute difference between measurements was 3.7 pp, 10.0 pp for KCNQ1II and 13.9 pp for PM20D1, while it was 1.7 pp for KCNJ11.

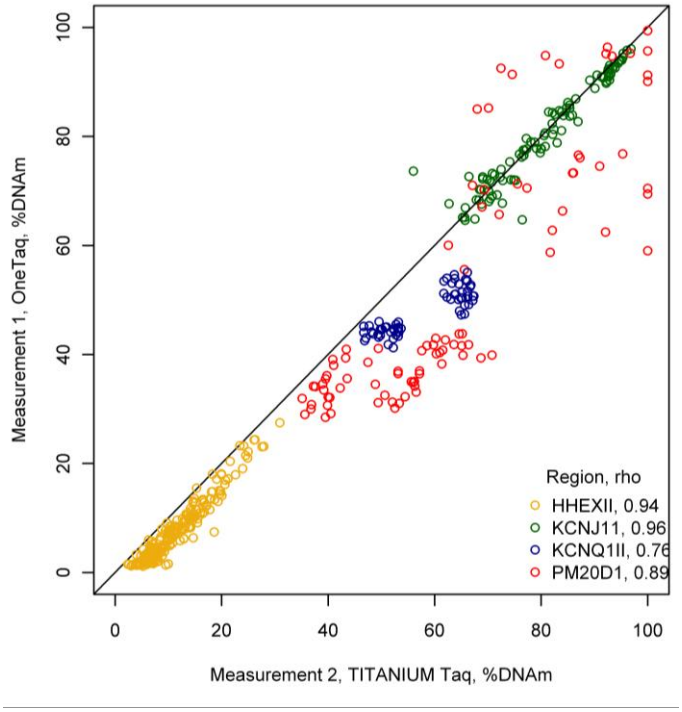


Figure 12. Test for DNA methylation measurement bias by polymerase type in the preceding PCR.
DNA was amplified using two different polymerases and the products analyzed on the pyrosequencer. The DNA methylation levels detected after amplification using the OneTaq polymerase (y-axis) were compared to those detected after amplification using the TITANIUM-Taq polymerase (x-axis) per CpG site. The dots represent a comparison of measured DNA methylation level from the two analyses. Data for each of the four assays tested, HHEXII (n=21), KCNJ11 (n=14), KCNQ1_2 (n=16) and PM20D1 (n=16), is represented in a distinct color (see legend).

Finally, an analysis was performed to test whether a measurement bias could occur when measuring DNAm on the pyrosequencer depending on whether a one-step or a nested PCR was performed on DNA samples prior to analysis on the pyrosequencer. Eight DNA samples were converted, PCR amplified and analyzed on the pyrosequencer in duplicate under the same conditions, except using a nested PCR approach in one case, and a one-step approach in a second case. This analysis was only conducted for the HHEXII assay. The measured DNAm levels from the two pyroruns for each sample were compared per CpG site as before, and their correlation calculated.

The results indicated that no bias was caused by the use of different PCR approaches. The average absolute difference between measurements was 1.0 pp and the Spearman's correlation coefficient was 0.90.

7.2 Protocols

7.2.1 Protocol for isolation of mononuclear and polymorphonuclear cells from whole blood

The following protocol was used to isolate MNCs and PMNCs from whole blood. For this process, Histopaque-1077 Ficoll medium and Accuspin Tubes were used, and the protocol is adapted from the manufacturer's instructions (Sigma-Aldrich).

1. A day before blood is to be drawn, 15 ml of the Histopaque-1077 Ficoll medium are poured into the 50 ml Accuspin Tubes and stored at 4°C.
2. Prior to pouring blood into the tube, the medium should have sunk through the porous membrane in the tube. If this is not the case, or if the tubes are to be prepared on the day of blood collection, they can be centrifuged briefly. Additionally, the blood and medium (in the tubes) should be at room temperature prior to proceeding further.
3. Blood is poured slowly into the tubes, on top of the membrane which separates it from the medium below.
4. The tubes are centrifuged for 15 minutes at 800 x g.
5. The MNC fraction can be observed as a cloudy layer between the plasma (topmost layer) and the medium. These cells are collected by dipping a transfer pipette through the plasma and by "sucking" the liquid just above the cell layer.
6. The PMNCs fraction can be observed as a pellet at the bottom of the tube. These cells are extracted by inserting a pipette into the tube and shifting the membrane carefully by pushing with the pipette's tip at the membrane/tube boundary, so that the pipette can reach the bottom. When the pipette tip reaches the bottom of the tube, the pellet is extracted with the pipette.
7. The fractions are transferred into separate tubes and kept at -20°C for DNA extraction.

7.2.2 Protocol for DNA extraction

The following protocol was used for extraction of DNA from whole blood and blood fractions. The protocol is based on an extraction method developed by Scotlab Bioscience (Coatbridge, Scotland, UK). Recipes for the reagents and buffers used are provided following the protocol.

1. Blood and all solutions used should be cold (4°C) before use, unless otherwise specified.
2. Reagent A, 30-40 ml, is added to 5-10 ml of EDTA blood and the solution is mixed by inverting its container briskly (e.g., tubes) for 2 minutes.
3. The solution is centrifuged at 1600 x g for 10 minutes.
4. The supernatant is decanted without disturbing the pellet which has formed at the bottom.
5. Two milliliters of Reagent B (room temperature) are added and the pellet dissolved in the liquid by vortexing.
6. The solution is transferred to 0.5 ml of 5 M sodium-perchlorate and incubated at 65°C for 25 minutes. During incubation, the solution is mixed shortly by inverting the container every five minutes.
7. The solution is cooled for ~ 15 minutes at 4°C and 2 ml of chilled (-20°C) chloroform added.
8. The solution is mixed by rotation for 10 minutes and subsequently centrifuged at 1600 x g for 10 minutes.
9. The supernatant is transferred to 5 ml of 96% (v/v) ethanol and mixed slowly by inverting the container.
10. The DNA can at this point be observed as a white substance floating around in the ethanol. It is transferred to ~1 ml of 70% (v/v) ethanol and finally to a tube containing 1 ml of TE buffer.
11. The DNA is reconstituted in the buffer by rotating the tube at room temperature for 24 hours and subsequently by incubation at 4°C for four weeks.

TE buffer (10X, 1 L)

1. 100 ml of 1 M Tris solution (pH=8)
2. 16.7 ml of 0.6 M EDTA solution
3. Add dH_2O to 1 liter
4. Autoclave

Reagent A (1X, 2 L)

1. 0.64 mol Sucrose (219.1 g of $C_{12}H_{22}O_{11}$)
2. 0.01 mol MgCl2 (2.04 g of $MgCl_2 \cdot 6H_20$)
3. 20 ml of 1 M Tris solution (pH=8)
4. 20 ml Triton X 100
5. Add dH_20 to 2 liters
6. Autoclave

Reagent B (1X, 0.5 L)

1. 200 ml of 1 M Tris solution (pH=8)
2. 50 ml of 0.6 M EDTA solution
3. 50 ml of 1.5 M NaCl solution
4. Add dH_2O to 450 ml
5. Autoclave
6. 50 ml of 10% (w/v) SDS solution added

7.2.3 Protocol for bisulfite conversion of DNA samples

The following protocol was used for bisulfite conversion of DNA samples in the study. It is adapted from the protocol provided by the manufacturer (Zymo Research) of the conversion kit used for this process (EZ-DNA Methylation kit). All buffers and plates (e.g., M-dilution buffer and the Conversion Plate respectively) referred to in the protocol are provided in the kit, and are not defined here.

1. Add 5 µl of M-Dilution Buffer to each DNA sample in a Conversion Plate and adjust the total volume to 50 µl with water. Mix each sample by pipetting up and down.
2. Incubate the Conversion Plate containing the samples at 37°C for 15 minutes in a thermal cycler.
3. After the above incubation, add 100 µl of the CT Conversion Reagent to each sample and mix by pipetting up and down.
4. Incubate the Conversion Plate in the dark at 50°C for 12 – 16 hours using a thermal cycler.
5. Incubate the sample at 0 - 4°C on ice for 10 minutes.
6. Add 400 µl of M-Binding Buffer to each well of a Zymo-Spin™ I-96 Binding Plate on a Collection Plate.
7. Load the samples (from Step 5) into the wells of the Zymo-Spin™ I-96 Binding Plate containing the M-Binding Buffer. Mix by pipetting up and down.
8. Centrifuge at 3,000 x g for 5 minutes. Discard the flow-through.
9. Add 500 µl of M-Wash Buffer to each well and centrifuge at ≥ 3,000 x g for 5 minutes.
10. Add 200 µl of M-Desulphonation Buffer to each well of and let stand at room temperature for 15 - 20 minutes. After the incubation, centrifuge at 3,000 x g for 5 minutes.
11. Add 500 µl of M-Wash Buffer to each well and centrifuge at 3,000 x g for 5 minutes. Add another 500 µl of M-Wash Buffer and centrifuge for 10 minutes.

12. Place the Zymo-Spin™ I-96 Binding Plate onto an Elution Plate. Add 15 µl of M-Elution Buffer directly to the binding matrix in each well. Centrifuge for 3 minutes at 3,000 x g to elute the DNA.
13. The DNA is ready for immediate analysis or can be stored at or below -20°C for later use. For long term storage, store at or below -70°C.

7.2.4 Protocol for preparation of amplicons for analysis on the pyrosequencer

The following protocol was used to capture and wash amplicons prior to analysis on the pyrosequencer.

1. Mix 20 µl of the PCR product, 2 µl Streptavidin Sepharose solution (GE Healthcare, cat.nr.; 19-5113-01), 40 µl PyroMark Binding Buffer (QIAGEN, cat.nr.: 979006) and 18 µl H_2O.
2. Agitate the above solution for 5-10 minutes at 1400 rpm and subsequently move to the PyroMark Q24 Vacuum Prep Workstation (QIAGEN, cat.nr.: 9001516).
3. The PCR products are captured with the Vacuum Prep Tool (part of the workstation) and moved through a series of solutions, aspirating each through the tool's filter probes.
4. First, 70% ethanol is flushed through apparatus for 5 seconds, second 0.2 M NaOH solution for 5 seconds, and third 10 mM Tris-Acetate buffer for 5 seconds (PyroMark Wash Buffer, QIAGEN, cat.nr.: 979008).
5. The PCR products are released onto a PyroMark Q24 Plate (QIAGEN, cat.nr.:979201) by turning the vacuum off and gently shaking the Vacuum Prep Tool with the filter probes positioned in the plates wells containing 25 µl of PyroMark Annealing Buffer (QIAGEN, cat.nr.: 979009) with 0.3 µM of the appropriate sequencing primer.
6. The plate is put in a PyroMark Plate Holder (QIAGEN, cat.nr.: 979205) and incubated on a thermal block at 80°C for 2 minutes. After the incubation, the plate is cooled down to room temperature by letting stand on the workbench for approximately 5 minutes.
7. The plate, and its contents, the PCR products, are ready for analysis on the pyrosequencer.

7.3 Supplementary tables and figures

Table 5. **Primer sequences for the PCR assays used in the study.**

Region	Inner/outer*	Sequencing primer	Forward primer	Reverse primer
KCNJ11	Outer		GTGTGTGGTTATTTGAGGTTTATTAG	AACCTAATAATCTACCCTCCTCAAC
KCNJ11	Inner	ATCACCCAAACCATACTATCC	GTTGTAGTTGTTTTTTTTGGATATAAAG	ACTCTACAATAAAACCCTAAACCAC
HHEXI	Outer		TGGATTGAAGATTGTATAGTTTTTGTT	CCCCTAAAAACTCCAAACACC
HHEXI	Inner	GAGTTCGTAGTATTTGAATTTTAGT/ AGGAATTTAGGGTA	GGTTTTTAAATGAAATTAGGTGGA	GTTTGGAAATAGTTGTTGTTATTT
HHEXII	Outer		TTTTTGGGTTATTGTTGGGAT	CAACCTTATACACACACAAACAAAC
HHEXII	Inner	GTTAGGATTGGAGGTTT	ATGTTGTTATAGTTTATGGGGTGGT	TTACCCCCTTAAATCTCCCTTAATA
HHEXIII	n/a	GGGGGTAAAAAGTTATGTATA	GGAGATTTAAGGGGGTAAAAAGT	CCTAAAACTAAATCCAAACATACCTTTAAC
HHEXIV	n/a	AAGTAATTTGATTATAAAATAAAG	GTAAAAAAAATGGTTAAAATGTGTTT	CAACAAAATCCAACCCCAATCA
KCNQ1I	n/a	GGGGGAGTTTTGTTTTA	TTGGTGTGGGGGAGTTTT	ACTTCCTTCCCTCCTCTACT
KCNQ1II	n/a	GGTTAGGTTGTATTGTTG	GTATTGTTTAGGTTAGGTTGTATTGT	ACCCTCCCCATCTCTCTAA
PM20D1	n/a	GTTGAATTGAGAAGGGAT	ATGAGTATAGGTGGGTGAAG	ACCCTAATAACTATACTACTCCTAATTTTC

*Inner or outer primer set.

Inner reverse primer was biotinilated in all cases, except for the KCNJ11 assay, where the inner forward primer was biotinilated.

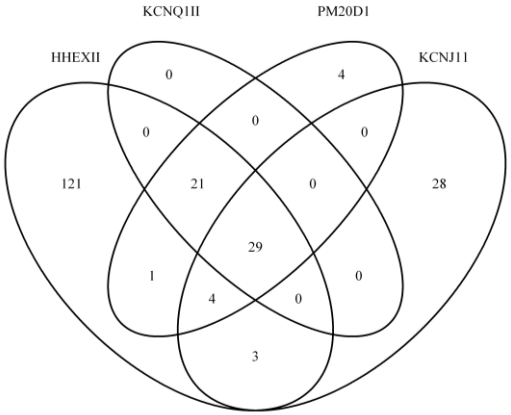


Figure 13. Venn diagram depicting the number of samples analyzed per region.
The diagram contains a set of 15 numbers that, when added together, represent the total number of whole blood DNA samples used for the study presented in section, 4.1.1, pg. 27 Each ellipse contains a set of numbers, that when added together represent the total number of samples analyzed for a specific region. Finally, some samples were analyzed for more than one region, and this is represented by the overlapping of ellipses.

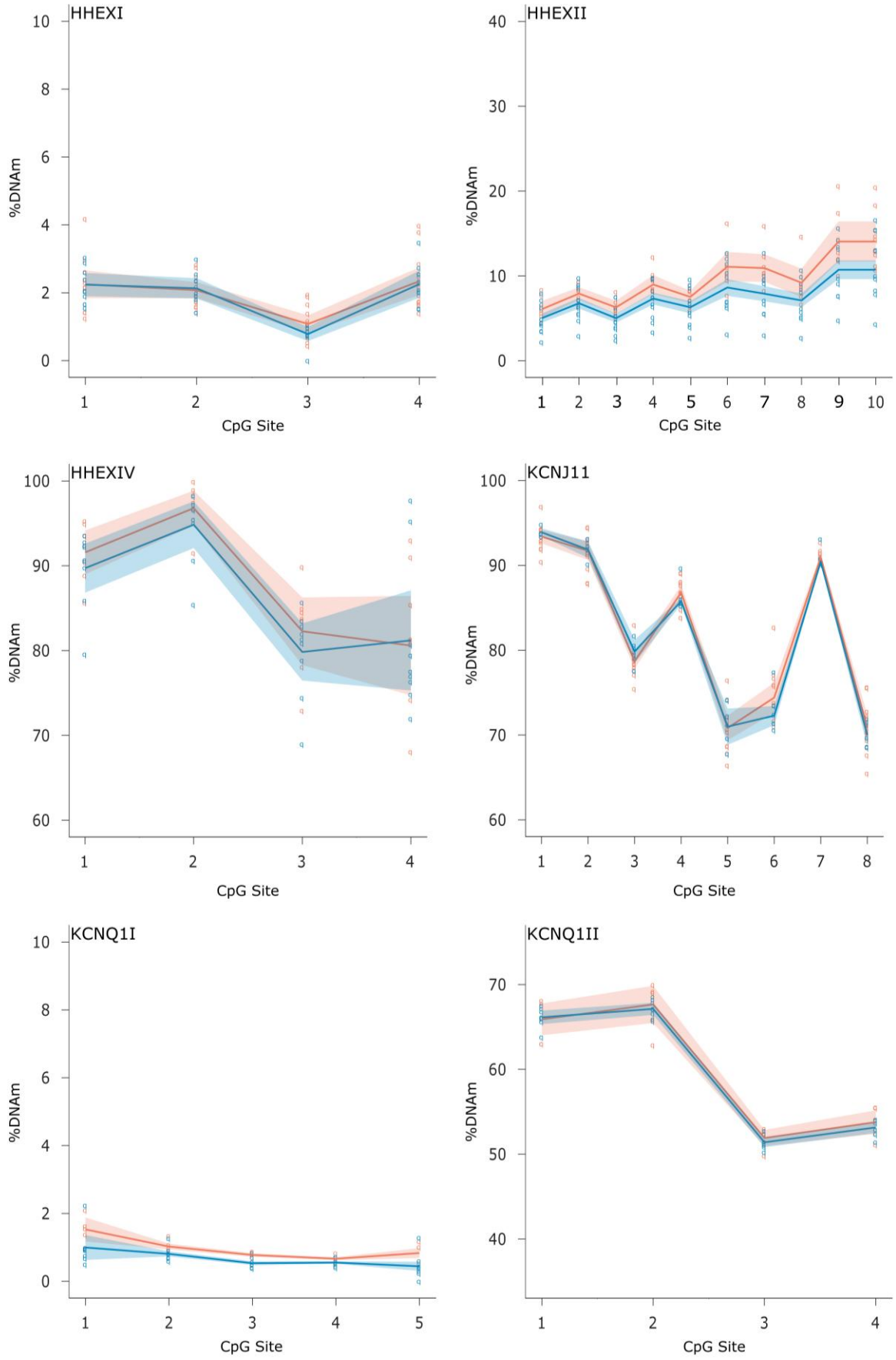


Figure 14. Comparison of DNAm levels in diabetics and controls in six regions in three genes. See legend on figure 6, pg. 34, where the same plots are shown. Here, the plots are provided on differing scales so as to provide better detail.

7.4 Other supplementary material

The following script was used to prepare figures 6, 7 and 14.

For a document “HHEXi.txt” containing a dataset in the following form (where variable S_ID is subject number, CPG1-4 is measured DNAm level per CpG site and group is a variable differentiating the groups of individuals being compared in the plot, e.g, with and without diabetes):

S_ID	CPG1	CPG2	CPG3	CPG4	group
123456	2.31	1.31	0.54	1.72	G1
123457	2.17	2.93	1.21	3.39	G2
123458	2.04	1.78	0.78	1.12	G2

…

The following script calculates variables that are required for plotting and saves them in separate dataframes:

```
hhexi <- read.delim("HHEXi.txt")
library(reshape2)
hhexim <- melt(hhexi, id=c("S_ID", "group"))
hheximo <- na.omit(hhexim)
CI <- function(x) sd(x)/sqrt(length(x))*1.96
o2 <- data.frame(matrix(c(1:4)), t(tapply(hheximo$value, list(hheximo$group, hheximo$variable), mean)), t(tapply(hheximo$value, list(hheximo$group, hheximo$variable), CI)), t(tapply(hheximo$value, list(hheximo$group, hheximo$variable), mean)) - t(tapply(hheximo$value, list(hheximo$group, hheximo$variable), CI)), t(tapply(hheximo$value, list(hheximo$group, hheximo$variable), mean)) + t(tapply(hheximo$value, list(hheximo$group, hheximo$variable), CI)))
colnames(o2) <- c("x", "mean", "mean", "ci", "ci", "ylo", "ylo", "yhi", "yhi")
ok <- rbind(subset(o2, select=c(1,2,4,6,8)), subset(o2, select=c(1,3,5,7,9)))
hhexi2 <- data.frame(group=matrix(c("G1","G1","G1","G1","G2","G2","G2","G2")), ok)
```

The following plots the data:

```
library(ggplot2)
ggplot() + geom_point(data=hhexim ,aes(variable,value, colour=factor(group))) + geom_line(data=hhexi2, aes(x=x, y=mean, colour=factor(group))) + geom_ribbon(alpha=0.25, data=hhexi2, aes(x=x, ymin=ylo, ymax=yhi, fill=group))
```

8 References

1. Mill J, Tang T, Kaminsky Z, Khare T, Yazdanpanah S, Bouchard L, et al. Epigenomic profiling reveals DNA-methylation changes associated with major psychosis. Am J Hum Genet. 2008;82:696–711.

2. Henikoff S, Matzke MA. Exploring and explaining epigenetic effects. Trends Genet. 1997;13:293–5.

3. Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, Tonti-Filippini J, et al. Human DNA methylomes at base resolution show widespread epigenomic differences. Nature. 2009;462:315–22.

4. Saxonov S, Berg P, Brutlag DL. A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. P Natl Acad Sci USA. 2006;103:1412–7.

5. Jabbari K, Bernardi G. Cytosine methylation and CpG, TpG (CpA) and TpA frequencies. Gene. 2004;333:143–9.

6. Suzuki MM, Bird A. DNA methylation landscapes: provocative insights from epigenomics. Nat Rev Genet. 2008;9:465–76.

7. Bird A. DNA methylation patterns and epigenetic memory. Gene Dev. 2002;16:6–21.

8. Ehrlich M, Gama-Sosa MA, Huang L-H, Midgett RM, Kuo KC, McCune RA, et al. Amount and distribution of 5-methylcytosine in human DNA from different types of tissues or cells. Nucleic Acids Res. 1982;10:2709–21.

9. Illingworth R, Kerr A, DeSousa D, Jørgensen H, Ellis P, Stalker J, et al. A novel CpG island set identifies tissue-specific methylation at developmental gene loci. PLoS Biol. 2008;6:e22.

10. Antequera F, Bird A. Number of CpG islands and genes in human and mouse. P Natl Acad Sci USA. 1993;90:11995–9.

11. Bestor TH, Ingram VM. Two DNA methyltransferases from murine erythroleukemia cells: Purification, sequence specificity, and mode of interaction with DNA. P Natl Acad Sci USA. 1983;80:5559–63.

12. Cheng X, Blumenthal RM. Mammalian DNA methyltransferases: a structural perspective. Structure. 2008;16:341–50.

13. Bestor TH. The DNA methyltransferases of mammals. Hum Mol Genet. 2000;9:2395–402.

14. Gu T-P, Guo F, Yang H, Wu H-P, Xu G-F, Liu W, et al. The role of Tet3 DNA dioxygenase in epigenetic reprogramming by oocytes. Nature. 2011;477:606–10.

15. Smallwood SA, Kelsey G. *De novo* DNA methylation: a germ cell perspective. Trends Genet. 2012;28:33–42.

16. Eckhardt F, Lewin J, Cortese R, Rakyan VK, Attwood J, Burger M, et al. DNA methylation profiling of human chromosomes 6, 20 and 22. Nat Genet. 2006;38:1378–85.

17. Fan S, Zhang X. CpG island methylation pattern in different human tissues and its correlation with gene expression. Biochem Bioph Res Co. 2009;383:421–5.

18. Irizarry RA, Ladd-Acosta C, Wen B, Wu Z, Montano C, Onyango P, et al. The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. Nat Genet. 2009;41:178–86.

19. De Bustos C, Ramos E, Young JM, Tran RK, Menzel U, Langford CF, et al. Tissue-specific variation in DNA methylation levels along human chromosome 1. Epigenetics & Chromatin. 2009;2:7.

20. Alberts B, Johnson A, Lewis J, Raff M, Roberts K, Walter P. Molecular biology of the cell. 4th ed. New York: Garland Science; 2002.

21. Kochanek S, Toth M, Dehmel A, Renz D, Doerfler W. Interindividual concordance of methylation profiles in human genes for tumor necrosis factors *α* and *β*. P Natl Acad Sci USA. 1990;87:8830–4.

22. Kochanek S, Radbruch A, Tesch H, Renz D, Doerfler W. DNA methylation profiles in the human genes for tumor necrosis factors *α* and *β* in subpopulations of leukocytes and in leukemias. P Natl Acad Sci USA. 1991;88:5759–63.

23. Wu H-C, Delgado-Cruzata L, Flom JD, Kappil M, Ferris JS, Liao Y, et al. Global methylation profiles in DNA from different blood cell types. Epigenetics. 2011;6:76–85.

24. Bernstein BE, Meissner A, Lander ES. The mammalian epigenome. Cell. 2007;128:669–81.

25. Kouzarides T. Chromatin modifications and their function. Cell. 2007;128:693–705.

26. Meissner A, Mikkelsen TS, Gu H, Wernig M, Hanna J, Sivachenko A, et al. Genome-scale DNA methylation maps of pluripotent and differentiated cells. Nature. 2008;454:766–70.

27. Rakyan VK, Down TA, Balding DJ, Beck S. Epigenome-wide association studies for common human diseases. Nat Rev Genet. 2011;12:529–41.

28. Laird PW. The power and the promise of DNA methylation markers. Nat Rev Cancer. 2003;3:253–66.

29. Frommer M, McDonald LE, Millar DS, Collis CM, Watt F, Grigg GW, et al. A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. P Natl Acad Sci USA. 1992;89:1827–31.

30. Xiong Z, Laird PW. COBRA: a sensitive and quantitative DNA methylation assay. Nucleic Acids Res. 1997;25:2532–4.

31. Ronaghi M, Uhlén M, Nyrén P. A sequencing method based on real-time pyrophosphate. Science. 1998;281:363, 365.

32. Uhlmann K, Brinckmann A, Toliat M, Ritter H, Nürnberg P. Evaluation of a potential epigenetic biomarker by quantitative methyl-single nucleotide polymorphism analysis. Electrophoresis. 2002;23:4072–9.

33. Tost J, Dunker J, Gut IG. Analysis and quantification of multiple methylation variable positions in CpG islands by Pyrosequencing. BioTechniques. 2003;35:152–6.

34. Ferguson-Smith AC. Genomic imprinting: the emergence of an epigenetic paradigm. Nat Rev Genet. 2011;12:565–75.

35. Brockdorff N, Turner BM. Epigenetics. 1st ed. New York: Cold Spring Harbor Laboratory Press; 2007.

36. Vardimon L, Kressmann A, Cedar H, Maechler M, Doerfler W. Expression of a cloned adenovirus gene is inhibited by *in vitro* methylation. P Natl Acad Sci USA. 1982;79:1073–7.

37. Stein R, Razin A, Cedar H. *In vitro* methylation of the hamster adenine phosphoribosyltransferase gene inhibits its expression in mouse L cells. P Natl Acad Sci USA. 1982;79:3418–22.

38. Mohandas T, Sparkes R, Shapiro L. Reactivation of an inactive human X chromosome: evidence for X inactivation by DNA methylation. Science. 1981;211:393–6.

39. Venolia L, Gartler S, Wassman E, Yen P, Mohandas T, Shapiro L. Transformation with DNA from 5-azacytidine-reactivated X chromosomes. P Natl Acad Sci USA. 1982;79:2352–4.

40. The ENCODE Project Consortium. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature. 2007;447:799–816.

41. Deng J, Shoemaker R, Xie B, Gore A, LeProust EM, Antosiewicz-Bourget J, et al. Targeted bisulfite sequencing reveals changes in DNA methylation associated with nuclear reprogramming. Nat Biotechnol. 2009;27:353–60.

42. Ball MP, Li JB, Gao Y, Lee J-H, LeProust EM, Park I-H, et al. Targeted and genome-scale strategies reveal gene-body methylation signatures in human cells. Nat Biotechnol. 2009;27:361–8.

43. Rauch TA, Wu X, Zhong X, Riggs AD, Pfeifer GP. A human B cell methylome at 100-base pair resolution. P Natl Acad Sci USA. 2009;106:671–8.

44. Hellman A, Chess A. Gene body-specific methylation on the active X chromosome. Science. 2007;315:1141–3.

45. Watt F, Molloy PL. Cytosine methylation prevents binding to DNA of a HeLa cell transcription factor required for optimal expression of the adenovirus major late promoter. Gene Dev. 1988;2:1136–43.

46. Bird AP, Wolffe AP. Methylation-induced repression - belts, braces, and chromatin. Cell. 1999;99:451–4.

47. Morgan HD, Sutherland HGE, Martin DIK, Whitelaw E. Epigenetic inheritance at the agouti locus in the mouse. Nat Genet. 1999;23:314–8.

48. Popp C, Dean W, Feng S, Cokus SJ, Andrews S, Pellegrini M, et al. Genome-wide erasure of DNA methylation in mouse primordial germ cells is affected by AID deficiency. Nature. 2010;463:1101–5.

49. Hajkova P, Erhardt S, Lane N, Haaf T, El-Maarri O, Reik W, et al. Epigenetic reprogramming in mouse primordial germ cells. Mech Develop. 2002;117:15–23.

50. Shoemaker R, Deng J, Wang W, Zhang K. Allele-specific methylation is prevalent and is contributed by CpG-SNPs in the human genome. Genome Res. 2010;20:883–9.

51. Hellman A, Chess A. Extensive sequence-influenced DNA methylation polymorphism in the human genome. Epigenetics & Chromatin. 2010;3:11.

52. Boks MP, Derks EM, Weisenberger DJ, Strengman E, Janson E, Sommer IE, et al. The relationship of DNA methylation with age, gender and genotype in twins and healthy controls. PloS ONE. 2009;4:e6767.

53. Fraga MF, Ballestar E, Paz MF, Ropero S, Setien F, Ballestar ML, et al. Epigenetic differences arise during the lifetime of monozygotic twins. P Natl Acad Sci USA. 2005;102:10604–9.

54. Relton CL, Davey Smith G. Epigenetic epidemiology of common complex disease: prospects for prediction, prevention, and treatment. PLoS Med. 2010;7:e1000356.

55. Feinberg A, Vogelstein B. Hypomethylation distinguishes genes of some human cancers from their normal counterparts. Nature. 1983;301:89–92.

56. Gama-Sosa MA, Slagel VA, Trewyn RW, Oxenhandler R, Kuo KC, Gehrke CW, et al. The 5-methylcytosine content of DNA from human tumors. Nucleic Acids Res. 1983;11:6883–94.

57. Feinberg AP, Tycko B. The history of cancer epigenetics. Nat Rev Cancer. 2004;4:143–53.

58. Robertson KD. DNA methylation and human disease. Nat Rev Genet. 2005;6:597–610.

59. Costello JF, Frühwald MC, Smiraglia DJ, Rush LJ, Robertson GP, Gao X, et al. Aberrant CpG-island methylation has non-random and tumour-type-specific patterns. Nat Genet. 2000;25:132–8.

60. Gonzalez-Zulueta M, Bender CM, Yang AS, Nguyen T, Beart RW, Van Tornout JM, et al. Methylation of the 5′ CpG island of the *p16/CDKN2* tumor suppressor gene in normal and transformed human tissues correlates with gene silencing. Cancer Res. 1995;55:4531–5.

61. Gupta A, Godwin AK, Vanderveer L, Lu A, Liu J. Hypomethylation of the *Synuclein γ* gene CpG island promotes its aberrant expression in breast carcinoma and ovarian carcinoma. Cancer Res. 2003;63:664–73.

62. Rakyan VK, Beyan H, Down TA, Hawa MI, Maslau S, Aden D, et al. Identification of type 1 diabetes–associated DNA methylation variable positions that precede disease diagnosis. PLoS Genet. 2011;7:e1002300.

63. Javierre BM, Fernandez AF, Richter J, Al-Shahrour F, Martin-Subero JI, Rodriguez-Ubreva J, et al. Changes in the pattern of DNA methylation associate with twin discordance in systemic lupus erythematosus. Genome Res. 2010;20:170–9.

64. Baranzini SE, Mudge J, van Velkinburgh JC, Khankhanian P, Khrebtukova I, Miller NA, et al. Genome, epigenome and RNA sequences of monozygotic twins discordant for multiple sclerosis. Nature. 2010;464:1351–6.

65. Nguyen A, Rauch TA, Pfeifer GP, Hu VW. Global methylation profiling of lymphoblastoid cell lines reveals epigenetic contributions to autism spectrum disorders and a novel autism candidate gene, *RORA*, whose protein product is reduced in autistic brain. FASEB J. 2010;24:3036–51.

66. Feinberg AP, Irizarry RA, Fradin D, Aryee MJ, Murakami P, Aspelund T, et al. Personalized epigenomic signatures that are stable over time and covary with body mass index. Sci Transl Med. 2010;2:49ra67.

67. Wang X, Zhu H, Snieder H, Su S, Munn D, Harshfield G, et al. Obesity related methylation changes in DNA of peripheral blood leukocytes. BMC Med. 2010;8:87.

68. Jüttermann R, Li E, Jaenisch R. Toxicity of 5-aza-2'-deoxycytidine to mammalian cells is mediated primarily by covalent trapping of DNA methyltransferase rather than DNA demethylation. P Natl Acad Sci USA. 1994;91:11797–801.

69. Terry MB, Delgado-Cruzata L, Vin-Raviv N, Wu HC, Santella RM. DNA methylation in white blood cells: Association with risk factors in epidemiologic studies. Epigenetics. 2011;6:828–37.

70. McCarthy MI. Genomics, type 2 diabetes, and obesity. New Engl J Med. 2010;363:2339–50.

71. Herder C, Roden M. Genetics of type 2 diabetes: pathophysiologic and clinical relevance. Eur J Clini Invest. 2010;41:679–92.

72. Lin Y, Sun Z. Current views on type 2 diabetes. J Endocrinol. 2010;204:1–11.

73. Stumvoll M, Goldstein BJ, van Haeften TW. Type 2 diabetes: principles of pathogenesis and therapy. Lancet. 2005;365:1333–46.

74. Danaei G, Finucane MM, Lu Y, Singh GM, Cowan MJ, Paciorek CJ, et al. National, regional, and global trends in fasting plasma glucose and diabetes prevalence since 1980: systematic analysis of health examination surveys and epidemiological studies with 370 country-years and 2.7 million participants. Lancet. 2011;378:31–40.

75. Wild SH, Forouhi NG. What is the scale of the future diabetes epidemic, and how certain are we about it? Diabetologia. 2007;50:903–5.

76. Albertsson V. Diabetes in Iceland. Diabetes. 1953;2:184–6.

77. Bergsveinsson J, Aspelund T, Guðnason V, Benediktsson R. Algengi sykursýki af týpu tvö á Íslandi 1967-2002. Læknablaðið. 2007;93:397–402.

78. Sladek R, Rocheleau G, Rung J, Dina C, Shen L, Serre D, et al. A genome-wide association study identifies novel risk loci for type 2 diabetes. Nature. 2007;445:881–5.

79. Hindorff L, MacArthur J, Wise A, Junkins H, Hall P, Klemm A, et al. A Catalog of Published Genome-Wide Association Studies [Internet]. [cited 2012 Apr 1];Available from: www.genome.gov/gwastudies

80. Gloyn AL, Weedon MN, Owen KR, Turner MJ, Knight BA, Hitman G, et al. Large-scale association studies of variants in genes encoding the pancreatic β-cell K_{ATP} channel subunits Kir6.2 (*KCNJ11*) and SUR1 (*ABCC8*) confirm that the *KCNJ11* E23K variant is associated with type 2 Diabetes. Diabetes. 2003;52:568–72.

81. Yasuda K, Miyake K, Horikawa Y, Hara K, Osawa H, Furuta H, et al. Variants in *KCNQ1* are associated with susceptibility to type 2 diabetes mellitus. Nat Genet. 2008;40:1092–7.

82. Matthews D, Hosker J, Rudenski A, Naylor B, Treacher D, Turner R. Homeostasis model assessment: insulin resistance and β-cell function from fasting plasma glucose and insulin concentrations in man. Diabetologia. 1985;28:412–9.

83. Wallace TM, Levy JC, Matthews DR. Use and abuse of HOMA modeling. Diabetes Care. 2004;27:1487–95.

84. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, et al. Finding the missing heritability of complex diseases. Nature. 2009;461:747–53.

85. McCarthy MI, Hirschhorn JN. Genome-wide association studies: potential next steps on a genetic journey. Hum Mol Genet. 2008;17:R156–65.

86. Bell CG, Finer S, Lindgren CM, Wilson GA, Rakyan VK, Teschendorff AE, et al. Integrated genetic and epigenetic analysis identifies haplotype-specific methylation in the *FTO* type 2 diabetes and obesity susceptibility locus. PloS ONE. 2010;5:e14040.

87. Toperoff G, Aran D, Kark JD, Rosenberg M, Dubnikov T, Nissan B, et al. Genome-wide survey reveals predisposing diabetes type 2-related DNA methylation variations in human peripheral blood. Hum Mol Genet. 2011;ddr472:1–13.

88. Pan XR, Li GW, Hu YH, Wang JX, X AZ, Hu ZX, et al. Effects of diet and exercise in preventing NIDDM in people with impaired glucose tolerance: The Da Qing IGT and Diabetes Study. Diabetes Care. 1997;20:537–44.

89. Li G, Zhang P, Wang J, Gregg EW, Yang W, Gong Q, et al. The long-term effect of lifestyle interventions to prevent diabetes in the China Da Qing Diabetes Prevention Study: a 20-year follow-up study. Lancet. 2008;371:1783–9.

90. Diabetes Prevention Program Research Group. Reduction in the incidence of type 2 diabetes with lifestyle intervention or metformin. New Engl J Med. 2002;346:393–403.

91. Martin GM. Epigenetic drift in aging identical twins. P Natl Acad Sci USA. 2005;102:10413–4.

92. Harris TB, Launer LJ, Eiriksdottir G, Kjartansson O, Jonsson PV, Sigurdsson G, et al. Age, Gene/Environment Susceptibility-Reykjavik Study: multidisciplinary applied phenomics. Am J Epidemiol. 2007;165:1076–87.

93. Kong A, Steinthorsdottir V, Masson G, Thorleifsson G, Sulem P, Besenbacher S, et al. Parental origin of sequence variants associated with complex diseases. Nature. 2009;462:868–74.

94. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, et al. The Human Genome Browser at UCSC. Genome Res. 2002;12:996–1006.

95. Trinklein ND, Aldred SJF, Saldanha AJ, Myers RM. Identification and functional analysis of human transcriptional promoters. Genome Res. 2003;13:308–12.

96. Down TA, Hubbard TJP. Computational detection and location of transcription start sites in mammalian genomic DNA. Genome Res. 2002;12:458–61.

97. Warnecke PM, Stirzaker C, Melki JR, Millar DS, Paul CL, Clark SJ. Detection and measurement of PCR bias in quantitative methylation analysis of bisulphite-treated DNA. Nucleic Acids Res. 1997;25:4422–6.

98. World Health Organization. Defenition and diagnosis of diabetes mellitus and intermediate hyperglycaemia: report of a WHO/IDF consultation [Internet]. Geneva: 2006. Available from: http://www.who.int/diabetes/publications/diagnosis_diabetes2006/en/index.html

99. King H, Aubert RE, Herman WH. Global burden of diabetes, 1995-2025: Prevalence, numerical estimates, and projections. Diabetes Care. 1998;21:1414–31.

100. Bjornsson H, Sigurdsson MI, Fallin MD, Irizarry RA, Aspelund T, Cui H, et al. Intra-individual change over time in DNA methylation with familial clustering. J Am Med Assoc. 2008;299:2877–83.

101. Fuke C, Shimabukuro M, Petronis A, Sugimoto J, Oda T, Miura K, et al. Age related changes in 5-methylcytosine content in human peripheral leukocytes and placentas: an HPLC-based study. Ann Hum Genet. 2004;68:196–204.

102. Stranges S, Rafalson LB, Dmochowski J, Rejman K, Tracy RP, Trevisan M, et al. Additional contribution of emerging risk factors to the prediction of the risk of type 2 diabetes: evidence from the Western New York Study. Obesity. 2008;16:1370–6.

103. Pascoe L, Frayling TM, Weedon MN, Mari A, Tura A, Ferrannini E, et al. Beta cell glucose sensitivity is decreased by 39% in non-diabetic individuals carrying multiple diabetes-risk alleles compared with those with no risk alleles. Diabetologia. 2008;51:1989–92.

104. Grarup N, Rose CS, Andersson EA, Andersen G, Nielsen AL, Albrechtsen A, et al. Studies of association of variants near the *HHEX*, *CDKN2A/B*, and *IGF2BP2* genes with type 2 diabetes and impaired insulin release in 10,705 Danish subjects: Validation and extension of genome-wide association studies. Diabetes. 2007;56:3105–11.

105. Staiger H, Stančáková A, Zilinskaite J, Vänttinen M, Hansen T, Marini MA, et al. A candidate type 2 Diabetes polymorphism near the *HHEX* locus affects acute glucose-stimulated insulin release in European populations: Results from the EUGENE2 study. Diabetes. 2008;57:514–7.

106. Talens RP, Boomsma DI, Tobi EW, Kremer D, Jukema JW, Willemsen G, et al. Variation, patterns, and temporal stability of DNA methylation: considerations for epigenetic epidemiology. FASEB J. 2010;24:3135–44.

107. Crompton MR, Bartlett TJ, MacGregor AD, Manfioletti G, Buratti E, Giancotti V, et al. Identification of a novel vertebrate homeobox gene expressed in haematopoietic cells. Nucleic Acids Res. 1992;20:5661–7.

108. Bedford FK, Ashworth A, Enver T, Wiedemann LM. *HEX*: a novel homeobox gene expressed during haematopoiesis and conserved between mouse and human. Nucleic Acids Res. 1993;21:1245–9.

109. Manfioletti G, Gattei V, Buratti E, Rustighi A, De Iuliis A, Aldinucci D, et al. Differential expression of a novel proline-rich homeobox gene (*Prh*) in human hematolymphopoietic cells. Blood. 1995;85:1237–45.

110. Kerkel K, Schupf N, Hatta K, Pang D, Salas M, Kratz A, et al. Altered DNA methylation in leukocytes with trisomy 21. PLoS Genet. 2010;6:e1001212.

111. Margolis KL, Rodabough RJ, Thomson CA, Lopez AM, McTiernan A. Prospective study of leukocyte count as a predictor of incident breast, colorectal, endometrial, and lung cancer and mortality in postmenopausal women. Arch Intern Med. 2007;167:1837–44.

112. Danesh J, Collins R, Appleby P, Peto R. Association of fibrinogen, C-reactive protein, albumin, or leukocyte count with coronary heart disease: Meta-analyses of prospective studies. J Am Med Assoc. 1998;279:1477–82.

113. Byun H-M, Siegmund KD, Pan F, Weisenberger DJ, Kanel G, Laird PW, et al. Epigenetic profiling of somatic tissues from human autopsy specimens identifies tissue- and individual-specific DNA methylation patterns. Hum Mol Genet. 2009;18:4808–17.

114. Ling C, Del Guerra S, Lupi R, Rönn T, Granhall C, Luthman H, et al. Epigenetic regulation of *PPARGC1A* in human type 2 diabetic islets and effect on insulin secretion. Diabetologia. 2008;51:615–22.

115. Yang BT, Dayeh TA, Kirkpatrick CL, Taneera J, Kumar R, Groop L, et al. Insulin promoter DNA methylation correlates negatively with insulin gene expression and positively with HbA_{1c} levels in human pancreatic islets. Diabetologia. 2011;54:360–7.

116. Bort R, Martinez-Barbera JP, Beddington RSP, Zaret KS. *Hex* homeobox gene-dependent tissue positioning is required for organogenesis of the ventral pancreas. Development. 2004;131:797–806.

117. Cui H, Cruz-Correa M, Giardiello FM, Hutcheon DF, Kafonek DR, Brandenburg S, et al. Loss of *IGF2* imprinting: a potential marker of colorectal cancer risk. Science. 2003;299:1753–5.

118. Bennett RG, Hamel FG, Duckworth WC. An insulin-degrading enzyme inhibitor decreases amylin degradation, increases amylin-induced cytotoxicity, and increases amyloid formation in insulinoma cell cultures. Diabetes. 2003;52:2315–20.

Supplement - Submitted Article

In the pages that follow a manuscript submitted to PLoS ONE on February 20th 2012 based on the results presented in the thesis is provided. The manuscript is displayed here in the form it was submitted to the journal.

Title

Heterogeneity in White Blood Cells Has Potential to Confound DNA Methylation Measurements

Authors and affiliations

Bjorn T Adalsteinsson[1,2], Haukur Gudnason[1], Thor Aspelund[1,2], Tamara B Harris[3], Lenore J Launer[3], Gudny Eiriksdottir[1], Albert V Smith[1,2] and Vilmundur Gudnason[1,2*]

1. Icelandic Heart Association, Kopavogur, Iceland, 2. University of Iceland, Reykjavik, Iceland, 3. National Institute on Aging, Laboratory for Epidemiology, Demography, and Biometry, NIH, Bethesda, Maryland, USA.

*Correspondence to Vilmundur Gudnason, email: v.gudnason@hjarta.is, telephone: +354 5351800, fax: +354 5351801.

Abstract

Epigenetic studies are commonly conducted on DNA from tissue samples. However, tissues are ensembles of cells that may each have their own epigenetic profile and therefore inter-individual difference in cellular heterogeneity may compromise these studies. Here, we explore the potential for such confounding on DNA methylation measurement outcomes when using DNA from whole blood. DNA methylation was measured using pyrosequencing based methodology in two white blood cell fractions, isolated using density gradient centrifugation. In three out of the four CGIs (CpG Islands) tested, we detected significant differential DNA methylation between the two fractions. The difference was very moderate in all but one CGI where the average absolute methylation difference per CpG site ranged between 3.4-15.7 percentage points. In this same CGI, inter-individual variation in cellular heterogeneity explained up to 35% ($p<0.0001$) of the variation in whole blood DNA methylation levels. In the examined CGIs, methylation levels were highly correlated between cell fractions. In summary, our analysis detects region-specific differential DNA methylation between white blood cell sub-types, which can confound the outcome of whole blood DNA methylation measurements. Finally, by demonstrating the high correlation between methylation levels in cell fractions, our results suggest a possibility to use a proportional number of a single white blood cell type to correct for this confounding effect in analyses.

Introduction

Tissue and cell specific methylation are well established in human DNA. In 2006 Eckhardt et al. presented data from the Human Epigenome Project (HEP) that suggest that tissue-specific differentially methylated regions (tDMRs) are very common in the genome [1]. The dataset describes DNA methylation of ~1.9 million CpG sites on chromosomes 6, 20 and 22 in 12 different tissues. Approximately 22% of the investigated amplicons were tDMRs and their average absolute methylation levels differed by up to 20% between tissues (or up to 15% if only somatic tissues are compared). Recently, Fan and Zhang analyzed DNA methylation in selected (CpG site coverage > 30%) CpG islands (CGIs) using the HEP dataset [2]. Similarly, their results indicate that a substantial proportion of CGIs (~18%) are tDMRs. Three recent independent studies using microarray based methods also identify tDMRs after interrogating CpG sites across the whole genome [3], in CGIs across the genome [4], and in non-CGI regions on chromosome 1 [5].

Relatively few studies have addressed the question whether different white blood cell types have specific DNA methylation levels or patterns. In two papers from 1990 and 1991 Kochanek et al. studied the methylation of *TNFα* and *TNFβ* genes in multiple white blood cell types [6,7]. Their results revealed gross differences in *TNFβ* methylation in lymphocytes versus granulo- and monocytes as well as minor distinctions in the *TNFα* gene between cell types. A comparison of DNA methylation levels in CD4+ and CD8+ lymphocytes was included in the HEP report which showed that these highly developmentally related cell types exhibit on average ~5% absolute difference in DNA methylation [1]. Finally, Wu et al. compared different methods and sources of DNA for measuring global DNA methylation in whole blood [8]. DNA derived from whole blood and two blood fractions (mononuclear cells (MNCs) and polymorphonuclear cells (PMNCs)) was measured using five assays; luminometric methylation assay (LUMA), [^{3}H]-methyl acceptance assay and MethyLight assays for long interspersed elements (LINE1), Sat2 and Alu repetitive elements. In four of

the five assays, global methylation levels in MNCs and PMNCs were not correlated, suggesting a widespread difference in methylation between the two cell groups.

As peripheral blood cell DNA is relatively easily accessible it has been an essential source for genetic experiments for the past decades. However whether it is appropriate material for studies on epigenetics has been debated [9] because inter-individual variation in the number of specific white blood cells in combination with cell specific methylation profiles could compromise measurement outcomes for DNA methylation carried out on cells from whole blood. This concern has largely been theoretical due to lack of experimental data. Recently, Talens et al. studied the effect of inter-individual differential white blood cell counts on methylation measurements using whole blood DNA [10]. For a majority of the 16 loci studied, cellular heterogeneity had no effect on variation in DNA methylation. However for one locus it explained 25-50% of the variation and in additional three loci the effect was borderline significant, accounting for up to 8% of the variation between individuals.

In the present study we aimed to investigate the potential confounding effect of cellular heterogeneity on DNA methylation measurement outcomes conducted using whole blood DNA in greater depth than has been done previously; first measuring methylation levels in whole blood DNA samples and estimating their association with cellular heterogeneity, and subsequently measuring and comparing DNA methylation levels in two whole blood cell fractions, MNCs and PMNCs, in order to verify whether any observed association was related to differential DNA methylation in the white blood cells. These analyses have been done in a general context, rather than in any disease-specific context, in order to understand the overall potential for confounding. A confounding effect may be region-specific, depending on two factors; first, the size of the difference in methylation level between cell types, and second due to the relative size of the difference compared to the variation in methylation levels caused by other factors. We therefore chose to analyze DNA methylation in four CGIs, (or more specifically, parts of CGIs, spanning 4-10 CpG sites in each), which represented a range of inter-individual variation in DNA methylation from very low to very high (in genes *HHEX* (Ensembl identifier: ENSG00000152804), *KCNJ11* (ENSG00000187486)*, KCNQ1*

(ENSG00000053918) and *PM20D1* (ENSG00000162877)). Our analysis detects region-specific differential DNA methylation in white blood cell fractions and suggests that such difference can confound DNA methylation measurement outcomes conducted on whole blood.

Results

DNA methylation in whole blood

DNA methylation was measured in DNA isolated from whole blood, in CGIs located in four genes; *HHEX*, *KCNJ11*, *KCNQ1* and *PM20D1*. The examined loci are referred to as the "*HHEX* CGI", " *KCNJ11* CGI", "*KCNQ1* CGI" and the "*PM20D1* CGI" in the text below, as the analysis is focused on the methylation levels of the CpG islands, rather than on the genes themselves. We analyzed the methylation levels in a total of 10 CpG sites for the *HHEX* CGI, 8 for the *KCNJ11* CGI, 4 for the *KCNQ1* CGI and 7 for the *PM20D1* CGI. In total, whole blood DNA methylation data was successfully obtained for 169 individuals for the *HHEX* CGI*,* 54 for the *KCNJ11* CGI, 49 for the *KCNQ1* CGI and 59 for the *PM20D1* CGI (after exclusion of individuals due to missing values and outliers, see materials and methods for details). Each CpG site was numbered sequentially on the basis of its distance from the forward primer. The exact genomic position and corresponding number assigned to each site is listed in table S1 and a gene-map for each locus, to indicate the approximate position of the CpG sites analyzed is shown in figure 1.

Whole blood DNA methylation levels differed between CpG sites within each CGI (figure 1), but were generally very low for the *HHEX* CGI (<20%), intermediate for the *KCNQ1* CGI (ranging between ~40-60%), intermediate to very high for the *KCNJ11* CGI (ranging between ~60-100%) and very low to very high for the *PM20D1* CGI (ranging between ~0-100%). The results also indicated that intra-individual variability in DNA methylation differed between CGIs; it was high for the *KCNJ11 and KCNQ1* CGIs, but low for the *HHEX and PM20D1* CGIs. In general, inter-individual variability was very low for the *KCNQ1* CGI, intermediate for the *HHEX* and *KCNJ11* CGIs and very high for the *PM20D1* CGI; the standard deviation per CpG site ranged between 1.4-1.9 percentage points (pp) for the *KCNQ1* CGI, 1.5-3.0 pp for the *HHEX* CGI, 1.3-3.4 pp for the *KCNJ11* CGI and 22.8-25.3 pp for the *PM20D1* CGI.

The inter-individual variability in whole blood DNA methylation level could in theory, at least partly, be explained in terms of differential white blood cell composition between the studied individuals. The numbers of white blood cells, neutrophils, lymphocytes, monocytes, eosinophils and basophils, were counted using an automated cell counter. In our samples (n=211) the cell counts varied considerably between individuals. The relative standard deviation for the neutrophil proportion was 14.7% (56.8% ± 8.3%), lymphocytes 25.8% (29.7 ± 7.7%), monocytes 30.8% (9.4 ± 2.9%), eosinophils 61.1% (3.6 ± 2.2%) and basophils 96.9% (0.5 ± 0.5%). We analyzed whether the variation in proportional numbers of specific white blood types were associated with variation in measured DNA methylation levels. Statistical analysis indicated that a significant proportion of the variability in the *HHEX* CGI could be explained by this factor, or up to 35% ($p<0.0001$). Additionally, DNA methylation levels in the *KCNJ11* CGI were suggestively associated with the basophil proportion, explaining 3% of the variation in the tested model ($p=0.04$), perhaps only owing to multiple testing. None of the five white blood cell ratios were significantly associated with measurement outcomes for the *KCNQ1* and *PM20D1* CGIs (table 1). These results were minimally affected by outliers and missing values, except for the association between DNA methylation in the *KCNJ11* CGI and the basophil proportion, which was not significant when this data was included.

DNA methylation in white blood cell fractions

To examine if the variability in measured methylation level at different CGIs in whole blood was attributable to differential methylation in the white blood cell types comprising whole blood, we fractionated whole blood samples from 20 individuals into mononuclear cells (MNCs, containing lymphocytes and monocytes) and polymorphonuclear cells (PMNCs, containing neutrophils, basophils and eosinophils), isolated DNA and measured the methylation levels at the four CGIs in each fraction. DNA was also isolated from whole blood for these individuals, methylation levels measured in each of the four CGIs and the data included in the analysis above.

We compared the methylation levels measured in MNCs and PMNCs and observed a higher average methylation in MNCs in 21 of the 29 CpG sites analyzed in total. Paired Wilcoxon signed rank test revealed that 18 of these CpGs were significantly differentially methylated in the two different cell fractions, located in the *HHEX*, *KCNJ11* and *KCNQ1* CGIs (figure 2). The average absolute difference between the two cell fractions was highest for the *HHEX* CGI. All ten CpGs studied at this CGI showed significantly higher methylation in MNCs. The absolute difference ranged between 3.4-15.7 pp (corresponding to ~2.3-4.0 fold higher methylation levels in MNCs per CpG site). Methylation in the *KCNJ11* CGI was also significantly higher in MNCs. The difference was more moderate, but nonetheless significant in 7 out of 8 CpGs, ranging between 0.4-6.1 pp (corresponding up to ~1.1 fold higher methylation levels). In the case of the *KCNQ1* CGI, only one CpG site was significantly differentially methylated between the two fractions (average absolute difference was 1.2 pp, corresponding to ~2% higher methylation levels). DNA methylation levels did not differ significantly in the *PM20D1* CGI.

DNA methylation levels are correlated between blood cell fractions

The results in figure 2 suggest that the methylation patterns between cell fractions are highly similar. To quantify this observation we analyzed the correlation between methylation levels for the two different fractions (figure 3). The correlation was very high in all CGIs, irrespective of whether methylation levels differed between cell fractions or not. Spearmans ρ was 0.72 for *the HHEX* CGI, 0.93 for the *KCNJ11* CGI, 0.80 for the *KCNQ1* CGI and 0.95 for the *PM20D1* CGI.

Discussion

Studies on DNA methylation using whole blood DNA frequently do not control for inter-individual variation in the cellular population from which the DNA is derived, the white blood cells; lymphocytes, neutrophils, eosinophils, basophils and monocytes. This has been criticized due to hypothesized potential for confounding effect when cellular heterogeneity is present in conjunction with cell type specific DNA methylation [9]. Here, we studied this hypothesis in a more comprehensive manner than has been done previously by first testing for an association between whole blood DNA methylation levels and cellular heterogeneity, and second to test whether differential methylation in two cell fractions might underlie the observed association. Our data indicated that indeed a locus specific association between measured DNA methylation levels and cellular heterogeneity in whole blood can be observed. Further, we observed significant differences in locus specific DNA methylation levels in two blood fractions, MNCs and PMNCs, suggesting that it could be the underlying cause of the observed association between DNA methylation levels and white blood cell counts. Finally, in all loci tested we observed that DNA methylation in MNCs and PMNCs is highly correlated independent of differential methylation levels in these fractions.

Up to 35% of the inter-individual variation in whole blood DNA methylation in the *HHEX* CGI was attributed to cellular heterogeneity, suggesting that a considerable confounding can affect measured levels of whole blood DNA methylation due to differences in the cellular population. We also detected a weak association between DNA methylation in the *KCNJ11* CGI and the basophil ratio of small effect size (3%). However, given very small proportion of basophils, as well as a suggestive p-value of 0.04, this result does not convincingly demonstrate an example of cellular heterogeneity confounding methylation measurements. Additionally, given the number of tests performed, a correction for multiple testing may be appropriate. Any such correction would presumably deem the association between the basophil ratio and *KCNJ11* CGI methylation not significant, while even a conservative correction (e.g., Bonferroni) would not affect the significance of the association

between cell fractions and DNA methylation in the *HHEX* CGI. No effect on measurements for the *KCNQ1* and *PM20D1* CGIs was observed, suggesting that this type of confounding does not affect DNA methylation outcomes universally throughout the genome, but may be locus-specific. These results are in concordance with a previous study [10] where out of a total of 16 loci assayed, only a single locus was affected in similar magnitude as the *HHEX* CGI. Together, these studies indicate that while measured DNA methylation levels in some loci may be affected by cellular heterogeneity, a substantial proportion of loci may not be affected by this confounding effect.

We detected that DNA methylation levels in PMNCs and MNCs differed significantly in three out of four CGIs examined; i.e., in all CpG sites analyzed in the *HHEX* CGI, 7 of 8 CpG sites analyzed in the *KCNJ11* CGI and 1 of 4 CpG sites analyzed in the *KCNQ1* CGI but not in the *PM20D1* CGI. The gross difference observed in the *HHEX* CGI may reflect the fact that the *HHEX* gene is differentially expressed in the various blood cells [12-14]. Just as in whole blood DNA methylation measurements, this analysis may have been confounded by cellular heterogeneity because PMNCs and MNCs both consist of groups of cells. However, the fractionation split up the two white blood cell groups that affected whole blood DNA methylation measurements and their numbers are so dominant relative to the other groups that the analysis is likely to be minimally affected. Kerkel et al. have previously studied methylation in these fractions, and identified multiple differentially methylated loci [15]. Their analysis was however not described in detail. Nonetheless, together these studies indicate that differential methylation between white blood cell types may be relatively common.

Analysis of DNA methylation both in whole blood and blood fractions has allowed evaluation of the hypothesis that measured DNA methylation levels in whole blood can be confounded by cellular heterogeneity due to differential methylation levels in the various white blood cell types. We observed differential methylation between cell fractions in the *HHEX*, *KCNJ11* and *KCNQ1* CGIs and not in the *PM20D1* CGI, but we were only able to detect a significant effect due to cellular heterogeneity on whole blood DNA methylation measurement outcomes for the *HHEX* CGI. However, the difference in DNA methylation

between fractions was very moderate in the *KCNJ11* and *KCNQ1* CGIs and in the *KCNQ1* CGI only one of four CpG sites was differentially methylated. It is therefore possible that the effect of cellular heterogeneity on measurement outcomes for the *KCNJ11* and *KCNQ1* CGIs, if any, is exceedingly subtle, and thus undetectable by the methods we employed. It is therefore our view that these results support the hypothesis, and that they suggest a need to control for cellular heterogeneity in the analysis of methylation in blood cells.

Since the confounding effect would only be observed when both the genomic region of interest is differentially methylated amongst white blood cell types, and when there is blood cell count heterogeneity in the individuals being compared, controlling for this problem may be addressed in different ways depending on available data. Differences in white blood cell composition may be assessed if blood cell counts for the individuals under investigation are available. Alternatively, subjects can be paired with controls that are concordant in terms of cellular composition prior to the analysis. Furthermore, whole blood can be fractionated to assess possible differential methylation in the area of interest. This may be done with the Ficoll medium method used here which is relatively easy to perform, but due to heterogeneity in the fractions, as noted previously, this approach may not be sufficient to address the problem. Finally, referring to the literature may be advisable to assess the risk of altered blood cell counts in the groups of individuals under study. For example white blood cell counts have been shown to be associated with the development of cancers [16] and coronary heart disease [17]. This raises the issue that whenever there is a difference in cell fractions associated with disease, an adjustment for blood cell proportions could be essential for better controlled analyses.

The different approaches may cause inconsistent results, and therefore it is important to standardize methods for this correction. As has been discussed previously [10], adjusting for white blood cell counts can be achieved with standard statistical approaches. Such an approach may be well suited for that purpose since such data is presumably readily available at many laboratories conducting experiments on whole blood DNA. This could be achieved in two ways: One is to use multiple variables accounting for the absolute number of each cell

type (commonly five; neutrophils, lymphocytes, monocytes, basophils and eosinophils) or alternatively use a single variable accounting for the proportion of one cell type. Using a single variable is more appealing because the other option would reduce the number of degrees of freedom. However, to be able to correct for the confounding effect of cellular heterogeneity in statistical models by using a variable accounting for the proportional number of one cell type, there needs to be a correlation between methylation levels in the different types of white blood cells. Our results indicate that in the analyzed CGIs, methylation patterns across the corresponding CpG sites within a CGI are very similar between the different cell types irrespective of demonstrable differences in the cell specific absolute methylation levels. Our analysis therefore suggests that use of a single variable to account for the proportional number of a single cell type (e.g., neutrophils or lymphocytes) in statistical analyses might be sufficient to correct for the confounding effect of cellular heterogeneity on DNA methylation measurements conducted using whole blood DNA.

Our findings may not only be relevant for methylation measurements using whole blood DNA. Other tissues are samples of different types of cells as well, so a similar problem could affect measurements in these tissues. Our data indicate that although methylation levels may differ between blood cell types in some loci, the methylation pattern may at the same time be very similar (as indicated by the high correlation between methylation levels). This is in agreement with previous studies which have shown that different cells and tissues, even from separate germ layers, generally have similar DNA methylation patterns [2,10,18]. If blood cell DNA methylation measurements could be used as surrogates for methylation in other tissues based on this feature, it might be preferable to use blood.

DNA methylation levels are sometimes assessed in a global manner, assaying CpG sites across the entire genome. Since our study was conducted in a gene-specific manner the results may not apply to global DNA methylation measurements. Indeed, in a previous study using LUMA to estimate global methylation, we report no association between methylation levels and white blood cell counts [19]. However, as mentioned above, Wu et al.

report that global methylation levels in PMNCs, as measured by LUMA, are significantly higher than in MNCs and are not correlated [8]. In the same study, results from three other assays for global DNA methylation showed no association between PMNCs and MNCs methylation levels. It is therefore possible that global methylation measurements are also confounded by cellular heterogeneity. A more detailed analysis, including comparison on the association between global methylation levels in whole blood and cellular composition, such as in the present study, should be conducted in order to extend these observations.

The results from the present study call for an analysis of larger number of CpG sites to reveal the full extent of how confounding effects may influence analyses on DNA methylation conducted using whole blood DNA. It is important to assess whether measured methylation levels at a considerable amount of loci are affected by this effect. Second, it would be of value to study whether methylation of CpGs positioned in certain genes is more prone to be affected by this factor than others (e.g., in genes that are differentially expressed in the different cell subtypes such as *HHEX*). Finally, it would be interesting to investigate whether certain sequences (e.g., introns, exons, CGIs, CGI shores, transcription start sites or promoter regions) are more likely to be affected by this confounding effect.

Materials and methods

Ethical statement

The Age, Gene/Environment Susceptibility (AGES)-Reykjavik [20] and the Risk Evaluation For Infarct Estimates (REFINE)-Reykjavik studies are approved by the Icelandic National Bioethics Committee (VSN: 05-112, VSN: 00-063) and the Data Protection Authority. All participants gave written informed consent on arrival to the clinic.

Samples

Samples used in the present study were obtained from two cohort studies conducted at the Icelandic Heart Association, the AGES-Reykjavik [20] and the REFINE-Reykjavik studies. Whole blood DNA samples, which were analyzed independently for each CGI assayed in the study, were obtained from both the AGES-Reykjavik and the REFINE-Reykjavik studies (n=191). Blood was collected from individuals taking part in the REFINE-Reykjavik study (n=20), and these samples subsequently used for DNA extraction from both whole blood and two whole blood cell fractions (see details in next section). Three DNA samples were therefore obtained from each blood sample. All three DNA samples from all the 20 individuals were analyzed for each of the four CGIs assayed in the study. Both the whole blood DNA samples we obtained and the blood samples we collected were randomly selected from apparently healthy men and women. The age range of all individuals included in the study (n=211) was 22-96 years, and ~45% were males. An overview of the total number of whole blood DNA samples analyzed per CGI, and their overlap is provided in figure S1

Briefly, AGES-Reykjavik study was the last of seven visits in the Reykjavik Study, a population-based cohort study initiated in 1967, inviting all Reykjavik inhabitants born between 1907 and 1935 to participate. In this final visit 5764 of the surviving members were recruited. REFINE-Reykjavik is a prospective study on risk factors and cause of atherosclerotic disease in a population of Icelandic people. The main goal of the study is to improve the predictability of cardiovascular disease risk estimates. The study was initiated in

2005 and recruitment of the first phase was completed in spring 2011 recruiting 6942 men and women born in the years 1936-1980 living in the Reykjavik city area.

DNA isolation

Whole blood was fractionated by density gradient centrifugation using Histopaque-1077 Ficoll medium and Accuspin™ Tubes (Sigma-Aldrich, catalog numbers (cat.nr.): 10771 and A1930 respectively). The mononuclear cell fraction was extracted from the serum/medium boundary and the polymorphonuclear cell fraction from the bottom of the tubes. The blood samples were processed as “fresh” as possible, never later than 4 hours after the blood draw.

A simple salting out method was used for DNA extraction, based on an extraction method developed by Scotlab Bioscience (Coatbridge, Scotland, UK). The DNA was dissolved in TE buffer and its concentration measured using UV absorbance quantification (260 nm) on a Spectramax M2 (Molecular Devices, Sunnyvale, CA, USA) microplate reader.

Blood cell counts

For all participants, white blood cells (monocytes, lymphocytes, eosinophils, basophils and neutrophils) were counted in whole blood by an automated cell counter, Coulter HmX AL Hematology Analyzer (Beckman Coulter, High Wycombe, England, UK).

Bisulfite conversion of DNA samples

Bisulfite conversion of DNA samples was carried out using the EZ DNA Methylation™ kit (Zymo Research, cat.nr.: D5004) following the manufacturer's instructions. When the DNA was not analyzed immediately following the conversion process it was stored at -20°C for later use. DNA from blood fractions and the corresponding whole blood DNA for each individual was converted in the same batch.

345 *Analysis of DNA methylation*

346 DNA methylation was analyzed at four CGIs chosen to represent regions with a range
347 of inter-individual variation in DNA methylation levels from very low to very high. The *HHEX*,
348 *KCNJ11* and *KCNQ1* CGIs had been studied previously at the IHA (unpublished data) and
349 were chosen to represent low to intermediate variability regions while the *PM20D1* CGI was
350 selected from our previous, published work to represent a highly variable region [21]. More
351 specifically, the CGIs were selected from a larger set of CGIs based on two criteria. First, on
352 basis of the size of the inter-individual variability present at each CGI so as to select CGIs
353 representing a spectrum of variability from very low to very high and second, on basis of
354 which CGI in each variability category had available data on DNA methylation in the largest
355 number of whole blood DNA .The assays were designed to analyze DNA methylation levels
356 in CGIs, located using the University of California, Santa Cruz genome browser (Human
357 March 2006 NCBI36/hg18 assembly) [11].

358 Primer sets (forward and reverse PCR primers, one tagged with biotin, and a
359 sequencing primer) were designed using PyroMark Assay Design software (version 2.0.1.15,
360 QIAGEN, Hilden, Germany). Primer sequences and genomic positions of the CpG sites
361 analyzed in each assay are listed in tables S1 and S2. A 30 µl PCR was carried out on a
362 2720 Thermal cycler (Applied Biosystems, Foster City, CA, USA) using 1X TITANIUM Taq
363 polymerase (Clontech, cat.nr.: 639220) or 3 units OneTaq™ Hot Start polymerase (New
364 England Biolabs, cat.nr.:M0481L), 1X Standard Taq Reaction Buffer (New England Biolabs,
365 cat.nr.: B90145), 0.2 mM dNTP (New England Biolabs, cat.nr.: N04465), 0,25 µM of each
366 primer (Sigma-Aldrich) and 3 µl of bisulfite converted DNA. PCR cycling conditions for all
367 assays were as follows; 2 minutes at 96°C, followed by 40 cycles of 90s at 96°C, 90s at 62°C
368 and 90s at 72°C and finally 72°C for 10 minutes after cycling.

369 The biotinilated sequencing template was extracted from the PCR product mixture by
370 annealing with streptavidin coated sepharose beads (Streptavidin Sepharose™ High
371 Performance, GE Healthcare, cat.nr.: 17-5113-01). The template was subsequently washed
372 in a series of steps using a Vacuum prep workstation (QIAGEN cat.nr.: 9001518) and finally

released onto a sequencing plate (QIAGEN, cat.nr.: 979201) containing annealing buffer (QIAGEN, cat.nr.: 979309) with the appropriate sequencing primer. The samples were analyzed for methylation at each CpG site using a PyroMark Q24 pyrosequencer (QIAGEN) and PyroMark™ Gold Q24 reagents (QIAGEN, cat.nr.: 97082).

Data analysis

Pyrograms from the pyrosequencing reactions were analyzed with the "PyroMark Q24 Software" (v1.0.10, QIAGEN). Methylation levels were calculated as the ratio between peak heights for methylated C's and the sum of methylated and unmethylated C's for each CpG site. Default software settings were used for quality assessment of the pyrograms per CpG site and measurements that failed the assessment were discarded when appropriate. Consequently, some individuals had missing values for one or more CpG site and were analyzed separately. To verify that the assays were robust, the measurements were partly replicated (analysis not shown). For the replicated data, average methylation from the two measurements was used in the subsequent analysis. We tested for batch effects introduced by use of two brands of polymerases and only pooled data acquired through use of the two polymerases lacking any significant batch effects (data not shown). Finally, outliers were defined as values outside mean ± 2.698s (where s is standard deviation) per CpG site. For a standard Gaussian distribution, this criterion defines 0.35% of the data farthest from the mean in both directions as outliers. Individuals with one or more outliers were analyzed separately

In total, whole blood DNA methylation data was obtained for 179 individuals for the *HHEX* CGI, 64 individuals for the *KCNJ11* CGI, 50 individuals for the *KCNQ1* CGI and 59 individuals for the *PM20D1* CGI (figure S1). For the *HHEX* CGI, one or more outliers were detected in the CpGs studied for seven individuals, and measurement of methylation at one or more CpG sites failed the quality assessment in an additional three samples. For the *KCNJ11* CGI, measurements for five samples failed quality assessment at one or more CpG sites and five outliers were present. For the *KCNQ1* CGI, a single outlier was present, but no

missing values. No outliers were present in the data for the *PM20D1* CGI and none of the measurements failed the quality assessment. Successful and reliable measurements for all corresponding CpG sites in 169 samples for the *HHEX* CGI, 54 for the *KCNJ11* CGI, 49 for the *KCNQ1* CGI and 59 for the *PM20D1* CGI were therefore obtained from whole blood DNA and used in the subsequent analysis. The proportion of variation in methylation levels between individuals explained by differential white blood cell counts was estimated from unadjusted mixed model analysis of the data using PROC MIXED in SAS Enterprise Guide version 4.2 using a random intercept term to account for the correlation within a person. Since R^2 cannot be obtained directly from such analysis, two models were applied, an intercept only model containing only CpG sites as fixed effects and a full model where additionally, the proportional number of a specific cell type was added to the intercept model as fixed effect. R^2 was then calculated from the residual variance (v_r) and variance of the random intercept (v_s) terms using the formula $R^2=(V_i-V_f)/V_i$ where $V_i=v_r+v_s$ for the intercept only model and $V_f=v_r+v_s$ for the full model.

Measurements on DNA from the two blood cell fractions, PMNC and MNCs, were conducted on the same 20 individuals (40 samples total) for all four CGIs. Measurements were successfully obtained for all CpG sites at all CGIs. A single outlier was present in the data for the *HHEX*, *KCNJ11* and *KCNQ1* CGIs, but none in the data for the *PM20D1* CGI. Due to the limited number of samples used in this analysis, this data was not excluded. Excluding the data did however not affect the analysis (analysis not shown). The data were analyzed using non-parametric statistics to avoid making a generalized assumption about the distribution of our data, which may differ between loci. Paired Wilcoxon signed rank test was used to assess statistical differences in methylation levels between the two cell populations and their correlation assessed with Spearman correlation coefficient using R version 2.12.2.

Acknowledgements

The researchers are indebted to the participants for their willingness to participate in the study.

References

1. Eckhardt F, Lewin J, Cortese R, Rakyan VK, Attwood J, et al. (2006) DNA methylation profiling of human chromosomes 6, 20 and 22. Nat Genet 38: 1378-1385.

2. Fan S, Zhang X (2009) CpG island methylation pattern in different human tissues and its correlation with gene expression. Biochem Biophys Res Commun 383: 421-425.

3. Irizarry RA, Ladd-Acosta C, Wen B, Wu Z, Montano C, et al. (2009) The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. Nat Genet 41: 178-186.

4. Illingworth R, Kerr A, DeSousa D, Jørgensen H, Ellis P, et al. (2008) A Novel CpG Island Set Identifies Tissue-Specific Methylation at Developmental Gene Loci. PLoS Biol 6: e22.

5. De Bustos C, Ramos E, Young JM, Tran RK, Menzel U, et al. (2009) Tissue-specific variation in DNA methylation levels along human chromosome I. Epigenetics Chromatin 2: 7.

6. Kochanek S, Toth M, Dehmel A, Renz D, Doerfler W (1990) Interindividual concordance of methylation profiles in human genes for tumor necrosis factors α and β. Proc Natl Acad Sci USA 87: 8830-8834.

7. Kochanek S, Radbruch A, Tesch H, Renz D, Doerfler W (1991) DNA methylation profiles in the human genes for tumor necrosis factors *α* and *β* in subpopulations of leukocytes and in leukemias. Proc Natl Acad Sci USA 88: 5759-5763.

8. Wu HC, Delgado-Cruzata L, Flom JD, Kappil M, Ferris JS, et al. (2011) Global methylation profiles in DNA from different blood cell types. Epigenetics 6: 76-85.

9. Martin GM (2005) Epigenetic drift in aging identical twins. Proc Natl Acad Sci USA 102: 10413-10414.

10. Talens RP, Boomsma DI, Tobi EW, Kremer D, Jukema JW, et al. (2010) Variation, patterns, and temporal stability of DNA methylation: considerations for epigenetic epidemiology. Faseb J 24: 3135-3144.

11. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, et al. (2002) The Human Genome Browser at UCSC. Genome Res 12: 996-1006.

12. Crompton MR, Bartlett TJ, MacGregor AD, Manfioletti G, Buratti E, et al. (1992) Identification of a novel vertebrate homeobox gene expressed in haematopoietic cells. Nucleic Acids Res 20: 5661-5667.

13. Bedford FK, Ashworth A, Enver T, Wiedemann LM (1993) *HEX*: a novel homeobox gene expressed during haematopoiesis and conserved between mouse and human. Nucleic Acids Res 21: 1245-1249.

14. Manfioletti G, Gattei V, Buratti E, Rustighi A, De Iuliis A, et al. (1995) Differential expression of a novel proline-rich homeobox gene (Prh) in human hematolymphopoietic cells. Blood 85: 1237-1245.

15. Kerkel K, Schupf N, Hatta K, Pang D, Salas M, et al. (2010) Altered DNA Methylation in Leukocytes with Trisomy 21. PLoS Genet 6: e1001212.

16. Margolis KL, Rodabough RJ, Thomson CA, Lopez AM, McTiernan A (2007) Prospective Study of Leukocyte Count as a Predictor of Incident Breast, Colorectal, Endometrial, and Lung Cancer and Mortality in Postmenopausal Women. Arch Intern Med 167: 1837-1844.

17. Danesh J, Collins R, Appleby P, Peto R (1998) Association of fibrinogen, C-reactive protein, albumin, or leukocyte count with coronary heart disease: meta-analyses of prospective studies. JAMA 279: 1477-1482.

18. Byun HM, Siegmund KD, Pan F, Weisenberger DJ, Kanel G, et al. (2009) Epigenetic profiling of somatic tissues from human autopsy specimens identifies tissue- and individual-specific DNA methylation patterns. Hum Mol Genet 18: 4808-4817.

19. Bjornsson HT, Sigurdsson MI, Fallin MD, Irizarry RA, Aspelund T, et al. (2008) Intra-individual Change Over Time in DNA Methylation With Familial Clustering. Jama 299: 2877-2883.

20. Harris TB, Launer LJ, Eiriksdottir G, Kjartansson O, Jonsson PV, et al. (2007) Age, Gene/Environment Susceptibility-Reykjavik Study: multidisciplinary applied phenomics. Am J Epidemiol 165: 1076-1087.

21. Feinberg AP, Irizarry RA, Fradin D, Aryee MJ, Murakami P, et al. (2010) Personalized Epigenomic Signatures That Are Stable Over Time and Covary with Body Mass Index. Sci Transl Med 2: 49ra67.

Figure Legends

Figure 1. Percent DNA methylation in whole blood samples.

Percent DNA methylation (y-axis) in whole blood DNA per CpG site (x-axis) in four CGIs located in the *HHEX* (n=169), *KCNJ11* (n=54), *KCNQ1* (n=49) and *PM20D1* (n=59) genes respectively. Data for each CGI is depicted in a separate boxplot. Below each boxplot is a gene-map which roughly indicates the position of the analyzed CpG sites (adapted from the UCSC genome browser) [11]. Genes are depicted in blue, the exons as blocks, the introns as thin lines connecting the blocks, and the 5' and 3' untranslated regions as thin blocks at each end. CGIs are shown as green blocks. The genomic position depicted for each CGI is; 10:94,439,661-94,445,388 (chromosome:first base-last base) for the *HHEX* CGI, chr11:17,363,372-17,366,783 for the *KCNJ11* CGI, chr11:2,422,797-2,826,916 for the *KCNQ1* CGI and chr1:204,063,776-204,085,881 for the *PM20D1* CGI. An arrow indicates the direction of transcription and the position of the transcription start site.

Figure 2. Percent DNA methylation in mononuclear and polymorphonuclear cells.

Percent DNA methylation (y-axis) in mononuclear and polymorphonuclear cells (MNCs and PMNCs) per CpG site (x-axis) in four CGIs located in the *HHEX*, *KCNJ11*, *KCNQ1* and *PM20D1* genes respectively (n=20 each). Data for each CGI is depicted in a separate boxplot where measurements for MNCs are shown in red and for PMNCs in blue. The dotted lines separating the boxes indicate that at each CpG site a pair of data is being compared (i.e., for MNCs and PMNCs). Significantly ($p<0.05$) differentially methylated CpG sites (MNCs versus PMNCs DNA methylation) are indicated with an asterisk.

Figure 3. Correlation between DNA methylation in mononuclear and polymorphonuclear cells.

Comparison of DNA methylation levels measured in two cell fractions, mononuclear cells (MNCs) and polymorphonuclear cells (PMNCs). Percent methylation in PMNCs (y-axis) is plotted against percent methylation in MNCs (x-axis). Each dot represents the two

measurements for a single CpG per individual. The Spearman ρ for correlation between measurements in MNCs and PMNCs for each CGI is shown in the legend.

Figure S1. Venn diagram depicting the number of samples analyzed per CGI.

The diagram contains a set of 15 numbers that, when added together, represent the total number of individuals analyzed with DNA from whole blood. Each ellipse contains a set of numbers, that when added together represent the total number of individuals analyzed for a specific CGI. Finally, some individuals were analyzed for more than one CGI, and this is represented by the overlapping of ellipses.

Table S1. Genomic positions of the CpG sites analysed per locus.

Table S2. Primer sequences for the PCR assays used in the study.

Tables

Table 1. Proportion of variation in measured DNA methylation level accounted for by cellular heterogeneity

	Variance explained by cell proportion (%)				
CGI	Lymphocytes	Monocytes	Neutrophils	Eosinophils	Basophils
HHEX	35**	0	26**	0	0
KCNJ11	0	0	0	0	3*
KCNQ1	1	0	0	0	0
PM20D1	0	0	0	0	0

*p<0.05, **p<0.0001

Figure 1

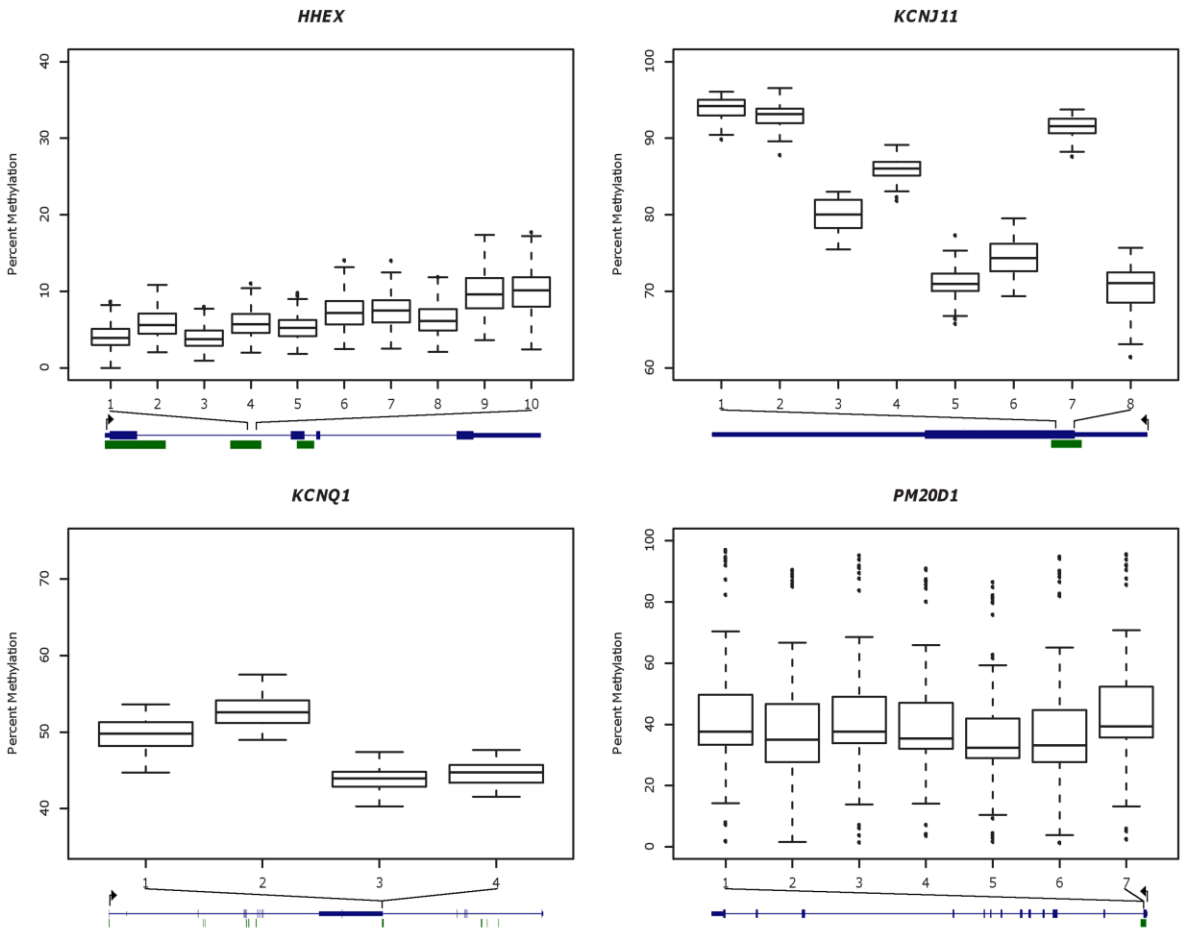


Figure 2

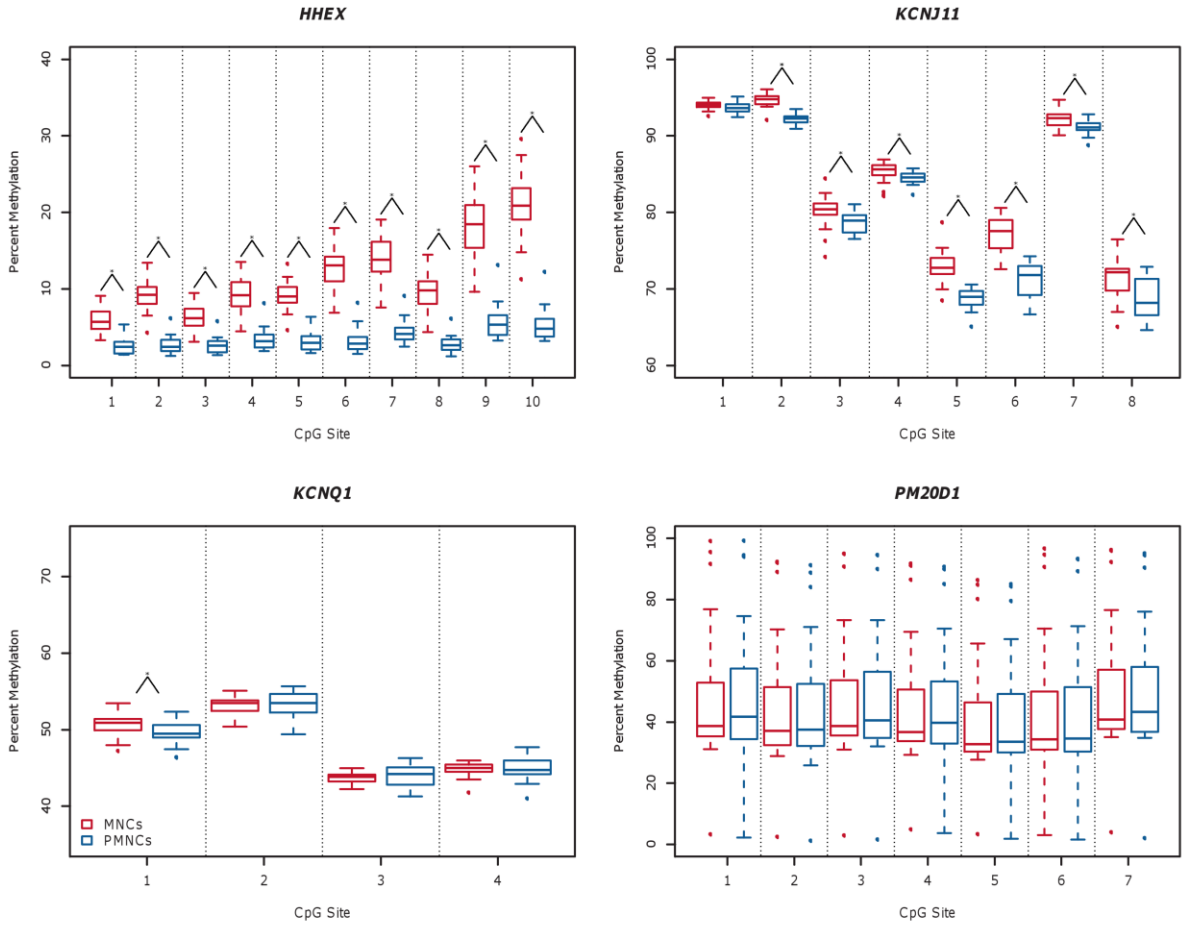
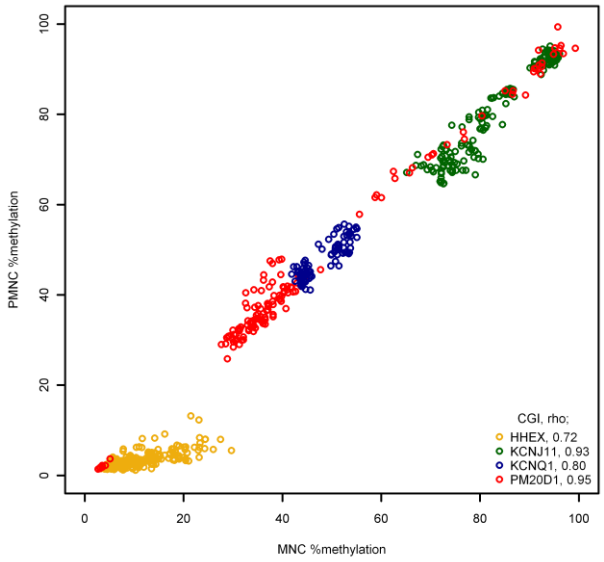

HHEX
KCNJ11
KCNQ1
PM20D1
Percent Methylation
CpG Site
MNCs
PMNCs

Figure 3


PMNC %methylation
MNC %methylation
0
20
40
60
80
100
CGI, rho;
HHEX, 0.72
KCNJ11, 0.93
KCNQ1, 0.80
PM20D1, 0.95

Supplementary materials

Figure S1:

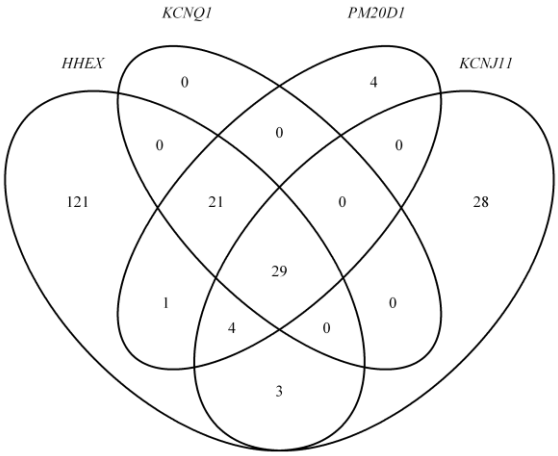


Table S1:

CGI	Chromosome	Position									
		1	2	3	4	5	6	7	8	9	10
HHEX	10	94441605	94441607	94441619	94441627	94441633	94441637	94441644	94441646	94441662	94441676
KCNJ11	11	17366204	17366187	17366178	17366168	17366135	17366129	17366123	17366114		
KCNQ1	11	2677095	2677111	2677115	2677117						
PM20D1	1	204085711	204085713	204085716	204085733	204085740	204085749	204085760			

* Human March 2006 NCBI36/hg18 assembly

Table S2:

CGI	Sequencing primer	Forward primer	Reverse primer
HHEX	GTTAGGATTGGAGGTTT	ATGTTGTTATAGTTTATGGGGTGGT	TTACCCCCTTAAATCTCCCTTAATA
KCNJ11	ATCACCCAAACCATACTATCC	GTTGTAGTTGTTTTTTTTGGATATAAAG	ACTCTACAATAAAACCCTAAACCAC
KCNQ1	GGTTAGGTTGTATTGTTG	GTATTGTTTAGGTTAGGTTGTATTGT	ACCCTCCCCATCTCTCTAA
PM20D1	GTTGAATTGAGAAGGGAT	ATGAGTATAGGTGGGTGAAG	ACCCTAATAACTATACTACTCCTAATTTTC