Chapter 7
Thesis Summary and discussion of results

Table 7.1: Summary of thesis results at candidate loci

	Characteristics of CpG methylation						Prenatal famine
	Level (%)		Cell population		Correlation		Period
Locus[a]	Mean	SD	Association[b]	Adjustable[c]	Longitudinal[d]	Blood vs. buccal[e]	P[f]
NR3C1	< 20	< 3.0	-				-
TNF	< 20	< 3.0	+	+			-
APOC1	< 20	3.0 – 6.0	+	+	+	+	-
IL10	20 – 40	> 6.0	+	-	-	-	↑
LEP	20 – 40	3.0 – 6.0	+	+	+	+	↑
KCNQ1OT1	20 – 40	< 3.0	-		-	-	-
ABCA1[m]	20 – 40	> 6.0	+	+			↑
GNASAB	20 – 40	3.0 – 6.0	-				-
GRB10	40 – 60	3.0 – 6.0	-				-
GNASAS	40 – 60	3.0 – 6.0	-				↑
MEG3	40 – 60	< 3.0	-				-
IGF2	40 – 60	3.0 – 6.0	-		+	-	↓[n]
CRH	60 – 80	> 6.0	+	+	+	+	-
IGF2R	60 – 80	> 6.0	-		+	+	-
INS	> 80	< 3.0	-		-	-	↓
FTO	> 80	< 3.0	-				-

a: Loci are ordered on mean level of methylation from lowest to highest

b: Significant influence of cell population heterogeneity on methylation variation: + = $p < 0.05$; - $p > 0.05$ (uncorrected for multiple testing)

c: The amount of influence of cellular heterogeneity is sufficiently small to allow statistical adjustment for it

d: Correlation of methylation variation between longitudinal blood samples (follow-up = 20 years): - = $\rho < 0.75$; + = $\rho > 0.75$

e: Correlation of methylation variation between blood and buccal cells: - = $\rho < 0.75$; + = $\rho > 0.75$

f: DNA methylation is associated with famine in the period around conception

g: DNA methylation is associated with famine in the last trimester of gestation (and early post natal life)

			Accumulated ageing related changes				Risk for CHD[l]	
	Specificity		Variation increase[j]		Environmental[k]			
L[g]	Sex[h]	Time[i]	Inter-ind-ivid-ual	Within-MZ-pair	Unique	Familial	Associa-tion	Sex specific
-		-						
-	-	+					-	-
-	M	-	++	++	+	-	-	-
-		-	-	+	+	-		
	-		+	++			-	-
↓	F	+	+	-	+	-	↑	F
-[n]	-[n]	+[n]	++	++	+	-	-	-
-		-	++	++	-	+		
-		-						
-	M	+	++	++			↑	F

h: The association with famine around conception is restricted to men (INS and LEP) or more pronounced in women (GNASAS)

i: The association with famine is restricted to only one period, or for GNASAS is in opposite direction between both periods

j: Significant (+) or substantial (++) increase of methylation variation (inter-individual and within pair discordance) calendar age

k: Attributability of the increase in methylation variation to the unique (individual) or the familial (shared) environment

l: Significant association of DNA methylation at the locus with risk of MI during 3 year follow-up is restricted to the women of the study

m:Information is on the methylated CpG sites at the 5'end of the assay (Chapter2)

n: These observations [14] are not included as a chapter of this thesis but were the study directly preceding Chapter 3

Epigenetic mechanisms regulate cellular gene expression potential without changing the genetic code [29]. Like the genetic sequence, epigenetic marks are faithfully transmitted during mitosis and are generally stable in differentiated cells, but in contrast with the static genome, the epigenome retains the capacity for dynamic changes in each individual cell. Epigenetic variation is therefore a topic of interest for research on ageing and its related common diseases [47,50]. In this thesis we focus on DNA methylation, which is the most studied layer of epigenetic information [28], and is correlated to the other epigenetic layers [21,25]. We used a combination of successive studies to investigate aspects of variation in DNA methylation, various sources generating such variation and its relation with risk for myocardial infarction (MI) at candidate loci for cardiovascular and metabolic diseases.

Thesis results

In **chapter two** we investigated the possibilities of epigenetic epidemiological research using samples stored in existing bio-repositories, most of which were originally designed for genetic research and usually contain DNA from blood drawn at baseline without longitudinal sampling. With reference epigenome browsers still under construction [88,89], we first established a reference of inter-individual variation in the methylation status of 104 CpG sites distributed over 16 loci (Table 7.1). These loci were selected because they map to candidate genes for cardiovascular and metabolic diseases and their epigenetic characteristics represent various potential targets for epigenetic regulation of gene expression (Chapter 1). We established that CpG methylation in the cell population of a tissue sample is a quantitative trait at these loci. We further found that, although both average methylation level and the amount of variation could vary substantially between adjacent CpG sites, the methylation of nearby CpG sites was generally correlated, which may have some practical analogy to the linkage disequilibrium blocks of genetic variation [159] and fits the common view that the regulatory effect of DNA methylation usually relies on the status of multiple CpG sites within a genomic region [94]. A recent whole methylome study confirmed the existence of such correlations and showed that they extent roughly a hundred base pairs to

either side [83]. We then investigated the concern that inter-individual variation in DNA methylation might be a reflection of leukocyte population heterogeneity [103]. We found that variation in DNA methylation at 10 loci was not influenced by leukocyte population heterogeneity, this influence was marginal at 5 loci, and it was substantial only at the *IL10* locus (Table 7.1). We described a statistical method to account for minor influences of cell type heterogeneity on DNA methylation variation. Recent studies have observed similar proportions of unaffected, marginally affected and strongly affected loci in a study on different candidate loci [246] and on a large set of individual CpG sites distributed throughout the genome [77].

Next we investigated the stability of variation in DNA methylation over time in longitudinal DNA samples of both blood (10-20 year time span) and buccal cells (2 – 8 year time span) in 34 individuals (Table 7.1). Although DNA methylation was not perfectly maintained at any locus, 5 of 8 loci assessed displayed a strong correlation ($\rho \geq 0.75$) between the time points in both tissues, indicating that prospective studies on DNA methylation are possible. Such locus specific temporal stability was also found in another longitudinal study [197]. DNA methylation levels of imprinted loci are commonly thought to be similar between tissues, since they originate in early development [69,106]. We finally investigated whether the variation in DNA methylation of blood may correspond to that of another tissue. We found that DNA methylation at 4 of 8 loci displayed a strong correlation ($\rho \geq 0.75$) between the recent blood and buccal cell samples, irrespective of imprinting status. A recent genome wide study reported indicated that DNA methylation patterns in blood could be a surrogate marker for cerebellum at some loci [234]. In all, depending on the locus of interest and with careful study design, some epigenetic research questions can be answered using DNA from existing biobanks.

In **chapter three** we investigated whether in utero exposure to the Dutch Famine of 1944-1945 is associated with persistent differences in epigenetic information at the candidate loci described in chapter 2, assessing the generality of our primary findings at the *IGF2* locus [14]. We compared DNA methylation between middle-aged individuals

who were born or conceived during the Dutch Famine and their same sex older or younger siblings. We found that DNA methylation was persistently altered in prenatally exposed individuals at 6 of the 15 loci assessed (Table 7.1). The direction of these exposure related changes was specific for the loci, with 5 showing higher, and 1 lower DNA methylation. Furthermore, the exposure related change in DNA methylation was sex-specific at 3 loci, and specific to the timing of the exposure at 5 loci. At 4 of these loci only famine exposure in the period around conception was associated with altered DNA methylation, whereas at the *GNASAS* locus famine exposure in both the period around conception and in the final three months of gestation resulted in altered DNA methylation in opposite directions. The combination of epigenetic changes specific to the locus, the type and timing of the exposure and the sex of the exposed can unlikely be explained solely by epigenetic damage due to a deficiency in methyl donors. Further research may reveal whether and which other mechanisms underlie these observation, including the proposed possibility of an adaptive response being recorded in the epigenome [178].

More prenatal conditions have been reported to produce persistent epigenetic changes, including maternal micronutrient deficiency [225], folate supplementation [120,219], maternal smoking [67,226], and placental morphology [217]. Such observations in humans resemble results from intervention studies in animal models [13,63,184]. In animal studies the epigenetic changes were also associated with phenotypic changes related to disease development, forwarding epigenetic alterations as a plausible molecular mechanism for the developmental origins hypothesis of adult disease, which poses that adverse conditions during development and early life increase the risk for age related common diseases like CHD [165]. However, since the Dutch Famine related epigenetic differences were observed in blood, it remains unclear whether and how they are involved in the phenotypic observations related to the Dutch Famine that indicate a predisposition to metabolic and cardiovascular diseases [167,168]. In any case, the persistence and specificity of epigenetic information found associated with early life exposures makes such marks attractive potential reporters for developmental conditions, for which the genetic sequence itself cannot provide any clue.

Further, epigenetic marks found associated with conditions around conception and early embryogenesis might even correlate across the tissues that developed from the same progenitor stem cells affected at the time of the exposure, thus potentially marking the epigenetic state of inaccessible tissues [69,94].

In **chapter four** we investigated the susceptibility of epigenetic information to changes during the whole adult life span, having observed in Chapter 2 that every locus assessed displayed some degree of epigenetic changeability over 20 years. This chapter aims to extend findings of previous studies investigating this question with cross-sectional and longitudinal designs on smaller sample sizes with narrower age ranges [130,197,247]. In 230 monozygotic twin pairs (MZ twins) ranging from young to elderly adults (18 – 88 years old) we compared the amount of inter-individual variation and within-pair discordance in both global DNA methylation and DNA methylation at 7 of the loci, assessed in chapter 2. Increases in these measures of epigenetic variation in the population were used as a proxy for combined stochastic and environmentally driven epigenetic changes in individuals and were both found to be largely independent from leukocyte population heterogeneity. Although at all ages global DNA methylation showed little variation, its variation increased significantly with age, and results at most candidate loci further indicated that epigenetic changes accumulate gradually with increasing calendar age. Recent genome-wide studies also report associations of DNA methylation with age [58,248]. The rate of accumulation appeared specific to the locus with some loci (e.g. *IGF2*, *INS*, *LEP*) displaying a substantial accumulation of (random) epigenetic changes throughout adult life, while inter-individual variation at *KCNQ1OT1* and within-pair discordance at *GNASAS* remained stable (Table 7.1).

At the *CRH* locus mean methylation levels were similar at all ages while increases in variation were substantial, potentially indicating that changing methylation marks may yield another trait in epigenetic research. Epigenetic instability is proposed to be both a stochastic initiator of disease development [50] and an important biological mediator of environmentally driven disease risk [68]. The age related increase in DNA methylation variation fits both

slightly contrasting views. Stochastic DNA methylation changes without an environmental component are likely related to imperfect epigenetic maintenance processes during mitosis [51,205]. However, the directed DNA methylation changes that result from environmental exposures related to lifestyle and living conditions [72,77] may likely accumulate as changes in (partly) opposite directions, since conditions change across a lifetime. Both processes will manifest as random epigenetic changes with age, which is commonly called epigenetic drift. We observed that most epigenetic variation and most of its age-related increase was generally attributable to individual factors (Table 7.1), which fits both causes for epigenetic instability. Remarkably, although longitudinal examination showed that the increase in methylation variation was continued into old age, there were still old twin pairs with virtually identical DNA methylation signatures, indicating that either the epigenome of some individuals is less changeable, or that these individuals are better capable of correcting epigenetic changes. The eventual build-up of dysfunctional epigenetic changes in each cell is thought to result in an age related increase in the proportion of cells in a tissue that are epigenetically dysregulated [50,205], a state that has been forwarded as a possible molecular mechanism in the process of ageing [58,59]. Our observations at old age may thus forward epigenetic marks in the elderly as a potential indicator of biological age [249], rather than calendar age [250]. Future research into the causes and phenotypic consequences of epigenetic instability may be able to shed light on such questions.

In this chapter we examined 5 of the 6 loci at which DNA methylation was found associated with prenatal famine [14]. We found that 4 of these loci showed considerable susceptibility to epigenetic drift, which would seem to contradict the observed persistence of the epigenetic changes induced by the Dutch Famine (Table 7.1). However, observations at the *IGF2* locus showed that the differences induced by the Dutch Famine were not similar for every exposed and non-exposed sibling pair, with the persistence of the difference manifested at the group level [14]. We observed in this chapter that there were still old twin pairs with similar methylation levels, and, although the loci show substantial increases in variation with age, changes

in the population averages seem proportionately smaller across a lifespan. A plausible interpretation of this would be that epigenetic differences related to the (prenatal) environment might be most pronounced in the population just after the event, and then slowly fade as epigenetic variation around the average increases with age through accumulated (random) epigenetic changes in each individual. An alternative explanation may view the Dutch Famine associated differences in DNA methylation as a persistently accelerated DNA methylation age in the prenatally exposed group, with the age related increase in epigenetic variation at these loci relating to an increased variation in DNA methylation age [251]. Both interpretations fit the developmental component and the stochastic variation of ageing [59,178].

In **chapter five** we investigated whether, in line with the developmental origins hypothesis [165], DNA methylation was related to risk for CHD at the 6 candidate loci, at which it was associated with prenatal famine [14] (Chapter 3). Within the placebo group of the PROSPER trial on Pravastatin and the risk of CHD [118,218], we compared all individuals who were event free at baseline and developed myocardial infarction (MI) during 3 years follow-up (n = 122) with a similar sized control group. Due to our results from the Dutch Famine study, we tested for and found a significant sex interaction for *INS* and *GNASAS*. Higher DNA methylation at these loci was associated with an increased risk for MI among women (Table 7.1) with the association strongest for individuals displaying hypermethylation at both loci. We observed no epigenetic associations among men, which could be due to an aspect of study design such as inclusion age [230] and / or may have a biological reason. This study was one of the first to show a locus specific epigenetic association with CHD, but more are emerging. A case-control study [252] on coronary artery disease (CAD) found an association with methylation of a candidate gene in the 9p21 locus, implicated in CHD by genome wide association studies [253]. A smaller retrospective study on familial hypercholesterolemia patients [254] reported an association between CAD and methylation at the exact same CpG sites of *ABCA1* as measured in this thesis. Our prospective study on MI risk found no association at *ABCA1*, the simplest explanation for which

would be differences in the study population (event free vs. hypercholesterolemic) and the study design (prospective vs. retrospective). The associations we found were independent of the well-established traditional risk factors of CHD (i.e. lipid profile, hypertension, Type 2 diabetes), suggesting that DNA methylation at specific loci may be an epigenetic marker for disease risk. This view has been substantiated by two recent longitudinal studies reporting associations in blood between persistent DNA methylation at candidate loci and obesity or type 2 diabetes [48,49]. With our investigation limited to blood, our results do not directly implicate an epigenetic mechanism in disease development.

Do results in this chapter reflect a developmental component of CHD? The original developmental origins hypothesis (Figure 7.1 A) was formulated on epidemiological studies in which various proxies for poor conditions in early life were associated with an increased incidence and severity of CHD in middle age [8,165]. In line with this hypothesis, epidemiological studies on both Dutch Famine cohorts [167,168] indicate that prenatally exposed individuals have a higher incidence of its phenotypic pre-stages (Figure 7.1 C). Results from a wealth of intervention studies on animal models [66] show that adverse early life conditions result in epigenetic changes specific to the insult and the locus, and that these associations coincide with proxies for the pre-stages of CHD (Figure 7.1 B). Both *INS* and *GNASAS* methylation is important in regulating fetal growth [180,201]. Chapter 3 showed that persistently different methylation at *INS* and *GNASAS* was associated with prenatal exposure to the Dutch Famine (Figure 7.1 C). In this chapter we observe that higher methylation at both loci is associated with risk for CHD independently of the traditional CHD risk factors (Figure 7.1 D). The associations at *GNASAS* are a complete match between prenatal (peri-conception) susceptibility and risk for CHD, both in terms of direction of the difference as in terms of sex-specificity (Table 7.1). Chapter 4 shows that methylation at *GNASAS* is mostly inert to epigenetic drift, making an unknown developmental component the most likely explanation for its association with CHD. The associations at *INS* are in opposite direction and in the other sex, prenatally exposed men have lower DNA methylation and women at risk for CHD have higher DNA methylation. Further, *INS* methylation is susceptible to changes, stochastic

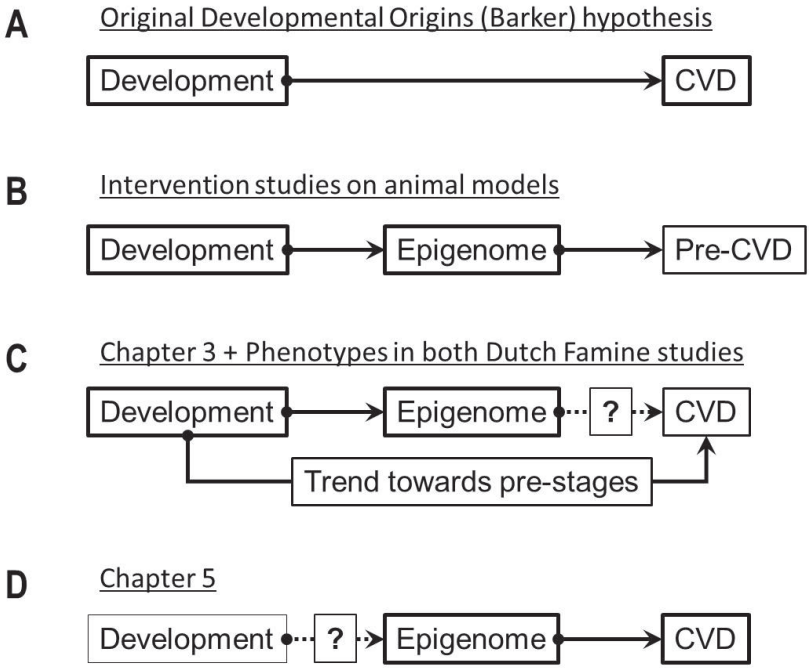

Figure 7.1: **Schematic summary of results related to DOHaD. A**) The original developmental origins hypothesis was formulated on epidemiological investigations that found a relation between birth and early childhood in poor areas and increased risk and severity of ischemic heart attack in middle age [8]. **B**) Many intervention studies on animal models provide convincing evidence that prenatal conditions affect epigenetic information at CHD candidate loci resulting in phenotypes that lead to CHD development in humans [65,66]. **C**) In Chapter 3 of this thesis we present evidence that prenatal exposure to the Dutch Famine of 1944-1945 affected DNA methylation at CHD candidate loci, but the study was not fit due to size and age of the participants to investigate a further association between DNA methylation and CHD directly. However, epidemiological studies on both Dutch Famine cohorts do indicate the prenatal exposure to it is associated with increased manifestation of the pre-stages of CHD [167,168], although DNA methylation is not indicated in these studies as a potential explanation. **D**) in Chapter 5 we find an association of risk for MI with DNA methylation at 2 of the loci that showed persistent changes in Chapter 3. Although we have no specific information on the early life conditions of the study participants, we do know that they were born in the same period as the populations, that were the basis of the developmental origins hypothesis.

and / or environmentally driven, during the adult period (Chapter 4). Thus, although a developmental explanation for its association with CHD risk is still plausible, it is equally plausible to explain its hypermethylation in cases with epigenetic events in adult life. This can only be resolved with data on both prenatal and adult life conditions of the PROSPER participants.

In **chapter six** we performed a genome-wide investigation on autopsy derived samples from 6 individuals (blood, subcutaneous fat, visceral fat, muscle, liver, pancreas, and spleen) using the recently launched Illumina 450k DNA methylation chip on the capacity of DNA methylation in blood to mark that of internal tissues. Principal component analysis (PCA) revealed that DNA methylation patterns remain stable within at least 24 hours after death, demonstrating that autopsy derived tissue samples can be used to create reference DNA methylomes of inaccessible tissues. PCA further showed that genome-wide variation in DNA methylation is primarily due to tissue differences, dismissing the use of surrogate tissues to mark whole methylome patterns of internal tissues. In contrast, we observed that average DNA methylation levels were similar between blood and each internal tissue at most CpG sites, and across all tissues at roughly a quarter of CpG sites. Both observations were in line with existing literature [100–102].

Chapter 2 showed that at specific loci DNA methylation in blood may correlate with DNA methylation in another tissue, which a recent genome wide study expanded upon [234]. We therefore inspected correlation of CpG methylation at individual sites across the tissues, focusing mostly on CpG rich areas for technical reasons, and found that over 30 % of CpG sites showed a strong correlation ($r \geq 0.75$) between blood and another tissue. At the majority of CpG sites such correlations were between blood and one tissue exclusively, although at a sizeable minority they were found across multiple tissues. At a few CpG sites we observed correlation across all tissues which likely reflects the effect of unknown genetic variation [92,243]. Strong correlation coefficients were not associated with average DNA methylation, its variation, or its difference between the tissues. Our observations were repeated in a recent study on the same dataset extended with additional samples and analyzed using different

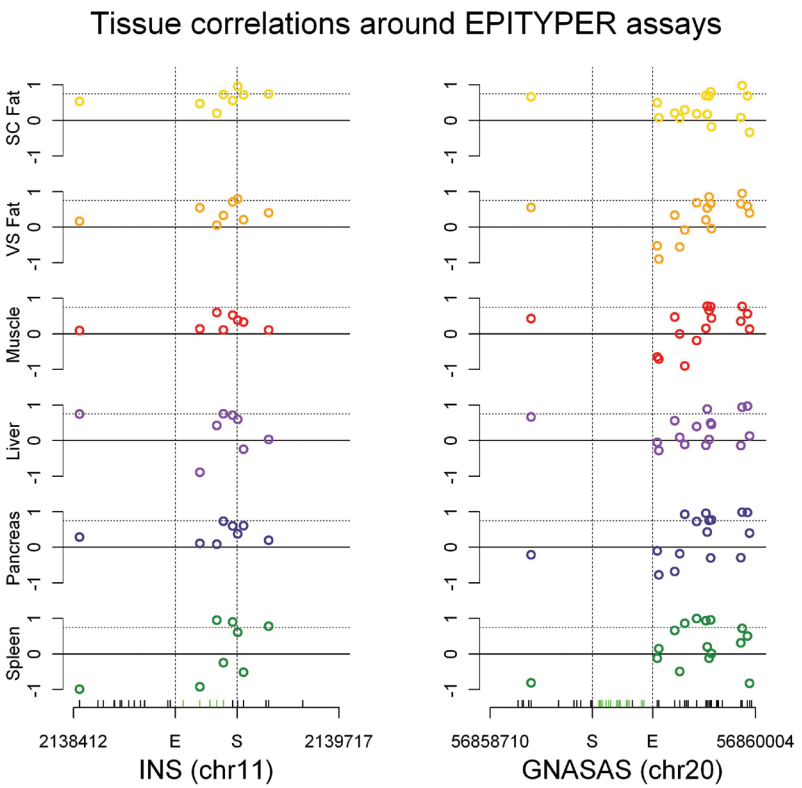

Figure 7.2: Correlation of CpG methylation between blood and spleen (green), pancreas (blue), liver (purple), skeletal muscle (red), visceral fat (orange), and subcutaneous fat (yellow) at CpG sites within and around 500 bp from either start (vertical dotted line at S) or end (vertical dotted line at E) of the SEQUENOM EPITYPER assays for the candidate loci *INS* (left) and *GNASAS* (right) measured in Chapters 2 to 5. The numbers on the x-axis correspond to the chromosomal coordinates (hg 18) of the assay +/- 500 bp, the tick-marks on the x-axis represent all the CpG sites, the green ticks indicate CpG sites of which the methylation was called using the EPITYPER platform. The colored dots represent correlation between blood and the tissues at CpG sites called using the 450K ILLUMINA BeadChip Array in Chapter 6. Horizontal dotted lines are for reference, representing a correlation coefficient of 0.75.

statistical methods [255]. Further, a recent whole methylome study suggested that blood and buccal cells can both be used as surrogate tissues for internal tissues [256]. There are more peripheral tissues that can be obtained with minimal (e.g. buccal cells and epidermis) [107,108] or moderate (e.g. biopsies of subcutaneous fat or muscle) [109,110] invasiveness. As the capacity of a surrogate tissue to mark a target tissue appears locus and tissue specific a reference of DNA methylation correlation across tissues will be of value in choosing the relevant surrogate tissues that mark the target tissue in future epigenetic epidemiological studies.

The current chapter is a pilot when it comes to finding locus specific surrogate tissues, but this data may already help interpret the associations found at *INS* and *GNASAS* in this thesis. We therefore plotted the correlation between blood and the internal tissues for the CpGs on and around ($^{+}/_{-}$ 500 bp) both Sequenom EPITYPER assays (amplicon) measured in chapters 2 to 5 (Figure 7.2). For *INS*, 3 of the CpG sites called in the amplicon were assessed in this chapter, as were 3 CpG site within 60 bp of the Start of the amplicon, and 1 CpG site 470 bp away from the End. The CpG sites around the Start of the *INS* amplicon show some promising correlations between blood and the other tissues. Insulin, under the influence of DNA methylation at *INS*, is produced in pancreatic β-cells [200], which constitute only a tiny fraction of the whole pancreas, yet 1 CpG site shows a correlation with some promise between blood and whole pancreas (Figure 7.2). For *GNASAS* none of the CpG sites called within the amplicon were assessed in this chapter and only 1 CpG site 300 bp from the Start of the amplicon, but within 500 bp from the End of the amplicon we assessed 15 CpG sites in this chapter, of which the CpGs 155 - 455 bp away show promising correlations between blood and other tissues. Higher DNA methylation at *GNASAS* is associated with higher expression of Gαs, a key component of the intracellular cAMP signaling, in adipose tissue (both brown and white), the pituitary and the thyroid gland [201,223]. A few CpG sites show promising correlations between blood and subcutaneous fat or visceral fat. Although there is evidence that small changes in percentage DNA methylation may explain almost half the variance in gene expression [188], and the cAMP and insulin signaling pathways are implicated in metabolic dysregulation [228], it still remains

unclear whether the observed differences associated with MI in chapter 5 indicate an epigenetic process involved in disease development. However, this possibility has not been disproven either, if anything our exploration warrants a bigger investigation with more individuals and samples of tissues involved in cardiovascular and metabolic diseases.

Sequential Study Design

In this thesis we applied a sequence of research questions, each with a specific study design and population cohort, to investigate the causes of epigenetic variation and its consequence for disease risk at the same loci. We found that epigenetic marks can be persistently altered by conditions of the prenatal environment (Chapter 3), that they may change dynamically throughout life (Chapter 4), and that they are associated with risk for CHD (Chapter 5). The summation of random and directed epigenetic variation with a heritable [243], or prenatally determined background fits both the observed randomness in onset of complex diseases, the influence of environmental components on disease development, and the association between conditions during prenatal and early life with risk for these diseases [257]. Unfortunately, there is no single study design that can readily isolate all components of epigenetic variation, nor will one population cohort likely accommodate all research questions involved in investigating the accumulated causes and consequences of epigenetic variation across a lifetime. However, careful formulation of specific research question and subsequent selection of appropriate study designs, may still promise meaningful epigenetic research on ageing and its related diseases [112], as demonstrated by sequentially carrying results across chapters 3, 4, and 5 in this thesis.

The sequential study design that we applied with chapters 3 to 5 combines a retrospective, a (mainly) cross-sectional, and a prospective study, developing epigenetic markers of a past environmental exposure, establishing their susceptibility to epigenetic drift, and using these markers to investigate disease risk in a clinical trial, respectively. This approach exploits the potential of epigenetic research, covering time periods and events from preconception till death that are beyond single population studies and providing insight into

the environmental and stochastic components of the common age related diseases. However, prospective and retrospective investigation necessitates using epigenetic marks with a stable variation for at least the duration of the study. Investigating longitudinal samples would be best for this [49], but, when not available, another form of evidence on temporal stability can be obtained. Similarly, isolating specific leukocyte cell types will be impractical for large studies, many of which even lack information on the proportions of the major leukocyte sub-populations (e.g. chapters 3 and 5), although in genome wide studies another method can be applied [104]. As, in humans, direct analysis of internal tissues is practically impossible in anything but a study on post mortem biosamples, establishing the correlation of methylation in peripheral tissues with these disease relevant tissues will be crucial for deeper interpretation of uncovered associations. Thus, effective epigenetic research requires information on the stability of epigenetic variation over time and its correspondence across tissues and cell types.

In this thesis we used a preliminary sub-study on longitudinal biosamples from healthy individuals (potential controls) to assess the amount of epigenetic variation at the candidate loci, the influence of leukocyte population heterogeneity and its temporal stability during a period covering the follow-up period of most clinical studies (chapter 2). As an auxiliary we used a (small) genome wide study on autopsy samples (Chapter 6) to interpret the associations of *GNASAS* and *INS* with disease risk in relation to the mechanism of disease development. With the eventual completion of epigenetic data bases, such as those of the International Human Epigenome Consortium [96] (http://www.ihec-epigenomes.org/) and the National Institutes of Health (NIH) Epigenomics Roadmap [97] (http://www.roadmapepigenomics.org/), such steps in the sequential design can be performed by a database search. Thus, from this thesis it follows that the design and interpretation of epigenetic associations studies into ageing related diseases requires data bases that record data from: **1)** studies on genome wide DNA methylation marks in longitudinal samples of peripheral tissues from healthy individuals to assess temporal stability (Chapter 2; [88,89]. **2)** studies on postpartum samples of extra-embryonic tissues from neonates [62] to acquire references for developmental DNA

methylation dynamics in humans from gametogenesis till birth; **3)** studies on post-mortem samples of peripheral and internal tissues to assess the influence of cell type population heterogeneity on DNA methylation marks for each tissue (Chapters 2 and 4), and the correlation of methylation marks across tissues (Chapter 6), and preferably to also provide a record of whole DNA methylomes from each cell type [104]; **4)** studies on mono-cultures of multiple cell types to assess the overlap between DNA methylation and the other layers of epigenetic information (i.e. DNA methylation and other base modifications, histone modifications, association with nuclear lamina, or CTCF hubs); **5**) studies on multiple cell cultures coupling epigenomic with transcriptomic, and perhaps also proteomic, analyses to assess the downstream effects of epigenetic signatures on cellular gene expression; **6**) classical twin studies comparing monozygotic, dizygotic twin pairs and unrelated individuals to assess the influence of heritable genetic sequence variation and environmental (prenatal and late) factors on the epigenetic status at a locus; and **7**) human intervention studies to investigate the epigenetic consequences of medication, therapy and lifestyle changes.

Statistical analyses

Studying the epigenetic basis of common disease not only requires specific study designs, but also particular statistical tools. We implemented linear mixed models to address issues for analyzing DNA methylation data. It uses all available methylation data, accounts for the correlated CpG methylation of a locus, actually using this information to handle data missing at random (Chapters 2, 3, 4, and 5) and achieve accurate imputation for further analyses such as logistic regression (Chapter 5). Besides powerful, the linear mixed model is also versatile and we used several adaptations of it in our various analyses throughout this thesis. The model simultaneously tests whether the observed effect is independent of included covariates (e.g. age, sex, blood parameters, exposures) and adjusts the effect for it. We used this aspect to make all relevant adjustments on the raw data simultaneously (Chapter 2, 3, 4, and 5). The model also calculates the amount of variation that remains unaccounted for by its covariate structure (i.e. residual variance). We used this residual variance in a nested model approach [156]

to assess the methylation variation explained by each of the major leukocyte sub types (Chapters 2 and 4). Random effect covariates allow estimation of changes in variation, which we used to investigate ageing related changes in methylation variation (Chapter 4). Further, by inclusion of multiple subject variables in the model, we assessed the separate effects of shared and unique environmental factors on age related variation in DNA methylation (Chapter 4). The crucial characteristic of a statistical model for analyzing CpG methylation at candidate loci is that it can accommodate the correlation in methylation between adjacent CpG sites (Chapter 2). The linear mixed model can, and is our model of choice for this type of data due to its statistical robustness and its analytical versatility. However, other models that can handle the correlation of adjacent CpG sites exist and can be used.

Future prospects

In this thesis we explored the limits and possibilities of the emerging research field of epigenetic epidemiology, investigating the kind of research questions on the developmental and environmental components of CHD that existing biobanks can be used for. We first described characteristics of variation in DNA methylation at 16 known CHD candidate loci with demonstrated or suspected epigenetic regulation. Then, using a sequence of studies with linked design, we found evidence that epigenetic signatures at some of these loci can persistently change due to environmental (prenatal) conditions, are susceptible to stochastic perturbations that gradually accumulate during life and that epigenetic signatures at two of these loci were associated with risk for CHD. Expanding on these results, epigenetic research promises the potential to reveal the molecular mechanisms that underlie the influence of the developmental period in CHD, the environmental components of CHD and the randomness in CHD development [112,257].

The epigenetic epidemiology of CHD and other complex diseases is still in its infancy, but a recent study reported an association of adult smoking with DNA methylation signatures persisting in blood years after the exposure [71] and in a second study of linked design the same marks were found

associated with risk for CHD [258]. Such results parallel results from the sequential study design in this thesis and present another model of a disease related exposure that is associated with epigenetic marks that are also associated with the outcome. Whether the associated epigenetic signatures mark the exposure or mediate between exposure and outcome remains to be revealed. Interestingly, another study reported DNA methylation associated with prenatal exposure to smoking [67]. In addition there is increasing speculation of transgenerational epigenetic inheritance, with some associations reported between environmental exposures and phenotypic changes across multiple generations [259,260]. However, no convincing empirical evidence has yet been reported on its importance for humans in context of CHD development [112].

Current epigenetic knowledge may help understand the functional relevance of major GWAS hits in non-coding genomic regions [252], but it only scratches at the surface of the epigenetic contribution to CHD development in later life. Assessing the phenotypic relevance of associations between epigenetic marks and CHD in this thesis and other studies [55,254] requires a much deeper understanding of the epigenome, the origin of epigenetic variation, and its causal relation with disease development. A technological hurdle to be cleared for epigenetic epidemiology is the creation of a platform that integrates the possibilities of whole genome bisulphite sequencing [83], with the simultaneous genome wide calling of other modified bases, such as hmC, and various histone modifications [261,262]. Such integrated whole epigenome scans will allow a better assessment of the function of disease associated epigenetic marks in relation to all other epigenetic marks at the locus and throughout the epigenome. A practical hurdle is the necessity to use peripheral tissues as surrogates for inaccessible tissues that are relevant for disease development. Current research has started to map the possibilities of epigenetic marks in accessible tissues for use as surrogate markers [255,256], but the potentially interesting epigenetic signatures of some tissues will likely remain out of reach. Methodological hurdles to be cleared are the establishment of an atlas of soma-wide epigenetic variation as reference [96,97] development of study designs that enable investigating the mechanisms behind the establishment and maintenance of epigenetic marks and the environmental factors that affect

these processes from preconception throughout life [15,251,263]; and the creation of an integrated functional –omics paradigm that combines genomic, epigenomic, and transcriptomic data [47,112]. Eventually such knowledge will aid understanding the downstream effects of disease associated epigenetic marks on gene expression of the locus in relation to cellular expression pathways, and the functionality of the tissue(s) involved in the process of disease development [94,263]