Chapter 1

General Introduction

1.1 Introduction

One of the major challenges in genetics today is to understand the causes of complex diseases. Complex diseases refer to disorders that do not follow Mendel's laws of inheritance (Mendel, 1950; Wang et al., 2005). Such diseases are considered to derive from multiple heritable and environmental factors. Cancer, schizophrenia, diabetes, lupus and cardiovascular diseases are a few representative examples. Identifying the basis of complex diseases will be of great medical relevance (Kiberstis, 2002; Kiberstis, 2002; L, 2002; Lander and Schork, 1994).

In recent years, a lot of attention has been given to the genetic components of such diseases. Large, collaborative studies have been focused on the identification of single nucleotide polymorphisms (SNPs) (The International HapMap Project, 2003; WTCCC, 2007) and copy number variation (CNVs) (Beckmann et al., 2007; Redon et al., 2006) that could be associated with human diseases and eventually lead to new therapies. However, to date only a few genetic variants have a significant and replicated association to complex diseases indicating that other factors in addition to genetic variation may contribute to complex phenotypes.

Recent advances in high-throughput epigenomics (mapping of genome-wide epigenetic modifications) (Bernstein et al., 2007), have led to the concept of epigenetic variation which is now also considered to play an important role in the development of complex diseases (Hatchwell and Greally, 2007; van Vliet et al., 2007). A number of studies, including this thesis aim to elucidate the role of epigenetics in the context of such phenotypes.

I used the major histocompatibility complex (MHC) region (Horton et al., 2004), a 4Mb region on human chromosome 6, as a model system to elucidate the role of DNA methylation (the best-studied epigenetic mark to date) in the regulation of MHC loci. The MHC has been chosen because it is associated with susceptibility to more complex

diseases than any other region within the human genome (Lechler, 2000). Unravelling the epigenetic code of the MHC will give further insights into the impact of epigenetics in complex phenotypes.

1.2. The Major Histocompatibility Complex - MHC

The major histocompatibility complex (MHC) is a 4Mb region on the short arm of human chromosome 6 (6p21.3) (Horton et al., 2004). It is one of the most gene-dense and highly polymorphic regions of the human genome and it is associated with many complex diseases including infectious, autoimmune and inflammatory diseases as well as cancer, and it is important in transplant medicine.

The classical MHC is divided into three classes: Class I, Class II and Class III. Figure 1.1 shows the gene map of the MHC region indicating the order of the three classes (I, III, II) and their relative sizes as well as the genes encoded within each class. The concept of the extended MHC (xMHC) (although not discussed any further within this thesis) was recently established based on the finding that linkage disequilibrium (LD) and MHC-related genes exist outside the boundaries of the classical 4Mb MHC region (Horton et al., 2004).

1.2.1 MHC encoded genes

The gene map of the MHC region was completed and reviewed recently (Horton et al., 2004). The classical MHC comprises 224 gene loci, of which more than 57% are thought to be expressed. At least 10% of the MHC genes have functions related to the immune system. The MHC is divided into three regions in the following order: telomere – class I, class III, class II - centromere (figure 1.1):

i. MHC class I region

This region has a size of about 2Mb and contains the three main MHC class I genes, *HLA-A*, *HLA-B* and *HLA-C*, which are highly polymorphic and members of the immunoglobin superfamily (Lawlor et al., 1990). They present intracellular antigen peptides to the T-cell receptors of cytotoxic T-cells. Antigens, in order to be presented to the cell surface, go through a pathway called the antigen presenting pathway (Hewitt, 2003). These genes are expressed by most somatic tissues at varying levels. The MHC class I region also harbours the non-classical class I genes, *HLA-E*, *-F* and *–G*, which are less polymorphic and display a more restricted tissue expression compared to the classical genes (Geraghty, 1993). *HLA-G* is the only class I gene expressed in foetal trophoblast cells and may play a role in the maternal tolerance of the foetus (Loke and King, 1991; Parham, 1996). The MHC class I like genes *MICA* and *MICB* (Stephens, 2001) and a plethora of pseudogenes (Geraghty, 1993; Le Bouteiller, 1994) are also encoded within this region.

ii. MHC class II region

The class II region, 880 kb in size, contains one gene every 40kb on average. This class, similar to class I, encodes members of the immunoglobin superfamily *HLA-DP*, *-DQ*, *–DR* and pseudogenes (the classical MHC class II genes) (Andersson et al., 1987). However, these genes are expressed as heterodimers only on specialised antigen-presenting cells such as macrophages, B cells and some T cells. The former engulf and internalise exogenous antigens which are then presented to helper T cells by class II members of the immunoglobin superfamily. The non-classical class II genes, *HLA-DM* and *–DO*, are not expressed on the cell surface, but form heterotetrameric complexes involved in peptide exchange and loading onto classical class II molecules (Alfonso and Karlsson, 2000).

The class II region also encodes *PSMB8*, *PSMB9*, *TAP1*, *TAP2* and *TAPBP*. The products of these genes are involved in the MHC class I antigen processing and presentation machinery (Androlewicz, 1999; Lehner and Trowsdale, 1998; van Endert, 1999).

iii. MHC class III region

This is a very gene dense region of about 100 kb and contains at least 70 genes with diverse functions (Aguado, 1996). Examples of class III genes are *C2* and *C4* which are members of the complement system. *C2* and *C4* gene products mediate phagocytosis and lysis of bacterially infected cells leading to inflammatory response. Members of the tumour necrosis family, *TNF-*α, *LTA* and *LTB*, which are cytokines that control inflammation, are also encoded within this region (Gruss and Dower, 1995). Three heat shock proteins (HSP) are also encoded in the class III region. They are involved in stress-induced signalling of immune responses mediating the elimination of damaged, infected or malignant cells (Gleimer and Parham, 2003)

1.2.2 MHC Polymorphism

HLA class I and class II genes are highly polymorphic (Robinson et al., 2003) and *HLA-B* has been reported to be the most polymorphic gene in the human genome (Mungall et al., 2003). The extensive polymorphism of the MHC loci is believed to enhance immune defence by broadening the array of antigenic peptides available for T-cell recognition. MHC encoded molecules govern immune responses by presenting antigen peptides to T-cells (figure 5.2). Genetic polymorphism within the MHC region also facilitates variable susceptibility to MHC-linked diseases as well as to pathogens (Goulder and Watkins, 2008). Single nucleotide polymorphisms are the most common type of variation within the MHC. Recently, four-independent re-sequencing projects have significantly expanded our knowledge of variation within the MHC (Horton et al., 2008; Raymond et

al., 2005; Shiina et al., 2006; Smith et al., 2006). However, polymorphism is not restricted to genetic variation (SNPs). Structural copy number variation (CNVs) (Redon et al., 2006) also exist within the MHC (Stewart et al., 2004). More specifically, two hyper-variable regions have been reported to have CNVs: (i). the RCCX region (within the MHC class III region) which contains duplications of *C4, TNXA, CYP21A1P* and *STK19P* pseudogenes (Chung et al., 2002; Yang et al., 1999), and (ii). the *DRB* locus (within the MHC class II region) between *HLA-DRB1* and *HLA-DRB9* which shows haplotype-specific rearrangements (Marsh, 2000) .

In addition, polymorphism resulting from the presence or absence of retroviral sequences (LINEs, Alu, HERV, LTR, MER and SVA) (Stewart et al., 2004) has been observed. The retroviral insertions are often located either in the *HLA-DR* region or near the MHC class I genes, possibly promoting molecular evolution. Recently a HERV-derived gene, *HCP5*, has been implicated in HIV-1-host interaction (Fellay et al., 2007).

1.2.3 MHC-linked diseases

The MHC is associated with many diseases including most if not all autoimmune diseases (Lechler, 2000). A representative list of MHC-linked diseases is given in table 1.1. In most of the cases listed, the disease causing mutation/variation is not yet known. Disease-causing and disease-associated genes are indicated in table 1.1.

The involvement of the MHC in many complex diseases was confirmed further by whole genome association (WGA) studies (WTCCC, 2007). A recent study has shown that MHC-class I mediated events, mainly involving a *HLA-B* locus, contribute to the aetiology of type I diabetes, which is an autoimmune disease (Nejentsev et al., 2007), whereas the role of three other polymorphic MHC loci (*HLA-B, HLA-C* and *ZNRD1*) were reported to influence the host response to HIV-1 (Fellay et al., 2007). The latter supports the role of the MHC in conferring resistance to infectious diseases.

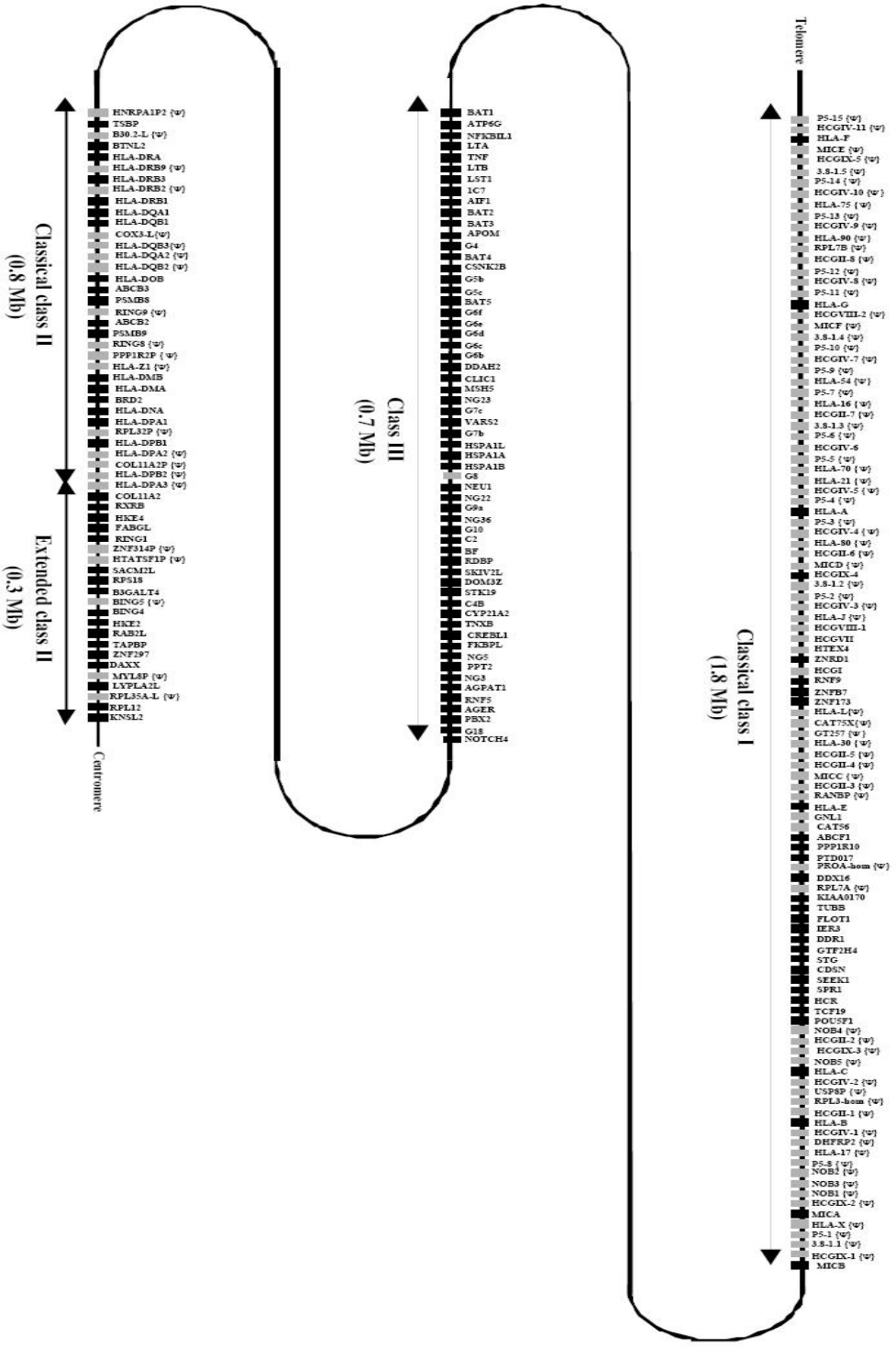

Figure 1.1. **Gene map of the human MHC**. Genes are displayed in order from telomere to centromere but are not drawn to scale. Solid black boxes indicate the loci that have been

investigated by the HEP (see below). {Ψ} indicates the presence of a pseudogene. Figure was taken from Novik et al. (Novik et al., 2002).

1.2.3.1 Future challenges in studying MHC-linked diseases

The MHC provides a prototype for the study of complex diseases. Although past studies have generated extensive data for the genetics of the MHC resulting in important contributions to medicine (de Bakker et al., 2006; Rioux and Abbas, 2005; Vyse and Todd, 1996), further studies are necessary to improve our understanding of the causes of MHC-linked diseases. As summarised in figure 1.2, complex MHC-linked diseases, like autoimmune diseases, are the result of complex interactions between genetic, epigenetic and environmental factors. Epigenetic factors include DNA methylation, histone modifications and non-coding RNAs (see below).

Elucidating the epigenetic code of the MHC can be expected to be highly beneficial to biomedical research.

1.2.4 MHC and Epigenetics – What is known so far

Emerging evidence suggests that epigenetic events are associated with the regulation of MHC gene expression. This is based on the below findings:

i. The MHC class II transactivator (CIITA) and the regulatory factor X (RFX) proteins serve as focal points for recruiting histone modifying enzymes to MHC class II promoters, whereby CIITA itself is regulated by DNA methylation, histone modifications and ncRNAs (Wright and Ting, 2006; Zika and Ting, 2005).

ii. Treatment of melanoma and esophageal cell lines with the DNA methylation inhibitor 5-aza-2'-deoxycytidine (see below) led to restoration of MHC class I expression (which is suppressed in these cell lines), implicating DNA

methylation in the expression of MHC class I genes (Maio et al., 2003; Nie et al., 2001; Serrano et al., 2001)

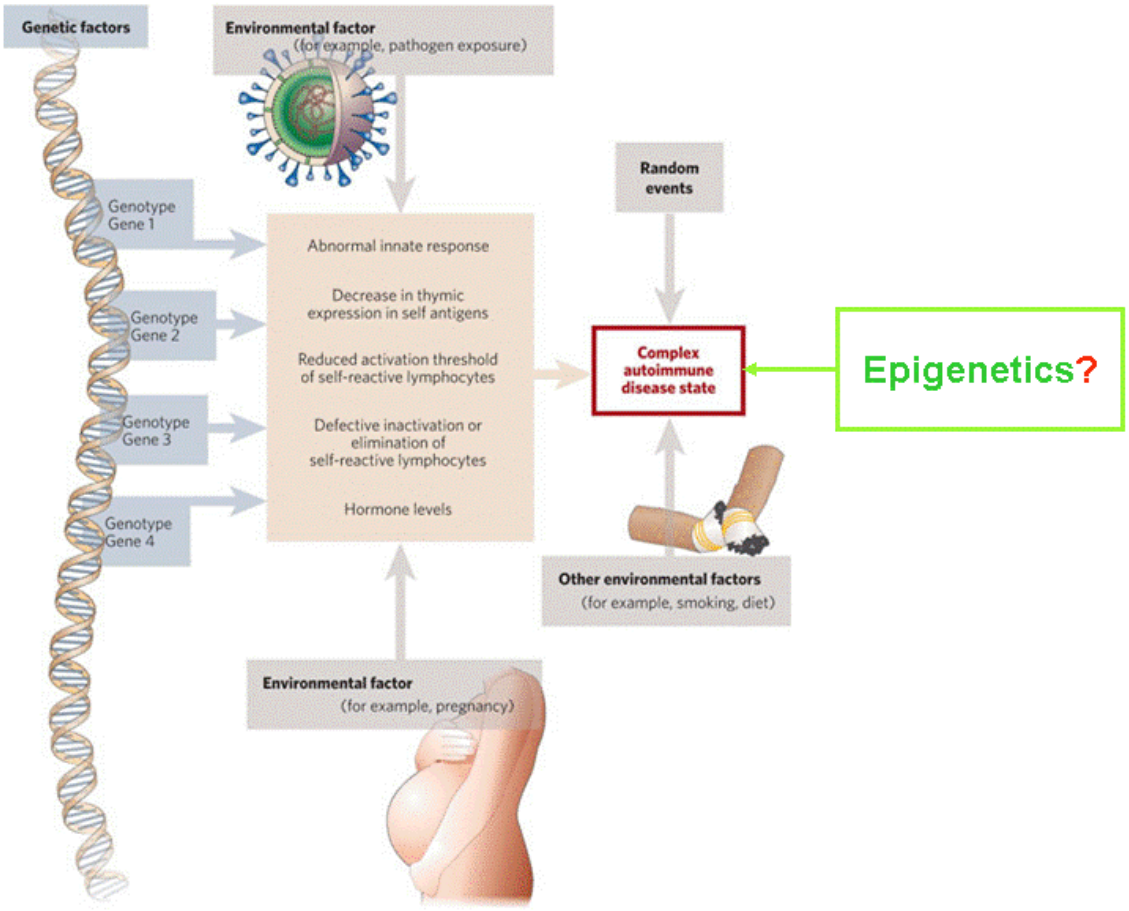

Figure 1.2. **Autoimmune diseases caused by complex traits.** In complex traits, the clinically recognised disease state results from interactions between multiple genotypes and the environment. Recently the role of epigenetics in such complex diseases has been implicated. The question mark next to 'epigenetics' reflects that the contribution of epigenetics is still not well understood. A lot of current studies aim to elucidate the impact of genetics, epigenetics and environmental factors in susceptibility, disease progression and clinical management. Figure was taken (with some modifications) from Rioux and Abbas (Rioux and Abbas, 2005).

MHC class I region	
HLA-G	Associated with *Pemphigus vulgaris* in Jewish patients
HLA-A	Associated with autoimmune diseases; for example, birdshot chorioretinopathy
HLA-E	Associated with type 1 *Diabetes mellitus*; also influences age of onset of disease
MDC1	Associated with inadequate DNA damage responses owing to MDC1-deficiency
CDSN	Causes hypotrichosis simplex of the scalp
PSORS1C1	Associated with psoriasis
PSORS1C2	Associated with psoriasis
C6orf18	Associated with psoriasis
HLA-C	Associated with autoimmune diseases; for example, psoriasis
HLA-B	Associated with autoimmune diseases; for example, ankylosing spondylitis or Behcet disease
MICA	Associated with autoimmune diseases; for example, rheumatoid arthritis and coeliac disease
MICB	Associated with coeliac disease
MHC class III region	
NFKBIL1	Associated with rheumatoid arthritis
LTA	Associated with myocardial infarction
TNF	Associated with septic shock, cerebral malaria
LTB	Associated with infective/inflammatory diseases
NCR3	Associated with impairment of NK cell function in HIV-1 infected patients
BAT2	Associated with influence on age at onset of type 1 *Diabetes mellitus*
NEU1	Causes type I and II sialidosis
C2	Causes C2 deficiency
C4B	Causes C4 deficiency
C4A	Causes C4 deficiency
CYP21A2	Causes several disorders owing to 21-hydroxylase deficiency
TNXB	Causes Ehlers–Danlos syndrome (hypermobility type) owing to tenascin X deficiency
AGER	Associated with amplification of inflammatory responses in rheumatoid arthritis
MHC class II region	
HLA-DR loci	Associated with autoimmune diseases; for example, rheumatoid arthritis, type 1 and type 2 *Diabetes mellitus*
HLA-DQ loci	Associated with autoimmune diseases; for example, narcolepsy
TAP2	Causes bare lymphocyte syndrome type I owing to TAP2-deficiency; associated with various diseases; for example, rheumatoid arthritis
TAP1	Causes bare lymphocyte syndrome type I owing to TAP1-deficiency; associated with various diseases; for example, vitiligo in Caucasian patients that are young in age at onset
BRD2	Associated with juvenile myoclonic epilepsy
HLA-DP loci	Associated with autoimmune diseases; for example, chronic beryllium disease

Table 1.1. **Genes in the MHC in which variation has a relationship to disease.** Table was taken from Horton et al. (Horton et al., 2004)

iii. The HEP study has shown that at least 10% of the MHC loci analysed show tissue-specific methylation patterns, implicating DNA methylation in tissue-specific expression of MHC genes (Rakyan et al., 2004).

iv. A non-coding RNA (microRNA), encoded by human cytomegalovirus (HCMV) during infection, might regulate the expression of a 'stressed-induced' MHC gene (*MICB*) (Stern-Ginossar et al., 2007).

v. It should be noted that extensive the genetic polymorphism within the MHC makes the latter an ideal region for studying interaction between the genome and the epigenome and the formation of hepitypes (see section 1.3.7.1). It can be postulated that SNPs that result to gain or loss of one or more critical CpG sites may affect the overall methylation profile of a locus. Alternatively, non-CpG SNPs located within an epigenetically sensitive regulatory element may also influence the epigenetic make-up of a region.

Based on the above observations I reasoned that epigenetics may be important in the regulation of genes encoded within in the MHC, and hence be associated with MHC-linked phenotypes.

In the following sections I introduce the concept of epigenetics, refer in detail to DNA methylation and epigenetic variation in the form of differentially methylated regions (DMRs), and discuss how DMRs can be linked to complex phenotypes. The rationale of my study, which aimed to identify DMRs within the MHC, is given in the last section of this chapter.

1.3 Epigenetics

1.3.1 Definition

The term epigenetics was first introduced by Conrad Waddington in 1942 (Waddington, 1942). It was used to describe the interactions of genes with their environment “to bring a phenotype into being”. Today epigenetics refers to mitotically and, in some cases, meiotically heritable states of gene expression that are not due to changes in the DNA sequence (Allis et al., 2007). The Greek prefix ‘epi-‘ implies features that are “in addition” to genetics, and this is reflected in the current definition.

1.3.2 Epigenetic Modifications in Mammalian Genomes

Epigenetic modifications are stable modifications of the DNA or chromatin that do not alter the primary nucleotide sequence. They can alter the functions of associated genes by modulating DNA accessibility, protein recruitment and chromatin structure (figure 1.3a).

Histones are the major protein component of chromatin. The core histones, including H2A, H2B, H3 and H4, make up the nucleosome and are subjected to post-translational modifications at specific positions within the amino-terminus of their tails. These modifications include for instance acetylation, methylation, phosphorylation and ubiquitination (figure 1.3b) and they are correlated with chromatin accessibility and transcriptional activity or repression (Berger, 2007; Kouzarides, 2007).

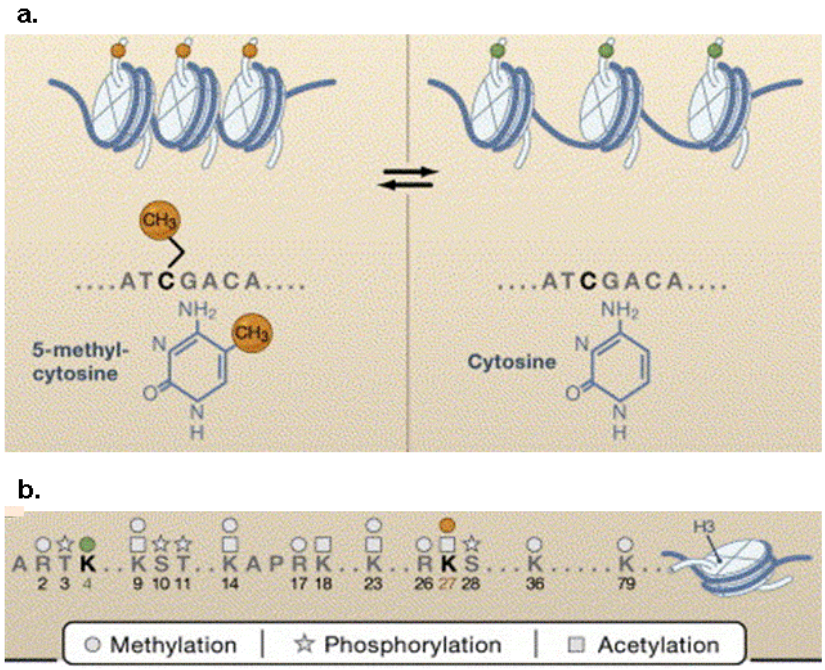

Figure 1.3. **DNA methylation and histone modifications**. Cytosine methylation is the only known covalent modification of DNA in mammals. In contrast, histones are subject to different combinations of modifications, including acetylation, methylation, phosphorylation and ubiquitination. Part a. illustrates the structure and effects of cytosine methylation (repressive/orange, activating/green). Part b. illustrates the diversity of histone H3 modifications. Figure was taken from Bernstein el al., (Bernstein et al., 2007).

DNA methylation is a covalent modification of the 5-carbon position of cytosine. The reaction involves the addition of a methyl group (figure 1.3a) and it is catalysed by DNA methyltransferases (DNMTs) with S-adenosyl-methionine (SAM) as the methyl donor (figure 1.4). In mammals this occurs predominantly in the context of cytidine-guanosine (CpG) dinucleotides (Bird, 2002) but non-CpG methylation has also been reported in certain cell types, and is common in plants (Finnegan and Kovac, 2000; Grandjean et al., 2007; Ramsahoye et al., 2000).

Recently, non-coding RNAs (ncRNAs) have been recognised as an additional component associated with epigenetic modulation and have been reported to be

involved in X-chromosome inactivation, chromatin structure, DNA imprinting and DNA demethylation (Costa, 2005).

DNA methylation is the epigenetic mark studied in this thesis and is discussed in more detail in the following sections.

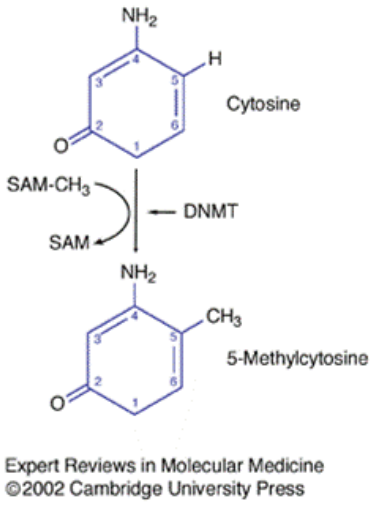

Figure 1.4. **Mechanism of DNA methylation.** 5-Methylcytosine is produced by the action of the DNA methyltransferases (DNMT 1, 3a or 3b), which catalyse the transfer of a methyl group (CH_3) from *S*-adenosylmethionine (SAM) to the 5-carbon position of cytosine.

1.3.3 DNA methylation in mammals

DNA methylation is the most stable epigenetic modification known to date. Cytosine methylation patterns are propagated through cell division and this involves the action of specific DNA methyltransferases (DNMT1) (Bird, 2002; Goll and Bestor, 2005). Methylation patterns are established during early mammalian development starting with the paternal genome undergoing active demethylation shortly after protamine-histone exchange in the male pro-nucleus. The maternal genome also undergoes demethylation probably through a passive DNA replication mechanism (Reik et al., 2001; Santos et al., 2002). Genome-wide methylation levels increase rapidly in the blastocyst by the action of the *de novo* DNA methyltransferases DNMT3A and DNMT3B (Bestor, 2000) which

ultimately lead in the formation of methylation patterns found in adult somatic cells. These physiological patterns of cytosine methylation can be disrupted to cause disease, the best-studied example being cancer, in which abnormal methylation is common and implicated in pathologic events such as silencing of tumour-suppressor genes (Robertson, 2005). In addition to DNMT1, DNMT3A and DNMT3B a fourth DNA methyltransferase, DNMT2, is known to date. DNMT2 has low methyltransferase activity *in vitro* and its absence has no discernable effect on DNA methylation levels (Bestor, 2000).

In mammalian somatic cells, DNA methylation occurs at the 5-carbon position of cytosine at CpG dinucleotide (5m-CpG) sites (figures 1.3a and 1.4). About 70% of CpGs are methylated (hypermethylated), amounting to about 1% of total DNA bases in the human genome (Ehrlich et al., 1982). In normal somatic cells, 5m-CpG sites predominantly occur in repetitive DNA elements, satellite DNAs, non-repetitive intergenic DNA and exons. Regions with high G+C and CpG content, termed CpG islands (Cross and Bird, 1995; Klose and Bird, 2006), have been considered to be mostly unmethylated. CpG islands cover about 0.7% of the human genome and contain 7% of the CpG sites. The unmethylated status of CpG islands, at least in the germ line, protects them from CpG depletion caused by spontaneous deamination of methylated cytosines. The mismatch repair system can accurately recognize and correct the deamination product of cytosine bases (uracil), but not the deamination product of methyl-cytosine (thymine). About 60% of human gene promoters are associated with CpG islands whereas there are about 8,500 autosomal non-promoter CpG islands within the human genome (Hubbard et al., 2007).

Recent studies indicate that a subset of promoter CpG islands are subjected to *de novo* methylation during normal development and tumourigenesis (Meissner et al., 2008), and this has been reported to be associated with repression of CpG island-promoters

(Eckhardt et al., 2006; Estecio et al., 2007; Jones and Baylin, 2002; Khulan et al., 2006; Rakyan, 2008; Weber et al., 2007). Concerning non-promoter CpG islands only 27% were found to be unmethylated compared to 67% of promoter CpG islands that are constitutively unmethylated (Rakyan et al., 2008). Correlation with expression data suggested that approximately half of the currently annotated non-promoter CpG islands are likely to have similar functions as promoter CpG islands (see below).

1.3.4 Function of DNA methylation

In mammals, DNA methylation plays a vital role in a diverse range of cellular functions, including tissue-specific gene expression, imprinting, X-chromosome inactivation, cell differentiation and the regulation of chromatin structure (Bird, 2002). It is also associated with many diseases including cancer and ageing (Robertson, 2005).

Mechanistically, a methylated cytosine (frequently referred as the fifth base) can be recognised by a number of regulatory proteins and alter the transcriptional potential of genomic regions. DNA methylation has been mainly implicated with transcriptional repression especially when it occurs within promoter regions or in close proximity to the transcription start sites (TSS) of genes (Bird, 2002). However, recent evidence suggest that it also facilitates transcription when it occurs downstream of promoter regions within gene bodies. Recent studies have shown a positive correlation between gene-body DNA methylation and gene expression (Eckhardt et al., 2006; Rakyan, 2008; Rakyan et al., 2004). This is consistent with data generated for the X-chromosome where hypomethylation at gene promoters and hypermethylation of gene bodies was associated with active transcription (Hellman and Chess, 2007). DNA methylation is also involved in the global maintenance of the genome, protection against mobile elements and inhibition of cryptic transcription. These will be discussed below.

i. *DNA methylation and gene expression silencing*

There are two basic models that underpin the relationship between DNA methylation of promoter regions and gene silencing (figure 1.5):

1. Direct model – DNA methylation represses transcription by directly blocking the recruitment of transcriptional activators to the cognate DNA sequence (Watt and Molloy, 1988).

2. Indirect model – additional factors, like the methyl binding domain (MBD) containing proteins, including MeCP2, MBD1 and MBD2, that bind to methylated DNA are required for the recruitment of transcription repression complexes (figure 1.5) (Ballestar et al., 2003; Ballestar and Wolffe, 2001).

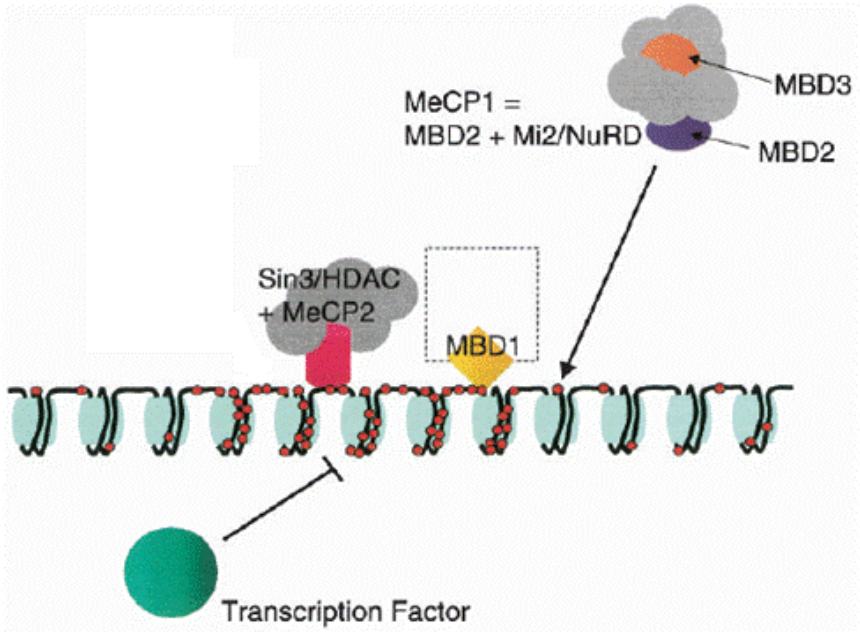

Figure 1.5. **Transcriptional repression by DNA methylation.** A stretch of nucleosomal DNA is shown with all CpGs methylated (red circles). Below the diagram is a transcription factor that is unable to bind its recognition site when it is methylated. The top of the diagram illustrates protein complexes that can be attracted by methylation, including the methyl-CpG binding protein MeCP2 (plus the Sin3A histone deacetylase complex), the MeCP1 complex comprising MBD2 plus the NuRD corepressor complex, and the uncharacterized MBD1 complex. MeCP2 and MBD1 are chromosome bound proteins, whereas MeCP1 may be less tightly bound (reviewed in (Bird, 2002). Figure was taken from Bird (Bird, 2002).

It is also worthy of mention that there are several lines of evidence supporting the notion that DNA methylation does not mediate silencing of active promoters but rather affects

genes with low transcriptional activity, suggesting that DNA methylation is a secondary event during the gene silencing process (Bird, 2002; Clark and Melki, 2002; Stirzaker et al., 2004; Turker, 2002). Examples of *de novo* methylation by DNMTs following gene inactivation include methylation of genes that are already silenced during X-inactivation, and hypermethylation of *GSTP1* CpG island promoter which is initiated by a combination of gene silencing and spreading of DNA methylation. This is consistent with a recent study looking for changes in methylation patterns upon differentiation proposing a "Use it or Lose it" model. According to this model, genes with low levels of transcriptional activity in a given cell type are likely to be locked in this state by DNA methylation (Meissner et al., 2008). This can be the result of changes in the balance of chromatin modifying enzymes (e.g. decreased H3K4 de-methylase activity) due to transcriptional silencing. It has been shown that DNA methylation is in inverse correlation with methylation of H3K4 (Meissner et al., 2008).

ii. *Inhibition of cryptic transcription initiation*

DNA methylation within coding regions has been confirmed by many recent studies (Eckhardt et al., 2006; Meissner et al., 2008; Rakyan et al., 2008; Rakyan et al., 2004; Weber et al., 2007). One potential role for intragenic methylation could be inhibition of cryptic transcription initiation outside gene promoters (Zilberman et al., 2007). It is possible that the transcription machinery itself disrupts chromatin structure and exposes cryptic initiation sites to be methylated (Carrozza et al., 2005).

iii. *Protection against mobile elements*

In mammalian genomes, most repetitive DNA sequences are found to be methylated (Rollins et al., 2006). Work in Dnmt1$^{-/-}$ mice supports the notion that methylation leads to silencing of repeats (Walsh et al., 1998), and hence to immobilization of mobile elements which is important to insure genomic integrity.

iv. Maintenance of genome stability

Several lines of evidence indicate that global hypomethylation can lead to increased genomic instability in mammalian cells. In human cell lines, deletion of DNMTs induces chromosomal abnormalities (Chen et al., 2007) and partial loss of DNMT3b is linked to the immunodeficiency (ICF) syndrome, which is characterised by chromosomal rearrangements in centromeric regions (Xu et al., 1999). Global hypomethylation in the gene bodies and repetitive elements is a well known characteristic of human cancer cells. This phenomenon has been linked to increased chromosomal instability and tumour progression (Eden et al., 2003), and it was also shown to precede copy number changes in gastrointestinal cancer (Suzuki et al., 2006).

1.3.5 DNA methylation inhibition assay

The correlation between loss and gain of DNA methylation, and activation and repression of transcription of the associated genes can be verified by the use of DNA methyltransferase inhibitors. One class of methylation inhibitors are nucleoside analogues which have a modified cytosine ring attached to either a ribose or deoxyribose moiety (Jones and Taylor, 1980) (figure 1.6a). These analogues can be metabolised into nucleotides, and hence incorporated into DNA and/or RNA (Li et al., 1970). Treatment of cultured cells with such analogues can lead to loss of DNMT activity, as the latter becomes irreversibly bound to the analogues, resulting in passive loss of DNA methylation (figure 1.6b). Two cytosine analogues, 5-Azacytidine (5-Aza-CR) and 5-Aza-2'-deoxycytidine (5-Aza-CdR) are commonly used in cell culture DNA methylation studies and they have been widely studied for cancer treatment (Christman, 2002; Yoo and Jones, 2006). Both 5-Aza-CR and 5-Aza-CdR were recently approved by the U.S. Food and Drug Administrator (FDA) (with the clinical names Vidaza and Decitabine respectively) for treatment of myelodysplastic syndrome, a preleukemic disease (Gal-Yam et al., 2008; Kaminskas et al., 2005; Kantarjian et al., 2007).

Cells cultured in the presence of 5-Aza-CdR incorporate it into DNA during DNA replication which leads into the formation of stable covalent complexes between the DNA molecule and DNMTs (Santi et al., 1983). The modification at the C5 position prevents the release of the enzyme (figure 1.6.c). This prevents further methylation of the genome and results to progeny cells with reduced DNA methylation.

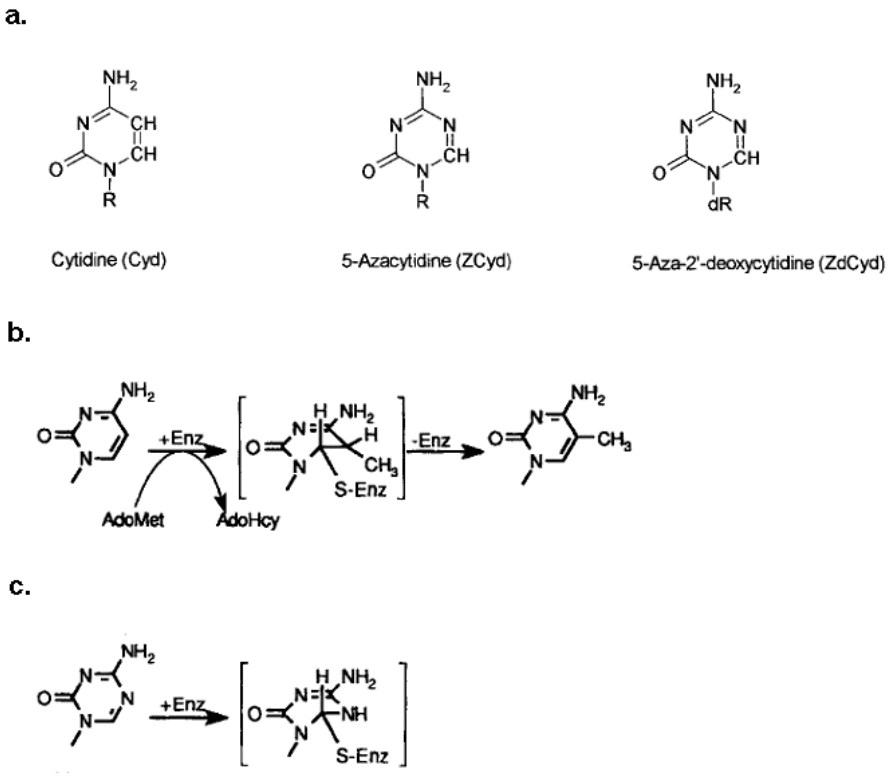

Figure 1.6. **DNA methylation inhibitors.** Cytidine analogues have a modified cytosine ring, preventing methylation. Part a. Structure of cytidine and its 5-aza analogues. R=ribose, dR=deoxyribose. Part b. Methylation of cytidine at the 5-carbon position. Part c. DNMTs bind irreversibly to cytidine analogues and block DNA methylation. Figure was taken from Christman (Christman, 2002).

1.3.6 Methodologies for detection of DNA methylation

Many methods of DNA methylation analysis have been developed over the years. DNA methylation detection approaches are based on one of three techniques: bisulphite conversion, digestion with methylation-sensitive restriction enzymes and affinity purification (reviewed in (Beck and Rakyan, 2008; Weber and Schubeler, 2007;

Zilberman and Henikoff, 2007). Bisulphite sequencing and the immunoprecipitation approach for capturing methylated DNA are introduced below and used extensively for the work described within this thesis.

1.3.6.1 Bisulphite Sequencing

The gold standard approach for methylation analysis is 'bisulphite sequencing'. Bisulphite sequencing involves treating DNA with sodium bisulphite to convert unmethylated cytosines to uracils (figure 1.7). Converted DNA is subjected to PCR amplification using primer sets corresponding to the regions of interest. Primers are specific for converted DNA and do not contain CpG sites. The last step involves conventional DNA sequencing; unmethylated cytosines will be read as thymine, while methylated cytosines will be read as cytosine (Frommer et al., 1992) (figure 1.7).

Typically in tissue samples, because they contain a mixture of different cells, methylation levels for a specific CpG site appear to be heterogeneous. Hence, it is necessary to quantify the proportion of methylated CpG sites under investigation after bisulphite sequencing. An algorithm called ESME was developed as part of the Human Epigenome Project (HEP) (see below) and was applied to estimate methylation levels from signal ratios of the corresponding sequence traces (Lewin et al., 2004). It has been demonstrated that this method can detect differences in methlyation rates of 20% highly accurately.

Recently bisulphite conversion has been adapted for large scale DNA methylation analysis. Meissner and colleagues have developed a bisulphite conversion-based method called reduced representation bisulphite sequencing (RRBS) (Meissner et al., 2005) which was successfully combined with next generation sequencing technology (Meissner et al., 2008). In a similar manner bisulphite converted DNA was subjected to deep sequencing using an Illumina Genetic Analyser (GA) for the analysis of the *A.*

thaliana methylome (methylC-seq) (Lister et al., 2008). Bisulphite conversion has also been combined with microarray platforms for large-scale methylation analysis (Adorjan et al., 2002; Gitan et al., 2002; Reinders et al., 2008).

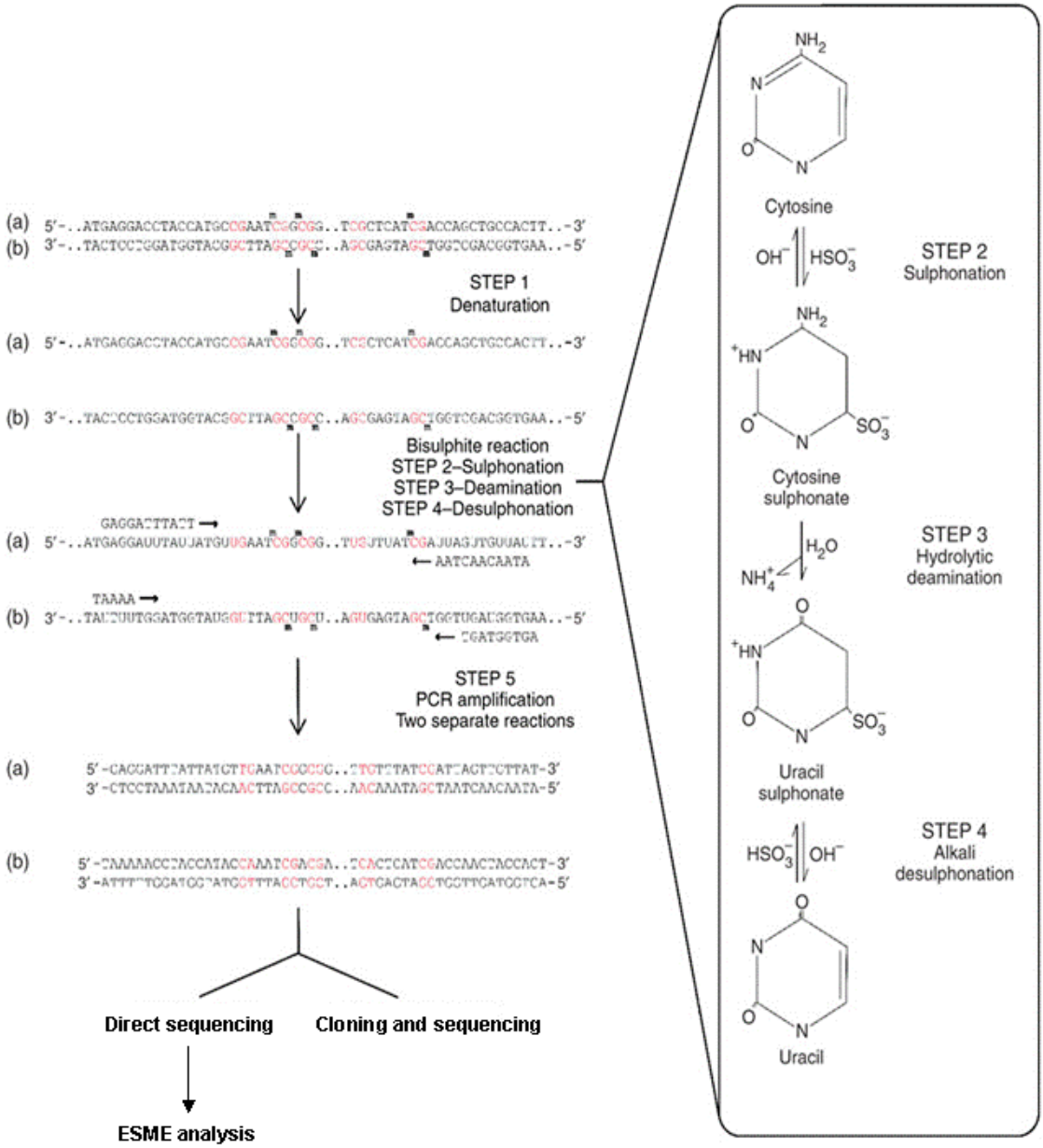

Figure 1.7. **Bisulphite conversion**. An example DNA sequence, 5' to 3' orientation, with the complementary plus (a) and minus (b) DNA strands is shown. The CpG sites are colored red and methylation of a CpG site is indicated by ^{m}CpG. After denaturation, the DNA is single stranded and each strand, a and b, can be amplified independently with strand-specific bisulphite-specific primers to determine

the methylation state of each strand. Example strand-specific and bisulphite-specific PCR primers are indicated above and below the DNA strands (in reality, primers are longer). In the forward primers, the cytosine bases are replaced by thymine bases and, in the reverse primers, the guanines (complementary base to cytosine) are replaced by adenine residues. Detailed design parameters of the bisulphite-specific PCR primers are given in section 2.2.3.1. After PCR amplification, methylation of the CpG sties in the target sequence can be determined by either direct PCR sequencing of the product or cloning and sequencing. Figure was taken (with some modifications) from Clark et al. (Clark et al., 2006).

1.3.6.2 Methylated DNA Immunoprecipitaton – MeDIP

Methylated DNA Immunoprecipitation (MeDIP) has been developed recently (Keshet et al., 2006; Weber et al., 2005) and since then has been used extensively especially for large-scale methylation studies (see below). In the MeDIP assay, a monoclonal antibody against methylated cytosines is used to enrich methylated DNA fragments. In brief, genomic DNA is fragmented to an average size between 300 and 600 bp and denatured to generate single-stranded DNA fragments. Methylated single-stranded DNA fragments are immunoprecipitated after incubation with an antibody that has affinity for the methyl group of methylated cytosines (figure 1.8).

The immunoprecipitated DNA can be used for the analysis of the methylation status of a particular genomic region by employing specific primers. However, the importance of this technique is underlined by the fact that MeDIP can be easily adapted for large-scale or even genome-wide methylation analysis studies (figure 1.8). MeDIP has already been combined with microarray technologies to generate methylation profiles in cancer and normal tissue samples (Illingworth et al., 2008; Keshet et al., 2006; Mohn et al., 2008; Rakyan et al., 2008; Weber et al., 2005; Weber et al., 2007; Zhang et al., 2008; Zhang et al., 2006; Zilberman et al., 2007). Recently MeDIP was combined with next-generation sequencing technology leading to the first mammalian methylome (Down et al., 2008). Human sperm DNA was used for this purpose. Such advances in methylation profiling are promising great potential for future studies of DNA methylation in humans.

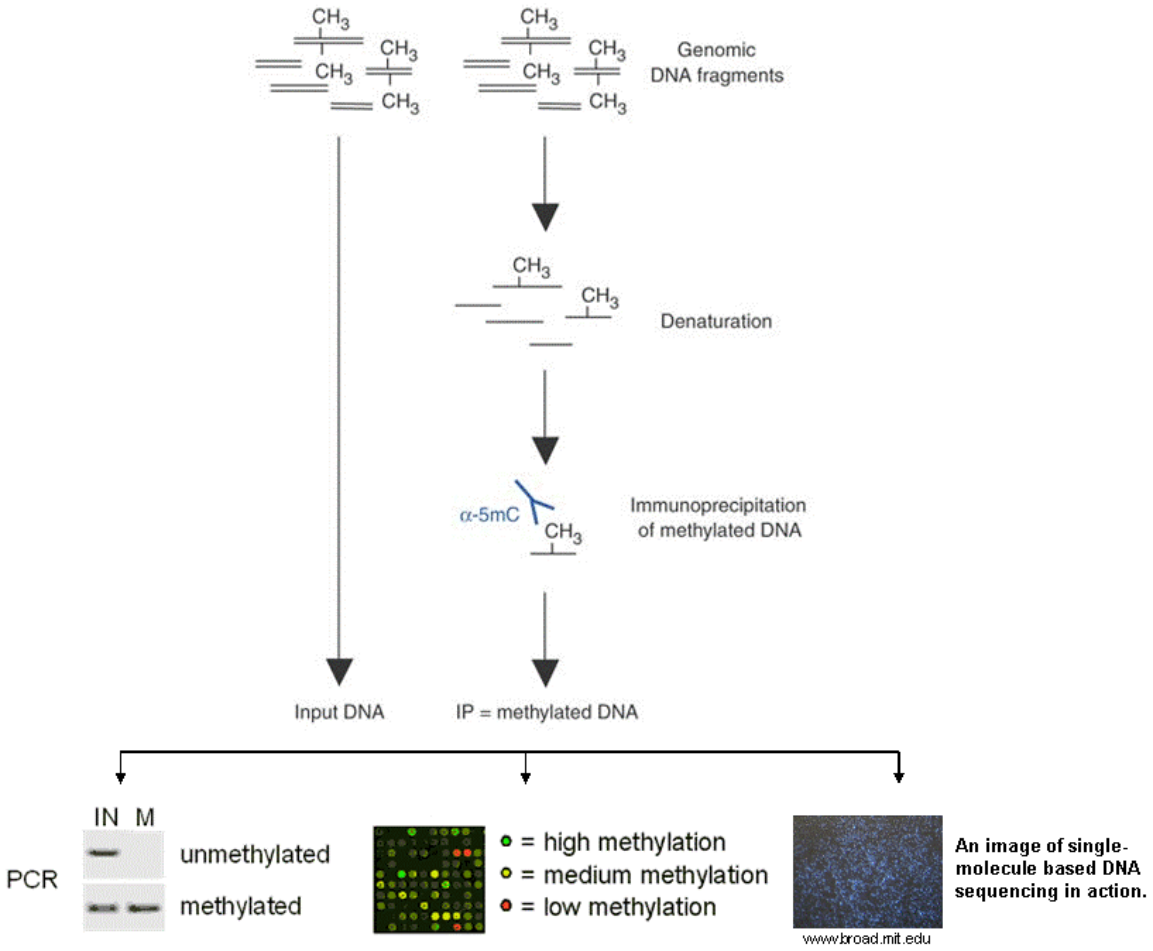

Figure 1.8. **Methylated DNA immunoprecipitation (MeDIP).** Denatured genomic DNA of desired fragment length (generated by restriction or sonication) is incubated with an antibody directed against 5-methyl-cytosine (a-5mC), and methylated DNA is isolated by immunoprecipitation. Enrichment of target sequences in the methylated fraction can be quantified by standard DNA detection methods such as PCR, by comparing input (IN) to MeDIP (M) DNA, microarrays (MeDIP-chip) or by next generation sequencing technology (MeDIP-seq). Figure was adapted from Weber et al (Weber et al., 2005).

1.3.7 Epigenetic variation in humans

The diversity of human phenotypes is the result of genetic and epigenetic variation and the interaction of these two biological variables with the environment (Hoffmann and Willi, 2008; Jaenisch and Bird, 2003). Although a number of studies have focused on studying genetic variation (Beckmann et al., 2007), the scale and significance of epigenetic variation has only begun to be elucidated (Rakyan and Beck, 2006).

The need to consider epigenetic alongside genetic variation in the context of complex diseases, has been highlighted by: (i). the finding of disease discordance in monozygotic twin studies (especially those living in the same environment) and (ii). the confirmation that epigenetic factors play a decisive role in the aetiology of many of human diseases (Robertson, 2005). To this effect the first systematic effort for cataloguing epigenetic variation was launched in 1999 (Beck et al., 1999). The resulting Human Epigenome Project (HEP) aimed to identify methylation variable positions (MVPs), which are akin to SNPs, promising to advance the understanding and diagnosis of human diseases. Following HEP, recent advances in epigenome mapping (Beck and Rakyan, 2008; Mendenhall and Bernstein, 2008) were beneficial in conducting large-scale comprehensive epigenetic studies looking for epigenetic variation. These studies are expected to have a high impact on our understanding, diagnosis and treatment of complex diseases in the next few years.

1.3.7.1 Differentially Methylated Regions - DMRs

The most frequent and stable form of epigenetic variation is differential DNA methylation. Alterations to the temporal or spatial patterns of DNA methylation which are indicative of local changes in genome functionality can lead to the formation of differentially methylated regions (DMRs). DMRs can vary in size from a few to hundreds or thousands of base pairs and, based on context, can be associated with: (i). specific tissues or cell types (tissue specific DMRs – tDMRs) and (ii). specific phenotypes or disease conditions (phenotype specific DMRs – pDMRs).

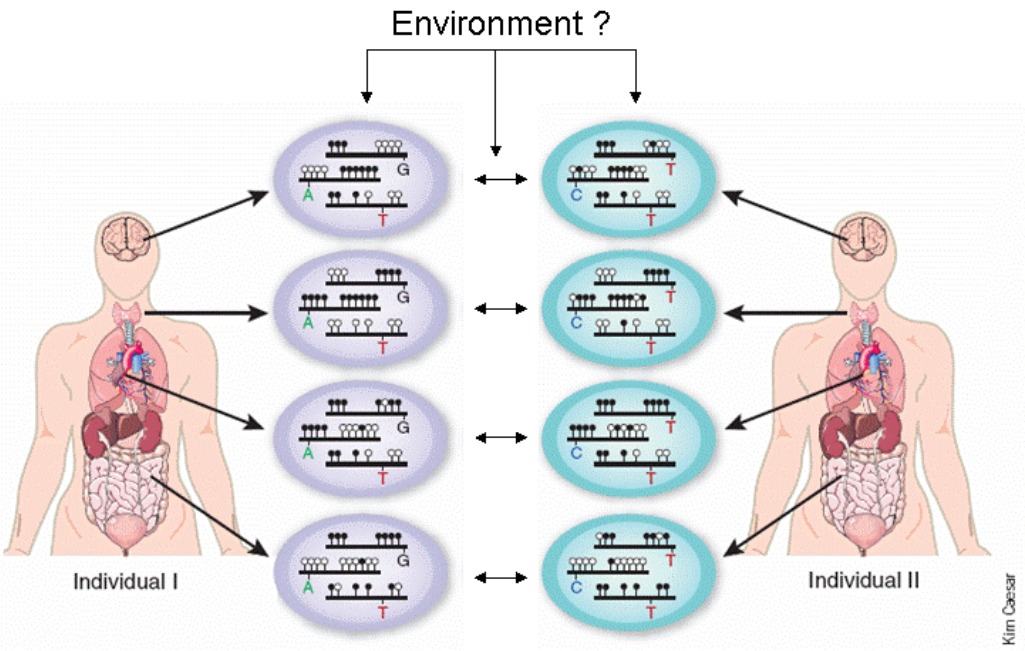

Figure 1.9. **DNA methylation heterogeneity among individuals and cell types**. Cell type–specific and tissue-specific DNA methylation are illustrated by organ-to-organ variations in the clusters of methylated CpGs within the same individual. Despite overall consistency in tissue-specific DNA methylation patterns, variations in these patterns exist among different individuals. Methylated CpGs are indicated by a filled circle and unmethylated CpGs by an open circle. SNPs are indicated by the corresponding base. The potential role of the environment in heterogeneous methylation patterns is also indicated. Figure was adapted from Brena et al. (Brena et al., 2006).

i. Tissue specific DMRs - tDMRs

tDMRs refer to methlyation differences between different cell types with otherwise identical genetic material (figure 1.9). Several large scale DNA methylation studies have started to catalogue such tissue-specific methylation differences (tDMRs) (Eckhardt et al., 2006; Illingworth et al., 2008; Rakyan et al., 2008; Rakyan et al., 2004; Shiota, 2004; Weber et al., 2007). Identification of tDMRs at a genome-wide scale will eventually lead to a better understanding of the role of DNA methylation in setting up and maintaining tissue-specific expression patterns. Comparison of tDMRs within gene loci and gene expression profiles suggested that tDMRs are involved in regulating tissue-specific gene expression. Interestingly, the majority of tDMRs were not within the 5'UTRs of genes but

rather in exons and introns of functionally diverse genes whereas a significant proportion of them found to overlap with evolutionary conserved, non-protein coding regions (ECRs). The latter supports the notion that tDMRs may have a functional role beyond the mere control of transcription via promoter methylation.

One interesting question arising from the existence of tDMRs is the mechanism by which they occur. Large-scale analysis using pluripotent cells suggest that they may arise at early stages of development processes. Comprehensive DNA methylation studies using ES and lineage committed cells (Bibikova et al., 2006; Farthing et al., 2008; Meissner et al., 2008; Mohn et al., 2008) will give insights into whether epigenetic marks in early development have a primary role in determining tissue-specific expression patterns and hence tissue-specific identities.

ii*. Phenotype specific DMRs – pDMRs*

pDMRs reflect epigenetic variation within cell types from the same origin between different individuals (figure 1.9). Based on the source of this variability inter-individual DMRs can be divided into two classes:

1. DMRs that are driven by genetic variation. Variation in DNA methylation levels can be affected by genetic variation directly by the introduction or removal of CpG sites, or indirectly by the introduction of sequences (e.g. repeat elements or transposons) (Lippman et al., 2004) that affect methylation in *cis*. It has been shown that in Beckwith-Wiedeman syndrome (BWS) patients, specific haplotypes within the IGF2 locus have been associated with loss of methylation (Murrell et al., 2004), supporting the notion that the genotype can act synergistically with the epigenotype. This led to the introduction of the term ‘hepitype’ which combines the contribution of a haplotype and an epitype to a given phenotype (Murrell et al., 2005). This concept was further supported by a recent study reporting allele-specific DNA methylation (ASM) at 16 SNP-tagged loci distributed across various chromosomes (Kerkel et al., 2008). The authors of this paper introduced

the term ‘epihaplotype’, which is akin to ‘hepitype’ (Murrell et al., 2005), to describe sequence-depended DNA methylation patterns. These findings can be useful for fine mapping and interpretation of non-coding RNAs and for their association with complex diseases.

2. DMRs that are generated by stochastic events independently of genetic variation. Such DMRs can arise due to errors during DNA replication. Stochastic events that lead to variation in DNA methylation patterns can occur in cell culture condition (*in vitro*) or *in vivo* during ageing. This is supported by a recent study reporting epigenetic differences between aging monozygotic twins (Fraga et al., 2005). However, as this study did not examine the same individuals serially over time, it is not clear if the methylation differences observed occurred over time or were present historically. A subsequent study found no age-related variation in DNA methylation, but again this study did not track the same individuals over time.(Eckhardt et al., 2006). Such DMRs can lead, in many cases, to cancer development (the single leading risk factor for cancer is age) and they may explain the adult onset of a number of complex, non-malignant diseases (Feinberg, 2004; Feinberg, 2008).

These stochastic events can be influenced by the environment. It has been reported that decreased grooming and nursing by rat mothers reduced DNA methylation at a glucocorticoid receptor gene promoter in the hippocampus of the offspring, resulting in increased stress response in later life (Weaver et al., 2004). This example demonstrates how environmental stimuli during childhood could affect phenotypic outcomes in later life through the epigenome. This might have great relevance to phenotypic differences observed between monozygotic twins growing up in different environments.

1.3.7.2 DMR identification goes global in the human genome

The first large scale study for DMR identification was launched in 1999 by Stephan Beck and colleagues (Beck et al., 1999). The Human Epigenome Project (HEP) aimed to generate methylation data for selected regions of the human genome in both normal and disease tissues by bisulphite sequencing (see above) using locus specific PCR primers. HEP has generated DNA methylation profiles for three human chromosomes (6, 20 and 22) in 12 different tissues revealing novel insights in tissue specificity of DNA methylation patterns (Eckhardt et al., 2006; Rakyan et al., 2004).

Since then, and as DMRs and epigenetic variability in general have been recognised as an important factor in determining genome functionality, other large-scale efforts have emerged aiming for systematic mapping of the human epigenome. Both bisulphite sequencing and MeDIP-technology (see above) have been adapted to be used with microarray and next-generation sequencing platforms. Application of MeDIP together with illumina GA platform led to the first human methylome (Down et al., 2008). Figure 1.10 summarises major landmarks form the launch of the HEP (1999) to the first human methylome (2008).

It is now clear that global DNA methylation profiling has come of age. Continuous efforts in comprehensive cataloguing of DMRs in multiple tissues and cell types as well as during differentiation and phenotype development will provide a broad basis for future studies aiming to understand how the (epi)genome functions in health and disease.

.

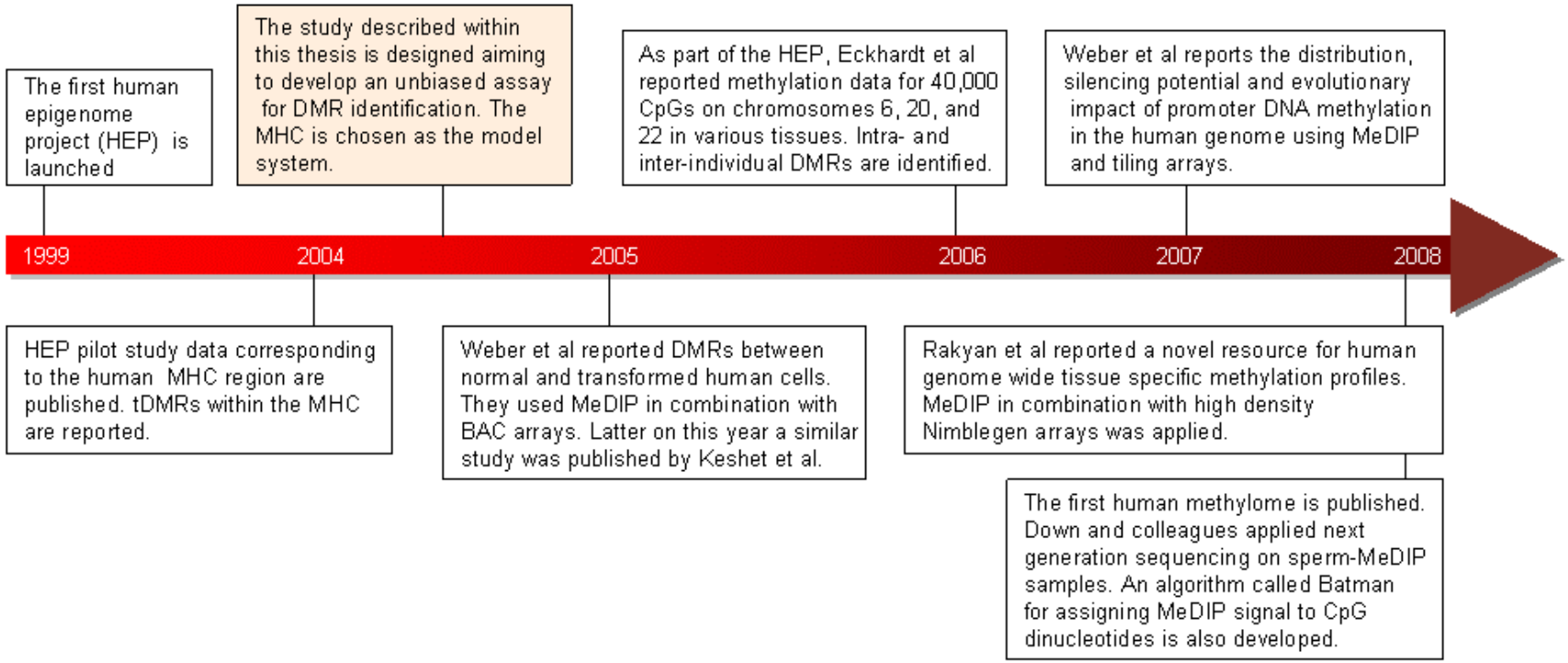

Figure 1.10. **Selected landmarks in large-scale DNA methylation studies and DMR identification in the human genome.**

1.4 Rationale of my thesis

The aim of this project was to identify and characterise differentially methylated regions (DMRs) that can contribute to phenotypic plasticity. This study was designed in April 2005. At that time the HEP pilot study had been published, reporting the existence of DMRs within the MHC region and Weber and colleagues had at this time developed MeDIP. MeDIP was used in combination with BAC arrays (100 kb resolution) for methylation profiling and, more specifically, for the identification of DMRs between normal and cancer samples (figure 1.10) (Weber et al., 2005).

In this context, I chose to use the MHC as a model system and develop a higher resolution array-based assay for the unbiased identification of DMRs. The assay involved MeDIP in combination with an MHC tiling array covering the whole MHC at 2kb resolution (chapter 3). This assay was used to perform two screens: (1). tDMR screen, looking for DMRs associated with specific tissues (chapter 4) aiming to give insights in the role of DNA methylation in tissue specific gene expression, and (2). pDMR screen, looking for DMRs associated with a particular phenotype (chapter 5) and aiming to elucidate the role of epigenetic variation in complex diseases. The phenotype I tested is the MHC class I down-regulation (MHC class I^- phenotype).

Both screens were successful, verifying further the implication of DMRs, and of epigenetic variation, in the regulation of MHC loci and hence in MHC-linked phenotypic plasticity.

In addition, I performed DNA methylation and gene expression analysis of genes that may be implicated in the MHC class I^- phenotype and are encoded outside the MHC region (chapter 6). The findings of this analysis verify further the complexity of MHC-linked phenotypes.

Finally I discuss and summarise my findings (chapter 7) and discuss future directions for studies to elucidate complex diseases associated with the MHC region.

In summary the work described within this thesis supports the notion that epigenetic variation plays a decisive role in the development of normal and aberrant phenotypes and hence it should necessarily be considered, together with genetic variation, as an important factor when studying complex diseases.