Epigenomics: Genome-Wide Study of Methylation Phenomena

K. L. Novik*, I. Nimmrich², B. Genc², S. Maier², C. Piepenbrock², A. Olek² and S. Beck¹

¹Wellcome Trust Sanger Institute, Genome Campus, Hinxton, Cambridge CB10 1SA, UK
²Epigenomics AG, Kastanienallee 24, 10435 Berlin, Germany

Abstract

Epigenetics is one of the key areas of future research that can elucidate how genomes work. It combines genetics and the environment to address complex biological systems such as the plasticity of our genome. While all nucleated human cells carry the same genome, they express different genes at different times. Much of this is governed by epigenetic changes resulting in differential methylation of our genome – or different epigenomes. Individual studies over the past decades have already established the involvement of DNA methylation in imprinting, gene regulation, chromatin structure, genome stability and disease, especially cancer. Now, in the wake of the Human Genome Project (HGP), epigenetic phenomena can be studied genome-wide and are giving rise to a new field, epigenomics. Here, we review the current and future potential of this field and introduce the pilot study towards the Human Epigenome Project (HEP).

Introduction

The field of epigenetics is enjoying a meteoric rise from the euphoria surrounding the human genome project. The free availability and subsequent publication of the human draft sequence generated the essential foundation from which to explore the greater genetic complexities of life (IHGSC, 2001). Attention has now turned from the identification of genes to the expression and function of genes; epigenetics must surely hold one of the master keys to unlock and understand the mechanisms governing these vital processes.

Conrad Waddington suggested the phrase “epigenetics” in 1942, to describe the “study of the processes by which genotype gives rise to phenotype” (reviewed by Wu and Morris, 2001). The definition has evolved over the years to keep pace with the ever-expanding knowledge, and the most current classification of epigenetics would be: “a study of the changes in gene expression that are mitotically and/or meiotically heritable and do not involve a change in the DNA sequence” (Wu and Morris, 2001). Epigenetic mechanisms are responsible for several phenomena including: X-inactivation – the random silencing of one of the X chromosomes in every normal somatic cell of female mammals (reviewed by Park and Kuroda, 2001) and genomic imprinting – the expression or repression of certain genes according to their parental origin (reviewed by Ferguson-Smith and Surani, 2001). From the latter half of the last century the simple addition of methyl groups, to DNA, was recognised as the main mechanism of epigenetics. It was soon accepted that DNA methylation could be variously associated with gene expression, genomic stability and disease (in particular the progression to malignancy), but proof remained elusive. The discovery of methylation-specific enzymes, within the past 20 years, provided much of that proof and established a direct link between DNA methylation, gene expression and changes in chromatin structure. The focus now is to understand global gene function and to probe the susceptibility to common diseases. Therefore, mechanisms that affect gene expression are at the forefront of current and future investigations. After years of struggling to be accepted as serious science, epigenetics - and in particular DNA methylation – has been drawn into the biological limelight. In order to reflect the fact that this branch of science is now being investigated genome-wide, it seems timely to add the term “epigenomics” to our scientific vocabulary.

Here we aim to provide an overview on DNA methylation, its function, and the potential association with malignancy and other common diseases. Furthermore, we provide an exploratory outlook of the novel field of epigenomics by introducing the pilot project that is spearheading the planned human epigenome project.

Cytosine methylation - the forgotten 5th base

The major target for methylation in the mammalian genome is cytosine; the enzymatic attachment of a methyl group to carbon-5 of the pyrimidine ring creates 5-methylcytosine (Doerfler, 1983; Santi et al., 1983; Bird, 1992). This often forgotten 5th base of genomic DNA is indistinguishable from cytosine in its complementary base pairing with guanine. Furthermore, the majority of 5-methylcytosine (5-MeC) occurs in the context of the dinucleotide CpG, although CpNG, CC(a/t)GG, CpA and CpT can also be methylated at a very low frequency (Clark et al., 1995; Woodcock et al., 1997; Lorincz and Groudine, 2001). It is worth mentioning that in plants and filamentous fungi, cytosine methylation at non-CpG sites occurs much more frequently than in animals (Martienssen and Colot, 2001). Furthermore, in prokaryotes such as E.coli and Agrobacterium tumefaciens the N-6 position of adenine can become methylated (Kahng and Shapiro, 2001). Prokaryotic cytosine and adenine methylation can influence...
gene transcription, affect cell viability and also serve the restriction-modification systems that protect the bacterial host DNA from cleavage by specific endonucleases (Kahng and Shapiro, 2001).

CpG dinucleotides are non-randomly distributed across the mammalian genome; a phenomenon that appears to be driven, at least in part, by their methylation status. 5-MeC can spontaneously deaminate to thymine and, as a consequence, C to T transitions accumulate over evolutionary time (Duncan and Miller, 1980; Bird, 1987). By contrast, non-methylated cytosines deaminate to uracil, which is recognised as a “non-DNA” base and is promptly removed (Duncan and Miller, 1980; Bird, 1987). The interplay of CpG density and methylation is an interesting relationship and has created some fascinating genomic features. Three major targets for cytosine methylation are: CpG islands, G+C isochores and CpG hotspots.

CpG islands

The tendency of 5-MeC to spontaneous deamination means that most CpGs are present only once in every 80 dinucleotides, representing a one-fifth reduction of that predicted from the average base composition (40% G+C), and they are usually methylated (Subak-Sharpe et al., 1966; Cooper and Krawczak, 1989). In contrast, a small proportion of total CpGs are clustered into “CpG islands”. These regions are on average 1-2 kb in length, with a CpG frequency close to that predicted from random base composition, and are unmethylated under normal circumstances (Bird et al., 1985; Bird, 1987; Gardiner-Garden and Frommer, 1987, IHGS Consortium, 2001). The recent completion of the human genome draft sequence identified almost 29,000 potential CpG islands, with most chromosomes harbouring 5-15 islands per Mb (IHGS Consortium, 2001). There is a fairly good correlation between the number of CpG islands and gene density, with chromosome 19 having by far the highest average density of CpG islands (IHGS Consortium, 2001). CpG islands often encompass the promoters and early exons of genes (Bird, 1987; Gardiner-Garden and Frommer, 1987) and it is no surprise that these transcription-controlling regions should normally be unmethylated, when considering that methylation often leads to transcriptional silencing. Genes with tissue-specific expression patterns may be transcriptionally silent at given times but their associated CpG islands can remain unmethylated e.g. the human α-globin gene (Bird, 1987). It has been speculated that promoters, expressed in early development when methylation memory is most active, may create these unmethylated CpG islands (Bird, 2002). Furthermore, if transgene promoter function is impaired in transgenic mice then the associated CpG island becomes methylated (Macleod et al., 1994). These results may be interpreted as the failure of transcription permitting de novo methylation (Bird, 2002). It is particularly noticeable during aging and the development of diseases, such as cancer (see below), that CpG island methylation increases (reviewed by Issa, 2000; Baylin and Herman, 2000). The mechanisms permitting CpG island de novo methylation remain elusive but may reflect cell-wide changes to the associated factors that regulate methylation (Jones and Laird, 1999).

G+C isochores

The human genome can be divided into large (>300 kb) mosaic structures, or isochores, which are compositionally fairly homogeneous and fall into distinct ranges of G+C content (for review see Bernardi, 2000; Bernardi, 2001). Five isochore families exist: L1 and L2 are G+C poor (~40%), H1 and H2 are G+C rich (45% and 50% respectively) and H3 is G+C very rich (53%). The recent completion of the draft human sequence (IHGSC, 2001) has provided a valuable resource for investigators to determine the precise isochore map of the entire genome (Pavlíček et al., 2002a). It has been demonstrated that the G+C content does equate to gene density, with H3 isochores containing the greatest number of genes, CpG islands and CG dinucleotides.

Some of the best-studied isochores in the human genome map to the major histocompatibility complex (MHC) on the short arm of chromosome 6. The MHC is divided into a number of different classes (see below) and the extended class II is an H2 isochrome, class II is an L-family isochrome and class III is an H3 isochrome (Fukagawa et al., 1995; Stephens et al., 1999). Recently it has been shown that the MHC isochores are highly conserved between human and mouse, suggesting that strong selection is responsible for maintaining these structures (Pavlíček et al., 2002b). Furthermore, a definitive correlation has been found between the timing of DNA replication and different isochrome families (Federico et al., 1998; Federico et al., 2000). Regions which are G+C rich, and therefore gene rich, replicate early in the cell cycle and condense late; the opposite is true for G+C poor regions. Precise switching of replication timing has been observed at the boundary between the isochores of the MHC classes II and III (Tenzen et al., 1997). At the heart of replication timing lies chromatin, which controls the packaging and accessibility of the DNA. It has also been shown that covalent modifications of the exposed histone tails of chromatin (including methylation, acetylation and phosphorylation) may influence the interactions with accessory proteins and ultimately determine the transcriptional fate of the underlying DNA (for review see Jenuwein and Allis, 2001). As discussed below, methylation signals on DNA attract methyl-binding proteins (MBDs), recruit histone deacyetylases (HDACs) and initiate local chromatin remodelling and condensation (reviewed by Robertson and Wolffe, 2000). Knowledge of methylation patterns across well-studied regions such as the MHC, would allow possible correlations between isochores, replication timing and temporal chromatin structure to be inferred; these all have major implications on the biological activity and stability of this region. Therefore, it is not surprising that the MHC has been chosen as the model system for such a study (see below).

CpG hotspots

CpG sites have been shown to act as hotspots for mutations. Up to 30 % of point mutations in the germline have been estimated to result from MeCpG → TpG transitions caused by the spontaneous deamination of 5-methylcytosine (Cooper and Youssoufian, 1988). The p53
gene represents a well-studied example of this type of mutation. It has been shown that the CpG at codon 248 and two flanking CpGs of p53, were methylated in all tissues and cell lines studied and could therefore potentially act as mutational hotspots (Magewu and Jones, 1994). Also, approximately 50% of inactivating mutations in p53, occurring in colorectal cancer, are at CpG dinucleotides (Fearon and Jones, 1992).

Methylated cytosines can also serve to direct some carcinogens to CpG dinucleotides, which are the preferential targets for G to T transversion mutations in smoking-associated lung cancers (Yoon et al., 2001). These point mutations are thought to result from polycyclic aromatic compounds such as benzo[a]pyrene and other bulky chemical carcinogens reacting with guanosines that are flanked by 5-MeC (Pfeifer et al., 2000). DNA methylation therefore promotes genetic alterations that may contribute towards carcinogenesis.

DNA methylating enzymes

Methyl groups are enzymatically added by DNA methyltransferases (DNMTs). Four mammalian DNMTs have been isolated and characterised to date: DNMT1 (Bestor et al., 1988), DNMT2 (Yoder and Bestor, 1998), DNMT3A and DNMT3B (Okano et al., 1998a). DNA methyltransferases have two well-characterised functions: the maintenance of methylation signals and de novo methylation. DNMT1 has been christened the global "maintenance methylase" because of an affinity for the hemi-methylated DNA that arises after DNA replication (Pradhan et al., 1997; Pradhan et al., 1999). This methylase ensures that the pre-replication methylation pattern is restored to the new DNA strand (Gruenbaum et al., 1982; Stein et al., 1982). Loss of DNMT1 is lethal; mouse embryos that are deficient in Dnmt1 die after gastrulation because of genome-wide demethylation (Li et al., 1992). There is no direct evidence to exclude Dnmt1 from a role in de novo methylation and indeed methyltransferases do display overlapping functions (Bestor, 2000). Interestingly, the overexpression of DNMT1 leads to de novo methylation of endogenous CpG islands in cancer cell lines (Vertino et al., 1996; Rhee et al., 2000).

The role of DNMT2 is not yet fully understood and although it contains the conserved motifs characteristic of cytosine methyltransferases, it seems to lack such an activity in vitro (Okano et al., 1998b). There is however some evidence for controlling of centromeric regions (Bestor, 2000).

DNMT3A and DNMT3B have de novo methylase activity and play a major role during early embryogenesis (Okano et al., 1999). Shortly after fertilisation, the genome of the mouse zygote becomes demethylated (except for certain imprinted genes) by a poorly understood mechanism (reviewed by Razin and Cedar, 1993; Reik et al., 2001). Genome-wide remethylation is orchestrated de novo by Dnmt3a and Dnmt3b at the time of implantation and is essential for normal mammalian development (Monk et al., 1987; Razin and Cedar, 1993; Reik et al., 2001). Mice that lack both Dnmt3a and Dnmt3b have abnormal morphology and die shortly after gastrulation (Okano et al., 1999). Interestingly, despite the lack of de novo methylation these mutants still maintain the methylation patterns that were pre-existing, presumably as dnmt1 remains functional (Okano et al., 1999). Dnmt3a and Dnmt3b also exhibit some non-overlapping functions in development and have distinct expression patterns. Mice lacking either Dnmt3a or Dnmt3b display quite different developmental defects and ages at death (Okano et al., 1999). Dnmt3b in particular is needed for the methylation of centromeric minor satellite repeats in the mouse. This function appears to be conserved evolutionarily; humans with the rare autosomal recessive ICF syndrome show hypomethylation of their pericentromeric heterochromatin and have been found to harbour mutations in DNMT3B (Jeanpierre, 1993; Hansen et al., 1999; Okano et al., 1999; Xu et al., 1999).

It is interesting to note that clonal cell populations show heterogeneity in their DNA methylation patterns, maybe suggestive of variable fidelity in maintaining the original methylation signals (Pollack et al., 1980; Silva et al., 1993). As a consequence of these variable epigenetic signals, genotypically identical cells may have quite different phenotypes (Beck et al., 1999). Furthermore, an enzyme with DNA demethylase activity has been reported recently but the biochemical characterisation is incomplete and the absence of the enzyme is still actively debated (Jost et al., 1995; Bhattacharya et al., 1999).

DNA methylation function

Two established functions of DNA methylation, that are not entirely mutually exclusive, involve the suppression of gene expression and the maintenance of genome integrity by the suppression of repetitive elements and heterochromatin.

Methylation modulates the interactions between DNA and proteins to bring about a change in chromatin structure that ultimately affects transcription (reviewed by Robertson and Wolffe, 2000; Jones and Takai, 2001). More often than not, methylation of the vertebrate genome results in transcriptional silencing. In some instances local methylation can directly interfere with the binding of transcription factors and thus prevent transcription (Iguchi-Ariga and Schaffner, 1989; Tate and Bird, 1993). Methylation of the regulatory sequences for transcription factors myc and AP-2 has been shown to inhibit their binding (Comb and Goodman, 1990; Prendergast and Ziff, 1991). Similarly, DNA methylation blocks the interaction between the zinc finger protein CTCF and the parentally imprinted Igf2/H19 insulator (Hark et al., 2000; Holmgren et al., 2001). However, many transcription factors are not fully influenced by the presence of 5-MeC in their recognition sequences or the sequences may not contain CpG sites at all. Sp1 controls the expression of many genes that have CpG rich promoter regions but binding is not influenced by the methylation of one CpG (Harrington et al., 1988). The binding is only inhibited when the sequence CpCpG has both cytosines methylated (Clark et al., 1997). Clearly an additional mechanism of methylation-mediated gene silencing was essential to account for global gene suppression.
In 1992, the first methyl-CpG binding protein, MECP2, was characterised (Lewis et al., 1992). This protein has a great affinity for binding to methylated CpGs and then recruiting co-repressor complexes such as histone deacetylases (HDACs) to the site (reviewed by Robertson and Wolffe, 2000). MeCP2 recruits transcriptional co-repressors and the histone deacetylase, HDAC1, via the bridging protein Sin3 (Jones et al., 1998; Nan et al., 1998).

HDACs deacetylase lysine residues on the histone tails, thus facilitating the interaction with neighbouring histones and the formation of a dense-packed chromatin structure that is repressive to gene expression. Additional methyl binding proteins (MBDs) have been identified over the last few years and have been called MBD 1-4 (Hendrich and Bird, 1998; reviewed by Wade, 2001). It is not yet clear whether hypoacetylation or methylation is the crucial step for gene silencing. Methylation induced gene silencing is prevented by treating cells with trichostatin A (TSA), a histone deacetylase inhibitor (Kass et al., 1997). However, in oestrogen receptor (ER) negative breast cancer cells, re-expression of the receptor is permissible only by synergistic inhibition of DNA methylation and histone deacetylation (Yang et al., 2000; Yang et al., 2001). As well as acetylation, histone tails can also be modified by methylation and phosphorylation (reviewed by Jenuwein and Allis, 2001). It was recently shown that the methylation of histones could control DNA methylation in the filamentous fungus Neurospora crassa (Tamaru and Selker, 2001). Evidence suggests that the methylation of lysine 9 of histone H3 is needed for the active recruitment of DNMTs in N. crassa (Tamaru and Selker, 2001). Similarly, in a mammalian model, a methylation marker on Lys9 of H3 causes the recruitment of heterochromatin proteins and the subsequent repression of transcription (Bannister et al., 2001). In both of these examples it would appear that methylation of histones could be the initiating event in changing the methylation status on the DNA level. Clearly, further studies are needed to unravel the complex interdependence of DNA methylation and histone modification in the pathway leading to transcriptional repression.

Closer observation of the mammalian genomic distribution of 5-MeC, reveals that the vast majority resides within retrotansposons; these repetitive sequences account for approximately 40% of the genome and are CpG rich (Yoder et al., 1997; Colot and Rossignol, 1999). It has been proposed that DNA methylation evolved as a host-defense mechanism to silence the expression and thereby limit the rogue-integration of these repetitive elements into transcriptional units (Yoder et al., 1997; Bestor, 1998). Expression of integrated repetitive elements, from their strong promoters, can result in the formation of aberrant transcripts from the targeted transcriptional unit, irrespective of the orientation of integration (Robertson and Wolffe, 2000). Furthermore, reduced methylation of repetitive elements can result in chromosomal instability. Embryonic stem (ES) cells that are homozygous mutant for DNMT1 show hypomethylation of their repetitive elements, which are subsequently transcribed (Walsh et al., 1998). There is also a 10-fold increase in the rate of mutations, involving gene rearrangements, compared to the wild type ES cells (Chen et al., 1998). Pericentromeric heterochromatin is also heavily methylated under normal situations and this appears crucial to maintaining chromosome conformation and preventing aberrant recombination (Ehrlich, 2000). Hypomethylation of the centromeric heterochromatin of chromosomes 1 and 16 leads to a number of gross chromosomal abnormalities such as isochromosomes, translocations and whole-arm deletions (Ehrlich, 2000). DNA methylation may suppress homologous recombination by blocking recombination initiation sites or by condensing the chromatin structure (Robertson, 2001).

DNA methylation and cancer

It is some 15 years since DNA methylation in tumor cells was recognised as being significantly different to the situation in normal cells (Goelz et al., 1985; Feinberg et al., 1988). These methylation alterations are often the earliest and most frequent events known to occur in human cancers (Jones, 1986). At least three different mechanisms exist in which DNA methylation contributes to tumor formation: mutational hotspots at CpG positions (as discussed above), general hypomethylation of genomic DNA and inactivation of tumor suppressor genes by hypermethylation of CpG islands.

Hypomethylation

Tumor cells can display global hypomethylation of the genome. Interestingly, this decrease in genomic methylation can be detected very early in malignancy, even before tumor formation (Christman et al., 1993, Pogribny et al., 1997). Furthermore, hypomethylation of oncogenes can generate increased and inappropriate gene expression. This has been shown for bcl-2 in lymphocytic leukemia (Hanada et al., 1993) and k-ras in lung and colon carcinomas (Feinberg and Vogelstein, 1983). Other regions frequently hypomethylated in tumors are transposons and repetitive elements such as SINES (short interspersed nuclear elements) and LINES (long interspersed nuclear elements). In urothelial cell carcinomas, LINE-1 retrotransposons and HERV-K (human endogenous retrovirus) proviral DNA are strongly hypomethylated compared to the respective normal tissue and transcription of these elements was observed (Florl et al., 1999). Several other studies have also demonstrated that repetitive elements show reduced methylation in tumors and that this effect correlates with the progression of disease (Qu et al., 1999). Additionally, tumors with global hypomethylation frequently display centromeric rearrangements of chromosomes 1 and 16; these centromeres are particularly rich in heterochromatin that would be methylated under the normal situation (Qu et al., 1999).

Hypermethylation

The most prominent alterations of DNA methylation in cancer involve hypermethylation of CpG islands that are
usually, but not exclusively, unmethylated in normal tissue. In tumors, CpG islands that span the promoter regions of certain genes are often methylated in contrast to the bulk DNA that is hypomethylated. This CpG hypermethylation in tumors is often inversely correlated with gene expression. Knudson’s “two hit” hypothesis of tumor generation assumes that for full inactivation of a tumor suppressor gene, two independent inactivating events are required; one at each of the specific alleles (Knudson, 1971). Initially, such “hits” were thought to be only intragenic mutations and/or loss of chromosomal DNA (loss of heterozygosity or LOH). It is now well accepted that promoter CpG methylation and the subsequent transcriptional repression of tumor suppressor genes, can also act as one of Knudson’s “hits” (reviewed by Jones and Laird, 1999; Baylin and Herman, 2000). In the last few years a number of genes that are silenced by promoter CpG hypermethylation, in different kinds of tumors, have been described. Table 1 provides a summary of the most important of these genes with their respective references. A prominent example is the gene CDKN2A, which encodes the cyclin-dependent kinase inhibitor p16-ink4A. As a “genomic gatekeeper”, p16 assumes that for full inactivation of a tumor suppressor gene, leads to a reduction of promoter methylation, reexpression of the gene and a partial abolition of the MIN+ phenotype (Herman et al., 1998).

Hitherto, we have only considered the association between DNA hypermethylation in cancer and transcriptional silencing. For those genes lacking CpG islands, the methylation density within the gene structure may correlate with transcriptional activity. Furthermore, methylation of downstream CpG sites is often correlated with increased transcription (Chan et al., 2000). In chronic myelogenous leukaemia (CML), a chromosomal translocation juxtaposes the ABL protooncogene coding region with the 5'-untranslated region of the BCR gene (reviewed by Laurent et al., 2001). The BCR-ABL fusion protein is therefore expressed from the BCR promoter. One of the ABL promoters can become embedded within the fusion gene and may be transcribed normally. However, as the disease progresses the ABL promoter becomes heavy de novo methylated and thus silenced; the fusion gene appears to be relatively overexpressed in comparison (Zion et al., 1994). A positive correlation between gene expression and de novo methylation has also been reported for exon 5 of PAX6 in colon and bladder cancer (Salem et al., 2000).

**Tumour classification by methylation patterns**

Several groups have shown that tumour classes can be determined by microarray-based expression profiling. Golub and colleagues (Golub et al., 1999) investigated the expression levels of some 6800 genes and discovered that the profiles of just 50 informative genes were sufficient to distinguish between acute lymphoblastic leukaemia (ALL) and acute myeloid leukaemia (AML). Using a similar gene expression approach, subclasses of diffuse large B-cell lymphoma could be classified according to their differences in response to therapy and disease outcome (Alizadeh et al., 2000).

A number of recent studies have also shown an association between the pattern of DNA methylation in a tumor cell and tumor development. Costello and colleagues
used restriction landmark genomic scanning to show that methylation patterns are tumour-type specific (Costello et al., 2000). Highly characteristic DNA methylation patterns have been shown for breast cancer cell lines (Huang et al., 1999). Unique hypermethylation profiles, in several types of cancer, have also been demonstrated for a limited number of genes (Eads et al., 2000; Esteller et al., 2001). Zion and colleagues showed that de novo methylation of the normal ABL promoter, embedded in the BCR-ABL fusion gene, could be used as a reliable marker for tracking the progression of CML (Zion et al., 1994). As with mRNA expression profiles, genome-wide methylation patterns should represent a molecular fingerprint of cancer tissues; determination of susceptibility to malignancy and prediction of outcome may be feasible using methylation profiling.

Methods to assess genome-wide methylation status have often employed the digestion of genomic DNA with methylation-sensitive enzymes; this unfortunately limits the analysis to sites for which these specific enzymes are available (Liang et al., 1998; Toyota et al., 1999; Yan et al., 2000). The methylation status of specific genes can be assessed by sodium bisulphite treatment of genomic DNA; bisulphite converts all unmethylated cytosines to uracil but methylated cytosines are unchanged (Hayatsu et al., 1970; Hayatsu, 1976). Several methods may then be employed to distinguish between the resulting sequence variants; e.g. bisulphite sequencing (Frommer et al., 1992), methylation-specific PCR (Herman et al., 1996b) and ‘MethylLight’ PCR (Eads et al., 2000). Some of these approaches have been applied to determining the methylation status of large patient populations (Eads et al., 2000; Esteller et al., 2001) but they cannot analyse many genes in parallel. Recently, two novel microarray-based assays has been introduced that can perform this parallel analysis (Gitan et al., 2001; Adorjan et al., 2002). In the assay described by Adorjan and colleagues, specific CpG positions are PCR amplified from bisulphite-treated genomic DNA. The resulting sequence differences, between originally methylated and unmethylated CpGs, are detected by hybridising the PCR products to oligonucleotides immobilised on a glass slide. These oligonucleotides detect the originally methylated (CG) or unmethylated (TG) form of the CpG. Figure 1 illustrates this novel assay. Using supervised and unsupervised learning algorithms to analyse the results, it was possible to distinguish between acute lymphoblastic leukaemia and acute myeloid leukaemia (Model et al., 2001; Adorjan et al., 2002). They could also differentiate between primary prostate carcinomas, primary clear-cell kidney carcinomas and their respective normal controls. Some 200 CpG dinucleotides, located in the regulatory regions of 50 genes, were evaluated for methylation status in this study (Adorjan et al., 2002). This certainly demonstrates that the DNA methylation patterns of a cell can be used to monitor tumorigenesis and they may be of clinical value if the prognostic and predictive relevance can be extrapolated.

DNA methylation and non-cancerous diseases

The sinister partnership between methylation and cancer is extensively documented but it should not be forgotten that abnormalities of the methylation pathway have also been linked with other human diseases. Genomic imprinting, which is not extensively discussed in this review, is intimately linked with DNA methylation. The best-characterised syndromes arising from defective imprinting are Beckwith-Wiedemann syndrome (BWS) and the Prader-Willi/Angelman syndromes (for reviews see Reik et al., 1997; Nicholls et al., 1998). ICF syndrome is another disorder associated with aberrant methylation and results from mutations in the de novo methylase, DNMT3B (Hansen et al., 1999; Okano et al., 1999; Xu et al., 1999). Rett syndrome, an X-linked neurodevelopmental disorder that occurs almost exclusively in females, is caused by mutations in the methyl-CpG binding protein, MECP2 (Hagberg et al., 1983; Amir et al., 1999). Mutations in the X-linked ATRX gene change the methylation patterns of highly repetitive sequences and as a consequence cause a syndrome characterised by alpha-thalassaemia and mental retardation (Gibbons et al., 2000). Fragile X syndrome is a common form of mental retardation, that is also X-linked, and caused by aberrations in the FMR1 gene (Kremer et al., 1991; Oberle et al., 1991). FMR1 has a polymorphic CGG repeat in the untranslated region of exon 1 that is prone to great expansion in fragile X patients. This expansion encourages de novo methylation of the repeat but unfortunately also methylates the promoter of FMR1 and therefore aberrantly silences the gene (Kremer et al., 1991; Oberle et al., 1991). The diverse phenotypes of these disorders suggest that methylation-mediated gene regulation is a highly complex process. Is there any evidence to suggest that aberrant methylation may be contributing towards the aetiology of more common diseases, such as diabetes mellitus?

A role for DNA methylation in diabetes?

Diabetes mellitus is classified into two principal groups. Type-1 diabetes mellitus (or juvenile diabetes) involves the autoimmune destruction of pancreatic islet cells by circulating autoantibodies, leading to the consecutive failure of insulin secretion. Type-2 diabetes mellitus comprises a heterogeneous group of mild forms of diabetes and the most prominent feature is the insulin resistance of target tissues such as liver, adipose tissue and muscle. This resistance leads to reactive hyperinsulinaemia and overt diabetes develops when the pancreatic beta cells can no longer maintain the increased insulin demands. Patients with type-2 diabetes can be grouped, on the basis of body weight, into obese and non-obese subtypes; further subclassification based on circulating insulin levels and other laboratory parameters is currently not possible. Up to 85% of patients with type-2 diabetes are obese, most frequently with an abdominal distribution of fat and a high waist-to-hip ratio. Genetic predisposition is a major risk factor for developing type-2 diabetes. It has been suggested that polymorphisms in genes involved in insulin secretion and response might modify individual disease susceptibility. However, in large population based studies only a few polymorphisms in such genes could be shown to influence the incidence of diabetes (Altshuler et al., 2000; Hara et al., 2000; Rissanen et al., 2000).
Figure 1. A novel microarray-based assay for analyzing DNA methylation (Adorjan et al., 2002).
(a) Illustration of the principle of the assay. Bisulphite-treated genomic DNA is amplified and the resulting fluorescently labelled PCR products are hybridized to oligonucleotides immobilized on a glass slide. For every CpG position there is a corresponding pair of oligonucleotides: one specific for the CpG position in its methylated (CG) and one in its unmethylated state (TG).
(b) A scan image, in false colours, resulting from a hybridized glass slide. The intensities of two adjacent spots (one containing the oligonucleotide specific for CG and the other one for TG), allow calculation of the methylation rate at the corresponding CpG position.
Recent studies have shown that development of obesity and type-2 diabetes is associated with changes in the expression levels of several genes (Goldfine et al., 1999; Smith et al., 1999; Andreelli et al., 2000; Huang et al., 2000; Nikoulina et al., 2000; Stepan et al., 2001). In a mouse model of type-2 diabetes and obesity, progression from a lean state to obesity and overt hyperglycemia was associated with changes in gene expression inverse to those seen in adipocyte differentiation (Nadler et al., 2000). In a similar study in humans, gene expression was analyzed in omental fat samples taken from lean and obese nondiabetic subjects and obese type-2 diabetic patients (Corominola et al., 2001). The authors identified some 2000 cDNAs that showed potential differential expression between each of these groups (Corominola et al., 2001). Furthermore, the upregulation of several genes, in response to insulin, was completely abrogated in type-2 diabetic patients compared to control subjects, insulin-resistant nondiabetic obese patients, and hyperglycemic type-1 diabetic subjects (Ducluzeau et al., 2001).

Interestingly, several oral anti-diabetic agents have been shown to correct the altered gene expression found in animal models of type-2 diabetes (Maier et al., 2000; Stepan et al., 2001). As a major regulator of transcriptional activity, DNA methylation may also be involved in the expression differences observed in type-2 diabetes.

Changes in methylation patterns have been described for several genes examined during the differentiation of various cell types (Takei et al., 1996; Tutt Landolfi et al., 1997; Yokomori et al., 1999; Rothenburg et al., 2001). Furthermore, DNA methylation represents a stable cellular memory for the maintenance of expression patterns, and may contain more information about the differentiation state of a population than the gene expression patterns themselves. T helper–1 (Th1) and –2 (Th2) cells exhibit striking differences in their patterns of cytokine expression; the expression is maintained in quiescent Th1 and Th2 cells even in the absence of further stimulation. The persistence of these expression patterns is thought to be due to epigenetic changes in chromatin structure, locus accessibility and DNA methylation (Agarwal et al., 1998; Rao et al., 2000). Also, methylation patterns in the IFN-gamma promoter exhibit long-term faithful inheritance in T cells and their progeny. The demethylated IFN-gamma promoter is faithfully inherited following the withdrawal of T cell stimulation and the loss of detectable IFN-gamma mRNA (Fitzpatrick et al., 1999). There is evidence that this “DNA methylation memory” is also involved in maintaining the gene expression patterns associated with insulin resistance in type-2 diabetes. Prenatal glucose and insulin levels influence the risk of developing type-2 diabetes in later life, quite independent of the maternal type of diabetes and genetic predisposition (Dabelea et al., 2000). This suggests the presence of a “cellular memory” in insulin target tissues such as adipose tissue, skeletal muscle and liver. Several glucose-metabolism genes have already been shown to exhibit differential DNA methylation in their promoters; GLUT4, the major glucose transporter in adipose and muscle tissues, and the mitochondrial uncoupling protein UCP2 are good examples (Yokomori et al., 1999; Carretero et al., 1998). Furthermore, recent insights into the pathogenesis of transient neonatal diabetes or TND (a rare subtype of diabetes that is characterized by transient hyperglycemia in the neonatal period and a predisposition to diabetes in adult life), provide a link between methylation, gene dosage effects and diabetes. TND results when gene dosage is doubled for the chromosomal region 6q24. Paternal uniparental isodism, duplication of the 6q24 region and loss of methylation in this imprinted region, all result in phenotypically indistinguishable TND (Temple et al., 2000).

In addition to the orchestrated DNA methylation changes during cellular differentiation and in response to external stimuli, random DNA methylation changes have been observed during the aging of different organisms (Post et al., 1999; Tohgi et al., 1999; Bormann et al., 2001). These age-related methylation changes are involved in the aetiology of a number of different diseases such as atherosclerosis and cancer. Type-2 diabetes is strongly age-related; the incidence is increased in older populations and the metabolic situation of individuals deteriorates over time. Both of these phenomena may be explained by inappropriate DNA methylation accumulating with increasing age.

The human epigenome project

Andrew Feinberg recently proposed the idea of the “methylome” – the complete set of DNA methylation modifications of a cell (Feinberg, 2001). He suggests that knowledge of the methylome will extend our understanding of the genome, development, disease, and may also define polymorphic variation in populations (Feinberg, 2001). The recent completion of the human genome draft sequence (IHGSC, 2001), followed closely by a whole genome map of single nucleotide polymorphisms (The International SNP Map Working Group, 2001), established a solid foundation for the Human Epigenome Project (HEP). The Human Epigenome Consortium (HEC) is a rapidly growing international collaboration aiming to gain more insight into development, disease susceptibility and genome stability, by identifying global patterns of methylation in the human genome (Beck et al., 1999). As a pilot study, partners of this consortium (including the authors of this review) are mapping methylation sites across the human major histocompatibility complex (MHC). This region was chosen because it is associated with more diseases (including most if not all autoimmune diseases) than any other region in the human genome and has a complex and interesting biology.

The human major histocompatibility complex (MHC)

The MHC is located on chromosome 6 (6p21.3). This locus contains genes with a high diversity of function, many of which are involved in the innate and adaptive immune systems, and is one of the most gene-dense regions of the human genome (for review see Rhodes and Trowsdale, 1999). The MHC has been subdivided into 5 different classes based loosely on the functional characteristics of the genes. The complete sequence of a “virtual” MHC haplotype was published in 1999 and was found to span...
Figure 2. 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 are being investigated for MVPs in the human epigenome pilot project. (Ψ) indicates the presence of a pseudogene. The extended class I, telomeric of the classical class I, is not included in the pilot MHC project and is therefore excluded from this figure.
about 8 Mb (MHC sequencing consortium, 1999). For historical reasons, classes I, II and III formed the classical MHC but analysis of flanking regions revealed that class II could be extended towards the centromere and class I towards the telomere. These regions are now referred to as the extended classes I and II. Figure 2 is a gene map of the MHC indicating the order of the classes, and the genes therein, and the relative size of each class. For the sake of clarity in this review, the telomeric extended class I has been omitted from Figure 2 but a complete list of genes within this class can be found at http://www.sanger.ac.uk/HGP/Chr6. There are currently some 240 identified gene loci and 160 pseudogenes in the MHC. Over 40% of these loci were located as a direct result of the sequencing effort.

A number of genes within class I of the MHC are members of the immunoglobulin superfamily (e.g. HLA-A, -B and -C). They encode specialised cell-surface molecules (the so-called MHC molecules) that present short fragments of mainly intracellular antigenic peptides to the T-cell receptor of cytotoxic T cells (for review see Monaco, 1992). The response of the T cell leads to the killing of the presenting cell. Class I genes are ubiquitously expressed on the surface of most nucleated cells. Additionally the class I harbours a plethora of pseudogenes.

The class II genes HLA-DP, -DQ and -DR, are also members of the immunoglobulin superfamily. The cell-surface molecules that they encode are expressed only on specialised antigen-presenting cells such as macrophages, B cells and some T cells. These cells have the ability to engulf and internalise exogenous antigen and the class II molecules then present the resulting peptides to the specific receptor of a helper T cell. An adaptive immune response is then elicited (for review see Pieters, 1997). Classical class I and II genes are extremely polymorphic to allow for the vast specificity to a wide range of antigens. Other class II genes encode subunits of the proteasome (PSMB9 and PSMB8) – a macromolecular complex that degrades proteins within the cytoplasm and produces peptides for presentation by class I molecules (Groettrup et al., 1996). Also the ABC2 and 3 genes together encode a heterodimeric molecule which spans the membrane of the endoplasmic reticulum (ER) and transports antigenic peptides from the cytoplasm into the ER for peptide loading of free class I molecules (Kelly et al., 1992). The extended class II region also harbours some quite interesting genes. TAPBP (tapasin) facilitates the loading of peptide onto class I molecules (Ortmann et al., 1997) and the DAXX gene product acts as an effector in an apoptotic pathway (Chang et al., 1998).

The class III region of the MHC is very gene-dense and its members have quite disparate functions. Many are involved in innate immune responses and inflammation. The gene products of C2, C4 and BF form part of the complement cascade that instigates phagocytosis and lysis of bacterially infected cells and initiates an inflammatory response. The genes that control inflammation are clustered at the telomeric end and include TNF, LTA, LTB, LST1 and IC7.

It was recently shown that the B and T cell receptors, which specifically recognise antigens presented by MHC molecules, have monoallelic expression and the choice of allele is controlled by epigenetic mechanisms (reviewed by Rada and Ferguson-Smith, 2002). Although this finding does not affect the MHC directly, it does illustrate that epigenetics can influence the immune system.

Autoimmune diseases and the MHC

The majority of human autoimmune diseases have been associated with the MHC but for many the exact causes remain unknown. Autoimmunity involves the destruction of self-tissue and/or cells by the immune system and these diseases do not conform to simple mendelian genetics. They have very complex segregation patterns, cannot be attributed to single causal genes and are likely to involve a substantial environmental element(s). “Etiological mutations” or allelic variants of genes controlling the immune system, combined on a background of environmental factors such as age, race and sex, predispose individuals to particular autoimmune diseases (Vyse and Todd, 1996). Identifying the responsible alleles is extremely difficult because of the very small risks conferred by the individual variants; the allelic combinations impose the full risk. A “needle-in-a-haystack” approach has historically been taken to elucidate the causative genes of multifactorial diseases. Numerous genome linkage and association studies have consistently identified correlations between autoimmune disorders and the MHC (Vyse and Todd, 1996; Das et al., 2000). Predisposition (and sometimes resistance) to insulin-dependent diabetes mellitus (type 1 diabetes), rheumatoid arthritis, celiac disease and multiple sclerosis, have all been associated with various alleles of the class II genes HLA-DQB1 and HLA-DRB1 (Vyse and Todd, 1996; Das et al., 2000). Alleles of some class I genes (in particular HLA-Cw6 and HLA-B57) have been found to increase susceptibility to the inflammatory skin disease, psoriasis (Russell et al., 1972; Trembath et al., 1997; Jenisch et al., 1998). Ankylosing spondylitis also has an association with a class I allele (HLA-B27) and the class III complement genes C4A and C2, are reported to affect susceptibility to systemic lupus erythematosus (Das et al., 2000). It should be mentioned that the MHC genotype alone is not always sufficient to cause autoimmunity and a number of non-MHC genes have also been implicated in susceptibility (Wandstrat and Wakeland, 2001). The association of class I and II alleles is unsurprising, given that an autoimmune response involves the action of T-cells which in turn rely upon presentation of an antigen by MHC molecules. However, it remains unclear how the MHC molecules allow autoantigen to be presented and how the target organ specificity is determined. One theory is that negative selection of T-cells in the thymus becomes imprecise, perhaps due to structural changes in the MHC molecules, and autoreactive T-cells are allowed to flourish (Das et al., 2000). It is possible that there is an epigenetic component to autoimmunity; regulation of MHC gene expression may play a role in producing the susceptibility to autoimmune diseases. It would therefore be interesting and justified to examine the CpG methylation status of MHC loci.
The human epigenome pilot project

The MHC epigenome pilot project, established by members of the HEC, is the first of its kind to investigate the methylome on a large scale (Beck et al., 1999). The project aims to map methylation variable positions (MVPs) (Beck, 2001) across the MHC in both the normal and the disease situation. Some 130 MHC loci (excluding the extended class I loci) have been selected for this study. These loci have a high CpG content and are therefore likely to contain MVPs. Figure 2 shows the selected loci as solid black boxes. Two or three PCR amplicons will be generated per MHC loci, with priority given to promoter regions and early exons. Amplicons are made from bisulphite-treated genomic DNA and the PCR products will be directly sequenced (Frommer et al., 1992). A number of different healthy tissues, cell lines and individuals are included in the study and their methylation status will be compared with the MHC reference sequence (MHC Sequencing Consortium, 1999). Psoriasis, an MHC-associated skin disorder (Russell et al., 1972; Trembath et al., 1997; Jenisch et al., 1998), has been chosen as the model disease in this pilot project and possible epigenetic components will be investigated.

Knowledge of the methylation signals across the MHC will add to the established data regarding the genomics, biology, and disease association of this locus; indeed it will be one of the most thoroughly investigated regions of the human genome. Hitherto unknown correlations between genomics and epigenomics may be uncovered, and it should highlight the functional interrelationships of genes in health and disease.

Outlook and summary

For a cell and for an organism, control is everything. The genome and its set of genes, like a deck of cards, is identical for all cells in a body. It is played in a different way each time giving rise to the high variability in protein expression patterns that are tightly controlled by “rules”. Since the human genome project has shown us the deck of cards, we race on to figure out the details of these “rules” and epigenetics will be one of the keys to understanding them. DNA methylation is the prototypical epigenetic parameter and its general properties are so well suited for regulating genes that it is no surprise that evolution selected it for this purpose. DNA methylation affects gene transcription, is heritable across cell division and there are enzymes to actively maintain or change it depending on environmental influences. These are the prerequisites to making cellular differentiation work and to lock cells into different differentiated states. In this way, DNA methylation represents an entire level of cellular information on top of the DNA sequence and provides a link between genotype and phenotype. Most importantly, DNA methylation is technically accessible and lends itself to analysis on a genomic scale. Sequencing and microarray technologies allow us to create an epigenomic map or the DNA methylation blueprint of the epigenome. Changes in DNA methylation patterns can therefore be monitored and correlated with the silencing of genes and pathways, imprinting of genes, environmental influences, effects of health and nutrition, age-related changes, and disease-specific alterations.

The human epigenome project has been proposed as a venture of a huge scale. To get the complete picture, it will be necessary to read the methylation state from the regulatory regions of all genes, in all major cell types, and in their diseased variants. This project is realistic using modern genomic throughput technologies. The Human Epigenome Consortium (HEC) has initiated this process with a targeted approach on the MHC region, one of the most gene-rich regions of the human genome. The results of this project will create a map that complements the results of other large-scale efforts that are linking our knowledge of gene sequence and cellular phenotypes: the DNA sequence, its variations (SNPs), the mRNA expression patterns, and proteomic analyses. The epigenome map – the DNA methylation blueprint – will lay the ground for understanding gene regulation and the intricate interactions between different genes in normal and disease situations. Not only will it offer new explanations for well-studied areas like cancer research, but it will also provide a basis for novel approaches to research on environmental effects, nutrition, and aging. Furthermore, combined with other genomic technologies, the benefits for pharmaceutical research will be immediate; DNA methylation screening promises new epigenetic markers for disease states and new targets for drug development. Already, some medical applications based on DNA methylation research are emerging in diagnostics (Cairns et al., 2001). Some companies are even trying to directly target the epigenomic machinery of cells with drugs, for the treatment of cancer; methyltransferase inhibitors (Besterman and McLeod, 2000; Lubbert, 2000) and deacetylase inhibitors (Munster et al., 2001) are entering clinical trials. These first attempts are using drugs with global effects on cells but more refined approaches will become possible with accumulating knowledge in the new field of epigenomics.

Acknowledgements

This work was supported in part by a grant from the European Union (QLRT-1999-30417). KLN and SB are also supported by the Wellcome Trust.

References


Giltan, R.S., Shi, H., Chen, C-M., Yan, P.S., and Huang, T.


Vyse, T., and Todd, J.A. 1996. Genetic analysis of
Further Reading

Caister Academic Press is a leading academic publisher of advanced texts in microbiology, molecular biology and medical research. Full details of all our publications at caister.com

- MALDI-TOF Mass Spectrometry in Microbiology
  Edited by: M Kostrzewa, S Schubert (2016)
  www.caister.com/malditof

- Aspergillus and Penicillium in the Post-genomic Era
  Edited by: RP Vries, IB Gelber, MR Andersen (2016)
  www.caister.com/aspergillus2

- The Bacteriocins: Current Knowledge and Future Prospects
  Edited by: RL Dorfl, SM Roy, MA Riley (2016)
  www.caister.com/bacteriocins

- Omics in Plant Disease Resistance
  Edited by: V Bhadauria (2016)
  www.caister.com/opdr

- Acidophiles: Life in Extremely Acidic Environments
  Edited by: R Quatrini, DB Johnson (2016)
  www.caister.com/acidophiles

- Climate Change and Microbial Ecology: Current Research and Future Trends
  Edited by: J Marxsen (2016)
  www.caister.com/climate

- Biofilms in Bioremediation: Current Research and Emerging Technologies
  Edited by: G Lear (2016)
  www.caister.com/biorem

- Microalgae: Current Research and Applications
  Edited by: MN Tsialoglou (2016)
  www.caister.com/microalgae

- Gas Plasma Sterilization in Microbiology: Theory, Applications, Pitfalls and New Perspectives
  Edited by: H Shintani, A Sakudo (2016)
  www.caister.com/gasplasma

- Virus Evolution: Current Research and Future Directions
  Edited by: SC Weaver, M Denison, M Roossinck, et al. (2016)
  www.caister.com/virusevol

- Arboviruses: Molecular Biology, Evolution and Control
  Edited by: N Vasilakis, DJ Gubler (2016)
  www.caister.com/arbo

- Shigella: Molecular and Cellular Biology
  Edited by: WD Picking, WL Picking (2016)
  www.caister.com/shigella

- Aquatic Biofilms: Ecology, Water Quality and Wastewater Treatment
  Edited by: AM Romani, H Guasch, MD Balaguer (2016)
  www.caister.com/aquaticbiofilms

- Alphaviruses: Current Biology
  Edited by: S Mahalingam, L Herrero, B Herring (2016)
  www.caister.com/alpha

- Thermophilic Microorganisms
  Edited by: F Li (2015)
  www.caister.com/thermophile

- Flow Cytometry in Microbiology: Technology and Applications
  Edited by: MG Wilkinson (2015)
  www.caister.com/flow

- Probiotics and Prebiotics: Current Research and Future Trends
  Edited by: K Venema, AP Carmo (2015)
  www.caister.com/probiotics

- Epigenetics: Current Research and Emerging Trends
  Edited by: BP Chadwick (2015)
  www.caister.com/epigenetics2015

- Corynebacterium glutamicum: From Systems Biology to Biotechnological Applications
  Edited by: A Burkovski (2015)
  www.caister.com/cory2

- Advanced Vaccine Research Methods for the Decade of Vaccines
  Edited by: F Bagnoli, R Rappuoli (2015)
  www.caister.com/vaccines

- Antifungals: From Genomics to Resistance and the Development of Novel Agents
  Edited by: AT Coste, P Vandeputte (2015)
  www.caister.com/antifungals

- Bacteria-Plant Interactions: Advanced Research and Future Trends
  www.caister.com/bacteria-plant

- Aeromonas
  Edited by: J Graf (2015)
  www.caister.com/aeromonas

- Antibiotics: Current Innovations and Future Trends
  Edited by: S Sánchez, AL Demain (2015)
  www.caister.com/antibiotics

- Leishmania: Current Biology and Control
  Edited by: S Adak, R Datta (2015)
  www.caister.com/leish2

- Acanthamoeba: Biology and Pathogenesis (2nd edition)
  Author: NA Khan (2015)
  www.caister.com/acanthamoeba2

- Microarrays: Current Technology, Innovations and Applications
  Edited by: Z He (2014)
  www.caister.com/microarrays2

- Metagenomics of the Microbial Nitrogen Cycle: Theory, Methods and Applications
  Edited by: D Marco (2014)
  www.caister.com/h2

Order from caister.com/order