The Epigenome

Molecular Hide and Seek

Edited by
S. Beck and A. Olek

WILEY-VCH GmbH & Co. KGaA

The Epigenome
Molecular Hide and Seek

Edited by
S. Beck and A. Olek

Further Titles of Interest

E. Engel, S. E. Antonarakis

Genomic Imprinting and Uniparental Disomy in Medicine

Clinical and Molecular Aspects

2001
ISBN 0-471-35126-1

S. Lorkowski, P. Cullen

Analysing Gene Expression

A Handbook of Methods. Possibilities and Pitfalls

2002
ISBN 3-527-30488-6

J. Licinio, M. L. Wong

Pharmacogenomics

The Search for Individualized Therapies

2002
ISBN 3-527-30380-4

C. W. Sensen

Essentials of Genomics and Bioinformatics

2002
ISBN 3-527-30541-6

T. Lengauer, R. Mannhold,
H. Kubinyi, G. Folkers

Bioinformatics

From Genomes to Drugs

2001
ISBN 3-527-29988-2

The Epigenome

Molecular Hide and Seek

Edited by
S. Beck and A. Olek

WILEY-VCH GmbH & Co. KGaA

Editors:

Dr. Stephan Beck
The Sanger Centre
Wellcome Trust Genome Campus
Hinxton, Cambridge CB 10 1SA
UK

Dr. Alexander Olek
Epigenomics AG
Kastanienallee 24
10435 Berlin
Germany

Library of Congress Card No.: applied for
A catalogue record for this book is available from the British Library.

Bibliographic information published by Die Deutsche Bibliothek
Die Deutsche Bibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data is available in the Internet at <http://dnb.ddb.de>.

Printed in the Federal Republic of Germany.
Printed on acid-free paper.

Composition ProSatz Unger, Weinheim
Printing betz-druck GmbH, Darmstadt
Bookbinding Großbuchbinderei J. Schäffer GmbH & Co. KG, Grünstadt

ISBN 3-527-30494-0

Preface

The Human Genome Project (HGP) delivered the location and sequence of all genes – the Human Epigenome Project (HEP) will help to uncover their temporal and spatial expression. By systematically analyzing the Epigenome for (cytosine) methylation patterns, epigenetic markers will be discovered that have the power to distinguish active from non-active genes and healthy from non-healthy tissues. This new knowledge will provide the missing link between genetics, disease and the environment and will be invaluable for the understanding of how our genome functions and for future health care.

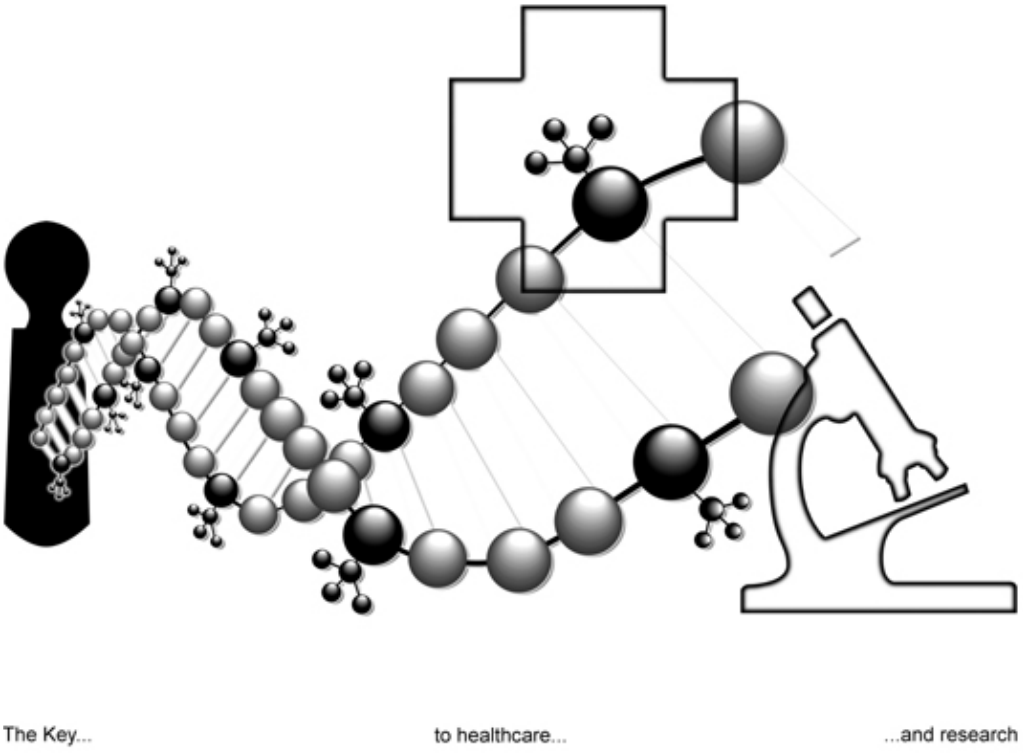
Written in popular science style, this book is aimed at anyone interested in the science arising from the human genome project and its implications for future biomedical research. In nine chapters, world-leading scientists discuss the history, current facts and future potential of this most exciting area in genome science at the interplay of nature and nurture. The topics covered stretch from the discovery of epigenetic signals to their use in molecular diagnostics and therapy but also explore their effects or dependency on development, gene regulation, disease, diet and aging. In addition to the described science, the central messages of all chapters have been captured in artistic illustrations that may well serve as future icons for these fast evolving fields of research.

Following the announcement of the human genome draft sequence in June 2000, a "spoof" editorial in a leading science magazine *(Nature Biotechnology)* suggested that a trilogy be required to complete this historic project: [I] "The Draft", [II] "The Closure" and [III] "The Epigenome Strikes Back". With the "Draft" delivered, the "Closure" on track for Spring 2003, the time has come to explore the next frontier, the "Epigenome".

Stephan Beck & Alexander Olek
January 2003

Contents

1
Five Not Four:
History and Significance of the Fifth Base



1
Five Not Four: History and Significance of the Fifth Base

Douglas S Millar, Robin Holliday, and Geoffrey W Grigg

Summary

The past 30 years have seen a tremendous surge of interest in the field of epigenomics. This work has led to exciting discoveries: new enzymes that can methylate mammalian DNA in both a maintenance and a de-novo manner, the bisulfite modification technique for the direct detection of methyl cytosine residues in genomic DNA, the role of abnormal methylation in disease and the aging process, and, more recently, the link between chromatin remodeling and methylation in the transcriptional repression of gene function. Over the next 30 years, epigenomics will be one of the boom fields of biotechnology with pharmaceutical companies waking up to the fact that the presence or absence of the 5th base has major implications in human health care and the treatment of an ever-increasing aged population.

One of the first major hurdles is the sequencing of the "methylome" or epigenome, which is unlikely to be achieved in the same manner as the sequencing of the human genome. This is because each individual tissue type in the human body has a unique methylation signature, and at least 180 cell types make up the human body. With advances in robotics and automation, however, this task does not seem quite as impossible as it would have been before the race to sequence the human genome began.

Sequencing the human epigenome aside, many questions have at best been only touched upon and many more are still unanswered. How are basic developmental pathways involving methylation established and later modified in a coordinated fashion? Can methylation directly cause gene inactivation or is the process a secondary phenomenon to stably repress gene function? How accurately is the original methylation pattern repaired in a region of DNA damaged by radiation, chemical reactions, or other environmental stimuli? In the next 30 years can the methylation machinery or methylation fingerprints be altered to therapeutically reverse the disease process or indeed the natural process of aging? These questions and many more need to be fully answered before we have true insight into the full impact that epigenomics has on basic human development and disease.

1.1 Historical Introduction

The modified base 5-methyl cytosine (5-mC) was first detected in DNA by Hotchkiss [1]. In subsequent years, many other modified bases were discovered in bacteria, and it became clear that they were an integral part of the modification-restriction systems that are widely distributed in prokaryotes. It was shown that restriction enzymes could distinguish between modified and unmodified DNA substrates.

In plants, approximately 80% of cytosines at CpG doublets are methylated; however, methylation also frequently occurs in CpNpG triplets [2]. Recent studies have also found that methylation of cytosines can occur in non-symmetrical sites such as CpTpA [3]. Despite the overall high global levels of 5-mC, plants contain regions of unmethylated DNA that resemble the *HpaII* tiny fragments characteristic of mammalian genomes. These areas are typically unmethylated in a wide variety of tissues, whether or not the associated gene is expressed [4].

In mammalian DNA, the major modified base is 5-mC, at a level of 2%–5% of all cytosines. Small amounts of other modified bases should not be ruled out, and there is one reliable report that 6-methyl adenine exists in mouse DNA [5]. 5-mC occurs predominantly, but not exclusively, in CpG doublets [6]. In the total mammalian genome, the CpG doublet is very significantly less than what would be expected from the overall base composition of the DNA. However, there are CpG islands associated with structural genes, where such depletion is not observed.

Riggs [7] and Holliday and Pugh [8] proposed that 5-mC in mammalian DNA might have an important role in the regulation of gene expression. They argued that, since restriction enzymes can distinguish between modified and unmodified sites, there might also be regulatory proteins that could make the same distinction at specific DNA sequences in promoter regions. Instead of cutting the DNA, they would recognize the specific methylation signal and would thereby control the presence or absence of transcription at an adjacent gene. These authors also proposed that the pattern of DNA methylation could be inherited, if there was a maintenance DNA methyltransferase that recognized hemimethylated DNA just after replication and methylated the new strand. The same enzyme would not act on nonmethylated CpG doublets. This provided a basis for the epigenetic inheritance of a given pattern of DNA methylation and therefore also of the specific controls of gene expression in given cell types. It could also account for those cases, such as X-chromosome inactivation in female eutherian mammals, where only one of two homologous genes in a diploid cell is active, while the other is inactive. It was already known that these states of activity are very stable in dividing and nondividing cells. In 1975, when these papers appeared, there was no direct evidence for the hypotheses proposed, and the authors therefore did not suggest that methylation, or lack of methylation, was associated with gene activity. However, it later became clear that methylation is associated with the inactivity of genes in almost all cases. For example, in the inactive X chromosome, CpG islands adjacent to genes are methylated, and these same islands are unmethylated on the active X chromosome. It also became clear that the pattern of DNA methylation could be stably inherited, so DNA methyl-maintenance

enzymes must exist in cells. However, DNA methyltransferases that were studied in vitro had some activity on nonmethylated substrates in vitro.

In most early experiments, methyl-sensitive restriction enzymes were used, together with an isoschizomer that recognized the same short sequence and cut it whether or not it was methylated. For example, the enzyme HpaII cuts the substrate CCGG, but not if the central CpG is methylated, whereas MspI cuts whether or not methylation is present. It is evident that in all such studies, only a proportion of all 5-mC bases are detected. In the example just cited this is less than 10% of all CpGs. Many published studies in which this method was used have concluded that lack of gene expression in the system studied had nothing to do with DNA methylation. However, it should have been obvious that a full sequence analysis of all methylated and nonmethylated cytosines was necessary before such a conclusion could be reached.

1.2 Sequencing 5-methylcytosine (5-mC) Residues in Genomic DNA

Historically, the first of the chemical methods for sequencing 5-mC arose as a by-product of Maxam and Gilbert's method for sequencing DNA [9]. This method relies on specific but partial cleavage by various reagents at particular base sequences. Hydrazine, one of the reagents used, reacts with cytosine but not with 5-mC, thus allowing differentiation between 5-mC and C sites. Its main drawback was that it required the use of purified radiolabeled restriction fragments of the region of interest.

This protocol was improved by adding to the chemically cleaved fragments an indirect end-labeling procedure [10]. This method, however, still required a relatively large amount of genomic DNA and was somewhat complicated to perform.

The Saluz and Jost procedure [11], which introduced a PCR amplification step and simplified some of the procedures, was a further significant improvement, but it too required 25–50 μg of genomic DNA to produce a signal from a single copy of a eukaryotic gene. This drawback was answered by the ligation-mediated polymerase chain reaction [12]. All these methods, however, recognize an individual cytosine residue by a negative observation – a gap in the sequencing ladder – which is possible to miss when reading the sequence.

The Bisulfite Method

The need for positive and sensitive detection of methylated cytosine residues in genomic DNA was addressed by the development of the bisulfite genomic sequencing protocol. This powerful method depended on the reaction of bisulfite with cytosines in single-stranded DNA which, in an aqueous milieu, are converted to uracil; whereas 5-methylcytosines (5-mC) are unreactive [13]. The modified DNA strands, which now had very few C residues, could be amplified by use of PCR and then either sequenced or, if single half-strand data were required, cloned before sequencing. Direct sequencing without a prior cloning step allowed the average methylation status of individual CpG sites to be obtained, whereas sequencing cloned DNA allowed the analysis of CpG sites on individual half-strand DNA molecules.

The bisulfite method was invented in 1988–1989 [14] and developed into a practical method by Frommer and associates [15]. Several significant improvements to the original method were described subsequently, which greatly improved the sensitivity and removed aberrations caused by secondary structure formation in single-stranded DNA and also demonstrated that the method was applicable even with very small numbers of cells in developmental studies [16–20]. Another study demonstrated that bias may be introduced by the amplification reaction, depending on the sequence of the DNA region of interest [21]; thus, precautions should be taken to prevent false-positive and -negative results.

The principal advantages of the bisulfite method over others are:

- It can provide good DNA sequence information from only a few cells.
- It is a completely general method and provides a positive readout for 5-mCs in the sequencing gel. Since all the C residues are missing from the template but all the mC residues remain, the positions of the latter appear as distinct bands, making them easy to score.
- Single DNA half-strand data on the 5-mC distribution can be determined readily if needed; thus, information about possible hemi-methylated sites or the methylation status of individual alleles useful in analyzing silencing of genes involved in imprinting, normal embryonic development, and cancer can be determined.

The bisulfite genomic sequencing protocol and modifications thereof have revolutionized the field of methylation analysis, yet confusion still comes from the assumption that several authors have made that *all* 5-mC bases are in some way related to gene expression. From the outset, however, this seemed unlikely, and it seemed probable that only a small subset might have a direct regulatory role, with the rest having some other function. One such function appears to be the inactivation of foreign DNA, such as mobile genetic elements, which become inserted into the genome. It is also likely that genes in transgenic animals that become inactivated by a so-called "position effect", are in fact inactivated by DNA methylation. Another function of methylation may be to suppress recombination between repeated genetic sequences in the genome [22]. Without such suppression, many chromosome abnormalities would be generated by crossing-over at meiosis and also during mitosis.

Many authors have argued that DNA methylation may not be important in development, because it was believed that some complex eukaryotes, such as *Drosophila*, have no 5-mC in their DNA. Recently, this belief was shown to be mistaken, because *Drosophila* DNA does contain 5-mC, particularly in the larval stage of development [23, 24]. Indeed, this may strengthen the case for thinking that only a minor fraction of all methylation in mammalian DNA has an important role in development.

1.3 Gene Silencing

It is frequently stated in the literature that DNA methylation is correlated with the inactivation of genes and is sometimes stated that the role of methylation is to "lock in" an inactive state, as a consequence of the silencing of a gene by some other mechanism. However, whatever the mechanism, this is an epigenetic change, spontaneous or induced, that has a specific phenotypic effect; so the term "epimutation" was coined to explain it. When mutations were first studied by geneticists, they had no knowledge of the chemical nature of the gene, let alone how genes produced a particular phenotype. With appropriate genetic analysis, they could deduce that a gene mutation caused a specific phenotype; it was not just correlated with that phenotype. The same applies to methylation changes and gene inactivation. Where a proper analysis can be done, it becomes clear that the methylation of a particular gene causes a particular phenotype, and demethylating agents reactivate that gene. Such analysis has been done in cultured mammalian cells [25], in control elements in maize [26], and in the fungus *Ascobolus* [27]. A study of the structural gene for adenine phospho ribosyl transferase (APRT) in CHO cells demonstrated that two types of inheritance exist: one is due to standard mutations, and the other is due to epimutations [28].

One method of silencing genes by epimutations is to use 5-methyl dCTP after permeabilization of the cells by electroporation [29, 30]. In the experiments with APRT, the epimutations were analyzed by the bisulfite procedure of genomic sequencing [28]. Usually, a single epimutation resulted in extensive methylation of the 16 CpG doublets in the region that was sequenced (within part of the CpG island spanning the promoter region). It is extremely unlikely that such extensive methylation would occur following 5-methyl dCTP treatment. What seems much more plausible is that a low initial level of methylation may be a sufficient trigger for further extensive methylation in the same region of DNA. A similar two-step mechanism may also occur in tumor-suppressor genes (see Sect. 1.5). Further analysis is needed to determine whether the initial epimutation is sufficient to produce the observed phenotype or whether the process of secondary methylation is also necessary. Much more information is also needed in the examples of genes without associated CpG islands. It is possible that a low level of methylation in the promoter region is sufficient to prevent transcription. Furthermore, there may be critical sites where 5-mC blocks gene expression [31, 32]. Some transcription factors are methyl sensitive, which strongly suggests that such sites exist.

It has often been stated that the CpG methylation observed in the silenced genes on the inactive X chromosome is a secondary phenomenon that stably "locks in" the inactive state of the chromosome. It is nevertheless possible that the initial epigenetic switch to inactivation is due to a low level of methylation, perhaps spreading from the inactivation center. What is not in doubt is the effect of another epimutagen, 5-azacytidine [33], which has been shown to reactivate silent genes in a variety of contexts. The evidence suggests that the analog is incorporated into DNA (in its deoxy form), and that the maintenance methyltransferase binds covalently to it [34].