CpG Islands

This text is mostly redundant with the outline that follows. The outline was made for my own purposes. I know that outlines can sometimes be difficult to follow and so I wrote this text.

CpG refers to two adjacent nucleotides that appear in a row on the same strand of DNA. Do not be confused about the "p", which refers to the phosphodiester bond between them. Cytosine can be methylated at the 5 position. When I say that the CpG is methylated, what I really mean is that the cytosine part of the CpG is methylated. The #5 position is within the cytosine nitrogenous base. Notice that a 5' CG 3' on one strand will have a 3' GC 5' on the other. That is, CpG is a palindrome. The palindromic nature of the CpG means that after replication of methylated DNA, one strand will be have a methyl CpG and other will have an unmethylated CpG. If only one strand's CpG is methylated, then it is said to be hemimethylated. This hemimethylated state is recognized by the methylating enzymes and converted to the completely methylated state. Thus, this CpG methylation can be "remembered" across cycles of DNA replication.

DNA methyltransferases are the class of enzymes that methylate CpG. Notice that I am very careful to specify CpG. This is because in a different context (think CpA, CpC, CpT), cytosine is not methylated [1].

5'-methyl-cytosine is sometimes called the "fifth nucleotide" because the cell treats it so differently. It is bound by different proteins and it has a different meaning to the cell than just a cytosine-containing nucleotide. Of course, it is not really a fifth deoxynucleotide.

CpGs are highly under represented in the vertebrate genome [2]. This is because, most CpG's are methylated. 5-methyl cytosine has the nasty property of deaminating to thymine at a relatively high rate. This means that CpG often turn into TpG, which represents the loss of CpG. WITH THE EXCEPTION OF CpG islands, the frequency of CpG is only 20% of the predicted value [3]. One usually sees that satellite DNA, transposons, exons, and turned off CpG islands are methylated.

CpG islands are runs of CpG. The CpG island is the place that UNMETHYLATED CpGs are usually found in vertebrates. These CpG islands are actually transcriptional promoters that can have enhancer elements interdigitated between some of the CpGs. In vertebrates, this is the most common type of transcriptional promoter. About 70% of known promoters are CpG islands [2].

OK, so what is going on here? Basically, the cell seems to methylate NON-PROMOTER CpGs to prevent them from being used as promoters. Another reason for all of this methylation is to keep our massive transposon load (40-50%) from becoming active. The methylation certainly helps to keep transposons turned off but it also causes them to mutate (C-->T) and some people believe that this helps to "kill" transposons [4].

Traditionally, CpG island methylation has been thought to always be involved in turning off promoters. However, this may prove to merely be the first aspect of the CpG methylation to be understood. Recent studies have shown that CpG methylation correlates with the activation of some genes [5]. Furthermore, an important but not terribly common use of CpG methylation is in imprinting.

But how does one detect 5'methyl CpG? The most common way these days is bisulfite sequencing. The experimenter uses sodium bisulfite to convert all cytosines in the DNA to uracils. However, 5' methyl C does not convert. Then the converted DNA and a sample of unconverted DNA is sequenced and compared. Every time that you see a C in the converted DNA, you know that it was really a 5' methyl C in the cell.

I) **CpG** The fifth nucleotide

CpG refers to the dinucleotide in which the cytosine-containing nucleotide is joined to a guanine-containing nucleotide. The 'p' in the middle stands for the phosphate. The other strand of the DNA will have a GpC because of Watson-Crick base pairing. **CpG dinucleotides are underrepresented in the vertebrate genome [2].**

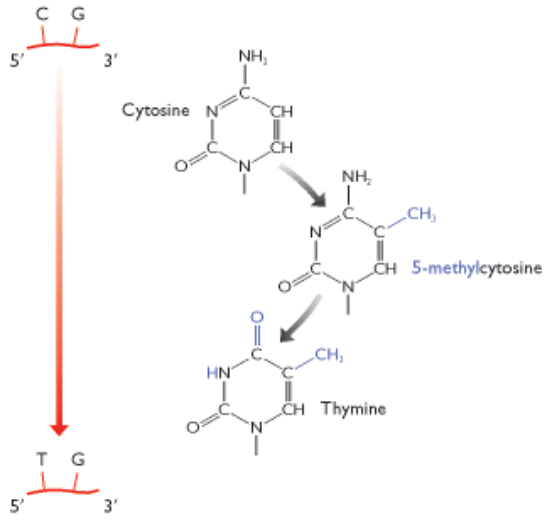
A) **In mammals in most CpGs the cytosine is methylated (5'-methyl-cytosine).**

B) **In vertebrates, DNA methylation of cytosine occurs when it appears in CpG not when it is alone.**

C) **The type of enzymes that methylate CpG are called DNA methyltransferases (DNMTs).** [6]

1) **DNMT1, DNMT3A and DNMT3B**
2) **DNMT3A/3B required for de novo methylation**
3) **DNMT1 is the maintenance methyl transferase**
    a) prefers hemimethylated DNA
    b) found at replication fork where it interacts with PCNA and UHRF1 (repair protein)

D) **The cost** - Increases the incidence of mutation
5-methyl-cytosine can lead to hard to identify mutations. Cytosine can spontaneously deaminate to uracil which is easy to identify as a mistake but 5-methyl cytosine deaminates to thymine which is hard to identify as a mistake.



From Genomes 2 2nd edition TA Brown

E) **The consequence** - Because vertebrates methylate most of their CpGs, these nucleotides have increased rates of mutation (C-->T). Over time the

increased rate of mutation repletes CpGs from the genomes. They are changed to TpGs. Vertebrates are CpG deficient because of the mutagenic quality of 5-meC. Outside of the CpG island, the frequency of CpG is only 20% of the predicted value [3].

II) CpG islands

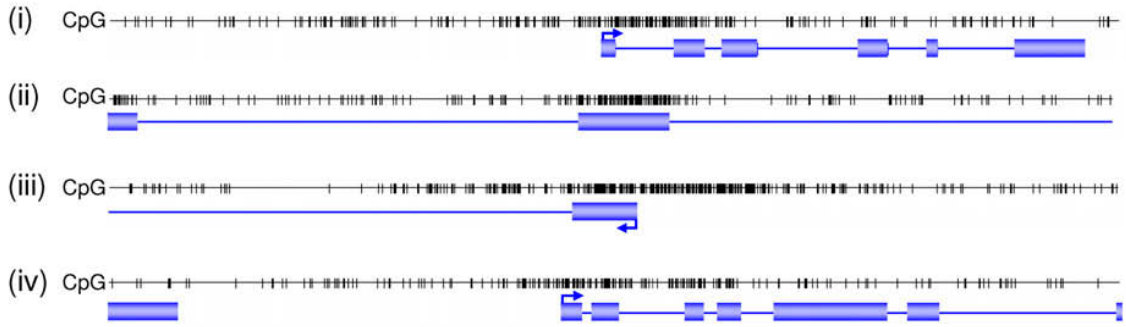
A) CpG islands are core promoter elements in mammals.

CpGs islands are usually not methylated.

They are often a true enhancer/core promoter mix. Enhancer elements can be interdigitated between CpGs or even include CpGs.

Methylation can hide both the core promoter qualities and the enhancer qualities of the CpG island. Some probably function as pure core promoters or pure enhancers.
This means that methylation of a CpG island usually reflects a long-term decision to not use that promoter.

CpG islands contain multiple GC boxes that can be bound by SP1 and other CG, GC binding proteins.
~50% are acting as core promoter elements (current guesstimate)[7,8]
+1 is usually within the island.



Illingworth, R. S., and Bird, A. P. (2009). CpG islands--'a rough guide'. FEBS Lett *583(11)*, 1713-1720.

Some (maybe many) are bidirectional.

B) Statistics

1) What is a CpG island?
    a) GC rich CpG rich non methylated [2]
2) How is it defined?
    a) 200-3,000 bp
    b) and greater than 60% CpG

c) There are about 45,000 CpG islands. [9]

3) Overall CpG is statistically under represented in mammalian genome[10]

4) About 2% of the total genome are CpG islands[6]

5) 40%-70% of CpGs are near or are mammalian promoters

The expected frequency of CpG can be calculated based on GC nucleotide composition of the genome. CpG appears at the expected frequency only in CpG islands. Elsewhere the frequency of CpG is only 20% of the predicted value.[3]

6) Methylation statistics

a) 60-90% of CpG dinucleotides are methylated in the mammalian genome

b) Most CpGs near active promoters are NOT methylated.

c) CpG that are NOT associated with an Island are almost always methylated.

7) About 70% of annotated gene promoters are associated with a CpG Island.

III) Which CpGs are usually methylated?

CpG's are usually methylated when found in exons, transposable elements and satellite DNA. Methylated CpG's tend to persist because the hemimethylated version is recognized by DNMT1 (maintenance methyl transferase) that then methylates the CpG on the other strand.

A) **About 60-90% of CpG's are methylated in mammals**

B) **Satellite DNAs, transposons, other repetitive DNA (probably dead transposons) and intergenic DNA**

C) **EXONS!**

D) **CpG islands in imprinted genes**

IV) Function of CpG methylation

A) **The ABSENCE of methylation of a CpG Islands indicates that it is a promoter. That is, functional CpG islands have low methylation.**

B) **Methylation of CpG islands is often a long-lasting decision to keep that promoter off.**
USUALLY CELLS DIFFER LITTLE IN CpG PATTERN. JUST BECAUSE A CpG ISLAND IS NOT METHYLATED DOES NOT MEAN THAT THE GENE WILL BE EXPRESSED.

1) **Idea that C-methylation of CpG could be used as a type of molecular memory dates to the mid-70's.** How would this work?
CpG are self-complementary and so after replication each daughter double

helix will be hemi-methylated. It was proposed that there was enzyme that recognizes hemi-methylated DNA and then methylates it.

2) CpG methylation is NOT an on off switch that is easily flicked.

It is usually a way to identify which bits should be used transcriptionally and which bits should never be used. As usual, exceptions will exist.

C) **But, in general, it is not usually the trigger for silencing a gene.** Instead CpG methylation is a way to help make a silencing decision stable (more permanent).

D) **Methylation of intragenic regions is thought to inhibit cryptic transcription.** Sometimes, DNA in coding regions just can't help looking like a transcription start site. A mark like this can be used to clear this up this possible misunderstanding.

E) **Defense mechanism to silence non-coding DNA** much of which is the product of invasion by transposons - defense against DNA of foreign origin.

This is a big deal. 40-50% of mammalian genome are transposable elements.[4]

Amongst different species, the amount of CpG methylation is directly proportional to the amount of non-coding DNA (transposons) that they have.

F) **Methylation of repetitive sequences and mobile elements to promote genome stability**
Instability refers to inappropriate recombination, the inability to identify or use centromeres or telomeres, and the activation of transposable elements.

V) **Gross methylation in mammals in done early in embryogenesis**
Seems to occur only in totipotent cells as they differentiate.

1) Development

A few hours after fertilization the paternal genome is de-methylated. This is an active process. Enzyme is unknown. Later in embryogenesis maternal genome is passively de-methylated (failure to methylate hemimethylated target). Most of the genome but not all is de-methylated.

The enzymes Dnmt3a and Dnmt3b re-methylate the genome after embryo implantation.

2) X chromosome inactivation
CpG islands methylated on on X chromosome in female mammals during embryogenesis

VI) **Location of CpG Islands**

There exist 5' CpG Islands, 3'CpG Islands, intragenic CpG Islands,and intergenic CpG Islands. [7]

About half over lap the transcription start site. [7,8]

Coding exons at 5' ends of genes often have CpG island.

VII) **What role do CpG's play in promoter recognition?**

A) **The CpG island is a core promoter element in mammals.**

They contain multiple GC boxes that can be bound by SP1.
~50% are acting as core promoter elements (current guesstimate) [7,8]
+1 is usually within the island.

Illingworth, R. S., and Bird, A. P. (2009). CpG islands--'a rough guide'. FEBS Lett *583(11)*, 1713-1720.

VIII) **What prevents the CpG's in front of promoters from being methylated?**
Ideas - there is not a single agreed upon mechanism. Remember, right now we are talking about preventing methylation just during early embryogenesis. After that the methylation pattern will be maintained.

Choose one or more from the list.

A) There is evidence that proteins bind them and deny access to the Dnmt's.

B) The CpG island has a shape or acts as an origin of DNA replication and this prevents methylation.

C) Nucleotide turnover at these sites is bizarrely high during early embryogenesis.

D) Early transcription from a CpG promoter prevents it.
This is likely to be used.

E) CpG Methylation is determined by histone methylation pattern.

Now we have to wonder how the correct histones are modified.

There is substantial evidence that the H3 methylation (meH3K9) can cause methylation of CpGs.

F) Antisense transcription to recruit Dnmt's

IX) **5meCpG is most often described as repressing gene expression**

A) Methyl group is in the major groove.
The methyl group of a CpG is positioned in the major groove of the DNA. Most transcription factors make heavy use of the major groove to read the bases. A methyl group here can make the enhancer unrecognizable. Determined using cell extracts and gel shifts.

B) MeCpG's recruit repressors

Specific proteins bind methylated CpGs.
These are called methyl CpG-binding proteins.

Examples: MBD1, MBD2 and MeCP2

MeCP2 is very popular. The binding domain is called methyl CpG binding domain (MBD). Quit a few different MeCP's exist.

MeCP2 reads the methylation state and recruits a other co-repressors

C) **CpG methylation and repressive histone modifications have an incestuous relationship** Reference: the DNA Methylation book by Doefler.

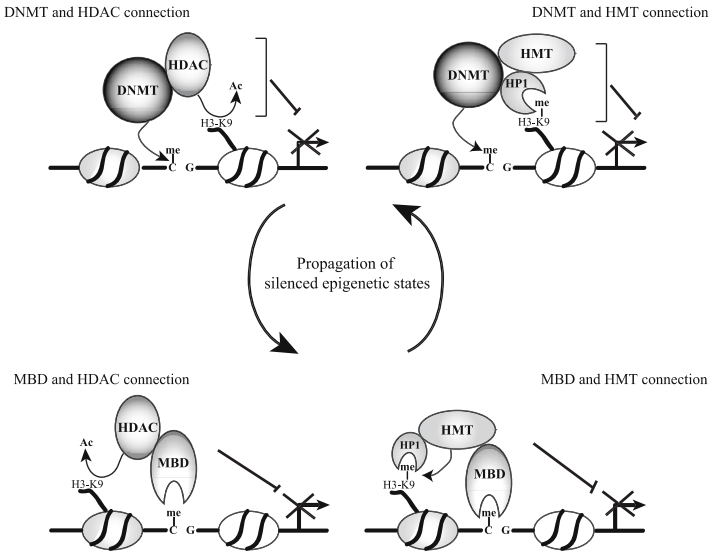


Fig. 2 DNA methylation and chromatin modifications interact intimately to bring about transcriptional silencing. In a first phase, the association of DNMTs with HDACs leads to histone deacetylation and, in some instances at least, to CpG methylation. This would lead to chromatin compaction and transcriptional silencing. Association of DNMTs with H3-K9 histone methyltransferase (HMT) and the HP1 adaptor protein would lead to a direct impact of the H3-K9 methylation state on the DNMTs. In a second phase, methylation of CpGs by DNMTs would allow binding of methyl-CpG binding domain proteins (MBD) to the DNA. MBD would in turn associate with HDAC and the H3-K9/HP1 system and favor histone deacetylation and H3-K9 methylation, respectively. This sequential process coupling DNA methylation with histone deacetylation and H3-K9 methylation may create a self-perpetuating epigenetic cycle for the maintenance of transcriptional repression. Ac, acetyl group; me, methylated group; H3-K9, Lys 9 of histone H3

X) **CpG methylation may ALSO be a common activator**

A) **In unbiased studies CpG methylation has been shown to be associated with gene activation.**

In fact is may be a more common event than repression. 85% of the surveyed genes showed a direct correlation between MeCP2 binding and gene activation.

Chahrour, M, Jung, SY, Shaw, C, Zhou, X, Wong, ST, Qin, J, Zoghbi, HY (2008) MeCP2, a key contributor to neurological disease, activates and represses transcription. Science, 320:1224–1229.

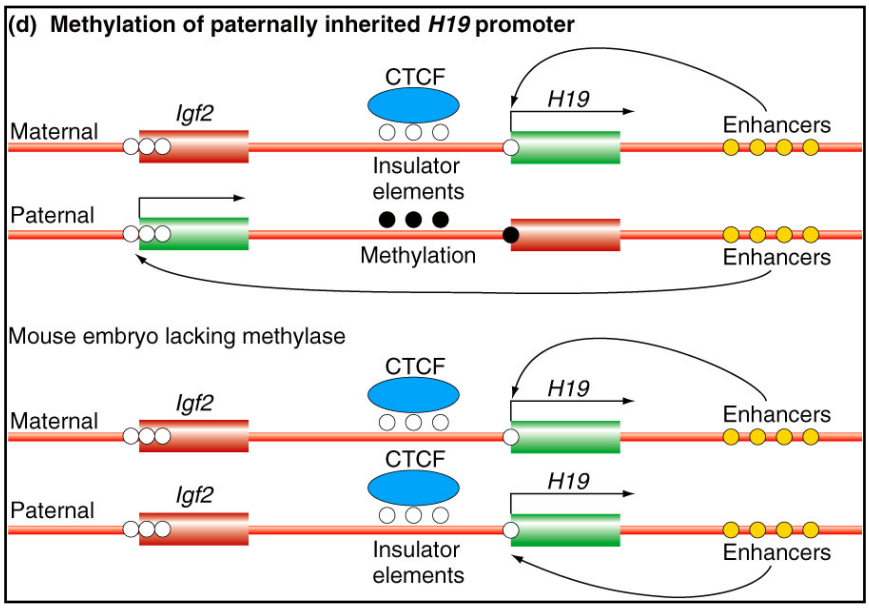
Sometimes results in repression and sometimes activation.
Luo, S.-W. et al. EMBO J. 28, 2568–2582 (2009).

B) **Imprinting**

Methylation activates a gene resulting in imprinting.

The CTCF protein is an insulator protein (boundary element protein, SARs protein). In this instance, in the maternal chromosome, it binds its targets and stops an enhancer from activating a promoter (Igf2 promoter). Therefore the maternal gene is off.

In the paternal chromosome, this sequence is methylated and CTCF CANNOT bind. Therefore, the activators that bind the enhancer turn on the promoter. The gene is on in the paternal chromosome.



From Genetics by Hartwell

XI) **What prevents the CpG's in front of promoters from being methylated?** Ideas – there is not a single agreed upon mechanism. Remember, right now we are talking about preventing methylation just during early embryogenesis. After that the methylation pattern will be maintained.

Choose one or more from the list.

A) There is evidence that proteins bind them and deny access to the Dnmt's.

B) The CpG island has a shape or acts as an origin of DNA replication and this prevents methylation.

C) Nucleotide turnover at these sites is bizarrely high during early embryogenesis.

D) Early transcription from a CpG promoter prevents it.
This is likely to be used.

E) CpG Methylation is determined by histone methylation pattern.

Now we have to wonder how the correct histones are modified.

There is substantial evidence that the H3 methylation (meH3K9) can cause methylation of CpGs.

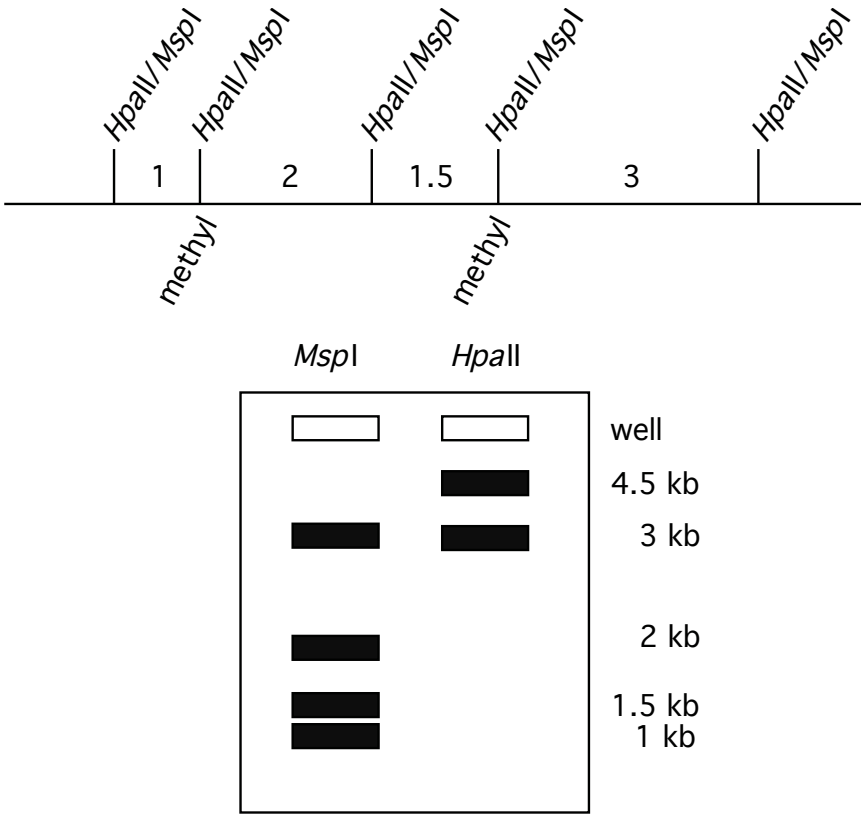
F) Antisense transcription to recruit Dnmt's

XII) **Detection**

A) DNA-methylation sensitive restriction enzymes. <-- Earliest method

*Hpa*II and *Msp*I recognition site is 5' CCGG 3'.

Isoschizomeric restriction enzymes e.g. *Hpa*II and *Msp*I can be used to test for CpG Methylation. *Hpa*II does not cut at 5'CCGG3' if the cytosines are methylated. MspI cuts at 5'CCGG3' whether they are methylated or not. 10 to 15% of CpGs are assayable by this technique.



B) **Bisulfite mapping** <-- Most common method.

Bisulfite sequencing allows the detection of methylated cytosine in anything that you can sequence.

Clark, S. J., Harrison, J., Paul, C. L., & Frommer, M. (1994). High sensitivity mapping of

methylated cytosines. *Nucleic Acids Res*, *22*(15), 2990-2997.

Sodium bisulphite converts cytosine to uracil in single stranded DNA but 5-methyl cytosine is unreactive ----> PCR amplify ----> sequence the modified DNA.

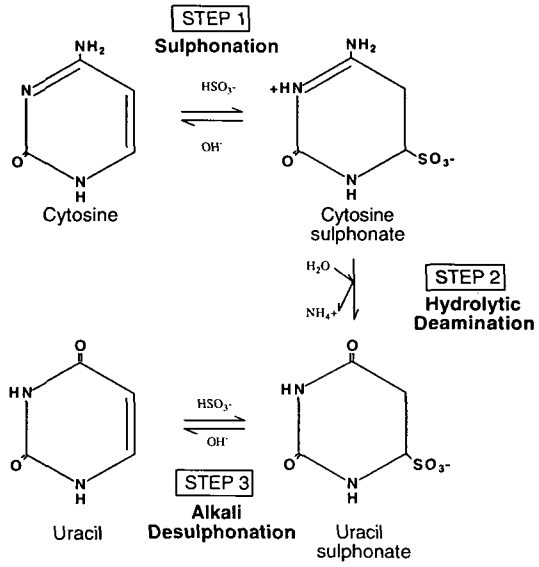
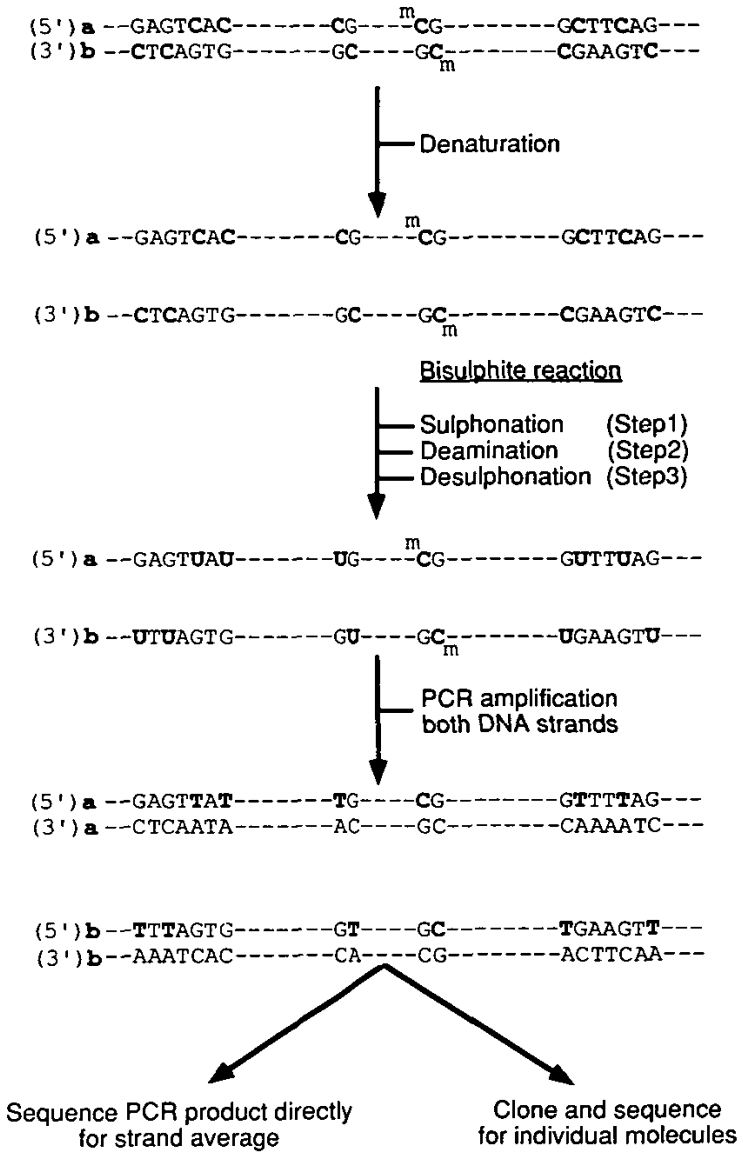


Figure 1. Schematic diagram of the bisulphite conversion reaction.



For individual genes you have a single PCR product. You then sequence this single product. Notice that there are two ways to go about the sequencing.

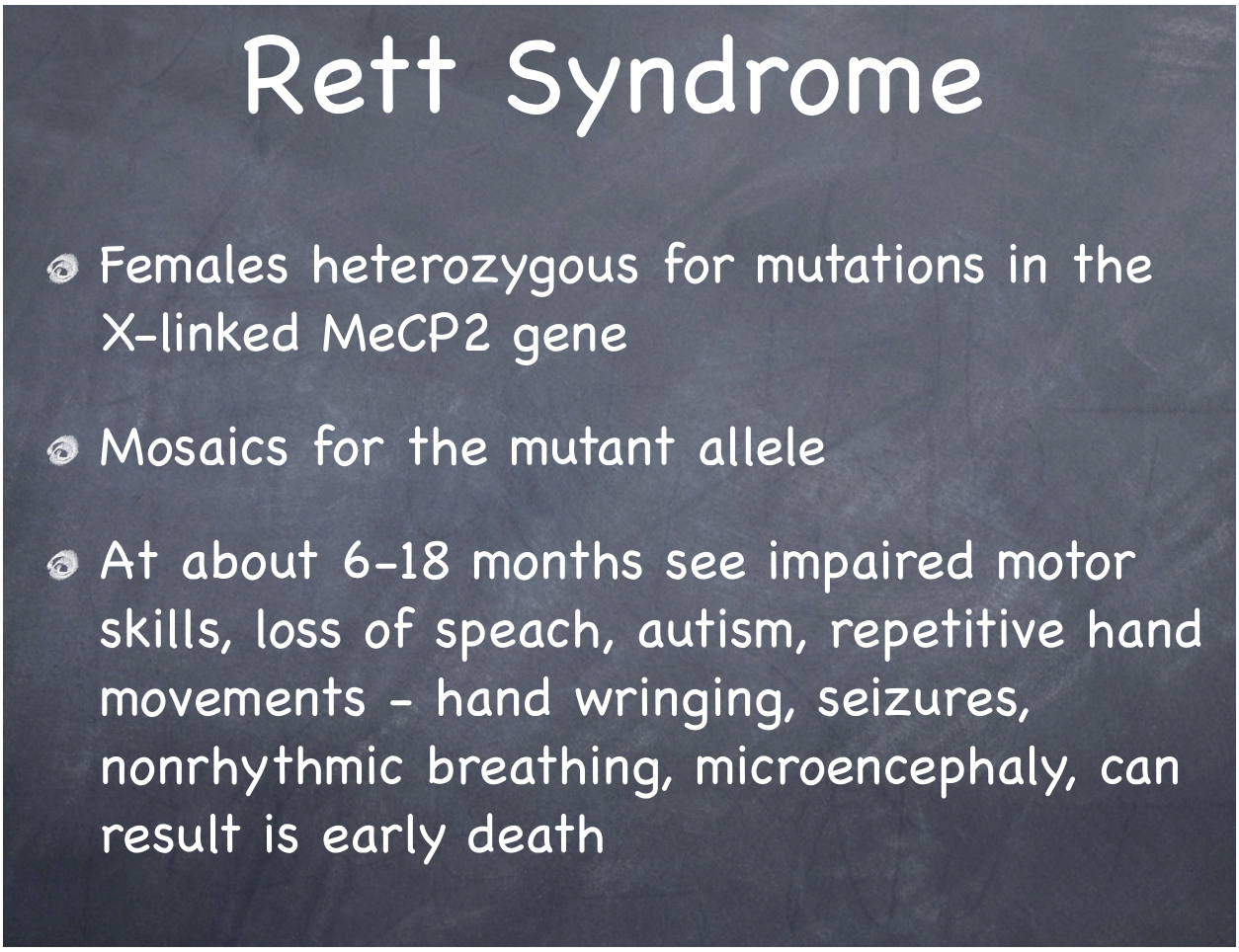
C) **Immunoprecipitation**

Use an antibody that binds 5meC, or a 5meC-binding protein and an antibody that binds the 5meC-binding protein.
Quantify the precipitated DNA.

XIII) **Rett Syndrome**

A) **Mutation in Mecp2. This gene is on the X. Only seen in females. Kills males.**

MeCP2 binds 5' methyl cytosine.



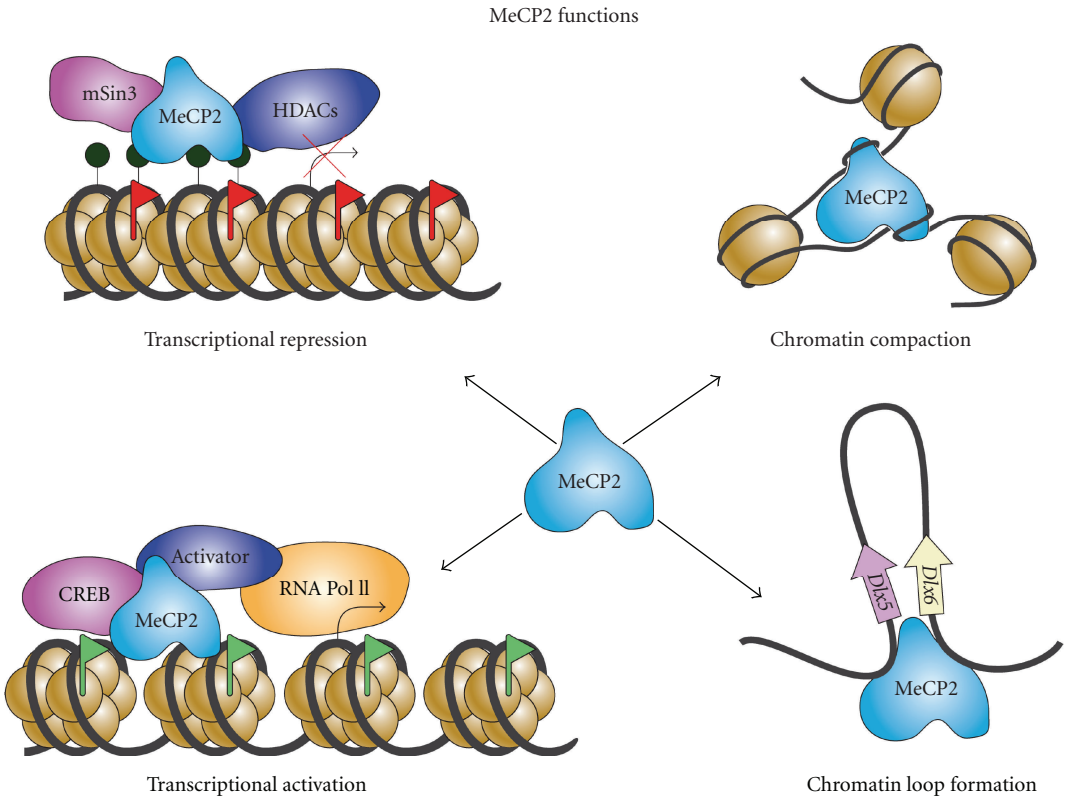


FIGURE 1: The diverse functions of MeCP2 in gene regulation and chromatin organization.

Zachariah, RM and Rastegar, M (2012). Linking epigenetics to human disease and Rett syndrome: the emerging novel and challenging concepts in MeCP2 research. Neural Plast 2012:415825.

XIV) Cancer

A) Epigenetic switching in the cancer genome [6]

1) Healthy cells have high methylation of CpGs that are not part of CpG islands and LOW methylation of CpGs that are part of CpG islands
2) Cancerous cells invert this pattern

B) Consequences

1) Loss of methylation of retrotransposons causes them to become active
2) Loss of methylation of repetitive elements (centromeres, satellite sequences, LINE1 elements telomeres, etc) can lead to genetic instability [4]
3) Loss of methylation results in bidirectional transcription.[4]
4) Hypermethylation of CpG islands involved in regulating genes that limit cell cycle[4]

5) HypOmethylation far exceeds HypErmethylation. [4]

C) **Epigenetic therapy for cancer** [4]

1) Too much CpG methylation of islands controlling cell cycling limiting genes.
2) Block DNMT enzyme can reverse this.
3) 5-azacytidine, 5-aza-2'-deoxyycytidine, 5'fluoro2-deoxycytidine zebularine
4) HDAC inhibitors can also reactive such genes

D) **Epigenomic biomarkers of cancer** [4]

1) Pattern of methylation changes appear to be diagnostic of specific types of cancer
2) May eventually be used as a premalignant marker, to monitor tumor progression and to monitor the progress of therapy.

E) **Mutation in Tet2 probably causes stochastic aberrant DNA methylation that can occasionally give cells a growth advantage.** [11] Thus, there is a selection for such events in cancer.

XV) 5-Hydroxymethyl-cytosine (5HmC) [11]

A) 5mC can be converted to 5HmC.

5mC are very similar between cells types but 5HmC vary quite a bit between tissue types.

B) 5HmC is interpreted differently than 5mC

So far, most 5mC binding proteins ignore 5hmC.
DNMT's work more poorly on 5HmC DNA than hemimethylated DNA.

XVI) Synthesis of 5HmC

A) Some members of TET protein family can 5meC-->5hmC.

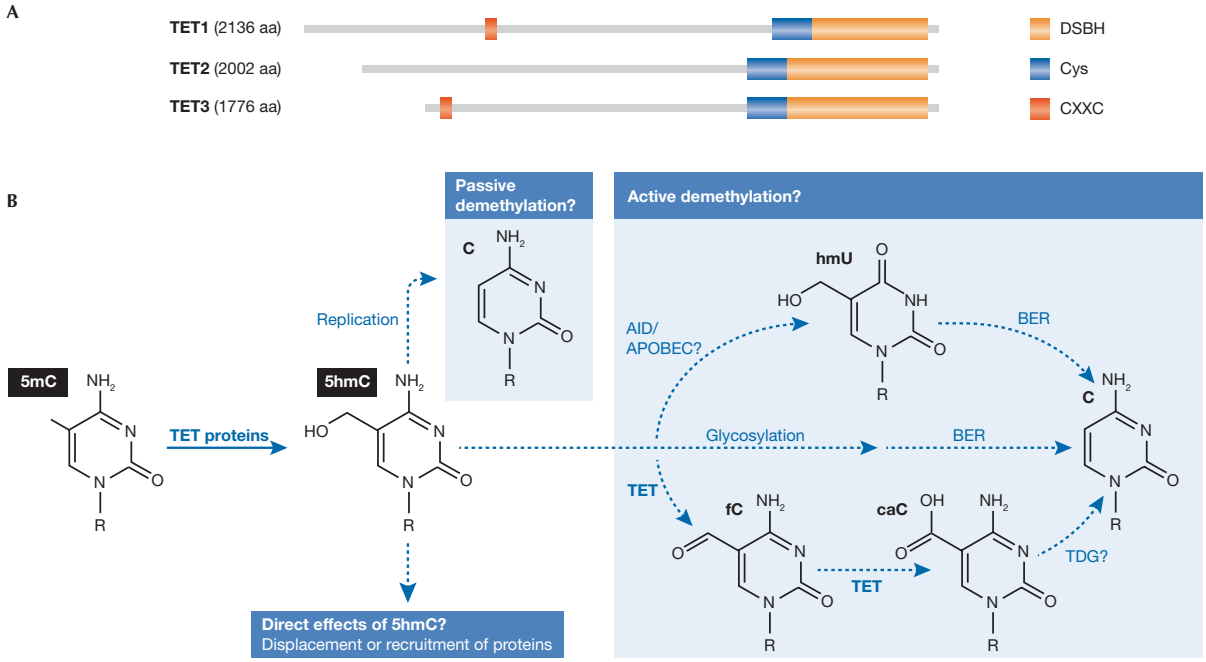


Fig 1 | Possible biological roles of TET proteins and 5hmC. (**A**) The domain structure of human TET proteins. TET1–3 contain a cysteine (Cys)-rich region followed by the double-stranded β-helix (DSBH) fold characteristic of the 2OG-Fe(II) oxygenases and required for catalytic activity. TET1 and TET3 also contain a CXXC domain. (**B**) Several biological consequences of the TET-mediated conversion of 5mC to 5hmC can be envisioned. 5hmC might facilitate a passive demethylation that is replication-dependent or could be converted to cytosine through an active demethylation pathway. Finally, 5hmC might also have direct effects by displacing or recruiting effector proteins. 5hmC, 5-hydroxymethylcytosine; 5mC, 5-methylcytosine.

B) TET2 frequently mutated in haematopoietic tumors

C) TET stands for Ten-Eleven Translocation

D) Have CXXC domain which is a CpG binding motif.

In stem cells, non-methylated CpG islands have been seen to be high in Tet1 and also H3K4me3.

E) Mutation in Tet2 probably causes stochastic aberrant DNA methylation that can occasionally give cells a growth advantage.

Thus, there is a selection for such events in cancer.

F) **KO of Tet gene does not effect transcription simply. Many genes show no change.**

G) **A late 2011 review on Tet proteins ends with a series of questions that clearly illustrate how fragile our understanding of these proteins are.** [11]

"(i) Do the TET proteins have major roles in regulating transcription?
(ii) Are the TET proteins regulating DNA methylation fidelity?
(iii) Do the TET proteins and 5hmC contribute to DNA demethylation?
(iv) Does 5hmC have a signalling function?
(v) How does loss of function of TET2 lead to cancer development?
(vi) How are TET proteins recruited to specific DNA-binding sites?
(vii) Do the TET proteins have functionally redundant functions?"

XVII) References

In addition to the references listed below and elsewhere in the text, I also drew on information read in Chapter 18 DNA Methylation in Mammals in Epigenetics textbook

and especially one of my favorite books:

Walter Doerfler, Petra Böhm (2006) DNA methylation: basics mechanisms. books.google.com and Genomes 2nd edition. Author T.A. Brown section 8.2

http://www.ncbi.nlm.nih.gov/books/bv.fcgi?highlight=CpG%20island&rid=genomes.section.6866#6890

1. Zachariah RM, Rastegar M (2012) Linking epigenetics to human disease and Rett syndrome: the emerging novel and challenging concepts in MeCP2 research. Neural Plast 2012: 415825.
2. Deaton AM, Bird A (2011) CpG islands and the regulation of transcription. Genes Dev 25: 1010-1022.
3. Antequera F (2003) Structure, function and evolution of CpG island promoters. Cell Mol Life Sci 60: 1647-1658.
4. Park YJ, Claus R, Weichenhan D, Plass C (2011) Genome-Wide Epigenetic Modifications in Cancer. In: Gasser S, Li E, editors. Epigenetics and Disease. Dordrecht: Springer. pp. 1 online resource (275 p.).
5. Chahrour M, Jung SY, Shaw C, Zhou X, Wong ST et al. (2008) MeCP2, a key contributor to neurological disease, activates and represses transcription. Science 320: 1224-1229.
6. Taberlay PC, Jones PA (2011) DNA Methylation and Cancer. In: Gasser S, Li E, editors. Epigenetics and Disease. Dordrecht: Springer. pp. 1 online resource (275 p.).
7. Medvedeva YA, Fridman MV, Oparina NJ, Malko DB, Ermakova EO et al. (2010) Intergenic, gene terminal, and intragenic CpG islands in the human genome. BMC Genomics 11: 48.
8. Illingworth RS, Bird AP (2009) CpG islands--'a rough guide'. FEBS Lett 583(11): 1713-1720.
9. NIH.GOV (2011) CpG Islands. National Library of Medicine - Medical Subject Headings 2011 MeSH, MeSH Descriptor Data Available: http://www.nlm.nih.gov/cgi/mesh/2011/MB_cgi?mode=&term=CpG+Islands via the Internet.
10. Hallgrímsson B, Hall BK (2011) Chaptr 9 The Role of Epigenetics in Nervous System Development.pdf. editors. Epigenetics: Linking Genotype and Phenotype in Development and Evolution. University of California Press. pp. 137-163.
11. Williams K, Christensen J, Helin K (2011) DNA methylation: TET proteins-guardians of

CpG islands? EMBO Rep 13: 28-35.