FEBS Letters 583 (2009) 1713–1720

journal homepage: www.FEBSLetters.org

Minireview

CpG islands – 'A rough guide'

Robert S. Illingworth *, Adrian P. Bird

Wellcome Trust Centre for Cell Biology, Michael Swann Building, University of Edinburgh, Mayfield Road, Edinburgh EH9 3JR, United Kingdom

ARTICLE INFO

Article history:
Received 9 March 2009
Revised 4 April 2009
Accepted 6 April 2009
Available online 18 April 2009

Edited by Miguel De la Rosa

Keywords:
CpG Island
Computational prediction
DNA methylation
Transcriptional regulation

ABSTRACT

Mammalian genomes are punctuated by DNA sequences containing an atypically high frequency of CpG sites termed CpG islands (CGIs). CGIs generally lack DNA methylation and associate with the majority of annotated gene promoters. Many studies, however, have identified examples of CGI methylation in malignant cells, leading to improper gene silencing. CGI methylation also occurs in normal tissues and is known to function in X-inactivation and genomic imprinting. More recently, differential methylation has been shown between tissues, suggesting a potential role in transcriptional regulation during cell specification. Many of these tissue-specific methylated CGIs localise to regions distal to promoters, the regulatory function of which remains to be determined.

1. Introduction

In mammals, the majority of CpG pairs are chemically modified by the covalent attachment of a methyl group to the C5 position of the cytosine ring. This modified residue is distributed throughout the majority of the genome including gene bodies, endogenous repeats and transposable elements and functions to repress transcription [1–3]. Methylcytosine spontaneously deaminates to thymine resulting in the under representation of CpG (21% of that expected in the human genome) [4]. The genome is punctuated however by non-methylated DNA sequences called CpG islands (CGIs) which have an elevated G + C content and little CpG suppression [5–7]. These conspicuous unique sequences are approximately 1 kb in length and overlap the promoter regions of 60–70% of all human genes [4,6,8–10].

CGIs have been shown to colocalise with the promoters of all constitutively expressed genes and approximately 40% of those displaying a tissue restricted expression profile [8,11]. CGI promoters appear to define a class of transcription start site (TSS) which can initiate from multiple positions. The more tissue restricted class of non-CGI promoters is generally associated with a single well defined initiation site (reviewed in [12]). Promoter association accounts for the uneven distribution of CGIs in the genome, showing preferential localisation to gene rich loci [4].

Consistent with promoter association, CGIs are generally characterised by a transcriptionally permissive chromatin state [10,13,14]. These findings suggest that CGIs may provide a means to distinguish gene promoter regions from the large proportion of transcriptionally irrelevant intergenic chromatin. Support for this idea was provided by an early study investigating the distribution of transcription factor (TF) binding sites in a small panel of human genes [15]. Whilst binding sites were slightly enriched in promoter proximal sequences, they were also highly abundant throughout the genome (approximately 16 sites per 100 bp). This study concluded that the presence of binding sites alone was insufficient to identify promoters, which supports the idea that CGIs may serve as TF "landing lights" in the darkness of the nucleus [15,16].

Not all CGIs localise to annotated TSSs (Fig. 1 – example iii), however it is interesting to note that detailed investigation of intragenic CGIs has led to the identification of previously unanticipated promoters [17–19]. This raises the possibility that all CGIs represent sites of transcriptional initiation many of which have yet to be characterised. Indeed it is possible that certain alternative transcriptional start sites are utilised in a highly tissue restricted fashion, and consequently have escaped annotation. Several transcripts initiate from intragenic CGIs and have been shown to be expressed during specific developmental stages [17,19].

Abbreviations: CGIs, CpG islands; TSS, transcription start site; ES cells, embryonic stem cells; CGBP, CpG-binding protein; MAGE, melanoma antigen encoding genes; RLGS, restriction landmark genome scanning; HOX, Homeobox; ncRNA, non-coding RNA; RACE, rapid amplification of cDNA ends

* Corresponding author. Fax: +44 0131 650 5379.
E-mail addresses: rillingw@staffmail.ed.ac.uk (R.S. Illingworth), a.bird@ed.ac.uk (A.P. Bird).

0014-5793/$36.00
doi:10.1016/j.febslet.2009.04.012

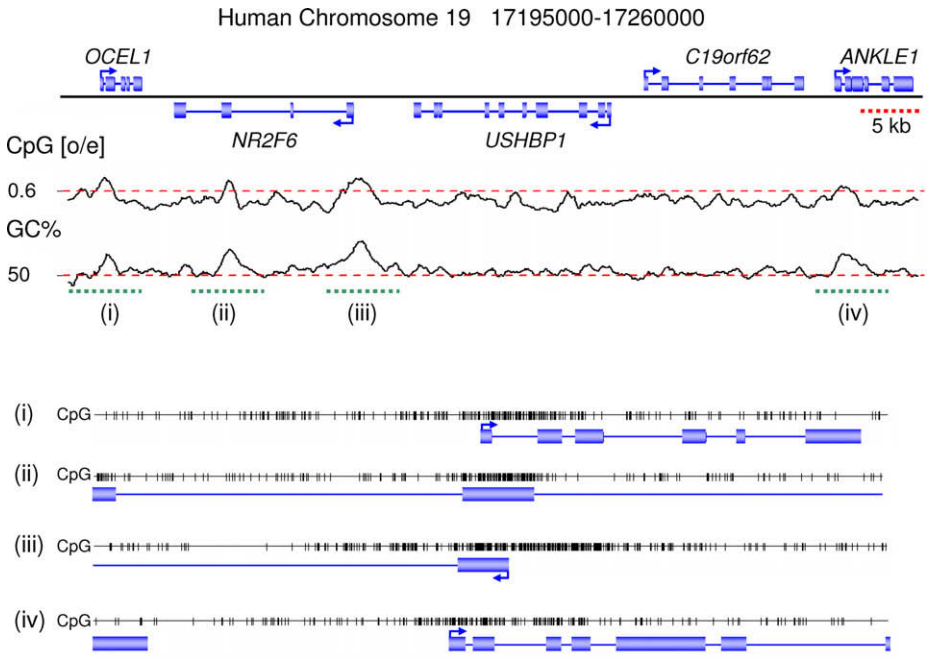

Fig. 1. CpG islands located within a region of human chromosome 19. The upper panel illustrates a 65 kb portion of human chromosome 19 (17195000–17260000) which contains five annotated genes (blue bars) and four CpG islands. The promoters of *OCEL1*, *NR2F6* and *ANKLE1* overlap with CGIs (i,iii and iv) and an additional CGI (ii) localises to the third exon of *NR2F6*. The classical sequence parameters applied to CGI prediction are illustrated (dashed red lines) for CpG (observed/expected; CpG[*o*/*e*] = 0.6) and G + C base composition (GC% = 50%). The lower panel represents an enlarged view of four 6 kb regions (i–iv) spanning each CGI and illustrates the distribution of CpG sites (vertical black strokes) relative to the annotated genes.

2. CGI identification

CGIs were first identified by digestion of mouse genomic DNA using the methyl-CpG sensitive restriction enzyme HpaII (CCGG recognition site). A small portion of the genome, composed of very highly fragmented DNA, was found to be derived from sequences containing clusters of non-methylated CpG sites [5,6,20]. Quantification of these digestion products, combined with sequence analysis and correction for contaminating DNA indicated that these were derived from approximately 26 300 discrete CGIs [21,22]. These sequences were characterised as at least 200 bp in length and with a G + C content of 50% and a CpG frequency (observed/expected; [*o*/*e*]) of 0.6 (Fig. 1) [7,8].

The completion of the human genome project in 2001 facilitated in silico CGI prediction [4]. Values for length and base composition similar to those identified by Gardiner-Garden and Frommer are routinely employed by the major genome browsers to annotate CGIs (Table 1). Thresholds are somewhat arbitrary however, and the effect of varying these values can profoundly alter prediction accuracy [23–25]. To reduce the extraneous inclusion of non-CGI sequences Takai and Jones investigated the effect of increasing the minimum length, CpG[*o*/*e*] and G + C composition to 500 bp, 0.65% and 55%, respectively. This increased stringency reduced the number of identified islands by approximately 90% and largely excluded contaminating Alu elements. This algorithm also reduced the number of gene promoter associated islands, suggesting that bona fide CGIs were also being discarded [24].

Repeat elements such as "young" Alus resemble the base composition of CGIs and significantly contribute to the number of false positives identified [24]. Preliminary computational analysis of the human genome sequence identified 50 267 CGIs, of which only 28 890 were unique [4]. Many of the multi copy sequences could be removed by screening against known classes of repeats identified in the Repbase database [26]. This database is subject to iterative improvements due to updating the repeat repertoire. Reanalysis of the human genome sequence in 2002 resulted in the loss of a further 1890 false positives suggesting a more conservative estimate of 27 000 CGIs [27]. The beneficial consequences of repeat masking can be illustrated by the example of a low copy repetitive element that is related to the adenovirus sequence located on human chromosomes 4 and 19 [28]. This element is identified as a single CGI or a tandem cluster of repeated CGIs by

Table 1
Overview of CpG island prediction algorithms.

Database/prediction	Length	G + C	CpG[*o*/*e*]	RM[a]	Comments	Reference
ENSEMBL	⩾400	⩾50%	⩾0.6	N	Stringent length constraint	[88]
NCBI relaxed	⩾200	⩾50%	⩾0.6	N	Total CGIs = 307 193	
NCBI strict	⩾500	⩾50%	⩾0.6	N	Total CGIs = 24 163	
USCS[b]	>200	⩾50%	>0.6	Y	Total CGIs = 28 226	[89]
EMBOSS	UD[c]	UD	UD	NA	Variable parameters	[90]
CpGProD	>500	>50%	>0.6	Y	Total CGIs = 76 793	[23]
CpGcluster	NA	NA	NA	N	Clustering Total = 197 727	[25]

[a] RM, repeat masked; Y, yes; N, no; NA, non applicable.
[b] Parameters used for CGI identification for the ENCODE project although totals vary due to repeat masking differences between hg17 and hg18 builds [87].
[c] UD, user defined.

ENSEMBL and NCBI, but is recognized as a repeat and eliminated by the algorithm employed by the USCS browser (Table 1).

The total number of predicted CGIs is highly variable depending on the exact sequence parameters applied. NCBI Mapview maintains two different permutations of these parameters to provide a relaxed and stringent identification of CpG islands (Table 1). "NCBI strict" predicts 24 163 unique CGIs whereas their relaxed criterion identifies more than 307 000. This variability arises due to the following factors: (1) the application of arbitrary thresholds, (2) no account being taken for the heterogeneity of CGIs and (3) the fact that DNA sequence based prediction methods necessarily ignore DNA methylation status.

To overcome these problems we recently developed a novel technique to select CGI sequences based on the empirical criterion of non-methylated CpG clustering [29]. A recombinant CXXC domain from murine MBD1 with specific affinity for non-methylated CpG pairs was used to purify CGIs from total genomic DNA [29–32]. The sequenced library identified in excess of 17 000 CGIs in human blood DNA. Extrapolating the identified CGI sequences to annotated genes suggests that the complete human somatic cell CGI complement is approximately 25 000 [29].

Most computational prediction and sequence selection techniques identify a CGI complement of between 24 000 and 27 000. Despite the apparent concordance between these methods, many identified CGIs are not common between the different sets. This inconsistency may be addressed by the incorporation of multiple layers of information into prediction methods, including DNA methylation status and chromatin modifications. CGIs generally associate with domains of chromatin containing hyperacetylated nucleosomes consistent with a transcriptionally permissive state [10,14,33]. In the future, epigenetic information may facilitate detection methods allowing current, somewhat arbitrary, thresholds to be replaced by accurate contextual information [34].

3. The origin of CpG islands

The mechanism by which CGIs remain hypomethylated during the period of global de novo methylation during early development remains unclear [35,36]. The characteristic clustering of CpG sites is a consequence of immunity against de novo methylation during the earliest stages of mammalian development. A simple suggestion would be that CGIs are intrinsically refractory to de novo methylation by DNA methyltransferases (DNMT) due to their DNA sequence (Fig. 2A). This seems unlikely however, as CGIs contain a substantially elevated density of CpG sites, the preferred substrate of the DNMT enzymes [37]. Moreover, CGIs located on the female inactive X chromosome and those of certain cultured mammalian cells readily acquire DNA methylation [2,38].

A second possibility is that CGIs are targeted by a DNA demethylation mechanism, which specifically removes the methyl moiety from the cytosine base (Fig. 2B) [39]. Various protein factors, including CGBP (CpG-binding protein) possess a CXXC domain, which can specifically bind to non-methylated CpG sites [40,41]. This protein has been shown to associate with the MLL complex, which mediates the formation of transcriptionally permissive chromatin via histone modifying activities [42]. It is possible, that an equivalent recruitment mechanism could target a demethylation activity to CGIs. However, no such demethylase activity has thus far been identified in somatic tissues.

A plausible alternative is that bound transcription factors sterically preclude DNMT association at CGI sequences (Fig. 2C) [43]. Evidence for such a mechanism is supported by mouse transgenic experiments in which ablation of binding sites for the ubiquitous transcription factor Sp1 was shown to facilitate de novo methylation of the *APRT* promoter CGI [44,45]. Consistent with this idea, α-globin is transcribed in the embryo and contains a promoter CGI whilst the related, but transcriptionally silent β-globin gene

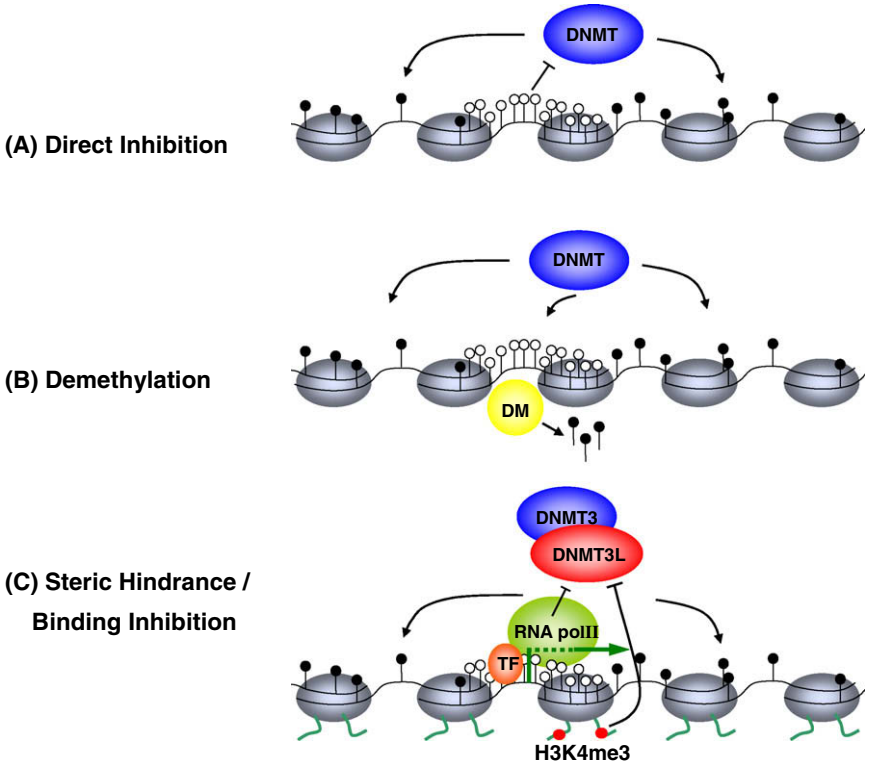

Fig. 2. Potential mechanisms leading to CGI hypomethylation. (A) CGIs remain hypomethylated via intrinsic sequence properties which exclude the action or association of DNA methyltransferases (DNMT; blue ovals). (B) CGIs acquire DNA methylation normally but are targeted by a demethylating activity (DM). (C) The basal transcriptional machinery (RNApolII and TF) and histone H3 lysine 4 trimethylation (H3K4me3) excludes the DNMTs from sites of transcriptional initiation (dashed green line). (A–C) Methylated and unmethylated CpGs are denoted by filled and open lollipops, respectively.

is not embryonically transcribed and has no CGI [46]. Analysis of a panel of genes expressed during mouse embryogenesis found that 93% are associated with a 5′ CGI [47]. These data raise the possibility that CGIs are footprints of the basal transcription machinery localised during embryogenic de novo methylation (Fig. 2C). Furthermore, global run on expression analysis indicated that bidirectional transcription initiation frequently occurs at gene promoters [48]. This observation could account for the relatively large region of steric hindrance that would be required to generate a typical CGI. However, this model does not account for the observation that CGIs are intrinsically sensitive to nuclease digestion and therefore more accessible than the majority of the genome [14,49]. Moreover, the majority of CGIs remain hypomethylated in terminally differentiated cells irrespective of transcriptional activity [1,8,10].

Recent studies investigating the methyltransferase like factor DNMT3L suggests a rather speculative mechanism for the persistence of hypomethylated CGIs. This protein associates with, and facilitates the action of the de novo methyltransferases [50–53]. However, DNMT3L cannot bind to chromatin in which the Histone H3 tails are tri-methylated at the lysine 4 position [54]. Genome wide determination indicated that the majority of protein coding gene promoters are occupied by RNA polymerase II and possess islands of trimethylated H3K4 even in the absence of transcriptional elongation in ES cells [13,55,56]. The presence of this active mark at CGI-promoters may be refractive to de novo methylation via repulsion of DNMT3L (Fig. 2C).

It is conceivable that more than one of these models is involved in the establishment of CGIs and the hypomethylation which usually persists during subsequent differentiation. Global analysis of chromatin modifications, transcriptional activity, transcription factor binding and DNA methylation analysis will help to determine the origin of CGIs and the mechanism that maintains them.

4. CpG island methylation

The majority of CGIs are hypomethylated, but a small percentage acquires methylation during normal development. Some of these examples are known to play a key role in X-inactivation and genomic imprinting [57,58]. Disruption of CGI methylation patterns has also been well-documented as a hallmark of neoplastic cells [59]. Recently, DNA "methylome" characterisation has been the basis for an increasing number of investigations due to significant advances in analytical technologies [1,2,10,29,60]. A major focus of this work has centered on CGIs as they represent a tractable fraction of the genome with obvious regulatory potential.

Several studies have recently improved our understanding of DNA methylation at CGI-promoters. This is of particular interest as it is known that hypermethylation of CGI promoters result in stable transcriptional repression [3]. Microarrays probed with DNA enriched for methyl-CpGs identified 3–4% of CGI-promoters as hypermethylated in a panel of somatic tissues [2,10,61]. Alternatively, promoters with relatively reduced CpG content were frequently found to be more often hypermethylated [10]. This is consistent with the observation that a methylated fraction purified from human whole blood was found to be enriched for DNA sequences with a CpG density intermediate between CGIs and bulk genomic DNA [62].

The above studies focused on gene promoter however CGIs distal to TSSs have also been implicated in transcriptional regulation [58,63]. Systematic analysis of all predicted CGIs (149) on the q arm of human chromosome 21 determined that 22% were hypermethylated in peripheral blood DNA [64]. An independent investigation characterised methylation at 2524 regions of human chromosomes 6, 20 and 22 across 12 tissues using high resolution bisulfite sequencing [1]. This study identified 9.2% of predicted CGIs as methylated at more than 80% of CpG sites in one or more somatic tissues [1]. More recently, global CGI methylation has been characterised through affinity purification of methylated DNA and microarray screening. MBD affinity purified (MAP) DNA identified 11.6% of islands as hypermethylated in a panel of somatic tissues [29]. In a similar manner, Rauch and coworkers identified approximately 25% of CGIs as being heavily methylated in human B cells [63].

These global studies indicate that sites of CGI methylation frequently localise to genomic regions distal to promoters. Consistent with this observation, bisulfite analysis identified 2.1% of promoter-associated CGIs as hypermethylated (>80% of CpGs) relative to more than 9% of the complete CGI complement [1]. However, despite this observation the exact proportion of hypermethylated CGIs varies widely between these studies (9–25%). The discrepancies between these studies may be attributed to three key experimental factors:

(1) *Variable detection:* The relative methylation levels required for detection differs between each of the analytical techniques. For example, bisulfite analysis provides single base pair resolution allowing the determination of intermediate levels of methylation (>20 and <80% ^{me}CpG) which is imperceptible by techniques such as Methyl-DNA Immunoprecipitation (MeDIP) [2].
(2) *Inconsistent CGI classification:* The number of CGIs identified can vary widely depending on the sequence parameters applied to their identification. This is illustrated for the study by Rauch et al., where the inclusion of CGIs with a relatively low CpG density, are frequently methylated relative to CGIs identified by more stringent criteria [1,7,63]. These sequences are arguably not bona fide CGIs.
(3) *Tissue specific CGI methylation:* A proportion of all CGIs are differentially methylated between tissues. Studies investigating multiple tissues will consequently identify a greater total number of hypermethylated CGIs.

Two recent studies have combined bisulfite conversion with next generation sequencing technology to characterise DNA methylation at CGIs [60,65]. This technology, although presently limited to the characterisation of a small fraction of the genome, provides unparalleled resolution and a greater insight into the distribution of DNA methylation in the mammalian genome.

5. Differential CGI methylation

A small but significant proportion of CGIs are differentially methylated between normal tissues and cell types [1,29,66–70]. Characterisation of these differences identifies the existence of tissue specific CGI methylation fingerprints which may demarcate cellular functions [69,71].

5.1. Germ line specific hypomethylation

A number of CGIs have been found to be unmethylated in cells of the germ line, but methylated in all tested somatic cell types. For example, germ line specific genes of the *MAGE* (melanoma antigen encoding genes) family acquire promoter-CGI methylation during embryogenesis and are silent in all somatic tissues [72]. Promoter demethylation correlates with the ectopic expression of these genes in various cancer cells suggesting that DNA methylation is the primary silencing mechanism [73]. Genome wide characterisation of a synthetic mouse differentiation model identified accumulation of de novo methylation and transcriptional silencing at the

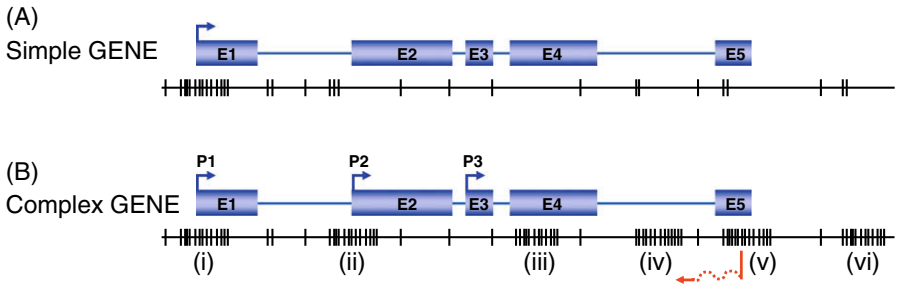

Fig. 3. Schematic representation of CGI gene association. (A) A simple mammalian gene with a single promoter associated CGI (high density of vertical black strokes). (B) A more complex gene structure including alternative promoters (P1–3), multiple intragenic CGIs (i–v), a single intergenic CGI and an antisense transcript (dashed red arrow).

promoters of many germ line specific genes [74]. Restriction landmark genome scanning (RLGS) in a panel of mouse tissues identified 5% of CGIs as differentially methylated [75]. Candidate analysis of 15 of these islands confirmed that 14 were specifically hypomethylated in mature sperm and heavily methylated in somatic cells [75]. A similar trend was identified in human tissues, where promoter arrays probed with methylated DNA from brain, testis and monocytes identified testis specific hypomethylation [67]. Furthermore, CGIs shown to be hypermethylated in human blood were also completely devoid of methylation in sperm DNA [61].

Sperm specific hypomethylation suggests that certain germline-specific genes may be irrevocably silenced via CGI methylation in somatic cells. Since mature sperm cells are transcriptionally inactive, the genes which are regulated by CGI methylation must be expressed during sperm maturation. Consistent with this suggestion, somatic acquisition of DNA methylation correlated well with transcriptional activity in human testis (containing primordial germ cells) and gene silencing in somatic cells [76]. Knowledge of specific gene expression profiles in immature germ cells will be required to determine if CGI methylation acts as a primary repressor of germ line specific genes in somatic cells.

5.2. Differential methylation in embryonic stem cells

Embryonic stem cells are pluripotent and might therefore be expected to lack any CGI methylation. Characterisation of DNA methylation patterns in mouse embryonic stem cells (ES cells) has however identified hypermethylation at approximately 3% of CGI-promoters. These included various developmental genes such as *Rhox2* and many genes involved in testis and oocyte specific functions [56]. *Rhox* genes are temporally and spatially regulated during post-implantation development in mice and are expressed specifically in the extraembryonic tissues [77]. The *Rhox* cluster has been shown to associate with embryo specific hypermethylation and transcriptional analysis in DNMT deficient embryonic cells indicates that this provides the primary silencing mechanism for these genes [56,77]. Minimal overlap between genes repressed by DNA promoter methylation and those targeted by PcG and Nanog/Oct4 suggests that there are multiple complementary regulatory mechanisms which maintain correct expression during mammalian embryogenesis [56].

5.3. Differential CGI methylation in somatic cells

Recent studies have revealed a significant fraction of CGIs that show tissue specific DNA methylation. It is tempting to hypothesise that this differential methylation serves to regulate gene expression during cellular differentiation. Consistent with this notion, the promoter-CGIs of r*SPHK1* and h*SLC6A8* have been shown to be specifically methylated in non-expressing tissues [68,78]. The CpG-rich promoter of the human gene *MASPIN* was shown to be differentially methylated in a panel of 10 somatic tissues and cell types. Although this promoter sequence represents a weak CGI, DNA methylation levels correlate well with the transcriptional activity of the gene [66].

Analysis of human chromosomes 6, 20 and 22 by bisulfite genomic sequencing identified eleven CGIs which were differentially methylated between 8 somatic tissues [1]. Interestingly, these genes displayed a relatively poor correlation with gene expression levels in these tissues. Global methylation studies also identified a limited concordance between differentially methylated CGIs and gene expression [29,63]. It is not yet clear whether this represents limited sensitivity of the transcriptional assays or an independent repression mechanism which functions irrespective of methylation status (discussed below).

A clearer understanding of the role of differential CGI methylation may be gained by characterising the function of specific genes in more detail. Strikingly, genes involved in developmental processes are frequently associated with differentially methylated CGIs. PAX6, OSR1 and various members of the Homeobox (HOX) super family have been shown to exhibit cell type-specific DNA methylation at CGIs [29,63,69,79]. HOX genes are highly conserved and function to dictate the positional identities of cells within the embryo, representing key regulators of mammalian development. Similarly, the *PAX6* transcription factor is required for neural and ocular development and its expression is temporally and spatially partitioned within the mammalian brain [19]. Further work is needed to test the hypothesis that tissue specific CGI methylation at genes of this kind plays an important role in cell type specification.

6. CGI methylation and transcription

6.1. CGI methylation and transcriptional regulation

There is extensive evidence to support a functional role for promoter-CGI methylation in transcriptional repression (see, for example [10,72,80]). DNA methylation of the CpG-rich promoters of *MASPIN* and *GATA2* correlates with tissue specific gene silencing [66,75]. In light of this evidence, it is tempting to hypothesise that the major function of CGI methylation is to repress transcription. However many genes display a relatively poor correlation between CGI hypermethylation and the transcriptional status of associated genes [29,63,81].

There are several potential explanations for this lack of correlation as illustrated in Fig. 3. In a simple example such as that depicted in Fig. 3A, hypermethylation of the single promoter associated CGI would lead to stable transcriptional silencing. The majority of methylated CGIs are located within intragenic regions where the effect on transcription is less clear [1,29,63].

Many genes can generate multiple transcripts by utilising alternative transcription starts sites. Rauch and colleagues identified expression of *PARP12* despite hypermethylation of its primary CGI promoter. Rapid amplification of cDNA ends (5′ RACE), however, identified transcription initiation from an intragenic promoter downstream of the methylated CGI [63]. Alternative promoters (e.g. P1–3 in Fig. 3B) could be inactivated by CGI methylation (Fig. 3B – CGIs (i) and (ii)).

Where intragenic islands do not associate with a known TSS, it is possible that their methylation could prevent spurious gene body transcription which could otherwise interfere with the correct expression of the parent gene (Fig 3B – CGI (iii and iv)). As yet there is no evidence for this conjecture.

There is evidence that Intragenic CGIs can localise to sites of antisense non-coding RNA (ncRNA) transcription initiation which negatively regulate the expression of the sense transcript (Fig. 3B – CGI (v)). Both the *Air* and *Tsix* ncRNA transcripts originate from CGIs and are involved in the regulation of the sense transcript [82–84]. The *HOXD* cluster is repressed in trans by the action of *HOTAIR, a ncRNA* transcribed from the HOXC locus [85]. In each case, CGI methylation results in the derepression of genes silenced by ncRNAs.

Many hypermethylated CGIs are located in intergenic DNA outside coding sequences and therefore have no obvious regulatory role in gene transcription (Fig. 3B – CGI (vi)). In the case of the *H19*/*IGF2* imprinted locus however, parent specific methylation at an intergenic CGI upstream of the *H19* ncRNA gene determines the expression of the imprinted locus. CGI methylation prevents the association of the insulator element CTCF and promotes expression of *IGF2* from the paternal allele [58]. This illustrates a potential mechanism whereby hypermethylation of intergenic CGIs can illicit a transcriptional effect.

These examples illustrate the complexity in determining the effect of DNA methylation at CGIs. Characterisation of transcription initiation using RNA polymerase chromatin immunoprecipitation and RACE will provide a better understanding of the function of CGI methylation at these sites.

6.2. Initiation or maintenance

Does hypermethylation of TSS associated CGIs act as the initial silencing mechanism or as a secondary event to provide stable, heritable gene repression? Several germ line and embryonic specific genes associate with methylated CGI promoters and can be reactivated by depletion of DNA methylation levels [56,72,73,77]. This observation indicates the former possibility; although it is conceivable that once silenced, the initial repressive event is lost and DNA methylation merely acts as a maintenance device. Several studies have identified differential CGI methylation between somatic tissues associated with constitutively repressed genes. This suggests that methylation is stochastically accumulated in different cell types in the absence of transcription. This fits with the observation that CGI methylation is a relatively late event during X-inactivation following gene repression [86].

Absence of TFs at silenced promoters could facilitate transient de novo methylation. This possibility would align with the notion that methylation may be regarded as the basal state of the genome and is excluded from specific regions by the presence of bound factors. Alternatively DNMT recruitment could be mediated by initial repressive events to target DNA methylation and irrevocably silence transcription of the associated gene. To dissect these possibilities it will be necessary to measure gene transcription levels, chromatin modification, transcription factor binding and DNA methylation during cellular differentiation to determine the order of events leading to transcriptional repression.

7. Concluding remarks

The completion of the human and mouse genome projects has revealed and unexpectedly small number of genes [4,27]. The mammalian transcriptome, however, is highly complex, with many genes generating multiple, often functionally distinct transcripts [12]. This is the result of many factors, including alternative splicing, differential promoter usage, TF availability, and the expression of regulatory ncRNAs. Several recent studies have identified differential patterns of DNA methylation across the genome. This evidence indicates that CGI methylation may provide an important epigenetic component of mammalian development and cellular differentiation. Interestingly, one recent study identified extensive tissue-specific methylation localised to regions which flank CGIs (<2 kb). Differential DNA methylation of these CGI "shores" correlates well with tissue specific gene expression [69]. These findings illustrate the complex role played by DNA methylation in the regulation of mammalian transcription.

There are many remaining questions. Do all CGIs colocalise to sites of transcriptional initiation? Do tissue-specific methylation patterns have a mechanistic role in "hard wiring" expression patterns in terminally differentiated cells? How prevalent is inter-individual differential CGI methylation? These questions must be addressed before we can begin to understand the role of CGIs in transcriptional regulation and consequently the aberrant events associated with disease.

References

[1] Eckhardt, F. et al. (2006) DNA methylation profiling of human chromosomes 6, 20 and 22. Nat. Genet. 38, 1378–1385.
[2] Weber, M., Davies, J.J., Wittig, D., Oakeley, E.J., Haase, M., Lam, W.L. and Schubeler, D. (2005) Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells. Nat. Genet. 37, 853–862.
[3] Bird, A. (2002) DNA methylation patterns and epigenetic memory. Genes Dev 16, 6–21.
[4] Lander, E.S. et al. (2001) Initial sequencing and analysis of the human genome. Nature 409, 860–921.
[5] Cooper, D.N., Taggart, M.H. and Bird, A.P. (1983) Unmethylated domains in vertebrate DNA. Nucleic Acids Res. 11, 647–658.
[6] Bird, A., Taggart, M., Frommer, M., Miller, O.J. and Macleod, D. (1985) A fraction of the mouse genome that is derived from islands of nonmethylated, CpG-rich DNA. Cell 40, 91–99.
[7] Gardiner-Garden, M. and Frommer, M. (1987) CpG islands in vertebrate genomes. J. Mol. Biol. 196, 261–282.
[8] Larsen, F., Gundersen, G., Lopez, R. and Prydz, H. (1992) CpG islands as gene markers in the human genome. Genomics 13, 1095–1107.
[9] Saxonov, S., Berg, P. and Brutlag, D.L. (2006) A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. Proc. Natl. Acad. Sci. USA 103, 1412–1417.
[10] Weber, M., Hellmann, I., Stadler, M.B., Ramos, L., Paabo, S., Rebhan, M. and Schubeler, D. (2007) Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. Nat. Genet. 39, 457–466.
[11] Zhu, J., He, F., Hu, S. and Yu, J. (2008) On the nature of human housekeeping genes. Trends Genet. 24, 481–484.
[12] Sandelin, A., Carninci, P., Lenhard, B., Ponjavic, J., Hayashizaki, Y. and Hume, D.A. (2007) Mammalian RNA polymerase II core promoters: insights from genome-wide studies. Nat. Rev. Genet. 8, 424–436.
[13] Guenther, M.G., Levine, S.S., Boyer, L.A., Jaenisch, R. and Young, R.A. (2007) A chromatin landmark and transcription initiation at most promoters in human cells. Cell 130, 77–88.
[14] Tazi, J. and Bird, A. (1990) Alternative chromatin structure at CpG islands. Cell 60, 909–920.
[15] Prestridge, D.S. and Burks, C. (1993) The density of transcriptional elements in promoter and non-promoter sequences. Hum. Mol. Genet. 2, 1449–1453.
[16] Bird, A.P. (1995) Gene number, noise reduction and biological complexity. Trends Genet. 11, 94–100.
[17] Macleod, D., Ali, R.R. and Bird, A. (1998) An alternative promoter in the mouse major histocompatibility complex class II I-Abeta gene: implications for the origin of CpG islands. Mol. Cell Biol. 18, 4433–4443.
[18] Gardiner-Garden, M. and Frommer, M. (1994) Transcripts and CpG islands associated with the pro-opiomelanocortin gene and other neurally expressed genes. J. Mol. Endocrinol. 12, 365–382.
[19] Kleinjan, D.A., Seawright, A., Childs, A.J. and van Heyningen, V. (2004) Conserved elements in Pax6 intron 7 involved in (auto)regulation and alternative transcription. Dev. Biol. 265, 462–477.

[20] Bird, A.P. (1987) CpG Islands as gene markers in the Vertebrate Nucleus. TIG 3, 342–347.
[21] Antequera, F. and Bird, A. (1993) Number of CpG islands and genes in human and mouse. Proc. Natl. Acad. Sci. USA 90, 11995–11999.
[22] Ewing, B. and Green, P. (2000) Analysis of expressed sequence tags indicates 35,000 human genes. Nat. Genet. 25, 232–234.
[23] Ponger, L. and Mouchiroud, D. (2002) CpGProD: identifying CpG islands associated with transcription start sites in large genomic mammalian sequences. Bioinformatics 18, 631–633.
[24] Takai, D. and Jones, P.A. (2002) Comprehensive analysis of CpG islands in human chromosomes 21 and 22. Proc. Natl. Acad. Sci. USA 99, 3740–3745.
[25] Hackenberg, M., Previti, C., Luque-Escamilla, P.L., Carpena, P., Martinez-Aroza, J. and Oliver, J.L. (2006) CpGcluster: a distance-based algorithm for CpG-island detection. BMC Bioinformatics 7, 446.
[26] Jurka, J., Kapitonov, V.V., Pavlicek, A., Klonowski, P., Kohany, O. and Walichiewicz, J. (2005) Repbase Update, a database of eukaryotic repetitive elements. Cytogenet. Genome Res. 110, 462–467.
[27] Waterston, R.H. et al. (2002) Initial sequencing and comparative analysis of the mouse genome. Nature 420, 520–562.
[28] Epstein, N.D., Karlsson, S., O'Brien, S., Modi, W., Moulton, A. and Nienhuis, A.W. (1987) A new moderately repetitive DNA sequence family of novel organization. Nucleic Acids Res. 15, 2327–2341.
[29] Illingworth, R. et al. (2008) A novel CpG island set identifies tissue-specific methylation at developmental gene loci. PLoS Biol. 6, e22.
[30] Birke, M., Schreiner, S., Garcia-Cuellar, M.P., Mahr, K., Titgemeyer, F. and Slany, R.K. (2002) The MT domain of the proto-oncoprotein MLL binds to CpG-containing DNA and discriminates against methylation. Nucleic Acids Res. 30, 958–965.
[31] Jorgensen, H.F., Ben-Porath, I. and Bird, A.P. (2004) Mbd1 is recruited to both methylated and nonmethylated CpGs via distinct DNA binding domains. Mol. Cell Biol. 24, 3387–3395.
[32] Lee, J.H., Voo, K.S. and Skalnik, D.G. (2001) Identification and characterization of the DNA binding domain of CpG-binding protein. J. Biol. Chem. 276, 44669–44676.
[33] Roh, T.Y., Cuddapah, S. and Zhao, K. (2005) Active chromatin domains are defined by acetylation islands revealed by genome-wide mapping. Genes Dev. 19, 542–552.
[34] Bock, C., Walter, J., Paulsen, M. and Lengauer, T. (2007) CpG island mapping by epigenome prediction. PLoS Comput. Biol. 3, e110.
[35] Antequera, F. (2003) Structure, function and evolution of CpG island promoters. Cell Mol. Life Sci. 60, 1647–1658.
[36] Antequera, F. and Bird, A. (1999) CpG islands as genomic footprints of promoters that are associated with replication origins. Curr. Biol. 9, R661–R667.
[37] Ramsahoye, B.H., Biniszkiewicz, D., Lyko, F., Clark, V., Bird, A.P. and Jaenisch, R. (2000) Non-CpG methylation is prevalent in embryonic stem cells and may be mediated by DNA methyltransferase 3a. Proc. Natl. Acad. Sci. USA 97, 5237–5242.
[38] Antequera, F., Boyes, J. and Bird, A. (1990) High levels of de novo methylation and altered chromatin structure at CpG islands in cell lines. Cell 62, 503–514.
[39] Frank, D., Keshet, I., Shani, M., Levine, A., Razin, A. and Cedar, H. (1991) Demethylation of CpG islands in embryonic cells. Nature 351, 239–241.
[40] Voo, K.S., Carlone, D.L., Jacobsen, B.M., Flodin, A. and Skalnik, D.G. (2000) Cloning of a mammalian transcriptional activator that binds unmethylated CpG motifs and shares a CXXC domain with DNA methyltransferase, human trithorax, and methyl-CpG binding domain protein 1. Mol. Cell Biol. 20, 2108–2121.
[41] Carlone, D.L., Hart, S.R., Ladd, P.D. and Skalnik, D.G. (2002) Cloning and characterization of the gene encoding the mouse homologue of CpG binding protein. Gene 295, 71–77.
[42] Ansari, K.I., Mishra, B.P. and Mandal, S.S. (2008) Human CpG binding protein interacts with MLL1, MLL2 and hSet1 and regulates Hox gene expression. Biochim. Biophys. Acta 1779, 66–73.
[43] Cuadrado, M., Sacristan, M. and Antequera, F. (2001) Species-specific organization of CpG island promoters at mammalian homologous genes. EMBO Rep. 2, 586–592.
[44] Brandeis, M. et al. (1994) Sp1 elements protect a CpG island from de novo methylation. Nature 371, 435–438.
[45] Macleod, D., Charlton, J., Mullins, J. and Bird, A.P. (1994) Sp1 sites in the mouse aprt gene promoter are required to prevent methylation of the CpG island. Genes Dev. 8, 2282–2292.
[46] Daniels, R., Lowell, S., Bolton, V. and Monk, M. (1997) Transcription of tissue-specific genes in human preimplantation embryos. Hum. Reprod. 12, 2251–2256.
[47] Ponger, L., Duret, L. and Mouchiroud, D. (2001) Determinants of CpG islands: expression in early embryo and isochore structure. Genome Res. 11, 1854–1860.
[48] Core, L.J., Waterfall, J.J. and Lis, J.T. (2008) Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. Science 322, 1845–1848.
[49] Antequera, F., Macleod, D. and Bird, A.P. (1989) Specific protection of methylated CpGs in mammalian nuclei. Cell 58, 509–517.
[50] Hata, K., Okano, M., Lei, H. and Li, E. (2002) Dnmt3L cooperates with the Dnmt3 family of de novo DNA methyltransferases to establish maternal imprints in mice. Development 129, 1983–1993.
[51] Suetake, I., Shinozaki, F., Miyagawa, J., Takeshima, H. and Tajima, S. (2004) DNMT3L stimulates the DNA methylation activity of Dnmt3a and Dnmt3b through a direct interaction. J. Biol. Chem. 279, 27816–27823.
[52] Gowher, H., Liebert, K., Hermann, A., Xu, G. and Jeltsch, A. (2005) Mechanism of stimulation of catalytic activity of Dnmt3A and Dnmt3B DNA-(cytosine-C5)-methyltransferases by Dnmt3L. J. Biol. Chem. 280, 13341–13348.
[53] Jia, D., Jurkowska, R.Z., Zhang, X., Jeltsch, A. and Cheng, X. (2007) Structure of Dnmt3a bound to Dnmt3L suggests a model for de novo DNA methylation. Nature 449, 248–251.
[54] Ooi, S.K. et al. (2007) DNMT3L connects unmethylated lysine 4 of histone H3 to de novo methylation of DNA. Nature 448, 714–717.
[55] Mikkelsen, T.S. et al. (2007) Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature 448, 553–560.
[56] Fouse, S.D., Shen, Y., Pellegrini, M., Cole, S., Meissner, A., Van Neste, L., Jaenisch, R. and Fan, G. (2008) Promoter CpG methylation contributes to ES cell gene regulation in parallel with Oct4/Nanog, PcG complex, and histone H3 K4/K27 trimethylation. Cell Stem Cell 2, 160–169.
[57] Reik, W. (2007) Stability and flexibility of epigenetic gene regulation in mammalian development. Nature 447, 425–432.
[58] Edwards, C.A. and Ferguson-Smith, A.C. (2007) Mechanisms regulating imprinted genes in clusters. Curr. Opin. Cell Biol. 19, 281–289.
[59] Esteller, M. (2007) Cancer epigenomics: DNA methylomes and histone-modification maps. Nat. Rev. Genet. 8, 286–298.
[60] Meissner, A. et al. (2008) Genome-scale DNA methylation maps of pluripotent and differentiated cells. Nature 454, 766–770.
[61] Shen, L. et al. (2007) Genome-wide profiling of DNA methylation reveals a class of normally methylated CpG island promoters. PLoS Genet. 3, 2023–2036.
[62] Brock, G.J., Charlton, J. and Bird, A. (1999) Densely methylated sequences that are preferentially localized at telomere-proximal regions of human chromosomes. Gene 240, 269–277.
[63] Rauch, T.A., Wu, X., Zhong, X., Riggs, A.D. and Pfeifer, G.P. (2009) A human B cell methylome at 100-base pair resolution. Proc. Natl. Acad. Sci. USA 106, 671–678.
[64] Yamada, Y. et al. (2004) A comprehensive analysis of allelic methylation status of CpG islands on human chromosome 21q. Genome Res. 14, 247–266.
[65] Zeschnigk, M. et al. (2009) Massive parallel bisulfite sequencing of CG-rich DNA fragments reveals that methylation of many X-chromosomal CpG islands in female blood DNA is incomplete. Hum. Mol. Genet. 18, 1439–1448.
[66] Futscher, B.W., Oshiro, M.M., Wozniak, R.J., Holtan, N., Hanigan, C.L., Duan, H. and Domann, F.E. (2002) Role for DNA methylation in the control of cell type specific maspin expression. Nat. Genet. 31, 175–179.
[67] Schilling, E. and Rehli, M. (2007) Global, comparative analysis of tissue-specific promoter CpG methylation. Genomics 90, 314–323.
[68] Imamura, T., Ohgane, J., Ito, S., Ogawa, T., Hattori, N., Tanaka, S. and Shiota, K. (2001) CpG island of rat sphingosine kinase-1 gene: tissue-dependent DNA methylation status and multiple alternative first exons. Genomics 76, 117–125.
[69] Irizarry, R.A. et al. (2009) The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. Nat. Genet. 41, 178–186.
[70] Shiota, K., Kogo, Y., Ohgane, J., Imamura, T., Urano, A., Nishino, K., Tanaka, S. and Hattori, N. (2002) Epigenetic marks by DNA methylation specific to stem, germ and somatic cells in mice. Genes Cells 7, 961–969.
[71] Ladd-Acosta, C. et al. (2007) DNA methylation signatures within the human brain. Am. J. Hum. Genet. 81, 1304–1315.
[72] De Smet, C., Lurquin, C., Lethe, B., Martelange, V. and Boon, T. (1999) DNA methylation is the primary silencing mechanism for a set of germ line- and tumor-specific genes with a CpG-rich promoter. Mol. Cell Biol. 19, 7327–7335.
[73] Honda, T., Tamura, G., Waki, T., Kawata, S., Terashima, M., Nishizuka, S. and Motoyama, T. (2004) Demethylation of MAGE promoters during gastric cancer progression. Br. J. Cancer 90, 838–843.
[74] Mohn, F., Weber, M., Rebhan, M., Roloff, T.C., Richter, J., Stadler, M.B., Bibel, M. and Schubeler, D. (2008) Lineage-specific polycomb targets and de novo DNA methylation define restriction and potential of neuronal progenitors. Mol. Cell 30, 755–766.
[75] Song, F., Smith, J.F., Kimura, M.T., Morrow, A.D., Matsuyama, T., Nagase, H. and Held, W.A. (2005) Association of tissue-specific differentially methylated regions (TDMs) with differential gene expression. Proc. Natl. Acad. Sci. USA 102, 3336–3341.
[76] Schilling, E. and Rehli, M. (2007) Global, comparative analysis of tissue-specific promoter CpG methylation. Genomics. 90, 314–323.
[77] Oda, M. et al. (2006) DNA methylation regulates long-range gene silencing of an X-linked homeobox gene cluster in a lineage-specific manner. Genes Dev. 20, 3382–3394.
[78] Grunau, C., Hindermann, W. and Rosenthal, A. (2000) Large-scale methylation analysis of human genomic DNA reveals tissue-specific differences between the methylation profiles of genes and pseudogenes. Hum. Mol. Genet. 9, 2651–2663.
[79] Bloushtain-Qimron, N. et al. (2008) Cell type-specific DNA methylation patterns in the human breast. Proc. Natl. Acad. Sci. USA 105, 14076–14081.
[80] Stein, R., Razin, A. and Cedar, H. (1982) In vitro methylation of the hamster adenine phosphoribosyltransferase gene inhibits its expression in mouse L cells. Proc. Natl. Acad. Sci. USA 79, 3418–3422.
[81] Oakes, C.C., La Salle, S., Smiraglia, D.J., Robaire, B. and Trasler, J.M. (2007) A unique configuration of genome-wide DNA methylation patterns in the testis. Proc. Natl. Acad. Sci. USA 104, 228–233.
[82] Panning, B. and Jaenisch, R. (1996) DNA hypomethylation can activate Xist expression and silence X-linked genes. Genes Dev. 10, 1991–2002.

[83] Sleutels, F., Zwart, R. and Barlow, D.P. (2002) The non-coding Air RNA is required for silencing autosomal imprinted genes. Nature 415, 810–813.
[84] Wutz, A., Smrzka, O.W., Schweifer, N., Schellander, K., Wagner, E.F. and Barlow, D.P. (1997) Imprinted expression of the Igf2r gene depends on an intronic CpG island. Nature 389, 745–749.
[85] Rinn, J.L. et al. (2007) Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. Cell 129, 1311–1323.
[86] Mermoud, J.E., Popova, B., Peters, A.H., Jenuwein, T. and Brockdorff, N. (2002) Histone H3 lysine 9 methylation occurs rapidly at the onset of random X chromosome inactivation. Curr. Biol. 12, 247–251.
[87] Birney, E. et al. (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature 447, 799-816.
[88] Birney, E. et al. (2004) An overview of Ensembl. Genome. Res. 14, 925–928.
[89] Karolchik, D. et al. (2008) The UCSC genome browser database: 2008 update. Nucleic Acids Res. 36, D773–D779.
[90] Rice, P., Longden, I. and Bleasby, A. (2000) EMBOSS: the European molecular biology open software suite. Trends Genet. 16, 276–277.