Computers in Biology and Medicine 68 (2016) 57–66

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/cbm

Genome wide classification and characterisation of CpG sites in cancer and normal cells

Mohammadmersad Ghorbani [a,c], Michael Themis [b], Annette Payne [a,*]

[a] Department of Computer Science, Brunel University, Uxbridge, Middlesex UB8 3PH, UK
[b] Department of Biosciences, Brunel University, Uxbridge, Middlesex UB8 3PH, UK
[c] Wellcome Trust - Medical Research Council Cambridge Stem Cell Institute

ARTICLE INFO

Article history:
Received 11 June 2015
Accepted 29 September 2015

Keywords:
Motif
Pattern identification
Methylation in cancer
Computational analysis
Pattern searching algorithm
CpG
DNA sequence

ABSTRACT

This study identifies common methylation patterns across different cancer types in an effort to identify common molecular events in diverse types of cancer cells and provides evidence for the sequence surrounding a CpG to influence its susceptibility to aberrant methylation. CpG sites throughout the genome were divided into four classes: sites that either become hypo or hyper-methylated in a variety cancers using all the freely available microarray data (HypoCancer and HyperCancer classes) and those found in a constant hypo (Never methylated class) or hyper-methylated (Always methylated class) state in both normal and cancer cells. Our data shows that most CpG sites included in the HumanMethylation450K microarray remain unmethylated in normal and cancerous cells; however, certain sites in all the cancers investigated become specifically modified. More detailed analysis of the sites revealed that majority of those in the never methylated class were in CpG islands whereas those in the HyperCancer class were mostly associated with miRNA coding regions. The sites in the Hypermethylated class are associated with genes involved in initiating or maintaining the cancerous state, being enriched for processes involved in apoptosis, and with transcription factors predicted to bind to these genes linked to apoptosis and tumourgenesis (notably including E2F). Further we show that more LINE elements are associated with the HypoCancer class and more Alu repeats are associated with the HyperCancer class. Motifs that classify the classes were identified to distinguish them based on the surrounding DNA sequence alone, and for the identification of DNA sequences that could render sites more prone to aberrant methylation in cancer cells. This provides evidence that the sequence surrounding a CpG site has an influence on whether a site is hypo or hyper methylated.

1. Introduction

DNA methylation involving the addition of methyl groups to CpG sequences is one of the mechanisms used by the cells to control gene expression, gene silencing being a major biological consequence of DNA methylation. This phenomenon, known as epigenetic control has been reported to be important to mammalian development, X inactivation and genomic imprinting [1]. Epigenetic changes have been shown to occur in both healthy cells, where it assists in regulating gene expression during development, and diseased cells, where it is associated with aberrant gene expression, most notably in oncogenesis [2]. Many studies have also shown that differentially methylated CpG sites can act as a biomarkers in identifying disease and specific CpG site methylation can be a signature for specific types of tumours [3–7]. In tumour development global DNA hypomethylation is often followed by hypermethylation at specific CpGs [8–12]. Closer inspection of these studies and the fact that the cancer phenotype is associated with aberrant expression of a significant number of the same genes e.g. TP53 and RB1, in different cancer types, would suggest that there are common pathways and molecular mechanisms that can be identified across the different types of cancer. Further the differentially methylated CpGs could be informative in discovering mechanisms leading to malignancy. Factors that influence CpG methylation include chromatin accessibility, which have been shown to modulate methylation, DNASE1 footprinting, transcription factor levels and CTCF binding, where higher levels and the act of binding protect DNA from methylation [13–17].

Particular DNA motifs have been identified in previous studies that may be used to predict the methylation status of DNA sequences in normal cells. Notably methylation is more prevalent

* Corresponding author.
E-mail address: annette.payne@brunel.ac.uk (A. Payne).

http://dx.doi.org/10.1016/j.compbiomed.2015.09.023
0010-4825/

in regions of low CpG density, with regions of intermediate density being most variably methylated [18]. Yamada and Satou [19] employed machine learning methods, specifically support vector machine and random forest methods, using previously reported methylation data, to analyse DNA sequence features to predict methylation status. They revealed that frequencies of sequences containing CG, CT or CA are different when they compared unmethylated and methylated CpG islands. Ali and Seker [20] used an adapted K-nearest neighbour classifier method to predict the methylation state on chromosomes 6, 20 and 22 in various tissues. They identified four feature sub-sets which showed that methylated CpG islands can be distinguished from unmethylated CpG islands based on DNA sequence. Lastly Previti et al. [21] used data mining in the absence of supervised clustering to predict the methylation status of CpG islands in different tissues. These studies showed that there are significant differences in the sequences of CpG islands (CGIs) that predisposed them to methylation. Other studies have identified that the density and spacing of CpGs, the histone code (methylation of histone 3 at Lysine 4 (H3K4)), CTCF protein binding and REST protein binding can influence DNA methylation [22–29]. In their review of computational epigenetics called "Computational Epigenetics", Bock and Lengauer [18] highlighted the fact that, although it is clear that much work has been done to document the epigenetic state of the genome (much of it reported in the ENCODE project [17]), to date, work in the area of de novo DNA methylation prediction is limited. One study however has shown that aberrant methylation has been shown to be associated with mutations where methylation in the MGMT promoter has been demonstrated to be closely associated with G:C to A:T mutations [30].

Thus whilst studies have identified motifs associated with normal methylation patterns few studies have attempted to search for motifs associated with aberrant methylation using computational techniques, one study by Feltus et al. [31] used Restriction Landmark Genome Scanning software to identify methylation resistant and methylation prone motifs based on DNA sequence and another by Lu et al. [32] has been carried out using word composition computation. Gorbani et al. [33] have suggested that the sequence surrounding a CpG can be used to predict aberrant methylation in trinucleotide repeat diseases using a pattern searching algorithm. Their results suggest that the sequence surrounding a CpG can be used to predict aberrant methylation. In another study by McCabe et al. [34] patterns were identified using machine learning techniques and used for pattern matching where DNA signatures and a co-occurrence with polycomb binding were found to predict aberrant CpG methylation in cancer cells. The reason for recruitment of the de novo DNA methyltransferases to specific genomic targets however remains largely unknown. Dnmt3 and certain transcription factors have been shown to interact with each other to target methylation Hervouet et al. [35] and recently it has been reported that DNMT3L and the lysine methyltransferase G9a are required for the initiation of proviral de novo DNA methylation [36,37]. Lastly Rowe et al. [38] have shown that ERV sequences are sufficient to direct rapid de novo methylation of a flanked promoter in embryonic stem (ES) cells.

In this study we have used a pattern searching algorithm to identify motifs in the DNA surrounding aberrantly methylated CpGs in the DNA of cancer cells from multiple cancer types and tissues so as to investigate whether common patterns of methylation across these different cancers can be identified. Previous studies have concentrated on one cancer or tissue type. Further most former studies that analysed surrounding DNA sequences are based on the sequences surrounding CpG islands or two classes of islands, methylation prone and methylation resistant. CpGs not associated with islands were not included [31]. With more data becoming publicly available about the methylation status around single CpG sites not associated with islands, it is now possible to investigate increasing numbers of sites and more additional classes of DNA methylation. In this study, we examined the DNA sequences surrounding CpG sites. We divided sites into four classes of DNA methylation: sites that either become hypo or hypermethylated in a variety cancers (HypoCancer and HyperCancer classes) and those found in a constant hypo (Never methylated class) or hyper-methylated (Always methylated class) state in both normal and cancer cells. Thus we have divided the CpG sites into four classes:

1. Never methylated in either cancer or normal cells (class NM).
2. Always methylated class in cancer and normal cells (class AM).
3. Hypomethylated in normal and hypermethylated in cancer (class HyperCancer).
4. Hypermethylated in normal and hypomethylated in cancer (class HypoCancer).

Then we investigated the DNA sequence flanking these sites to investigate if we could find common sequences or motifs in each class. We have carried out this work in an attempt to better understand a possible influence of DNA sequence on aberrant methylation.

1.1. Objectives of this work

1. Identify four classes of CpG sites based on data from diverse cancer types and normal tissue.
2. Identify methylation sites that could act as biomarkers.
3. Analyse the genes and DNA features associated with differentially methylated CpG sites to identify any links with carcinogenesis.
4. To identify DNA motifs in the DNA sequence surrounding a CpG that could render a CpG prone to aberrant methylation in cancer.
5. Using these motifs, suggest prediction criteria that could be used to identify CpG sites that are differentially methylated in normal and cancer cells in silico.

2. Results

2.1. CpG sites and their classes

Using the method described 653 CpG sites were identified that could be divided into the four classes according to their methylation status: 447 CpG sites in the Never methylated class (class NM), 148 sites in the Always methylated class (class AM), 51 hypomethylated in normal and hypermethylated in cancer (class HyperCancer) and 7 sites hypermethylated in normal and hypomethylated in cancer (class HypoCancer). We mapped the positional relationship of the CpG sites to CpG islands in the UCSC browser. 81 CpG sites were not in any positional relationship with a CpG Island. Never methylated sites are predominantly within islands. Most of the CpGs in the two classes of variably methylated sites have no relationship to any CpG islands. Always methylated CpGs are spread among the different positional relationships to UCSC CpG islands. These results are shown in Figs. 1 and 2.

2.2. MicroRNA results

The UCSC table browser was used in order to find out if methylation of these CpG sites could interfere with the expression of microRNA coding regions since miRNAs are suggested to interact with epigenetic machinery [39] and are important regulators of gene expression that are aberrantly regulated in cancer through