Genomics 93 (2009) 434–440

Contents lists available at ScienceDirect

Genomics

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

SNP-based prediction of the human germ cell methylation landscape

Hehuang Xie [1], Min Wang [1], Jared Bischof, Maria de Fatima Bonaldo, Marcelo Bento Soares *

Cancer Biology and Epigenomics Program, Children's Memorial Research Center, 2300 Children's Plaza, Box 220; Department of Pediatrics, Feinberg School of Medicine, Northwestern University, Chicago, IL 60614-3394, USA

ARTICLE INFO

Article history:
Received 25 August 2008
Accepted 16 January 2009
Available online 3 February 2009

Keywords:
SNP
Germ cells
DNA methylation
Methylation associated SNP
Homology-dependent methylation
Segmental duplication
Tandem repeat
CpG island
Alu elements

ABSTRACT

Base substitution occurs at a high rate at CpG dinucleotides due to the frequent methylation of CpG and the deamination of methylated cytosine to thymine. If these substitutions occur in germ cells, they constitute a heritable mutation that may eventually rise to polymorphic frequencies, hence resulting in a SNP that is methylation associated. In this study, we sought to identify clusters of methylation associated SNPs as a basis for prediction of methylation landscapes of germ cell genomes. Genomic regions enriched with methylation associated SNPs, namely "methylation associated SNP clusters", were identified with an agglomerative hierarchical clustering algorithm. Repetitive elements, segmental duplications, and syntenic tandem DNA repeats were enriched in methylation associated SNP clusters. The frequency of methylation associated SNPs in Alu Y/S elements exhibited a gradient pattern suggestive of linear spreading, being higher in proximity to methylation associated SNP clusters and lower closer to CpG islands. Interestingly, methylation associated SNP clusters were over-represented near the transcriptional initiation sites of immune response genes. We propose a *de novo* DNA methylation model during germ cell development whereby a pattern is established by long-range chromatic interactions through syntenic repeats combined with regional methylation spreading from methylation associated SNP clusters.

Introduction

DNA methylation is an epigenetic modification involved in many biological processes, including development, aging, and tumorigenesis. It is a dynamic process during development consisting of at least two genome reprogramming phases: first during gametogenesis and subsequently at preimplantation [1]. Considering the data generated from mouse models, the paternal genome is actively demethylated prior to DNA replication after fertilization while the maternal genome is passively demethylated with cleavage divisions [2,3]. *De novo* DNA methylation occurs prior to implantation, to establish the tissue specific methylation patterns of somatic cells. During the migration in the genital ridge, the genomes of primordial germ cells lose methylation until reaching the gonad. Imprinted regions arise after another round of *de novo* methylation at gonadal sex determination for the male and at assembly of primordial follicles in the female [1,4,5].

In spite of decades of effort, genome-wide DNA methylation patterns and underlying DNA methylation mechanisms remain largely unknown, principally due to technical limitations and inaccessibility of human tissues. High-throughput approaches, such as microarray hybridization, large-scale epigenomic sequencing, and methylation-sensitive enzyme associated approaches have been recently exploited to construct global DNA methylation landscapes [6–8]. However, the epigenome of human germ cells is mostly unknown. Several computational approaches have been developed to predict global methylation patterns, albeit based on limited datasets [9–11]. Most of these efforts have focused on the methylation patterns of CpG islands, which are stretches of DNA with high GC contents and unusually rich in CpG dinucleotides. In contrast to CpG islands, which are known to be predominantly hypomethylated, most repetitive elements are believed to be hypermethylated in somatic tissues. Homology dependent methylation has been found to be a mechanism which initiates *de novo* DNA methylation and transmits methylation pattern in *Neurospora* [12]. DNA methylation has been proposed to be a result of the interactions between homologous DNA:DNA or DNA:RNA pairings [12,13]. Both homologous pairings have been associated with repetitive elements, which may serve as "way-stations" [14,15]. Since little experimental evidence has been provided, genome-wide studies aimed at uncovering the methylation patterns of repetitive elements are greatly needed in that they would improve our understanding of the establishment and maintenance of DNA methylation in vertebrates. However, the low complexity characteristics of repeat sequences leads to cross-hybridization in array-based experiments and to inaccurate mapping or alignment in sequence-based approaches. Indeed, methylation studies of repetitive elements are extremely challenging with current high-throughput approaches.

Abbreviations: SNP, single nucleotide polymorphism; CGI, CpG island; LINE, long interspersed elements; SINE, short interspersed elements; LTR, long terminal repeat.

* Corresponding author. Fax: +1 773 755 6551.
E-mail address: mbsoares@childrensmemorial.org (M.B. Soares).

[1] These authors contributed equally to this work.

0888-7543/$ – see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.ygeno.2009.01.005

In addition to its effects on gene expression and chromosomal organization, methylation of cytosine leads to a high rate of cytosine to thymine transition. Spontaneous deamination of unmethylated cytosine to uracil occurs at a frequency as high as 100–500 events per cell per day [16]. The hydrolytic deamination rate occurring at methylated cytosine is two to three-fold higher than that of unmethylated cytosine [17]. In addition, the repair of a T:G mismatch by thymine DNA glycosylase is much less efficient than the repair of a U:G mismatch by uracil DNA glycosylase [16]. It has been estimated that C to T (or G to A) transitions at CpG sites occur at frequencies that are at least ten-fold higher than those of other nucleotide substitutions [18–20]. Nucleotide substitutions occurring in somatic tissues are not transmitted to the offspring. In contrast, deamination of a methylated cytosine in germ cells constitutes a heritable mutation that may eventually rise to polymorphic frequencies, hence resulting in a SNP that is methylation associated. Thus, SNP data can be used to estimate mutation rates that result from deamination of methylated cytosines [21].

In this study, we extracted C/T and G/A SNPs occurring at CpG dinucleotides in the human genome ([C/T]G or C[G/A]), and developed a clustering algorithm to identify methylation associated SNP clusters. More than 25,000 methylation associated SNP clusters were thus identified and utilized to generate a putative DNA methylation landscape of human germ cells. In addition, based on the analysis of the methylation associated SNP clusters herein identified, we propose a model for the establishment of genome-wide DNA methylation patterns.

Results

SNP classification and chromosome distribution

To reveal the underlying methylation information embedded in the SNP data, we developed an approach to classify and cluster SNPs (Fig. 1). Over 11 million RefSNP clusters were downloaded from the NCBI SNP database [22]. 9,229,281 bi-allelic RefSNP clusters that mapped unambiguously to single genomic loci were extracted for further analysis. Among them, a total of 2,002,619 [C/T] or [G/A] SNPs occurring at CpG dinucleotides were identified and defined as methylation associated SNPs. Considering that there are approximately 30 million CpG dinucleotides in the human genome, this analysis indicates the average methylation associated SNP occurrence rate be of the order of one per fifteen CpG dinucleotides (one SNP per thirty bases). This rate is ten-fold higher than the overall SNP occurrence rate (one SNP per 300 bases). Such overrepresentation of SNPs at CpG sites was also observed in previous genome-wide SNP studies [23]. It is important to note that one cannot distinguish C to T transitions as results of deamination of methylated cytosines from those derived from normal mutation based exclusively on SNP data. However, the former occurs at a much higher frequency. Indeed, as mentioned above, when compared with other types of nucleotide substitutions, C to T transitions resulting from deamination of methylcytosines occur at a rate that is at least ten-fold higher [18–20]. Therefore in our study, when we classified all [C/T] and [G/A] SNPs occurring at CpG dinucleotides as methylation associated SNPs, we expected that less than 10% of such SNPs were falsely predicted to be associated with methylation.

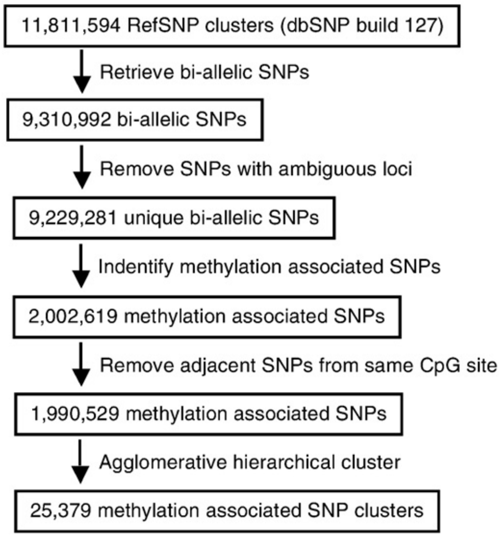
11,811,594 RefSNP clusters (dbSNP build 127)
Retrieve bi-allelic SNPs
9,310,992 bi-allelic SNPs
Remove SNPs with ambiguous loci
9,229,281 unique bi-allelic SNPs
Indentify methylation associated SNPs
2,002,619 methylation associated SNPs
Remove adjacent SNPs from same CpG site
1,990,529 methylation associated SNPs
Agglomerative hierarchical cluster
25,379 methylation associated SNP clusters

Fig. 1. Data processing flow chart for the identification of methylation associated SNP clusters. Detailed descriptions of each step were provided in the methods. Briefly, after downloading dbSNP database, bi-allelic SNPs were first retrieved. SNPs that were mapped to more than one genomic locus were removed. The remaining unique bi-allelic SNPs were then classified based on the types of immediate adjacent nucleotides. [C/T] or [G/A] SNPs occurring at CpG dinucleotide sites were identified as methylation associated SNPs. Adjacent SNPs occurring at the same CpG dinucleotide site were considered as one methylation associated SNP in order to be consistent with the number of CpG sites that can undergo methylation. Finally, an agglomerative hierarchical clustering algorithm was implemented to identify genomic regions with a high density of methylation associated SNPs, which were defined as methylation associated SNP clusters.

As a first step, distributions of CpG dinucleotides, SNPs, and methylation associated SNPs on each chromosome were examined (Table 1). Significant biases were observed on chromosome Y as CpG dinucleotides were significantly underrepresented. On chromosome Y, there were only 3.8 CpG dinucleotides per kb, in contrast to the genome-wide average of approximately 10 CpG dinucleotides per kb. Similarly, SNPs were also underrepresented on the Y chromosome: 0.53 SNPs per kb as opposed to the average genome-wide rate of 3 SNPs per kb. Consequently, methylation associated SNPs occur at a particularly low frequency on the Y chromosome: 2% of the CpG sites contain a methylation associated SNP, compared with a genome-wide frequency of 7%. Moreover, 16% of all SNPs on chromosome Y are methylation associated SNPs, in contrast to a genome-wide rate of 22%.

Methylation associated SNP clustering

To identify genomic regions with a high density of methylation associated SNPs, an agglomerative hierarchical clustering algorithm was implemented [24]. One of the most complicated problems in clustering algorithm is to determine the number of iterations to execute. Also challenging in this particular case was to decide the number of methylation associated SNPs to be in a cluster. Unfortunately, no existing experimental data can be used to justify the requirements to define a “way-station” in terms of the density of methylated CpG dinucleotides and the length of the region. We identified approximately two million methylation associated SNPs in the human genome, with an occurrence rate of one in 1500 bases on average. To ensure a high density of methylation associated SNPs, we arbitrarily stopped clustering iterations at 20 with 537 bp as the maximum spacing between methylation associated SNPs within a cluster. We further required that methylation associated SNP clusters contain at least six methylation associated SNPs. As a result, the methylation associated SNP clusters defined in our study had a three-fold enrichment for methylation associated SNPs compared to the average level in the human genome.

After 20 clustering iterations, 1,189,794 clusters of methylation associated SNPs had been generated. Among them, 25,379 clusters (2.1%) encompassed six or more methylation associated SNPs, representing over 17.8% of all methylation associated SNPs. These 25,379 clusters were defined as methylation associated SNP clusters (Supplementary table). As discussed above, we assumed that less than 10% of the methylation associated SNPs might not be associated with cytosine methylation at CpG dinucleotides. Therefore, the statistical likelihood decreases exponentially for having more methylation associated SNPs falsely predicted within a methylation associated