Zhang MQ, Smith AD. Challenges in understanding genome-wide DNA methylation. JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY 25(1): 26–34 Jan. 2010

Challenges in Understanding Genome-Wide DNA Methylation

Michael Q. Zhang[1,2] (张奇伟) and Andrew D. Smith[3], *Member*, *ACM*

[1] *Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, U.S.A.*

[2] *Bioinformatics Division, TNLIST and MOE Key Lab of Bioinformatics, Tsinghua University, Beijing 100084, China*

[3] *Department of Biological Sciences, University of Southern California, Los Angeles, California, U.S.A.*

E-mail: mzhang@cshl.edu; andrewds@usc.edu

Received November 19, 2009; revised November 24, 2009.

Abstract DNA methylation is a chemical modification of the bases in genomes. This modification, most frequently found at CpG dinucleotides in eukaryotes, has been identified as having multiple critical functions in broad and diverse species of animals and plants, while mysteriously appears to be lacking from several other well-studied species. DNA methylation has well known and important roles in genome stability and defense, its pattern change highly correlates with gene regulation. Much evidence has linked abnormal DNA methylation to human diseases. Most prominently, aberrant DNA methylation is a common feature of cancer genomes. Elucidating the precise functions of DNA methylation therefore has great biomedical significance. Here we provide an update on large-scale experimental technologies for detecting DNA methylation on a genomic scale. We also discuss new prospect and challenges that computational biologist will face when analyzing DNA methylation data.

Keywords DNA methylation, epigenome, computational epigenomics

1 Introduction

Our capacity for collecting data about DNA methylation has surged forward alongside recent advances in sequencing technology (for more details on advances in sequencing technology, see the article “New generations: Sequencing machines and their computational challenges” by Shwartz and Waterman[1]). Making sense of DNA methylation data presents many challenges for computational biologists who must design novel algorithmic and statistical methods for analyzing the data. In this article we review the biology of DNA methylation, explain the experimental technologies for collecting this data, and then review the computational problems associated with analyzing DNA methylation data.

2 Background

DNA methylation was first postulated over 30 years ago to be a heritable modification capable of affecting gene expression[2-3]. The addition of a methyl group (-CH_3) to the cytosine base (mC) does not alter the primary DNA sequence and is therefore considered to be an epigenetic modification, literally meaning to act “on top of” or “in addition” to genetics. Three DNA methyltransferase (Dnmt1, Dnmt3a and Dnmt3b) are required for the establishment and maintenance of DNA methylation patterns[4] and two additional enzymes (Dnmt2 and Dnmt3L) may also have more specialized but related functions[5]. Genomes in different cell types have different epigenetic codes written at least partially in patterns of DNA methylation on top of a common underlying genetic code giving rise different cellular phenotypes.

Epigenetic gene-silencing (e.g., methylation of the gene promoter) can be equivalent to a null mutation phenotypically. DNA methylation may have developed as a host defense against expression of parasitic DNA insertions[6] and it has also been used as a weapon mediating sexual conflict[7].

Since mammalian genomic DNA methylation exists primarily as mCpG and 5-methylcytosine is prone to spontaneous deamination and point mutation to thymine[8], CpG dinucleotides are consequently depleted almost 5-fold in the genome during human evolution[9]. Although 5-mCpG represents only 1% bases in the human genome, CpG dinucleotide is involved in 1/3 of point mutations causing human genetic disorders[10] and a similar proportion of SNPs detected in coding regions[11]. DNA methylation changes in cancer appear to be 10∼100 times more frequent events

Survey

This work is supported by NIH under Grant Nos. ES017166 and HG001696.

than genetic mutations[12].

Due to its importance, the Human Reference Epigenome Project was initiated in Europe in 2003, aiming to "identify, catalog and interpret genome-wide DNA methylation patterns of all human genes in all major tissues"[13]. In 2009, the NIH Roadmap Initiative started supporting large-scale Human Reference Epigenome Mapping efforts that will produce epigenomic maps (including DNA methylation, Histone modifications, ncRNA transcripts) in many cell types[14].

3 Experimental Methods

Biologists seeking to investigate DNA methylation have several options available for extracting the methylation data. These technologies depend on specific characteristics of methylated and unmethylated cytosines that can be used to positively identify the presence of the modification. Here we describe the three major classes of experimental methods for detecting DNA methylation. While each of these general methods can be implemented in several ways, when coupled with second-generation sequencing technologies they become especially powerful.

3.1 Methylation-Sensitive Restriction

Restriction enzymes recognize specific DNA sequences and cut DNA molecules at or near those sites. Certain restriction enzymes have been discovered to cut DNA at CpG sites in a methylation sensitive manner. Enzymes like McrBC recognize sites with methylated cytosines[15]. Others, including *HpaII* and *NotI*, recognize unmethylated sites[16-17]. Digesting DNA with these enzymes (alone or in combinations) can yield much information about methylation. Early experiments were able to quantify overall levels of genomic methylation by restriction with methylation-sensitive enzymes and then measuring the distribution of sizes for the resulting fragments. A genome with very little DNA methylation would result in large fragments when cut with an enzyme that recognizes methylated sites. Restriction-based methods have also been adapted to provide information about localization of DNA methylation, for example when combined with microarrays to identify which genomic fragments have been targeted for restriction[18].

3.2 Immunoprecipitation

Antibodies that detect either methylated cytosines or methyl-CpG-binding domain (MBD) proteins can be used to detect the presence of methylated cytosines[19]. Immunoprecipitation with one of these antibodies results in a sample enriched in DNA fragments containing methylated cytosines. Then the sample is interrogated, either with an array-based method or with direct sequencing (e.g., using a high-throughput second-generation sequencing technology). As with methods based on methylation-sensitive restriction, immunoprecipitation-based methods have limited resolution. Sophisticated analytical methods have been designed to help increase the resolution of immunoprecipitation methods[20].

3.3 Bisulfite Treatment

Treatment of DNA with sodium bisulfite has the effect of converting cytosine, through deamination, to uracil. If the treatment is followed by PCR amplification, those uracil bases are converted into thymine. Methylated cytosines, however, are left unconverted. If some method is then applied to interrogate the sequences, identifying cytosines at genomic CpG positions indicates methylation, while identifying thymine at the same position indicates lack of methylation in the molecule.

The sample interrogation can be done in a variety of ways, including PCR[21], array hybridization[22] and even mass-spec[23] (C and U yield different mass signals). However, the combination of DNA sequencing with bisulfite treatment, termed bisulfite sequencing (BS-seq), has been especially powerful. When coupled with second-generation sequencing, it becomes feasible to conduct BS-seq experiments on entire mammalian genomes. The main tradeoff in using second-generation sequencing for BS-seq is between the length of the reads and the accuracy of those reads: short reads will generally contain fewer CpGs, and therefore analysis that seeks to understand relationships between CpGs in the same chromosome will be more difficult (see Section 8).

There are several complications in bisulfite sequencing. First, DNA methylation does not survive PCR, which constrains the amount of starting DNA and makes it difficult to conduct DNA methylation on those important but rare cells of early development or cancer stem cells. The main tradeoff is the between the degree of bisulfite conversion and the loss of DNA observed due to the harshness of the bisulfite treatment.

Currently the only technology that can produce data about methylation of individual CpGs is bisulfite sequencing; regional levels of DNA methylation can also be obtained by restriction or ChIP-based methods. In later sections we will generally assume that the bisulfite sequencing method has produced the DNA methylation data.

4 Mapping Bisulfite Treated Reads

The first technical challenge to emerge from the coupling of bisulfite sequencing with short-read technology

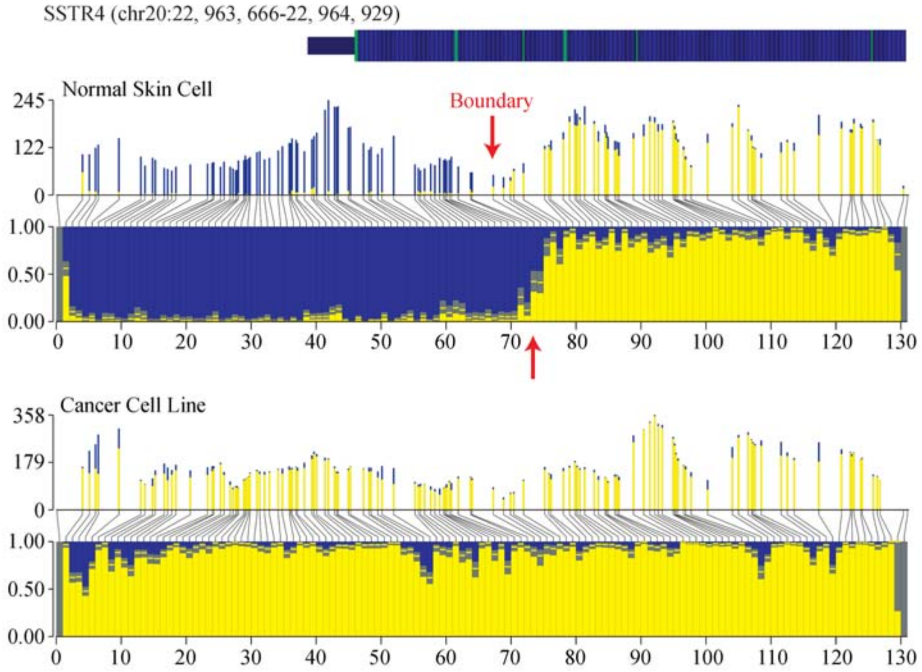

Fig.1. Methylation profile from bisulfite sequencing through a CpG island covering the TSS and first exon of the SSTR4 gene in the human assembly hg18. Two profiles are presented: one for a normal skin cell line and the other for a breast cancer cell line. Each plot shows the counts of methylated (yellow) and unmethylated (blue) reads mapping over each CpG (top; normalized frequencies appear below, where grey indicates 90% confidence interval). A distinct boundary is evident in the normal skin cell line, but this boundary appears to have been lost in the cancer cell line.

was the need for read mapping algorithms capable of dealing with the $C \rightarrow T$ conversion. To identify the genomic origin of reads produced in bisulfite sequencing experiments requires mapping those reads back to a reference genome without penalizing when a T in a read aligns over a genomic C. This relation is not symmetric: Cs observed in reads must have originated from a genomic C, and therefore should be penalized when mapping over a genomic T. The read mapping problem in general is a difficult approximate string matching problem[24], and the T/C wild-card required of bisulfite sequencing makes the problem substantially more difficult. One approach is to convert all Cs in reads and all Cs in the reference genome to T, and then to map all reads using only a 3-letter alphabet[25]. The first problem with the 3-letter approach is that the loss of information makes most mapping approaches slower. Mapping algorithms often require that some positions in the reads match the reference genome exactly in order for a full comparison to be computed between the read and the corresponding position in the reference genome. Efficiency of this approach depends on the complexity of the underlying sequence: lower complexity increases the number of exact matches during the initial stage because a random position is more likely to match.

The second problem is that any additional mapping specificity in those remaining Cs will not be used if the 3-letter alphabet is assumed. Downstream analysis in bisulfite sequencing experiments usually requires that reads are mapped to unique locations in the genome. Substantial portions of mammalian genomes contain sequences that are exactly repeated at some other location in the genome, and reads mapping to any copy of such a repeat will necessarily map equally well to all copies. Those reads are called ambiguously mapping, and cannot usually be used in bisulfite sequencing experiments. This problem is made much worse by the loss of complexity resulting from bisulfite treatment, and any cytosines that survived the bisulfite treatment can be used to resolve ambiguities in the mapping locations of reads. One subtle consideration is in using unconverted cytosines at C positions to assist in mapping. If reads indicating a greater number of methylated CpGs (i.e., having more Cs at positions corresponding to CpG dinucleotides) are more likely to be mapped uniquely, then estimating the frequency of methylation based on mapped reads becomes problematic. For mapping mCpG, one solution is to allow Cs in reads to map over Ts in the genome only if the genomic T is part of a TpG dinucleotide. With each additional wild-card rule, the mapping process becomes more complex and inevitably less efficient.

In addition to the T/C wildcard, certain protocols will also have sequencing adaptors attached to strands complementary to those that underwent bisulfite treatment[26-27], for example if adaptors are added following bisulfite treatment. Following bisulfite treatment, use of PCR amplification results in one strand of DNA having those Gs complementary to unmethylated Cs converted into As. In these cases, each read will either require mapping while assuming the T/C wild-card or and A/G wild-card — with these two cases being exclusive for a given read.

We developed the RMAPBS algorithm[28] to implement the wild-card matching and to leverage randomly unconverted cytosines while not biasing mappability towards a particular methylation state. At present these wild-card mapping strategies have yet to be implemented using the most efficient mapping strategies based on highly-compressed representations for the reference genome[29-30].

Non-CpG methylation has also been reported in human cells in a recent Human Reference Epigenome Mapping effort, it tends to occur in stem cells and in non-genic regions, especially in many enhancer regions[31]. Mapping such non-CpG methylation is more challenging both because of less read depth or coverage and because of more difficulties in computational techniques caused by the loss of denucleotide constraint.

5 Determining Methylation State

Since most of methylation in mammals occur in CpG dinucleotides, we will mainly discuss mCpG detection. At any given instant, an individual cytosine will either be methylated or not, and at the level of individual DNA molecules methylation of a cytosine is a Boolean variable. Under some extremely simplifying assumptions, the most basic question we can ask from DNA methylation data is whether or not a particular CpG is methylated. If we are working with bisulfite sequencing data, then a methylated CpG would be identified with unconverted cytosines in reads mapping over that position. The reality is not quite so simple, and noise in the data (mapping errors, sequencing errors, incomplete bisulfite conversion) can introduce errors, so that a methylated cytosine may have both C's and T's mapping over it. Naturally we would conclude that a particular CpG is methylated if a great majority of the relevant reads indicate methylation. We then require some method of determining when sufficient information is available (i.e., enough informative reads) to make a conclusion about whether or not a CpG is methylated. The statistical problem of determining whether or not a CpG is methylated, when there might be noise in the system, is similar to the problem of identifying single nucleotide polymorphisms (SNPs). In reality it will not often make sense to assume that cytosine methylation is a Boolean variable: the biological sample being studied may contain some intermediate proportion of molecules with methylation at a particular CpG, and the appropriate questions concern estimates of this proportion and our confidence in those estimates. When a sufficient number of reads is available, we may assume that the methylation states observed in those reads, for any given CpG, follow a binomial distribution. Confidence intervals on the binomial can then be used to estimate our confidence in the methylation proportion at any given CpG[32].

There remain substantial challenges associated with this assignment of methylation levels to either individual CpGs or genomic regions. In the near-term, genome-wide DNA methylation projects in mammals will not produce enough data for simple methods (such as Binomial confidence intervals mentioned above) to produce accurate results. To get the most useful information from low-coverage methylation profiles will require more sophisticated models that can be trained on a few high-coverage methylomes. We know that CpG methylation is generally autocorrelated along genomes[33], and we also know that sequence context influences CpG methylation[34]. Such prior information can be incorporated into models, for example as priors in Bayesian models, and can be leveraged to help understand lower-coverage methylomes. In general, any methods that can be applied to infer methylation status from less data can help drive down the cost of experiments and are highly valuable.

6 Features of Methylation Profiles

As indicated in Section 2, in mammals DNA methylation accumulates *de novo* beginning at the earliest stages of development, and continues through adulthood. Certain genomic regions are differentially methylated between tissues, which suggests some information encoded in the genome to guide methylation to those loci. A fundamental biological problem is to identify those factors controlling methylation in these differentially methylated regions (DMRs). Taking the critical steps towards unlocking these mysteries first requires accurate computational methods for identifying those regions that are differentially methylated between two datasets.

It is also well known that DNA methylation spreads along the genome, which has been attributed to the processivity of the methyltransferase enzymes[35]. Without control on this spreading behavior, methylation might continue through regions for which proper function depends on lack of methylation. Fortunately methylation seems not to spread past particular regions which behave as *boundaries*. Currently there is very little understanding of these boundaries: we know very little about where these boundaries are, let alone the mechanisms of how they are specified and maintained. Current experimental projects to produce full-genome methylation profiles will eventually provide the data needed for detailed examinations of methylation boundaries. However, making use of this data will require computational techniques for identifying those precise genomic loci where methylation levels change.

In the simple case, the task is to segment the genome into two kinds of regions: hypomethylated and hypermethylated. Consider the data as a sequence of methylation measurements associated with individual CpGs. Examples of important considerations include: How to model the sizes of the segments? What are the appropriate statistical models for describing the data (e.g., read depths at individual CpGs)? How should we measure confidence in the identification of a boundary between consecutive segments? More generally, the problem is to identify those contiguous genomic regions that appear to have some homogeneity of DMA methylation. Genomic segmentation problems have arisen previously in several contexts, for example in studying copy number variation[36]. The demands of this particular computational problem in the context of identifying methylation boundaries are no greater than for studies of copy number variation, and in fact both such tasks will require more powerful methods to fully leverage the data when the underlying technology is second-generation sequencing. Although these analysis problems will be of particular interest to the signal processing community, efficiency concerns are also present and attention must also be paid to computational efficiency of these analyses.

7 Methylated Sequence Analysis

Methylated cytosine in mammals has been referred to as the "5th base"[37-38], and clearly this 5th base can encode distinct information from unmethylated cytosine, suggesting that in some contexts we should treat these two as distinct letters in our sequence alphabet. When one considers that much differential methylation, particularly in mammals, occurs in regulatory regions such as promoters, the 5-letter alphabet becomes even more important. It has been found that individual methylated cytosines can affect binding of transcription factors[39]. The transcription factor CTCF has been found to functionally bind at sites lacking DNA methylation, but methylation through the site prevents CTCF binding[40]. In addition, the deeply conserved MBD family of proteins bind to specific DNA sequences that include methylated CpGs[41-42]. There is evidence that MBD proteins have functions related to the regulation of chromatin, hence, transcription.

The problem of discovering transcription factor binding sites (often called motif discovery) has received much attention from both practical[43] and theoretical perspectives[44]. The abstract task is, given a set of sequences, to identify a sequence pattern that is degenerate and that is enriched in the set of sequences vs. some control set. From the perspective of algorithmic complexity, there is no difference between an alphabet of 4-letter alphabet and with 5 letters, similar to the well studied task of predicting functional transcription factor binding sites. A substantial complication arises, however, when one considers that the methylation data will not generally be discrete, but will instead provide a methylation level associated with individual cytosines.

Modern genome-wide DNAseI hyper-sensitive site mapping data can be valuable for both types of analyses mentioned above. Although TFs have been thought to provide sequence specificity in initial setup of DNA methylation, more evidences have also pointed to ncRNAs that can also provide such roles, especially in initiation of silencing in imprinted regions.

8 Understanding Heterogenous Samples

Interesting computational problems arise from experiments where the data may contain multiple DNA methylation through individual genomic regions. The simplest case is allelic DNA methylation, where two distinct methylation patterns are present through a given region, even in data produced from homogenous cell samples. Diploid cells have two copies of each chromosome, and therefore possibly two distinct alleles for each gene. In many cases DNA methylation has been associated with the silencing of one allele, and this is the major mechanisms of silencing an X chromosome in females having two X chromosomes. This is also the only mechanism yet identified in genomic imprinting. How would data look in the case of allelic methylation? For a profile of methylation frequencies at individual CpGs, we expect to observe roughly 50% methylation through a region having allelic methylation. However, simply observing 50% methylation is not sufficient for a conclusion of two "epi-alleles." Depending on the resolution of the data (related to the amount of data available and the resolution of the experimental technology) we might be observing 50% methylation as a result of 50% of CpGs being fully methylated in both alleles. The information required to distinguish these two possibilities is the relationships between methylation states of multiple CpGs in the same molecule. The individual reads from bisulfite sequencing experiments can provide this information over short distances — a limitation associated with read lengths — but it remains unclear how best to resolve these two scenarios. One may also use SNPs to identify alleles.

A more complex situation can arise when we profile DNA methylation in heterogeneous samples. It is well known that tumors consist of many cell types, with some exhibiting normal and healthy phenotypes, but many cells in the tumor will have highly unusual and aggressive phenotypes[45]. The epigenomic features of these cells are expected to differ significantly. One major challenge in cancer research is to understand the phenotypic composition of tumor samples according to the epigenomics of each phenotype.

Given a dataset originating from a mixture of cell types, the fundamental computational problem is to 1) identify the methylation characteristics of the individual cell types, and 2) quantify the relative frequencies of those cell types. This problem is analogous to a problem faced in metagenomic sequence assembly[46], where DNA is sequenced from heterogeneous populations of microorganisms and all sequenced DNA can be mapped to a single reference, presumed to be sufficiently close to each organism. The task is to identify the individual species present. This computational problem is reminiscent of sequence assembly, with two major differences: 1) the order and orientation of fragments are known, and 2) rather than a single sequence to assemble, we must assemble several.

To obtain a more formal abstraction, let $X = \{x_1, \ldots, x_n\}$ be the set of methylation states in a set of n reads from a bisulfite sequencing experiment. We assume that $|x_i| = w$ for all i, indicating that each read maps over the same number of CpGs (not generally a realistic assumption). We also assume that the reads have been mapped to a reference genome, so we have a mapping function $p : X \mapsto \{1, \ldots, m - w + 1\}$, such that $p(x_i) = j$ indicates that read i is mapped over CpGs beginning with the j-th CpG, where m is the number of total CpGs. The most basic computational problem asks, for a given k, does there exist a set $\mathcal{S} = \{S_1, \ldots, S_k\} \subseteq \{0, 1\}^m$ of strings such that for each $x \in X$, there exists some $S \in \mathcal{S}$ such that

$$x(j) = S(p(x) + j - 1)$$

for $1 \leqslant j \leqslant w$, where $x(j)$ indicates the j-th position in the binary string x and similarly for S. This algorithmic problem can be solved through an elegant transformation to the problem of partition a poset into chains, and a further transformation to bipartite graph matching[47-48]. Variants of this problem arise when errors are allowed (i.e., some mismatches are permitted between x_i and members of $\mathcal{S}$), when cast as an optimization problem under various objectives and when different aspects of the underlying experiments are considered (e.g., the use of paired-end reads). Any of these problem variants may emerge as important in different contexts — both in the analysis of methylation data and other “meta-assembly” problems.

9 DNA Methylation in Regulatory Networks

In silico construction of regulatory networks is a field still in its infancy, and there are many challenges remaining even when the networks are restricted to traditional forms of data[49-50]. Because DNA methylation functions so frequently to regulate transcription it seems natural to include this information in transcriptional regulatory networks. Research into regulatory networks has mainly focused on transcriptional regulation, and is generally based on gene expression data[51] and often uses information about transcription factor localization[52]. The goal is to identify transcriptional regulatory relationships between genes, where the regulators are transcription factors, and when possible we try to identify direct relationships indicating a physical interaction (e.g., a transcription factor binding at a promoter of the target gene). Other forms of regulation have been included, for example at the posttranscriptional or proteomic levels[53]. There remains essentially no research directed towards how best to include epigenomic information, either for histone modifications/variants or DNA methylation. We remark that the methods for incorporating DNA methylation data into regulatory network construction will likely lead to methods that can also be applied to histone data.

Let us first consider a possible scenario where the absence of epigenomic information in regulatory networks can lead to significant problems. Suppose the network construction effort has available all relevant expression data, and also has information about the presence of functional transcription factor binding sites in regulatory regions (including which TFs can bind at each site). If a particular region is not accessible for binding by transcription factors, then the regulatory interaction will not take place, so clearly heterochromatic silencing of genes mediated by epigenomic modifications is an essential piece of information. One might argue that epigenomic state is controlled by DNA binding proteins, the expression of which could suggest epigenomic state if the function of those proteins is sufficiently well understood. Such reasoning is false because of the critical fact that epigenomic information is mitotically heritable (and with a known mechanism in the case of DNA methylation; see Section 2). The genes responsible for regulating the epigenomic state at any locus may no longer be expressed.

To integrate epigenomic information into regulatory networks the first step is identifying genomic regions with functional differential methylation. We discussed this in Section 6. Next these differentially methylated regions must be associated with specific genes. This association is not likely a trivial task, and will almost certainly require efforts to model expression based on features of both sequence and epigenomic state of promoters and enhancers. A first step here would build on work modeling expression using genomic elements[54-56]: if differential methylation through a region can predict differential expression of a gene, then we may assume that region is relevant to the gene. It is also clear from studies of imprinting that methylation through individual regions can have regulatory influence over several genes[57]. We must also determine the direction of regulation by differential methylation. Although methylation through promoters is generally associated with silencing of the associated genes, methylation through more distal regulatory regions can have either activating or repressing effects by blocking the activity of either repressing or activating transcription factors.

In terms of representation and visualization of DMRs in regulatory networks, it might be sufficient to simply annotate genes in a way that indicates their epigenomic regulation. From a different perspective, one may include DMRs as primary elements (nodes) in regulatory networks, which make sense as these elements can regulate and be regulated, and individual DMRs can influence multiple target genes.

10 DNA Methylation and Somatic Trees

Computational problems associated with evolutionary trees have attracted attention from computational biologists for decades[58-60], due as much to their mathematical appeal and challenges as to their biological importance. Just as evolutionary trees relate species according to common descent and ancestral relationships, developmental trees can relate the cells in a multicellular organism. The developmental tree is rooted at some omnipotent stem cell (e.g., the zygote in case of sexual reproduction), and terminally differentiated cells form the leaves. Unlike evolutionary trees, developmental trees are generally predetermined: while many somatic changes do happen randomly, most important changes along any tree branch are precisely regulated. Epigenomic modifications play a large role in development, but most details remain poorly understood. As DNA methylation becomes easier to profile, it will be possible to examine the role of methylation in guiding development, but doing so will depend on computational methods for processing epigenomic marks (including DNA methylation) relative to the structures of developmental trees.

The initial analysis question in understanding methylation changes through somatic trees is to identify the DMRs, but computational methods are required to answer much more sophisticated biological questions. For example, which features of DNA methylation through individual loci appear to be “cladistic” in developmental trees? Such features may indicate modes of regulation that can stably restrict phenotypes to a particular lineage, but to identify cladistic features requires measures for cladisticity and fast algorithms for testing this property. When multiple cells along a single lineage are available, is it possible to identify progressive changes? Evidence exists for “lineage priming”, the promiscuous expression of lineage specific genes in uncommitted progenitors prior to their expression at lineage specific levels[61], and highly-sensitive methods for identifying subtle yet progressive changes in methylation along individual lineages. This knowledge will also be crucial for understanding iPS cell reprogramming and fate determination.

Computation involving DNA methylation changes through somatic trees has recently been used for the highly-specific problem of tracing the clonal evolution of tumors and stem cell populations[62-63]. Aberrant DNA methylation is a general feature of cancer genomes, with methylation being disrupted early in tumorigenesis and contributing to tumor progression. Cancer begins with the transformation of a single cell, and tumors grow as clonal expansions originating from the transformed cell. These clonal cell populations are related by an ancestral tree, and the random changes that occur at genomic regions where DNA methylation is unregulated can be used as molecular clocks. Using DNA methylation patterns through these randomly methylated loci, it is possible to infer ancestral relationships for cells sampled at distinct physical regions of tumors. Existing studies of methylation dynamics during clonal evolution have yet to fully leverage second-generation sequencing technology, and doing so will undoubtedly require novel methods that can integrate data from multiple loci.

Analysis of epigenomic data relative to somatic lineage trees (whether known or inferred) presents multiple challenges and opportunities for computational biologists. As experimental methods for profiling DNA methylation and other epigenomic marks mature, data analysis problems will place computational biology at the center of elucidating the epigenomics of development and differentiation.

11 Conclusion

Understanding the functions of DNA methylation and other epigenomic marks presents many opportunities for computational biologists. These opportunities range from developing computational technology to address fundamental problems of massive datasets to the design of higher-level analysis methods for sophisticated *in silico* experiments. As genome-scale experimental technologies for DNA methylation become increasingly accessible, the critical steps in major discoveries will gradually shift towards data analysis. The result is a more central role for computational scientists in these projects. A deep understanding of both the relevant biological questions and experimental technologies will enable computer scientists to identify exciting and challenging analytical problems as they emerge from the field of epigenomics.

References

[1] Schwartz D C, Waterman M S. New generations: Sequencing machines and their computational challenges. *J. Comput. Sci. & Technol.*, 2010, 25(1): 3-9.

[2] Holliday R, Pugh J E. DNA modification mechanisms and gene activity during development. *Science*, 1975, 187(4173): 226-232.

[3] Riggs A. X inactivation, differentiation, and DNA methylation. *Cytogenet. Cell. Genet.*, 1975, 14(1): 9-25.

[4] Bird A. DNA methylation patterns and epigenetic memory. *Genes & Development*, 2002, 16(1): 6-21.

[5] Bestor T H. The DNA methyltransferases of mammals. *Human Molecular Genetics*, 2000, 9(16): 2395-2402.

[6] Yoder J A, Walsh C P, Bestor T H. Cytosine methylation and the ecology of intragenomic parasites. *Trends in Genetics*, 1997, 13(8): 335-340.

[7] Bestor T H. Cytosine methylation mediates sexual conflict. *Trends in Genetics*, 2003, 19(4): 185-190.

[8] Gonzalgo M L, Jones P A. Rapid quantitation of methylation differences at specific sites using methylation-sensitive single nucleotide primer extension (Ms-SNuPE). *Nucleic Acids Research*, 1997, 25(12): 2529-2531.

[9] Simmen M W. Genome-scale relationships between cytosine methylation and dinucleotide abundances in animals. *Genomics*, 2008, 92(1): 33-40.

[10] Cooper D N, Youssoufian H. The CpG dinucleotide and human genetic disease. *Human Genetics*, 1988, 78(2): 151-155.

[11] Jiang C, Zhao Z. Mutational spectrum in the recent human genome inferred by single nucleotide polymorphisms. *Genomics*, 2006, 88(5): 527-534.

[12] Wood L D, Parsons D W, Jones S *et al.* The genomic landscapes of human breast and colorectal cancers. *Science*, 2007, 318(5853): 1108-1113.

[13] Human Epigenome Consortium. http://www.epigenome.org/, Accessed Sept. 16, 2009.

[14] Epigenomics — Overview. Division of Program Coordination, Planning, and Strategic Initiatives, National Institutes of Healt. http://nihroadmap.nih.gov/epigenomics/, Accessed Sept. 16, 2009.

[15] Raleigh E A. Organization and function of the mcrBC genes of Escherichia coli K-12. *Molecular Microbiology*, 1992, 6(9): 1079-1086.

[16] Bird A P. Use of restriction enzymes to study eukaryotic DNA methylation: II. The symmetry of methylated sites supports semi-conservative copying of the methylation pattern. *Journal of Molecular Biology*, 1978, 118(1): 49-60.

[17] Gruenbaum Y, Cedar H, Razin A. Restriction enzyme digestion of hemimethylated DNA. *Nucl. Acids Res.*, 1981, 9(11): 2509-2515.

[18] Lippman Z, Gendrel A V, Colot V, Martienssen R. Profiling DNA methylation patterns using genomic tiling microarrays. *Nature Methods*, 2005, 2(3): 219-224.

[19] Weber M, Davies J J, Wittig D, Oakeley E J, Haase M, Lam W L, Schubeler D. Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells. *Nat. Genet.*, 2005, 37(8): 853-862.

[20] Down T A, Rakyan V K, Turner D J *et al.* A Bayesian deconvolution strategy for immunoprecipitation-based DNA methylome analysis. *Nat. Biotech.*, 2008, 26(7): 779-785.

[21] Xiong Z, Laird P W. COBRA: A sensitive and quantitative DNA methylation assay. *Nucleic Acids Research*, 1997, 25(12): 2532-2534.

[22] Zhou D, Qiao W, Yang L, Lu Z. Bisulfite-modified target DNA array for aberrant methylation analysis. *Analytical Biochemistry*, 2006, 351(1): 26-35.

[23] Ehrich M, Nelson M R, Stanssens P *et al.* Quantitative high-throughput analysis of DNA methylation patterns by base-specific cleavage and mass spectrometry. *Proc. Natl. Acad. Sci. USA*, 2005, 102(44): 15785-15790.

[24] Smith A D, Xuan Z, Zhang M Q. Using quality scores and longer reads improves accuracy of Solexa read mapping. *BMC Bioinformatics*, 2008, 9: 128.

[25] Meissner A, Mikkelsen T S, Gu H *et al.* Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature*, 2008, 475(7205): 766-770.

[26] Ball M P, Li J B, Gao Y *et al.* Targeted and genome-scale strategies reveal gene-body methylation signatures in human cells. *Nature Biotechnology*, 2009, 27(4): 361-368.

[27] Deng J, Shoemaker R, Xie B *et al.* Targeted bisulfite sequencing reveals changes in DNA methylation associated with nuclear reprogramming. *Nature Biotechnology*, 2009, 27(4): 353-360.

[28] Smith A D, Chung W, Hodges E, Kendall J, Hannon G, Hicks J, Xuan Z, Zhang M Q. Updates to the RMAP short-read mapping software. *Bioinformatics*, 2009, 25(21): 2841-2842.

[29] Li R, Li Y, Kristiansen K, Wang J. Soap: Short oligonucleotide alignment program. *Bioinformatics*, 2008, 24(5): 713-714.

[30] Langmead B, Trapnell C, Pop M, Salzberg S. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, 2009, 10(3): R25.

[31] Lister R, Ecker J, Ren B. 2009. (Personal Communication)

[32] Hodges E, Smith A D, Kendall J *et al.* High definition profiling of mammalian DNA methylation by array capture and single molecule bisulfite sequencing. *Genome Research*, 2009, 19(9): 1593-1605.

[33] Eckhardt F, Lewin J, Cortese R *et al.* DNA methylation profiling of human chromosomes 6, 20 and 22. *Nat. Genet.*, 2006, 38(12): 1378-1385.

[34] Das R, Dimitrova N, Xuan Z *et al.* Computational prediction of methylation status in human genomic sequences. *Proc. Natl. Acad. Sci. USA*, 2006, 103(28): 10713-10716.

[35] Vilkaitis G, Suetake I, Klimasauskas S, Tajima S. Processive methylation of hemimethylated CpG sites by mouse Dnmt1 DNA methyltransferase. *J. Biol. Chem.*, 2005, 280(1): 64-72.

[36] Sebat J, Lakshmi B, Troge J *et al.* Large-scale copy number polymorphism in the human genome. *Science*, 2004, 305(5683): 525-528.

[37] Model F, Adorjan P, Olek A, Piepenbrock C. Feature selection for DNA methylation based cancer classification. *Bioinformatics*, 2001, 17(Suppl. 1): S157-S164.

[38] Lister R, Ecker J R. Finding the fifth base: Genome-wide sequencing of cytosine methylation. *Genome Research*, 2009, 19(6): 959-968.

[39] Watt F, Molloy P L. Cytosine methylation prevents binding to DNA of a HeLa cell transcription factor required for optimal expression of the adenovirus major late promoter. *Genes & Development*, 1988, 2(9): 1136-1143.

[40] Bell A C, Felsenfeld G. Methylation of a CTCF-dependent boundary controls imprinted expression of the Igf 2 gene. *Nature*, 2000, 405(6785): 482-485.

[41] Lewis J D, Meehan R R, Henzel W J *et al.* Purification, sequence, and cellular localization of a novel chromosomal protein that binds to Methylated DNA. *Cell*, 1992, 69(6): 905-914.

[42] Klose R J, Sarraf S A, Schmiedeberg L, McDermott S M, Stancheva I, Bird A P. DNA binding selectivity of MeCP2 due to a requirement for A/T sequences adjacent to Methyl-CpG. *Molecular Cell*, 2005, 19(5): 667-678.

[43] Tompa M, Li N, Bailey T L *et al.* Assessing computational tools for the discovery of transcription factor binding sites. *Nat. Biotechnol.*, 2005, 23(1): 137-144.

[44] Li M, Ma B, Wang L. On the closest string and substring problems. *Journal of the ACM*, 2002, 49(2): 157-171.

[45] Reya T, Morrison S J, Clarke M F, Weissman I L. Stem cells, cancer, and cancer stem cells. *Nature*, 2001, 414(6859): 105-111.

[46] Riesenfeld C S, Schloss P D, Handelsman J. Metagenomics: Genomic analysis of microbial communities. *Annu. Rev. Genet.*, 2004, 38: 525-552.

[47] Ford L, Fulkerson D. Flows in Networks. Princeton University Press, 1962.

[48] Eriksson N, Pachter L, Mitsuya Y *et al.* Viral population estimation using pyrosequencing. *PLoS Comput. Biol.*, May 2008, 4(5): e1000074.

[49] Babu M M, Lang B, Aravind L. Methods to reconstruct and compare transcriptional regulatory networks. *Methods Mol. Biol.*, 2009, 541: 163-180.

[50] Hecker M, Lambeck S, Toepfer S, van Someren E, Guthke R. Gene regulatory network inference: Data integration in dynamic models — A review. *Biosystems*, 2009, 96(1): 86-103.

[51] Bar-Joseph Z, Gerber G K, Lee T I *et al.* Computational discovery of gene modules and regulatory networks. *Nature Biotechnology*, 2003, 21(11): 1337-1342.

[52] Lee T I, Rinaldi N J, Robert F *et al.* Transcriptional regulatory networks in saccharomyces cerevisiae. *Science*, 2002, 298(5594): 799-804.

[53] Schwikowski B, Uetz P, Fields S. A network of protein-protein interactions in yeast. *Nature Biotechnology*, 2000, 18(12): 1257-1261.

[54] Beer M A, Tavazoie S. Predicting gene expression from sequence. *Cell*, 2004, 117(2): 185-198.

[55] Smith A D, Sumazin P, Xuan Z, Zhang M Q. DNA motifs in human and mouse proximal promoters predict tissue-specific expression. *Proc. Natl. Acad. Sci. USA*, 2006, 103(16): 6275-6280.

[56] Pennacchio L A, Loots G G, Nobrega M A, Ovcharenko I. Predicting tissue-specific enhancers in the human genome. *Genome Research*, 2007, 17(2): 201-211.

[57] Verona R I, Mann M R W, Bartolomei M S. Genomic imprinting: Intricacies of epigenetic regulation in clusters. *Annual Review of Cell and Developmental Biology*, 2003, 19(1): 237-259.

[58] Felsenstein J. Evolutionary trees from DNA sequences: A maximum likelihood approach. *Journal of Molecular Evolution*, 1981, 17(6): 368-376.

[59] Sankoff D. Computational complexity of inferring phylogenies by compatibility. *Systematic Zoology*, 1986, 35(2): 224-229.

[60] Gusfield D. Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology. Cambridge University Press, 1997.

[61] Miyamoto T, Iwasaki H., Reizis B, Ye M, Graf T, Weissman I L, Akashi K. Myeloid or lymphoid promiscuity as a critical step in hematopoietic lineage commitment. *Developmental Cell*, 2002, 3(1): 137-147.

[62] Yatabe Y, Tavaré S, Shibata D. Investigating stem cells in human colon by using methylation patterns. *Proc. Natl. Acad. Sci. USA*, 2001, 98(19): 10839-10844.

[63] Kim J Y, Tavaré S, Shibata D. Counting human somatic cell replications: Methylation mirrors endometrial stem cell divisions. *Proc. Natl. Acad. Sci. USA*, 2005, 102(49): 17739-17744.

Michael Q. Zhang obtained the B.S. degree in mech. eng. from Univ. Sci. & Tech. China in 1981 and Ph.D. degree in physics from Rutgers University in 1987. He studied statistical mechanics and integrable systems as a postdoctoral fellow at Courant Institute of Mathematical Sciences, NYU for three years and then moved to Cold Spring Harbor Laboratory for twenty years. He is now a professor at Watson School of Biological Sciences at Cold Spring Harbor Laboratory in New York. He has also been a guest professor at Tsinghua University in Beijing, China since 2003. He has also been an adjunct professor at Stony Brook University since 1997. He has associated with the editorial board for Nucleic Acids Research, Bioinformatics, BMC Journals, etc. and served as chairman/section chair or program committee member for CSHL Meetings, ISMB, RECOMB, APBC, etc. Dr. Zhang is one of the pioneers in human genome research and made important contributions to computational genomics and epigenomics.

Andrew D. Smith received the B.A. degree in psychology and the B.C.S. degree (Bachelor of Computer Science) in 2000 and the Ph.D. degree in computer science from University of New Brunswich in 2004. Dr. Smith studied computational biology and genomics at Cold Spring Harbor Laboratory until 2008 at which time he moved to University of Southern California where he is currently assistant professor of biological sciences.