Hereditary Genetics

Su et al., Hereditary Genet 2012, 1:2
DOI: 10.4172/2161-1041.1000107

Review Article Open Access

Advances in Bioinformatics Tools for High-Throughput Sequencing Data of DNA Methylation

Jianzhong Su[1]*, Dan Huang[2], Haidan Yan[2], Hongbo Liu[2] and Yan Zhang[2]

[1]*Academy of Fundamental and Interdisciplinary Science, Harbin Institute of Technology, Harbin 150080, China*
[2]*College of Bioinformatics Science and Technology, Harbin Medical University, Harbin 150081, China*

Abstract

DNA methylation plays crucial roles in regulating gene expression during cellular development and differentiation. Recently, Next-generation sequencing (NGS) technologies spurred a revolution in investigating global DNA methylation profiles. Analysis of DNA methylation patterns on a genome-wide scale is essential to understanding the underlying mechanisms of DNA methylation. Here, we reviewed several next-generation sequencing techniques coupled with different pretreatment methods (endonuclease digestion, affinity enrichment and bisulfite conversion) and summarized the relative bioinformatics tools and resources for further analysis of DNA methylation.

Keywords: DNA methylation; Next-generation sequencing technologies; Bioinformatics tools

Introduction

Epigenetics, the study of mitotically or meiotically heritable regulatory changes in gene function that occur without changing the DNA sequence, has flourished in recent years. One of the most well-studied epigenetic phenomenon is DNA methylation that occurs mainly at the 5' position of cytosine in the sequence context CpG, CpHpG and CpHpH where H is A, T or C and has been widely observed in animals, plants and fungi [1]. Methylation patterns are constantly changing over evolutionary time [2]. The most common pattern in invertebrate animals is 'mosaic methylation', whereby stable methylated domains are interspersed with methylation-free regions. In contrast, vertebrate genomes are globally methylated, with the exception of CpG islands (genomic regions of CpGs enrichment) [3]. In comparison with the relative stability of genomic DNA sequences, DNA methylomes dynamically change among different cells and even vary along with the change of conditions in a single cell [4].

As an important epigenetic mark involved in a diverse range of biological processes, DNA methylation is well-known for its roles in stable transcriptional gene silencing, X inactivation [5] and genomic imprinting [6]. Current evidence indicates that DNA methylation also plays a significant role in maintaining cellular function and development of autoimmunity and aging [7,8]. Aberrant DNA methylation may be associated with the disorder of gene expression in carcinogenesis [9]. What's more, DNA methylation is adapted for a specific cellular memory function in development supported by the heritability and the secondary nature of DNA methylation states [2]. Therefore, it is of great significance to study the biological function of DNA methylation and its underlying mechanism in various organisms. Fortunately, the development of sequencing technology makes it easier to measure the genome-wide DNA methylation profiling, which is a premise of understanding the role of methylation in development and disease. This review mainly focuses on bioinformatics tools for processing and analysis of high-throughput DNA methylation data generated by the next-generation sequencing technologies (Figure 1).

NGS-based technologies for detecting DNA methylation

Next-generation sequencing (NGS), the latest and most promising methodology for genome-wide analysis of DNA methylation, may be used as an alternative for DNA methylation analysis. Very large amounts of sequence information produced by NGS provide a quantitative measure of DNA methylation abundance. Meanwhile, sequencing-based analysis will increase the efficiency and resolution of the detection of DNA methylation as it may use less input DNA to obtain high coverage sequencing of whole genomes and avoid biases that affect hybridization such as sequence composition [1,10]. What's more, the ability to interrogate CpGs in repetitive elements gives the sequencing-based approaches a distinct advantage over microarrays [11]. Nearly all technologies of sequence-specific DNA methylation analysis are based on one of three main pretreatment approaches of DNA samples: endonuclease digestion, affinity enrichment and bisulfite conversion. The different combination of pretreatment methods and subsequent molecular biology techniques, such as DNA microarrays and high-throughput sequencing, generates a plethora of techniques for mapping DNA methylation feasible on a genome-wide scale. In the following section, we reviewed the various NGS-based technologies for detecting DNA methylation, as summarized in Table 1.

Endonuclease digestion

Methylation sensitive restriction endonucleases (HpaII, MspI and HhaI) and their corresponding isoenzymes (not sensitive to methylation) are widely used to distinguish methylated from unmethylated cytosines [10]. Over the past decade, several methyl-sensitive restriction based methods have been developed. By using NGS to analyze the output of the HELP assay (HpaII-tiny fragment enrichment by Ligation-mediated PCR), HELP-seq [12], is more sensitive than array-based HELP in identifying hypomethylated loci. Methyl-sensitive cut counting (MSCC) is a cost-effective approach to detect unmethylated CpGs at single-base resolution using a flanking cut with a type-IIs restriction enzyme (MmeI) and adaptor ligation after HpaII digestion [13-15]. Another method, Methyl-seq, may sequence the fragments digested by HpaII or MspI other than randomly sheared fragments [16].

Affinity enrichment

Affinity purification of methylated DNA, the most recent and simplest way to enrich methylated DNA , was first proved with the

***Corresponding author:** Jianzhong Su, Academy of Fundamental and Interdisciplinary Science, Harbin Institute of Technology, Harbin 150080, China, E-mail: jianzhongsu@yahoo.cn or yanyou1225@yahoo.com.cn

Received February 27, 2012; **Accepted** April 22, 2012; **Published** April 26, 2012

Citation: Su J, Huang D, Yan H, Liu H, Zhang Y (2012) Advances in Bioinformatics Tools for High-Throughput Sequencing Data of DNA Methylation. Hereditary Genet 1:107. doi:10.4172/2161-1041.1000107

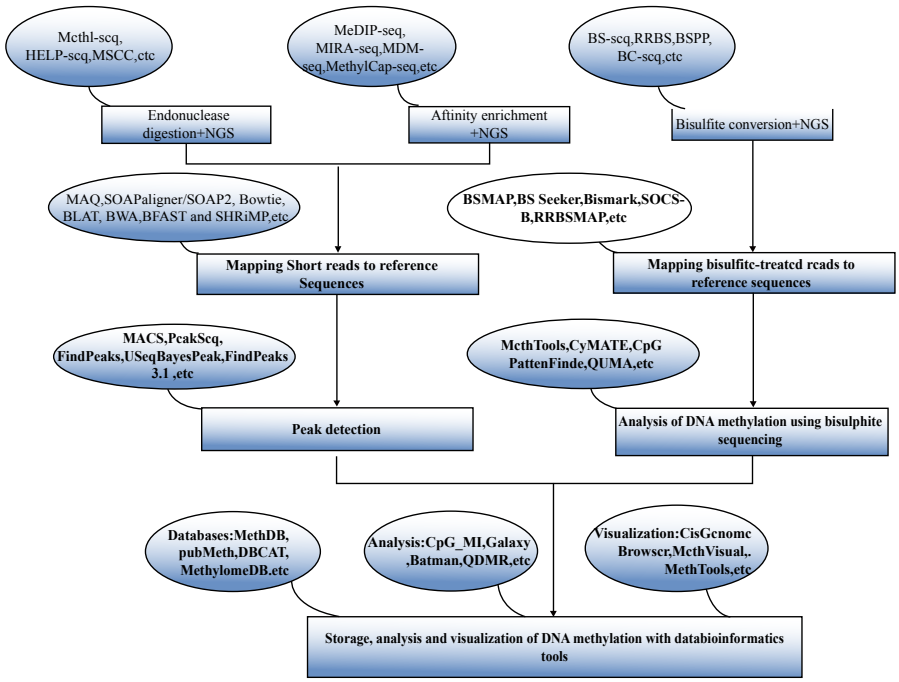

Figure 1: Workflow of high-throughput DNA methylation analysis. The workflow shows the basic process of high-throughput DNA methylation analysis by existing bioinformatics tools. The bioinformatics tools are shown in the ellipses and the corresponding functions in the neighboring boxs.

Pretreatment method	Genome coverage	NGS-based analysis	Application	Ref
Endonuclease digestion	Moderate	Methl-seq	Assay a range of genomic elements; allowing a broader survey of regions than classic methylation studies limited to CpG islands and promoters	[16]
		HELP-seq	Measurement of repetitive sequences, copy-number variability, allele-specific and smaller fragments (<50bp) ;Sensitivity of detection of hypomethylated loci	[12, 75]
		MSCC	Identification of the unmethylated region of a genome by pinpointing unmethylated CpGs at single base-pair resolution	[13, 14]
Affinity enrichment	Moderate	MeDIP-seq	Generation of unbiased, cost-effective, and full-genome methylation levels without the limitations of restriction sites or CpG islands;	[1, 76]
		MIRA-seq	Analyze recovered or double-stranded methylated DNA on a genome-wide scale; Applicable to various clinical and diagnostic situations.	[19]
		MDB-seq	Applied to any biological settings to identify differentially methylated regions at the genomic scale	[18]
		MethylCap-seq	Detection of differentially methylated regions with high genome coverage; Detect DMRs in clinical samples	[20, 77]
Bisulfite conversion	High	BS-seq	Sensitively measure cytosine methylation on a genome-wide scale within specific sequence contexts	[48, 78]
		RRBS	Analyze a limited number of gene promoters and regulatory sequence elements in a large number of samples; Analyzing and comparing genomic methylation patterns	[11, 28]
		BSPP	Focus sequencing on the most informative genomic regions ;exon capturing and SNP genotyping; Detecting methylation in large genomes	[29]
		BC-seq	Detect site-specific switches in methylation; Determine DNA methylation frequencies in CGIs sampled from a variety of genomic settings including promoters, exons, introns, and intergenic loci	[30]

Table 1: NGS-based technologies for detecting DNA methylation.

methyl-binding protein MECP2 [17]. Enrichment based methods are widely employed to survey DNA methylation pattern. Those methods use NGS to sequence methylated DNA fragments obtained by methyl-CpG binding domain (MBD) proteins (MBD-seq [18], MIRA [19], MethylCap-seq [20] or immunoprecipitation with specific antibodies for methylated cytosine (MeDIP-seq [21]. MeDIP-seq and MBD-seq are similar in concept where fragmented DNA is enriched based on its methylation content [22]. MIRA-seq (methylated CpG island recovery assay) consists of the capture of fragment double-stranded methylated DNA using the MBD2b/MDB3L1 complex and subsequent next-generation sequencing of eluted DNA. In MethylCap-seq, a methyl-binding domain protein is utilized to enrich DNA fractions with similar methylation levels and the avidity of the DNA-MBD interaction relies on the local methyl-CpGs density [20,23,24].

Bisulfite conversion

Having found the fact that after sodium bisulfite treatment, unmethylated cytosines in single-stranded DNA are deaminated to give uracil while leaving methylated cytosine intact, a revolution of DNA methylation analysis was spurred since the 1990s [25,26]. Bisulfite-

Resources	URL	Ref
MAQ	http://maq.sourceforge.net/	[33]
SOAPaligner/soap2	http://soap.genomics.org.cn/index.html	[36]
Bowtie	http://bowtie-bio.sourceforge.net/index.shtml	[32]
BLAT	http://genome.ucsc.edu/cgi-bin/hgBlat	[79]
BWA	http://maq.sourceforge.net/	[80]
BFAST	http://bfast.sourceforge.net	[81]
SHRiMP	http://compbio.cs.toronto.edu/shrimp.	[82]

SOAP, Short Oligonucleotide Analysis Package; MAQ, Mapping and Assembly with Quality; BLAT, BLAST-like alignment tool, BWA, Burrows-Wheeler Alignment; BFAST, BLAT-like Fast Accurate Search Tool; SHRiMP, The Short-Read Mapping Package.

Table 2: Alignment tools for short reads.

converted DNA is particularly well suited for sequencing-based approaches especially with the application of next-generation sequencing platforms [1]. The whole genome bisulfite-sequencing (BS-seq) is a powerful technique to measure the methylation state genome-wide at single-base resolution, based on the combination of the treatment of DNA with sodium bisulfite and NGS [27]. Nevertheless, mapping high-throughput bisulfite reads to the reference genome remains a great challenge. A modified version of BS-seq, Reduced Representation Bisulfite Sequencing (RRBS) [28], has been developed, which is based on size selection of restriction fragments. However, although RRBS can reduce the sequence redundancy, it is limited to methylation at restriction sites. To address this limitation , an approach called Bisulfite padlock probes (BSPP) [29], is a choice, which employs padlock probes to capture an arbitrary set of sequencing targets from bisulfite-converted DNA in highly parallel manner. Another bisulfate based method, BC-seq , is applied for DNA methylation profiling in genomic regions spanning tens of millions of bases through a combination of bisulfite conversion with hybrid selection technologies and deep sequencing [30].

Tools for mapping short reads to references and peak detection

Since millions to billions of short reads are generated through the combination of the NGS and the pretreatment methods, mapping the sheer volume of NGS data to the reference genome faces several challenges such as the alignment accuracy of short reads, time-consuming and high cost. Recently, several tools have been developed to address the above issues, such as MAQ, SOAPaligner/SOAP2, Bowtie, BLAT, BWA, BFAST and SHRiMP (Table 2) [31].These methods can be divided into two categories: hash table-based methods and Burrows–Wheeler based methods. All of HRiMP, MAQ, SOAP, BFAST and BFAST are the hash table-based tools which has either the reads or references by constructing a hash table of short oligomers [32]. Among these hash table-based tools, MAQ is an accurate, efficient, versatile, and user-friendly tool, which has been wildly used in aligning short reads from a single individual [33]. However, as other hash table-based aligner, MAQ also suffers two defects: A large memory is required to build an index for the human genome. In addition, it is unsuitable for the alignment of longer read because it is unable to support the gapped alignment for single-end reads [34]. Several tools, like Bowite, BWA and SOAP2, have been developed based on Burrows–Wheeler indexing. Bowtie is a short-read alignment tool based on Burrows-Wheeler and has comparable speed and high accuracy on aligning single-end reads rather than paired-end reads [32]. The SOAP2 alignment tool is a significantly improved version of the original SOAP [35,36]. By implementing Burrows–Wheeler indexing, SOAP2 not only reduces computer memory usage but also increases alignment speed at an unprecedented rate [31]. Compared with Bowite, SOAP2 may be applied to both single-end and paired-end reads. And BWA may also perform gapped alignments of short reads by sacrificing the speed of alignments. To sum up, mapping tens of millions of short reads to a reference genome efficiently with these tools open the doors to further investigate DNA methylation patterns.

To further detect significant functional regions with different methylation patterns from NGS data, several peak detection algorithms for chip-seq data have been developed in many studies, which have be applied to detect significantly methylated or unmethylated regions from NGS data of DNA methylation [37,38]. To our knowledge, more than 14 special tools for peak detection have developed as summarized in Table 3. Sliding window is the commonly used technique in peak detection algorithms, such as MACS, PeakSeq, FindPeaks and USeq. After identifying windows, these methods use different approaches to determine which windows are the true enriched regions [39]. For example, FindPeaks simply calculates significance of genomic regions without the control sample based on the assumed Poisson distribution followed by the reads [40], while MACS uses a control sample to more accurately model the background distribution of the reads and empirically estimates the false discovery rate (FDR) for each detected peak [41]. There are also several statistical algorithms, like BayesPeak [42], based on a fully Bayesian hidden Markov model and chromaSig

Algorithm	Description	Availability	Ref
MACS	Model-based analysis of ChIP-Seq	http://liulab.dfci.harvard.edu/MACS/	[41]
ChIPseeqer	in-depth analysis of ChIP-seq datasets	http://physiology.med.cornell.edu/faculty/elemento/lab/CS_files/ChIPseeqer-2.0.tar.gz	[83]
HPeak	A HMM-based algorithm for defining read enriched regions	www.sph.umich.edu/csg/qin/HPeak	[84]
CASSys	ChIP-seq data Analysis Software System	http://localness.zbh.uni-hamburg.de/~ProjektChipSeq/cgi-bin/login.rb	[85]
PeakSeq	A general scoring approach for ChIP-seq data analysis.	http://info.gersteinlab.org/PeakSeq	[86]
Sole-Search	Integrated peak-calling and analysis software	http://chipseq.genomecenter.ucdavis.edu/cgi-bin/chipseq.cgi	[87]
SISSRS	Precise identification of binding sites from short reads generated from ChIP-Seq experiment	http://sissrs.rajajothi.com/	[88]
BayesPeak	Bayesian analysis of ChIP-seq data	http://bioconductor.org/packages/release/bioc/html/BayesPeak.html	[42]
PeakRanger	A cloud-enabled peak caller for ChIP-seq data	http://www.modencode.org/software/ranger/	[89]
FindPeaks 3.1	A tool for identifying areas of enrichment	http://www.bcgsc.ca/platform/bioinfo/software/findpeaks	[40]
Sole-Search	An integrated analysis program for peak detection and functional annotation using ChIP-seq data	http://chipseq.genomecenter.ucdavis.edu/cgi-bin/chipseq.cgi	[87]
PeakAnalyzer	Genome-wide annotation of chromatin binding and modification loci	http://www.bioinformatics.org/peakanalyzer/wiki/	[90]
chromaSig	A Probabilistic Approach to Finding Common Chromatin Signatures	http://bioinformatics-renlab.ucsd.edu/rentrac/wiki/ChromaSig	[43]
Fish the ChIPs	A pipeline for automated genomic annotation of ChIP-Seq data	http://bio.ifom-ieo-campus.it/ftc/	[91]

Table 3: Peak detection algorithms.

Resources	Purpose	URL	Ref
RRBSMAP	A fast, accurate and user-friendly alignment tool for reduced representation bisulfite sequencing	http://rrbsmap.computational-epigenetics.org/	[92]
BSMAP	Ahole genome bisulfite sequence mapping	http://code.google.com/p/bsmap/	[48]
BS Seeker	Precise mapping for bisulfite sequencing	http://pellegrini.mcdb.ucla.edu/BS_Seeker/BS_Seeker.html	[46]
Bismark	Map and determine the Methylation state of BS-Seq read	http://www.bioinformatics.bbsrc.ac.uk/projects/bismark/	[47]
SOCS-B	An alignment algorithm for bisulfite sequencing using the Applied Biosystems SOLiD System	http://solidsoftwaretools.com/gf/project/socs/	[93]
BRAT	Bisulfite-treated reads analysis tool	http://compbio.cs.ucr.edu/brat/	[94]
BISMA	Analysis of bisulfite Sequencing data from both unique and repetitive sequences	http://biochem.jacobs-university.de/BDPC/BISMA/	[95]
BiQAnalyzerHT	Locus-specific analysis of DNA methylation by high-throughput bisulfite sequencing	http://biq-analyzer-ht.bioinf.mpi-inf.mpg.de/	[49]
CpGviewer	Sequence analysis and editing for bisulphite genomic sequencing projects	http://xserve1.leeds.ac.uk/~iancarr/cpgviewer	[50]
CpG PatternFinder	Windows-based program for bisulphite DNA	-	[52]
CyMATE	Bisulphite-based analysis of plant genomic DNA	http://www.gmi.oeaw.ac.at/CyMATE	[51]
GenomeStudio Software	Analyzing data generated from Illumina assays	-	-
MethMarker	Design, optimize and validate DNA methylation biomarkers for a given DMR	http://methmarker.mpi-inf.mpg.de/	[96]
BDPC	Bisulfite sequencing Data methylation analysis.	http://biochem.jacobs-university.de/BDPC	[97]
MethylCoder	Software pipeline for bisulte-treated sequences	https://github.com/brentp/methylcode	[98]
QUMA	Quantification tool for methylation analysis	http://quma.cdb.riken.jp/	[53]

CyMATE ,Cytosine Methylation Analysis Tool for Everyone; BSMAP, Bisulphite Sequence Mapping Program, BISMA, Bisulfite Sequencing DNA Methylation Analysis; BRAT,bisulte-treated reads analysis tool; DMR, differentially methylated DNA region; QUMA, Quantification Tool For Methylation Analysis.

Table 4: Analysis of bisulfite sequencing data with computational tools.

[43], on unsupervised learning method. Some significant comparisons of these algorithms have been made by Wilbanks and Facciotti [44] and Pepke et al. [45]. However, these peak detection algorithms could not taken CpG density of genomic fragments into consideration in detecting peaks of tag density. Therefore, the new algorithm for identifying the significant peak without the effect of CpG density will help in detecting the potential methylated or unmethylated regions of the genomes from the NGS data.

Analysis of bisulfite sequencing data with computational tools

The alignment approaches of short reads could not be directly applied to bisulfite sequencing data because the pretreatment method of bisulfite-conversion converts unmethylated cytosines to thymines. Fortunately, several analysis tools have been developed for analyzing bisulfite sequencing data as shown in Table 4. Most of the alignment tools apply a combination of the strategy for the asymmetric C/T conversion and the following mapping algorithms based on the previous short read mapping programs such as Bowtie (BS Seeker [46], Bismark [47], and SOAP (BSMAP [48]. Meanwhile, several software tools may be currently available for further bisulfite sequencing analysis. several tools such as BiQ AnalyzerHT [49] and CpGviewer [50], can accept raw bisulfite sequences as input data, while CyMATE [51], and CpG PatternFinder [52], need to use aligned sequences as input data. Furthermore, the tool of QUMA is an interactive web-based tool for quantitative methylation analysis and includes most of data-processing functions for the analysis of bisulfite sequences [53]. By the special tools for bisulfite sequencing data, the users may determine the methylated level of CpGs at single-base resolution.

DNA Methylation databases

In order to storage the vast amount of data generated by previous mentioned NGS-based technologies, several useful methylation databases have been available for researchers who might use the data as input for further research (Table 5). There are several primary methylation databases, like MethDB [54], designed to store heterogeneous data from different kinds of experiments and NGSmethDB [55], established for storage and retrieval of methylation data derived from NGS. MethylomeDB [56], includes published DNA methylation data which is related to the brain development and function, and MethyCancer [57], is a openly reachable database for human DNA methylation and cancer. What's more, having incorporated gene methylation data derived from cross-data set analysis for disease and normal samples, DiseaseMeth [58], a human disease methylation database , can be used for identifying differentially methylated genes deeply and investigating the relationship between gene and disease. In sum, following the further study about methylation, more databases will be produced and then more information about methylation will be known.

CpG islands

CpG islands (CGIs) are genomic regions with high frequency of CpGs which typically occur in the promoter regions [59]. Due to their importance as genomic markers in promoter regions and as epigenetic regulatory regions associated with promoter activity, the identification of CGIs becomes indispensible. CGIs can be identified either through experimental [60,61], or computational methods [62]. Here, we merely introduce some computational methods (Table 5). Three sequence parameters (length, GC content, and ratio of the observed over the expected CpGs(Obs_{CpG}/Exp_{CpG})) are commonly used as the criteria in sliding window in the identification algorithms(CpG IE [63], CpG IS [64], CpGProD [65]). However, the traditional CGIs criteria mostly identify repetitive sequences which are generally highly methylated. Although CpGProD and CpGIS use more stringent criteria to solve this problem, a portion of functional CGIs could be missed due to ad hoc thresholds. In addition , the window size and step size may limit the number and length of CGIs found by those methods [66]. As a result, rather than revise presented base compositional criteria further, several other methods focus on statistical property in a sequence. For example, CpGcluster [67], identifies CGIs based on the physical distances between neighboring CpGs and CG clusters [68], obtains CG-dense fragments based on empirical species-specific CG cluster definition. Nevertheless, compared with the sequence-criteria-based methods, CpGcluster has a high false positive rate and the proportion of promoter-associated CGIs in CG clusters is slightly lower [69]. Among the current tools for identification of CpG islands, CpG_MI [66], obtained highest prediction accuracy of functional CpG islands by fully utilizing the cumulative mutual information of physical distances between two neighboring CpGs. These algorithms for identification of CpG islands provide the functional regions for the studies of DNA methylation.

Other analysis tools of methylation data

Several other methods for bioinformatics analysis of DNA methylation are also listed in Table 5.Besides CpG islands, DMR (differen-

Tools	Purpose	UTR	Ref
MethDB	Database for DNA methylation data	http://www.methdb.de	[54]
MethyCancer Database	Database of cancer DNA methylation data	http://methycancer.psych.ac.cn/	[57]
PubMeth	Database of DNA methylation literature	http://www.pubmeth.org/	[99]
NGSmethDB	Database for DNA methylation data at single-base resolution	http://bioinfo2.ugr.es/NGSmethDB/gbrowse/	[55]
DBCAT	Database of CpG islands and analytical tools for identifying comprehensive methylation profiles in cancer cells	http://dbcat.cgm.ntu.edu.tw/	[100]
MethylomeDB	Database of DNA methylation profiles of the brain	http://epigenomics.columbia.edu/methylomedb/index.html	[56]
DiseaseMeth	Human disease methylation database	http://bioinfo.hrbmu.edu.cn/diseasemeth	[58]
CpG IE	Identification of CpG islands	http://bioinfo.hku.hk/cpgieintro.html	[63]
CpG IS	Identification of CpG islands	http://cpgislands.usc.edu/	[64]
CG clusters	Identification of CpG islands	http://greallylab.aecom.yu.edu/cgClusters/	[68]
CpGcluster	Identification of CpG islands	http://bioinfo2.ugr.es/CpGcluster	[67]
CpGIF	Identification of CpG islands	http://www.usd.edu/~sye/cpgisland/CpGIF.htm	
CpG_MI	Identification of CpG islands	http://bioinfo.hrbmu.edu.cn/cpgmi	[66]
CpGProD	Identification of CpG islands	http://pbil.univ-lyon1.fr/software/cpgprod.html	[65]
EpiGRAPH	Genome scale statistical analysis	http://epigraph.mpi-inf.mpg.de/WebGRAPH	[71]
Galaxy	General purpose analysis	http://main.g2.bx.psu.edu/	[102]
QDMR	Identification of differentially methylated regions	http://bioinfo.hrbmu.edu.cn/qdmr.	[103]
Batman	MeDIP DNA methylation analysis tool	http://td-blade.gurdon.cam.ac.uk/software/batman	[23]
CisGenome Browser	A flexible tool for genomic data visualization	http://biogibbs.stanford.edu/~jiangh/browser/	[70]
MethVisual	Visualization and exploratory statistical analysis of DNA methylation profiles from bisulfite sequencing	http://methvisual.molgen.mpg.de/	[104]
MethTools	A toolbox to visualize and analyze DNA methylation data	http://genome.imb-jena.de/methtools/	[105]

Table 5: Bioinformatics tools.

tially methylated region) is another focus in recent DNA methylation studies. Compared with other methods based on statistics or counting, QDMR [58], (quantitative differentially methylated regions) is an effective tool to quantify methylation difference and identify DMRs across multiple samples by adapting Shannon entropy. CisGenome Browser is a wide application tool for data visualization [70]. The comprehensive tools of EpiGRAPH [71], and Galaxy [72], performed analysis of the genomic and epigenomic data, such as genome sequences, conservation scores, methylation data or any signal associated with genomic loci or regions generated by biological experiments.

Conclusion

In recent years, researchers pay more and more attention to the studies of DNA methylation associated with embryonic development [73], as well as cancer [74]. Coupled with different pretreatment approaches, numerous next-generation sequencing based technologies are available for detecting DNA methylation. Although compared with bisulfite-based methods, enzyme-based and affinity enrichment-based DNA methylation analysis technologies are relatively simple and cheap, the single-base resolution of bisulfite sequencing makes it possible for researchers to extract methylation information of CpGs genome-wide. A great quantity of data generated by these methods shift the bottleneck in DNA methylation advances from data generation to data analysis [72]. In this review, we summarized the useful alignment methods and peek-detection algorithms as well as bioinformatics tools for storage, analysis and visualization of DNA methylation data. In brief, along with the advances in measuring technologies of DNA methylation in the future , more computational tools and resources for DNA methylation analysis will be available, which may facilitate the users to explore the mechanism of DNA methylation patterns.

Acknowledgement

JS and DH contributed equally to this work and are regarded as co-first authors. This work was supported partly by National Natural Science Foundation of China (61075023), Science Foundation of Heilongjiang Province (C201012 and QC2011C061) and Scientific Research Fund of Heilongjiang Provincial Education Department (12511272).

References

1. Laird PW (2010) Principles and challenges of genome-wide DNA methylation analysis. Nat Rev Genet 11: 191-203.
2. Bird A (2002) DNA methylation patterns and epigenetic memory. Genes Dev 16: 6-21.
3. Tweedie S, Charlton J, Clark V, Bird A (1997) Methylation of genomes and genes at the invertebrate-vertebrate boundary. Mol Cell Biol 17: 1469-1475.
4. Suzuki MM, Bird A (2008) DNA methylation landscapes: provocative insights from epigenomics. Nat Rev Genet 9: 465-476.
5. Anderson CL, Brown CJ (2002) Variability of X chromosome inactivation: effect on levels of TIMP1 RNA and role of DNA methylation. Hum Genet 110: 271-278.
6. Li E, Beard C, Jaenisch R (1993) Role for DNA methylation in genomic imprinting. Nature 366: 362-365.
7. Richardson BC (2002) Role of DNA methylation in the regulation of cell function: autoimmunity, aging and cancer. J Nutr 132: 2401S-2405S.
8. Su J, Shao X, Liu H, Liu S, Wu Q, et al. (2012) Genome-wide dynamic changes of DNA methylation of repetitive elements in human embryonic stem cells and fetal fibroblasts. Genomics 99: 10-17.
9. Liu H, Su J, Li J, Liu H, Lv J, et al. (2011) Prioritizing cancer-related genes with aberrant methylation based on a weighted protein-protein interaction network. BMC Syst Biol 5: 158.
10. Zilberman D, Henikoff S (2007) Genome-wide analysis of DNA methylation patterns. Development 134: 3959-3965.
11. Harris RA, Wang T, Coarfa C, Nagarajan RP, Hong C, et al. (2010) Comparison of sequencing-based methods to profile DNA methylation and identification of monoallelic epigenetic modifications. Nat Biotechnol 28: 1097-1105.
12. Oda M, Glass JL, Thompson RF, Mo Y, Olivier EN, et al. (2009) High-resolution genome-wide cytosine methylation profiling with simultaneous copy number analysis and optimization for limited cell numbers. Nucleic Acids Res 37: 3829-3839.
13. Colaneri A, Staffa N, Fargo DC, Gao Y, Wang T, et al. (2011) Expanded methyl-sensitive cut counting reveals hypomethylation as an epigenetic state that highlights functional sequences of the genome. Proc Natl Acad Sci 108: 9715-9720.
14. Ball MP, Li JB, Gao Y, Lee JH, LeProust EM, et al. (2009) Targeted and genome-scale strategies reveal gene-body methylation signatures in human cells. Nat Biotechnol 27: 361-368.
15. Berman BP, Weisenberger DJ, Laird PW (2009) Locking in on the human methylome. Nat Biotechnol 27: 341-342.
16. Brunner AL, Johnson DS, Kim SW, Valouev A, Reddy TE, et al. (2009) Distinct DNA methylation patterns characterize differentiated human embryonic stem cells and developing human fetal liver. Genome Res 19: 1044-1056.

17. Cross SH, Charlton JA, Nan X, Bird AP (1994) Purification of CpG islands using a methylated DNA binding column. Nat Genet 6: 236-244.

18. Serre D, Lee BH, Ting AH (2010) MBD-isolated Genome Sequencing provides a high-throughput and comprehensive survey of DNA methylation in the human genome. Nucleic Acids Res 38: 391-399.

19. Mitchell N, Deangelis JT, Tollefsbol TO (2011) Methylated-CpG Island Recovery Assay. Methods Mol Biol 791: 125-133.

20. Brinkman AB, Simmer F, Ma K, Kaan A, Zhu J, et al. (2010) Whole-genome DNA methylation profiling using MethylCap-seq. Methods 52: 232-236.

21. Palmke N, Santacruz D, Walter J (2011) Comprehensive analysis of DNA-methylation in mammalian tissues using MeDIP-chip. Methods 53: 175-184.

22. Hirst M, Marra MA (2010) Next generation sequencing based approaches to epigenomics. Brief Funct Genomics 9: 455-465.

23. Down TA, Vardhman KR, Daniel JT, Flicek P, Li H, et al. (2008) A Bayesian deconvolution strategy for immunoprecipitation-based DNA methylome analysis. Nat Biotechnol 26: 779-785.

24. Weber M, Davies JJ, Wittig D, Oakeley EJ, Haase M, et al. (2005) Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells. Nat Genet 37: 853-862.

25. Frommer M, McDonald LE, Millar DS, Collis CM, Watt F, et al. (1992) A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. Proc Natl Acad Sci 89: 1827-1831.

26. Clark SJ, Harrison J, Paul CL, Frommer M (1994) High sensitivity mapping of methylated cytosines. Nucleic Acids Res 22: 2990-2997.

27. Krueger F, Kreck B, Franke A, Andrews SR (2012) DNA methylome analysis using short bisulfite sequencing data. Nat Methods 9: 145-151.

28. Meissner A, Gnirke A, Bell GW, Ramsahoye B, Lander ES, et al. (2005) Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. Nucleic Acids Res 33: 5868-5877.

29. Deng J, Shoemaker R, Xie B, Gore A, LeProust EM, et al. (2009) Targeted bisulfite sequencing reveals changes in DNA methylation associated with nuclear reprogramming. Nat Biotechnol 27: 353-360.

30. Hodges E, Smith AD, Kendall J, Xuan Z, Ravi K, et al. (2009) High definition profiling of mammalian DNA methylation by array capture and single molecule bisulfite sequencing. Genome Res 19: 1593-1605.

31. Lee HC, Lai K, Lorenc MT, Imelfort M, Duran C, et al. (2011) Bioinformatics tools and databases for analysis of next-generation sequence data. Brief Funct Genomics 11: 12-24.

32. Langmead B, Trapnell C, Pop M, Salzberg SL et al. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10: R25.

33. Li H, Ruan J, Durbin R (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res 18: 1851-1858.

34. Li H, Durbin R (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics 26: 589-595.

35. Li R, Li Y, Kristiansen K, Wang J (2008) SOAP: short oligonucleotide alignment program. Bioinformatics 24: 713-714.

36. Li R, Chang Yu, Yingrui Li, Tak-Wah Lam, Siu-Ming, et al. (2009) E SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics 25: 1966-1967.

37. Lan X, Adams C, Landers M, Dudas M, Krissinger D, et al. (2011) High resolution detection and analysis of CpG dinucleotides methylation using MBD-Seq technology. PLoS One 6: e22226.

38. Kallio MA, Tuimala JT, Hupponen T, Klemelä P, Gentile M, et al. (2011) Chipster: user-friendly analysis software for microarray and other high-throughput data. BMC Genomics 12: 507.

39. Malone BM, Tan F, Bridges SM, Peng Z (2011) Comparison of four ChIP-Seq analytical algorithms using rice endosperm H3K27 trimethylation profiling data. PLoS One 6: e25260.

40. Fejes AP, Robertson G, Bilenky M, Varhol R, Bainbridge M, et al. (2008) FindPeaks 3.1: a tool for identifying areas of enrichment from massively parallel short-read sequencing technology. Bioinformatics 24: 1729-1730.

41. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, et al. (2008) Model-based analysis of ChIP-Seq (MACS). Genome Biol 9: R137.

42. Spyrou C, Stark R, Lynch AG, Tavaré S (2009) BayesPeak: Bayesian analysis of ChIP-seq data. BMC Bioinformatics 10: 299.

43. Hon G, Ren B, Wang W (2008) ChromaSig: a probabilistic approach to finding common chromatin signatures in the human genome. PLoS Comput Biol 4: e1000201.

44. Wilbanks EG, Facciotti MT (2010) Evaluation of algorithm performance in ChIP-seq peak detection. PLoS One 5: e11471.

45. Pepke S, Wold B, Mortazavi A (2009) Computation for ChIP-seq and RNA-seq studies. Nat Methods 6: S22-S32.

46. Chen PY, Cokus SJ, Pellegrini M (2010) BS Seeker: precise mapping for bisulfite sequencing. BMC Bioinformatics 11: 203.

47. Krueger F, Andrews SR (2011) Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. Bioinformatics 27: 1571-1572.

48. Xi Y, Li W (2009) BSMAP: whole genome bisulfite sequence MAPping program. BMC Bioinformatics 10: 232.

49. Lutsik P, Lars Feuerbach, Julia Arand, Thomas Lengauer, Jorn Walter, et al. (2011) BiQ Analyzer HT: locus-specific analysis of DNA methylation by high-throughput bisulfite sequencing. Nucleic Acids Res 39: W551-W556.

50. Carr IM, Valleley EMA, Cordery SF, Markham AF, Bonthron DT (2007) Sequence analysis and editing for bisulphite genomic sequencing projects. Nucleic Acids Res 35: e79.

51. Hetzl J, Foerster AM, Raidl G, Mittelsten SO (2007) CyMATE: a new tool for methylation analysis of plant genomic DNA after bisulphite sequencing. Plant J 51: 526-536.

52. Xu YH, Manoharan HT, Pitot HC (2007) CpG PatternFinder: a Windows-based utility program for easy and rapid identification of the CpG methylation status of DNA. Biotechniques 43: 334, 336-340, 342.

53. Kumaki Y, Oda M, Okano M (2008) QUMA: quantification tool for methylation analysis. Nucleic Acids Res 36: W170-W175.

54. Grunau C, Renault E, Rosenthal A, Roizes G (2001) MethDB--a public database for DNA methylation data. Nucleic Acids Res 29: 270-274.

55. Hackenberg M, Barturen G, Oliver JL (2011) NGSmethDB: a database for next-generation sequencing single-cytosine-resolution DNA methylation data. Nucleic Acids Res 39: D75-D79.

56. Xin Y, Chanrion B, O'Donnell AH, Milekic M, Costa R, et al. (2012) MethylomeDB: a database of DNA methylation profiles of the brain. Nucleic Acids Res 40: D1245-D1249.

57. He X, Suhua Chang, Jiajie Zhang, Qian Zhao, Haizhen Xiang, et al. (2008) MethyCancer: the database of human DNA methylation and cancer. Nucleic Acids Res 36: D836-D841.

58. Lv J, Hongbo Liu, Jianzhong Su, Xueting Wu, Hui Liu, et al. (2012) DiseaseMeth: a human disease methylation database. Nucleic Acids Res 40: D1030-1035.

59. Saxonov S, Berg P, Brutlag DL (2006) A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. Proc Natl Acad Sci 103: 1412-1417.

60. Illingworth R, Kerr A, DeSousa D, Jørgensen H, Ellis P, et al. (2008) A novel CpG island set identifies tissue-specific methylation at developmental gene loci. PLoS Biol 6: e22.

61. Heisler LE, Torti D, Boutros PC, Watson J, Chan C, et al. (2005) CpG Island microarray probe sequences derived from a physical library are representative of CpG Islands annotated on the human genome. Nucleic Acids Res 33: 2952-2961.

62. Gardiner-Garden M, Frommer M (1987) CpG islands in vertebrate genomes. J Mol Biol 196: 261-282.

63. Wang Y, Leung FC (2004) An evaluation of new criteria for CpG islands in the human genome as gene markers. Bioinformatics 20: 1170-1177.

64. Takai D, Jones PA (2003) The CpG island searcher: a new WWW resource. In Silico Biol 3: 235-240.

65. Ponger L, Mouchiroud D (2002) CpGProD: identifying CpG islands associated with transcription start sites in large genomic mammalian sequences. Bioinformatics 18: 631-633.

66. Su J, Zhang Y, Jie Lv, Hongbo L,Tang X, et al. (2010) CpG_MI: a novel approach for identifying functional CpG islands in mammalian genomes. Nucleic Acids Res 38: e6.

67. Hackenberg M, Previti C, Luque-Escamilla PL, Carpena P, Martínez-Aroza J, et al. (2006) CpGcluster: a distance-based algorithm for CpG-island detection. BMC Bioinformatics 7: 446.

68. Glass JL, Thompson RF, Khulan B, Figueroa ME, Olivier EN, et al. (2007) CG dinucleotide clustering is a species-specific property of the genome. Nucleic Acids Res 35: 6798-6807.

69. Zhao Z, Han L (2009) CpG islands: algorithms and applications in methylation studies. Biochem Biophys Res Commun 382: 643-645.

70. Jiang H, Wang F, Dyer NP, Wong WH (2010) CisGenome Browser: a flexible tool for genomic data visualization. Bioinformatics 26: 1781-1782.

71. Bock C, Halachev K, Büch J, Lengauer T (2009) EpiGRAPH: user-friendly software for statistical analysis and prediction of (epi)genomic data. Genome Biol 10: R14.

72. Bock C, Von Kuster G, Halachev K, Taylor J, Nekrutenko A, et al. (2010) Web-based analysis of (Epi-) genome data using EpiGRAPH and Galaxy. Methods Mol Biol 628: 275-296.

73. Reik W, Dean W, Walter J (2001) Epigenetic reprogramming in mammalian development. Science 293: 1089-1093.

74. Jones PA, Baylin SB (2002) The fundamental role of epigenetic events in cancer. Nat Rev Genet 3: 415-428.