Published online 18 October 2016 *Nucleic Acids Research, 2017, Vol. 45, Database issue* ***D85–D89***
doi: 10.1093/nar/gkw950

MethSMRT: an integrative database for DNA N6-methyladenine and N4-methylcytosine generated by single-molecular real-time sequencing

Pohao Ye[1,†], Yizhao Luan[1,†], Kaining Chen[1], Yizhi Liu[1], Chuanle Xiao[1,*] and Zhi Xie[1,2,3,*]

[1]State Key Laboratory of Ophthalmology, Guangdong Provincial Key Lab of Ophthalmology and Visual Science, Zhongshan Ophthalmic Center, Sun Yat-sen University, Guangzhou 510060, China, [2]Center for Precision Medicine, Collaborative Innovation Center for Cancer Medicine, Sun Yat-sen University, Guangzhou 510000, China and [3]Collaborative Innovation Center of Genetics and Development, Fudan University, Shanghai 200438, China

Received August 11, 2016; Revised September 30, 2016; Editorial Decision October 07, 2016; Accepted October 10, 2016

ABSTRACT

DNA methylation is an important type of epigenetic modifications, where 5- methylcytosine (5mC), 6-methyadenine (6mA) and 4-methylcytosine (4mC) are the most common types. Previous efforts have been largely focused on 5mC, providing invaluable insights into epigenetic regulation through DNA methylation. Recently developed single-molecule real-time (SMRT) sequencing technology provides a unique opportunity to detect the less studied DNA 6mA and 4mC modifications at single-nucleotide resolution. With a rapidly increased amount of SMRT sequencing data generated, there is an emerging demand to systematically explore DNA 6mA and 4mC modifications from these data sets. MethSMRT is the first resource hosting DNA 6mA and 4mC methylomes. All the data sets were processed using the same analysis pipeline with the same quality control. The current version of the database provides a platform to store, browse, search and download epigenome-wide methylation profiles of 156 species, including seven eukaryotes such as *Arabidopsis*, *C. elegans*, *Drosophila*, mouse and yeast, as well as 149 prokaryotes. It also offers a genome browser to visualize the methylation sites and related information such as single nucleotide polymorphisms (SNP) and genomic annotation. Furthermore, the database provides a quick summary of statistics of methylome of 6mA and 4mC and predicted methylation motifs for each species. MethSMRT is publicly available at http://sysbio.sysu.edu.cn/methsmrt/ without use restriction.

INTRODUCTION

DNA methylation is an important type of epigenetic modifications, which greatly expands the information content of DNA. The most common types of DNA methylation are 5-methylcytosine (5mC), 6-methyadenine (6mA) and 4-methylcytosine (4mC) (1). In eukaryotes, 5mC is the dominant type, playing an important role in gene regulation, transposon suppression and genomic imprinting (2). Aberrant 5mC patterns have been associated with many diseases and cancers (3). Take retinoblastoma for example, DNA hypermethylation silenced gene expression of RAS-associated domain family 1A in tumor but not in normal tissue (4). In prokaryotes, 6mA and 4mC are the most prevalent DNA modifications that are primarily used for distinguishing host DNA from foreign pathogenic DNA (5). In contrast, 6mA and 4mC are suggested to be minimal and only detectable by highly sensitive technologies in eukaryotes (5). Until recently, several studies reported the epigenome-wide patterns of 6mA in eukaryotes, including *Chlamydomonas*, *C. elegans* and *Drosophila*, showing wide existence of 6mA in eukaryotes and its important functions in regulating gene regulation and development (6–8).

To date, many DNA methylation databases had been constructed, providing invaluable resources for the epigenetic community. MethDB is the first database that stores DNA methylation profiles and associated gene expression information (9). NGSMethDB hosts DNA methylation profiles generated from bisulfite sequencing technique (10). MethBank focuses on methylome changes during embryonic development (11) while MethyCancer and MENT focus on cancers (12,13). PubMeth is another cancer methylation database, based on text-mining of published literature (14). All these databases hosted DNA 5mC profiles and no database provided DNA 6mA or 4mC information so far.

*To whom correspondence should be addressed. Tel: +86 20 87335131; Email: xiezhi@gmail.com
Correspondence may also be addressed to Chuanle Xiao. Tel: +86 20 87335131; Email: xiaochuanle@126.com
†These authors contributed equally to this work as the first authors.

Recently developed single-molecule real-time (SMRT) sequencing technology allows detection of individual molecules of DNA in real time without amplification process, which is also called 'third generation sequencing' (15). By monitoring kinetic 'signature' of a base during the normal course of sequencing, the presence of the base modification can also be directly detected as a measurement of increased inter-pulse duration (IPD) compared to unmodified DNA bases, where IPD is defined as space between fluorescence pluses (15). All the three major types of DNA methylation can be detected from SMRT sequencing data. In particular, 6mA and 4mC provide highly sensitive kinetic signals therefore require only 25-fold coverage to obtain high confident levels of detection (16). Because the present high-throughput techniques for DNA methylation mainly focus on 5mC modification, SMRT sequencing technology provides a unique opportunity to detect the less studied DNA 6mA and 4mC modifications (16).

With a rapidly increased amount of SMRT sequencing data generated in the past several years, there is an emerging demand to systematically explore DNA 6mA and 4mC modifications from these data sets. Here, we present MethSMRT, the first database for DNA 6mA and 4mC methylomes, generated from the publicly available SMRT sequencing data sets. The database provides a platform to host, analyze, browse, search and download 6mA and 4mC methylomes for 156 species, including seven eukaryotes and 149 prokaryotes. It also offers a genome browser to visualize the methylation profiles and related information such as single nucleotide polymorphisms (SNP) and gene annotation. In addition, the database provides a quick summary of statistics of methylomes of 6mA and 4mC and predicted methylation motifs for each species. Because all the data sets in the database were processed using the same analysis pipeline with the same quality control, it is feasible for a consistent comparison of methylomes between different species and data sets. MethSMRT is publicly available at http://sysbio.sysu.edu.cn/methsmrt/ without use restriction.

MATERIALS AND METHODS

Data sources

The SMRT sequencing data sets were downloaded from NCBI Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo/) (17) and Sequence Read Archive (http://www.ncbi.nlm.nih.gov/sra) (18). DNA 6mA and 4mC sites were detected from the SMRT sequencing data sets using a unified pipeline as described in the next section. The current version of MethSMRT includes a total of 156 species, including 7 eukaryotes such as *Arabidopsis*, *C. elegans*, *Drosophila*, Mouse and Yeast, as well as 149 prokaryotes. The associated SNP data were downloaded from dbSNP (19) or species-specific databases. The genome references were downloaded from NCBI or IMG databases (See Supplementary Table S1). Available 5mC modification sites of *Arabidopsis* and mouse were downloaded from NGSMethDB (10).

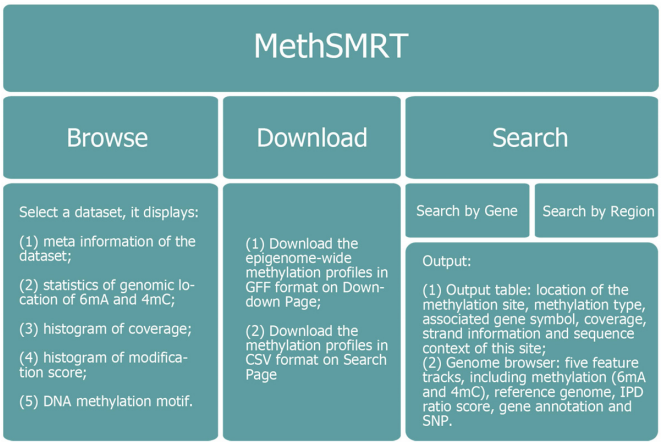

Figure 1. Functionality of MethSMRT.

DNA modification detection

We used the PacBio SMRT analysis platform (version: 2.3.0) for DNA 6mA and 4mC modification detection (http://www.pacb.com/products-and-services/analytical-software/smrt-analysis/analysis-applications/epigenetics/). Briefly, raw data files in the h5 format were downloaded. Raw reads first were filtered using SFilter to remove sequencing adapters, short reads, defined as read length less than 50-nucleotide (nt) or reads with a low quality region (read score < 0.75 by default). The filtered reads were aligned to the reference genome using pbalign (version: 0.2.0.1). Kinetic analysis of the aligned DNA sequencing data were used to identify DNA 6mA and 4mC modifications using the default parameters of the SMRT analysis platform. Followed the recommendations by the PacBio, the 6mA or 4mC sites with less than 25-fold coverage per strand were removed from further analysis. The modification score is defined as phred-transformed *P*-value that a kinetic deviation exists at the position according to the PacBio manual. To gain a reliable modification site, the sites with modification score < 20 have been filtered out.

Data analysis

To obtain distributions of 4mC and 6mA sites on genomes, we classified the modification sites onto exon, intron, intergenic, promoter and UTR regions for eukaryotes. For prokaryotes, the modification sites were classified onto exon, intergenic and promoter regions. The genomic annotation of exon, intron, UTR and intergenic regions were defined according to the annotation from NCBI RefSeq (20) or JGI database (21). The genomic regions were extracted by BEDTools (v2.19.1) (22). To define promoters of eukaryotic genes, we used regions between 500-nt upstream of transcription starting site to 500-nt downstream of transcription starting site (23). For promoters of prokaryotic genes, we used regions between 100-nt upstream of the starting site of coding DNA sequence (CDS) to that of 50-ntdownstream (24). To infer DNA modification motif, we first extracted sequences between 4-nt upstream to 4-nt downstream of the modification site. Duplicated sequences were excluded and MEME (version: 4.11.1) was used to pre-

Figure 2. Screenshot of the genome browser. (**A**) Five information tracks available in the browser: (1) 6mA/4mC track, (2) reference genome track, (3) IPD track, (4) gene annotation track, (5) 5mC track and (6) SNP track. Note that 5mC track is only available for *Arabidopsis* and mouse. (**B**) Zoom in a methylation site, with reference genome sequence and IPD signal. (**C**) Detailed information of the methylation site is available when clicking the methylation site in the browser. (**D**) Detailed information of the gene annotation is available when clicking the gene annotation track.

dict enriched motifs (25). For 5mC profiles of *Arabidopsis* or mouse, we combined all the available methylation sites of different samples downloaded from NGSMethDB. Methylation score was defined as the number of methylated reads on a given position divided by the number of total mapped reads of that position. Among multiple reported 5mC events on the same position, the maximum of methylation scores was used.

Database implementation

The database was organized using MySQL and queried using PHP scripts. The web pages were constructed using HTML5 with JavaScript. To provide a smooth and friendly users interface, we used Bootstrap framework from the front-end toolkit (http://getbootstrap.com/). The graphs in the Browse pages were produced by D3.js library (https://d3js.org/). Jbrowser was used to browse genomes and visualize modification sites (26).

RESULTS

Usage and access

The main functionality of MethSMRT is shown in Figure 1, including browsing, searching, visualizing and downloading DNA 6mA and 4mC modifications in single-nucleotide resolution for 156 species.

Search

Users can query the 6mA and 4mC profiles using gene name or genomic location. For searching by a gene name, users first select a species and enter a gene symbol or Ensembl ID in the search box. Alternatively, users can also search the database by entering a genomic region in a chromosome. The output page returns 6mA and 4mC profiles of a given gene or region. By default, the output page displays genomic location of the methylation site, methylation type, associated gene symbol, coverage of reads, strand information and sequence context of this site (Supplementary Figure S1). Some features of usage of the search page include: (i) Users can sort the table by clicking the column names; (ii) Users can selectively display the methylation sites based on the coverage or methylation type; (iii) Users can selectively display the columns of the table; and (iv) Users can export the output page in CSV format.

Genome browser

MethSMRT provides an interactive genome browser to query and visualize the 6mA and 4mC profiles and some relevant information. By clicking the 'view' icon in the JBrowse column on the most left side of the output table, the genome browser starts (Figure 2). The top left section of the browser provides click-box of six different feature tracks, including 6mA/4mC track, reference genome, IPD ratio score, gene annotation, SNP and 5mC track, where 5mC track is only available for *Arabidopsis* or mouse in the current version of MethSMRT The 6mA/4mC track shows the methylation sites and by clicking the site users can obtain the detailed attributes of the site such as coverage, sequence context of the site, exact position and IPD score. The reference genome track shows the genomic sequence and the gene annotation track shows the genomic structure of genes. The SNP track displays the SNP information and the IPD ratio track indicates the kinetic signature of the methylation site. The 5mC track displays previously identified 5mC sites.

Browse

The browse page displays meta-information of each SMRT data set and summary of the 6mC and 4mA profiles (Supplementary Figure S2). The meta-information of a given SMRT data set includes size and the number of runs of the data, SRA ID, Pubmed ID and a quick link to the processed file in GFF format. The page also displays statistics of genomic location of the methylation sites, histogram of coverage in log2 scale and histogram of modification score in log2 scale (see Materials and Methods section for details). In addition, the page displays the consensus sequence motif the methylation sites.

Download

MethSMRT provides two ways to download the methylation profiles. First, the download page provides GFF files that contain epigenome-wide 6mA and 4mC modification sites. Because there are over 150 species in the database, we provide a convenient search box for users to quickly find out the species using species name, SRA ID or taxon ID. In addition, users can also download the search results on the Search page in CSV format.

DISCUSSION

MethSMRT is the first resource of DNA 6mA and 4mC methylomes, generated from SMRT sequencing technology. With delivery of the new SMRT sequencing platform, the Sequel system, from PacBio Biosciences, the sequencing costs further reduced and we expect that more SMRT data sets will be generated with an increasing pace. MethSMRT will continue updating and incorporating new data sets. It should be noted that although the SMRT data sets of human and gorilla genomes have been recently reported (27,28), the current PacBio SMRT analysis platform requires extensive size of RAM (estimated $>$ 2 Tb) to process the data sets and we are in process to parallelize the analysis platform. We will include the 6mA and 4mC methylomes for human and gorilla in the near future.

To sum up, MethSMRT integrates and visualizes single-nucleotide resolution of DNA 6mA and 4mC methylomes. Together with many other useful DNA modification databases, MethSMRT will be an important epigenetic resource that facilitates discovery of methylation events and understanding of its biological functions.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors thank the members of Zhi Xie's lab for providing valuable suggestions.

FUNDING

National Natural Science Foundation of China [31471232 to Z.X.]; Joint Research Fund for Overseas Natural Science of China [3030901001222 to Z.X.]; Major Program of Science and Technology of Guangzhou [201607020001to Z.X.]; Science and Technology Planning Projects of Guangdong Province [2014B030301040 to Z.X.]; The Basic Research Funds of SYSU [15ykjc23d] and Guangdong Natural Science Function [2015A0303131127 to C.L.X.]. Funding for open access charge: National Natural Science Foundation of China [31471232 to Z.X.].
Conflict of interest statement. None declared.

REFERENCES

1. Chen,K., Zhao,B.S. and He,C. (2016) Nucleic acid modifications in regulation of gene expression. *Cell Chem. Biol.*, **23**, 74–85.
2. Suzuki,M.M. and Bird,A. (2008) DNA methylation landscapes: provocative insights from epigenomics. *Nat. Rev. Genet.*, **9**, 465–476.
3. Jones,P.A. (2012) Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nat. Rev. Genet.*, **13**, 484–492.
4. Harada,K., Toyooka,S., Maitra,A., Maruyama,R., Toyooka,K.O., Timmons,C.F., Tomlinson,G.E., Mastrangelo,D., Hay,R.J., Minna,J.D. *et al.* (2002) Aberrant promoter methylation and silencing of the RASSF1A gene in pediatric tumors and cell lines. *Oncogene*, **21**, 4345–4349.
5. Heyn,H. and Esteller,M. (2015) An adenine code for DNA: A second life for N6-methyladenine. *Cell*, **161**, 710–713.
6. Fu,Y., Luo,G.Z., Chen,K., Deng,X., Yu,M., Han,D., Hao,Z., Liu,J., Lu,X., Dore,L.C. *et al.* (2015) N6-methyldeoxyadenosine marks active transcription start sites in Chlamydomonas. *Cell*, **161**, 879–892.
7. Greer,E.L., Blanco,M.A., Gu,L., Sendinc,E., Liu,J., Aristizabal-Corrales,D., Hsu,C.H., Aravind,L., He,C. and Shi,Y. (2015) DNA methylation on N6-Adenine in C. elegans. *Cell*, **161**, 868–878.
8. Zhang,G., Huang,H., Liu,D., Cheng,Y., Liu,X., Zhang,W., Yin,R., Zhang,D., Zhang,P., Liu,J. *et al.* (2015) N6-methyladenine DNA modification in Drosophila. *Cell*, **161**, 893–906.
9. Grunau,C., Renault,E., Rosenthal,A. and Roizes,G. (2001) MethDB–a public database for DNA methylation data. *Nucleic Acids Res.*, **29**, 270–274.
10. Hackenberg,M., Barturen,G. and Oliver,J.L. (2011) NGSmethDB: a database for next-generation sequencing single-cytosine-resolution DNA methylation data. *Nucleic Acids Res.*, **39**, D75–D79.
11. Zou,D., Sun,S., Li,R., Liu,J., Zhang,J. and Zhang,Z. (2015) MethBank: a database integrating next-generation sequencing single-base-resolution DNA methylation programming data. *Nucleic Acids Res.*, **43**, D54–D58.
12. He,X., Chang,S., Zhang,J., Zhao,Q., Xiang,H., Kusonmano,K., Yang,L., Sun,Z.S., Yang,H. and Wang,J. (2008) Methycancer: the database of human DNA methylation and cancer. *Nucleic Acids Res.*, **36**, D836–D841.
13. Baek,S.J., Yang,S., Kang,T.W., Park,S.M., Kim,Y.S. and Kim,S.Y. (2013) MENT: methylation and expression database of normal and tumor tissues. *Gene*, **518**, 194–200.
14. Ongenaert,M., Van Neste,L., De Meyer,T., Menschaert,G., Bekaert,S. and Van Criekinge,W. (2008) PubMeth: a cancer methylation database combining text-mining and expert annotation. *Nucleic Acids Res.*, **36**, D842–D846.
15. Flusberg,B.A., Webster,D.R., Lee,J.H., Travers,K.J., Olivares,E.C., Clark,T.A., Korlach,J. and Turner,S.W. (2010) Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nat. Methods*, **7**, 461–465.
16. Davis,B.M., Chao,M.C. and Waldor,M.K. (2013) Entering the era of bacterial epigenomics with single molecule real time DNA sequencing. *Curr. Opin. Microbiol.*, **16**, 192–198.
17. Clough,E. and Barrett,T. (2016) The gene expression omnibus database. *Methods Mol. Biol.*, **1418**, 93–110.
18. Leinonen,R., Sugawara,H. and Shumway,M. (2011) The sequence read archive. *Nucleic Acids Res.*, **39**, D19–D21.
19. (2016) Database resources of the national center for biotechnology information. *Nucleic Acids Res.*, **44**, D7–D19.
20. O'Leary,N.A., Wright,M.W., Brister,J.R., Ciufo,S., Haddad,D., McVeigh,R., Rajput,B., Robbertse,B., Smith-White,B., Ako-Adjei,D. *et al.* (2016) Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.*, **44**, D733–D745.
21. Nordberg,H., Cantor,M., Dusheyko,S., Hua,S., Poliakov,A., Shabalov,I., Smirnova,T., Grigoriev,I.V. and Dubchak,I. (2014) The genome portal of the Department of Energy Joint Genome Institute: 2014 updates. *Nucleic Acids Res.*, **42**, D26–D31.
22. Quinlan,A.R. and Hall,I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.
23. Kanhere,A. and Bansal,M. (2005) Structural properties of promoters: similarities and differences between prokaryotes and eukaryotes. *Nucleic Acids Res.*, **33**, 3165–3175.
24. Blow,M.J., Clark,T.A., Daum,C.G., Deutschbauer,A.M., Fomenkov,A., Fries,R., Froula,J., Kang,D.D., Malmstrom,R.R., Morgan,R.D. *et al.* (2016) The epigenomic landscape of prokaryotes. *PLoS Genet.*, **12**, e1005854.
25. Bailey,T.L., Boden,M., Buske,F.A., Frith,M., Grant,C.E., Clementi,L., Ren,J., Li,W.W. and Noble,W.S. (2009) MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.*, **37**, W202–W208.
26. Buels,R., Yao,E., Diesh,C.M., Hayes,R.D., Munoz-Torres,M., Helt,G., Goodstein,D.M., Elsik,C.G., Lewis,S.E., Stein,L. *et al.* (2016) JBrowse: a dynamic web platform for genome visualization and analysis. *Genome Biol.*, **17**, 66–70.
27. Chaisson,M.J., Huddleston,J., Dennis,M.Y., Sudmant,P.H., Malig,M., Hormozdiari,F., Antonacci,F., Surti,U., Sandstrom,R., Boitano,M. *et al.* (2015) Resolving the complexity of the human genome using single-molecule sequencing. *Nature*, **517**, 608–611.
28. Gordon,D., Huddleston,J., Chaisson,M.J., Hill,C.M., Kronenberg,Z.N., Munson,K.M., Malig,M., Raja,A., Fiddes,I., Hillier,L.W. *et al.* (2016) Long-read sequence assembly of the gorilla genome. *Science*, **352**, aae0344.