D836–D841 *Nucleic Acids Research, 2008, Vol. 36, Database issue* *Published online 21 September 2007*
doi:10.1093/nar/gkm730

MethyCancer: the database of human DNA methylation and cancer

Ximiao He[1,2], Suhua Chang[1], Jiajie Zhang[1,2], Qian Zhao[1,2], Haizhen Xiang[1], Kanthida Kusonmano[1,3], Liu Yang[4,5], Zhong Sheng Sun[4], Huanming Yang[1] and Jing Wang[1,*]

[1]Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 101300, China, [2]Graduate University of the Chinese Academy of Sciences, Yuquan Road 19A, Beijing 100039, China, [3]Bioinformatics Program, School of Bioresources and Technology and School of Information Technology, King Mongkut's University of Technology Thonburi, Bangkok 10140, Thailand, [4]Behavioral Genetics Center, Institute of Psychology, Chinese Academy of Sciences, Beijing 101300, China and [5]James D. Watson Institute of Genome Sciences, Zhejiang University, Hangzhou 310007, China

Received August 15, 2007; Accepted September 3, 2007

ABSTRACT

Cancer is ranked as one of the top killers in all human diseases and continues to have a devastating effect on the population around the globe. Current research efforts are aiming to accelerate our understanding of the molecular basis of cancer and develop effective means for cancer diagnostics, treatment and prognosis. An altered pattern of epigenetic modifications, most importantly DNA methylation events, plays a critical role in tumorigenesis through regulating oncogene activation, tumor suppressor gene silencing and chromosomal instability. To study interplay of DNA methylation, gene expression and cancer, we developed a publicly accessible database for human DNA Methylation and Cancer (MethyCancer, http://methycancer.genomics.org.cn). MethyCancer hosts both highly integrated data of DNA methylation, cancer-related gene, mutation and cancer information from public resources, and the CpG Island (CGI) clones derived from our large-scale sequencing. Interconnections between different data types were analyzed and presented. Furthermore, a powerful search tool is developed to provide user-friendly access to all the data and data connections. A graphical MethyView shows DNA methylation in context of genomics and genetics data facilitating the research in cancer to understand genetic and epigenetic mechanisms that make dramatic changes in gene expression of tumor cells.

INTRODUCTION

As the world's second biggest killer after cardiovascular disease, cancer causes millions of deaths per year and is a worldwide threat to human health. Cancer has been considered as a genetic disease that involves in molecular, cellular and systematic dysfunction. Thus, the imperative task in cancer research field is to understand the mechanism of these dysfunctions, thereby facilitate to develop more effective means for cancer diagnostics, treatment and prognosis. Currently, there are several comprehensive projects devoted for such purpose, including The Cancer Genome Atlas (TCGA) in US (1) and Cancer Genome Project (CGP) in UK (http://www.sanger.ac.uk/genetics/CGP/). China recently initiated the Cancer Genome/Epigenome Project focusing on certain types of cancers that have high incidence and mortality rate in the Chinese population.

Since the discovery of epigenetic modification involved in human cancer in 1983 (2), scientists have been engaged in dissecting the roles of epigenetic abnormalities in tumorigenesis. It has been well recognized that both genetic and epigenetic modifications function at all stages of cancer development (3). An altered pattern of epigenetic modifications, most importantly DNA methylation events, plays a critical role in tumorigenesis through regulating oncogene activation (2,4–7), tumor-suppressor gene silencing (8–10) and chromosomal instability (11). Recent technological advances allow cancer epigenetics to be studied on genomic scale (12). The study of global methylation patterns of human genome and altered DNA methylation patterns in tumorigenesis offer great potential for developing strategies to provide molecular

*To whom correspondence should be addressed. Tel: +86 10 80485492; Fax: +86 10 80498676; Email: wangjing@genomics.org.cn
The authors wish it to be known that, in their opinion, the first four authors should be regarded as joint First Authors.

screening for cancer risk, early detection, prevention and treatment.

Currently there are several resources available for cancer research, such as the Cancer Genome Anatomy Project (CGAP) (13), Catalog of Somatic Mutations in Cancer (COSMIC) (14) and Atlas of Genetics and Cytogenetics in Oncology and Haematology (AGCOH) (15), all of which provide useful information regarding the cancer genetics. For study of DNA methylation, the DNA Methylation Database (MethDB) is a well-maintained resource to store DNA methylation data and to make the data readily available to the public (16). Although the Human Epigenome Project (HEP) is aiming to identify, catalog and interpret genome-wide DNA methylation patterns of all human genes in all major tissues (17,18), there is no specialized database focusing on the correlation among DNA methylation, gene expression and cancer types. On the other hand, as resources of both cancer/cancer gene and DNA methylation are complementary to each other, there is a growing need to integrate the data to provide a comprehensive dataset for convenience of in-depth data mining. MethyCancer is thus developed to study the interesting interplay between DNA methylation, gene expression and cancer through large-scale data integration, production and mining. Our MethyCancer contains both highly integrated data from public resources with manual curation, and experimental data produced from the Cancer Epigenome Project in China. As a bridge to study both DNA methylation and cancer, MethyCancer could function both as an information resource and analysis platform for study of CGI distribution in human genes, alteration of DNA methylation patterns in promoter CGIs, identification of novel cancer genes altered by DNA methylation alone or in combination with genetic events, and discovery of novel epigenetic targets.

DATA CONTENT AND DATA INTEGRATION

There are mainly four types of data included in MethyCancer: (i) CGI clones and global CGI predictions, (ii) DNA methylation data, (iii) cancer information, genes and mutations and (iv) correlation among DNA methylation, gene expression and cancer.

In order to study the global CGI distribution in human genome and alteration of DNA methylation patterns in promoter CGIs, we have generated approximately 20 000 CGI clone sequences by large-scale sequencing of the CGI library from the Sanger Institute (19). The number continues to increase with development of the new sequencing technologies. In addition, 17 132 CGI clone sequences downloaded from UHN Human CpG Island Microarray Database (20) were also integrated. Taken together, currently a total of 34 738 CGI clones have been included in MethyCancer database. By mapping the clones onto human genome (NCBI Build36) using BLAT and BLAST, all clones were clustered into 18 240 genomic loci. To verify the experimental CGI sequences and identify novel candidate CGIs, CGI predictions on the whole human genome were performed using CpGi130 (21) with criteria of: (i) length greater than 500 bp, (ii) GC content no less than 55%, and (iii) the ratio of observed CpG frequency over the expected frequency exceeds 0.65 (22). CGI predictions from UCSC (23) are also included in MethyCancer.

The DNA methylation data were mainly integrated from MethDB, HEP and Methylation Landscape of Human Genome at Columbia University (Columbia) (24). The data from MethDB contains the human DNA methylation patterns, profiles and total methylation content data for different tissues and phenotypes coming from thousands of experiments. The HEP data were produced by bisulfite DNA sequencing. The data types include analysis, trace, CpG variation and sample, where 'analysis' is a cluster of traces and 'CpG variation' refers to the position of the C of each CpG. Data from Columbia include genome-scale dataset of methylated domains and unmethylated domains in human brain tissues. Since the data from different resources are heterogeneous, we first created a uniform format to describe the diverse data and then integrated the data by mapping the corresponding sequences onto the human genome. Less than 5% of data without hits on the genome were discarded. A total of 197 493 sequences (including CGI clones and predictions) are mapped onto human genome, and these sequences were further clustered into 64 681 distinct genomic loci, which we termed MethyLoci. Here we use MethyLoci to represent both CGI and methylation data for convenience of data explanation since altered CGI methylation patterns contribute to tumorigenesis.

The data of cancer classification, pathology and cytogenetics, cancer-related genes, as well as mutations are mainly from CGAP, the website of National Cancer Institute (NCI, http://www.cancer.gov/), CGP Cancer Gene Census (25), COSMIC, AGCOH, OMIM, and hundreds of paper reviews for the top 10 cancers listed in WHO, respectively. To identify novel cancer genes, additional human genes and transcripts from Ensembl (26) were also integrated. After normalization, non-redundant gene sets were grouped into three categories with the estimates of confidence of their relationship with cancer: (i) annotated cancer gene, (ii) candidate cancer gene and (iii) other gene. All mutations were mapped to genes and thus the correlation of cancer types, cancer genes and mutations are well established.

In order to elucidate the interplay of DNA methylation, gene expression and cancer types, we annotated MethyLoci to genes by genomic mapping, and linked methylation data with cancers via cancer genes. All genes were further classified into nine categories based on this analysis, e.g. annotated/candidate cancer gene with experimental/predicted methylation data or without methylation data and other gene with experimental/predicted methylation data or without methylation data. Moreover, gene promoters with experimental validation were extracted from the Eukaryotic Promoter Database (EPD) (27), which facilitates users to study DNA methylation, especially methylation patterns in promoter CGIs, in context of precise gene structures. Gene expression data from CGAP and UniGene (28) was

Table 1. MethyCancer data content and statistics as of 1 August 2007

Data content	Data statistics
Methylation	
Methylation data	199 607
MethyLoci[a]	64 681
Tissue	145
Sample	15 005
Experiment	15 658
Cancer and gene	
Cancer	511
Cancer-related gene	7100
Annotated cancer gene	485
with methylation data	437
Candidate cancer gene	6615
with methylation data	5598
Other gene	24 020
with methylation data	11 561
Other integrated data	
EPD[b]	1794
GO[c]	6193
SNP[d]	57 790
Pathway	502
Repeat	5 012 310
Reference	27 547

[a]MehyLoci: distinct genomic loci of clustered methylation sequences.
[b]EPD: Eukaryotic Promoter Database.
[c]GO: Gene Ontology.
[d]SNP: Single Nucleotide Polymorphism.

integrated as well to provide a clue for the study of the possible changes in gene expression by alteration of methylation patterns. Accurate correlations between DNA methylation and expression profiles are difficult to establish. But this will become evident as more DNA methylation patterns with particular states of gene activity emerge (29). The inclusion of functional classifications of genes by Gene Ontology (GO) (30), proteins, pathways, mutations and references help users to interpret the data so as to identify novel cancer genes or select candidate genes for further experimental confirmation. Genome-wide repetitive sequences were predicted and included in MethyCancer for study of CGI distribution and methylation profiling since the genome of the cancer cell undergoes global hypomethylation at repetitive sequences (31).

The current version of MethyCancer contains 485 annotated cancer genes, of which 323 are supported by experimentally validated methylation data and 114 matched with CGI predictions, 6615 candidate cancer genes, of which 3698 and 1900 are supported by experimental methylation data and CGI predictions, respectively and 24 020 other genes. The data statistics dated 1 August 2007 are shown in Table 1. More detailed statistical figures and tables are available on the MethyCancer website. All data can be downloaded free of charge from our FTP site (http://methycancer.genomics.org.cn/FTP.do).

DATABASE USAGE AND ACCESS

MethyCancer provides users a powerful search engine to query different data types and data interactions housed in the database. Besides the simple keyword search, MethyCancer offers advanced searches, namely Methylation Search, Gene Search, Cancer Search, Clone Search and Repeat Search. Taken the Methylation Search as an example, users can specify and combine the query options such as methylation type (pattern, profile, content, domain), data source (BIG, UHN, MethDB, HEP, Columbia), experimental method, sample information (tissue, sex, age, phenotype) and chromosomal positions. Gene Search and Cancer Search would be the recommended starting points to surf MethyCancer, as users will be more familiar with the corresponding terms.

'Methy&Cancer' is a special module dedicated to study the relationship amongst DNA methylation, gene and cancer. Users can query the interconnections between the three data types and study the methylation status of certain cancer genes in a specific tissue. For example, if a user is interested in breast cancer and wants to know how many breast cancer genes with supportive experimental methylation data from BIG/UHN and HEP, all he or she needs to do is select two gene categories of 'Annotated/Candidate cancer genes with experimental methylation data' from the pull-down menu, then type 'breast tumor' in the search field of Cancer Name, and select BIG/UHN and HEP in Methylation Data Source. Users may also define specific expression tissues and genes with methylation data in promoter regions. A list of genes and methylation data will be returned and a factual detailed report could be shown upon demand.

As an important and efficient visualization tool, MethyView is developed to facilitate the users to browse methylation data in the context of existing genome annotations. MethyView is composed of three layers of sub-viewers in hierarchical architecture, namely ChroView, MethyLoci Survey and MethyLoci Detail View. ChroView contains an outline of a chromosome including statistics of cancer genes, methylation-related genes and methylation data. ChroView allows users to center the map onto a specific chromosome band to expand for more detailed views. MethyLoci Survey shows where the MethyLoci are located in chromosomal regions and the distribution of its sourcing data (BIG/UHN, MethDB, HEP, Columbia, CGI predictions). By clicking on MethyLoci or original methylation data from different resources, users can switch to the MethyLoci Detail View, which displays the methylation levels in the form of a matrix. Each color-coded square within the matrix represents one CpG site and the color coding represents the different methylation level (from yellow to blue, it represents from 0% to 100% methylation and gray indicates CpG sites for which methylation levels could not be determined). Clicking on a square reveals the level of methylation observed at that particular CpG site. Each row represents all the CpG sites of one methylation record shown at the bottom of the MethyLoci Detail View. Rows of squares are grouped by different data sources and tissues. If one continues to zoom in on the MethyLoci, the absolute position of each CpG site will be shown. At each layer of MethyView, gene distribution is plotted and chromosome coordinates and transcripts are also shown. By enlarging a specific gene, the methylation data will be

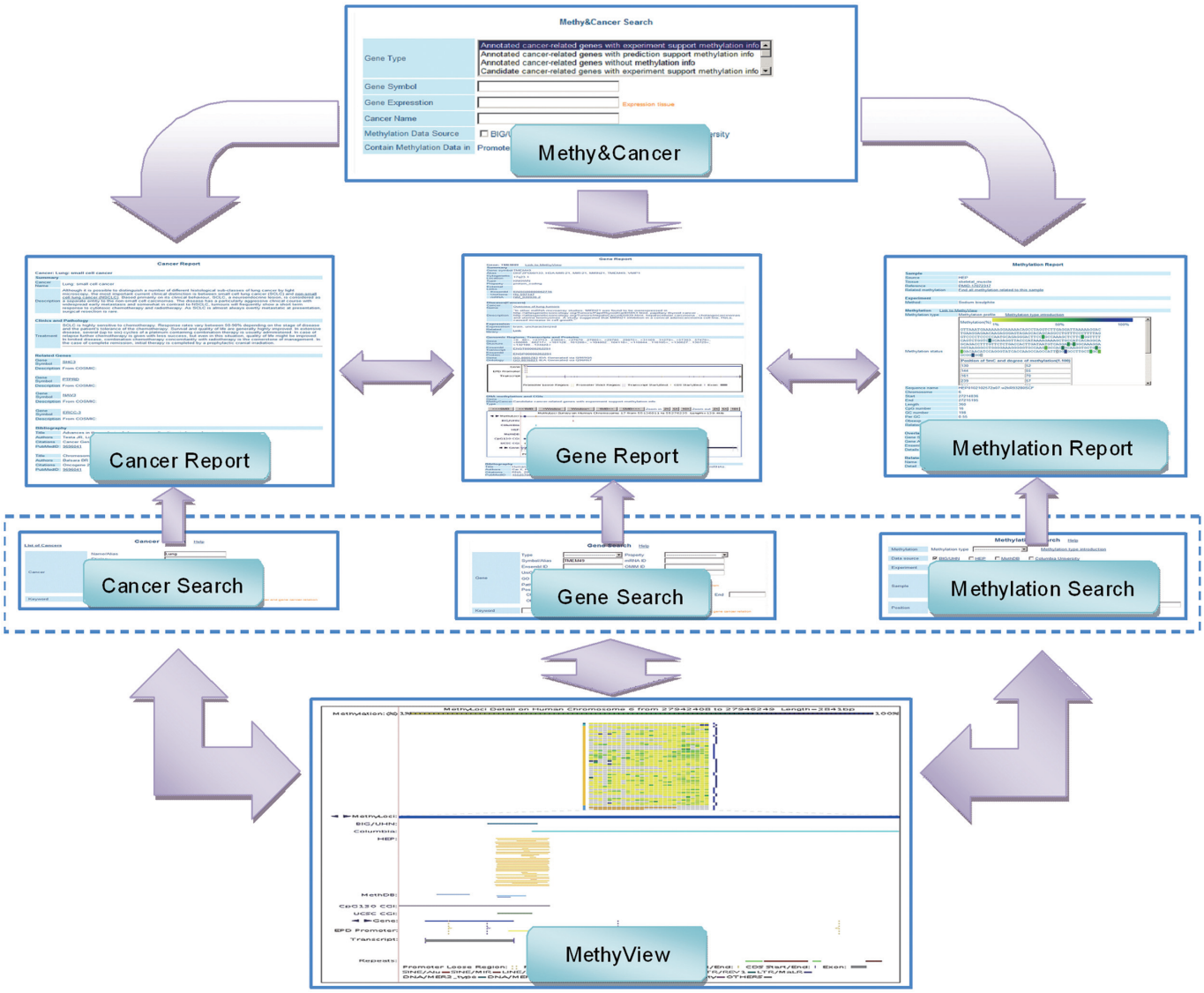

Figure 1. Screenshot showing the interrelation of MethyCancer data and tools. Users access the data through search engine and study the correlation of DNA methylation, gene expression and cancer types through Methy&Cancer. All data and data connections can be viewed through MethyView.

shown along the precise gene structure of promoters, exons and introns, as well as distribution of repetitive sequences. A factual report for each element contained in the visualization system is displayed automatically on demand. Links to MethyView are also available in various search result pages.

As shown in Figure 1, all data and tools housed in MethyCancer are crosslinked, and the whole website forms an interrelated network, which help users to utilize, analyze and understand the data more efficiently.

SYSTEM DESIGN AND IMPLEMENTATION

MethyCancer was developed using our established pipeline for biological database systems (32–35). It consists of three hardware components: a Word Wide Web server, a database server and a computational server for sequence analysis. The system is based on a MySQL relational database, and the front end consists of a set of JSP scripts running on an Apache Tomcat web server. The web services were developed using Apache Struts, a Java web application framework and Hibernate, a high-performance object/relational persistence and query service for Java, both of which help improve the software quality and ensure the stability of the web services. The back-end data analysis programs were written by PERL, which are available upon request. The BLAST and CGI prediction tool run on clusters of supercomputers.

FUTURE DEVELOPMENT

Aiming to build an integrated research platform for studying the interplay of DNA methylation, gene expression and cancer, continued efforts will be made to update

the MethyCancer data, improve data quality and database functionality. MethyCancer will also serve as a platform by which we would like to share data and analytical results from the Cancer Genome/Epigenome Project in China with colleagues all over the world. Newly sequenced CGI clones and data from our CGI array and global DNA methylation profiling study will be added as soon as they become available. Results from our gene expression profiling project will also be incorporated, and targeted to predict human transcriptome based on transcript sampling data. Considering that DNA methylation plays an important role not only in cancer but also other diseases and biological processes, such as loss of imprinting and aging, we plan to extend the scope of the diseases, and include other methylation-related diseases. Since histone modification is also recognized as hallmarks of gene silencing and involved in tumorigenesis, as more data generated from histone modification study (36), we would extend the research scope and integrate histone modification data into MethyCancer. Finally, we will improve the current tools for CGI predictions based on our comprehensive CGI clone set, and will integrate other tools for predicting methylation status of CGI and promoters. We hope our continuous efforts in MethyCancer will contribute to the improvement of global human health.

ACKNOWLEDGEMENTS

This work was sponsored by the National High Technology Research and Development Program of China (863 Program) (Grant No. 2006AA02A304) from the Ministry of Science and Technology of the People's Republic of China. Funding to pay the Open Access publication charges for this article was also provided by the 863 grant.

Conflict of interest statement. None declared.

REFERENCES

1. Collins,F.S. and Barker,A.D. (2007) Mapping the cancer genome. Pinpointing the genes involved in cancer will help chart a new course across the complex landscape of human malignancies. *Sci. Am.*, **296**, 50–57.
2. Feinberg,A.P. and Tycko,B. (2004) The history of cancer epigenetics. *Nat. Rev. Cancer*, **4**, 143–153.
3. Jones,P.A. and Baylin,S.B. (2007) The epigenomics of cancer. *Cell*, **128**, 683–692.
4. Cheah,M.S., Wallace,C.D. and Hoffman,R.M. (1984) Hypomethylation of DNA in human cancer cells: a site-specific change in the c-myc oncogene. *J. Natl Cancer Inst.*, **73**, 1057–1065.
5. De Smet,C., De Backer,O., Faraoni,I., Lurquin,C., Brasseur,F. and Boon,T. (1996) The activation of human gene MAGE-1 in tumor cells is correlated with genome-wide demethylation. *Proc. Natl Acad. Sci. USA*, **93**, 7149–7153.
6. Lee,T.S., Kim,J.W., Kang,G.H., Park,N.H., Song,Y.S., Kang,S.B. and Lee,H.P. (2006) DNA hypomethylation of CAGE promotors in squamous cell carcinoma of uterine cervix. *Ann. NY Acad. Sci.*, **1091**, 218–224.
7. Cho,B., Lee,H., Jeong,S., Bang,Y.J., Lee,H.J., Hwang,K.S., Kim,H.Y., Lee,Y.S., Kang,G.H. *et al.* (2003) Promoter hypomethylation of a novel cancer/testis antigen gene CAGE is correlated with its aberrant expression and is seen in premalignant stage of gastric carcinoma. *Biochem. Biophys. Res. Commun.*, **307**, 52–63.
8. Greger,V., Passarge,E., Hopping,W., Messmer,E. and Horsthemke,B. (1989) Epigenetic changes may contribute to the formation and spontaneous regression of retinoblastoma. *Hum. Genet.*, **83**, 155–158.
9. Sakai,T., Toguchida,J., Ohtani,N., Yandell,D.W., Rapaport,J.M. and Dryja,T.P. (1991) Allele-specific hypermethylation of the retinoblastoma tumor-suppressor gene. *Am. J. Hum. Genet.*, **48**, 880–888.
10. Clark,S.J. and Melki,J. (2002) DNA methylation and gene silencing in cancer: which is the guilty party? *Oncogene*, **21**, 5380–5387.
11. Esteller,M. and Almouzni,G. (2005) How epigenetics integrates nuclear functions. Workshop on epigenetics and chromatin: transcriptional regulation and beyond. *EMBO Rep.*, **6**, 624–628.
12. Esteller,M. (2007) Cancer epigenomics: DNA methylomes and histone-modification maps. *Nat. Rev. Genet.*, **8**, 286–298.
13. Hess,J.L. (2003) The Cancer Genome Anatomy Project: power tools for cancer biologists. *Cancer Invest.*, **21**, 325–326.
14. Bamford,S., Dawson,E., Forbes,S., Clements,J., Pettett,R., Dogan,A., Flanagan,A., Teague,J., Futreal,P.A. *et al.* (2004) The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website. *Br. J. Cancer*, **91**, 355–358.
15. Huret,J.L., Dessen,P. and Bernheim,A. (2003) Atlas of Genetics and Cytogenetics in Oncology and Haematology, year 2003. *Nucleic Acids Res.*, **31**, 272–274.
16. Amoreira,C., Hindermann,W. and Grunau,C. (2003) An improved version of the DNA Methylation database (MethDB). *Nucleic Acids Res.*, **31**, 75–77.
17. Eckhardt,F., Lewin,J., Cortese,R., Rakyan,V.K., Attwood,J., Burger,M., Burton,J., Cox,T.V., Davies,R. *et al.* (2006) DNA methylation profiling of human chromosomes 6, 20 and 22. *Nat. Genet.*, **38**, 1378–1385.
18. Bradbury,J. (2003) Human epigenome project – up and running. *PLoS Biol.*, **1**, 316–319.
19. Cross,S.H., Charlton,J.A., Nan,X. and Bird,A.P. (1994) Purification of CpG islands using a methylated DNA binding column. *Nat. Genet.*, **6**, 236–244.
20. Heisler,L.E., Torti,D., Boutros,P.C., Watson,J., Chan,C., Winegarden,N., Takahashi,M., Yau,P., Huang,T.H. *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.
21. Takai,D. and Jones,P.A. (2003) The CpG island searcher: a new WWW resource. *In Silico Biol.*, **3**, 235–240.
22. Takai,D. and Jones,P.A. (2002) Comprehensive analysis of CpG islands in human chromosomes 21 and 22. *Proc. Natl Acad. Sci. USA*, **99**, 3740–3745.
23. Kuhn,R.M., Karolchik,D., Zweig,A.S., Trumbower,H., Thomas,D.J., Thakkapallayil,A., Sugnet,C.W., Stanke,M., Smith,K.E. *et al.* (2007) The UCSC genome browser database: update 2007. *Nucleic Acids Res.*, **35**, D668–D673.
24. Rollins,R.A., Haghighi,F., Edwards,J.R., Das,R., Zhang,M.Q., Ju,J. and Bestor,T.H. (2006) Large-scale structure of genomic methylation patterns. *Genome Res.*, **16**, 157–163.
25. Futreal,P.A., Coin,L., Marshall,M., Down,T., Hubbard,T., Wooster,R., Rahman,N. and Stratton,M.R. (2004) A census of human cancer genes. *Nat. Rev. Cancer*, **4**, 177–183.
26. Hubbard,T.J., Aken,B.L., Beal,K., Ballester,B., Caccamo,M., Chen,Y., Clarke,L., Coates,G., Cunningham,F. *et al.* (2007) Ensembl 2007. *Nucleic Acids Res.*, **35**, D610–D617.
27. Schmid,C.D., Perier,R., Praz,V. and Bucher,P. (2006) EPD in its twentieth year: towards complete promoter coverage of selected model organisms. *Nucleic Acids Res.*, **34**, D82–D85.
28. Wheeler,D.L., Church,D.M., Federhen,S., Lash,A.E., Madden,T.L., Pontius,J.U., Schuler,G.D., Schriml,L.M., Sequeira,E. *et al.* (2003) Database resources of the National Center for Biotechnology. *Nucleic Acids Res.*, **31**, 28–33.
29. Murrell,A., Rakyan,V.K. and Beck,S. (2005) From genome to epigenome. *Hum. Mol. Genet.*, **14**(Spec No. 1), R3–R10.
30. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.

31. Nishiyama,R., Qi,L., Lacey,M. and Ehrlich,M. (2005) Both hypomethylation and hypermethylation in a 0.2-kb region of a DNA repeat in cancer. *Mol. Cancer Res.*, **3**, 617–626.
32. Zhao,W., Wang,J., He,X., Huang,X., Jiao,Y., Dai,M., Wei,S., Fu,J., Chen,Y. *et al.* (2004) BGI-RIS: an integrated information resource and comparative analysis workbench for rice genomics. *Nucleic Acids Res.*, **32**, D377–D382.
33. Wang,J., He,X., Ruan,J., Dai,M., Chen,J., Zhang,Y., Hu,Y., Ye,C., Li,S. *et al.* (2005) ChickVD: a sequence variation database for the chicken genome. *Nucleic Acids Res.*, **33**, D438–D441.
34. Wang,J., Xia,Q., He,X., Dai,M., Ruan,J., Chen,J., Yu,G., Yuan,H., Hu,Y. *et al.* (2005) SilkDB: a knowledgebase for silkworm biology and genomics. *Nucleic Acids Res.*, **33**, D399–D402.
35. Chang,S., Zhang,J., Liao,X., Zhu,X., Wang,D., Zhu,J., Feng,T., Zhu,B., Gao,G.F. *et al.* (2007) Influenza Virus Database (IVDB): an integrated information resource and analysis platform for influenza virus research. *Nucleic Acids Res.*, **35**, D376–D380.
36. Barski,A., Cuddapah,S., Cui,K., Roh,T.Y., Schones,D.E., Wang,Z., Wei,G., Chepelev,I. and Zhao,K. (2007) High-resolution profiling of histone methylations in the human genome. *Cell*, **129**, 823–837.