D1030–D1035 *Nucleic Acids Research, 2012, Vol. 40, Database issue* *Published online 1 December 2011*
doi:10.1093/nar/gkr1169

DiseaseMeth: a human disease methylation database

Jie Lv[1,2], Hongbo Liu[1], Jianzhong Su[1], Xueting Wu[1], Hui Liu[1], Boyan Li[1], Xue Xiao[1], Fang Wang[1], Qiong Wu[2,*] and Yan Zhang[1,*]

[1]College of Bioinformatics Science and Technology, Harbin Medical University, Harbin 150081 and [2]State Key Laboratory of Urban Water Resource and Environment, Department of Life Science and Engineering, Harbin Institute of Technology, Harbin 150001, China

Received August 13, 2011; Revised November 12, 2011; Accepted November 13, 2011

ABSTRACT

DNA methylation is an important epigenetic modification for genomic regulation in higher organisms that plays a crucial role in the initiation and progression of diseases. The integration and mining of DNA methylation data by methylation-specific PCR and genome-wide profiling technology could greatly assist the discovery of novel candidate disease biomarkers. However, this is difficult without a comprehensive DNA methylation repository of human diseases. Therefore, we have developed DiseaseMeth, a human disease methylation database (http://bioinfo.hrbmu.edu.cn/diseasemeth). Its focus is the efficient storage and statistical analysis of DNA methylation data sets from various diseases. Experimental information from over 14 000 entries and 175 high-throughput data sets from a wide number of sources have been collected and incorporated into DiseaseMeth. The latest release incorporates the gene-centric methylation data of 72 human diseases from a variety of technologies and platforms. To facilitate data extraction, DiseaseMeth supports multiple search options such as gene ID and disease name. DiseaseMeth provides integrated gene methylation data based on cross-data set analysis for disease and normal samples. These can be used for in-depth identification of differentially methylated genes and the investigation of gene–disease relationship.

INTRODUCTION

DNA methylation is one of the enzymatic modifications in mammalian genomes (1,2). The methylated cytosines are almost exclusively in a CpG dinucleotide sequence. DNA methyltransferases (DNMTs) are the main enzymes that catalyze CpG methylation. DNA methyltransferase 1 (DNMT1) is responsible for the post-replicative copying of preexisting CpG methylation patterns, while DNMT3A and DNMT3B are responsible for *de novo* DNA methylation (3). Previous studies suggested that DNA methylation abnormality is one of the most frequent epigenetic events in human diseases, and DNA methylation patterns in disease tissue are different from those in their normal counterparts (4,5). Aberrant hypomethylation may lead to genome instability, transcriptional activation of oncogenes, loss of imprinting, while hypermethylation in local regions may be related to the selective advantage for cancer cells (6–8). Because of the important roles of promoter methylation in functional regulation, DNA methylation has been studied extensively in diseases such as neurodevelopmental disorders, neurodegenerative and neurological diseases, autoimmune diseases and cancers. Some important cases have been reported, they include neprilysin (NEP) in Alzheimer's disease, frataxin (FXN) in Friedreich's ataxia (9), survival of motor neuron (SMN2) in spinal muscular atrophy (9), methylguanine-DNA methyltransferase (MGMT) in colorectal cancer (5,10), prolactin receptor (PRLR) in breast cancer (11), methyl CpG binding protein (MeCP2) in Rett syndrome (12) and imprinting controlled region at 15q11–q13 in Prader–Willi and Angelman syndromes (13).

Locus-specific approaches, like methylation specific PCR, pyrosequencing and bisulfite sequencing, were widely used in laboratories. Recently, high-throughput approaches based on array and next-generation sequencing for genome-wide analysis have been favored (14). A collection of the disease methylation data produced by these techniques will be useful and could be used to explore the potential methylation markers/phenotypes from whole methylomes in human disease states.

*To whom correspondence should be addressed. Tel/Fax: +86 451 86667543; Email: yanyou1225@yahoo.com.cn
Correspondence may also be addressed to Qiong Wu. Tel/Fax: +86 451 86403181; Email: kigo@hit.edu.cn

The authors wish it to be known that, in their opinion, the first three authors should be regarded as joint First Authors.

In general, there are two types of methylation databases, experimental evidence databases and large-scale databases. The earlier methylation databases include MethDB (15), MethPrimerDB (16), MethCancerDB (17), PubMeth (18) and MeInfoText (19). MethDB holds information about the occurrence of methylated cytosines in the DNA assay information obtained from multiorganism CpG methylation analysis. MethPrimerDB stores confirmed primer and experimental information on four PCR methods for CpG methylation analysis. MethCancerDB provides documentation of pre-existing information regarding DNA methylation in various cancers that includes study size, type of cancer and method used. PubMeth and MeInfoText are based on text-mining of Medline/PubMed abstracts to extract information on methylation in cancer. In contrast to these databases, there are two large-scale methylation databases MethyCancer (20) and NGSmethDB (21). MethyCancer hosts large-scale methylation reference data, cancer-related genes and cancer information from public data sources. NGSmethDB hosts several sequence-based reference methylation data sets that can be used to gain gene specific and differential methylation information. However, there is a need for a database that can integrate the dispersed data and provide a convenient way for in-depth data mining. To this end, we developed DiseaseMeth to combine experimental methylation information from loci-specific technologies with inferred gene-centric methylation states from methylation profiling technology.

Various laboratories have profiled the methylomes of a few human diseases [breast cancer (22,23) and leukemia (24,25)], producing data that could be integrated to gain further knowledge. It should be possible to identify differentially methylated genes in diseases by integrating data sets of the same disease. Cross-data set analysis for specific diseases is useful because it is difficult and costly for experimental biologists to discover potentially novel genes/regions in diseases. Furthermore, results are often contradictory and need further confirmation.

To tackle these challenges, DiseaseMeth was developed to store and mine data efficiently through a user-friendly extraction interface. The current release of DiseaseMeth incorporates 72 disease types. Moreover, DiseaseMeth stores many reference methylation data sets derived from normal tissues/cells that can be used to identify aberrantly methylated genes, and genomic data such as CpG islands, histone modifications and annotated genes. In addition, DiseaseMeth provides: (i) search options that can be used to statistically identify gene-centric methylation differences, extract detailed information of differentially methylated genes in disease compared with normal tissue, and calculate the significance of a specific differentially methylated gene; (ii) tools to calculate the correlation of DNA methylation with the pairwise relationships of gene–gene, gene–disease and disease–disease, which could help in the discovery of disease-specific and disease-consistent genes/markers; and (iii) a genome methylation browser and customized views that display gene-centric disease methylation information combined with genomic information on the genomic scale. In brief, DiseaseMeth hosts comprehensive disease methylation data and provides tools to explore the relationships between diseases and DNA methylation.

OVERVIEW AND DATA PROCESSING

DiseaseMeth includes literature-based experimental information and large-scale methylation data. Over 14 000 entries for experimental information were collected by text mining from more than 25 000 published paper of DNA methylation research in PubMed. DiseaseMeth also holds 175 large-scale methylation data sets for 50 diseases, which were primarily collected from various websites and institutes, such as ArrayExpress, Gene Expression Omnibus (GEO), UCSC. Detailed and updated statistics of the number of disease types is shown on the DiseaseMeth homepage. A summary of the publications sourced and the links to the sites from which the data were downloaded is maintained and updated on the download page. The download page lists detailed information about the data sets including name/ID, disease, data analysis, publication link, experimental platform, sample size and download link. For raw high-throughput data, only information about methylation within the promoter region of RefSeq genes was kept for analysis. We defined the promoter region as the region from 1.5 kb upstream of the transcription start site (TSS) to 500 bp downstream of the TSS of RefSeq genes. For data from assemblies other than the UCSC March 2006 human reference sequence (hg18, NCBI build 36.1), we used the LiftOver tool from UCSC (26) to convert the coordinates from other assemblies to hg18. Normalized and standardized data (0–100%) was used directly. Unnormalized data was transformed according to the common procedures. All normalized methylation data were subsequently transformed into a consistent interval [0,100] by percentile normalization reflecting relevant methylation levels before being finally stored. All data available for download is stored in GFF (General Feature Format) files. The basic operations in the DiseaseMeth database are search, view, download and analyze (Figure 1). A flexible search engine based on a MySQL backend is provided to allow user-friendly data mining and downloading. The methylation information in DiseaseMeth with a few other annotations can be viewed using the visualization module based on the Perl Bio::Graphics package.

DATABASE USAGE AND ACCESS

Using the search tool to retrieving the methylation states of promoters

All methylation states for a given chromosomal region can be retrieved, when the start and end chromosome coordinates are provided. The data for a selected disease, tissue, cell line, technology and gene ID can also be retrieved. The more query parameters that are provided the narrower the range of the entries that are retrieved. A valid RefSeq ID (NM*) or an official gene

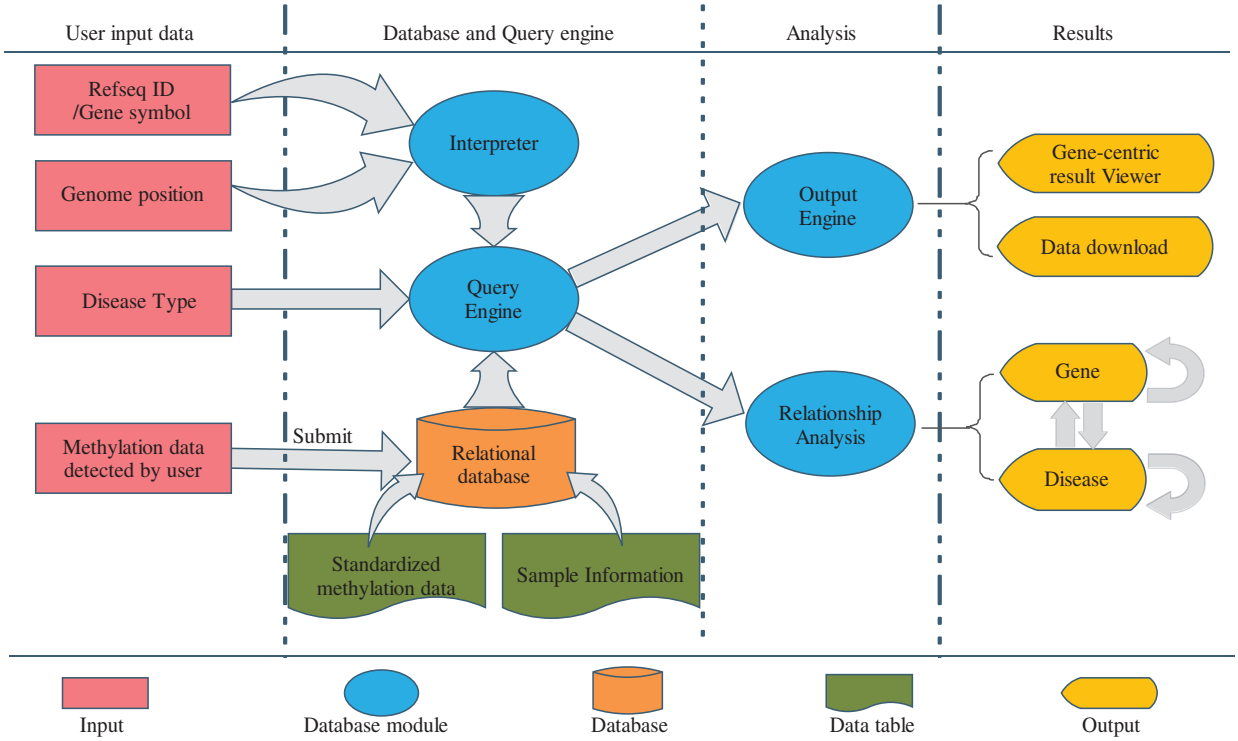

Figure 1. Overview of structure and workflow of DiseaseMeth. Users can input Refseq ID, Gene symbol or genomic position to the query engine to gain the methylation pattern of these regions in different samples. The terms imported by users would be transformed into the genomic coordinates which are further used to search the relational database of DiseaseMeth. Users also can restrict the disease type. The query results can be viewed in the gene-centric result viewer, and downloaded as flat format. The relationship analysis of module is provided for users to investigate the relationships among genes and diseases.

symbol is needed to obtain the gene-centric methylation information. The output is displayed by default as an overview table that summarizes the methylation states of the genes and gives disease and other information. The table contains links to generate gene-centric methylation information panels for disease and normal samples based on the specified search parameters. As an example, we analyzed the promoter region of the gene *RASSF1* (Figure 2). The results show that the promoter of the transcript is differentially methylated between the disease and control states. To facilitate viewing genes and other relevant information, gene-specific links to other resources such as the GeneCards database are displayed in the overview panel and in the gene-centric information table. The full results of a query can be downloaded from a link in the overview panel.

Genomic methylation viewer

DiseaseMeth includes a user-friendly and configurable genome browser through which multiple genomic and epigenomic resources can be visualized simultaneously (Figure 2). The genomic methylation viewer connecting to a MySQL backend is used to show the methylation landscape for disease methylation information, other genomic annotations [GC content, genes and CpG islands (27)] and epigenomic information (methylation data sets and histone modifications). Features of the viewer include the ability to zoom through the given regions, to enter a region by specifying the genomic coordinates, to change the order of tracks and to show and hide certain feature tracks and configure the appearance of the displayed information. Currently, a few epigenomic tracks can be configured in the viewer: (i) HAIB RRBS tracks from ENCODE/HudsonAlpha tracks; (ii) RRBS tracks from BI Human Reference Epigenome Mapping Project; (iii) MeDIP-chip; (iv) MeDIP-seq; and (v) Histone modification. In the viewer, the methylation values do not indicate strand-specific information. A color gradient from white (methylation value = 0) to red (methylation value = 100) is used to display the numeric methylation states of the cytosines/regions. The browser can also link out to other epigenomic databases such as MethyCancer (20) and HHMD (28).

Analysis tools to explore the relationships between genes and diseases

In disease, a few genes, including tumor suppressor genes, are silenced by promoter CpG island methylation (13). However, only a subset of colorectal cancers has been documented as exhibiting promoter methylation and is referred to as the CpG island methylator phenotype (CIMP) (29). In addition, gene–gene and disease–disease relationships have been characterized using available gene expression data (30). Because the availability of DNA methylation data is low, it is inefficient to analyze the relationships of genes and diseases using DNA methylation. However, with the data stored in DiseaseMeth, it is easy to determine a preliminary profile of the relationship

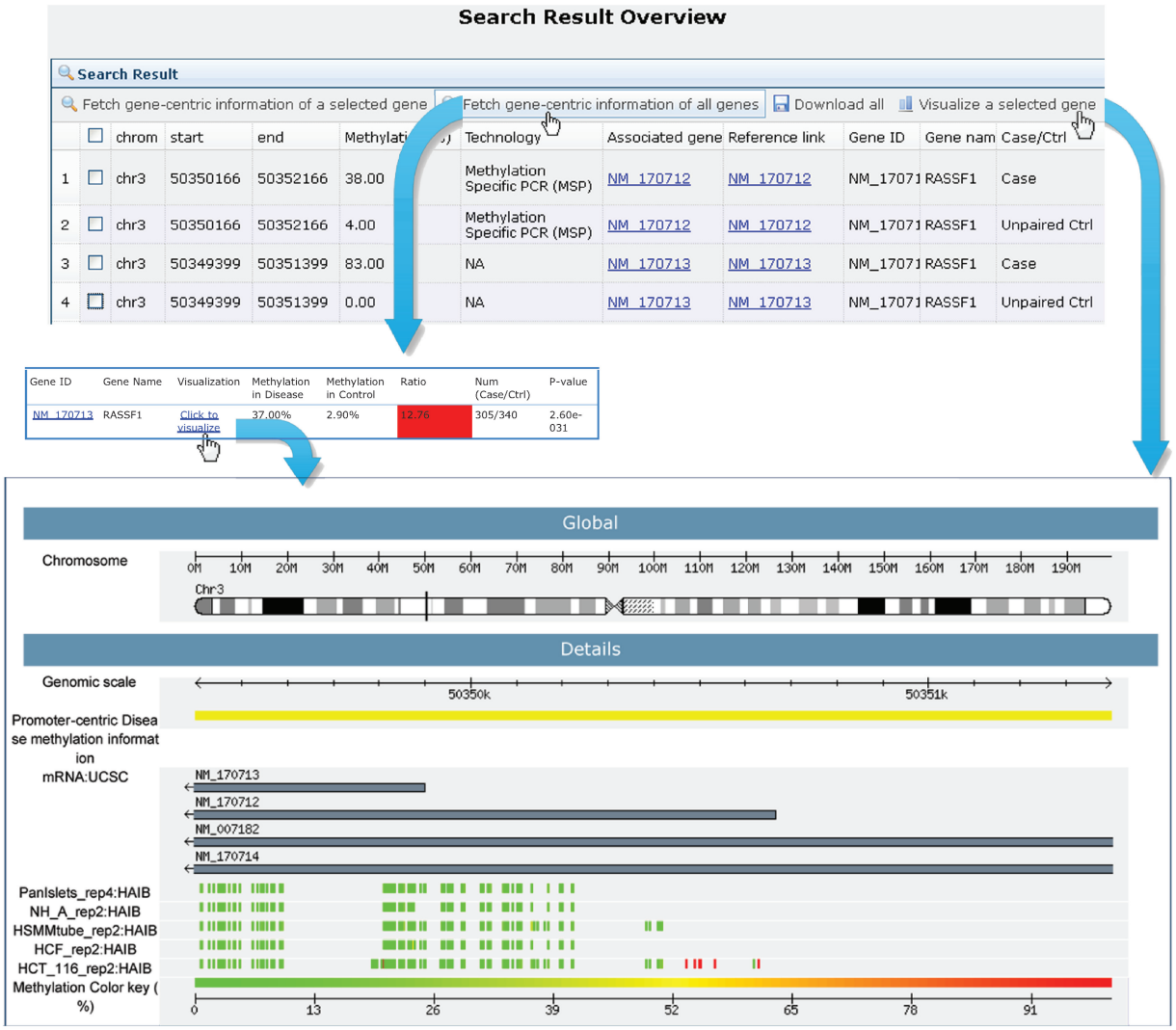

Figure 2. A screen shot of the DiseaseMeth search results for the gene *RASSF1*. The default view generated by the search tool is shown. Clicking the 'Fetch gene-centric information of all genes' button in the toolbar displays the gene-centric results, where the gene ID, gene Name, methylation level (from 0% to 100%), the number of relevant data in the database, and the significance of the methylation difference between disease and normal data sets for the genes are shown. In addition, the relevant reference links are also included in the overview panel. Concurrent searching of multiple genes is supported. In the gene-centric panel, a link (Visualization) is available to display the epigenomic data in the genomic context. There is also a 'Visualize a selected gene' button in the toolbar in the default view that does the same task. The whole of the search results can be downloaded by clicking the 'Download all' button in the toolbar.

between genes and diseases using the quantitative tools that have been developed: (i) disease–gene relationship analysis tool; (ii) disease–disease relationship analysis tool; and (iii) gene–gene relationship analysis tool. One of the merits of these tools is that they are highly customizable for analyzing given regions, diseases and so on; thus, facilitating specific analysis focusing on continuous regions such as imprinting clusters (31). The three analysis tools are available for download.

SYSTEM DESIGN AND IMPLEMENTATION

DiseaseMeth consists of three major software components: an Apache HTTP server, a MySQL database and a Perl installation using the Bio::Perl, Bio::Graphics and DBI packages. The backend data analysis programs were written in Perl and deployed as CGI programs. The Perl programs for the analysis tools are available on the website.

FUTURE DEVELOPMENT

To build a DNA methylation database focusing on human diseases, continued efforts will be made to update the DiseaseMeth data and improve the genomic methylation viewer and database functionality. As the DNA methylation data becomes available, we will continuously collect the latest disease data sets to keep DiseaseMeth up-to-date. Because of the usefulness of reference methylation maps in human DNA methylation analysis, we will include more methylation maps of normal cell lines/tissues in DiseaseMeth to help with comparative studies of

disease-specific from normal methylomes. We will develop new data processing algorithms to handle the large-scale nature of DNA methylation sequencing data. Because of the importance of integrative analysis, we will regularly collect data from new sources to enhance the analytical depths of DiseaseMeth. We will also encourage new data to be submitted directly to DiseaseMeth to keep DiseaseMeth updated and to make it comprehensive. The genomic methylation viewer will be improved to display more (epi)genomic resources and will be extended to include more configurable functionalities. Finally, novel analysis tools will be developed to provide better integration and to enhance the data mining capabilities. As a resource to study the potential regulatory function of DNA methylation, DiseaseMeth can be extended to include more data sets and tools can be developed for the identification of disease-related DNA methylation markers for candidate genes using an integrated differential methylation identification algorithm (32). We expect that the continuous efforts to use and improve DiseaseMeth will contribute to our understanding of DNA methylation driven human diseases.

ACKNOWLEDGEMENTS

The authors would like to thank Dr Yaoping Lei for revising the manuscript.

FUNDING

Funding for open access charge: National Natural Science Foundation of China (61075023 and 30971645); Natural Science Foundation of Heilongjiang Province (C201012); State Key Laboratory of Urban Water Resource and Environment (2010TS05) and Scientific Research Fund of Heilongjiang Provincial Education Department (12511272).

Conflict of interest statement. None declared.

REFERENCES

1. Kouzarides,T. (2007) Chromatin modifications and their function. *Cell*, **128**, 693–705.
2. Kim,J.K., Samaranayake,M. and Pradhan,S. (2009) Epigenetic mechanisms in mammals. *Cell. Mol. Life Sci.*, **66**, 596–612.
3. Jiang,Y.H., Bressler,J. and Beaudet,A.L. (2004) Epigenetics and human disease. *Annu. Rev. Genomics Hum. Genet.*, **5**, 479–510.
4. Feinberg,A.P. (2007) Phenotypic plasticity and the epigenetics of human disease. *Nature*, **447**, 433–440.
5. Esteller,M. (2008) Epigenetics in cancer. *N. Engl. J. Med.*, **358**, 1148–1159.
6. Brena,R.M. and Costello,J.F. (2007) Genome-epigenome interactions in cancer. *Hum. Mol. Genet.*, **16(Spec 1)**, R96–R105.
7. Szyf,M. (2003) Targeting DNA methylation in cancer. *Ageing Res. Rev.*, **2**, 299–328.
8. Lv,J., Su,J., Wang,F., Qi,Y., Liu,H. and Zhang,Y. (2010) Detecting novel hypermethylated genes in breast cancer benefiting from feature selection. *Comput. Biol. Med.*, **40**, 159–167.
9. Urdinguio,R.G., Sanchez-Mut,J.V. and Esteller,M. (2009) Epigenetic mechanisms in neurological diseases: genes, syndromes, and therapies. *Lancet Neurol.*, **8**, 1056–1072.
10. Shen,L., Kondo,Y., Rosner,G.L., Xiao,L., Hernandez,N.S., Vilaythong,J., Houlihan,P.S., Krouse,R.S., Prasad,A.R., Einspahr,J.G. *et al.* (2005) MGMT promoter methylation and field defect in sporadic colorectal cancer. *J. Natl. Cancer Inst.*, **97**, 1330–1338.
11. Esteller,M. (2007) Cancer epigenomics: DNA methylomes and histone-modification maps. *Nat. Rev. Genet.*, **8**, 286–298.
12. Shahbazian,M.D. and Zoghbi,H.Y. (2001) Molecular genetics of Rett syndrome and clinical spectrum of MECP2 mutations. *Curr. Opin. Neurol.*, **14**, 171–176.
13. Robertson,K.D. (2005) DNA methylation and human disease. *Nat. Rev. Genet.*, **6**, 597–610.
14. Feinberg,A.P. (2010) Genome-scale approaches to the epigenetics of common human disease. *Virchows Arch.*, **456**, 13–21.
15. Amoreira,C., Hindermann,W. and Grunau,C. (2003) An improved version of the DNA methylation database (MethDB). *Nucleic Acids Res.*, **31**, 75–77.
16. Pattyn,F., Hoebeeck,J., Robbrecht,P., Michels,E., De Paepe,A., Bottu,G., Coornaert,D., Herzog,R., Speleman,F. and Vandesompele,J. (2006) methBLAST and methPrimerDB: web-tools for PCR based methylation analysis. *BMC Bioinformatics*, **7**, 496.
17. Lauss,M., Visne,I., Weinhaeusel,A., Vierlinger,K., Noehammer,C. and Kriegner,A. (2008) MethCancerDB–aberrant DNA methylation in human cancer. *Br. J. Cancer*, **98**, 816–817.
18. 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.
19. Fang,Y.C., Huang,H.C. and Juan,H.F. (2008) MeInfoText: associated gene methylation and cancer information from text mining. *BMC Bioinformatics*, **9**, 22.
20. 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.
21. 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.
22. Fang,F., Turcan,S., Rimner,A., Kaufman,A., Giri,D., Morris,L.G., Shen,R., Seshan,V., Mo,Q., Heguy,A. *et al.* (2011) Breast cancer methylomes establish an epigenomic foundation for metastasis. *Sci. Transl. Med.*, **3**, 75ra25.
23. Flanagan,J.M., Cocciardi,S., Waddell,N., Johnstone,C.N., Marsh,A., Henderson,S., Simpson,P., da Silva,L., Khanna,K., Lakhani,S. *et al.* (2010) DNA methylome of familial breast cancer identifies distinct profiles defined by mutation status. *Am. J. Hum. Genet.*, **86**, 420–433.
24. Davidsson,J., Lilljebjorn,H., Andersson,A., Veerla,S., Heldrup,J., Behrendtz,M., Fioretos,T. and Johansson,B. (2009) The DNA methylome of pediatric acute lymphoblastic leukemia. *Hum. Mol. Genet.*, **18**, 4054–4065.
25. Rahmatpanah,F.B., Carstens,S., Guo,J., Sjahputera,O., Taylor,K.H., Duff,D., Shi,H., Davis,J.W., Hooshmand,S.I., Chitma-Matsiga,R. *et al.* (2006) Differential DNA methylation patterns of small B-cell lymphoma subclasses with different clinical behavior. *Leukemia*, **20**, 1855–1862.
26. Fujita,P.A., Rhead,B., Zweig,A.S., Hinrichs,A.S., Karolchik,D., Cline,M.S., Goldman,M., Barber,G.P., Clawson,H., Coelho,A. *et al.* (2011) The UCSC Genome Browser database: update 2011. *Nucleic Acids Res.*, **39**, D876–D882.
27. Su,J., Zhang,Y., Lv,J., Liu,H., Tang,X., Wang,F., Qi,Y., Feng,Y. and Li,X. (2010) CpG_MI: a novel approach for identifying functional CpG islands in mammalian genomes. *Nucleic Acids Res.*, **38**, e6.
28. Zhang,Y., Lv,J., Liu,H., Zhu,J., Su,J., Wu,Q., Qi,Y., Wang,F. and Li,X. (2010) HHMD: the human histone modification database. *Nucleic Acids Res.*, **38**, D149–D154.
29. Samowitz,W.S., Albertsen,H., Herrick,J., Levin,T.R., Sweeney,C., Murtaugh,M.A., Wolff,R.K. and Slattery,M.L. (2005) Evaluation

of a large, population-based sample supports a CpG island methylator phenotype in colon cancer. *Gastroenterology*, **129**, 837–845.
30. Linghu,B., Snitkin,E.S., Hu,Z., Xia,Y. and Delisi,C. (2009) Genome-wide prioritization of disease genes and identification of disease-disease associations from an integrated human functional linkage network. *Genome Biol.*, **10**, R91.
31. Ferguson-Smith,A.C. (2011) Genomic imprinting: the emergence of an epigenetic paradigm. *Nat. Rev. Genet.*, **12**, 565–575.
32. Zhang,Y., Liu,H., Lv,J., Xiao,X., Zhu,J., Liu,X., Su,J., Li,X., Wu,Q., Wang,F. *et al.* (2011) QDMR: a quantitative method for identification of differentially methylated regions by entropy. *Nucleic Acids Res.*, **39**, e58.