D842–D846 *Nucleic Acids Research, 2008, Vol. 36, Database issue* *Published online 11 October 2007*
doi:10.1093/nar/gkm788

PubMeth: a cancer methylation database combining text-mining and expert annotation

Maté Ongenaert*, Leander Van Neste, Tim De Meyer, Gerben Menschaert, Sofie Bekaert and Wim Van Criekinge

Department of Molecular Biotechnology, Faculty of Bioscience Engineering, Laboratory for Bioinformatics and Computational Genomics, Ghent University, B-9000 Ghent, Belgium

Received August 13, 2007; Revised September 15, 2007; Accepted September 17, 2007

ABSTRACT

Epigenetics, and more specifically DNA methylation is a fast evolving research area. In almost every cancer type, each month new publications confirm the differentiated regulation of specific genes due to methylation and mention the discovery of novel methylation markers. Therefore, it would be extremely useful to have an annotated, reviewed, sorted and summarized overview of all available data. PubMeth is a cancer methylation database that includes genes that are reported to be methylated in various cancer types. A query can be based either on genes (to check in which cancer types the genes are reported as being methylated) or on cancer types (which genes are reported to be methylated in the cancer (sub) types of interest). The database is freely accessible at http://www.pubmeth.org.

PubMeth is based on text-mining of Medline/PubMed abstracts, combined with manual reading and annotation of preselected abstracts. The text-mining approach results in increased speed and selectivity (as for instance many different aliases of a gene are searched at once), while the manual screening significantly raises the specificity and quality of the database. The summarized overview of the results is very useful in case more genes or cancer types are searched at the same time.

INTRODUCTION

DNA methylation represents a modification of DNA by addition of a methyl group to a cytosine, also referred to as the fifth base (1). This reaction uses *S*-adenosylmethionine as a methyl donor and is catalysed by a group of enzymes, the DNA methyltransferases (DNMTs). In humans and other mammals, this epigenetic modification is almost exclusively imposed on cytosines that precede a guanosine in the primary DNA sequence (often called a CpG dinucleotide). The frequency of these CpGs in the genome is much lower than would be expected as a methylated cytosine often is subject to deamination thereby forming thymidine. However, in some regions, dense clusters of CpGs can be identified: these regions are referred to as CpG islands (2).

DNA methylation is an epigenetic change: it does not alter the primary DNA sequence and might contribute to overall genetic stability and maintenance of chromosomal integrity. Consequently, it facilitates the organization of the genome into active and inactive regions with respect to gene transcription (3). Genes with CpG islands in their promoter region are generally unmethylated in normal tissues. Upon DNA hypermethylation, transcription of the affected genes may be blocked, resulting in gene silencing. In neoplasia, abnormal patterns of DNA methylation have been recognized. Hypermethylation is now considered one of the important mechanisms resulting in silenced expression of tumor suppressor genes, i.e. genes responsible for control of normal cell differentiation and/or inhibition of cell growth. In the last few years, new hypermethylated biomarkers have been used in cancer research and diagnostics (4).

MethDB (5), one of the few databases that focus on DNA methylation, is general and sample oriented. But it is not optimized to cancer-related queries because this type of query requires a summarized overview. However, in MethDB querying multiple genes or cancer types is not supported and data is always handled as a separate sample. Another database, MethPrimerDB (6), has a focus on detection methodologies (e.g. MSP primer design). Both databases discussed here, depend on submissions by administrators or users, which guarantees the required quality of the databases, but consequently they are not always complete and up to date. The databases are neither designed to rank and summarize cancer-related information [genes and cancer (sub)types

*To whom correspondence should be addressed. Tel: +32 9 264 99 22; Fax: +32 9 264 62 19; Email: mate.ongenaert@ugent.be

© 2007 The Author(s)
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.0/uk/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

involved], although this is crucial in applied methylation research in the cancer field.

Hereby we present PubMeth, a database that combines a text-mining approach (fast, intelligent to search multiple aliases and textual variants of these aliases, querying multiple keyword lists at once) with a manual reviewing and annotation step. The latter one drastically improves specificity and annotation quality. The interface is able to rank, summarize and represent data, making the information the database contains easily accessible.

The reviewing step also heavily depends on the text-mining step that sorts abstracts, highlights terms and provides links to different sources. This way, the reviewing step can be done fast and accurate enough to process all abstracts, electronically published until now in PubMed. In addition, using this approach, an update strategy can be more easily implemented.

DNA methylation in cancer research has evolved to a mainstream research topic. Methylation profiles are successfully used in early detection and personalized treatment. However, more and more data is available, especially with the availability of large-scale screening techniques. All the information taken together determines the knowledge of the 'cancer methylome'. Ultimately, the epigenome of all cancer tissues, including those of different stage and grade, could be mapped out. Epigenetic states differ widely among tissues, and changes are far more varied and much more frequent per tumor than DNA mutations. 'Each differentiated cell has a different epigenome', said Jones (7). In this perspective, it is very useful to extract which genes are already reported in which cancer types from literature. This information might be used as positive controls, to check the same genes in other (related) cancer types, to screen for markers that could be used as early diagnostic utility or in the context of personalized medicine and to deepen the knowledge of the mechanisms of methylation.

PubMeth tries to contain and summarize as many available literature data and presents them in a easy to use graphical interface. It speeds up the process of searching relevant literature, many aliases and keywords are searched at the same time and the results are reliable as they are manually reviewed as one would do when performing a manual literature search.

FILLING UP THE DATABASE

Abstracts, related to epigenetics and methylation, are downloaded in XML-format through NCBI E-Utils (E-fetch) using more than 15 methylation-related keywords (such as methylation, DNA methylation, methylated, epigenetic and a range of variants, as well as detection technologies). The aliases, symbols and descriptions of human genes, associated with an Ensembl ID, are obtained using a Perl-script. This queries the GeneCards database (8) that already combines different genetic databases such as Ensembl and Entrez Gene. Different textual variations of all aliases are generated to be as complete as possible (e.g. variants for BRCA1 include BRCA 1, BRCA-1, BRCAI, BRCA I and BRCA-I). To avoid counting and highlighting aliases that are also common English words, an alias is rejected if more than 100 000 PubMed abstracts are retrieved. A list (http://www.wordcount.org) of frequently used English words is searched at the same time.

Cancer-related keywords were obtained from a list of the National Cancer Institute (http://www.cancer.gov/cancertopics/alphalist) and keywords related with detection methodologies were manually compiled.

One by one, abstracts are searched for aliases and their variants, methylation-related keywords, sentences with both an alias and such a keyword. In addition, terms related with cancer and detection methodology are also highlighted and counted. This information is stored in a MySQL 5 relational database using Perl-DBI. Based on the information in this database, abstracts are ranked. This ranking is based on a large number of parameters such as the number of aliases, the number of different genes, the number of different aliases per gene, the number of sentences with both an alias and a methylation-related keyword, the presence of detection methodology and cancer-related keywords.

Abstracts are then manually reviewed, taking into account the order after ranking, with the aid of highlighting the different keyword lists, aliases and sentences with alias and methylation-related keyword in different colors. Aliases are linked with gene information using hover-over effects generated with JavaScript and CSS. After manual reviewing, the information in the database only has to be minimally updated or corrected. A schematic overview of the complete process is given in Figure 1. This process is still in progress; due to the ranking system the most important publications are currently in the database. The remaining abstracts will be reviewed soon, and an accurate update strategy will be developed.

QUERYING THE DATABASE

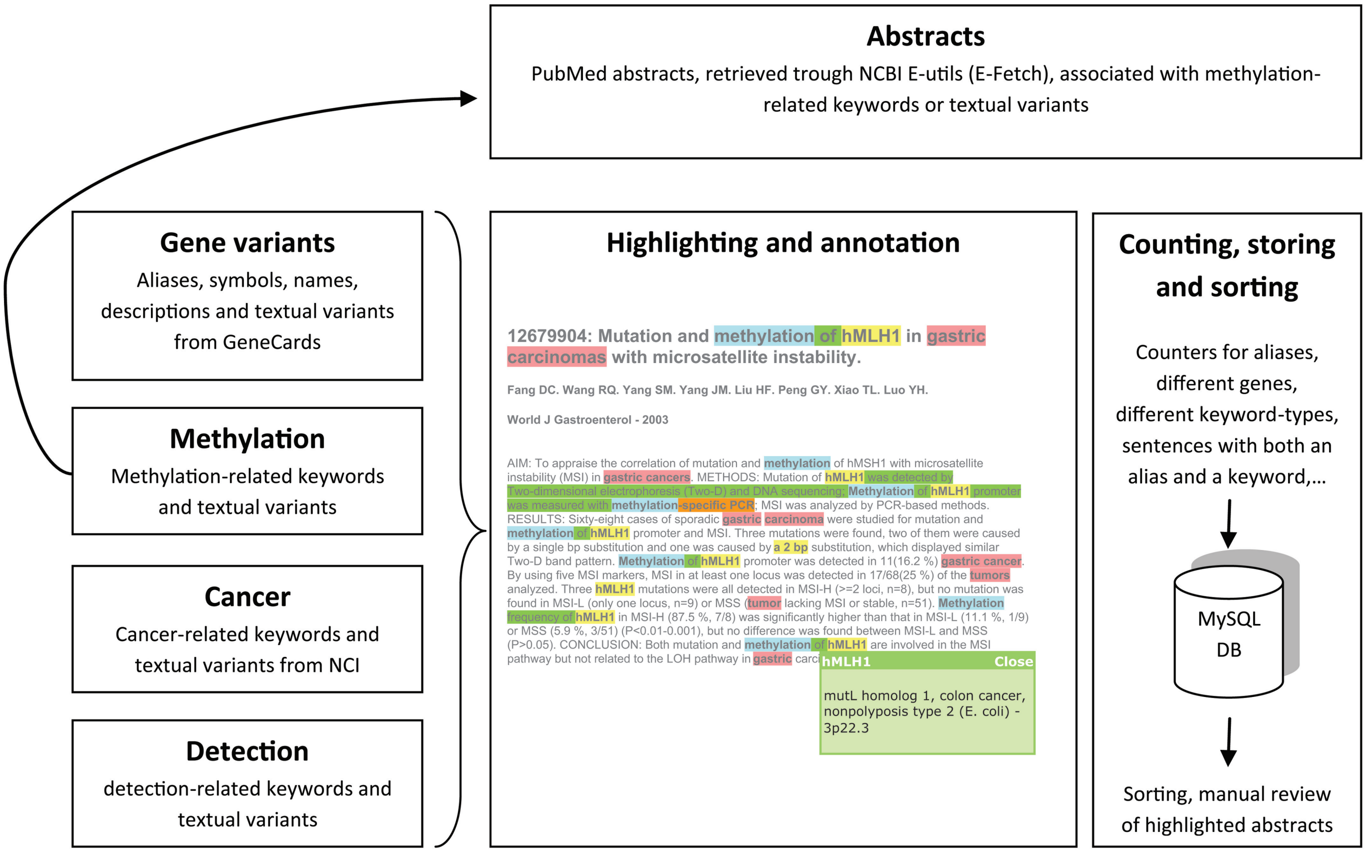
A record in the database contains information about the source of publication, the gene, the cancer type and subtypes if specified. It includes the number of primary cancer samples where methylation is analysed in, as well as the number of analysed cell lines and the number of normal tissues. For all these three categories the methylation frequency (the percentage of the samples that show methylation) is also available. Other information includes the detection technologies used and an 'evidence sentence' where most of the information in this record came from.

PubMeth can be queried using the web-interface at http://www.pubmeth.org in two ways, depending on the researcher's focus:

(1) Gene-related: in which cancer types (and subtypes) the genes of interest are reported to be methylated
(2) Cancer-related: which genes are reported to be methylated in the cancer types/subtypes

Gene-centric query

A query is created in two easy steps. In the first step, the user provides a list of genes (different identifiers are



Abstracts

PubMed abstracts, retrieved trough NCBI E-utils (E-Fetch), associated with methylation-related keywords or textual variants

Gene variants

Aliases, symbols, names, descriptions and textual variants from GeneCards

Methylation

Methylation-related keywords and textual variants

Cancer

Cancer-related keywords and textual variants from NCI

Detection

detection-related keywords and textual variants

Highlighting and annotation

12679904: Mutation and methylation of hMLH1 in gastric carcinomas with microsatellite instability.

Fang DC. Wang RQ. Yang SM. Yang JM. Liu HF. Peng GY. Xiao TL. Luo YH.

World J Gastroenterol - 2003

AIM: To appraise the correlation of mutation and methylation of hMSH1 with microsatellite instability (MSI) in gastric cancers. METHODS: Mutation of hMLH1 was detected by Two-dimensional electrophoresis (Two-D) and DNA sequencing; Methylation of hMLH1 promoter was measured with methylation-specific PCR; MSI was analyzed by PCR-based methods. RESULTS: Sixty-eight cases of sporadic gastric carcinoma were studied for mutation and methylation of hMLH1 promoter and MSI. Three mutations were found, two of them were caused by a single bp substitution and one was caused by a 2 bp substitution, which displayed similar Two-D band pattern. Methylation of hMLH1 promoter was detected in 11(16.2 %) gastric cancer. By using five MSI markers, MSI in at least one locus was detected in 17/68(25 %) of the tumors analyzed. Three hMLH1 mutations were all detected in MSI-H (>=2 loci, n=8), but no mutation was found in MSI-L (only one locus, n=9) or MSS (tumor lacking MSI or stable, n=51). Methylation frequency of hMLH1 in MSI-H (87.5 %, 7/8) was significantly higher than that in MSI-L (11.1 %, 1/9) or MSS (5.9 %, 3/51) (P<0.01-0.001), but no difference was found between MSI-L and MSS (P>0.05). CONCLUSION: Both mutation and methylation of hMLH1 are involved in the MSI pathway but not related to the LOH pathway in gastric carci

mutL homolog 1, colon cancer, nonpolyposis type 2 (E. coli) - 3p22.3

Counting, storing and sorting

Counters for aliases, different genes, different keyword-types, sentences with both an alias and a keyword,…

Sorting, manual review of highlighted abstracts

Figure 1. Scheme that illustrates the initial filling up of database using text-mining. Aliases of genes and different keyword lists (methylation, cancer and detection-related) are highlighted in the abstract. At the same time, different parameters are counted and stored in a MySQL relational database. Afterwards, the data is ranked and manually reviewed.

accepted: gene symbol or name, RefSeq, Ensembl ID, . . .). The query is analysed using local symbol/alias lists, generated using GeneCards, and suggestions are presented to the user. In the second step, the user reviews the selections made (most likely the genes selected due to intelligent sorting in the background are correct) and submits his choices.

At this point, the results will be generated and the main result page is presented to the user. This main result page ranks the genes, based on the number of references to the gene in the database. A graphical summary representation of the number of references, the number of primary samples and the mean methylation frequency within different cancer types is also given (example in Figure 2).

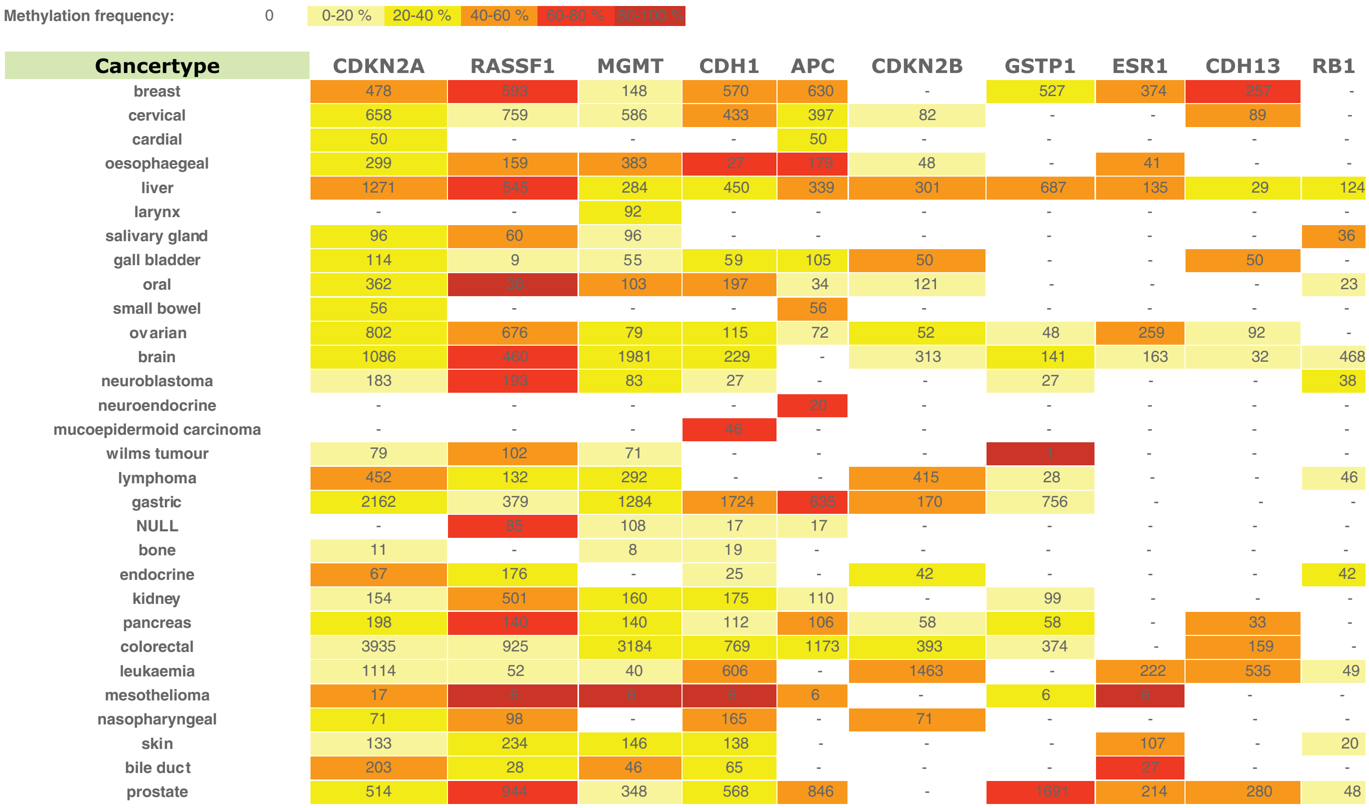
The summary is very useful if multiple genes are searched at once; this feature is what distinguishes this database from previous efforts. One practical usage example would be that, using a pharmacologic demethylation approach in cell lines, 50 candidate genes are selected. The question then is to subselect genes to verify in primary cancer samples, often based on time-consuming literature searches. This selection is facilitated by the summarization view of PubMeth.

From this main page, one can go to the detailed pages, focusing on a selected gene in a certain cancer type. On such a detailed page, graphical representations of the number of references in the database, the total number of samples and the mean methylation frequency are displayed for the different cancer types and their subtypes. The complete individual records, linked with their original PubMed record, are shown. Users can also choose to browse a precomputed genelist. Advantage is that the user can browse all genes in PubMeth without having to query the database, which is significantly faster. However, the summary view is not available.

Cancer-centric query

A cancer-centric query is executed in one easy step: the user selects cancer types (and/or subtypes up to three levels—e.g. lymphoma, non-Hodgkin lymphoma, b-cell lymphoma and diffuse large B-cell). An overview (in the same style as the gene-centric searching approach) of the genes that are most commonly described as methylated in the selected cancer types, as well as the total number of samples and the mean methylation frequency is returned. From this summary page, navigating to detailed pages is intuitive.

This type of search is meant to get a quick overview of the genes that are reported in the methylation context in the cancer (sub) types of interest and in which frequency, to explore methylation in the cancer types of interest, to compare experimental results with or to perform, in a next



Methylation frequency: 0 | 0-20 % | 20-40 % | 40-60 % | 60-80 % | 80-100 %

Cancertype	CDKN2A	RASSF1	MGMT	CDH1	APC	CDKN2B	GSTP1	ESR1	CDH13	RB1
breast	478	593	148	570	630	-	527	374	257	-
cervical	658	759	586	433	397	82	-	-	89	-
cardial	50	-	-	-	50	-	-	-	-	-
oesophaegeal	299	159	383	27	179	48	-	41	-	-
liver	1271	545	284	450	339	301	687	135	29	124
larynx	-	-	92	-	-	-	-	-	-	-
salivary gland	96	60	96	-	-	-	-	-	-	36
gall bladder	114	9	55	59	105	50	-	-	50	-
oral	362	36	103	197	34	121	-	-	-	23
small bowel	56	-	-	-	56	-	-	-	-	-
ovarian	802	676	79	115	72	52	48	259	92	-
brain	1086	460	1981	229	-	313	141	163	32	468
neuroblastoma	183	193	83	27	-	-	27	-	-	38
neuroendocrine	-	-	-	-	20	-	-	-	-	-
mucoepidermoid carcinoma	-	-	-	46	-	-	-	-	-	-
wilms tumour	79	102	71	-	-	-	1	-	-	-
lymphoma	452	132	292	-	-	415	28	-	-	46
gastric	2162	379	1284	1724	635	170	756	-	-	-
NULL	-	85	108	17	17	-	-	-	-	-
bone	11	-	8	19	-	-	-	-	-	-
endocrine	67	176	-	25	-	42	-	-	-	42
kidney	154	501	160	175	110	-	99	-	-	-
pancreas	198	140	140	112	106	58	58	-	33	-
colorectal	3935	925	3184	769	1173	393	374	-	159	-
leukaemia	1114	52	40	606	-	1463	-	222	535	49
mesothelioma	17	6	6	6	6	-	6	6	-	-
nasopharyngeal	71	98	-	165	-	71	-	-	-	-
skin	133	234	146	138	-	-	-	107	-	20
bile duct	203	28	46	65	-	-	-	27	-	-
prostate	514	944	348	568	846	-	1691	214	280	48

Figure 2. Summary page of a gene-centric query. The different colors represent the frequency of methylation of the gene in the different cancer types (what percentage of the samples showed methylation), while the numbers indicate the total number of primary samples tested for methylation.

step, a gene-centric search on these genes for full details in all cancer (sub) types.

A screencast that dynamically shows how to query PubMeth is available on the PubMeth website.

PERFORMANCE OF PUBMETH, DISCUSSION AND FUTURE

To evaluate the performance of PubMeth, we tested how well the database performed in comparison with a careful manual literature search. Therefore, we selected a very recent review, focusing on DNA methylation in breast cancer (9). This article contains a table where the authors provide a list of 39 genes, known to be hypermethylated in breast cancer and their literature references. The genes in this list are entered into the gene-centric search of PubMeth: 27 genes are listed in PubMeth, 11 are not listed and 1 gene could not be associated with a gene symbol. Of the 27 genes listed in PubMeth, 20 are described in breast cancer. Breast cancer is listed first on the results page due to the background sorting mechanisms, but 18 genes are not associated with breast cancer in PubMeth.

On the other hand, the review article lists 39 different genes, while a cancer-centric search for breast cancer returns 94 genes. Important to mention: the genes both in PubMeth and the review (the shared group) are associated with a high number of primary samples in PubMeth. If all 94 genes were ranked according to the number of primary samples in decreasing order, most members of the shared group are on top of this ranking, almost the complete top 10 is present in the shared group (except numbers 7 and 8).

This example clearly shows the power of PubMeth as well as its weaknesses. First of all, doing such a literature search manually usually takes multiple hours, while PubMeth presents its summary within minutes. PubMeth is in most cases able to find more references than a manual search would, using the different keywords and alias lists. On the contrary, often abstracts do not mention any of the genes in question, and these abstracts are not taken into account for consideration in PubMeth. Examples of such articles are large-scale studies with multiple genes or reviews. The articles generally are easily found by manual searches but not using our text-mining approach that is only able to screen abstracts.

As long as there is no universal or centralized system to be able to screen full text articles or a mainstream open access strategy, solution for this would be to leave the restriction that the abstract has to contain a gene out and to do more general searches. Other possibility is to allow users to enter their suggestions for inclusion into PubMeth; such a submission system would allow to combine the power of both submissions by users and an automated text-mining approach that demonstrates to be very powerful dealing with different keyword lists and gene name variants. The latter is available on the PubMeth website: articles related with DNA methylation in cancer that could not be found with PubMeth, can be suggested for inclusion.

Currently, PubMeth only focuses on hypermethylation; however, the inclusion of hypomethylated genes would be useful as well for some users. In a next update, keywords related with hypomethylation will be added.

Other future database updates should take into account different originating tissues (for example, clearly separate between primary cancer tissue and serum) and the different types of normals (surrounding tissue in tumor patient versus samples from healthy person). However, different articles use different terminologies and often this information is not easily extractable.

Improvements to the interface should represent the above described separation of samples. Currently, only the mean of the methylation frequency in primary cancers is given. This could be extended to give an idea of the degree of variation in the different experiments and the different methods, the difference between cancer and normal tissue and the frequency in cell lines. However, it is a real challenge to present all this useful information in a clear interface that is easy to overview and browse.

ACKNOWLEDGEMENTS

The authors would like to thank all initial test users of PubMeth for their detailed comments and suggestions for improvement. Many thanks to all the people who helped correcting and improving the manuscript. Funding to pay the Open Access publication charges for this article was provided by Ghent University.

Conflict of interest statement. None declared.

REFERENCES

1. Doerfler,W., Toth,M., Kochanek,S., Achten,S., Freisemrabien,U., Behnkrappa,A. and Orend,G. (1990) Eukaryotic DNA methylation—facts and problems. *FEBS Lett*., **268**, 329–333.
2. Herman,J.G. and Baylin,S.B. (2003) Gene silencing in cancer in association with promoter hypermethylation. *N. Engl. J. Med*., **349**, 2042–2054.
3. Robertson,K.D. (2002) DNA methylation and chromatin—unraveling the tangled web. *Oncogene*, **21**, 5361–5379.
4. Esteller,M. (2003) Cancer epigenetics: DNA methylation and chromatin alterations in human cancer. *New Trends in Cancer for the 21St Century*, **532**, 39–49.
5. Amoreira,C., Hindermann,W. and Grunau,C. (2003) An improved version of the DNA methylation database (MethDB). *Nucleic Acids Res*., **31**, 75–77.
6. Pattyn,F., Hoebeeck,J., Robbrecht,P., Michels,E., De,P.A., Bottu,G., Coornaert,D., Herzog,R., Speleman,F. *et al*. (2006) methBLAST and methPrimerDB: web-tools for PCR based methylation analysis. *BMC Bioinformatics*, **7**, 496.
7. Garber,K. (2006) Momentum building for human epigenome project. *J. Natl. Cancer Inst*., **98**, 84–86.
8. Rebhan,M., Chalifa-Caspi,V., Prilusky,J. and Lancet,D. (1998) GeneCards: a novel functional genomics compendium with automated data mining and query reformulation support. *Bioinformatics*, **14**, 656–664.
9. Agrawal,A., Murphy,R.F. and Agrawal,D.K. (2007) DNA methylation in breast and colorectal cancers. *Mod. Pathol*., **20**, 711–721.