Supporting Material for

DDMGD: the database of text-mined associations between genes methylated in diseases from different species

Arwa Bin Raies[1], Hicham Mansour[2], Roberto Incitti[1] and Vladimir B. Bajic[1,*]

[1]Computational Bioscience Research Center (CBRC), Computer, Electrical and Mathematical Sciences and Engineering Division, King Abdullah University of Science and Technology, Thuwal, Saudi Arabia
[2] Bioscience Core Laboratories, King Abdullah University of Science and Technology, Thuwal, Saudi Arabia
*To whom correspondence should be addressed: Tel: +966-5447-00088; Fax: +966 (0) 12-808-2386; Email: vladimir.bajic@kaust.edu.sa

Table of Contents

All references are from the main manuscript

Data Sources

We used scientific literature (PubMed and Open Access Subset of PubMed Central) as the main source of information about genes methylated in diseases in any species. To identify articles of relevance to our study, we searched on August 19th, 2014 the PubMed Central database using keywords such as disease, diseases, cancer, cancers, tumor, tumors, tumour, tumours, methylation, hypermethylation, hypomethylation, unmethylation, demethylation, methylated, hypermethylated, hypomethylated, unmethylated, demethylated, gene, genes and DNA, and we extracted those that are included in the Open Access Subset. We found 23,572 full-text articles related to genes methylated in diseases. In addition to these, we found 27,395 abstracts from PubMed, which are related to genes methylated in diseases but were different from the abstracts of the full-text articles in the analysed Open Access Subset.

DEMGD Text Mining System

DEMGD Overview

DEMGD is a text mining system that can extract in its original version (9) associations between methylated human genes and diseases from any free text without restriction to specific diseases. DEMGD consists of four components. The first is the text pre-processing component in which DEMGD splits free text into sentences, extracts genes, diseases and methylation words using dictionaries. The second component is the structured data representation organizing the text into structured format using the document-term matrix (DTM) and the position weight matrix (PWM) formulations. The third component is a classification module in which two random forest machine learning models based on DTM and PWM are used to determine if the genes are methylated and associated with the diseases, as mentioned in the same sentence. Finally, the association extraction module, organizes the extracted association in summary tables and full reports (explained in the next section).

An example of a sentence that includes an association is “In humans, BRCA1 was found frequently methylated in breast cancer patients” (sentence 1) in which BRCA1 methylation is associated with breast cancer in human. However, an example of a sentence that does not include an association is “This study aims to identify if BRCA1 methylation is a potential biomarker for breast cancer prognosis in human patients” (sentence 2), where the sentence mentions only a potential association between BRCA1 and breast cancer. DEMGD should classify the link of “BRCA1 methylation” and “breast cancer” from the first sentence as association, and from the second sentence as non-association. Considering that both sentences provide useful information about DNA methylation, we included both sentences in DDMGD, but each sentence is assigned a different confidence score based on its classification by the classification model (see the next section).

DEMGD requires genes diseases and methylation words to appear in the same sentence, in order to extract the associations. It can extract multiple associations from the same sentence. For example, the sentence “BRCA1 was methylated in breast cancer patients while FILIP1L was heavily methylated in human ovarian cancer” includes two associations between BRCA1 and breast cancer, and FILIP1L and ovarian cancer. A detailed description of machine learning models development can be found in (9), and the system is available for online text mining at: www.cbrc.kaust.edu.sa/demgd/.

Integrating DEMGD with DDMGD

For the purpose of this study, we extended DEMGD to be able to extract associations between genes methylated in diseases from various species, without restriction to specific species. First, we extended and additionally curated dictionaries of genes and diseases to include genes and diseases from various species. To compile the genes dictionary, we extended our previously developed human gene dictionary

by other genes from NCBI Gene database (16). However, we excluded those genes that are very ambiguous and significantly degrade the accuracy of genes recognition based on the following rules:

- Gene name/symbol is similar to common English words (e.g., HAND, TIME, AGE, AGO, etc.).
- Gene name/symbol consists of only digits without letters.
- Gene name/symbol length is shorter than three characters.
- Gene name/symbol mentioned frequently in the methylation context but represents entity that is not a gene (e.g., MSP, PCR, etc.).
- Gene name/symbol is the same as disease name/symbol.

To compile the diseases dictionary we searched for lists of diseases in animals and plants from different sources:

- http://www.cfsph.iastate.edu/DiseaseInfo/
- http://www.daff.gov.au/animal-plant-health/pests-diseases-weeds/animal/notifiable
- http://www.daff.qld.gov.au/animal-industries/animal-health-and-diseases/a-z-list
- http://www.defra.gov.uk/ahvla-en/disease-control/animal-diseases-a-z/
- http://www.fli.bund.de/en/startseite/services/national-reference-laboratories/list-of-notifiable-diseases-and-notifiable-animal-diseases.html
- http://www.daff.qld.gov.au/plants/health-pests-diseases/a-z-significant

We appended these lists to our previously developed human disease dictionary. However, we excluded those disease names/symbols that are similar to common English words or gene names/symbols for the same reason as mentioned above regarding gene names/symbols. Moreover, for the species dictionary we used scientific names of all species from NCBI Taxonomy Database (17) in addition to common names and synonyms of the species that are mentioned in MethDB.

Post-processing

We implemented a post-processing step to extract the species information. The post-processing is performed after an association between a methylated gene and a disease is extracted from a sentence by searching for the species names in the sentence. However, we found that genes, diseases, species, and methylation words do not appear very frequently in the same sentence. Therefore, the species information was not extracted for some of the associations.

In addition, we implemented pattern-matching rules to extract genes expression information such as “increased expression”, “decreased expression”, etc. to help us identify how gene methylation affects gene expression. Given that only a subset of sentences include gene expression information, we could

not extract gene expression information from such sentences. Additionally, we provide links to Expression Atlas, OMIM Gene Map, ArrayExpress, OMIM and GEO for more information about gene expression.

Moreover, we extracted associations between gene methylation and disease progression such as “gene methylation contributes to disease progression”, etc. using pattern-matching rules to determine how gene methylation affects disease progression. However, we could not extract disease progress information for the same reasons mentioned above for gene expression. But we could not find any resources that can provide information about disease progression.

Once the associations are extracted, DEMGD organizes the associations in summary tables that include names/symbols of genes, diseases, methylation words, species, evidence sentences that include the association, PubMed Central ID of the article where the sentence is mentioned, PubMed ID of the abstract where the association is mentioned, and a confidence score generated by DEMGD system. The confidence score, which ranges from 0 to 1, is given by the DEMGD classification model, and it indicates the likelihood that the gene found in the sentence is methylated in the disease found in the same sentence. A confidence score of ‘1’ suggests the highest confidence that an association exists between the methylated gene and the disease, while the smaller the confidence score, the less likely it is that the association exists. For example, sentence 1 from the previous section will be given a high score (0.5 to 1.0), while sentence 2 will be given a low score (0.0 to 0.49), because sentence 1 is classified as an association while sentence 2 is classified as non-association.

After the associations were extracted, we manually curated and removed ambiguous extracted genes, diseases and species from the database. Additionally, we manually assessed information from 1000 entries (500 predicted to be associations and 500 predicted to be non-associations) relative to the sentences from which these entries were extracted. These 1000 entries we name SET1 (SET1 is available at: www.cbrc.kaust.edu.sa/ddmgd/download.php). We used SET1 to identify wrongly extracted genes, diseases and species and they were removed from their corresponding dictionaries. We also identified genes, diseases and species that were mentioned in sentences, but were not extracted by the text mining system, and we added them to their corresponding dictionaries.

Furthermore, we developed post-processing rules based on these entries to filter false positive associations (entries that are wrongly predicted to be associations). Then, we re-processed the full text articles and the abstracts as described in the previous sections using the modified dictionaries and post-processing filtering rules.

DDMGD Database

DDMGD database stores three types of information: the associations between methylated genes and diseases, users' accounts, and normalization of names of genes, diseases and species. The first type of information includes the information on genes methylated in diseases from the summary tables that are provided by the DEMGD text mining system as explained previously. Moreover, the database stores users' accounts if users want to have an account (more details are available in the next section).

Finally, the database stores genes IDs, diseases IDs, species IDs to allow normalizing search queries that involve genes, diseases and species, respectively. Considering that genes, diseases and species can be written in multiple ways, normalizing these names using standard identifications is necessary to ensure compatibility with existing systems and allow for convenient access to required information. For example, if the user restricts search to a specific disease (e.g., breast cancer), the user will get results for all diseases in our database that have the same disease ID as the selected disease (e.g., breast tumor, breast carcinoma and breast tumors). We used NCBI Taxonomy Database, NCBI Gene Database, and Comparative Toxicogenomics Database (18) for species ID, genes ID, and diseases ID, respectively.

In normalizing the genes names, we also considered the species, because the same gene can have different gene IDs depending on the species with which it is associated. For example, if an association includes BRCA1 in human, the gene ID is 672, but in mice the gene ID is 12189. However, sine the species was not identified for many associations (as explained in the previous section), in such cases we used the gene ID for humans, led by the fact that humans are the most commonly studied species.

Evaluation of Database's Content

As benchmark lists, we used three published lists about DNA methylation. Each benchmark list has different characteristics, which allows for evaluating DDMGD from different perspectives. The first benchmark list from (23) includes a set of methylated genes in colorectal cancer in human, the second benchmark list from (24) contains a set of genes methylated in various diseases in human, and the last benchmark list from (25) includes a set of genes methylated in different cancers in mice.

The first benchmark list includes a set of 58 genes that are methylated in colorectal cancer (the genes are mentioned in a review article (Table 1 in (23))). First, we looked at how many genes were found in the database. Out of the 58 genes, DDMGD found 56. It did not find PAPSS2 and UNC5A. Then we looked for the genes in the benchmark list found to be methylated in colorectal cancer. DDMGD contains information on 56 such methylated genes in colorectal cancer. We found that it did not contain such information for PAPSS2 and UNC5A genes. Thus DDMGD contains 56 correct associations from this benchmark.

The second benchmark list is a recent survey about genes methylated in various diseases (autoimmune diseases) (Table 1 from (24)). The survey was published in 2013 and lists 14 genes (CD4, CD11a (ITGAL), CD70 (TNFSF7), Perforin (PRF1), CD40L (TNFSF5), LINE-1, IL-6, DR3, P14ARF, p15, p21, p16INK4a, SHP-1 and Insulin) that are associated with seven autoimmune diseases (systemic lupus erythematosus, rheumatoid arthritis, systemic sclerosis, primary Sjögren's syndrome, primary biliary cirrhosis, psoriasis, and type 1 diabetes). There are 20 associations between the methylated genes and these diseases (24). DDMGD contains all the 14 genes, all seven diseases (however, type 1 diabetes is mentioned as autoimmune diabetes) and 15 out of the 20 associations. It does not contain the four associations between p14ARF, p15, p21 and SHP-1 genes and psoriasis, and one more association between Insulin gene and type 1 diabetes.

Finally, the third benchmark list is a recent list that was published in 2010, which includes genes methylated in various cancers in mice (Table 1 from (25)). The list contains 16 associations that include 12 genes (Cdkn2a, Slc5a1 , Fhit , Rarb , Dapk , Mgmt , Id4 , Mlt1 , Tslc-1 , Prdx1 , Igfbp3 and Cxcr4), which are found to be associated with 10 types of cancers (lung, prostate, pancreatic, glioma, bladder, tongue, skin, leukaemia, liver and hepatocellular) in mice. DDMGD contains 9 out of 12 genes in mice. It does not contain Slc5a1, Igfbp3 and Cxcr4. Additionally, DDMGD contains 9 types of cancers in mice except tongue cancer. Regarding the actual information on genes methylated in these 10 cancers in mice, DDMGD contains six of the 16 associations in mice: (Cdkn2a, lung cancer), (RARB, lung cancer), (DAPK, lung cancer), (MGMT, skin cancer), (ID4, leukaemia), (MLT1, liver cancer).

Evaluation Metrics

We used the following metrics to evaluate the performance of our system, where TP, FP, TN, and FN stand for True Positive, False Positive, True Negative and False Negative, respectively:

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$\text{Specificity} = \frac{TN}{TN + FP}$$

$$\text{F-score} = 2\frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN}$$

In the context of evaluating performance of the named entity recognition, TP, FN, and FP correspond to the number of named entities that were correctly recognized, named entities that were not recognized, and non-named entities that were falsely recognized as named entities, respectively. However, in the context of association extraction, TP, FN, TN, and FP correspond to, respectively: associations that were predicted as associations, associations that were predicted as non-associations, non-associations predicted as non-associations, and non-associations predicted as associations.

Comparison of databases' features

We did thorough comparison of the features of DDMGD and other databases that include information about genes methylated in diseases to highlight the current properties DDMGD with respect to these databases. We compared 9 databases with DDMGD using 23 features. Some of the features are general, so that they can be applicable to all the databases. We used only features that can be objectively assessed for all databases. The comparison results are split into two tables. Table S1(A) shows the comparison between DDMGD and DiseaseMeth (13), MeInfoText 2.0 (8), MethDB (15,16) and MethyCancer (11) databases, while Table S1(B) shows the comparison of DDMGD with MethylomeDB (12), NGSmethDB (17,18), PubMeth (14), MENT (19) and CMS (20) databases.

The features 1-5 compare the factual content of the databases. DDMGD is the most comprehensive (to date) database on this topic. DiseaseMeth, MeInfoText2.0, MethyCancer, PubMeth, MENT, and CMS contain information limited to humans, MethylomeDB focuses on humans and mice, and NGSmethDB includes information on genes methylated in diseases only in five species. Most of these databases include information on genes methylated in diseases that involve different methylation types such as methylation, hyper-methylation or hypo-methylation. However, PubMeth focuses on methylation and hyper-methylation, but not hypo-methylation, while MethylomeDB does not provide the methylation type. Additionally, MeInfoText2.0, MethDB, MethyCancer, PubMeth, MENT and CMS focus on human cancers, but not other diseases, while MethylomeDB focuses on the brain tissues. MethDB is manually-curated database that aims to include information on genes methylated in diseases related to many species. Only three databases, DDMGD, MeInfoText 2.0 and MethyCancer, include methylation information related to miRNA.

Features 6-9 focus on the data sources used to extract information and populate databases. DDMGD, DiseaseMeth, MeInfoText2.0, NGSmethDB, MethDB and PubMeth provide reference to scientific literature from which the information was extracted. DDMGD and MeInfoText 2.0 are automatically compiled via text mining, PubMeth used text mining to rank the abstracts followed by manual curation, and the rest of the databases used manual curation. DDMGD used abstracts and full-text articles to extract the required information, while MeInfoText 2.0 and PubMeth used only abstracts.

Features 10-12 are about flexibility in using the database. Most of databases allow users to search using one gene, disease and/or species. However, batch queries allow users to search for several genes, diseases and/or species. MENT does not implement batch queries. Additionally, DDMGD, DiseaseMeth, MethDB, MethylomeDB, and MENT allow users to sort the search results.

Features 13-18 focus on the search results. Most of the databases show gene, disease/tissue, species and sometimes methylation type. However, some databases display additional information. DDMGD, DiseaseMeth, MeInfoText 2.0, NGSmethDB, PubMeth, MENT, and CMS present additional statistics, such as the number of diseases/tissues that are associated with the selected genes, methylation

frequency, etc. Additionally, most of databases include integrated analysis tools and links to other databases (e.g., NCBI Entrez Gene database) to allow researchers to get additional information. However, MethylomeDB and MENT do not provide links to other databases. Also, all databases show supporting information such as PubMed IDs of the abstracts. However, only DDMGD, MeInfoText 2.0 and PubMeth display color-highlighted evidence sentences.

Features 19-24 are about options for users' interaction with the database. DDMGD and MethDB allow users to create optional and free user accounts that provide additional functionality. Also, DDMGD, DiseaseMeth, MethDB, MethyCancer and PubMeth allow researchers to submit new information to the databases. However, DDMGD permits users to edit the information submitted, and allows them to save the search results in their account for later access. Additionally, DDMGD, DiseaseMeth, MethylomeDB, MENT and CMS allow users to download the search results. However, DDMGD, DiseaseMeth, MethDB, MethyCancer, MethylomeDB and NGSmethDB allow users to download the whole database.

Tables S1(A) and S1(B) show the features that are implemented by at least one database. DDMGD is the only one that implements all of these features. In addition, the features: processing full-text articles, edit submitted data and save search results are only implemented in DDMGD.

Table S1(A). Comparison with DiseaseMeth, MeInfoText 2.0, MethDB and MethyCancer Databases

#	Criteria	DDMGD	Disease Meth	MeInfoText 2.0	MethDB	MethyCancer
1	Any gene	Yes	No	No	Yes	No
2	Any disease	Yes	Yes	No	Yes	No
3	Any species	Yes	No	No	Yes	No
4	Methylation type	Yes	Yes	Yes	Yes	Yes
5	MicroRNA	Yes	No	Yes	No	Yes
6	Text data source	Yes	Yes	Yes	Yes	No
7	Automated compilation	Yes	No	Yes	No	No
8	Abstracts	Yes	No	Yes	No	No
9	Full-text	Yes	No	No	No	No
10	batch query	Yes	Yes	Yes	Yes	Yes
11	Sorting the results	Yes	Yes	No	Yes	No
12	Ease of use	Yes	Yes	Yes	Yes	Yes
13	Statistics	Yes	Yes	Yes	No	No
14	Evidence sentence	Yes	No	Yes	No	No
15	Highlighted evidence	Yes	No	Yes	No	No

	sentence					
16	Links to other databases	Yes	Yes	Yes	Yes	Yes
17	Integrated tools	Yes	Yes	Yes	Yes	Yes
18	Supporting information	Yes	Yes	Yes	Yes	Yes
19	User accounts	Yes	No	No	Yes	No
20	Submit new data	Yes	Yes	No	Yes	Yes
21	Edit submitted data	Yes	No	No	No	No
22	Save search results	Yes	No	No	No	No
23	Download search results	Yes	Yes	No	No	No
24	Download the database	Yes	Yes	No	Yes	Yes

Table S1(B). Comparison with MethylomeDB, NGSmethDB, PubMeth, MENT and CMS Databases

#	Criteria	DDMGD	MethylomeDB	NGSmethDB	PubMeth	MENT	CMS
1	Any gene	Yes	No	No	No	No	No
2	Any disease	Yes	No	No	No	No	No
3	Any species	Yes	No	No	No	No	No
4	Methylation type	Yes	No	Yes	No	Yes	No
5	MicroRNA	Yes	No	No	No	No	No
6	Text data source	Yes	No	Yes	Yes	No	No
7	Automated compilation	Yes	No	No	Yes	No	No
8	Abstracts	Yes	No	No	Yes	No	No
9	Full-text	Yes	No	No	No	No	No
10	batch query	Yes	Yes	Yes	Yes	No	Yes
11	Sorting the results	Yes	Yes	No	No	Yes	No
12	Ease of use	Yes	No	Yes	Yes	Yes	Yes
13	Statistics	Yes	No	Yes	Yes	Yes	Yes
14	Evidence sentence	Yes	No	No	Yes	No	No
15	Highlighted evidence sentence	Yes	No	No	Yes	No	No
16	Links to other	Yes	No	Yes	Yes	No	Yes

	databases						
17	Integrated tools	Yes	Yes	Yes	Yes	Yes	Yes
18	Supporting information	Yes	Yes	Yes	Yes	Yes	Yes
19	User accounts	Yes	No	No	No	No	No
20	Submit new data	Yes	No	No	Yes	No	No
21	Edit submitted data	Yes	No	No	No	No	No
22	Save search results	Yes	No	No	No	No	No
23	Download search results	Yes	Yes	No	No	Yes	Yes
24	Download the database	Yes	Yes	Yes	No	No	No

Comparison of databases' content

We compared the content of DDMGD with different databases to demonstrate that DDMGD extends existing and introduces additional information not provided by other resources. It also represents one way of evaluating the utility and quality of DDMGD. First, we compared DDMGD with another automatically-compiled database, MeInfoText 2.0. Second, considering that MeInfoText 2.0 is restricted only to DNA methylation information in only cancers, we also compared DDMGD with another database, DiseaseMeth, that includes information about other diseases. DiseaseMeth is the most comprehensive manually-curated database for DNA methylation information in human diseases. Finally, we compared the quality of automatically-compiled and manually-curated databases in this field, so we compared DDMGD with MethDB, because MethDB has the most similar scope to DDMGD.

As benchmark lists, we used three published lists about DNA methylation. Each benchmark list has different characteristics, which allows for evaluating DDMGD from different perspectives. The first benchmark list from (23) includes a set of methylated genes in colorectal cancer, the second benchmark list from (24) contains a set of genes methylated in various diseases, and the last benchmark list from (25) includes a set of genes methylated in different cancers and from mice.

Comparison with MeInfoText 2.0: The authors of MeInfoText 2.0 evaluated MeInfoText 2.0 using a set of 58 genes that are methylated in colorectal cancer (the genes are mentioned in a review article (Table 1 in (23))). We used the same list of genes as a benchmark for comparison of DDMGD and MeInfoText. Table S2 summarizes the comparison results.

First, we looked at how many genes were found in the two databases. Out of the 58 genes, DDMGD found 56. It did not find PAPSS2 and UNC5A. On the other hand, MeInfoText 2.0 found all 58 genes, with the genes Cyclin A1, RASSF2A and RIL found with the alternative names CCNA1, RASSF2 and PDLIM4, respectively, instead of their symbols in the benchmark list.

Then we looked for the genes in the benchmark list that are found to be methylated in colorectal cancer. DDMGD contains information on 56 such methylated genes in colorectal cancer. It did not contain such information for PAPSS2 and UNC5A genes. However, in MeInfoText 2.0, only 42 genes from the benchmark list are found to be methylated in colorectal cancer. Genes that are not found in MeInfoText 2.0 database methylated in this cancer are B4GALT1, BNC1, CD133, CDH4, Cyclin A1, DKK-3, EphA1, MAL, MSX1, NDRG2, NTRK2, PAPSS2, PTGIS, RIL, TUBG2 and UNC5A. The RASSF2A gene is found with its alternative symbol RASSF2 instead of its symbol in the benchmark list.

Comparison with DiseaseMeth: The authors of DiseaseMeth did not use any test data to evaluate DiseaseMeth. Therefore, we used a recent survey about genes methylated in various diseases (autoimmune diseases) (Table 1 from (24)). The survey was published in 2013 and lists 14 genes (CD4, CD11a(ITGAL), CD70(TNFSF7), Perforin(PRF1), CD40L(TNFSF5), LINE-1, IL-6, DR3, P14ARF, p15,

p21, p16INK4a, SHP-1 and Insulin) that are associated with seven autoimmune diseases (systemic lupus erythematosus (SLE), rheumatoid arthritis (RA), systemic sclerosis (SSc), primary Sjögren's syndrome (pSS), primary biliary cirrhosis (PBC), psoriasis, and type 1 diabetes(T1D)). There are 20 associations between the methylated genes and these diseases (24). Table S3 summarizes the criteria we used to compare DiseaseMeth and DDMGD.

DDMGD contains all the 14 genes, all seven diseases (however, type 1 diabetes is mentioned as autoimmune diabetes) and 15 out of the 20 associations. It does not contain the associations between p14ARF, p15, p21 and SHP-1 with psoriasis, and Insulin and type 1 diabetes. However, DiseaseMeth contains only four genes (CD4, ITGAL, CD70, and PRF1), and two diseases (systemic lupus erythematosus and rheumatoid arthritis), but none of the 20 associations.

Comparison with MethDB: The authors of MethDB did not use any data to evaluate MethDB. Therefore, we used a recent list that was published in 2010 (Table 1 from (25)). The list contains 16 associations that include 12 genes (Cdkn2a, Slc5a1, Fhit, Rarb, Dapk, Mgmt, Id4, Mlt1, Tslc-1, Prdx1, Igfbp3 and Cxcr4), which are found to be associated with 10 types of cancer (lung, prostate, pancreatic, glioma, bladder, tongue, skin, leukemia, liver and hepatocellular) in mice. We used this list as a benchmark to compare between DDMGD and MethDB. Table S4 summarizes the comparison results.

DDMGD contains 9 out of 12 genes in mice. It does not contain Slc5a1, Igfbp3 and Cxcr4. However, MethDB does not contain any of the 12 genes in mice. Additionally, DDMGD contains 9 types of cancers in mice except tongue. However, MethDB contains only two cancers: lung and liver in mice.

Regarding the actual information on genes methylated in these ten cancers in mice, DDMGD contains six of the 16 associations in mice: (Cdkn2a, lung), (rarb, lung), (DAPK, lung), (MGMT, skin), (ID4, leukemia), (MLT1, liver). However, MethDB does not contain any of the associations in mice.

Table S5 provides a summary of databases coverage with respect of the three benchmark lists

Table S2. Comparison between DDMGD and MeInfoText 2.0 using the first benchmark list

Criteria	DDMGD	MeInfoText
Number of genes (out of the 58 genes) are in the database	56	58
Number of genes (out of the 58 genes) are associated with colorectal cancer in the database	56	42

Table S3. Comparison between DDMGD and DiseaseMeth using the second benchmark list

Criteria	DDMGD	DiseaseMeth
Number of genes (out of the 14 genes) are in the database	14	4
Number of diseases (out of the 7 diseases) are in the database	7	2

Number of associations (out of 20 associations) are in the database	15	0

Table S4. Comparison between DDMGD and MethDB using the third benchmark list

Criteria	DDMGD	MethDB
Number of genes (out of the 12 genes) are in the database	9	0
Number of cancers (out of the 10 cancers) are in the database	9	2
The number of associations (out of the 16 associations) are in the database	6	0

Table S5. Summary of databases coverage with respect of the three benchmark lists

Criteria	Database	% of extracted genes	% of extracted diseases	% of extracted associations
Benchmark List 1	DDMGD	96.55%	100.00%	96.55%
	MeInfoText	100.00%	100.00%	72.41%
Benchmark List 2	DDMGD	100.00%	100.00%	75.00%
	DiseaseMeth	28.57%	28.57%	0.00%
Benchmark List 3	DDMGD	75.00%	90.00%	47.50%
	MethDB	0.00%	20.00%	0.00%
All benchmark lists combined	DDMGD	94.04%	94.44%	81.91%