D856–D862 *Nucleic Acids Research, 2020, Vol. 48, Database issue* *Published online 10 October 2019*
doi: 10.1093/nar/gkz830

DNMIVD: DNA methylation interactive visualization database

Wubin Ding[1,†], Jiwei Chen[1,†], Guoshuang Feng[2,3,†], Geng Chen[1], Jun Wu[1], Yongli Guo[2,3,*], Xin Ni[2,3,*] and Tieliu Shi[1,2,4,*]

[1]Center for Bioinformatics and Computational Biology, and the Institute of Biomedical Sciences, School of Life Sciences, East China Normal University, Shanghai 200241, China, [2]Big Data and Engineering Research Center, Beijing Key Laboratory for Pediatric Diseases of Otolaryngology, Head and Neck Surgery, the Ministry of Education Key Laboratory of Major Diseases in Children, Beijing Pediatric Research Institute, Beijing Children's Hospital, Capital Medical University, National Center for Children's Health, Beijing 100045, China, [3]Beijing Advanced Innovation Center for Big Data-Based Precision Medicine, Beihang University & Capital Medical University, Beijing 100083, China and [4]Biological Targeting Diagnosis and Therapy Research Center, Guangxi Medical University, Nanning 530021, China

Received August 13, 2019; Revised September 09, 2019; Editorial Decision September 16, 2019; Accepted September 28, 2019

ABSTRACT

Aberrant DNA methylation plays an important role in cancer progression. However, no resource has been available that comprehensively provides DNA methylation-based diagnostic and prognostic models, expression–methylation quantitative trait loci (emQTL), pathway activity-methylation quantitative trait loci (pathway-meQTL), differentially variable and differentially methylated CpGs, and survival analysis, as well as functional epigenetic modules for different cancers. These provide valuable information for researchers to explore DNA methylation profiles from different aspects in cancer. To this end, we constructed a user-friendly database named DNA Methylation Interactive Visualization Database (DNMIVD), which comprehensively provides the following important resources: (i) diagnostic and prognostic models based on DNA methylation for multiple cancer types of The Cancer Genome Atlas (TCGA); (ii) meQTL, emQTL and pathway-meQTL for diverse cancers; (iii) Functional Epigenetic Modules (FEM) constructed from Protein-Protein Interactions (PPI) and Co-Occurrence and Mutual Exclusive (COME) network by integrating DNA methylation and gene expression data of TCGA cancers; (iv) differentially variable and differentially methylated CpGs and differentially methylated genes as well as related enhancer information; (v) correlations between methylation of gene promoter and corresponding gene expression and (vi) patient survival-associated CpGs and genes with different endpoints. DNMIVD is freely available at http://www.unimd.org/dnmivd/. We believe that DNMIVD can facilitate research of diverse cancers.

INTRODUCTION

Multiple studies have identified DNA methylation biomarkers for cancer diagnosis and prognosis (1,2). In addition to methylated CpG sites in the promoter region, methylated CpGs in the gene body can also serve as cancer diagnostic markers (3). Therefore, recent reports have investigated global DNA methylation profiles to develop powerful screening markers for cancer diagnosis and prognosis (4,5). Several new algorithms and concepts have been proposed to facilitate cancer epigenetic research, such as Functional Epigenetic Modules (FEM) (6), expression–methylation quantitative trait loci (emQTL) (7), and differentially variable and differentially methylated CpGs (DVMCs) (8). Although several databases regarding DNA methylation have been launched, these databases are mainly focused on one aspect of DNA methylation, such as the Pancan-meQTL database (9) on methylation quantitative trait loci (meQTL), Lnc2Meth (10) on lncRNA–DNA methylation associations, and MethSurv (11) for survival analysis. However, no database has been available that simultaneously provides CpG-based diagnostic and prognostic models, emQTL, pathway activity-meQTL (pathway-meQTL), DVMCs, survival analysis, as well as FEM for different cancers, This information would aid researchers in exploring DNA methylation profiles from different aspects.

[*]To whom correspondence should be addressed. Tel: +86 21 54345020; Email: tieliushi@yahoo.com
Correspondence may also be addressed to Xin Ni. Email: nixin@bch.com.cn
Correspondence may also be addressed to Yongli Guo. Email: guoyongli@bch.com.cn
[†]The authors wish it to be known that, in their opinion, the first three authors should be regarded as Joint First Authors.

For this purpose, we constructed a user-friendly database named the DNA Methylation Interactive Visualization Database (DNMIVD), which comprehensively provides the following important resources: (i) diagnostic and prognostic models based on DNA methylation for 14 and 23 cancer types of TCGA, respectively; (ii) meQTL, emQTL and pathway-meQTL for diverse cancers; (iii) FEM constructed from Protein–Protein Interactions (PPI) and Co-Occurrence and Mutual Exclusive (COME) network by integrating DNA methylation and gene expression data of TCGA cancers; (iv) DVMCs, differentially methylated genes (DMGs) as well as related enhancer information; (v) correlations between methylation level of gene promoters and corresponding gene expression and (vi) patient survival associated CpGs and genes with different endpoints (overall survival (OS), disease-free interval (DFI) and progression-free interval (PFI)). These abundant DNA methylation resources and useful diagnostic and prognostic models will allow users to obtain valuable information for research.

DNMIVD is freely available at http://www.unimd.org/dnmivd/ and is convenient for building models, browsing, searching and downloading data of methylation information. Importantly, DNMIVD is the first database to allow researchers to build molecular models for cancer diagnosis and prognosis based on DNA methylation as well as to visualize the methylation profile of CpGs and gene promoters from various aspects. We believe that DNMIVD provides valuable DNA methylation resources and will facilitate the research of diverse cancers.

MATERIALS AND METHODS

Data collection

The gene expression and DNA methylation data of TCGA were downloaded from UCSC Xena (http://xena.ucsc.edu) and preprocessed as previously described (4). Clinical data with different endpoints were downloaded from the TCGA Pan-Cancer Clinical Data Resource (TCGA-CDR) (12). The R/Bioconductor package DESeq2 (13) was used to normalize raw RNA-seq read count. MeQTL was downloaded from the Pancan-meQTL database (9). Enhancer and related information were downloaded from the HACER database (14) (Figure 1A). The cancer types included in DNMIVD are listed in Supplementary Table S1.

Definition of DMGs between tumor and normal samples

We first assigned DNAm values for each gene with the average beta value of the probes mapped to the promoter region, including TSS200 (region from –200 bp upstream to the transcription start site (TSS) itself), 1stExon (the first exon), TSS1500 (from –200 to –1500 bp upstream of TSS), and 5′UTR in order (6,15). To define DMGs, we first defined β-difference as the difference between the mean β value of tumor and normal samples. An unpaired t-test was performed and *P*-value was adjusted by Benjamini/Hochberg method. DMGs were defined by |β-difference| > 0.2 and a false discovery rate (FDR) corrected *P*-value (Benjamini/Hochberg) ≤ 0.05 (16).

DVMCs

DVMCs were defined as previously described (8). Briefly, we used Bartlett's test to identify differentially variable CpGs and independent Student's *t*-test for differentially methylated CpGs between tumor and normal samples. CpGs with Bartlett's test adjusted *P*-value (Benjamini/Hochberg method) ≤ 0.05, |beta difference| > 0.2 and Student's *t*-test adjusted *P*-value (Benjamini/Hochberg method) ≤ 0.05 are considered as DVMCs. DVMCs have been proven to be useful to identify epigenetic field defects in breast cancer (8).

emQTL

emQTL was first defined by Fleischer *et al.* (7), who identified distinct breast cancer lineages with emQTL. We calculated Pearson correlation between each CpG with an interquartile range (IQR) > 0.1 and all genes; CpG-gene pairs with Bonferroni corrected *P*-value < 0.05 and | Pearson *r* | > 0.3 were considered as significant emQTL.

Pathway-meQTL

We first calculated pathway activities with DESeq2 normalized read count for each pathway from the Molecular Signatures Database (MSigDB) (17,18), as described by Zhang *et al.* (19) and used a previously published method to assess the pathway activities in each sample (20). Similar to emQTL, Pearson correlation for each CpG with IQR > 0.1 and all pathway activities among all samples were calculated. Pathway-CpG pairs with Bonferroni corrected *P*-value < 0.05 and | Pearson *r* | > 0.3 were considered as pathway-meQTL.

To illustrate the association between pathway-meQTL and cancer, we identified cancer hallmark associated pathway-meQTLs (available on the download page of DNMIVD) based on the context of cancer hallmark associated pathways from Zhang *et al.* (19). We used the previously described method (7) to perform enhancer and transcription factor binding site (TFBS) enrichment analysis. We found that the CpGs involved in cancer hallmark-associated pathway-meQTLs were significantly enriched in enhancer regions in 14 out of 18 cancer types (Supplementary Table S2). We checked whether the above CpGs were enriched in TFBS in 24 cancer types and found that some transcription factors (TFs) among the top 50 frequently enriched TFs in cancer hallmark-associated pathway-meQTLs (Supplementary Figure S1 and Supplementary Table S3) were closely associated with cancer. For example, the oncogene cFos, a member of the Fos family, was enriched in all 24 cancer types and is associated with tumorigenesis (21) and survival (22). Finally, we focused on two CpGs cg04396850 and cg25351606 that were identified as potential DNA methylation biomarkers for the diagnosis of pan-cancer (4). The hallmark associated pathway-meQTL result showed that the two CpGs are highly correlated with cancer hallmark pathway activities (Supplementary Figure S2), such as telomere maintenance, unwinding of DNA, double-strand break repair and cell cycle, which was consistent with the previous result that was identified by enrichment analysis of emQTL genes (4).

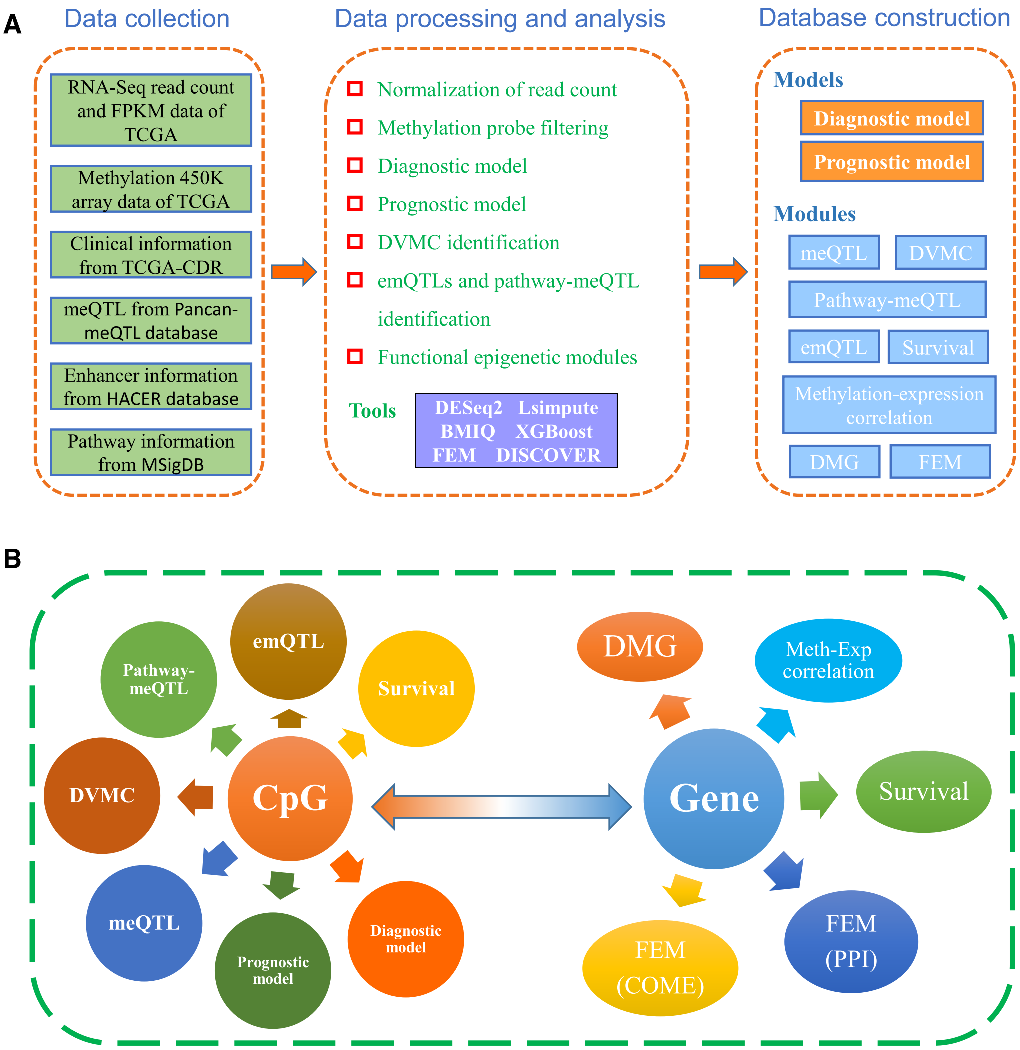
A
Data collection
RNA-Seq read count and FPKM data of TCGA
Methylation 450K array data of TCGA
Clinical information from TCGA-CDR
meQTL from Pancan-meQTL database
Enhancer information from HACER database
Pathway information from MSigDB
Data processing and analysis
Normalization of read count
Methylation probe filtering
Diagnostic model
Prognostic model
DVMC identification
emQTLs and pathway-meQTL identification
Functional epigenetic modules
Tools
DESeq2 Lsimpute
BMIQ XGBoost
FEM DISCOVER
Database construction
Models
Diagnostic model
Prognostic model
Modules
meQTL
DVMC
Pathway-meQTL
emQTL
Survival
Methylation-expression correlation
DMG
FEM
B
emQTL
Pathway-meQTL
Survival
DVMC
CpG
meQTL
Prognostic model
Diagnostic model
DMG
Meth-Exp correlation
Gene
Survival
FEM (COME)
FEM (PPI)

Figure 1. Overview of DNMIVD. (**A**) Data collection, processing and database construction. (**B**) CpG is composed of eleven categorized resources, and gene includes five categorized annotations.

Identification of COME gene pairs

We used DISCOVER (23), a novel statistical independence test, to assess COME by counting how many tumors have an alteration in two genes and comparing the observed alterations to the number of tumors expected to have such an overlap by chance if these alterations were independent. In order to screen COME events that are associated with tumor, fisher exact test was performed in each cancer type to screen the COME events that are significantly enriched in tumor patients or normal samples. To build a reliable network of COME events in each type of cancer, we further screened frequently occurred events from the above COME events, as frequently occurred events, we considered events that were detected in at least three different cancer types.

FEM models

The FEM algorithm (6) is a functional supervised algorithm that uses a network of PPI or COME to identify subnetworks in which a significant number of genes are associated with a phenotype of interest (POI; in our case, differential expression and differential methylation). The PPI network used in this study was downloaded from the supplementary data of a previous publication (24).

Statistical analysis

All statistical analyses were performed with Python3.5.2 on anaconda3–4.0.0. Kruskal–Wallis *H*-test and Chi-square test were performed with Python package Scipy (25). Python package lifelines (http://lifelines.readthedocs.io/en/latest/index.html) and Cox's proportional hazard model was implemented in Cox regression analysis.

Database implementation

The database was organized with MySQL (version 5.7.25) and Django (version 2.0.7). The web interface was developed using HTML with JavaScript. The interactive graph was constructed with Python package plotly (https://plot.ly, version 3.1.0). Venn diagrams in 'Browse by cancer type' module were implemented with jvenn (26).

DATABASE OVERVIEW

DNMIVD collected the RNA-Seq, methylation and clinical data from TCGA as well as meQTL, enhancer and pathway information. For each cancer type, the data were processed and analyzed using a series of filtering steps and tools (27–29) (Figure 1A). DNMIVD contains 396 000 CpG sites involving 20 982 distinct genes; each CpG and gene was connected to seven and five categories of annotations respectively (Figure 1B). For CpG, we provide seven categories of information: (i) the diagnostic model, which helps users screen diagnostic markers to distinguish tumor samples from normal samples; this would be useful for early cancer diagnosis, especially with DNA methylation profiles of blood samples; (ii) the prognostic model, from which, users can obtain related prognostic markers to predict the outcome of cancer patients; (iii) meQTL; (iv) DVMC; (v) pathway-meQTL, which we proposed and was proven to be closely associated with cancer (see Materials and methods); (vi) emQTL and (vii) survival analysis. Regarding genes, DNMIVD offers information on DMGs, methylation-expression correlation, survival analysis and FEM.

USER INTERFACE

The web-based interface of DNMIVD can be freely accessed at http://www.unimd.org/dnmivd/ and allows users to build model, browse, search and download data.

Diagnostic and prognostic model

On 'Model' page, DNMIVD provides two different online models: a diagnostic model and a prognostic model. Input for these two models is a list of CpG sites (HM450k array probes) or gene symbols. For input type of CpG, users can input differentially methylated CpGs (DMCs), DVMCs or CpGs. For gene symbols, a list of differentially expressed genes (DEGs), DMGs or other genes of interest can be imported to build these two kinds of models. If the input is a list of gene symbols, then the gene would be converted to a list of CpG sites located within these genes. Users can also choose different functional regions (TSS200, TSS1500, 1 exon, intron, gene body, TSS10kb, and TTS10kb) in which CpGs are located.

With the diagnostic model, users can predict whether the sample is normal, a benign tumor or malignant, and obtain basic information of the inputted CpGs, including methylation status and important scores. The important score for each CpG is calculated by XGBoost as previously explained (4) and the bar plot of important score will then be displayed. Next, a logistic regression model constructed from CpGs with an important score > 0 is implemented to train the diagnostic model. The Receiver Operating Characteristic (ROC) curve and Area Under Curve (AUC) will be displayed in the result page. Finally, an interactive unsupervised hierarchical clustering and heatmap associated with the methylation profile of screened CpGs is presented (Figure 2, diagnostic model).

With the prognostic model, in addition to basic information of CpGs, the result page also includes results of the univariate and multivariate proportional hazards regression model; the method implemented in the prognostic model was described previously (4). Moreover, a heatmap of DNA methylation profile of retained CpGs, distribution of partial hazard with user-selected endpoint and Kaplan-Meier plot are also generated in the result page (Figure 2, prognostic model). In addition, in the prognostic model, the user can also select different clinical outcome among OS, DFI and PFI.

Browse module

In the 'Browse' page, users can browse DNMIVD by five different panels: 'Cancer Type', 'DMG', 'FEM(PPI)', 'FEM(COME)' and 'pathway-meQTLs'. In the 'Cancer Type' panel, by clicking a cancer type, users can view a Venn diagram of CpGs and genes overlapping in different categories in this cancer (Figure 2, Browse). Users can click on the intersection numbers to show the specific or shared CpGs/genes in the table. In other panels, the corresponding tables with different kinds of hyperlinks are provided.

Search module

Users can search DNMIVD by CpG or by gene symbol, leading to two different result pages.

In the CpG result page, there are six different panels: 'Summary', 'meQTL', 'DVMC', 'pathway-meQTL', 'emQTL' and 'Survival'. The 'Summary' panel offers detailed annotations of CpG and enhancer information overlapped with CpG. In the 'meQTL' panel, users can click buttons in the table to link to the Pancan-meQTL database (9). The 'DVMC' panel shows the significance level and scatter plot of DVMC (Figure 2, DVMC). In the 'pathway-meQTL' and 'emQTL' panel, a table or heatmap is presented for one type of cancer or pan-cancer, respectively. The 'Survival' panel shows Kaplan–Meier for CpG with different endpoint and patient groups (median or cutoff by intervals of 0, 0.3, 0.7 and 1) (Figure 2, Survival).

The gene result page offers six different panels: 'Summary', 'DMG', 'Meth-Exp correlation', 'Survival', 'FEM (PPI)' and 'FEM (COME)'. The 'Summary' panel displays basic information of a gene, followed by a boxplot of DEG and table of CpGs located within the gene. The 'DMG' panel offers a boxplot of DMG (Figure 2, DMG), and the 'Meth-Exp correlation' panel provides the Pearson and Spearman correlation scatter plot between gene expression and methylation level of the gene promoter (Figure 2, Correlation). In the 'FEM (PPI)' and 'FEM (COME)' panel, a table and graph of a network constructed from PPI or COME are presented (Figure 2, FEM).

DISCUSSION AND FUTURE DIRECTIONS

In this study, we developed a database, DNMIVD, to collect and visualize different categories of DNA methylation resources, including models to screen diagnostic and prognostic markers, DVMC, emQTL, pathway-meQTL, FEM and survival analysis. Our database includes 23 cancer types and provides enriched information for users to investigate DNA methylation and gene expression pattern across distinct TCGA cancers. For example, in a

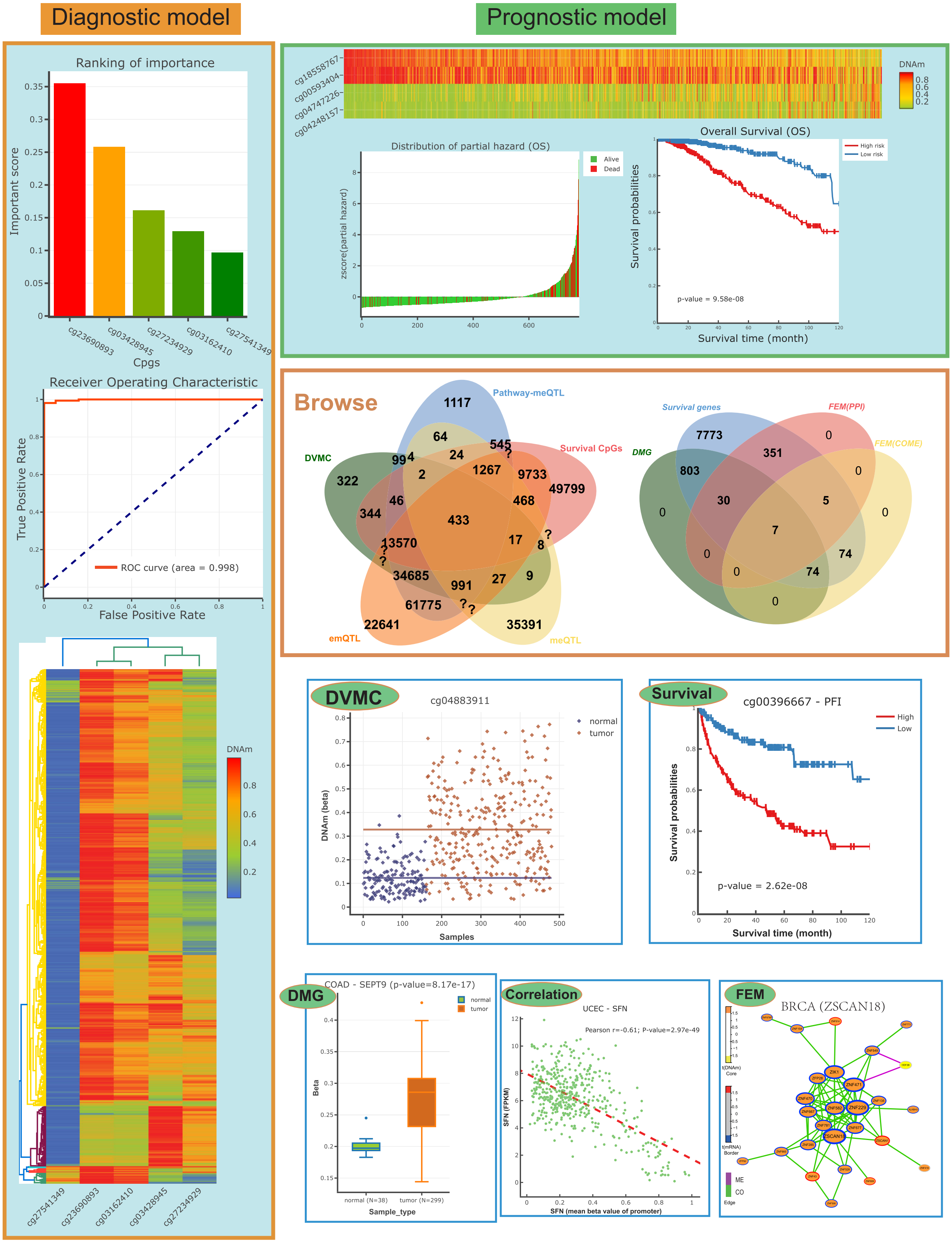

Figure 2. Content and user interface of DNMIVD. (i) Diagnostic model. The results page of the diagnostic model includes a bar plot of feature importance, ROC curve and the heatmap of DNA methylation profile for diagnostic markers in tumor and normal samples. (ii) Prognostic model. The results page of the prognostic model is composed of the heatmap of DNA methylation profile for the screened prognostic markers, bar plot for the distribution of partial hazard and survival Kaplan–Meier curve to display the result of the prognostic model. (iii) Browse by cancer types. If the number in the Venn diagram is clicked, detailed information of selected CpGs of genes will be shown. If the overlapping area is too small to accommodate the number, then a question mark "?" will replace the number. (iv) DVMC, survival, DMG, correlation and FEM panel for the search module.

diagnostic model, four CpGs (cg23690893, cg03428945, cg27234929, cg03162410, cg27541349) were found to be potential signatures for the diagnosis of breast cancer (with AUC > 0.99, diagnostic model in Figure 2). Regarding the prognostic model, four prognostic signatures (cg18558767, cg00593404, cg04747226, cg04248157) were obtained from our previous publication (4), and a multivariate proportional hazards regression model built with the four CpGs could be used to classify patients into high-risk and low-risk groups with different outcome (*P*-value < 0.001, prognostic model in Figure 2). DNMIVD also includes DMG and FEM, such as Septin 9 (SEPT9), the first FDA-approved methylation assay for colorectal cancer, is also observed to be differentially methylated in colon adenocarcinoma (COAD) in DNMIVD (DMG in Figure 2). For survival analysis, DNMIVD provides different endpoints to be chosen, such as OS, DFI and PFI.

We plan to continue to maintain and update the content in DNMIVD by the following strategies: (i) integrating public DNA methylation and gene expression dataset from Gene Expression Omnibus and other public sources besides TCGA. (ii) expanding DNA methylation datasets detected by WGBS, reduced representation bisulfite sequencing, and other approaches; and (iii) adding new algorithms to better integrate DNA methylation and gene expression data. We believe that DNMIVD can provide new insights into regulatory mechanisms and applications of DNA methylation in cancer.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank TCGA, ENCODE and other projects for generating and sharing the data used in this paper.

FUNDING

China Human Proteome Project [2014DFB30010, 2014DFB30030]; National Science Foundation of China [31671377, 31771460, 31801118]; Beihang University & Capital Medical University Advanced Innovation Center for Big Data-Based Precision Medicine Plan [BHME-201801]; Shanghai 111 Project [B14019].

Conflict of interest statement. The authors declare no potential conflicts of interest.

REFERENCES

1. Semaan,A., van Ellen,A., Meller,S., Bergheim,D., Branchi,V., Lingohr,P., Goltz,D., Kalff,J.C., Kristiansen,G., Matthaei,H. *et al.* (2016) SEPT9 and SHOX2 DNA methylation status and its utility in the diagnosis of colonic adenomas and colorectal adenocarcinomas. *Clinical Epigenet.*, **8**, 100.
2. Wei,J.H., Haddad,A., Wu,K.J., Zhao,H.W., Kapur,P., Zhang,Z.L., Zhao,L.Y., Chen,Z.H., Zhou,Y.Y., Zhou,J.C. *et al.* (2015) A CpG-methylation-based assay to predict survival in clear cell renal cell carcinoma. *Nat. Commun.*, **6**, 8699.
3. Wang,Y.W., Ma,X., Zhang,Y.A., Wang,M.J., Yatabe,Y., Lam,S., Girard,L., Chen,J.Y. and Gazdar,A.F. (2016) ITPKA gene body methylation regulates gene expression and serves as an early diagnostic marker in lung and other cancers. *J. Thorac. Oncol.*, **11**, 1469–1481.
4. Ding,W., Chen,G. and Shi,T. (2019) Integrative analysis identifies potential DNA methylation biomarkers for pan-cancer diagnosis and prognosis. *Epigenetics*, **14**, 67–80.
5. Hao,X., Luo,H., Krawczyk,M., Wei,W., Wang,W., Wang,J., Flagg,K., Hou,J., Zhang,H., Yi,S. *et al.* (2017) DNA methylation markers for diagnosis and prognosis of common cancers. *PNAS*, **114**, 7414–7419.
6. Jiao,Y., Widschwendter,M. and Teschendorff,A.E. (2014) A systems-level integrative framework for genome-wide DNA methylation and gene expression data identifies differential gene expression modules under epigenetic control. *Bioinformatics*, **30**, 2360–2366.
7. Fleischer,T., Tekpli,X., Mathelier,A., Wang,S., Nebdal,D., Dhakal,H.P., Sahlberg,K.K., Schlichting,E. and Oslo Breast Cancer Research, C.Oslo Breast Cancer Research, C. and Borresen-Dale,A.L. *et al.* (2017) DNA methylation at enhancers identifies distinct breast cancer lineages. *Nat. Commun.*, **8**, 1379.
8. Teschendorff,A.E., Gao,Y., Jones,A., Ruebner,M., Beckmann,M.W., Wachter,D.L., Fasching,P.A. and Widschwendter,M. (2016) DNA methylation outliers in normal breast tissue identify field defects that are enriched in cancer. *Nat. Commun.*, **7**, 10478.
9. Gong,J., Wan,H., Mei,S., Ruan,H., Zhang,Z., Liu,C., Guo,A.Y., Diao,L., Miao,X. and Han,L. (2019) Pancan-meQTL: a database to systematically evaluate the effects of genetic variants on methylation in human cancer. *Nucleic Acids Res.*, **47**, D1066–D1072.
10. Zhi,H., Li,X., Wang,P., Gao,Y., Gao,B., Zhou,D., Zhang,Y., Guo,M., Yue,M., Shen,W. *et al.* (2018) Lnc2Meth: a manually curated database of regulatory relationships between long non-coding RNAs and DNA methylation associated with human disease. *Nucleic Acids Res.*, **46**, D133–D138.
11. Modhukur,V., Iljasenko,T., Metsalu,T., Lokk,K., Laisk-Podar,T. and Vilo,J. (2018) MethSurv: a web tool to perform multivariable survival analysis using DNA methylation data. *Epigenomics*, **10**, 277–288.
12. Liu,J., Lichtenberg,T., Hoadley,K.A., Poisson,L.M., Lazar,A.J., Cherniack,A.D., Kovatich,A.J., Benz,C.C., Levine,D.A., Lee,A.V. *et al.* (2018) An Integrated TCGA Pan-Cancer clinical data resource to drive high-quality survival outcome analytics, *Cell*, **173**, 400–416.
13. Love,M.I., Huber,W. and Anders,S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.*, **15**, 550.
14. Wang,J., Dai,X., Berry,L.D., Cogan,J.D., Liu,Q. and Shyr,Y. (2019) HACER: an atlas of human active enhancers to interpret regulatory variants. *Nucleic Acids Res.*, **47**, D106–D112.
15. Sharma,P., Bhunia,S., Poojary,S.S., Tekcham,D.S., Barbhuiya,M.A., Gupta,S., Shrivastav,B.R. and Tiwari,P.K. (2016) Global methylation profiling to identify epigenetic signature of gallbladder cancer and gallstone disease. *Tumour Biol.*, **37**, 14687–14699.
16. Naumov,V.A., Generozov,E.V., Zaharjevskaya,N.B., Matushkina,D.S., Larin,A.K., Chernyshov,S.V., Alekseev,M.V., Shelygin,Y.A. and Govorun,V.M. (2013) Genome-scale analysis of DNA methylation in colorectal cancer using Infinium HumanMethylation450 BeadChips. *Epigenetics*, **8**, 921–934.
17. Subramanian,A., Tamayo,P., Mootha,V.K., Mukherjee,S., Ebert,B.L., Gillette,M.A., Paulovich,A., Pomeroy,S.L., Golub,T.R., Lander,E.S. *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *PNAS*, **102**, 15545–15550.
18. Liberzon,A., Subramanian,A., Pinchback,R., Thorvaldsdottir,H., Tamayo,P. and Mesirov,J.P. (2011) Molecular signatures database (MSigDB) 3.0. *Bioinformatics*, **27**, 1739–1740.
19. Zhang,H., Deng,Y., Zhang,Y., Ping,Y., Zhao,H., Pang,L., Zhang,X., Wang,L., Xu,C., Xiao,Y. *et al.* (2017) Cooperative genomic alteration network reveals molecular classification across 12 major cancer types. *Nucleic Acids Res.*, **45**, 567–582.
20. Levine,D.M., Haynor,D.R., Castle,J.C., Stepaniants,S.B., Pellegrini,M., Mao,M. and Johnson,J.M. (2006) Pathway and gene-set activation measurement from mRNA expression data: the tissue distribution of human pathways. *Genome Biol.*, **7**, R93.
21. Milde-Langosch,K. (2005) The Fos family of transcription factors and their role in tumourigenesis. *Eur. J. Cancer*, **41**, 2449–2461.
22. Mahner,S., Baasch,C., Schwarz,J., Hein,S., Wolber,L., Janicke,F. and Milde-Langosch,K. (2008) C-Fos expression is a molecular predictor of progression and survival in epithelial ovarian carcinoma. *Br. J. Cancer*, **99**, 1269–1275.

23. Canisius,S., Martens,J.W. and Wessels,L.F. (2016) A novel independence test for somatic alterations in cancer shows that biology drives mutual exclusivity but chance explains most co-occurrence. *Genome Biol.*, **17**, 261.
24. Menche,J., Sharma,A., Kitsak,M., Ghiassian,S.D., Vidal,M., Loscalzo,J. and Barabasi,A.L. (2015) Disease networks. Uncovering disease-disease relationships through the incomplete interactome. *Science*, **347**, 1257601.
25. Jones,E., Oliphant,T. and Peterson,P. (2014) SciPy: Open source scientific tools for Python. http://www.scipy.org/.
26. Bardou,P., Mariette,J., Escudie,F., Djemiel,C. and Klopp,C. (2014) jvenn: an interactive Venn diagram viewer. *BMC Bioinformatics*, **15**, 293.
27. Ji,X., Tong,W., Ning,B., Mason,C.E., Kreil,D.P., Labaj,P.P., Chen,G. and Shi,T. (2019) QuaPra: Efficient transcript assembly and quantification using quadratic programming with Apriori algorithm. *Sci. China. Life Sci.*, **62**, 937–946.
28. Jin,Y., Chen,G., Xiao,W., Hong,H., Xu,J., Guo,Y., Xiao,W., Shi,T., Shi,L., Tong,W. *et al.* (2019) Sequencing XMET genes to promote genotype-guided risk assessment and precision medicine. *Sci. China. Life Sci.*, **62**, 895–904.
29. Liu,D., Bu,D., Shi,T., Quan,J., Wang,D., Shi,Y., Bo,X.C. and Han,W. (2018) Biological data processing based on bio-processor unit (BPU), a new concept for next generation computational biology. *Sci. China. Life Sci.*, **61**, 597–598.