D890–D895 *Nucleic Acids Research, 2020, Vol. 48, Database issue* *Published online 4 October 2019*
doi: 10.1093/nar/gkz840

EWAS Data Hub: a resource of DNA methylation array data and metadata

Zhuang Xiong[1,2,3,4,†], Mengwei Li[1,2,3,4,†], Fei Yang[1,2,3,4,†], Yingke Ma[1,2,3], Jian Sang[1,2,3,4], Rujiao Li[1,2,3], Zhaohua Li[1,2,3,5], Zhang Zhang[1,2,3,4,5,*] and Yiming Bao[1,2,3,4,5,*]

[1]National Genomics Data Center, Beijing 100101, China, [2]BIG Data Center, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100101, China, [3]CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100101, China, [4]College of Life Sciences, University of Chinese Academy of Sciences, Beijing 100049, China and [5]School of Future Technology, University of Chinese Academy of Sciences, Beijing 100049, China

Received August 15, 2019; Revised September 09, 2019; Editorial Decision September 16, 2019; Accepted October 01, 2019

ABSTRACT

Epigenome-Wide Association Study (EWAS) has become an effective strategy to explore epigenetic basis of complex traits. Over the past decade, a large amount of epigenetic data, especially those sourced from DNA methylation array, has been accumulated as the result of numerous EWAS projects. We present EWAS Data Hub (https://bigd.big.ac.cn/ewas/datahub), a resource for collecting and normalizing DNA methylation array data as well as archiving associated metadata. The current release of EWAS Data Hub integrates a comprehensive collection of DNA methylation array data from 75 344 samples and employs an effective normalization method to remove batch effects among different datasets. Accordingly, taking advantages of both massive high-quality DNA methylation data and standardized metadata, EWAS Data Hub provides reference DNA methylation profiles under different contexts, involving 81 tissues/cell types (that contain 25 brain parts and 25 blood cell types), six ancestry categories, and 67 diseases (including 39 cancers). In summary, EWAS Data Hub bears great promise to aid the retrieval and discovery of methylation-based biomarkers for phenotype characterization, clinical treatment and health care.

INTRODUCTION

Epigenome-Wide Association Study (EWAS) has become an effective strategy to explore epigenetic basis of complex traits, such as aging (1–4), body mass index (BMI) (5,6), smoking (7,8) and diseases (9,10), accordingly leading to massive amounts of epigenetic data. Among different types of epigenetic data, DNA methylation is the most abundant and widely characterized one, primarily owing to the rapid advancement in DNA methylation profiling technologies, especially Infinium HumanMethylation450 (450K) and MethylationEPIC (850K) arrays (11,12). Therefore, comprehensive integration of DNA methylation array data and metadata is of fundamental significance to systematically characterize and investigate methylation states across different experimental conditions and explore epigenetic mechanisms associated with diverse traits.

Over the past several years, several databases have been developed to host DNA methylation array data (13–19), primarily including Gene Expression Omnibus (GEO) (20), ArrayExpress (21), The Cancer Genome Atlas (TCGA) (22), Encyclopedia of DNA Elements (ENCODE) (23), Firehose and The cBio Cancer Genomics Portal (cBioPortal) (24). Although they made valuable efforts to help users conduct methylation studies, these databases have three significant drawbacks. First, they lack an effective and unified normalization method to remove batch effects among different datasets, which may exert severe negative influences on downstream analysis (25,26). Second, different databases use different metadata standards, making it challenging to integrate methylation data across diverse conditions and samples. Third, as a result, none of them provides standardized and normalized DNA methylation profiles across different tissues, sexes, ancestry categories and diseases. In short, these databases are designed mainly for archiving raw data, without value-added curation for data normalization and metadata standardization.

To address these drawbacks, we develop EWAS Data Hub (https://bigd.big.ac.cn/ewas/datahub), for collecting and normalizing DNA methylation array data as well as archiving associated metadata. More than just rehosting

*To whom correspondence should be addressed. Tel: +86 10 84097858; Fax: +86 10 84097720; Email: baoym@big.ac.cn
Correspondence may also be addressed to Zhang Zhang. Tel: +86 10 84097261; Fax: +86 10 84097720; Email: zhangzhang@big.ac.cn
†The authors wish it to be known that, in their opinion, first three authors should be regarded as Joint First Authors.

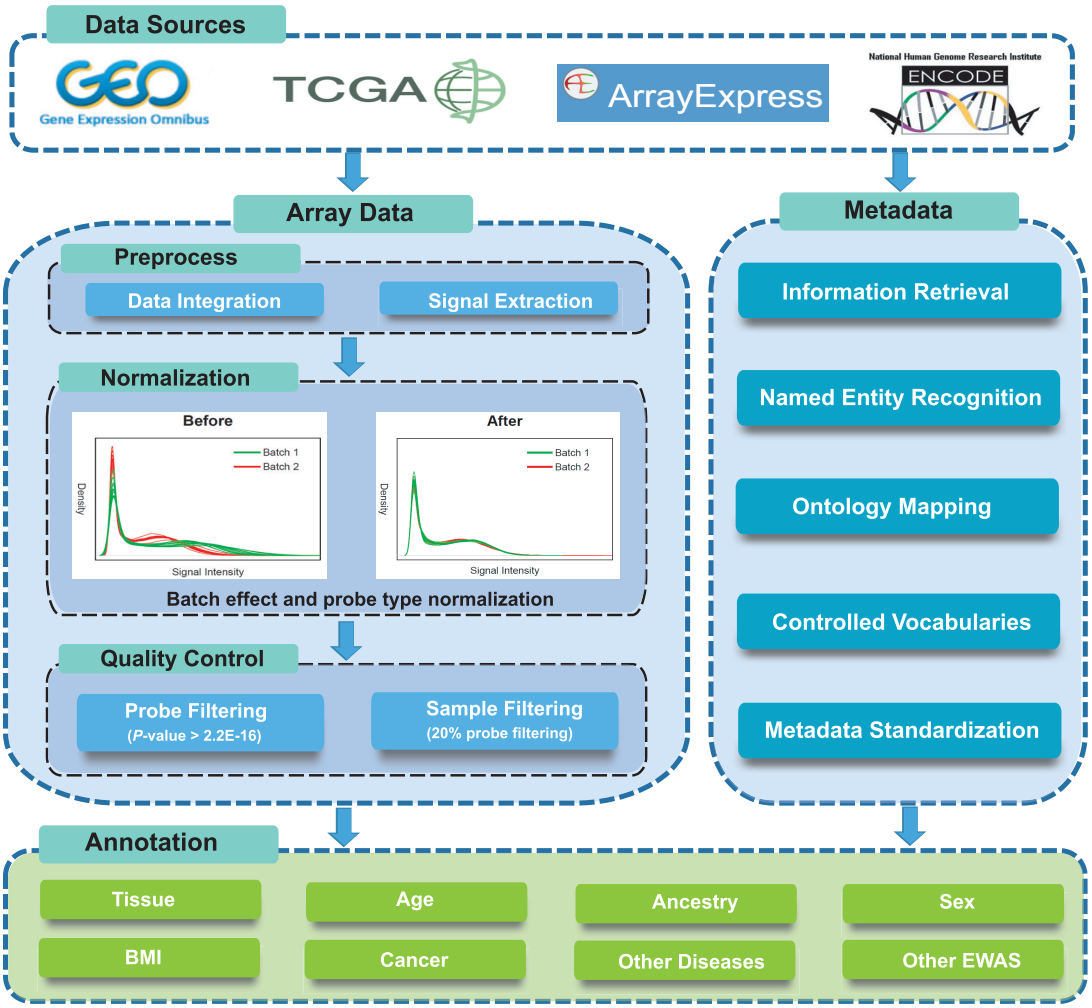

Figure 1. Schematic overview of data processing workflow.

datasets as they appear in public databases, a pipeline optimized for data normalization and metadata curation is employed to remove batch effects and standardize metadata across different datasets. Thus, EWAS Data Hub not only provides normalized methylation data and standardized metadata but also integrates a comprehensive collection of high-quality methylation profiles across different contexts.

IMPLEMENTATION

EWAS Data Hub is implemented using Spring Boot (http://spring.io), a prevailing and easy-to-configure Model-View-Controller (MVC) framework, deployed in a Centos Linux 7.4 environment. Thymeleaf (https://www.thymeleaf.org), which is integrated with the Spring Framework, is used to render the HTML interface. In the back-end part, metadata and reference data are stored in MySQL (https://www.mysql.com). Front-end interfaces are built using Bootstrap (https://getbootstrap.com) with jQuery (https://jquery.com) to provide responsive and user-friendly web pages. The documentation is generated by docsify (https://docsify.js.org). HighCharts (https://www.highcharts.com) and plotly (https://plot.ly/) are used to provide interactive charting and data visualization.

DATA CURATION AND DATABASE CONTENTS

We download all available datasets generated by Infinium HumanMethylation450 or MethylationEPIC arrays from GEO, TCGA, ArrayExpress and ENCODE. If raw data are available, signal intensities are extracted using the Minfi package in Bioconductor (27). A series of curation processes are applied to remove batch effects and improve the quality of data (Figure 1). First, we normalize signal intensities of type I probes between arrays using an in-house reference-based method (for details see the Documentation in the database website). Second, Beta-Mixture Quantile Normalization (BMIQ) is employed to correct the bias associated with technical differences between Type I and Type II array designs (28). Previous studies have shown that setting a stringent detection *P*-value threshold (10^{-16}) is able to significantly reduce the proportion of outlying values (due to a low signal-to-noise ratio of fluorescence intensities) and accordingly achieve improved calling (29,30). Therefore, we perform rigorous quality control to filter out probes with

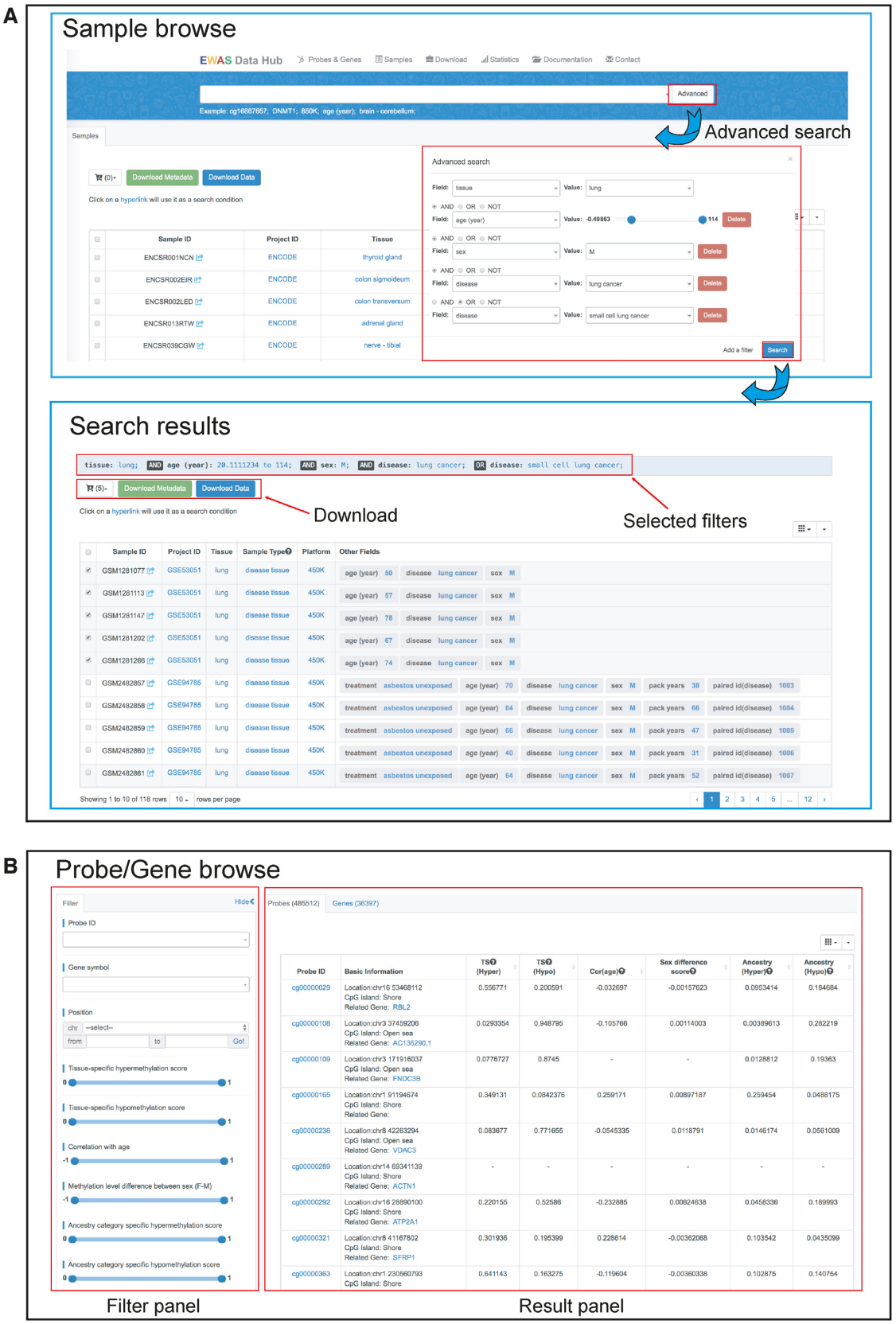

Figure 2. Screenshots of the 'Browse' pages. (**A**) An example of advanced search and its results, (**B**) the 'Browse' page of probe/gene.

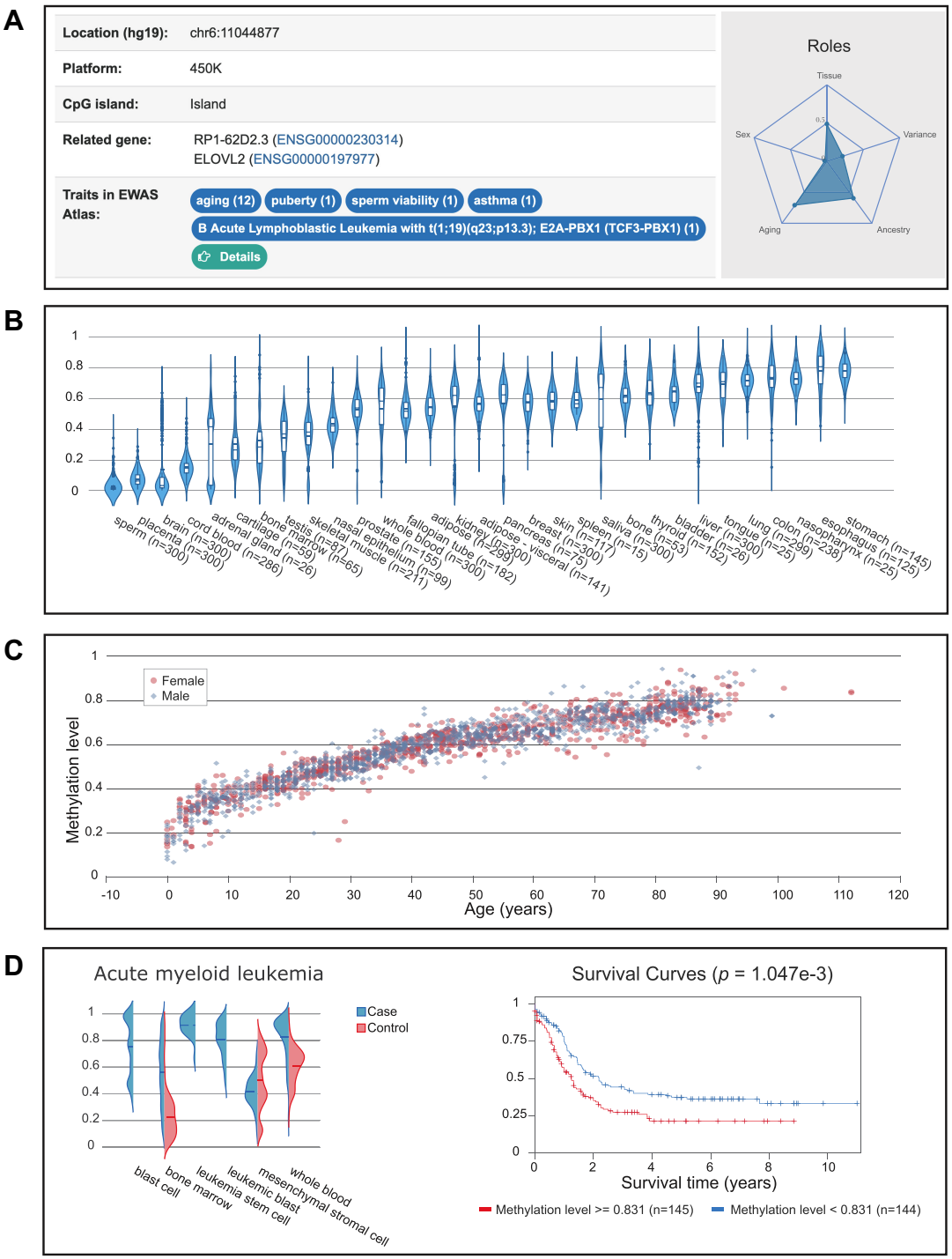

Figure 3. Reference data of probe 'cg16867657'. (**A**) The 'Basic' panel, (**B**) the 'Tissue' panel, (C) the 'Age' panel and (**D**) the 'Cancer' panel.

high detection *P*-values (by default, the threshold is set at 2.2×10^{-16}, which is the smallest number that can be stored by the floating system in R program) and remove samples with more than 20% of the probes with high detection *P*-values (31). To standardize the metadata, we develop a curation model that summarizes 178 fields, including four common fields (sample ID, project ID, tissue/cell and sample type which refers to health status of the sample) and 174 other fields (sex, age, disease, etc.) (see the online Documentation for details). We standardize the values of all fields, if applicable, by setting up controlled vocabularies or mapping them to experimental factor ontology (EFO) that combines parts of several biological ontologies, such as Disease Ontology (DO) and Gene Ontology (GO).

Based on the standardized curation process, EWAS Data Hub obtains a comprehensive collection of DNA methyla-

tion data as well as associated metadata from 75 344 samples (Figure 2), including 470 tissues/cell types, 306 diseases and other conditions. In order to find the sample(s) with specific characteristics, a set of advanced filters, such as tissue, age, sex and platform, are provided to facilitate users to query and narrow down the searched results (Figure 2A). After retrieval, users can download data and metadata of retrieved samples. Importantly, EWAS Data Hub integrates a curated collection of 485 512 probes in association with 36 397 genes (Figure 2B). For each probe/gene, it provides a series of relevant estimates, including tissue-specificity, age correlation, sex difference and ancestry-specificity. In addition, EWAS Data Hub is equipped with multiple filters, allowing users to easily find probes and genes of interest. Specifically, these filters include tissue-specificity score estimated across all collected tissues, correlation coefficient with age, methylation level difference between sex, and ancestry-specificity score. Taken together, EWAS Data Hub features a complete collection of curated probes/genes as well as standardized sample metadata.

For each probe/gene, EWAS Data Hub provides reference DNA methylation profiles cross different contexts, making it possible to systematically characterize and investigate the landscape of methylation states across a wide range of experimental conditions. To facilitate data presentation, taking the probe 'cg16867657' (https://bigd.big.ac.cn/ewas/datahub/probe/cg16867657) as an example, all these data are organized into different panels in terms of basic, tissue, sex, age, ancestry category, BMI, cancer, disease and public EWAS, respectively (Figure 3).

The 'Basic' panel not only provides fundamental information such as genomic location, position relative to CpG island, associated phenotypic traits in EWAS Atlas (32), but also summarizes the correlation of DNA methylation level with tissue, sex, age and ancestry (Figure 3A). Besides, a genome browser is presented to visualize related data in an interactive manner. The 'Tissue' panel contains DNA methylation profiles across 31 tissues, 25 brain parts and 25 blood cell types (Figure 3B). For each probe, its methylation profiles across different tissues/brain parts/blood cell types are depicted in a violin plot, which can greatly help users explore the methylation pattern in various conditions. The 'Sex' panel provides DNA methylation profiles across different sexes, which would be helpful to investigate the heterogeneity of DNA methylation in male and female. The 'Age' and 'BMI' panels provide the distribution of DNA methylation by chronological age and BMI, respectively (Figure 3C). Following a previous study (28), six ancestry categories are adopted in our study. Therefore, the 'Ancestry' panel contains six categories by grouping all datasets into different ancestries. The 'Cancer' and 'Other disease' panels provide DNA methylation profiles across 39 cancers and 28 diseases in both case and control samples. For cancers, Kaplan-Meier survival analyses of overall survival of patients according to the DNA methylation status are conducted (Figure 3D). Moreover, the relationships between DNA methylation and expression of proximal genes are presented in a scatter plot. The Public EWAS panel provides detailed information of public EWAS associations. In partnership with EWAS Atlas (32), EWAS Data Hub automatically retrieves related traits for each probe/gene and provides users with convenient links to EWAS Atlas. Moreover, all datasets collected in this study are publicly available at https://bigd.big.ac.cn/ewas/datahub/download.

DISCUSSION AND FUTURE DEVELOPMENTS

Considering the significance of DNA methylation as one of the most promising cancer diagnostic and therapeutic targets and also a key link between environmental factors and phenotypes (33–38), EWAS Data Hub provides great opportunities to dissect epigenetic mechanisms underlying complex biological traits by integrating and normalizing large amounts of DNA methylation array data as well as curating and standardizing the corresponding metadata. With the ever-growing volume of DNA methylation data and the rapid development of methylation profiling technology, EWAS Data Hub will be updated regularly to integrate more DNA methylation array data, especially those from 850K. Accordingly, the reference DNA methylation profiles will be updated and expanded to include more phenotypic traits, making it possible to conduct a meta-analysis for probes and genes from multiple studies of the same trait. In addition, considering that DNA methylation in combination with gene expression pattern has been frequently used to explore the molecular mechanisms between epigenetics and phenotype (39,40), the relationships between DNA methylation and expression of proximal genes in more phenotypes will be added to EWAS Data Hub. Moreover, online tools to visualize and analyze DNA methylation array data will be developed.

ACKNOWLEDGEMENTS

We thank a number of users for reporting bugs and providing suggestions and two anonymous reviewers for their valuable comments.

FUNDING

National Key Research and Development Program of China [2016YFE0206600, 2017YFC0908403, 2017YFC0907502]; Strategic Priority Research Program of the Chinese Academy of Sciences [XDA19050302, XDB13040500]; 13th Five-year Informatization Plan of Chinese Academy of Sciences [XXH13505-05], International Partnership Program of the Chinese Academy of Sciences [153F11KYSB20160008]; '100-Talent Program' of Chinese Academy of Sciences, and the Open Biodiversity and Health Big Data Initiative of IUBS. Funding for open access charge: Strategic Priority Research Program of the CAS [XDA19050302].

Conflict of interest statement. None declared.

REFERENCES

1. Klutstein,M., Nejman,D., Greenfield,R. and Cedar,H. (2016) DNA Methylation in cancer and aging. *Cancer Res.*, **76**, 3446–3450.
2. Rönn,T., Volkov,P., Gillberg,L., Kokosar,M., Perfilyev,A., Jacobsen,A.L., Jørgensen,S.W., Brøns,C., Jansson,P.-A. and Eriksson,K.-F. (2015) Impact of age, BMI and HbA1c levels on the genome-wide DNA methylation and mRNA expression patterns in human adipose tissue and identification of epigenetic biomarkers in blood. *Hum. Mol. Genet.*, **24**, 3792–3813.

3. Jones,M.J., Goodman,S.J. and Kobor,M.S. (2016) DNA methylation and healthy human aging. *Aging Cell*, **14**, 924–932.
4. Richardson,B. (2003) Impact of aging on DNA methylation. *Ageing Res. Rev.*, **2**, 245–261.
5. Wahl,S., Drong,A., Lehne,B., Loh,M., Scott,W.R., Kunze,S., Tsai,P.-C., Ried,J.S., Zhang,W. and Yang,Y. (2017) Epigenome-wide association study of body mass index, and the adverse outcomes of adiposity. *Nature*, **541**, 81.
6. Dick,K.J., Nelson,C.P., Tsaprouni,L., Sandling,J.K., Aïssi,D., Wahl,S., Meduri,E., Morange,P.-E., Gagnon,F. and Grallert,H. (2014) DNA methylation and body-mass index: a genome-wide analysis. *Lancet North Am. Ed.*, **383**, 1990–1998.
7. Wan,E.S., Weiliang,Q., Andrea,B., Carey,V.J., Helene,B., Rennard,S.I., Alvar,A., Wayne,A., Lomas,D.A. and Demeo,D.L. (2012) Cigarette smoking behaviors and time since quitting are associated with differential DNA methylation across the human genome. *Hum. Mol. Genet.*, **21**, 3073.
8. Tsaprouni,L.G., Tsun-Po,Y., Jordana,B., Dick,K.J., Stavroula,K., James,N., Ana,V.U., Elin,G., Nelson,C.P. and Eshwar,M. (2014) Cigarette smoking reduces DNA methylation levels at multiple genomic loci but the effect is partially reversible upon cessation. *Epigenetics.*, **9**, 1382–1396.
9. Rakyan,V.K., Down,T.A., Balding,D.J. and Beck,S. (2011) Epigenome-wide association studies for common human diseases. *Nat. Rev. Genet.*, **12**, 529.
10. Hung,C.-S., Wang,S.-C., Yen,Y.-T., Lee,T.-H., Wen,W.-C. and Lin,R.-K. (2018) Hypermethylation of CCND2 in lung and breast cancer is a potential biomarker and drug target. *Int. J. Mol. Sci.*, **19**, 3096.
11. Sandoval,J., Heyn,H., Moran,S., Serramusach,J., Pujana,M.A., Bibikova,M. and Esteller,M. (2016) Validation of a DNA methylation microarray for 450,000 CpG sites in the human genome. *Epigenetics*, **6**, 692–702.
12. Moran,S., Arribas,C. and Esteller,M. (2016) Validation of a DNA methylation microarray for 850,000 CpG sites of the human genome enriched in enhancer sequences. *Epigenomics*, **8**, 389–399.
13. Xiong,Y., Wei,Y., Gu,Y., Zhang,S. and Zhang,Y. (2017) DiseaseMeth version 2.0: a major expansion and update of the human disease methylation database. *Nucleic Acids Res.*, **45**, D888–D895.
14. Liu,D., Zhao,L., Wang,Z., Zhou,X., Fan,X., Li,Y., Xu,J., Hu,S., Niu,M. and Song,X. (2018) EWASdb: epigenome-wide association study database. *Nucleic Acids Res.*, **47**, D989–D993.
15. Zou,D., Sun,S., Li,R., Liu,J., Zhang,J. and Zhang,Z. (2014) MethBank: a database integrating next-generation sequencing single-base-resolution DNA methylation programming data. *Nucleic Acids Res.*, **43**, D54–D58.
16. He,X., Chang,S., Zhang,J., Qian,Z. and Xiang,H. (2008) MethyCancer: the database of human DNA methylation and cancer. *Nucleic Acids Res.*, **36**, D836–D841.
17. Huang,W.Y., Hsu,S.-D., Huang,H.-Y., Sun,Y.-M., Chou,C.-H., Weng,S.-L. and Huang,H.-D. (2015) MethHC: a database of DNA methylation and gene expression in human cancer. *Nucleic Acids Res.*, **43**, 856–861.
18. Xin,Y., Chanrion,B., O'Donnell,A.H., Milekic,M. and Haghighi,F.G. (2012) MethylomeDB: a database of DNA methylation profiles of the brain. *Nucleic Acids Res.*, **40**, D1245–D1249.
19. Li,R., Liang,F., Li,M., Zou,D., Sun,S., Zhao,Y., Zhao,W., Bao,Y., Xiao,J. and Zhang,Z. (2017) MethBank 3.0: a database of DNA methylomes across a variety of species. *Nucleic Acids Res.*, **46**, D288–D295.
20. Barrett,T., Wilhite,S.E., Ledoux,P., Evangelista,C., Kim,I.F., Tomashevsky,M., Marshall,K.A., Phillippy,K.H., Sherman,P.M. and Holko,M. (2012) NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Res.*, **41**, D991–D995.
21. Brazma,A., Parkinson,H., Sarkans,U., Shojatalab,M., Vilo,J., Abeygunawardena,N., Holloway,E., Kapushesky,M., Kemmeren,P. and Lara,G.G. (2003) ArrayExpress—a public repository for microarray gene expression data at the EBI. *Nucleic Acids Res.*, **31**, 68–71.
22. Hutter,C. and Zenklusen,J.C. (2018) The cancer genome atlas: Creating lasting value beyond its data. *Cell*, **173**, 283.
23. Davis,C.A., Hitz,B.C., Sloan,C.A., Chan,E.T., Davidson,J.M., Idan,G., Hilton,J.A., Kriti,J., Baymuradov,U.K. and Narayanan,A.K. (2018) The Encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic Acids Res.*, **46**, D794–D801.
24. Ethan,C., Jianjiong,G., Ugur,D., Gross,B.E., Selcuk Onur,S., Bülent Arman,A., Anders,J., Byrne,C.J., Heuer,M.L. and Erik,L. (2012) The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov.*, **2**, 401–404.
25. Buhule,O.D., Minster,R.L., Hawley,N.L., Medvedovic,M., Sun,G., Viali,S., Deka,R., Mcgarvey,S.T. and Weeks,D.E. (2014) Stratified randomization controls better for batch effects in 450K methylation analysis: a cautionary tale. *Front. Genet.*, **5**, 354.
26. Harper,K.N., Peters,B.A. and Gamble,M.V. (2013) Batch effects and pathway analysis: two potential perils in cancer studies involving DNA methylation array analysis. *Cancer Epidemiol. Biomarkers Prev.*, **22**, 1052–1060.
27. Aryee,M.J., Jaffe,A.E., Corrada-Bravo,H., Ladd-Acosta,C., Feinberg,A.P., Hansen,K.D. and Irizarry,R.A. (2014) Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics*, **30**, 1363–1369.
28. Teschendorff,A.E., Francesco,M., Matthias,L., Thomas,B., Jesper,T., David,G.C. and Stephan,B. (2013) A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data. *Bioinformatics*, **29**, 189–196.
29. Heiss,J.A. and Just,A.C. (2019) Improved filtering of DNA methylation microarray data by detection *P* values and its impact on downstream analyses. *Clin. Epigenet.*, **11**, 15.
30. Lehne,B., Drong,A.W., Loh,M., Zhang,W., Scott,W.R., Tan,S.-T., Afzal,U., Scott,J., Jarvelin,M.-R. and Elliott,P. (2015) A coherent approach for analysis of the Illumina HumanMethylation450 BeadChip improves data quality and performance in epigenome-wide association studies. *Genome Biol.*, **16**, 37.
31. Heiss,J.A. and Just,A.C. (2018) Identifying mislabeled and contaminated DNA methylation microarray data: an extended quality control toolset with examples from GEO. *Clin. Epigenet.*, **10**, 73.
32. Li,M., Zou,D., Li,Z., Gao,R., Sang,J., Zhang,Y., Li,R., Xia,L., Zhang,T. and Niu,G. (2018) EWAS Atlas: a curated knowledgebase of epigenome-wide association studies. *Nucleic Acids Res.*, **47**, D983–D988.
33. White,A.J., Kresovich,J.K., Keller,J.P., Xu,Z., Kaufman,J.D., Weinberg,C.R., Taylor,J.A. and Sandler,D.P. (2019) Air pollution, particulate matter composition and methylation-based biologic age. *Environ. Int.*, **132**, 105071.
34. Sutton,L.P., Jeffreys,S.A., Phillips,J.L., Taberlay,P.C., Holloway,A.F., Ambrose,M., Joo,J.-H.E., Young,A., Berry,R. and Skala,M. (2019) DNA methylation changes following DNA damage in prostate cancer cells. *Epigenetics*, **14**, 989–1002.
35. Capper,D., Jones,D.T., Sill,M., Hovestadt,V., Schrimpf,D., Sturm,D., Koelsche,C., Sahm,F., Chavez,L. and Reuss,D.E. (2018) DNA methylation-based classification of central nervous system tumours. *Nature*, **555**, 469.
36. Ryan,L., Mattia,P., Dowen,R.H., R David,H., Gary,H., Julian,T.F., Nery,J.R., Leonard,L., Zhen,Y. and Que-Minh,N. (2009) Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature*, **462**, 315–322.
37. Yeshurun,S. and Hannan,A.J. (2019) Transgenerational epigenetic influences of paternal environmental exposures on brain function and predisposition to psychiatric disorders. *Mol. Psychiatry*, **24**, 536–548.
38. Forno,E., Wang,T., Qi,C., Yan,Q., Xu,C.-J., Boutaoui,N., Han,Y.-Y., Weeks,D.E., Jiang,Y. and Rosser,F. (2019) DNA methylation in nasal epithelium, atopy, and atopic asthma in children: a genome-wide study. *Lancet Respir. Med.*, **7**, 336–346.
39. Hall,E., Nitert,M.D., Volkov,P., Malmgren,S., Mulder,H., Bacos,K. and Ling,C. (2018) The effects of high glucose exposure on global gene expression and DNA methylation in human pancreatic islets. *Mol. Cell Endocrinol.*, **472**, 57–67.
40. Bell,J.T., Pai,A.A., Pickrell,J.K., Gaffney,D.J., Pique-Regi,R., Degner,J.F., Gilad,Y. and Pritchard,J.K. (2011) DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines. *Genome Biol.*, **12**, R10.