Mathematical Biosciences and Engineering

MBE, 18(6): 8951–8961.
DOI: 10.3934/mbe.2021441
Received: 15 August 2021
Accepted: 08 October 2021
Published: 18 October 2021

http://www.aimspress.com/journal/MBE

Protocol

Purimeth: an integrated web-based tool for estimating and accounting for tumor purity in cancer DNA methylation studies

Nana Wei[1], Hanwen Zhu[1], Chun Li[2,3,4,*] and Xiaoqi Zheng[1,*]

[1] Department of Mathematics, Shanghai Normal University, Shanghai, China
[2] School of Mathematics and Statistics, Hainan Normal University, Haikou, China
[3] Key Laboratory of Data Science and Intelligence Education, Ministry of Education, Hainan Normal University, Haikou, China
[4] Key Laboratory of Computational Science and Application of Hainan Province, Haikou, China

* **Correspondence:** Email: lichunlizhg@163.com, xqzheng@shnu.edu.cn.

Abstract: Proportion of cancerous cells in a tumor sample, known as "tumor purity", is a major source of confounding factor in cancer data analyses. Lots of computational methods are available for estimating tumor purity from different types of genomics data or based on different platforms, which makes it difficult to compare and integrate the estimated results. To rectify the deviation caused by tumor purity effect, a number of methods for downstream data analysis have been developed, including tumor sample clustering, association study and differential methylation between tumor samples. However, using these computational tools remains a daunting task for many researchers since they require non-trivial computational skills. To this end, we present Purimeth, an integrated web-based tool for estimating and accounting for tumor purity in cancer DNA methylation studies. Purimeth implements three state-of-the-art methods for tumor purity estimation from DNA methylation array data: InfiniumPurify, MEpurity and PAMES. It also provides graphical interface for various analyses including differential methylation (DM), sample clustering, and purification of tumor methylomes, all with the consideration of tumor purities. In addition, Purimeth catalogs estimated tumor purities for TCGA samples from nine methods for users to visualize and explore. In conclusion, Purimeth provides an easy-operated way for researchers to explore tumor purity and implement cancer methylation data analysis. It is developed using Shiny (Version 1.6.0) and freely available at http://purimeth.comp-epi.com/.

Keywords: tumor purity; differential methylation; tumor sample clustering; DNA methylation analysis; webserver

1. Introduction

Tumor purity is defined as the proportion of cancer cells in tumor tissues [1,2]. As a source of confounding factor, tumor purity has also been recognized to have significant impact on a variety of high-throughput data analyses based on gene expression or DNA methylation data. In this case, estimating tumor purity in the admixture of cells constituting tumor microenvironment is an important step to perform cancer genomic or epigenetic analyses. Inaccurate estimation of tumor purity would jeopardize downstream analyses such as clustering, association study and differential analysis between tumor samples [3]. Traditionally, tumor purity is generally estimated by pathologists using experimental methods, for example, using Immuno-histochemistry (IHC). In recent years, a number of computational methods have been developed to estimate tumor purity from different types of genomic data such as DNA copy numbers [4], gene expression [5], and DNA methylation [1,6–8]. These methods are comprehensively reviewed and compared [9,10]. Among all those types of genomic data, DNA methylation is deemed to be one of the most suitable data for purity estimation due to the following reasons: 1) DNA methylation is a long-term, and more stable biomarker than gene expression in detecting cancers [11]; 2) Nearby CpG sites are highly co-methylated under the mechanism of methyltransferase enzymes, which potentially reduces the random noises and increases inferring accuracy; 3) CpG sites in each individual cell are either methylated (methylation level = 1) or unmethylated (methylation level = 0), so methylation ratio of a tumor tissue intrinsically reflects proportion of some certain cell types. Under these considerations, we proposed InfiniumPurify, a tool for estimating purities of tumor samples based on Illumina Infinium 450 k methylation microarray [7,12]. It estimates the tumor purity by exploiting the beta value distribution of the most differential DNA methylation sites (informative differentially methylated CpG sites, iDMCs). Along this line, PAMES updated the selection of iDMCs by taking advantage of highly clonal cancer specific CpG sites [6]. MEpurity uses a beta mixture model to estimate tumor purity from only tumor methylation data [8].

In spite of the advances in this field, it is still technically challenging for biological and clinical researchers to take advantage of the methodological development. A main reason is that the aforementioned tools are developed based on various platforms, for example, InfiniumPurify [12] and PAMES are R packages, while MEpurity was written in C++. Using these tools could be a daunting task for many researchers since they require non-trivial computational skills, thus, there is an urgent need for a set of more accessible and intuitive tools. In this work, we develop Purimeth, an interactive and user-friendly web-based tool for analyzing cancer DNA methylation data, which can implement purity estimation and downstream data analyses accounting for tumor purity in a few clicks. By uploading beta value matrices of both tumor and normal samples, purity of tumor samples can be obtained by using three state-of-the-art tools for purity estimation from DNA methylation array data, i.e., InfiniumPurify, MEpurity and PAMES. Based on the purity estimates, users can perform a series of DNA methylation analyses accounting for tumor purity, including differential methylation analysis, clustering tumor samples into subtypes and purification of tumor methylomes. Users are also allowed to download and explore purities of 9364 tumor samples from The Cancer Genome Atlas (TCGA) using

nine methods including ESTIMATE [5], ABSOLUTE [4], LUMP, IHC, CPE [1], InfiniumPurify [12], PAMES [6], MEpurity [8] and Consensus (see the Result section for detail).

2. Materials and implementation

2.1. Workflow

The Purimeth webserver system consists of five major modules: GetPurity, Differential methylation (DM), Clustering, Purification and TCGA purity exploration. The general workflow of Purimeth in a typical data analysis is illustrated in Figure 1. First of all, it requires a matrix of methylation levels (in beta values) for tumor samples, and optionally a matrix of methylation levels for normal samples or cancer type as input (Figure 1A). After inputting these data, users can then estimate tumor purity using one of the three methods, i.e., InfiniumPurify, MEpurity and PAMES (Figure 1B). Finally, with estimated purities, users can perform variety of downstream data analyses including differential methylation (DM), clustering, or purification of tumor methylomes (Figure 1C).

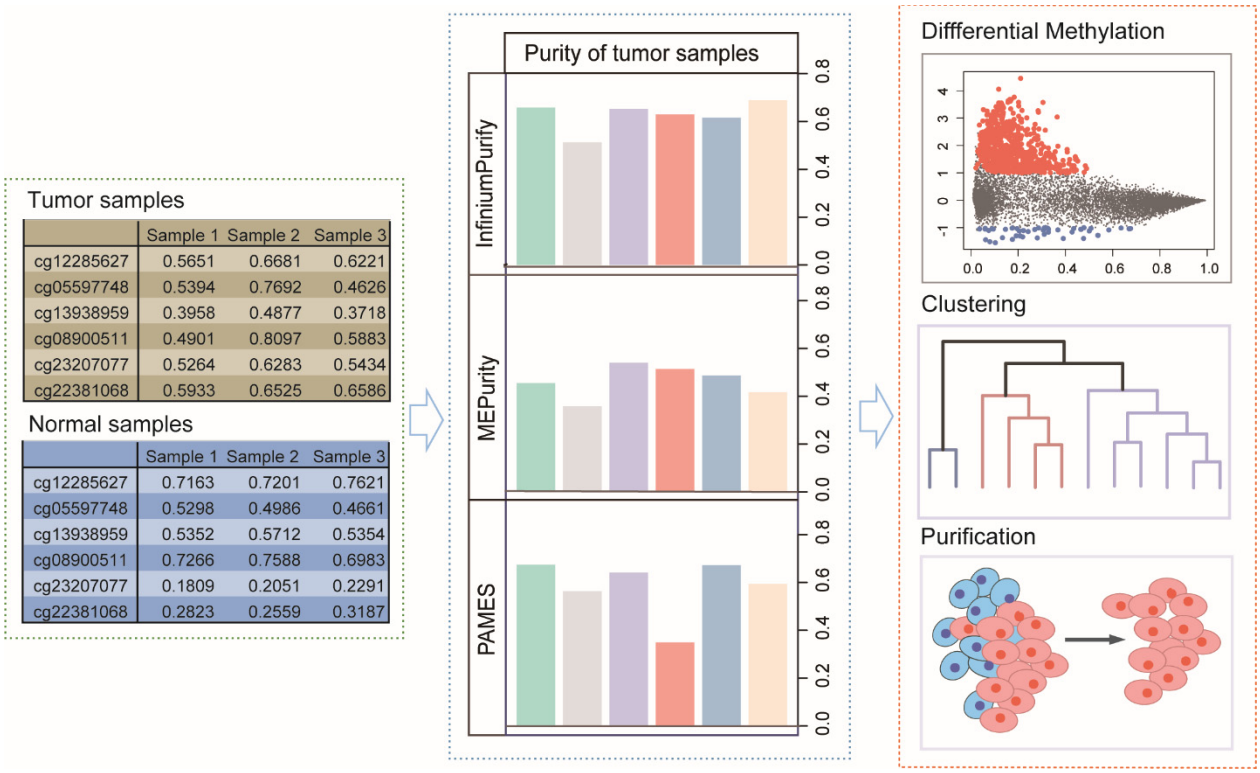

Figure 1. Workflow of Purimeth on tumor purity estimation and downstream data analyses. (A) Examples of input data format for tumor and normal samples, where rows are CpG sites and columns are samples. (B) Three existing tools for tumor purity estimation from DNA methylation data, i.e., InfiniumPurify, MEpurity and PAMES are implemented. (C) Using the purities estimated from the previous step or other related tools, three downstream modules (differential methylation, clustering and purification) for DNA methylation analyses accounting for tumor purity are implemented.

2.2. Data

Purimeth allows users to upload methylation profiles of tumor and normal sample from either 450 k or 850 k array in the form of a .txt, .zip or .gz file, where rows and columns represent the CpG sites and samples respectively. In some modules, users need to upload purity file (in same format) for tumor samples as well. For the convenience of users, we provide example data files for breast tumor samples and matched normal samples from TCGA and their corresponding tumor purity file on the website.

3. Results

3.1. Purity estimation

Increasing attention has been devoted to the relationship between tumor purity and various studies on tumor samples. For the purpose of obtaining the purity of the tumor sample fast and sound, the "GetPurity" module allows users to estimate purities of tumor samples by InfiniumPurify, MEpurity or PAMES on the same page according to the method selected. InfininumPurify estimates purity from the probability density of methylation levels of iDMCs from cancer-normal comparisons. MEpurity estimates tumor purity based on tumor-only Illumina Infinium 450 k methylation microarray data using a beta mixture model-based algorithm. PAMES uses the methylation level of a few dozens of highly clonal tumor type specific CpG sites to estimate the purity of tumor samples, and only works for 450 k array data in its current edition. For MEpurity users only need to upload DNA methylation matrix (where rows are for CpG sites and columns for samples). And for InfiniumPurify and PAMES, besides the beta value matrix of tumor samples, either cancer type that can be specified by the select button 'Cancer Type' or normal sample data should be inputted. In addition, the cancer type should be specified for InfiniumPurify if the inputted normal samples are insufficient for iDMC identification (less than 20). Once a file is uploaded, the first six rows of the data will automatically be shown so that the user can confirm whether the file is correct (same for other modules). With a click on the "Run" button, a table with the estimated purities of tumor samples will be displayed in the Result panel (Figure 2A). Meanwhile, a barplot will be shown in the Plot panel for visualizing the estimated purities (Figure 2B). To provide users a reference, we tested the running time of three methods using a typical example data of 20 tumor samples, which are shown in Figure 2C. When inputting only tumor samples, MEpurity takes more than 10 seconds to get the result, while InfinumPurify and PAMES take only less than 1 second. When both tumor and matched normal samples are input, MEpurity does not work in this case, while InfiniumPurity still runs faster than PAMES.

A

	purity
TCGA.EW.A1PH	0.8232
TCGA.A7.A26F	0.539
TCGA.A2.A0YL	0.47
TCGA.AC.A6NO	0.7455
TCGA.AO.A0JB	0.7782
TCGA.A2.A04R	0.8498
TCGA.BH.A1FC	0.6383
TCGA.AR.A2LJ	0.5059
TCGA.GM.A2DA	0.5968
TCGA.D8.A27R	0.6024

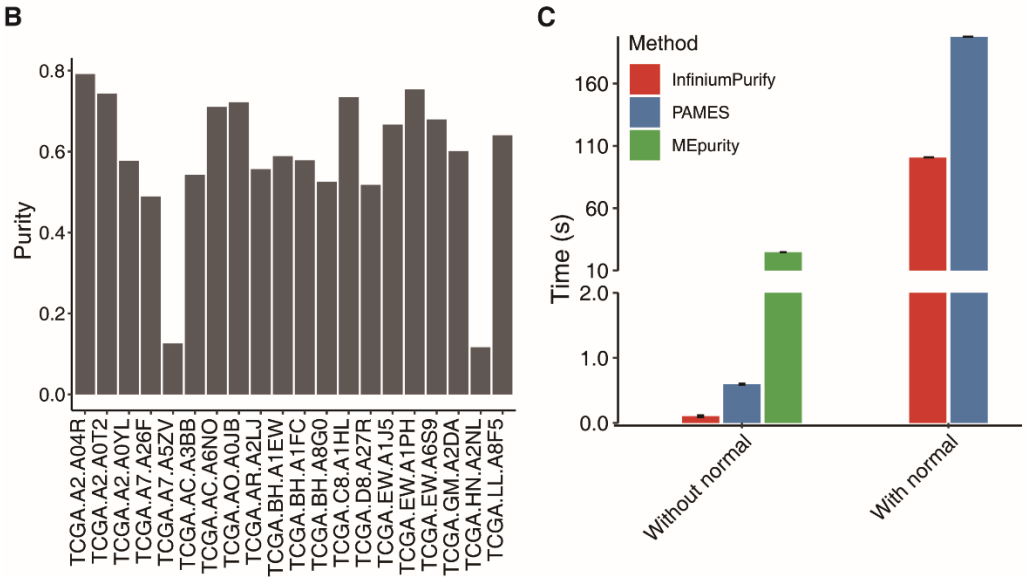

Figure 2. An illustration of the 'GetPurity' module. (A) A table or (B) barplot of estimated purities for tumor samples. (C) Running time of three methods on an example data.

3.2. Differential methylation

Differential methylation (DM) between tumor and normal samples, or between two groups of tumor samples showing different phenotypes is a central task in cancer epigenomics research. The differentially methylated CpG sites (DMCs) or regions (DMRs) could potentially serve as diagnostic biomarkers or therapeutic targets [13–16]. In Purimeth, DM module contains two submodules, "Tumor vs Normal" and "Tumor1 vs Tumor2", allowing users to infer the differentially methylated CpG sites accounting for tumor purity. These two modules correspond to our two previous works [7,17], both of which are based on the generalized linear regression model and Wald test to call DM sites. For 'tumor vs normal', users are needed to input beta value matrices of tumor and normal samples, as well as

tumor purity file for tumor samples, which could be obtained from the "GetPurity" module. And for "tumor1 vs tumor2", beta value matrices and tumor purity files (obtained from the first module) for both subtypes of tumor samples are required. Purimeth will return a list of differentially methylated CpG sites sorted by their q-values (Figure 3A). Meanwhile, besides a heat map showing the top N differentially methylated CpG sites (N could be set by users) (Figure 3B), it will also provide a scatter plot illustrating log2 fold change of average corrected methylation level between two sample groups for CpG sites (Figure 3C).

A

	pval	qval	tumor.mean	normal.mean	log2FC
cg20395967	0	0	0.6909	0.0933	2.8878
cg00970396	0	0	0.6769	0.0694	3.2863
cg17899889	0	0	0.3799	0.8642	-1.1857
cg26132320	0	0	0.707	0.1223	2.5314
cg27170427	0	0	0.3763	0.9198	-1.2895
cg09165441	0	0	0.6879	0.1174	2.5503
cg11957331	0	0	0.6241	0.0475	3.715
cg23900203	0	0	0.6255	0.1468	2.0909
cg03940848	0	0	0.622	0.0632	3.2991
cg05951296	0	0	0.3883	0.8984	-1.2101

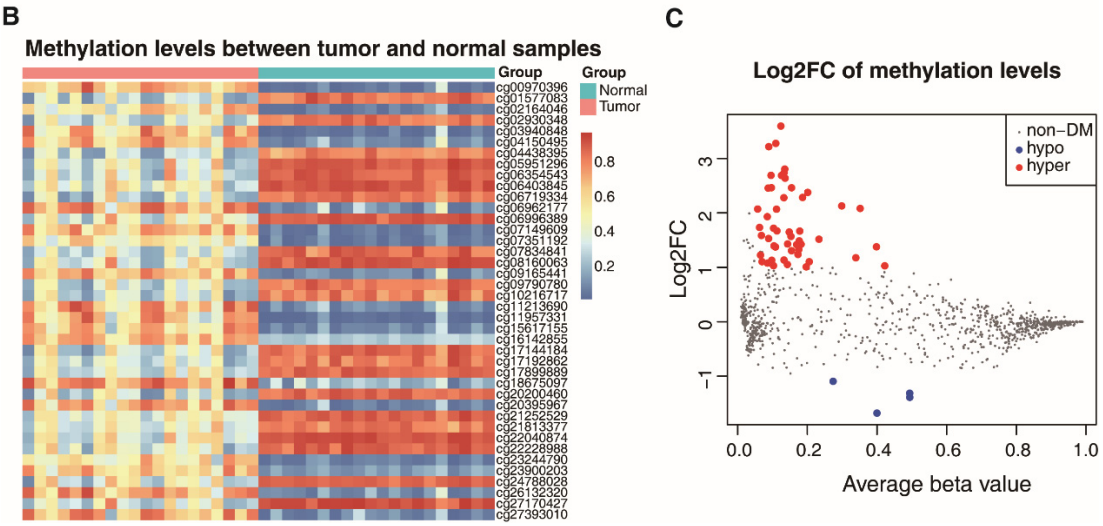

Figure 3. An illustration of outputs for 'DM' module. (A) Differential methylated CpG sites ranked by the q-values between tumor and normal samples. Users can also sort the result by clicking the column name. (B) A heatmap displaying DNA methylation levels of most differential methylated CpG sites between two groups. (C) MA-plot of the differentially methylated result between two groups of samples.

3.3. Clustering of tumor samples

The identification of tumor subtypes is of great significance for the early diagnosis and clinical

treatment of cancer. Given both DNA methylation profiles of tumor samples and tumor purities, the "Clustering" module allows users to cluster tumor samples into different subtypes. It models the subtype of a tumor mixture sample as a latent variable in a statistical model and solves it by the Expectation-Maximization (EM) algorithm [18]. The purity file inputted can be obtained from "GetPurity" module, other purity estimation tools or pathologists. This module performs with the adjustment of several parameters including the number of clusters, the maximum number of iterations and tolerance for convergence of EM iterations. The clustering result shows the predicted subtype for each sample (Figure 4A) and visualizes all samples by plotting the first two principal components of the data (Figure 4B).

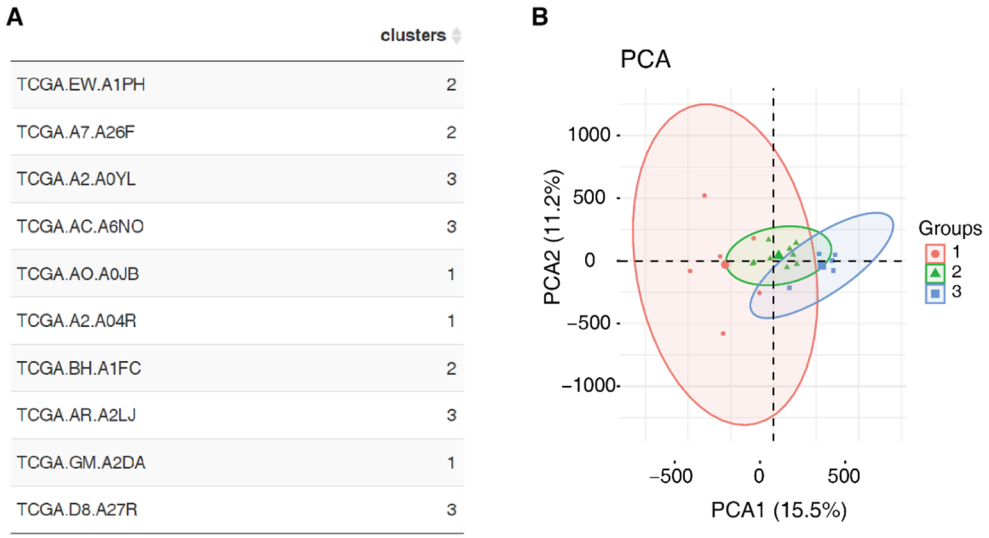

Figure 4. An illustration of the 'Clustering' module. (A) A table displaying the clustering results of the breast samples by 'Clustering' module, where number of clusters, maximum number of iterations and tolerance for convergence of EM iterations are 3, 100, 0.001 respectively. (B) A scatter plot visualizing the clustering result by using two principal components of tumor data.

3.4. Purification of tumor samples

Methylation profiles of pure cancer cells can be hardly obtained from real tumor tissues which are always mixtures of normal and cancer cells. In this situation, the "Purification" module aims to infer methylation profiles of pure cancer cells from tumor mixture samples, matched normal samples and tumor purities. This module implements a regression-based model to get rid of the normal cell signals and obtain pure cancer cell methylomes. After uploading the data and clicking the 'Run' button, Purimeth will report purified methylation profiles in a table (Figure 5A), and show the boxplots of tumor, normal and purified tumor data for 4 example CpG sites (Figure 5B). Users can also show barplots for any CpG sites (sorted by the average methylation difference between tumor and purified tumor samples) of interest by using the input box "Choose a CpG(s) site for plot".

A

	TCGA.EW.A1PH	TCGA.A7.A26F	TCGA.A2.A0YL	TCGA.AC.A6NO
cg00000029	0.0536	0.2089	0.0874	0.114
cg00000165	0.6971	0.797	0.3631	0.2615
cg00000236	0.8853	0.912	0.9133	0.9135
cg00000289	0.6023	0.718	0.5676	0.705
cg00000292	0.5173	0.591	0.5777	0.4218
cg00000321	0.472	0.7432	0.4169	0.5872
cg00000363	0.1313	0.3904	0.548	0.2683
cg00000622	0.013	0.0174	0.0144	0.0141
cg00000658	0.8274	0.8802	0.9013	0.7952

B

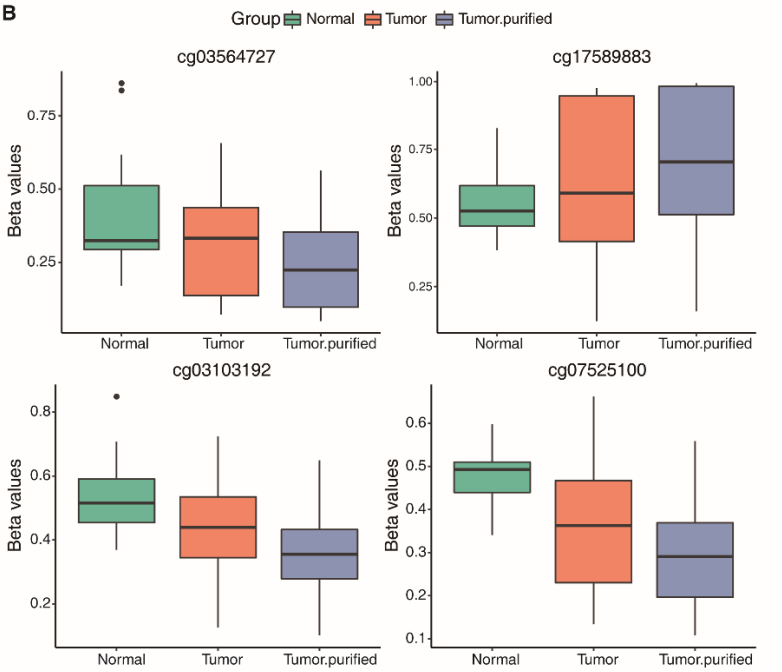

Figure 5. Application of 'Purification' module. (A) Purified tumor methylomes by the consideration of tumor purity. (B) Comparison of normal, tumor and purified tumor methylations of a few example CpG sites.

3.5. Purity estimation for TCGA tumor samples

A number of tools have been proposed to estimate tumor purity for TCGA tumor samples from different types of genomics data by using different underline models. However, the estimates for the same samples vary by method. Thus comparison and integration of estimates from different methods are needed. Motivated by Aran et al. [1], we created a consensus purity estimate (named "Consensus") by taking the median of purities estimated from five available methods including ABSOLUTE, ESTIMATE, LUMP, IHC and InfiniumPurify after normalization. Compared with the original CPE method, our update method includes the purities of InfiniumPurify. In the last module, Purimeth integrates tumor purity estimates of TCGA tumor samples using Consensus and the following eight state-of-the-art tools, i.e., ESTIMATE, ABSOLUTE, LUMP, IHC, CPE, InfiniumPurify, PAMES and MEpurity. Users can select any cancer types, methods and samples of interest to obtain their

corresponding purities which will be shown in the result panel (Figure 6A). If multiple samples are inputted, each sample should be separated by a comma. Based on the number of samples from the same cancer type, the plot panel will display two different figures. If there is only one sample for a cancer type, a bar plot will be displayed for this cancer type (Figure 6B). Otherwise, a box plot will be generated for each selected cancer type or method (Figure 6C). To compare the performance of different methods, we calculate the Pearson's correlation between each of the two methods on 21 cancer types and all merged samples. As shown in Figure 6D, InfiniumPurify and Consensus methods show the highest overall consistence for all cancer types compared to other methods.

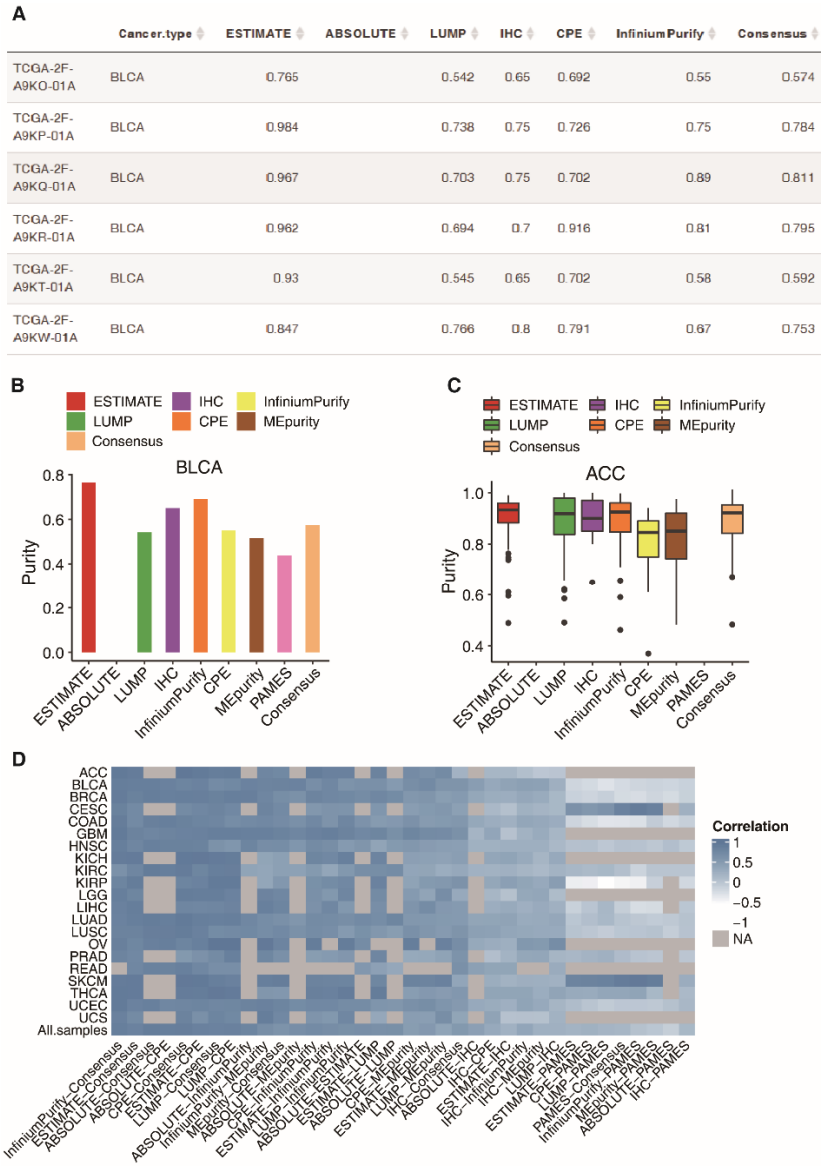

Figure 6. Application of 'TCGApurity' module. (A) A table of tumor purities estimated by six different methods for four cancer types. (B) A barplot showing tumor purity estimates for two samples of BRCA and LUAD. (C) Tumor purity distributions of a set of tumor samples by six methods. (D) Pairwise Pearson's correlations between all purity estimation methods for 21 cancer types and all merged samples, where grey cells mean that data are not available from both methods.

4. Discussion and conclusions

Estimating and accounting for tumor purity from DNA methylation data are hot topics in cancer research. In recent years, multiple purity estimation tools have been developed by different algorithms and software platforms. Using these tools could be a daunting task for many researchers since those methods require non-trivial computational skills. In this work, we developed Purimeth, an integrated web-based tool for estimating and accounting for tumor purity in DNA methylation studies. Besides Infinium 450 k array data, our tool was also tested on the latest EPIC bead chip (850 k array) data. Since the methylation profiles measured by microarray and bisulfite sequencing are highly consistent, our tool designed for microarray data also works for sequencing data including WGBS, RRBS and HMST-seq. For a given cancer type, users only need to extract methylation levels of its informative DMC sites (iDMCs), and upload it according to the example file format. We provided the iDMCs for 32 cancer types as a download link in the GetPurity module. As an example, we also provided a demo (in supplementary file) to use Purimeth on WGBS data of colon cancer samples, including purity estimation and differential methylation analysis. Overall, our study provides a comprehensive web tool for researchers to perform DNA methylation data analyses regarding tumor purities.

Purimeth was developed by Shiny (Version 1.6.0) on Tencent cloud server, which enables better stability and scalability for computing resources. The computational times for purity estimation, DM and purification are less than 2 minutes for a typical data set of 20 tumor and 20 normal samples, while the Clustering module is more time-consuming which will take 2 to 10 minutes to get the clustering results depending on the number of tumor samples and necessary steps for iteration given.

Acknowledgments

We thank Shijiang Wang for the suggestions on codes and web server design.

This work was supported by the National Key R & D Program of China [2018YFA0900600 to X.Z.]; National Natural Science Foundation of China [61972257 to X.Z.]; Key Laboratory of Data Science and Intelligence Education (Hainan Normal University), Ministry of Education [DSIE202002 to X.Z.] and the Hainan Province Natural Science Foundation [No. 2019RC184 to C.L.].

Conflict of interest

All authors declare there is no conflicts of interest.

References

1. D. Aran, M. Sirota, A. J. Butte, Systematic pan-cancer analysis of tumour purity, *Nat. Commun.*, **6** (2015), 8971.
2. J. K. Rhee, Y. C. Jung, K. R. Kim, J. Yoo, J. Kim, Y. J. Lee, et al., Impact of tumor purity on immune gene expression and clustering analyses across multiple cancer types, *Cancer Immunol. Res.*, **6** (2018), 87–97.
3. A. E. Jaffe, R. A. Irizarry, Accounting for cellular heterogeneity is critical in epigenome-wide association studies, *Genome Biol.*, **15** (2014), 1–9.

4. S. L. Carter, K. Cibulskis, E. Helman, A. McKenna, H. Shen, T. Zack, et al., Absolute quantification of somatic DNA alterations in human cancer, *Nat. Biotechnol.*, **30** (2012), 413–421.
5. K. Yoshihara, M. Shahmoradgoli, E. Martínez, R. Vegesna, H. Kim, W. Torres-Garcia, et al., Inferring tumour purity and stromal and immune cell admixture from expression data, *Nat. Commun.*, **4** (2013), 2612.
6. M. Benelli, D. Romagnoli, F. Demichelis, Tumor purity quantification by clonal DNA methylation signatures, *Bioinformatics*, **34** (2018), 1642–1649.
7. X. Zheng, N. Zhang, H. J. Wu, H. Wu, Estimating and accounting for tumor purity in the analysis of DNA methylation data from cancer studies, *Genome Biol.*, **18** (2017), 1–14.
8. B. W. Liu, X. F. Yang, T. J. Wang, J. D. Lin, Y. Y. Kang, P. Jia, et al., MEpurity: estimating tumor purity using DNA methylation data, *Bioinformatics*, **35** (2019), 5298–5300.
9. S. Haider, S. Tyekucheva, D. Prandi, N. S. Fox, J. Ahn, A. W. Xu, et al., Systematic assessment of tumor purity and its clinical implications, *JCO Precis. Oncol.*, **4** (2020), 995–1005.
10. F. Wang, N. Zhang, J. Wang, H. Wu, X. Zheng, Tumor purity and differential methylation in cancer epigenomics, *Briefings Funct. Genomics*, **15** (2016), 408–419.
11. A. Paziewska, M. Dabrowska, K. Goryca, A. Antoniewicz, J. Dobruch, M. Mikula, et al., DNA methylation status is more reliable than gene expression at detecting cancer in prostate biopsy, *Br. J. Cancer*, **111** (2014), 781–789.
12. N. Zhang, H. J. Wu, W. Zhang, J. Wang, H. Wu, X. Zheng, Predicting tumor purity from methylation microarray data, *Bioinformatics*, **31** (2015), 3401–3405.
13. M. T. Gyparaki, E. K. Basdra, A. G. Papavassiliou, DNA methylation biomarkers as diagnostic and prognostic tools in colorectal cancer, *J. Mol. Med.*, **91** (2013), 1249–1256.
14. B. N. Lasseigne, T. C. Burwell, M. A. Patil, D. M. Absher, J. D. Brooks, R. M. Myers, DNA methylation profiling reveals novel diagnostic biomarkers in renal cell carcinoma, *BMC Med.*, **12** (2014), 235.
15. G. Nikolaidis, O. Y. Raji, S. Markopoulou, J. R. Gosney, J. Bryan, C. Warburton, et al., DNA methylation biomarkers offer improved diagnostic efficiency in lung cancer, *Cancer Res.*, **72** (2012), 5692–5701.
16. A. Farooq, S. Grønmyr, O. Ali, T. Rognes, K. Scheffler, M. Bjørås, et al., HMST-Seq-Analyzer: A new python tool for differential methylation and hydroxymethylation analysis in various DNA methylation sequencing data, *Comput. Struct. Biotechnol. J.*, **18** (2020), 2877–2889.
17. W. Zhang, Z. Li, N. Wei, H. J. Wu, X. Zheng, Detection of differentially methylated CpG sites between tumor samples with uneven tumor purities, *Bioinformatics*, **36** (2020), 2017–2024.
18. W. Zhang, H. Feng, H. Wu, X. Zheng, Accounting for tumor purity improves cancer subtype classification from DNA methylation data, *Bioinformatics*, **33** (2017), 2651–2657.

AIMS Press