frontiers | Frontiers in Genetics

TYPE Original Research
PUBLISHED 13 March 2024
DOI 10.3389/fgene.2024.1367731

OPEN ACCESS

EDITED BY
Xuefeng Cui,
Shandong University, China

REVIEWED BY
Advait Balaji,
Occidental Petroleum Corporation, United States
Wei Lan,
Guangxi University, China

*CORRESPONDENCE
Le Zhang,
zhangle06@scu.edu.cn

RECEIVED 09 January 2024
ACCEPTED 08 February 2024
PUBLISHED 13 March 2024

CITATION
Xiao M, Xiao Y, Yu J and Zhang L (2024), PCGIMA: developing the web server for human position-defined CpG islands methylation analysis.
Front. Genet. 15:1367731.
doi: 10.3389/fgene.2024.1367731

PCGIMA: developing the web server for human position-defined CpG islands methylation analysis

Ming Xiao[1,2], Yi Xiao[1], Jun Yu[3,4] and Le Zhang[1,5,6]*

[1]College of Computer Science, Sichuan University, Chengdu, China, [2]Tianfu Engineering-oriented Numerical Simulation and Software Innovation Center, Chengdu, China, [3]CAS Key Laboratory of Genome Sciences and Information, Chinese Academy of Sciences, Beijing Institute of Genomics, Beijing, China, [4]University of Chinese Academy of Sciences, Beijing, China, [5]Key Laboratory of Systems Biology, Chinese Academy of Sciences, Hangzhou Institute for Advanced Study, University of Chinese Academy of Sciences, Hangzhou, China, [6]Key Laboratory of Systems Health Science of Zhejiang Province, Hangzhou Institute for Advanced Study, University of Chinese Academy of Sciences, Hangzhou, China

Introduction: CpG island (CGI) methylation is one of the key epigenomic mechanisms for gene expression regulation and chromosomal integrity. However, classical CGI prediction methods are neither easy to locate those short and position-sensitive CGIs (CpG islets), nor investigate genetic and expression pattern for CGIs under different CpG position- and interval-sensitive parameters in a genome-wide perspective. Therefore, it is urgent for us to develop such a bioinformatic algorithm that not only can locate CpG islets, but also provide CGI methylation site annotation and functional analysis to investigate the regulatory mechanisms for CGI methylation.

Methods: This study develops Human position-defined CGI prediction method to locate CpG islets using high performance computing, and then builds up a novel human genome annotation and analysis method to investigate the connections among CGI, gene expression and methylation. Finally, we integrate these functions into PCGIMA to provide relevant online computing and visualization service.

Results: The main results include: (1) Human position-defined CGI prediction method is more efficient to predict position-defined CGIs with multiple consecutive (d) values and locate more potential short CGIs than previous CGI prediction methods. (2) Our annotation and analysis method not only can investigate the connections between position-defined CGI methylation and gene expression specificity from a genome-wide perspective, but also can analysis the potential association of position-defined CGIs with gene functions. (3) PCGIMA (http://www.combio-lezhang.online/pcgima/home.html) provides an easy-to-use analysis and visualization platform for human CGI prediction and methylation.

Discussion: This study not only develops Human position-defined CGI prediction method to locate short and position-sensitive CGIs (CpG islets) using high performance computing to construct MR-CpGCluster algorithm, but also a

novel human genome annotation and analysis method to investigate the connections among CGI, gene expression and methylation. Finally, we integrate them into PCGIMA for online computing and visualization.

KEYWORDS

position-defined CGIs, DNA methylation, genome annotation, high performance computing, genome analysis

1 Introduction

CpG island (CGI) methylation is one of the key epigenomic mechanisms for gene expression regulation and chromosomal integrity (Dor and Cedar, 2018). Especially, recent studies have reported that position-sensitive CGI co-methylation mechanism is essential for such functions that are related to histone modification (Ming, et al., 2021). However, it is neither easy for current commonly used classical CGI island prediction methods (Gardinergarden and Frommer, 1987; Han et al., 2008; Takahashi et al., 2017) to locate those short and position-sensitive CGIs which called CpG islets (Hackenberg et al., 2006) due to the length limitation, nor investigate relationship among CGI density, methylation, and gene expression specificity. Therefore, it is urgent for us to develop such a bioinformatic algorithm that not only can locate short and position-sensitive CGIs (CpG islets), but also provide CGI methylation site annotation and functional analysis to investigate the regulatory mechanisms for CGI methylation (http://www.combio-lezhang.online/pcgima/home.html).

For CGI perdition method, we usually employ the unsupervised clustering methods such as CpGCluster (Hackenberg et al., 2006) and CPG_MI (Su et al., 2009) to locate CGIs with shorter length than the supervised (Bock et al., 2007; Ning et al., 2017), since these unsupervised algorithms do not need consider the constraints of CGI length and content ratio (Hackenberg et al., 2010). However, these methods are not only time-consuming for the big dataset, but also cannot investigate the genetic characteristics of CGIs under different CpG interval parameters. Therefore, our first scientific question is how to develop a novel CGI prediction method with CpG interval parameters selective feature and high-performance computing, and investigate the differences in genetic characteristics such as CpG coverage, CGI length, and GC content of CGIs under various CpG interval parameters.

Several previous studies have interrogated the connections between methylation and CGI (Reik, 2007; Smith et al., 2012; Liu et al., 2016; El-Maarri, 2019; Acton et al., 2021). For example, Ziller et al. (2013) have turned out that not only the hypermethylation of promoter CGI is related to gene expression, but also CGI methylation in the gene body is positively correlated with gene expression. However, these studies usually interrogate the methylation characteristics of CGI from partial sequence regions rather than genome-wide perspective. Meanwhile, although our previous studies (Zhang et al., 2018; Zhang et al., 2021a) have analyzed the relationship between CGI density and gene expression after annotating genome-wide CGI-related genes (CGI+) into high-CGI (HCGI), intermediate-CGI (ICGI), and low-CGI (LCGI) genes based on the classification of CGI density (Weber et al., 2007; Zhu et al., 2008), we are still unclear the relationship between CpG methylation and gene expression. Thus, our second scientific question is how to build up a human genome-wide CGI-based methylation and gene expression annotation and analysis method to investigate the relationship among CGI density, methylation, and gene expression specificity.

Meanwhile, although several CpG methylation online service are already available (Raney et al., 2010; Di et al., 2018; Xiong et al., 2019), most of them only focus on CpG island prediction and data downloading, but not provide visualization and analysis for the distribution of CGI in different sequence regions and the connections between methylation status of CGIs and gene expression. Therefore, our third scientific question is how to establish an easy-to-use web service for fast CGIs prediction and visualization of the connections between CGIs and methylation.

For these reasons, we propose three major innovations to answer the above scientific questions.

Firstly, we develop an unsupervised clustering-based CGI prediction method (Human position-defined CGI prediction), which not only employs high performance computing technology to accelerate its predictive speed, but also offers a parameter selective option that can help us to locate short CGIs (position-defined CGIs) with unique location- or sequence-sensitive features and explore the differences in the genetic characteristics of CGIs under various CpG interval parameters.

Secondly, we build up a novel human genome annotation and analysis function (Human position-defined CGI annotation and analysis), which not only can study the methylation characteristics of CGIs from a genome-wide perspective by computing the methylation level of all CpG sites in the human genome, but also improve the previous CpG-Island-based human gene expression annotation and analysis method (Zhang et al., 2021a) by integrating genome-wide methylation annotation to further investigate the connections among CGI density, gene expression and methylation.

Thirdly, we establish an easy-to-use web service "Position-defined CGI methylation analysis (PCGIMA)" with relevant CGI prediction, annotation, and data analysis functions, which provides us an online platform for further study on the regulation mechanism of CGI and methylation.

In conclusion, we develop a bioinformatic algorithm and web service to investigate the regulatory mechanism of CGI methylation. The main results include: 1) Human position-defined CGI prediction method is more efficient to predict position-defined CGIs with multiple consecutive (d) values and locate more potential short CGIs than previous CGI prediction methods; 2) Our annotation and analysis method not only can investigate the connections between position-defined CGI methylation and gene expression specificity from a genome-wide perspective, but also can analyze the potential association of position-defined CGIs with gene functions; (3) PCGIMA provides an easy-to-use analysis and visualization platform for human CGI prediction and methylation.

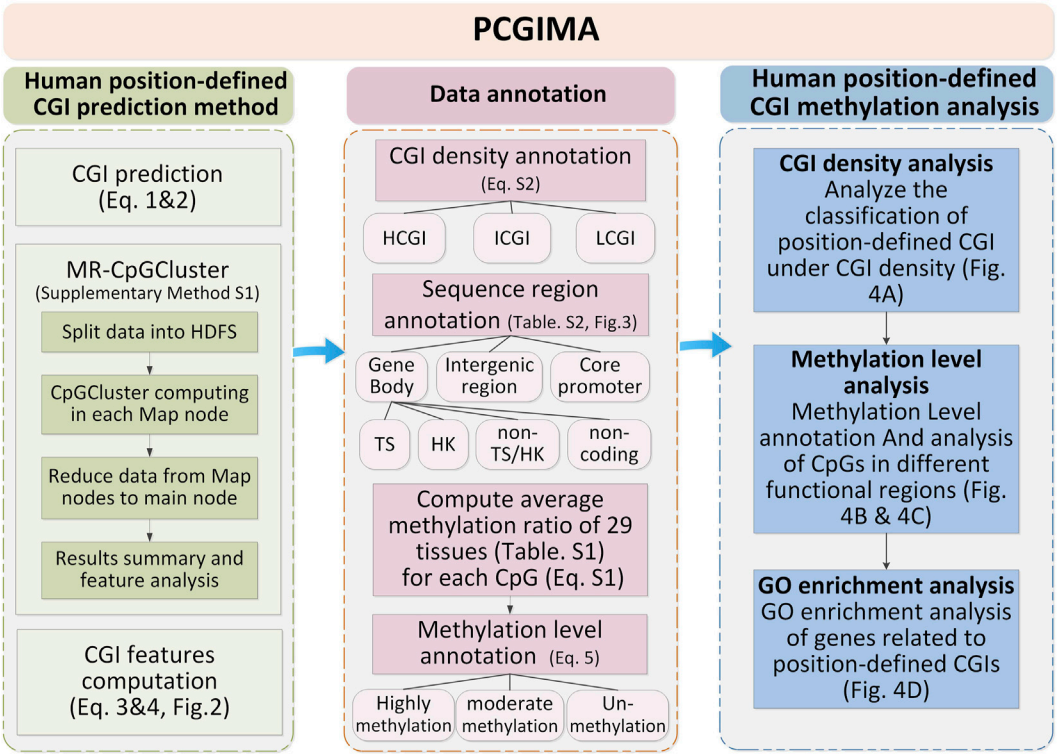

FIGURE 1
Workflow of the study.

2 Materials and methods

This study downloads human genome data from GRCh38 assembly (Schneider et al., 2017) at NCBI (Pruitt et al., 2005). To classify CGIs into density-defined and position-defined groups, we download human CGIs data and annotations from UCSC (Casper et al., 2018). Next, we use human genome annotated data (release 24) in GenBank GBFF format (Clark et al., 2016) from GENCODE (Wright et al., 2016) to define different sequence regions. Finally, to study the methylation level of CpG sites in different sequence regions, we obtain all CpG methylation data of 29 human tissues (Supplementary Table S1), including heart, spleen, lung and esophagus, from ENCODE databases (Harrow et al., 2006). In order to ensure data consistency, the above-listed annotation and methylation data are all annotated according to GRCh38 (Schneider et al., 2017). Figure 1 describes the workflow of the study with three essential steps: Human position-defined CGI prediction (left side of Figure 1), Data annotation (right side of Figure 1), and Human position-defined CGI methylation analysis (Bottom side of Figure 1).

Here, we describe the key equations as follows:

(1) CGI prediction: We employed Eq. 1 to define CpGs clusters (Hackenberg et al., 2006) at the start. Next, we consider these CpG clusters with small p-values (Eq. 2) as CGIs (Hackenberg et al., 2006).

$$d_i = x_{i+1} - x_i - 1 \tag{1}$$

Here, x and I represent the position and index of a CpG, respectively.

$$\mathbf{P}(d) = (1-p)^{d-1}p \tag{2}$$

P(d) represents the probability to find a distance d between neighboring CpGs. p corresponds to the probability of CpGs in the sequence. Since our previous studies (Zhang et al., 2018; Zhang et al., 2021a) has led to a conclusion that LAUPs (Lineage-associated underrepresented permutations) are closely related to CGIs and the shortest LAUPs of mammals range from 10bp to 14bp in length, here we use the intermediate value of d = 12bp.

(2) MR-CpGCluster: We develop a MR-CpGCluster algorithm (Supplementary Figure S1) to speed up CGI predict procedure based on MapReduce (Dittrich and Quiané-Ruiz, 2012) and Hadoop Streaming (Dede et al., 2016) techniques detailed by Supplementary Method S1 for Human position-defined CGI prediction method. Finally, our method computes the CGI features of the position-defined CGIs for subsequent analysis.

(3) CGI features computation: To compare the CGIs under different CpG distance intervals (Eq. 1), we compute CGI length, CG content, CpG O/E ratio (Gardinergarden and Frommer, 1987) (Eq. 3) and CpG density (Eq. 4) for each CGI (Hackenberg et al., 2006).

$$\mathbf{O/E} = \frac{\mathbf{CpNum}}{\mathbf{CNum} \times \mathbf{GNum}} \times \mathbf{N} \tag{3}$$

$$\mathbf{CpGdensity} = \frac{\mathbf{CpNum}}{\mathbf{N}} \tag{4}$$

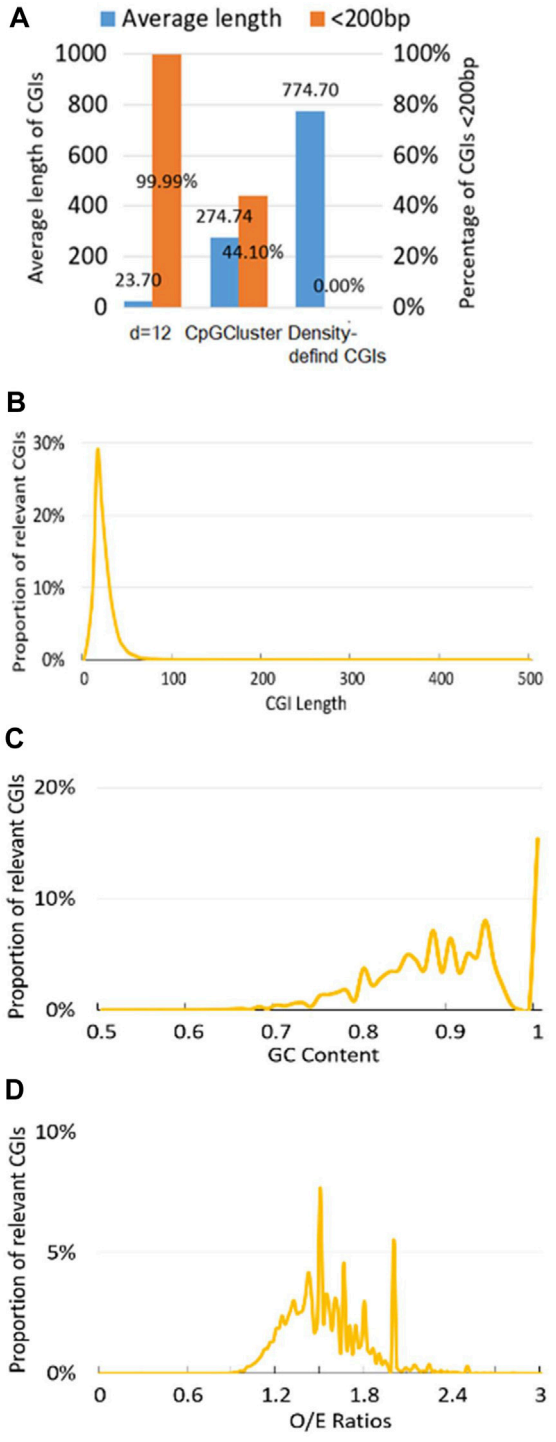

FIGURE 2
Position-defined CGI prediction and analysis. **(A)** CGIs comparative analysis. The proportion distribution of CGI at different **(B)** CGI length, **(C)** GC content, and **(D)** CpG O/E ratios (Eq. 3) under CpG interval d = 12bp.

Here, N is the length of the CGI, CpGNum, CNum and GNum represent the number of CpG, number of C, number of G respectively.

(4) Methylation level annotation: Eq. 5 classifies methylation ratio into three levels with respect to the definition (Ziller et al., 2013).

$$\text{Methylation level}(chr,p) = \begin{cases} 1, \text{highly mehylated} & \text{methylation\_ratio}(chr,p) > 0.75 \\ 2, \text{unmethylated} & \text{methylation\_ratio}(chr,p) < 0.1 \\ 3, \text{moderate methylated} & \text{otherwise} \end{cases} \quad (5)$$

Here, chr and p represent the chromosome and position of a CpG site, respectively.

3 Results

3.1 Human position-defined CGI prediction method

Indicated by previous study (Hackenberg et al., 2006), we consider CGIs as potentially functionally short islands (CpG islets) if length of CGIs is less than 200bp. Here, Figure 2A demonstrates that Human position-defined CGI prediction method not only can locate the shortest average length (23.7bp) under CpG interval d = 12bp, but also the percentage of CGIs <200bp for Human position-defined CGI prediction method are greater than both CpGCluster method (Hackenberg et al., 2006) and density-defined CGI prediction method (Weber et al., 2007; Zhang et al., 2021a).

Also, since proportion distribution of CGI features is closely related to the regulatory mechanisms for CGI methylation (Hackenberg et al., 2010), Human position-defined CGI prediction method can describe the proportional distribution of the predicted CGIs at different CGI length (Figure 2B), GC content (Figure 2C), and O/E (Figure 2D). Here, we employ default setup for CpG interval, d = 12bp (Zhang et al., 2018; Zhang et al., 2021a).

It should be noted that Human position-defined CGI prediction method can parallel carry out position-defined CGI prediction and comparative analysis for multiple CpG intervals (d) by MR-CpGCluster.

3.2 Data annotation

Data annotation is described by the right side of Figure 1. Firstly, the position-defined CGIs are classified into different densities by Supplementary Eq. S2. And then, we classify each CpG methylation site of CGIs into different gene functional regions by Supplementary Table S2. Lastly, we classify the CpG sites into three methylation levels by Eq.5.

Data annotation can help us investigate the distribution of all CpG sites in different structural and functional categories of genome sequences (Figure 3; Supplementary Table S3). For example, we not only can compare the distribution of the number of CpG sites in each region of the predicted CGIs under different CpG interval(d) (Figure 3A), but also visualize the density of CpG sites in different functional regions (Figure 3B).

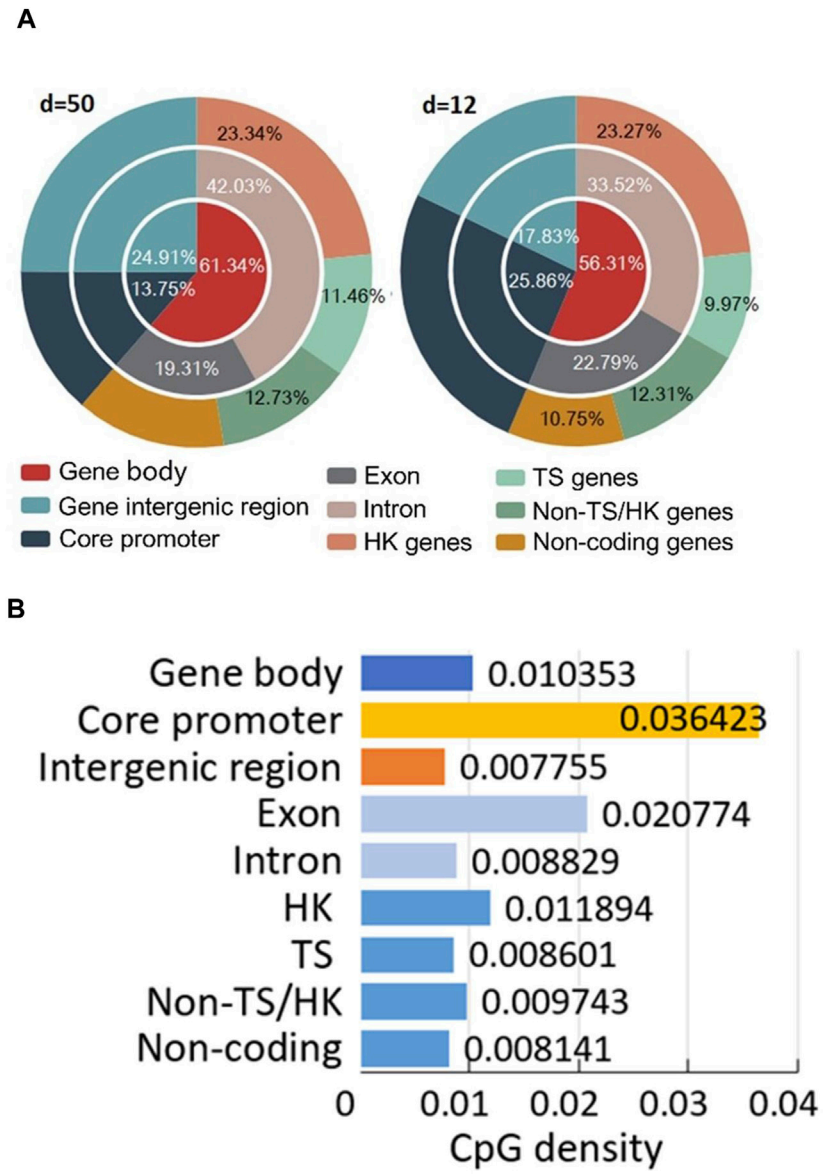

FIGURE 3
Position-defined CGI annotation results. **(A)** Distribution of all CpG sites in different structural and functional categories of genome sequences. **(B)** CpG density (Eq. 4) of different gene and sequence categories.

3.3 Human position-defined CGI methylation analysis

The position-defined CGI methylation analysis is described by the bottom side of Figure 1 with three functions.

First is "CGI density analysis" (Figure 4A), which is used to analyze the classification of position-defined CGI under various CGI density (Weber et al., 2007; Zhu et al., 2008) and CpG interval (d).

Second is "Methylation level analysis," which not only can analyze the specificity of methylation level for CpG sites under different annotation categories and CpG interval (d) (Figure 4B), but also allows the visualization and comparative analysis of methylation level of position-defined CGIs at the genome-wide perspective (Figure 4C).

The third is "GO enrichment analysis," which employs clusterProfiler (Yu et al., 2012) to make GO enrichment analysis (Liu et al., 2020) for the CGI + genes (Coding genes that at least one of its TSSs is located in the CGI) (Weber et al., 2007; Zhang et al., 2021a) of position-defined CGIs. Here, Figure 4D shows GO enrichment analysis for the CGI + genes under CpG interval d = 12bp.

3.4 Algorithm performance comparison

Firstly, As shown in Figure 5; Supplementary Figure S2, we compare the computing speed for Human position-defined CGI prediction method with MR-CpGCluster and this method without MR-CpGCluster with three commonly used

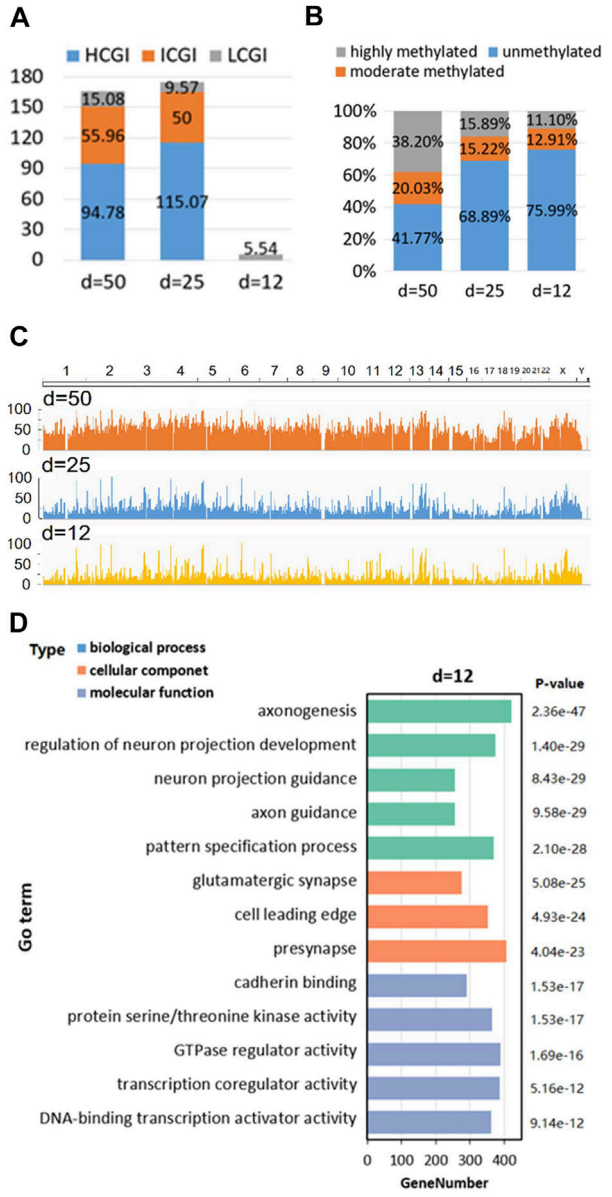

FIGURE 4
Human position-defined CGI methylation analysis. **(A)** The classification of position-defined CGI under CGI density. **(B)** CpG density of different gene and sequence categories. **(C)** Comparison of the methylation ratio of position-defined CGIs. The horizontal and vertical axes represent genomic chromosome position and the methylation rate of the CpG site at corresponding position, respectively. **(D)** GO enrichment analysis.

standards: Speedup, Scaleup and Sizeup (Schatz, 2009). Figure 5 shows that the Speedup is positively related to the number of nodes and the size of dataset. For example, when using 8 nodes for a 3920 MB dataset, the ratio between the actual and ideal Speedup is 6.00/8 = 75%, while with 6 nodes for a 980 MB dataset, this ratio is 2.88/6 = 49.17%.

Next, we compare the computing efficiency for Human position-defined CGI prediction method with commonly used density-defined CGIs prediction method (Weber et al., 2007; Zhang et al., 2021a) and another two classical distance-based CGI

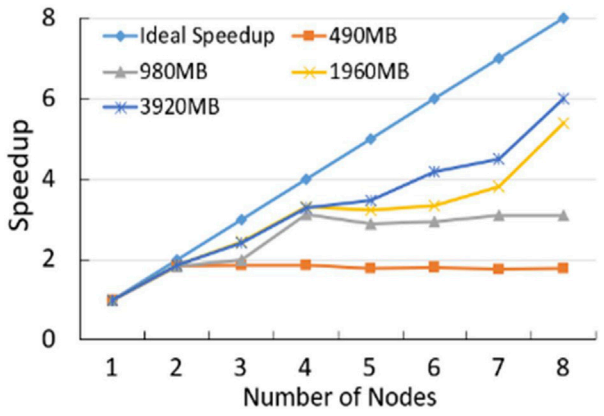

FIGURE 5
The speedup ratio of MR-CpGCluster.

prediction methods such as WordCluster (Hackenberg et al., 2011) and CpGProD (Ponger and Mouchiroud, 2002) by CGI length, GC content, and O/E ratio (Eq. 3), which are three broadly used standards (Wang and Leung, 2004; Hackenberg et al., 2010).

Table 1; Supplementary Figure S5 not only demonstrate that the average length of CGIs of Human position-defined CGI prediction method (23.7 ± 11.5bp) is statistically shorter, but also the average GC content (89.3% ± 7.5%) and O/E value (1.54 ± 0.27) of Human position-defined CGI prediction method are statistically greater than other prediction methods by statistical test (Zhang et al., 2021b; Zhang et al., 2021d; Gao et al., 2021; Liu et al., 2021; Lai et al., 2022; Song et al., 2022).

Note: Here, we employ default setup for CpG interval, d = 12bp (Zhang et al., 2018; Zhang et al., 2021a).

3.5 Web service construction

Figure 6 shows the technical architecture of PCGIMA (http://www.combio-lezhang.online/pcgima/home.html), which consists of three modules: "Human position-defined CGI prediction," "CpG sites annotation analysis," and "CGI methylation analysis."

PCGIMA employs MR-CpGCluster to predict the position-defined CGI for multiple consecutive (d) values. To compare and analyze the CpG methylation levels in different genome regions, we integrate the JavaScript version of IGV (Integrative Genomics Viewer) (Thorvaldsdottir et al., 2013) into our Web service. PCGIMA also imports the genome annotation information and analysis results into the MySQL database (Xia et al., 2010) and use eCharts (Bond and Goguen, 2002) to visualize CGI-related analysis results.

"Human position-defined CGI prediction" module provides two functions (Figure 2). One is "Position-defined CGI prediction," which can online predict position-defined CGI for the human genome or a particular chromosome with multiple consecutive (d) values. The other is "Position-

10.3389/fgene.2024.1367731

TABLE 1 CGI prediction methods comparison.

CGI prediction methods	CGI number	Average length ± standard deviation	Average GC ± standard deviation	Average O/E ± standard deviation	Average CpG Density ± standard deviation
Human position-defined CGI prediction method	89,063	23.7 ± 11.5	89.3% ± 7.5%	1.54 ± 0.27	0.294 ± 0.066
CpGCluster	198,445	274.7 ± 249.8	63.8% ± 7.6%	0.86 ± 0.27	0.087 ± 0.04
WordCluster	198,703	273.2 ± 246.4	63.8% ± 7.5%	0.86 ± 0.27	0.087 ± 0.04
CpGProD	76,793	1,043.8 ± 761.7	54.6% ± 6.1	0.64 ± 0.1	0.047 ± 0.016
Density-defined CGIs	30,477	774.7 ± 826.9	66.5% ± 4.7%	0.86 ± 0.14	0.094 ± 0.018

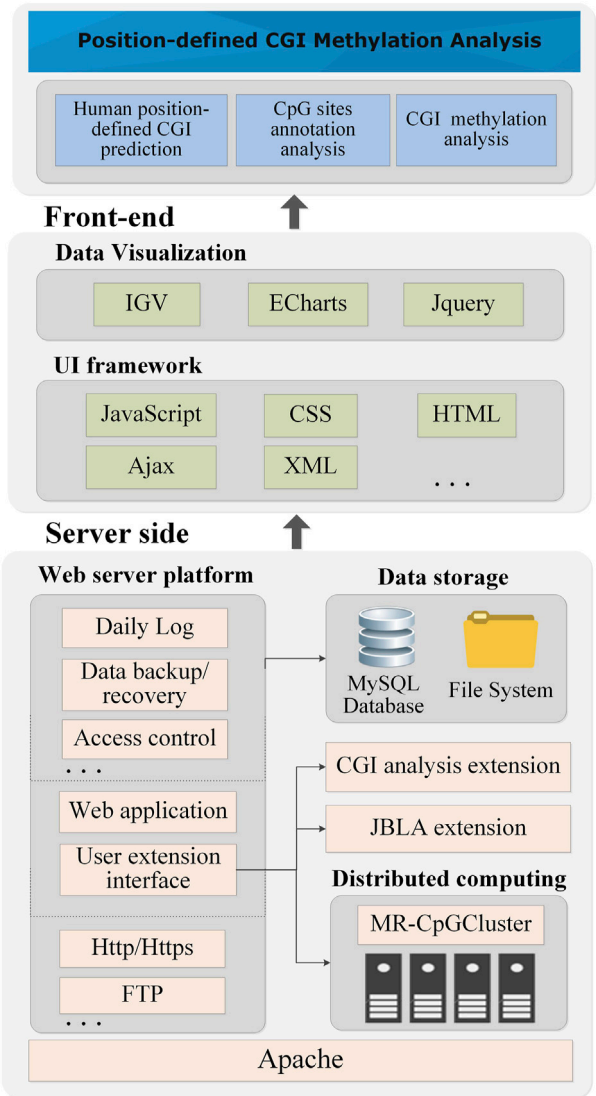

FIGURE 6
The technical architecture of PCGIMA.

defined CGI features analysis," which can describe the connection between the proportion distribution of CGI and CGI features.

"CpG sites annotation analysis" module consists of two functions. First is "Human CpG sites Distribution analysis," which can analyze the distribution of CpG methylation sites in different structural and functional categories of genomic sequences (Figure 3). Second is "Human CpG sites permutation analysis" module (Supplementary Method S4), which can analyze the CpG permutation patterns (Zhang et al., 2018) of density- and position-defined CGIs.

"CGI methylation analysis" module also provides two functions. One is "Position-defined CGI methylation analysis," which can analyze the specificity of methylation level for CpG sites under different annotation categories (Figure 4B). The other is "GO enrichment analysis," which can make GO enrichment analysis for the CGI + genes of position-defined CGIs (Figure 4D). Meanwhile, PCGIMA also provides related source code and data download services. The function descriptions are detailed in Supplementary Method S4.

4 Discussion and conclusion

This study not only develops Human position-defined CGI prediction method to locate short and position-sensitive CGIs (CpG islets) using high performance computing to construct MR-CpGCluster algorithm (Figure 1), but also a novel human genome annotation and analysis method to investigate the connections among CGI, gene expression and methylation. Finally, we integrate them into PCGIMA for online computing and visualization.

For Human position-defined CGI prediction method, it not only can efficiently locate CpG islets (Figure 2A; Table 1), but also it can parallel predict position-defined CGIs with multiple consecutive (d) values and investigate the genetic characteristics of position-defined CGIs under different CpG interval parameters (Figures 2B–D; Supplementary Datas S1–S3).

For annotation method, it can investigate the connections between position-defined CGI methylation and gene expression specificity from a genome-wide perspective by considering functional regions (core promoters and gene bodies) and the distribution of methylation sites of genes for different expression breadth (Figure 3). Our annotation method (Figure 3A) reveals that the distribution proportion of methylation sites in TS genes for short positional-defined CGIs (d = 12) is 9.97%, which is less than that for long positional-defined CGIs (d = 50, 11.46%).

For Human position-defined CGI methylation analysis, not only CGI density analysis (Figure 4A) finds an interesting phenomena that short position-defined CGIs (CpG islets) are closer to LCGI by classifying the position-defined CGI under various CGI density (Weber et al., 2007; Zhu et al., 2008) and CpG interval (d), but also methylation levels analysis demonstrates that the average methylation levels are obviously low for CpG islets from overall scale and genome-wide perspective, respectively (Figures 4B, C) as well as Go enrichment analysis implies that the position-defined CGI-related genes could be associated with unique gene regulatory functions (Figure 4D; Supplementary Figure S4).

For Algorithm performance comparison, Figure 5 turns out that MR-CpGCluster method is faster than classical CpGCluster for the big dataset, which implies Human position-defined CGI prediction method can parallel process the big CGI data.

Moreover, previous studies indicate that CGIs with length less than 200 bp may be functional CpG islets (Hackenberg et al., 2006) and high GC content and O/E values represent enrichment of methylation sites (Gardinergarden and Frommer, 1987; Takai and Jones, 2002). Since Table 1 demonstrates that the average CGI length of the Human position-defined CGI prediction method is much less than 200bp (column 3 of Table 1), and the average GC content and O/E value are statistically greater than other prediction methods (column 4 and 5 of Table 1), we can conclude that Human position-defined CGI prediction method can locate more potential short CGIs with special functions than previous CGI prediction methods (Takai and Jones, 2002; Takahashi et al., 2017).

Lastly, Figure 6 shows that since we utilize the MR-CpGCluster to speed up CGI prediction and incorporate extensive visualization methods to increase user usability, PCGIMA provides an easy-to-use analysis and visualization platform for human CGI prediction and methylation. It should be noted that since the human genome annotation and analysis results have been computed and imported into the database in advance, it is fast (about 2–3 min) for PCGIMA to show the analysis results except the "Human position-defined CGI prediction."

Although our study already made great progress in CGI prediction, annotation, analysis, and visualization, it still needs further improving. Firstly, we should make detail annotations for human position-defined CGIs in terms of functional and structural features. Secondly, we should interrogate the lineage-based and function-based subsets for CGIs and their regulatory implications (Blackledge et al., 2013). Finally, we should employ advanced high performance computing technology (Jiang et al., 2015; Zhang et al., 2021c; Xiao et al., 2021) to improve PCGIMA in the distant future.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding author.

Ethics statement

Ethical approval was not required for the study involving humans in accordance with the local legislation and institutional requirements. Written informed consent to participate in this study was not required from the participants or the participants' legal guardians/next of kin in accordance with the national legislation and the institutional requirements.

Author contributions

MX: Conceptualization, Data curation, Formal Analysis, Funding acquisition, Investigation, Methodology, Project administration, Software, Supervision, Visualization, Writing–original draft, Writing–review and editing. YX: Writing–original draft, Writing–review and editing, Formal Analysis, Visualization, Data curation, Methodology, Resources. JY: Writing–original draft, Writing–review and editing, Conceptualization, Investigation, Methodology, Project administration, Supervision. LZ: Writing–original draft, Writing–review and editing, Formal Analysis, Funding acquisition, Investigation, Methodology, Supervision, Validation.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This work was supported by grants from National Science and Technology Major Project (2021YFF1201200, China), National Science and Technology Major Project (2018ZX10201002, China), National Natural Science Foundation of China (Grant No. 62372316, China), China Postdoctoral Science Foundation (2020M673221, China), Fundamental Research Funds for the Central Universities (2020SCU12056, China), Sichuan Science and Technology Program (2022YFS0048), Chongqing Technology Innovation and Application Development Project (CSTB2022TIAD-KPX0067).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2024.1367731/full#supplementary-material

References

Acton, R., Yuan, W., Gao, F., Xia, Y., Bourne, E., Wozniak, E., et al. (2021). The genomic loci of specific human tRNA genes exhibit ageing-related DNA hypermethylation. *Nat. Commun.* 12, 2655. doi:10.1038/s41467-021-22639-6

Blackledge, N. P., Thomson, J. P., and Skene, P. J. (2013). CpG island chromatin is shaped by recruitment of ZF-CxxC proteins. *Cold Spring Harb. Perspect. Biol.* 5 (11), a018648. doi:10.1101/cshperspect.a018648

Bock, C., Walter, J., Paulsen, M., and Lengauer, T. (2007). CpG island mapping by epigenome prediction. *PLOS Comput. Biol.* 3 (6), e110. doi:10.1371/journal.pcbi.0030110

Bond, G. W., and Goguen, H. (2002). "ECharts: balancing design and implementation", in: Proceedings of the 6 th IASTED International Conference on Software Engineering and Applications, 149–155.

Casper, J., Zweig, A. S., Villarreal, C., Tyner, C., Speir, M. L., Rosenbloom, K. R., et al. (2018). The UCSC Genome Browser database: 2018 update. *Nucleic Acids Res.* 46, D762–D769. Database issue). doi:10.1093/nar/gkx1020

Clark, K., Karschmizrachi, I., Lipman, D. J., Ostell, J., and Sayers, E. W. (2016). *GenBank. Nucleic Acids Res.* 44, D67–D72. Database issue). doi:10.1093/nar/gkv1276

Dede, E., Sendir, B., Kuzlu, P., Weachock, J., Govindaraju, M., and Ramakrishnan, L. (2016). Processing cassandra datasets with hadoop-streaming based approaches. *IEEE Trans. Serv. Comput.* 9 (1), 46–58. doi:10.1109/tsc.2015.2444838

Di, l., Linna, Z., Zhaoyang, W., Xu, Z., Xiuzhao, F., Yong, L., et al. (2018). EWASdb: epigenome-wide association study database. *Nucleic Acids Res.* D1, D1. doi:10.1093/nar/gky942

Dittrich, J., and Quiané-Ruiz, J. A. (2012). Efficient big data processing in Hadoop MapReduce. *Proc. Vldb Endow.* 5 (12), 2014–2015. doi:10.14778/2367502.2367562

Dor, Y., and Cedar, H. (2018). Principles of DNA methylation and their implications for biology and medicine. *Lancet* 392 (10149), 777–786. doi:10.1016/s0140-6736(18)31268-6

El-Maarri, O., Olek, A., Balaban, B., Montag, M., van der Ven, H., Urman, B., et al. (2019). Methylation levels at selected CpG sites in the factor VIII and FGFR3 genes, in mature female and male germ cells: implications for male-driven evolution. *Am. J. Hum. Genet.* 63 (4), 1001–1008. doi:10.1086/302065

Gao, J., Liu, P., Liu, G. D., and Zhang, L. (2021). Robust needle localization and enhancement algorithm for ultrasound by deep learning and beam steering methods. *J. Comput. Sci. Technol.* 36 (2), 334–346. doi:10.1007/s11390-021-0861-7

Gardinergarden, M., and Frommer, M. (1987). CpG islands in vertebrate genomes. *J. Mol. Biol.* 196 (2), 261–282. doi:10.1016/0022-2836(87)90689-9

Hackenberg, M., Barturen, G., Carpena, P., Luque-Escamilla, P., Previti, C., and Oliver, J. (2010). Prediction of CpG-island function: CpG clustering vs sliding-window methods. *BMC genomics* 11, 327. doi:10.1186/1471-2164-11-327

Hackenberg, M., Carpena, P., Bernaola-Galván, P., Barturen, G., Alganza, Á. M., and Oliver, J. L. (2011). WordCluster: detecting clusters of DNA words and genomic elements. *Algorithms Mol. Biol.* 6 (1), 2. doi:10.1186/1748-7188-6-2

Hackenberg, M., Previti, C., Luqueescamilla, P. L., Carpena, P., Martínezaroza, J., and Oliver, J. L. (2006). CpGcluster: a distance-based algorithm for CpG-island detection. *BMC Bioinforma.* 7 (1), 446. doi:10.1186/1471-2105-7-446

Han, L., Su, B., Li, W. H., and Zhao, Z. (2008). CpG island density and its correlations with genomic features in mammalian genomes. *Genome Biol.* 9 (5), R79–R12. doi:10.1186/gb-2008-9-5-r79

Harrow, J., Denoeud, F., Frankish, A., Reymond, A., Chen, C. K., Chrast, J., et al. (2006). GENCODE: producing a reference annotation for ENCODE. *Genome Biol.* 7 (Suppl. 1), 1–9. doi:10.1186/gb-2006-7-s1-s4

Jiang, B., Dai, W., Khaliq, A., Zhou, X., and Zhang, L. (2015). Novel 3D GPU based numerical parallel diffusion algorithms in cylindrical coordinates for health care simulation. *Math. Comput. Simul.* 109, 1–19. doi:10.1016/j.matcom.2014.07.003

Lai, X., Zhou, J., Wessely, A., Heppt, M., Maier, A., Berking, C., et al. (2022). A disease network-based deep learning approach for characterizing melanoma. *Int. J. Cancer* 150 (6), 1029–1044. doi:10.1002/ijc.33860

Liu, B., Du, Q., Chen, L., Fu, G., Li, S., Fu, L., et al. (2016). CpG methylation patterns of human mitochondrial DNA. *Sci. Rep.* 6 (1), 23421. doi:10.1038/srep23421

Liu, G. D., Li, Y. C., Zhang, W., and Zhang, L. (2020). A brief review of artificial intelligence applications and algorithms for psychiatric disorders. *Engineering* 6 (4), 462–467. doi:10.1016/j.eng.2019.06.008

Liu, S., You, Y., Tong, Z., and Zhang, L. (2021). Developing an embedding, koopman and autoencoder technologies-based multi-omics time series predictive model (EKATP) for systems biology research. *Front. Genet.* 12, 761629. doi:10.3389/fgene.2021.761629

Ming, X., Zhu, B., and Li, Y. (2021). Mitotic inheritance of DNA methylation: more than just copy and paste. *J Genet Genomics.* 48 (1), 1–13. doi:10.1016/j.jgg.2021.01.006

Ning, Y., Xuan, G., Alexander, Z., and Pan, Y. (2017). GaussianCpG: a Gaussian model for detection of CpG island in human genome sequences. *BMC Genomics* 18 (S4), 392. doi:10.1186/s12864-017-3731-5

Ponger, L. C., and Mouchiroud, D. (2002). CpGProD: identifying CpG islands associated with transcription start sites in large genomic mammalian sequences. *Bioinformatics* 18 (4), 631–633. doi:10.1093/bioinformatics/18.4.631

Pruitt, K. D., Tatiana, T., and Maglott, D. R. (2005). NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.* 33, D501–D504. Database issue. doi:10.1093/nar/gki025

Raney, B. J., Cline, M. S., Rosenbloom, K. R., Dreszer, T. R., Katrina, L., Barber, G. P., et al. (2010). ENCODE whole-genome data in the UCSC genome browser (2011 update). *Nucleic Acids Res.* 39, D871–D875. doi:10.1093/nar/gkq1017

Reik, W. (2007). Stability and flexibility of epigenetic gene regulation in mammalian development. *Nature* 447 (7143), 425–432. doi:10.1038/nature05918

Schatz, M. C. (2009). CloudBurst: highly sensitive read mapping with MapReduce. *Bioinformatics* 25 (11), 1363–1369. doi:10.1093/bioinformatics/btp236

Schneider, V. A., Graves-Lindsay, T., Howe, K., Bouk, N., Chen, H. C., Kitts, P. A., et al. (2017). Evaluation of GRCh38 and *de novo* haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Res.* 27, 849–864. doi:10.1101/gr.213611.116

Smith, Z. D., Chan, M. M., Mikkelsen, T. S., Gu, H., Gnirke, A., Regev, A., et al. (2012). A unique regulatory phase of DNA methylation in the early mammalian embryo. *Nature* 484 (7394), 339–344. doi:10.1038/nature10960

Song, H., Chen, L., Cui, Y., Li, Q., Wang, Q., Fan, J., et al. (2022). Denoising of MR and CT images using cascaded multi-supervision convolutional neural networks with progressive training. *Neurocomputing* 469, 354–365. doi:10.1016/j.neucom.2020.10.118

Su, J., Zhang, Y., Lv, J., Liu, H., Tang, X., Wang, F., et al. (2009). CpG_MI: a novel approach for identifying functional CpG islands in mammalian genomes. *Nucleic acids Res.* 38, e6. doi:10.1093/nar/gkp882

Takahashi, Y., Wu, J., Suzuki, K., Martinez-Redondo, P., Li, M., Liao, H. K., et al. (2017). Integration of CpG-free DNA induces *de novo* methylation of CpG islands in pluripotent stem cells. *Science* 356 (6337), 503–508. doi:10.1126/science.aag3260

Takai, D., and Jones, P. A. (2002). Comprehensive analysis of CpG islands in human chromosomes 21 and 22. *Proc. Natl. Acad. Sci. U. S. A.* 99 (6), 3740–3745. doi:10.1073/pnas.052410099

Thorvaldsdottir, H., Robinson, J. T., and Mesirov, J. P. (2013). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief. Bioinform* 14 (2), 178–192. doi:10.1093/bib/bbs017

Wang, Y., and Leung, F. C. C. (2004). An evaluation of new criteria for CpG islands in the human genome as gene markers. *Bioinformatics* 20 (7), 1170–1177. doi:10.1093/bioinformatics/bth059

Weber, M., Hellmann, I., Stadler, M. B., Ramos, L., Pääbo, S., Rebhan, M., et al. (2007). Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. *Nat. Genet.* 39 (4), 457–466. doi:10.1038/ng1990

Wright, J. C., Mudge, J., Weisser, H., Barzine, M. P., Gonzalez, J. M., Brazma, A., et al. (2016). Improving GENCODE reference gene annotation using a high-stringency proteogenomics workflow. *Nat. Commun.* 7, 11778. doi:10.1038/ncomms11778

Xia, X. Q., McClelland, M., and Wang, Y. (2010). TabSQL: a MySQL tool to facilitate mapping user data to public databases. *BMC Bioinforma.* 11, 342. doi:10.1186/1471-2105-11-342

Xiao, M., Liu, G., Xie, J., Dai, Z., Wei, Z., Ren, Z., et al. (2021). 2019nCoVAS: developing the web service for epidemic transmission prediction, genome analysis, and psychological stress assessment for 2019-nCoV. *IEEE/ACM Trans. Comput. Biol. Bioinform* 18 (4), 1250–1261. doi:10.1109/TCBB.2021.3049617

Xiong, Z., Li, M., Yang, F., Ma, Y., Sang, J., Li, R., et al. (2019). EWAS Data Hub: a resource of DNA methylation array data and metadata. *Nucleic acids Res.* 48, D890–D895. doi:10.1093/nar/gkz840

Yu, G., Wang, L. G., Han, Y., and He, Q. Y. (2012). clusterProfiler: an R Package for comparing biological Themes among gene clusters. *Omics-a J. Integr. Biol.* 16 (5), 284–287. doi:10.1089/omi.2011.0118

Zhang, L., Dai, Z., Yu, J., and Xiao, M. (2021a). CpG-island-based annotation and analysis of human housekeeping genes. *Briefings Bioinforma.* 22 (1), 515–525. doi:10.1093/bib/bbz134

Zhang, L., Liu, G., Kong, M., Li, T., Wu, D., Zhou, X., et al. (2021b). Revealing dynamic regulations and the related key proteins of myeloma-initiating cells by integrating experimental data into a systems biological model. *Bioinformatics* 37 (11), 1554–1561. doi:10.1093/bioinformatics/btz542

Zhang, L., Xiao, M., Zhou, J., and Yu, J. (2018). Lineage-associated underrepresented permutations (LAUPs) of mammalian genomic sequences based on a Jellyfish-based LAUPs analysis application (JBLA). *Bioinformatics* 34 (21), 3624–3630. doi:10.1093/bioinformatics/bty392

Zhang, L., Zhang, L., Guo, Y., Xiao, M., Feng, L., Yang, C., et al. (2021c). MCDB: a comprehensive curated mitotic catastrophe database for retrieval, protein sequence alignment, and target prediction. *Acta Pharm. Sin. B* 11 (10), 3092–3104. doi:10.1016/j.apsb.2021.05.032

Zhang, L., Zhao, J., Bi, H., Yang, X., Zhang, Z., Su, Y., et al. (2021d). Bioinformatic analysis of chromatin organization and biased expression of duplicated genes between two poplars with a common whole-genome duplication. *Hortic. Res.* 8 (1), 62. doi:10.1038/s41438-021-00494-2

Zhu, J., He, F., Song, S., Wang, J., and Yu, J. (2008). How many human genes can be defined as housekeeping with current expression data? *BMC Genomics* 9 (1), 172. doi:10.1186/1471-2164-9-172

Ziller, M. J., Gu, H., Muller, F., Donaghey, J., Tsai, L. T., Kohlbacher, O., et al. (2013). Charting a dynamic DNA methylation landscape of the human genome. *Nature* 500 (7463), 477–481. doi:10.1038/nature12433