

CpGene: a web application for epigenetic signature identification from DNA methylation arrays

Konstantinos Lazaros^{1,2,*}, Souzana Logotheti², Christopher Logothetis³, Vasiliki Tzelepi², Panagiotis Vlamos¹, Aristidis G. Vrahatis¹

¹Bioinformatics and Human Electrophysiology Laboratory, Department of Informatics, Ionian University, Corfu 49100, Greece

²Department of Pathology, School of Medicine, University of Patras, Patras 26504, Greece

³Department of Genitourinary Medical Oncology, The University of Texas MD Anderson Cancer Center, Houston, Texas 77030, United States

*Corresponding author. Department of Informatics, Ionian University, Plateia Tsirigoti 7, Corfu 49100, Greece. E-mail: lakonstant@ionio.gr

Associate Editor: Jianlin Cheng

Abstract

Motivation: DNA methylation (DNAm) is the best studied epigenetic mechanism that plays pivotal role in tissue differentiation and epigenetic disruption has been correlated to diverse disease types (e.g. cancer, metabolic disorders). While various DNAm array platforms have been discovered, data analysis remains a challenging task which often requires in-depth bioinformatic expertise. Here, we developed a user-friendly web-based application for data analysis and visualization that accommodates users ranging from early-career basic/translational researchers to experienced bioinformaticians.

Results: CpGene is a web application for analyzing DNA methylation array data. It supports Illumina 450K, EPIC, and EPICv2 methylation array platforms and processes .idat files with integrated preprocessing, normalization, and quality control. Biomarker discovery is available through either classic differential methylation point analysis or machine learning-based feature selection as well as gene enrichment analysis. Results are summarized with clear visualizations, to aid interpretation. By combining these functions in a unified interface, CpGene streamlines methylation analysis and helps identify CpG sites and genes with biological and clinical relevance.

Availability and implementation: CpGene is openly accessible as a web service through <http://cpgene.duckdns.org:8001/> and its source code is available on <https://github.com/kostaslaazaros/cpgenene>.

1 Introduction

DNA methylation (DNAm) at CpG dinucleotides is an important epigenetic mechanism that regulates gene expression without altering the DNA sequence. It undergoes dynamic changes during development and differentiation, guiding cell identity and, when disrupted DNAm accounts for development and progression of various disease types (Chen et al. 2014, Hop et al. 2022, Logotheti et al. 2023). Beyond this, DNA methylation influences immune function, brain development, and cardiovascular health, with aberrant patterns linked to autoimmune diseases, neurological conditions, and heart disease (Fodder et al. 2023, Desiderio et al. 2024). Altered DNA methylation is also a hallmark of cancer carcinogenesis and cancer progression, where interplays with genomic instability in numerous haematological and solid tumors biomarkers (Chen et al. 2014, Papanicolaou-Sengos and Aldape 2022).

DNA methylation arrays are widely used to study DNAm across the genome, providing important insights into epigenetic regulation of gene expression and cell behavior. Illumina platforms

remain the gold standard, with the 450K array covering about 450 000 CpG sites, the EPIC array expanding this to 850 000, and the latest EPICv2 surpassing 900 000 by adding probes for key cis-regulatory regions such as enhancers and CTCF binding sites (Morris and Beck 2015, Mansell et al. 2019, Noguera-Castells et al. 2023). These arrays rely on conversion of methylated cytosines, which distinguishes them from unmethylated cytosines, enabling precise detection of methylation patterns that often extend across neighboring CpG sites in coordinated “in-phase” regions (Wang et al. 2019).

Although DNA methylation arrays are powerful tools for profiling epigenetic modifications at a large scale, they also present important limitations. From a computational standpoint, datasets derived from such arrays, typically contain significantly more features than samples, a problem known as the Hughes phenomenon or the curse of dimensionality (Mora 2023). This imbalance not only increases the complexity of the analysis but can also lead to biased or misleading results if not carefully addressed (Clark-Boucher et al. 2023).

Received: 10 December 2025. Revised: 1 March 2026. Accepted: 13 March 2026

© The Author(s) 2026. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

1.1 Related work

Several software packages have been developed to support pre-processing, visualization, and differential methylation analysis of DNA methylation array data. RnBeads (Müller et al. 2019) provides a modular and comprehensive framework for preprocessing, normalization, and DMP identification, supporting large-scale analyses across Illumina platforms. ChAMP (Tian et al. 2017) emphasizes flexibility and automation, offering integrated modules for normalization, batch correction, and DMP analysis accompanied by effective visualization tools. ENmix (Xu et al. 2021) focuses on noise reduction and background correction using advanced statistical modeling, which enhances the accuracy of methylation estimates and improves downstream analyses. SeSAmE (Zhou et al. 2018) extends functionality to include preprocessing, DMP analysis, and probe-level bias correction, addressing signal artifacts and improving the reliability of differential methylation detection. Watermelon (Pidsley et al. 2013) offers a streamlined framework for data normalization and quality assessment, providing multiple preprocessing options suited for Illumina arrays. MethyLumi (Davis and Bilke 2010) represents one of the earliest tools for Illumina methylation data, facilitating data import, quality control, and basic analytical tasks.

Although these tools have substantially advanced the computational analysis of DNA methylation, they are primarily implemented as R packages and require programming proficiency. None of them provide a fully integrated, web-based environment for end-to-end analysis, which limits accessibility for researchers without computational experience. In addition, current frameworks lack artificial intelligence-based CpG ranking algorithms, which can be particularly valuable in addressing the growing dimensionality and complexity of DNA methylation datasets. Integrating such methods can facilitate feature prioritization and enhance the identification of biologically meaningful biomarkers.

1.2 The CpGene application

Here, we introduce a web application named CpGene, that provides a user-friendly interface to facilitate the analysis of DNA methylation array data. The tool accommodates Illumina 450K, EPIC, and EPICv2 platforms, handling .idat files with built-in methods for preprocessing, normalization, and quality control. It provides options for biomarker discovery through either differential methylation analysis or machine learning-driven feature ranking, while functional interpretation is also supported through the use of EnrichR's API. Results are delivered through intuitive visualizations that enhance interpretability. By combining these capabilities within a single environment, CpGene offers a practical solution for uncovering CpG sites and genes of biological and clinical importance. Ultimately, the CpGene platform is designed to be accessible and beneficial to researchers at all levels, from early-stage investigators such as undergraduate and master's students, as well as PhD candidates, to senior researchers across a variety of scientific fields—including basic and translational scientists, clinicians, and experienced bioinformaticians. It offers a fast and reliable tool for epigenetic biomarker identification.

2 Materials and methods

2.1 Preprocessing and quality control

Preprocessing and quality control in CpGene are performed using the minfi R package (Aryee et al. 2014). Input consists of .idat files and a sample sheet, which are assembled into a single object for analysis. Low-quality probes are filtered out using detection *P*-values ($>.01$; the default for minfi), which identify signals that cannot be reliably distinguished from background fluorescence. Background noise correction is carried out with NOOB (Triche et al. 2013), followed by normalization steps that use principal component analysis of control probes to adjust for technical variation between samples, known as batch effects. Cross-reactive probes are then removed, and beta values are calculated to represent methylation levels at each CpG site. Finally, CpG sites of hemimethylated beta values (0.31–0.59), which cannot be clearly characterized as hypo- or hypermethylated, are excluded to refine the dataset (Nissenbaum et al. 2013). The outcomes of preprocessing are visualized with a bar plot of detection *P*-values, a dot plot for bisulfite conversion efficiency across red and green channels, a scatter plot for sample signal quality, and a beta value distribution plot after normalization. Users are also able to download the resulting processed files and all QC figures in high-resolution 300 dpi PNG format.

2.2 AI-based CpG ranking

Feature ranking is implemented through AI-based methods built on the scikit-learn Python library (Nelli 2023), offering flexibility depending on the strategy chosen by the user. The application incorporates wrapper-based approaches, embedded techniques, model-agnostic explainability methods, and neural-network-based strategies, ensuring that ranking can be performed under different modeling assumptions. Users may apply these methods either in a binary classification setting, comparing two conditions, or in a multi-class setting, involving more than two conditions. In addition, the interface allows users to interactively select the number of top-ranked features of preference prior to further investigation. The results are visualized through 2D PCA plots, where samples are shown as dots colored according to their target variable, with one plot generated before feature ranking and another after, using only the top *k* features specified by the user.

2.3 Differential methylation analysis

Differentially methylated position (DMP) analysis is supported through integration with the limma R package (Budkina et al. 2023), enabling comparisons between two conditions to identify CpG sites with significant methylation differences. Users can adjust delta beta and *P*-value thresholds directly through the interface to control stringency, and they may further select the number of top features to retain, ensuring that downstream analyses focus on the most relevant signals. The results are visualized with a volcano plot, where hypermethylated CpG sites are highlighted in red, hypomethylated sites in green, and non-significant sites in gray, providing a clear overview of differential

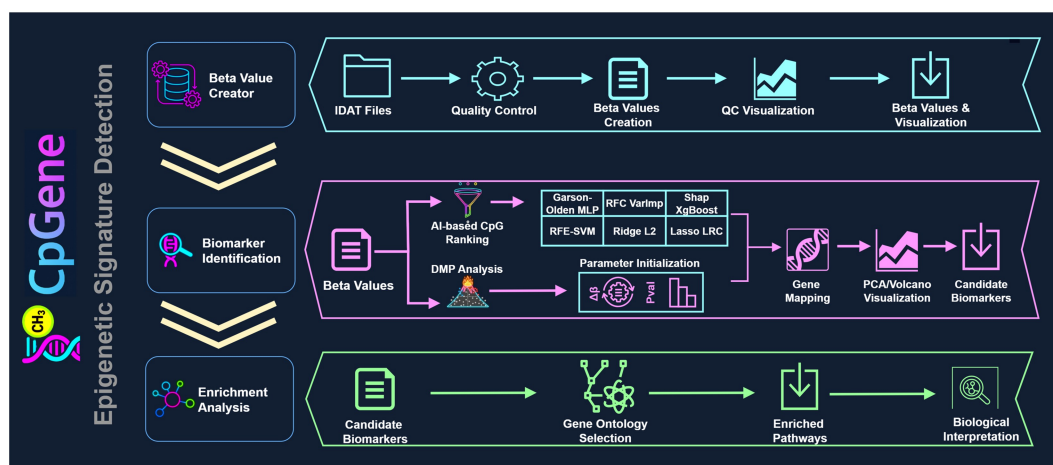


Figure 1 Overview of the CpGene workflow. The pipeline consists of three main stages: Beta Value Creator, Biomarker Identification, and Enrichment Analysis. Starting from Illumina IDAT files, CpGene performs preprocessing and quality control to generate normalized beta values and corresponding QC visualizations. These values are used for biomarker discovery through either machine learning-based feature selection or differential methylation point (DMP) analysis. During feature selection, sample distributions are visualized via principal component analysis (PCA) plots generated both before and after the selection process, whereas in DMP analysis, methylation differences across CpG sites are illustrated using volcano plots. Identified CpG sites are subsequently mapped to their corresponding genes. The final stage conducts pathway and term enrichment through Enrichr-based functional annotation, facilitating biological interpretation and the identification of enriched molecular pathways of potential clinical relevance.

Table 1 Comparison of CpGene with existing DNA methylation array analysis tools.^{a,b}

	CpGene	RnBeads	ChAMP	SeSAME	ENmix	wateRmelon	Methylumi
450K	✓	✓	✓	✓	✓	✓	✓
EPIC	✓	✓	✓	✓	✓	✓	×
EPICv2	✓	✓	×	✓	✓	×	×
Preprocessing/QC	✓	✓	✓	✓	✓	✓	✓
AI-based CpG ranking	✓	×	×	×	×	×	×
Differential methylation analysis	✓	✓	✓	✓	✓	✓	×
Gene enrichment analysis	✓	✓	✓	×	×	×	×
Web application	✓	×	×	×	×	×	×

^a CpGene integrates support for all Illumina array generations (450K, EPIC, EPICv2), preprocessing and quality control, differential methylation analysis, AI-based CpG ranking for biomarker discovery, and gene enrichment analysis within a user-friendly web application. In contrast, existing tools offer subsets of these functionalities, with limited or no support for advanced AI methods and web-based accessibility.

^b ✓ indicates support/availability; × indicates not supported.

methylation patterns (supplementary material). Fig. 1 provides a graphical abstract with the core utilities of the application.

2.4 Gene enrichment analysis

CpG sites selected through either feature ranking or differential methylation analysis are mapped to their corresponding gene symbols based on the array type, using the appropriate manifest file. Gene set enrichment analysis is then carried out through the Enrichr API (Xie et al. 2021), which provides access to a wide range of molecular pathway databases. The results are visualized through the use of a horizontal bar plot which can be sorted either by *P*-value or by combined enrichment score.

3 Conclusion

CpGene is a practical, efficient, and accessible tool for the analysis of DNA methylation array data. Table 1 offers a concise

comparison of CpGene's features with other well-established tools, highlighting its distinctive strengths. By integrating preprocessing, normalization, quality control, feature ranking, differential methylation analysis, and gene set enrichment within a single interface, it provides a systematic environment for uncovering biologically and clinically relevant CpG sites and genes. The interactive design and multiple analysis strategies render CpGene a versatile application that supports robust biomarker discovery and facilitates the exploration of disease-associated epigenetic alterations. In addition, the platform's comprehensive visualization capabilities—ranging from bar plots and scatter plots to volcano plots and enrichment charts—not only aid in the interpretation of complex genomic data but also enhance user engagement and understanding. This comprehensive approach streamlines epigenetic studies by minimizing technical barriers and maximizing analytic flexibility, positioning CpGene as a valuable resource for the scientific community in the ongoing pursuit of actionable epigenetic biomarker identification and insights into disease mechanisms.

Author contribution

Konstantinos Lazaros (Conceptualization [equal], Methodology [equal], Software [lead], Visualization [equal], Writing—original draft [equal], Writing—review & editing [equal]), Souzana Logotheti (Conceptualization [equal], Methodology [equal], Validation [lead], Visualization [equal], Writing—original draft [equal], Writing—review & editing [equal]), Christopher Logotheti (Conceptualization [equal], Validation [equal], Writing—original draft [equal], Writing—review & editing [equal]), Vasiliki Tzelepi (Conceptualization [equal], Funding acquisition [lead], Validation [equal], Writing—original draft [equal], Writing—review & editing [equal]), Panagiotis Vlamos (Conceptualization [equal], Validation [equal], Writing—original draft [equal], Writing—review & editing [equal]), and Aristidis G. Vrahatis (Conceptualization [equal], Methodology [equal], Software [lead], Visualization [equal], Writing—original draft [equal], Writing—review & editing [equal])

Supplementary material

[Supplementary material](#) is available at *Bioinformatics* online.

Conflict of interests

The authors declare no conflicts of interest.

Funding

This work was supported in part by funds from the National Science Foundation and Ronald and Rita McAulay Foundation (RRMF) funding grant [82690 to V.T.].

Data availability

There are no new data associated with this article.

References

- Aryee MJ, Jaffe AE, Corrada-Bravo H *et al.* Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics* 2014; **30**:1363–9.
- Budkina A, Medvedeva YA, Stupnikov A. Assessing the differential methylation analysis quality for microarray and ngs platforms. *Int J Mol Sci* 2023; **24**:8591.
- Chen Q, Zhu X, Li Y *et al.* Epigenetic regulation and cancer. *Oncol Rep* 2014; **31**:523–32.
- Clark-Boucher D, Zhou X, Du J *et al.* Methods for mediation analysis with high-dimensional DNA methylation data: possible choices and comparisons. *PLoS Genet* 2023; **19**:e1011022.
- Davis S, Bilke S. An introduction to the Methylumi package. *Biocond Package* 2010; **10**:B9.
- Desiderio A, Pastorino M, Campitelli M *et al.* DNA methylation in cardiovascular disease and heart failure: novel prediction models? *Clin Epigenet* 2024; **16**:115.
- Fodder K, de Silva R, Warner TT *et al.* The contribution of DNA methylation to the (DYS) function of oligodendroglia in neurodegeneration. *Acta Neuropathol Commun* 2023; **11**:106.
- Hop PJ, Zwamborn RAJ, Hannon E *et al.*; Brain MEND Consortium. Genome-wide study of DNA methylation shows alterations in metabolic, inflammatory, and cholesterol pathways in ALS. *Sci Transl Med* 2022; **14**:eabj0264.
- Logotheti S, Papadaki E, Zolota V *et al.* Lineage plasticity and stemness phenotypes in prostate cancer: harnessing the power of integrated “omics” approaches to explore measurable metrics. *Cancers (Basel)* 2023; **15**:4357.
- Mansell G, Gorrie-Stone TJ, Bao Y *et al.* Guidance for DNA methylation studies: statistical insights from the Illumina epic array. *BMC Genomics* 2019; **20**:366.
- Mora A. Machine learning models of epigenetic dynamics driving cell fate. 2023. <https://espace.library.uq.edu.au/view/UQ:8606709>
- Morris TJ, Beck S. Analysis pipelines and packages for Infinium humanmethylation450 beadchip (450k) data. *Methods* 2015; **72**:3–8. <https://doi.org/10.1016/j.ymeth.2014.08.011>
- Müller F, Scherer M, Assenov Y *et al.* Rnbeads 2.0: comprehensive analysis of DNA methylation data. *Genome Biol* 2019; **20**:55.
- Nelli F. Machine learning with scikit-learn. In: *Python Data Analytics: with Pandas, numpy, and Matplotlib*. Springer; 2023, 259–87.
- Nissenbaum J, Bar-Nur O, Ben-David E *et al.* Global indiscriminate methylation in cell-specific gene promoters following reprogramming into human induced pluripotent stem cells. *Stem Cell Rep* 2013; **1**:509–17.
- Noguera-Castells A, García-Prieto CA, Álvarez-Erriro D *et al.* Validation of the new epic DNA methylation microarray (900k epic v2) for high-throughput profiling of the human DNA methylome. *Epigenetics* 2023; **18**:2185742.
- Papanicolaou-Sengos A, Aldape K. DNA methylation profiling: an emerging paradigm for cancer diagnosis. *Annu Rev Pathol* 2022; **17**:295–321.
- Pidsley R, Y Wong CC, Volta M *et al.* A data-driven approach to preprocessing Illumina 450k methylation array data. *BMC Genomics* 2013; **14**:293.
- Tian Y, Morris TJ, Webster AP *et al.* Champ: updated methylation analysis pipeline for Illumina beadchips. *Bioinformatics* 2017; **33**:3982–4.
- Triche TJ, Weisenberger DJ, Van Den Berg D *et al.* Low-level processing of Illumina Infinium DNA methylation beadarrays. *Nucleic Acids Res* 2013; **41**:e90.
- Wang W, Auer P, Spellman SR *et al.* Epigenomic signatures in myelodysplastic syndrome patients as predictors of donor compatibility and transplant outcome. *Blood* 2019; **134**:4557.
- Xie Z, Bailey A, Kuleshov MV *et al.* Gene set knowledge discovery with enrichr. *Curr Protoc* 2021; **1**:e90.
- Xu Z, Niu L, Taylor JA. The enmix DNA methylation analysis pipeline for Illumina beadchip and comparisons with seven other preprocessing pipelines. *Clin Epigenetics* 2021; **13**:216.
- Zhou W, Triche TJ, Jr, Laird PW *et al.* Sesame: reducing artifactual detection of DNA methylation by Infinium beadchips in genomic deletions. *Nucleic Acids Res* 2018; **46**:e123.