Computational and Structural Biotechnology Journal 21 (2023) 5446–5450

Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

journal homepage: www.elsevier.com/locate/csbj

Software/web server article

eDAVE – Extension of GDC data analysis, visualization, and exploration tools

Jan Bińkowski, Olga Taryma-Leśniak, Katarzyna Ewa Sokolowska, Patrycja Kamila Przybylowicz, Melanie Staszewski, Tomasz Kazimierz Wojdacz *

Independent Clinical Epigenetics Laboratory, Pomeranian Medical University in Szczecin, Unii Lubelskiej 1, Szczecin 70-204, Poland

ARTICLE INFO

Keywords:
Epigenomics
Transcriptomics
Methylation

ABSTRACT

Publicly available repositories such as Genomic Data Commons or Gene Expression Omnibus are a valuable research resource useful for hypothesis driven research as well as validation of the results of new experiments. Frequently however, the use of those opulent resources is challenging because advanced computational skills are required to mine deposited data. To address this challenge, we have developed eDAVE, a user-friendly, web and desktop interface enabling intuitive and robust analysis of almost 12 000 methylomes and transcriptomes from over 200 types of cells and tissues deposited in the Genomic Data Commons repository. The application is implemented in Python, supported for major browsers and available at: https://edave.pum.edu.pl/

1. Introduction

There is no doubt that epigenetic changes, such as aberrations of DNA methylation, and alterations of gene expression associated with those changes, play a key role in the pathology of disease [1]. Moreover, the significance of disease-related methylation changes as biomarkers applicable in personalized medicine is rapidly increasing, as best exemplified by the fact that the newest liquid biopsy early cancer detection tests target mainly methylation biomarkers [2]. With growing interest in a translational aspect of disease specific methylation changes, large amounts of methylomics and transcriptomics data are generated and deposited in publicly available repositories, such as Gene Expression Omnibus (GEO) or Genomic Data Commons (GDC). However, mining of those repositories requires advanced computational skills because data deposited in them can be unstructured and poorly integrated.

Tools such as GEO2R [3], GEOexplorer [4], TCGAbiolinks [5] or GDC Data Analysis, Visualization, and Exploration Tools (DAVE) [6] have been developed to facilitate integration of data from different repositories but the use of these tools still requires a rather high level of bioinformatics skills or is limited to only certain parts of deposited data.

To address the general limited access to bioinformatics expertise that hampers the use of big data repositories we have developed eDAVE, a platform that enables fast and intuitive analysis of transcriptomics and methylomics data curated according to uniform protocol and deposited in the GDC.

2. Materials and methods

2.1. Data curation

The data transfer between eDAVE and the GDC database utilizes two distinct communication channels: automatic programming interface (API) and GDC downloading tool. The samples list along with the clinical description is extracted with API, however the API-based communication is not sufficiently efficient to transfer large files, therefore, to transfer methylation or expression profiling data, eDAVE uses the GDC data downloading tool. To avoid the communication issues and reduce the time of processing during real-time data analysis, eDAVE is based on a local but regularly updated data repository with methylomics and transcriptomics datasets curated as described in GDC documentation (https://docs.gdc.cancer.gov/Data/Introduction/). The sample groups in the repository are annotated to categories based on clinical information available in GDC, including sample type, primary diagnosis, and tissue or organ of origin. For example, category named "Primary Tumor_Kidney_Papillary adenocarcinoma" refers to profiles generated for primary tumor samples of kidney papillary adenocarcinoma. Due to the limited computational power of our local IT facilities, the web version of eDAVE is currently limited to analyze maximum (n = 50) number of samples per category. These samples are randomly selected from the GDC database using BitGenerator (PCG64) with fixed random seed to ensure reproducibility and representativeness of results.

* Corresponding author.
E-mail address: tomasz.wojdacz@pum.edu.pl (T.K. Wojdacz).

https://doi.org/10.1016/j.csbj.2023.10.057
Received 6 September 2023; Received in revised form 31 October 2023; Accepted 31 October 2023
Available online 4 November 2023
2001-0370/© 2023 Published by Elsevier B.V. on behalf of Research Network of Computational and Structural Biotechnology. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Additionally, we consider categories with less than 5 samples too small to be representative and therefore not include them in the local data repository. The desktop version of eDAVE does not have those limitations and minimum/maximum thresholds can be modified by user (precise instruction of usage of desktop version of eDAVE, is described in documentation deposited in the GitHub repository, see Section 2.9).

2.2. eDAVE modules

The data analysis in eDAVE is based on four modules: differential features explorer, single probe/gene explorer, methylation-expression association explorer and cluster explorer.

2.3. Differential features (DEGs/DMPs) explorer

This module was designed to identify differentially expressed genes (DEGs) or differentially methylated positions (DMPs) between two categories of samples. The module requires user to choose: the data type for the analysis (methylation or expression), categories of samples to compare, statistical significance level (alpha), and minimum effect size for methylation analysis: |delta| (defined as absolute difference of average methylation levels between compared groups) or for the expression analysis: |log2(FC)| (defined as absolute value of $\log_2$ transformed ratio of average expression levels between compared groups). After submission of the request, the application returns a volcano plot describing the effect size metric (change of methylation or expression) versus $\log_2$-transformed FDR-corrected p-value for each compared DEG or DMP (an example is shown in Fig. 1), pie chart showing fraction of significant and non-significant hits and data frame with the results of the analysis for each analyzed feature. The data frame includes: mean methylation/expression value, fold change, $\log_2$-transformed fold change, delta, absolute delta, delta adjusted for variance (Hedges' *g*), type of statistical test, p-value, $\log_{10}$-transformed p-value, FDR-corrected p-value, $\log_{10}$-transformed FDR-corrected p-value as well as information if the feature is DEG/DMP according to the significance level and minimum effect size chosen by the user (exemplary output in Supplementary Table 1).

2.4. Single probe/gene explorer

This module allows the user to compare the methylation levels at a specific CpG site or expression of a gene across multiple categories of samples. The module requires user to select: the data type (methylation or expression) and the name of the analyzed feature (gene name or CpG identifier from the Illumina EPIC or 450 K manifests). It gives options to select: scaling method for the analysis ($\log_2$, ln, $\log_{10}$, zero-mean and unit-variance or none), the type of visualization plot (box, violin or scatter) as well as statistical significance level of the analysis. After submission of the request, visualization of the methylation or expression levels at the analyzed feature is displayed (example shown in Fig. 2), together with the data frame that includes parameters of the statistical analysis (as described in Section 2.3, exemplary output in Supplementary Table 3).

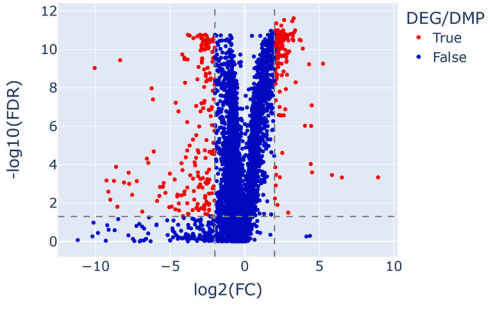

Fig. 1. Volcano plot illustrating results of analysis of gene expression differences between normal breast tissue (n = 50) and breast cancer infiltrating duct carcinoma (n = 50). Dashed horizontal and vertical lines indicate fold-change and FDR-corrected significance thresholds respectively. Dots above these thresholds (red) indicate genes differentially expressed while dots bellow thresholds (blue) indicate genes that do not meet the significance level of differential expression between compared tissues.

2.5. Methylation-expression association explorer

This module computes two types of analysis that describe association between the methylation levels of a specific CpG site and the expression of specified gene in a chosen sample category. The first analysis is regression-based, for which the degree of polynomial transformation can be chosen from default linear regression (degree of 1) to non-linear regression (degree > 1). The second type of analysis is bin-based, where methylation levels are converted to n-number of equally sized bins (from two to four as chosen by the user), then the expression levels of specified gene are compared between these bins. Both analyses require choice of the statistical significance for the comparisons and optionally scaling method. The output of the analysis includes: regression statistics (exemplary output in Supplementary Table 4), fitted model parameters table (exemplary output in Supplementary Table 5), and plot with the regression fit (exemplary output in Fig. 3). In the case of bin-based analysis, the application returns a box plot describing expression levels of the analyzed gene between bins (exemplary output in Fig. 4), the count of samples in each bin, as well as a summary statistics table with parameters of statistical analysis of expression levels differences observed between bins (exemplary output in Supplementary Table 6).

2.6. Cluster explorer

This module was designed to identify subgroups (clusters) of similar samples within datasets of interest and to assess the discriminative power of methylation or expression changes at a subset of CpG sites or genes, respectively. The module requires user to choose data type for the analysis (methylation or expression), paste the list of CpG sites or genes for which the analysis will be performed, and select: samples categories, decomposition algorithm (PCA, t-SNE) as well as the number of dimensions to project data in (two or three). Additionally for the t-SNE method perplexity parameter needs to be selected (perplexity is related to the number of nearest neighbors used in the analysis). The module returns PCA or t-SNE plots that illustrate the results of the cluster analysis colored by the category of sample and by the predicted cluster (exemplary output in Fig. 5). The cluster prediction process that this

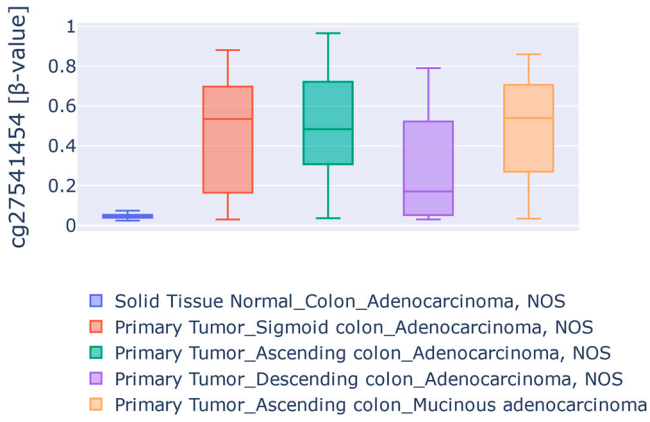

Fig. 2. Box plot illustrating methylation levels at cg27541454 in healthy colon tissue (n = 12) and four types of colon cancer: sigmoid colon adenocarcinoma (n = 50), ascending colon adenocarcinoma (n = 48), descending colon adenocarcinoma (n = 10), and ascending colon mucinous adenocarcinoma (n = 12).

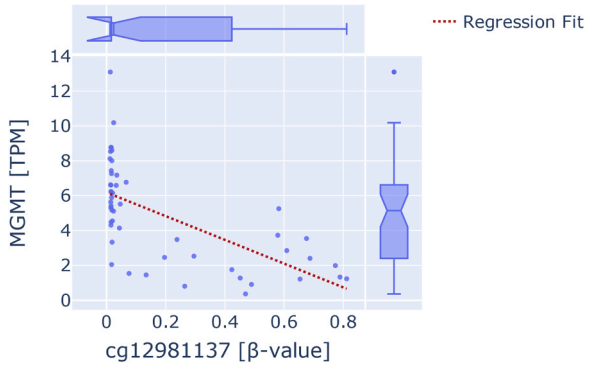

Fig. 3. Scatter plot illustrating association between methylation of cg12981137 and *MGMT* gene expression in glioblastoma samples (n = 50). Regression fit line (red dotted line) represents linear regression model fitted to data. Additionally horizontal and vertical boxplots illustrate distribution of methylation and expression levels, respectively.

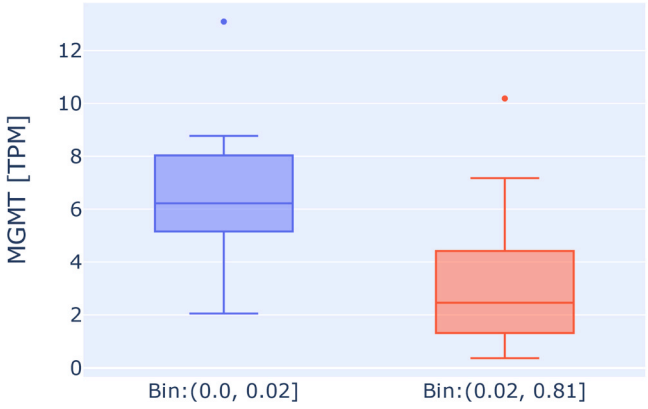

Fig. 4. Box plot illustrating results of bin-based analysis of association between methylation at cg12981137 and *MGMT* gene expression in glioblastoma samples. The blue box shows expression levels in glioblastoma samples (n = 25) in which methylation levels of cg12981137 range from 0.0 to 0.02 and red box displays expression levels in samples (n = 25) with methylation levels of cg12981137 range from 0.02 to 0.81.

module utilizes, consist of five steps. Firstly, the requested dataset is standardized by removing the mean and scaling to unit variance. The standardized data set is projected into two- or three-dimensional planes using PCA or t-SNE techniques. Then, iteratively optimal number of clusters for specific data set is selected (from two to ten) using the Ward clustering method based on geometric distance metric. The effectiveness of the clustering for each tested number of clusters is measured using the Calinski-Harabasz (CH) metric, which by definition is higher for dense and well-separated clusters. Finally, an optimal number of clusters within the dataset is defined as a number-maximizing CH metric.

2.7. DEGs / DMPs identification procedure

The differential features (DEGs/DMPs) explorer, single probe/gene explorer, and the bin-based approach in the methylation-expression explorer share uniform statistical significance assessment strategy. The procedure automatically selects the appropriate statistical test depending on the data distribution of the analyzed variable and minimizes risk of statistical assumptions violation. The procedure of statistical test selection is based on guidelines described in documentation of "pingouin" statistical library [7].

2.8. Results export formats

All plots generated by eDAVE include parameters of statistical analyses of the analyzed features that can be displayed by hovering the cursor over the feature. The plots can be downloaded in high resolution vector format (suitable for publication) by clicking the camera icon present in the upper-right corner of each figure. The tables with the analysis results can be exported in.csv format using "export" function. Additionally, all modules return "sample count table" with information about the number of samples in each analyzed category which can also be exported. All data sets accessible using eDAVE can be downloaded from the data explorer module available from the "More" dropdown menu.

2.9. Implementation

eDAVE was implemented in Python 3.10 using Dash web framework. Python code of application, script used to build local data repository, docker file as well as technical documentation are available from: https://github.com/ClinicalEpigeneticsLaboratory/eDAVE.

3. Results

3.1. Validation of performance of eDAVE

To verify the robustness and usability of eDAVE, we conducted a series of independent experiments and compared generated results with the results reported in previous publications.

3.2. Differential features (DEGs/DMPs) explorer

We first used eDAVE to perform a differential expression analysis between healthy breast tissue and breast infiltrating duct carcinoma and the application identified a set of 369 DEGs (|log2 normalized FC $\geq$ 2 and FDR $\leq$ 0.05, Supplementary Table 1, Fig. 1). We than confirmed using FUMA platform [8] that the expression of identified genes is Breast_Mammary_Tissue specific (Supplementary Figure 1), and those genes are involved in processes that have been associated with breast and soft tissue cancers (Supplementary Table 2).

3.3. Single probe/gene explorer

The hypermethylation of cg27541454 has been shown to be a potential colon cancer biomarker in both liquid and solid biopsy [9]. We used eDAVE to analyze methylation levels at this CpG site in healthy colon and four different colon cancer tissues. The analysis performed using single probe/gene explorer (Fig. 2, Supplementary Table 3) confirmed previously described statistically significant (FDR $\leq$ 0.05) increase of methylation levels in all analyzed colon cancer samples in comparison to healthy colon samples.

3.4. Methylation-expression association explorer

It has been shown that *MGMT* gene expression is regulated by methylation and the methylation of that gene is used to guide glioblastoma therapy [10,11]. We tested the association between *MGMT* promoter methylation and expression in glioblastoma samples accessible via eDAVE and confirmed a statistically significant negative association between methylation levels of cg12981137 located in the *MGMT* promoter and *MGMT* expression (Figs. 3–4, Supplementary Tables 4–6).

3.5. Cluster explorer

We used a set of 100 CpGs (Supplementary Table 7) previously identified to display different methylation levels between colon cancer and healthy colon tissue [12] to test the accuracy of the cluster explorer.

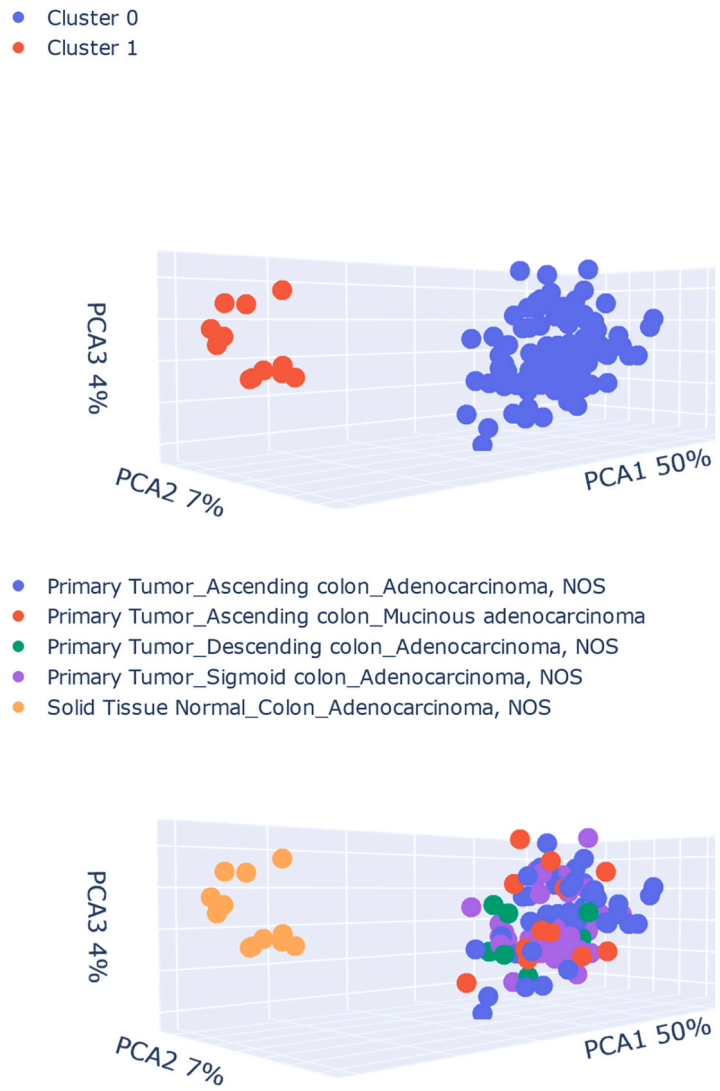

Fig. 5. Scatter plot of 3-D PCA generated using eDAVE's cluster explored module and based on 100 DMPs that previously has been shown to display different methylation levels in healthy colon and colon cancer tissues. Upper plot: results of unsupervised analysis showing that cluster explorer correctly identifies two homogenous and well separated categories of samples (cluster 0 n = 120 and cluster 1 n = 12). Bottom plot: visualization of the samples from the upper plot but samples are colored here by tissue type, including healthy colon tissue (n = 12), sigmoid colon adenocarcinoma (n = 50), ascending colon adenocarcinoma (n = 48), descending colon adenocarcinoma (n = 10), and ascending colon mucinous adenocarcinoma (n = 12).

The PCA three-dimensional visualization based on those probes correctly identified two distinct clusters consisting of cancer and normal samples (Fig. 5).

4. Discussion

We have developed an intuitive and robust tool for the analysis of methylation and gene expression data deposited in the GDC repository. The tool allows researchers without programming expertise easy access to the database resources comprised of large amounts of methylomics and transcriptomic data sets from various types of cancer and healthy tissues. The tests of eDAVE's performance have shown its utility in both hypothesis-driven and biomarker validation research.

Funding

This study was funded by Polish Returns grant program from Polish National Agency for Academic Exchange, grant ID: PPN/PPO/2018/1/00088/U and OPUS22 grant from National Science Centre, grant ID: 2021/43/B/NZ2/02979.

Author agreement statement

Authors confirm that the manuscript has been read and approved by all named authors and that there are no other persons who satisfied the criteria for authorship but are not listed. They further confirm that the order of authors listed in the manuscript has been approved by all of the authors.

Declaration of Competing Interest

The authors declare no conflict of interest.

References

[1] Taryma-Lesniak O, Sokolowska KE, Wojdacz TK. Short history of 5-methylcytosine: from discovery to clinical applications. J Clin Pathol 2021;74(11):692–6.
[2] Luo H, Wei W, Ye Z, Zheng J, Xu R-H. Liquid biopsy of methylation biomarkers in cell-free DNA. Trends Mol Med 2021;27(5):482–500.
[3] Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, et al. NCBI GEO: archive for functional genomics data sets–update. Nucleic Acids Res 2013;41:D991–5 (Database issue).

[4] Hunt GP, Grassi L, Henkin R, Smeraldi F, Spargo TP, Kabiljo R, et al. GEOexplorer: a webserver for gene expression analysis and visualisation. Nucleic Acids Res 2022; 50(W1):W367–w74.
[5] Colaprico A, Silva TC, Olsen C, Garofano L, Cava C, Garolini D, et al. TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. Nucleic Acids Res 2016;44(8):e71.
[6] Grossman RL, Heath AP, Ferretti V, Varmus HE, Lowy DR, Kibbe WA, Staudt LM. Toward a shared vision for cancer genomic data. N Engl J Med 2016;375(12): 1109–12.
[7] Vallat R. Pingouin: statistics in Python. J Open Source Softw 2018;3(31):1026.
[8] Watanabe K, Taskesen E, van Bochoven A, Posthuma D. Functional mapping and annotation of genetic associations with FUMA. Nat Commun 2017;8(1):1826.
[9] Fang Q, Yuan Z, Hu H, Zhang W, Wang G, Wang X. Genome-wide discovery of circulating cell-free DNA methylation biomarkers for colorectal cancer detection. Clin Epigenetics 2023;15:1.
[10] Donson AM, Addo-Yobo SO, Handler MH, Gore L, Foreman NK. MGMT promoter methylation correlates with survival benefit and sensitivity to temozolomide in pediatric glioblastoma. Pedia Blood Cancer 2007;48(4):403–7.
[11] Everhard S, Tost J, El Abdalaoui H, Crinière E, Busato F, Marie Y, et al. Identification of regions correlating MGMT promoter methylation and gene expression in glioblastomas. Neuro Oncol 2009;11(4):348–56.
[12] Baharudin R, Ishak M, Muhamad Yusof A, Saidin S, Syafruddin SE, Wan Mohamad Nazarie WF, et al. Epigenome-wide DNA methylation profiling in colorectal cancer and normal adjacent colon using infinium human methylation 450K. Diagnostics 2022;12(1):198.