



Assessing Differential Variability of High-Throughput DNA Methylation Data

Hachem Saddiki¹ · Elena Colicino¹ · Corina Lesseur¹

Accepted: 18 July 2022 / Published online: 30 August 2022
© The Author(s), under exclusive licence to Springer Nature Switzerland AG 2022

Abstract

Purpose of Review DNA methylation (DNAm) is essential to human development and plays an important role as a biomarker due to its susceptibility to environmental exposures. This article reviews the current state of statistical methods developed for differential variability analysis focusing on DNAm data.

Recent Findings With the advent of high-throughput technologies allowing for highly reliable and cost-effective measurements of DNAm, many epigenome studies have analyzed DNAm levels to uncover biological mechanisms underlying past environmental exposures and subsequent health outcomes. These studies typically focused on detecting sites or regions which differ in their mean DNAm levels among exposure groups. However, more recent studies highlighted the importance of identifying differentially variable sites or regions as biologically relevant features.

Summary Currently, the analysis of differentially variable DNAm sites has not yet gained widespread adoption in environmental studies; yet, it is important to examine the effects of environmental exposures on inter-individual epigenetic variability. In this article, we describe six of the most widely used statistical approaches for analyzing differential variability of DNAm levels and provide a discussion of their advantages and current limitations.

Keywords DNA methylation · Differential variability · Differential methylation · Variability test · Mean and variance test

Introduction and Background

DNA methylation (DNAm) is one of the most studied epigenetic marks known to play a key role in human health and development [1]. DNAm marks are chemical modifications of DNA which occur predominantly as 5-methylcytosine (5mC) at CpG dinucleotides in mammals. DNAm profiles are relatively stable and inherited during cellular division, but they are susceptible to environmental exposures such as smoking and diet [2]. Once established, DNAm alterations can persist even in the absence of the factors that induced them. Therefore, DNAm has unique properties to serve as a biomarker of past and concurrent environmental exposures

[3]. In addition, DNAm patterns have been shown to predict disease status and to be considered mediators of the effect of environmental factors on human diseases [4].

A common technology to profile DNAm uses microarrays to measure DNAm levels at approximately half to one million CpG sites for each DNA sample. These technologies for high-throughput DNAm measurements are highly reliable with a good trade-off between cost and coverage compared to other epigenetic marks [5]. These platforms have been widely used in epigenome studies in which DNAm levels are analyzed to uncover biological mechanisms that underlie prior environmental factors and subsequent health outcomes. However, most studies analyzing DNAm levels at individual CpG sites or regions of the epigenome have been focused on detecting differentially methylated levels, i.e., sites or regions different in their mean DNAm levels among exposure groups or concentrations. Only a few studies evaluated the variance in DNAm in relation to extrinsic environmental stimuli and health outcomes. Figure 1 illustrates three simulated scenarios of methylation levels at a given CpG site: (1a) difference in methylation mean, (2b) difference in

This article is part of the Topical Collection on *Environmental Epigenetics*

✉ Corina Lesseur
corina.lesseur@mssm.edu

¹ Department of Environmental Medicine and Public Health, Icahn School of Medicine at Mount Sinai, New York, NY, USA

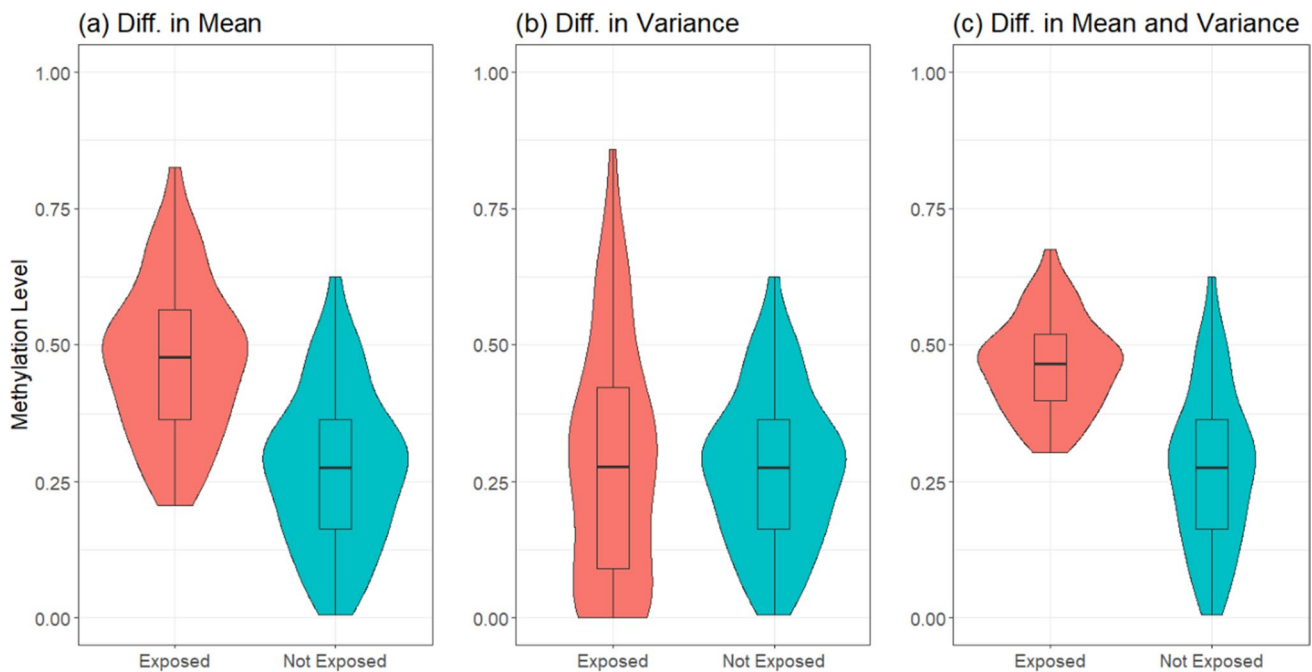


Fig. 1 Violin plots illustrating simulated examples of **a** difference in methylation mean, **b** difference in methylation variance, and **c** difference in both methylation mean and variance

methylation variance, and (3c) difference in both methylation mean and variance.

Recent findings from cancer epigenomic studies highlight the importance of identifying differentially variable CpG sites as biologically relevant features for understanding and predicting phenotypes of interest [6–9]. Importantly, increased DNAm variability in tumors has been suggested to be linked to the environmental adaptive capacity of cancer cells and can predict neoplastic transformation of epithelial cells [7, 10]. The idea is linked to stochastic variation which suggests that certain genetic variants, which do not change the mean phenotype, could alter the variability of the phenotype through epigenetic mechanisms [6]. In fact, epigenetically hypervariable regions throughout the genome can distinguish different tissue types including normal versus cancer; thus, epigenetic variability could be driving both cellular differentiation and Darwinian selection at the tissue which could potentially underlie cancer and other diseases [8]. This leads to the proposition that exposure to environmental risk factors may be driving this epigenetic variation of key genes contributing to disease phenotypes [11, 12].

While analytical methods used to detect differences in mean methylation levels across individual CpG sites and genomic regions are well-established in the literature [13, 14], a small number of approaches have been developed to identify differentially variable methylation. Statistical analysis methods based on differential variability have been shown to improve the detection of risk biomarkers in the

context of cancer genomic and epigenomic studies [7, 12]. The majority of studies examining DNAm variability have been centered in cancer outcomes [6–10]. More recently, some studies have evaluated interpersonal DNAm variability in other outcomes, including body mass index [11], age [15], type 1 diabetes [16], depression in monozygotic twins [17], Alzheimer’s disease [18], and inflammatory bowel disease [19].

Yet, the analysis of differentially variable DNAm sites has not yet gained widespread adoption in environmental studies. To date, variability in DNAm has only been analyzed in the context of tobacco smoking [20], where differential variability between never-smokers and current smokers was compared revealing 14 differentially variable CpG sites associated with smoking exposure, with a 50% (7/14) overlap found between differentially methylated and differentially variable sites; arsenic [21], where 23 differentially variable sites associated with high arsenic exposure through drinking water were identified in leukocytes (PBMCs) and buccal cells; trichloroethylene [22], where blood DNA methylation was compared among high and low trichloroethylene (TCE) exposed workers and a control group, the high and low exposed groups were found to be significantly different in terms of the global DNA methylation variance, with 288 differentially variable sites identified across the three comparison groups after filtering out sites matched with population-specific SNPs; and lead exposure [23], where increased variability in DNA methylation at 16 CpG sites were shown

to be significantly associated with neonatal lead exposure in dried bloodspot samples collected from a cohort of children.

Our goal in this paper is to shed light on the current state of statistical methods developed for differential variability analysis focusing on DNAm data, with the aim of providing a convenient and concise reference to expand their statistical analysis toolkit.

Methods for Differential Variability Analysis

In this section, we describe the most widely used statistical approaches for analyzing differential variability of DNAm levels: *F*-test, Bartlett’s test, Brown-Forsythe, and DiffVar. Next, we describe more recent methods which jointly test differences in mean and variance: penalized Exponential Tilt Model (pETM), and Joint Location and Scale score test (JLSsc). The reader is referred to Table 1 for a concise summary.

The classical *F*-test approach is used to test for equality of variances between two groups (e.g., cases and controls) using the *F*-statistic based on the ratio of variances from each group [7, 25]. This method is sensitive to outliers and departure from normality assumptions on DNAm levels, and covariate adjustment is not straightforward. One way to mitigate sensitivity to outliers is to perform a pre-processing step whereby outliers are removed from the original data set before performing the *F*-test. Other approaches include using variations of the *F*-statistic based on median absolute deviations instead of standard deviation for more robust statistical inference [25].

The Bartlett’s test extends the *F*-test for equality of variances across multiple groups against the alternative

hypothesis that variances are unequal for at least two groups [35]. This method inherits the same shortcomings as the *F*-test in terms of sensitivity to outliers and departure from normality assumptions; and it does not support covariate adjustment.

The Epigenetic Variable Outliers for Risk Prediction Analysis (EVORA) is a popular algorithm using Bartlett’s test as a feature selection step to identify significant CpGs. EVORA assigns each sample a scale independent score determining if it is an outlier with respect to each CpG. Then, a risk score for each sample is calculated based on the proportion of risk CpGs with that particular sample as an outlier [10]. An important EVORA assumption is that sensitivity to outliers in the Bartlett’s test is considered a feature rather than a limitation. The authors argue that variability due to outliers is of biological importance, particularly in the context of pre-cancerous lesions. More recently, a regularized version called iEVORA was developed whereby the initial set of CpG identified via Bartlett’s test are re-ranked according to a specified test statistic based on differences in average methylation [26]. This approach assigns higher importance to more differentially variable CpGs which exhibit significant differences in mean methylation.

The Brown-Forsythe is a Levene’s test variation. Levene’s test is a Bartlett’s test alternative which also tests for equality of variances across multiple groups; however, it has the advantage of being more robust to departures from normality assumptions [36]. Therefore, Levene’s test is preferred when the investigator has strong evidence that the DNAm data are generated from a non-normal distribution. The Brown-Forsythe test uses either the DNAm median or the trimmed median in addition to the mean to calculate the test statistic, as opposed to the original Levene’s test which uses only the

Table 1 Summary table of methods for assessing differential variability

Method	Exposure type	Covariate adjustment	R function	Example Study
<i>F</i> -test	Binary	No	<i>var.test()</i> base R [24]	[25]
Bartlett’s test	Multiple categories/groups	No	<i>bartlett.test()</i> stats package [24]	[10]
EVORA	Multiple categories/groups	No	<i>DoEvora()</i> evora package [12]	[26]
iEVORA	Multiple categories/groups	No	<i>row_ievora()</i> matrixTests package [27]	[28]
Brown-Forsythe test	Multiple categories/groups	No	<i>bf.test()</i> onewaytests package [29]	[30]
DiffVar	Multiple categories/groups	Yes	<i>varFit()</i> missMethyl package [31]	[21]
pETM	Binary	Yes	<i>pETM()</i> pETM package [32]	[33]
JLSsc	Continuous/multiple categories	Yes	<i>jlssc()</i> jlst package [34]	[34]

mean [37]. The Brown-Forsythe test was shown to perform favorably against competing tests in multiple simulation scenarios using real DNA methylation array data, and it is robust to deviations from normality assumptions [38]. However, this test does not support direct covariate adjustment.

The DiffVar method tests for equality of variances across multiple groups using a method inspired by Levene's test statistic. Specifically, DiffVar calculates absolute (or squared) deviations from the DNAm mean in each group. The rationale is that more variable groups will have larger deviations on average, while more consistent groups will have smaller deviations; and testing for equality between the average deviations across groups is equivalent to testing for equality between the variability across the same groups. DiffVar also uses moderated t-statistics instead of ordinary t-statistics to perform multiple comparisons across the large number of CpGs in order to mitigate the issue of false positives [39]. Additionally, the use of linear models within DiffVar allows for the inclusion of adjustment covariates.

The penalized Exponential Tilt Model (pETM) is part of a more recent approach to differential methylation analysis where both the average and variance of the DNAm levels are jointly modeled within one statistical framework. In particular, the penalized exponential tilt model uses network-based regularization which takes into account the correlation among CpG sites within a genomic region to identify differences in both methylation mean and variance between two groups [32]. The approach detects either differences in mean, differences in variance, or differences in both means and variance. Covariate adjustment is possible with this method, and the user can even specify grouping parameters to control the correlation between CpG sites within a genetic region.

The JLSsc test is a novel approach to combine location and scale tests; in other words, it uses the results from running a differential analysis based on mean value (location test) and one based on variability (scale test) and then combines them in a way that accounts for correlations between the results of the two tests [34]. The combination is performed via the residuals from fitting linear regression models, which allows for the inclusion of adjustment covariates and supports continuous or categorical exposures. Another advantage of the JLSsc test is that it allows the user to specify their choice of methods for conducting the difference in mean and variance tests.

Conclusion

The identification of differentially methylated CpG sites and regions associated with environmental factors and health outcomes has been the core of epigenetic analyses in population studies. However, it is also important to examine the

effects of environmental exposures on inter-individual epigenetic variability, as these changes may provide mechanistic insight into adaptive responses of the epigenome to environmental stimuli. Statistically, differences in mean in the absence of a difference in variance means that the distribution is simply shifted in a certain direction, while a difference in variance implies a stretching or change in the shape of the distribution. Thus, novel, adaptable, and robust statistical tools—able to capture differences in DNAm variability—are necessary to adequately infer conclusions and extract biologically sound insight from DNAm information.

In this manuscript, we reviewed six methods (*F*-test, Bartlett's test, Browne-Forsythe's test, DiffVar, pETM, and JLSsc test) which have been used in the context of genomic and epigenomic studies to identify inter-individual biomarker variability. In general, these approaches are easy to implement in R, and most of them, with the exception of pETM, assume normality of the biomarkers, implying the use of *M*-values. Importantly, the majority of these methods lack key features such as covariate adjustments and support for identifying contrasts with continuous exposures. Covariate adjustments are vital to accommodate variables like age, sex, race, ethnicity, and cell type proportions that are known to contribute to inter-individual epigenetic variability. Similarly, continuous exposure measures are quite common in epidemiological studies. Some studies have overcome those issues using residuals of linear regressions after correcting for cell type proportions and categorizing exposures [34, 39]. However, this 2-step approach is more prone to bias and variability than a direct approach, especially when the covariates being adjusted are confounder of the exposure-outcome relationship or if they are not linearly associated with the outcome [40]. Novel methods should aim to overcome these limitations.

Another major challenge in variability analysis is statistical power. Indeed, in order to accurately capture differences in variability, the sample size of the study needs to be larger than that required for analyzing differences in mean methylation levels. This is further exacerbated by the presence of technical artefacts introduced during sample collection, sample handling, and measurement procedures, thus making it harder to disentangle biological and technical variability [41]. Robust methods to account for batch effects such as ComBat are particularly useful to help reduce technical variability [42]; however, one must be careful to check for presence of batch effects before applying such methods to avoid over-correction which could inadvertently impact biological variability.

In the context of DNAm variability, some studies have shown that an association between mean DNA methylation levels and variance could potentially introduce unwarranted confounding bias in the analysis of differential variability [43, 44]. One possible solution to mitigate this

issue is to rely on methods, such as pETM and JLSc, which jointly model the mean and the variance; another solution involves the introduction of an additional measure which corrects for the dependency of the variability on the mean methylation levels [44].

With this paper, we aim to bring more attention to analysis on differential variability, especially in the context of environmental exposures. We designed this work to be a concise and handy reference for investigators wishing to uncover differential variability patterns in their DNAm studies, while highlighting state-of-the-art methods along with the current challenges facing variability analysis studies.

Funding During the preparation of this manuscript, EC was supported by the National Institute of Environmental Health Science (NIEHS): R01ES032242, 5U2CES026555-03, and P30ES023515. CL was supported by the Eunice Kennedy Shriver National Institute of Child Health and Human Development (NICHD): R00HD097286 and NIEHS P30ES023515.

Compliance with Ethical Standards

Conflict of Interest The authors declare no competing interests.

Human and Animal Rights and Informed Consent This article does not contain any studies with human or animal subjects performed by any of the authors.

References

- Baylin SB, Jones PA. A decade of exploring the cancer epigenome - biological and translational implications. *Nat Rev Cancer*. 2011;11(10):726–34.
- Fraga MF, et al. Epigenetic differences arise during the lifetime of monozygotic twins. *Proc Natl Acad Sci U S A*. 2005;102(30):10604–9.
- Nwanaji-Enwerem JC, Colicino E. DNA methylation-based biomarkers of environmental exposures for human population studies. *Curr Environ Health Rep*. 2020;7(2):121–8.
- Petronis A. Epigenetics as a unifying principle in the aetiology of complex traits and diseases. *Nature*. 2010;465(7299):721–7.
- Beck S. Taking the measure of the methylome. *Nat Biotechnol*. 2010;28(10):1026–8.
- Feinberg AP, Irizarry RA. Evolution in health and medicine Sackler colloquium: stochastic epigenetic variation as a driving force of development, evolutionary adaptation, and disease. *Proc Natl Acad Sci U S A*. 2010;107(Suppl 1):1757–64.
- Hansen KD, et al. Increased methylation variation in epigenetic domains across cancer types. *Nat Genet*. 2011;43(8):768–75.
- Issa JP. Epigenetic variation and cellular Darwinism. *Nat Genet*. 2011;43(8):724–6.
- Jaffe AE, et al. Significance analysis and statistical dissection of variably methylated regions. *Biostatistics*. 2012;13(1):166–78.
- Teschendorff AE, et al. Epigenetic variability in cells of normal cytology is associated with the risk of future morphological transformation. *Genome Med*. 2012;4(3):24.
- Feinberg AP, et al. Personalized epigenomic signatures that are stable over time and covary with body mass index. *Sci Transl Med*. 2010;2(49):49ra67.
- Teschendorff AE, Widschwendter M. Differential variability improves the identification of cancer risk markers in DNA methylation studies profiling precursor cancer lesions. *Bioinformatics*. 2012;28(11):1487–94.
- Piao Y, et al. Comprehensive evaluation of differential methylation analysis methods for bisulfite sequencing data. *Int J Environ Res Public Health*. 2021;18(15):7975.
- Teschendorff AE, Relton CL. Statistical and integrative system-level analysis of DNA methylation data. *Nat Rev Genet*. 2018;19(3):129–47.
- Slieker RC, et al. Age-related accrual of methylomic variability is linked to fundamental ageing mechanisms. *Genome Biol*. 2016;17(1):191.
- Paul DS, et al. Increased DNA methylation variability in type 1 diabetes across three immune effector cell types. *Nat Commun*. 2016;7:13555.
- Cordova-Palomera A, et al. Epigenetic outlier profiles in depression: a genome-wide DNA methylation analysis of monozygotic twins. *PLoS ONE*. 2018;13(11):e0207754.
- Huo Z, et al. DNA methylation variability in Alzheimer's disease. *Neurobiol Aging*. 2019;76:35–44.
- Agliata I, et al. The DNA methylome of inflammatory bowel disease (IBD) reflects intrinsic and extrinsic factors in intestinal mucosal cells. *Epigenetics*. 2020;15(10):1068–82.
- Ambatipudi S, et al. Tobacco smoking-associated genome-wide DNA methylation changes in the EPIC study. *Epigenomics*. 2016;8(5):599–618.
- Bozack AK, et al. Exposure to arsenic at different life-stages and DNA methylation meta-analysis in buccal cells and leukocytes. *Environ Health*. 2021;20(1):79.
- Phillips RV, et al. Human exposure to trichloroethylene is associated with increased variability of blood DNA methylation that is enriched in genes and pathways related to autoimmune disease and cancer. *Epigenetics*. 2019;14(11):1112–24.
- Montrose L, et al. Neonatal Lead (Pb) Exposure and DNA methylation profiles in dried bloodspots. *Int J Environ Res Public Health*. 2020;17(18):6775.
- Team R.C. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2013.
- Ho JW, et al. Differential variability analysis of gene expression and its application to human diseases. *Bioinformatics*. 2008;24(13):i390–8.
- Teschendorff AE, Jones A, Widschwendter M. Stochastic epigenetic outliers can define field defects in cancer. *BMC Bioinformatics*. 2016;17:178.
- Teschendorff A, et al. DNA methylation outliers in normal breast tissue identify field defects that are enriched in cancer. *Nat Commun*. 2016;7:10478.
- Webster AP, et al. Increased DNA methylation variability in rheumatoid arthritis-discordant monozygotic twins. *Genome Med*. 2018;10(1):64.
- Dag O, Dolgun A, Konar NM. onwaytests: an R package for one-way tests in independent groups designs. *R J*. 2018;10(1):175–99.
- Yang C, et al. Differentially variable genes of oral squamous cell carcinoma. In: International Conference on Crowd Science and Engineering. Association for Computing Machinery; 2018.
- Phipson B, Maksimovic J, Oshlack A. missMethyl: an R package for analysing methylation data from Illumina's HmanMethylation450 platform. *Bioinformatics*. 2016;15(32):286–8.

32. Sun H, et al. pETM: a penalized Exponential Tilt Model for analysis of correlated high-dimensional DNA methylation data. *Bioinformatics*. 2017;33(12):1765–72.
33. Wang Y, et al. Accounting for differential variability in detecting differentially methylated regions. *Brief Bioinform*. 2019;20(1):47–57.
34. Staley JR, et al. A robust mean and variance test with application to high-dimensional phenotypes. *Eur J Epidemiol*. 2022;37:377–87.
35. Bartlett MS. Properties of sufficiency and statistical tests. *Proc R Soc Lond Ser A*. 1937;160(901):268–82.
36. Levene H. Robust tests for the equality of variances. In: Olkin I, editor. *Contributions to probability and statistics*. Palo Alto: Stanford University Press; 1960.
37. Brown MB, Forsythe AB. Robust tests for the equality of variances. *J Am Stat Assoc*. 1974;69(346):364–7.
38. Li X, et al. A comparative study of tests for homogeneity of variances with application to DNA methylation data. *PLoS ONE*. 2015;10(12): e0145295.
39. Phipson B, Oshlack A. DiffVar: a new method for detecting differential variability with application to methylation in cancer and aging. *Genome Biol*. 2014;15(9):465.
40. Ceyhan E, Goad CL. A comparison of analysis of covariate-adjusted residuals and analysis of covariance. *Commun Stat - Simul Comput*. 2009;38:2019–38.
41. Ecker S, et al. Epigenetic and transcriptional variability shape phenotypic plasticity. *Bioessays*. 2018;40(2):1700148.
42. Chen C, et al. Removing batch effects in analysis of expression microarray data: an evaluation of six batch adjustment methods. *PLoS ONE*. 2011;6(2): e17238.
43. Du P, et al. Comparison of beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinformatics*. 2010;11:587.
44. Ecker S, et al. Genome-wide analysis of differential transcriptional and epigenetic variability across human immune cell types. *Genome Biol*. 2017;18(1):18.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.