BMC Bioinformatics

RESEARCH

Open Access

Impact of influential data on screening epigenome-wide data

Samia Sultana[1], Hongmei Zhang[1], Yu Jiang[1], Mohammad Nahian Ferdous Abrar[1], Hasan Arshad[2,3], Lu Xie[1] and Meredith Ray[1*]

*Correspondence:
Meredith Ray
maray@memphis.edu
[1]Division of Epidemiology, Biostatistics, and Environmental Health, University of Memphis, Memphis, TN, USA
[2]Allergy and Clinical Immunology Department, University of Southampton, Southampton, UK
[3]The David Hide Asthma and Allergy Research Centre, Isle of Wight, UK

Abstract

Background ttScreening (TT) is an effective high-dimensional screening algorithm to identify important cytosine-phosphate-guanine dinucleotide (CpG) sites associated with DNA methylation. Via simulations, we aimed to examine the impact of influential outliers on TT's performance.

Methods We simulated K = 2,000 and 10,000 CpG sites across n = 100 and 200 subjects, linearly associated with a continuous outcome, x_1, and other latent variables with the error term following a normal or Cauchy distribution. Among the K CpGs, 10 were associated with x_1 (informative CpGs) while the remaining sites were not associated with x_1 (non-informative CpGs). We artificially created 1 to 5 influential points in one informative and one non-informative CpG site and compared TT's accuracy to Bonferroni and false discovery rate (FDR)-based approaches.

Results TT performed as well as or better than the FDR and Bonferroni-based approaches, across all degrees of influentiality. When focusing on non-informative CpG detection, regardless of sample size, all approaches had high accuracy (above 85%, overall) at their optimal cutoff for a single influential point. Among the CpG sites with a higher number of influential points (five points) and a normal error term, TT required a minimum cutoff of 70% for accuracy $> 0\%$ compared to FDR and Bonferroni, both of which had an accuracy of 0% for n = 100 and 200. However, increasing the TT cutoff to 80% increased accuracy to 20% and 24%, respectively, and further increased to 97% and 99% for a 90% cutoff, respectively (among K = 2000). We observed the same patterns for 10,000 CpGs and informative CpG detection. When Cauchy error terms were applied, the same patterns held, but with a higher magnitude of accuracy for all approaches, and thus TT required lower cutoffs to achieve 100% accuracy. In summary, in the presence of influential data, we recommend a more conservative cutoff of 70–90% compared to the default cutoff of 50–70% suggested by Ray et al. (in Biomed Res Int 2016(1):2615348, 2016).

Conclusion TT, Bonferroni, and FDR are capable approaches for type 1 protection when screening high-dimensional data. However, in the presence of influential data, TT is likely to be the most robust approach.

Keywords Methylation, ttScreening, FDR and Bonferroni-based methods, Influential data

Background

In epigenetic studies, DNA methylation data's widely known high-dimensionality feature has driven the need to screen cytosine-phosphate-guanine-dinucleotide (CpG) sites [5]. Due to the large number of CpG sites, efforts have been made to improve statistical power to detect the true associations these sites may have with covariates of interest [4, 16], such as single-nucleotide polymorphisms (SNPs) or maternal smoking [3, 7]. One such approach is with the R package, ttScreening (TT) [12, 13], which has been developed to select CpG sites that utilize randomly selected training and testing data and can account for the effects of latent variables derived from surrogate variables. In Ray et al. [13], through intensive simulations, the authors compared TT's performance to that of Bonferroni and False Discovery Rate (FDR)-based multiple-testing correcting methods [2, 6], and demonstrated that TT had the ability to control for both type I and type II error at the same time. The TT screening approach yielded similar sensitivity findings to the FDR results and similar specificity to the Bonferroni results, consequently leading to a generally lower number of incorrectly identified CpG sites. Although the results tend to support the TT approach universally, it does come at a cost; the TT approach requires greater computational time and, therefore, can slow analyses down.

Although the TT approach was built with genetic/epigenetic data as the motivation, the approach can be directly applied to other high-dimensional continuous data. As a caveat with any continuous data, it is possible that extreme outliers, particularly influential points, may exist. Influential observations can result in biased or incomplete findings. In epigenetic association studies, such influential points can cause false positives among identified CpG sites on their association with a variable of interest [18]. The magnitude of such an impact on the performance of the TT process is unknown. In particular, to what extent do influential points and the frequency of influential points affect the performance of TT? To fill this gap, in this study, via simulations, we examined the impact of influential points and their frequency in a data set on the accuracy and efficiency of TT in detecting underlying true associations between DNA methylation at a set of CpG sites and a variable of interest. We compared the results from TT with those based on the commonly used Bonferroni and the FDR-based multiple testing adjustment methods [2, 6].

Methods

In total, we simulated 2,000 and 10,000 CpG sites across 100 and 200 subjects. The first ten CpG sites were selected as informative sites and were simulated as a linear association with a continuous variable, x_1, and five latent covariates (s_1–s_5, surrogate variables), i.e.,

$$y_i = 1.5 + 1.7 \times x_1 + 1.75 \times s_1 + 2 \times s_2 + 1.5 \times s_3 + 1.45 \times s_4 + 1.35 \times s_5 + \epsilon_i, \quad (1)$$

for $i = 1, \ldots, 10$, $x_1 \sim N(1, 1)$, $s_1 \sim N(0, 1)$, $s_2 \sim N(3, 1)$, $s_3 \sim N(0, 4)$, $s_4 \sim N(2, 3)$, $s_5 \sim N(0, 2)$, and $\epsilon_i \sim N(0, 1)$ or $\epsilon_i \sim \text{Cauchy}(0, 0.05)$.

The remaining CpG sites were deemed unimportant or non-informative CpG sites and were simulated as a linear association with the five latent covariates only (s_1–s_5, surrogate variables), i.e.,

$$y_i = 1.5 + 1.75 \times s_1 + 2 \times s_2 + 1.5 \times s_3 + 1.45 \times s_4 + 1.35 \times s_5 + \epsilon_i, \quad (2)$$

for $i = 11, \ldots, 2,000$ or 10,000 and $\epsilon_i \sim N(0,1)$ or $\epsilon_i \sim \text{Cauchy}(0, 0.05)$. Under this structure, the DNA methylation (y_i) was interpreted as a logit-transformed proportion of methylation. We considered the Cauchy distribution, in addition to the normal distribution, for the error terms because of its long-tail properties, which introduced additional variation to the simulated methylation data. It is worth noting that, with the Cauchy distribution, the normality assumption in linear regression is also violated. Thus, this setting also allows us to examine the robustness of TT.

We assessed the impact of influential data by intentionally including $j = 1, \ldots, 5$ influential points in the first informative CpG site (y_1) and in the first non-informative CpG site (y_{11}). To do this, we randomly selected *j* observation(s) and assigned random methylation value(s) from the range:

$$(\max(y_i) + 2\,\text{standard deviations}(y_i) - 5, \max(y_i) + 2\,\text{standard deviations}(y_i) + 5) \quad (3)$$

where $i = 1$ or 11. This forced the influential methylation values to be on the extreme tails of the observed data, but with some variability. Similarly, we assigned these same observations' x_1 value(s) to be a random value in the lower 25th percentile for x_1. Cook's D plots were used to examine and ensure that the points were influential.

In summary, we had 40 simulation settings, 20 settings where we focused on the performance of identifying influential data in a non-informative CpG site, and 20 settings focusing on the performance of identifying influential data in an informative CpG site. Within each of these 20 settings, 10 allowed methylation to be simulated with a normal error term, and the other 10 settings had methylation simulated with a Cauchy error term. Within each set of 10, 5 had a total of 2,000 CpG sites, and the other 5 had 10,000. Each of these 5 had 1 to 5 influential points, respectively, within the designated informative or non-informative CpG site.

In order to assess the TT screening's performance in the presence of influential points, we performed intensive simulations and compared the results to those of Bonferroni and FDR-based methods. To implement TT, several user options must be specified. We elected to use all default settings: the training dataset contained $\frac{2}{3}$ data, and the testing dataset contained $\frac{1}{3}$; 100 iterations, a two-step surrogate variable analysis algorithm, and ordinary least squares regression. As described below, we let the cutoff or the threshold for the minimum frequency across the 100 iterations for a CpG site to be selected as an informative vary from 30% to 90%. Both training and testing significance levels were set to 0.05. The FDR and Bonferroni significance levels were also set to $\alpha = 0.05$. A CpG site was considered informative if the covariate of interest x_1 (defined previously) was statistically significant (if selected at both the training and testing datasets more than the threshold cutoff for TT and after FDR and Bonferroni adjustments).

We generated 100 datasets for each setting. To evaluate the quality of identifying potentially informative CpGs, we calculated the accuracy defined as the number of times the informative or non-informative CpG site with influential points was correctly identified across the 100 datasets. For example, accuracy was the number of times y_1 (informative CpG site with influential data) was correctly identified as a significant CpG site across the 100 datasets. Each of the three approaches had a respective accuracy measure. As mentioned above, we applied the ttScreening package, which applies the TT approach and both FDR and Bonferroni methods. We let the TT cutoff range from 30 to 90% so that we could explore how the default cutoff value may be impacted.

Results

Overall, the ttScreening approach performed as well as or better than the traditional approaches in the presence of influential outliers.

Normal error term

For settings where we simulated the error term with a normal distribution, we see the accuracy of the TT method for correctly not selecting the non-informative CpG site as significant among 2,000 and 10,000 CpG sites compared to the FDR and Bonferroni-based method was generally higher (Fig. 1; Additional Files 1 and 2). Among simulations

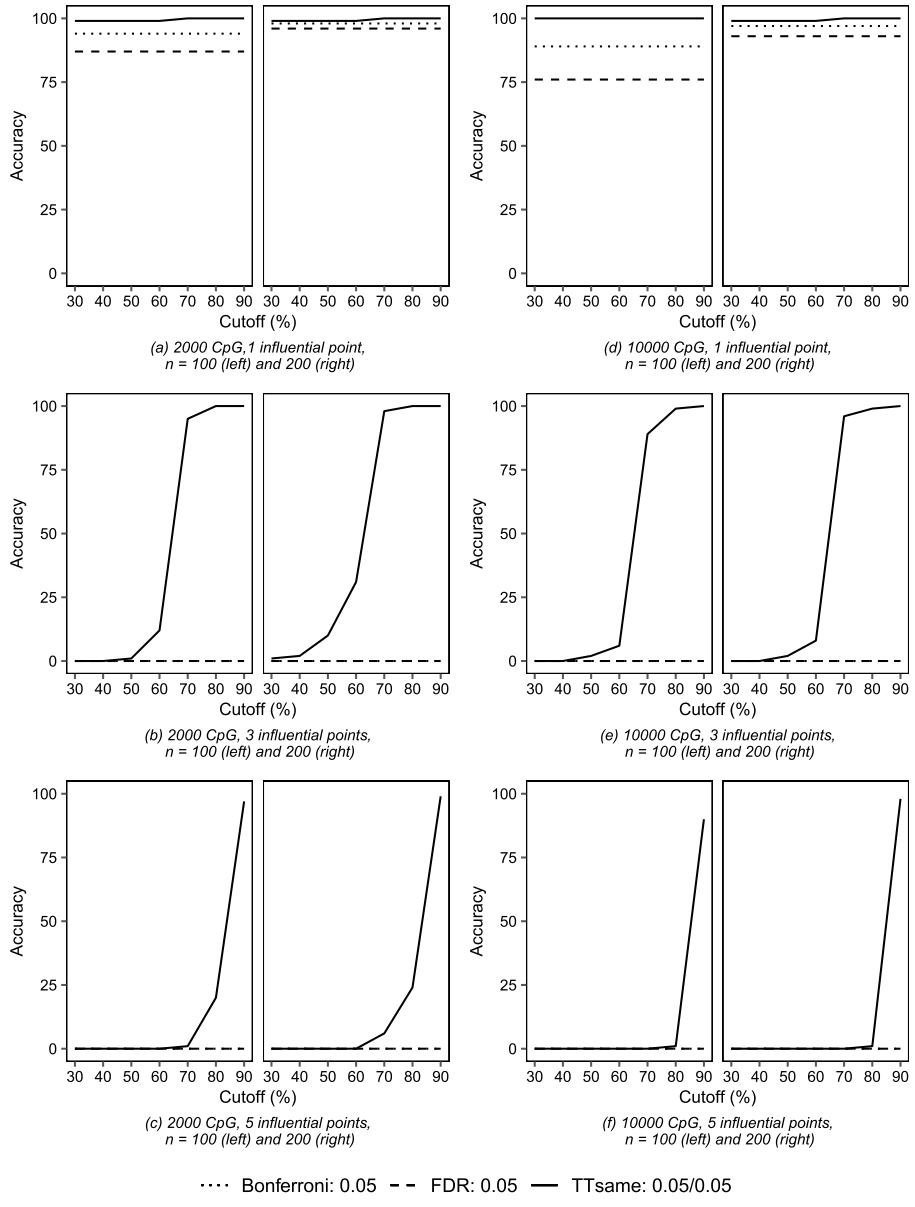

Fig. 1 Accuracy of ttScreening compared to Bonferroni and FDR-based methods with normal error term. Legend: This figure depicts the accuracy comparison of not selecting a non-informative CpG site versus cutoff proportions (from 30% to 90%). The total number of CpG sites considered and the number of influential points incorporated are **a** 2,000, 1; **b** 2,000, 3; **c** 2,000, 5; **d** 10,000, 1; **e** 10,000, 3; and **f** 10,000, 5 across 100 and 200 subjects, respectively. For the TT screening method, the significance levels of the training and testing data were set to 0.05. The FDR and Bonferroni significance levels were also set to $\alpha = 0.05$

for 100 subjects with one artificial influential point in the non-informative CpG site, the TT approach had an accuracy of 99% at a 30% cutoff and 100% accuracy at a 90% cutoff. Meanwhile, the accuracy of FDR and Bonferroni-based methods were 87% and 94%, respectively, for this same setting (Fig. 1, Additional File 1). For 100 subjects and five incorporated influential points in the non-informative CpG site, the accuracy of the TT screening method was 0% at a 30% cutoff, 1% at a 70% cutoff, and 97% at a 90% cutoff. Whereas the accuracy of FDR and Bonferroni-based methods was 0% (Fig. 1; Additional File 1). Similar patterns held when we increased the sample size to 200, but we noticed a slightly lower cutoff, which resulted in similar accuracy. FDR and Bonferroni also had slightly improved accuracy. For example, using the same two settings referenced above, when we simulated 200 subjects and one influential point in the non-informative CpG site, the accuracy of the TT screening method was 99% at a 30% cutoff and 100% at 80% cutoff; the accuracy of FDR and Bonferroni-based methods was 96% and 98%, respectively. When the number of influential points was increased to five (n = 200 subjects), the accuracy of TT screening method was 0% at a 30% cutoff, 6% at a 70% cutoff, and 99% at an 90% cutoff; whereas the accuracy of FDR and Bonferroni-based methods was both 0% (Fig. 1; Additional File 1). The remaining results for 2–4 influential points can be seen in Additional File 1. For our approach, the "optimal" cutoff proportion was determined as the cutoff that produced accuracies (n = 100, 200) higher than the accuracy of the FDR and Bonferroni-based methods. Therefore, for K = 2,000 with error having a normal distribution, we found the optimal cutoff to be 30% for one influential point, 50% for three influential points, and 70% for five influential points. When we increased the number of CpG sites to 10,000, patterns and results were consistent with the 2,000 CpG setting (Additional File 2). In terms of optimal cutoff, for one influential point, it was 30%; for three influential points, it was 50% as above, but it increased to 80% for five influential points. Hence, our overall recommendation for the optimal cutoff is 70–80% for these settings.

To ensure that influential data had minimal impact on informative CpG sites, we assessed the accuracy again across all three approaches for 100 and 200 subjects and 2,000 and 10,000 CpG sites. Accuracy remained at 100% across all settings and approaches; 100% accuracy for TT at all cut points. This data is provided in the supplemental material (Additional Files 3 and 4).

Cauchy error term

When we simulated the error term with a Cauchy distribution, we saw overall higher accuracy rates than with a normal distribution but similar patterns of performance (Fig. 2; Additional Files 5 and 6). Focusing on 2,000 CpG sites across 100 subjects, the accuracy of the TT screening method for not selecting the non-informative CpG site with a single influential point was 100% at all cutoffs and 99% for FDR and 100% for Bonferroni. When we increased up to five influential points, the accuracy for TT was 37% at a 30% cutoff, and 100% at a 90% cutoff, the accuracy for FDR and Bonferroni were 81% and 85%, respectively. Upon increasing our sample size to 200, our accuracy was 100% across the board for all cutoffs and all approaches for a single influential point. When we increased to five influential points, TT accuracy was 91% at a 30% cutoff and 100% at a 70% cutoff, while accuracy for FDR and Bonferroni were 97% and 100%, respecively, (Fig. 2; Additional File 5). Regarding the optimal cutoff (for n = 100 and n = 200), we

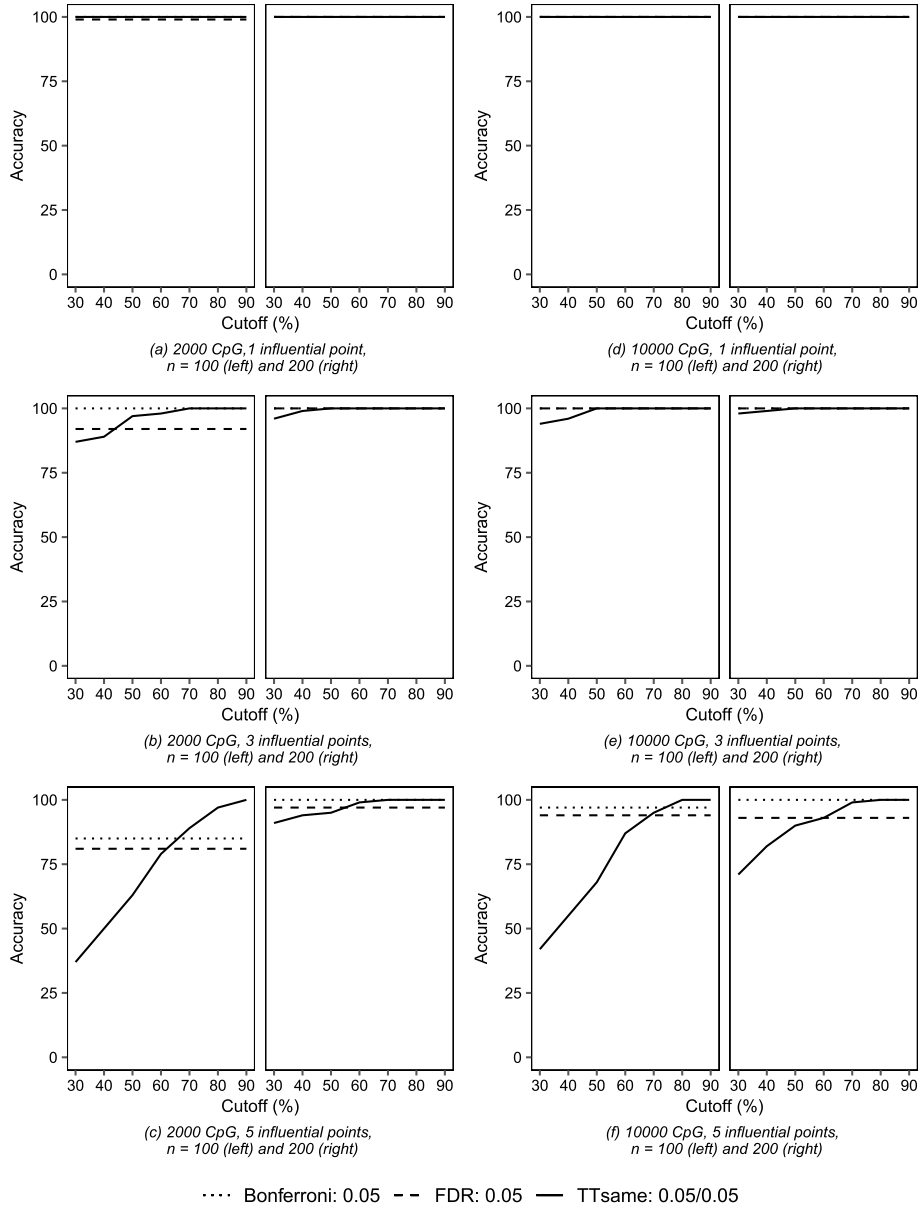

Fig. 2 Accuracy of ttScreening compared to Bonferroni and FDR-based methods with Cauchy error term. Legend: This figure depicts the accuracy comparison of not selecting a non-informative CpG site versus cutoff proportions (from 30% to 90%). The total number of CpG sites considered and the number of influential points incorporated are **a** 2,000, 1; **b** 2,000, 3; **c** 2,000, 5; **d** 10,000, 1; **e** 10,000, 3; and **f** 10,000, 5 across 100 and 200 subjects, respectively. For the TT screening method, the significance levels of the training and testing data were set to 0.05. The FDR and Bonferroni significance levels were also set to $\alpha = 0.05$

found it to be 30% for one influential point, 50% for three influential points, and 70% for five influential points. When we increased our CpG sites to 10,000, all approaches performed exceptionally well at 100% accuracy (for all cutoffs and for n = 100 and n = 200) for 1 and 2 influential points. We began to see a slight dip in accuracy for the lower cutoffs for TT screening with an increasing number of influential points. For 200 subjects, TT screening accuracy was 71% at a 30% cutoff, but 100% at a 80% cutoff, FDR accuracy was 93%, and Bonferroni was 100% (Additional File 6). Similar to the normal error term settings for K = 10,000, the optimal cutoff was 30% for one influential point, 50% for

three influential points, and 80% for five influential points. As such, for these settings, we again recommend 70–80% as the optimal threshold.

Unlike the normal distribution settings, we did see a decrease in the sensitivity of the performance of all approaches when examining the impact of influential data in the informative CpG site (Additional Files 7 and 8). Among the 2,000 CpG sites, as the number of influential points increased, TT required a higher cutoff percentage for greater accuracy. For example, among a sample size of 100 and one influential point, FDR and Bonferroni had accuracies of 25% and 22%, respectively, while the TT had a 6% accuracy at a 30% cutoff and 36% at 60% cutoff, and 65% accuracy at a 90% cutoff. When we increased to five influential points (n = 100), TT accuracy was 0% at 30% to 40% cutoff and 88% at a 90% cutoff, while FDR accuracy was 78% and Bonferroni was 78% (Additional File 7). Similar patterns were observed for n = 200 and K = 10,000 CpG sites; TT screening required a higher cutoff proportion to have as good or better accuracy than FDR and Bonferroni (Additional File 8). Therefore, our recommendation for the optimal cutoff would be 80–90%.

Real data analysis

To demonstrate the proposed technique for the detection of potentially informative CpGs in the presence of influential points, we utilized data collected in the Isle of Wight (IOW) birth cohort established in 1989–1990. The IOW birth cohort aimed to study the natural history of allergic diseases, including asthma [1]. DNA methylation in whole blood was measured at ages 10 and 18 years, representing pre- and post-adolescence, using the Illumina Infinium HumanMethylation450 BeadChip or the Infinium MethylationEPIC (850K) BeadChip. Adolescence's transition into maturity is accompanied by substantial changes, both physically and mentally. It is thus expected that relations between CpGs are likely to be different between pre- and post-adolescence. To this end, we randomly selected 1741 CpGs from those noted in the study by Han et al. [8], and of these 1741 CpGs, 40% showed a statistically significant difference in DNA methylation between ages 10 and 18 years in 366 subjects. To illustrate the concept of epigenetic changes from pre- to post-adolescence and to demonstrate our proposed method, we picked one CpG site, cg00002033, and assessed its association in DNA methylation with the remaining 1740 sites at age 10 and again at age 18. A full list of CpG sites, summary statistics, and the number of potentially influential points determined by Cook's D are available in (Additional File 9) (due to the large size, the summary table was uploaded as a CSV file). Cook's D essentially measures the distance between the estimated regression coefficient with and without a potentially influential point. A commonly applied rule of thumb is that data points with Cook's D $> 4/n$ (or $4/(n-1)$) are treated as influential points, where n denotes the sample size.

We applied the proposed ttScreening technique, using all default values (cutoff of 50%, training/testing significance level set to 0.05, use of surrogate variable estimation) to assess associations at ages 10 and 18 years, respectively. We compared the results from the proposed TT method with the Bonferroni and FDR-based approaches. Upon identifying candidate CpG sites, we performed enrichement analysis to examine the biological pathways these CpGs are involved with using the gometh() function, which mitigates probe-number bias, in the *missMethyl* package

in R. Outcomes with less than two CpGs selected as candidate sites were excluded from the analysis.

At age 10 years, the TT method identified three CpG sites showing an association with cg00002033: cg01571001, cg16277214, and cg26904215. Bonferroni and FDR-based approaches both missed cg16277214 but selected the other two (Additional File 10). At age 18 years, the TT method identified nine CpG sites: cg06639320, cg07110801, cg13391244, cg13392258, cg13593284, cg14041194, cg16732626, cg25150878, and cg26904215. The FDR-based approach identified four (cg06639320, cg13593284, cg25150878, and cg26904215), all of which were among the nine from the TT method. The Bonferroni-based approach was the most conservative and only identified three: cg13593284, cg25150878, and cg26904215 (Additional File 11). Again, all three of these were among the ones identified using the TT and FDR-based approach. Regardless of the approach implemented, DNA methylation at cg26904215 was associated with methylation at cg00002033 at both ages.

Gene and CpG locations and enrichment analysis for candidate sites for ages 10 and 18 are available as supplemental tables in the Online Appendix, (Tables A1–B8). At the age of 10, three CpG sites were found using the training and testing (TT) screening approach, two of which were also detected by both FDR and Bonferroni procedures (Table A1 in Online Appendix). The Gene Ontology (GO) enrichment analysis of these three CpGs identified various immune-related and cell cycle-associated terms (e.g., spindle organization, B cell differentiation, cilium assembly); however, none retained statistical significance following multiple testing correction (all FDR and Bonferroni p-values = 1) (Table B3 in Online Appendix). KEGG pathway analysis also revealed no enriched pathways (all adjusted p-values = 1). For the CpG site (cg16277214) identified by the TT screening method only, no enrichment was performed because the TT-only set contained a single CpG (n = 1) (Table B4 in Online Appendix).

At 18 years of age, nine CpG sites were found by the TT screening method, of which five were also recognized by FDR and three by the Bonferroni method (Table A2 in Online Appendix). The Gene Ontology (GO) enrichment analysis of these nine CpGs revealed multiple terms associated with DNA methylation, epigenetic regulations, and cell-death (e.g., DNA methylation on cytosine within a CG sequence, DNA methylation on cytosine, regulation of phenotypic switching). Nonetheless, none of the GO keywords maintained statistical significance after multiple testing correction (all FDR and Bonferroni p-values = 1) (Table B5 in Online Appendix). KEGGage, nine CpG sites were found by the TT screening method, of which pathway analysis revealed nominal enrichment in pathways including viral carcinogenesis and osteoclast differentiation, with raw p-values of 0.0019 and 0.0006, respectively. Similarly to the GO data, no KEGG pathway retained significance following multiple testing correction (minimum adjusted FDR = 0.204) (Table B6 in Online Appendix). However, when considering the five CpG sites identified by the TT approach only, the Gene Ontology (GO) enrichment study revealed nominal relationships with terms related to ion binding, DNA methylation, and cell death, such as transition metal ion binding, zinc ion binding, DNA methylation on cytosine within a CG sequence, and programmed cell death; none of these items retained statistical significance after correcting for multiple testing (all FDR and Bonferroni

p-values = 1) (Table B7 in Online Appendix). KEGG pathway analysis indicated minimal enrichment in pathways such as mitophagy - animal, cysteine and methionine metabolism, and immune-related signaling, including the RIG-I-like receptor signaling pathway, IL-17 signaling pathway, Hepatitis C, and apoptosis. Analogous to the GO analysis, none of these pathways maintained significance following multiple testing correction (minimum adjusted FDR = 0.5035) (Table B8 in Online Appendix).

Discussion

We aimed to examine TT screening's performance for identifying informative CpG sites in the presence of influential data. We found that TT screening was not only able to identify possible informative CpG sites but also outperformed traditional approaches of FDR and Bonferroni. Our simulations indicate that a higher cutoff, e.g., 70%, is preferred. However, to avoid missing potentially important CpGs, in practice, we would suggest starting from the default, 50%, as the cutoff, and then evaluating the selected sites for possible influential points to evaluate the need of increasing the cutoff.

We applied the TT screening to a set of CpG sites previously studied for examining changes in DNA methylation for adolescent transition. Among 10-year-olds, TT screening identified three important CpG sites, two of which were also selected by FDR and Bonferroni. The CpG site not selected by the traditional approaches, cg16277214, was selected by the TT iterative process 52 times across the 100 iterations. We examined this site to determine if influential data was present, thus possibly requiring a larger cutoff. This site had 18 influential data points (18 Cook's D values greater than 4/n), or about 5%. We elected to both graphically assess the influential data present as well as to rerun the regression (including the surrogate variables) after removing influential points to determine their collective impact. The influential points were uniformly distributed across all values of cg00002033 but occurred at the more extreme values for cg16277214 (Fig. 3). The *p*-value for cg00002033 ($p = 0.0004$) remained significant even after the removal of the 18 influential points ($p = 0.0026$). Thus, we conclude that the 50% cutoff for TT screening for CpG site cg16277214 was appropriate and that this site was missed by traditional approaches. Among 18-year-olds, we repeated the same process. TT identified nine CpG sites, with five (cg07110801, cg13391244, cg13392258, cg14041194, and cg16732626) not selected by FDR or Bonferroni and having selection proportions ranging from 51 to 61%. We again considered the amount, degree, and location of influential data points within each site via graphical assessment and by re-running the regression analysis, excluding potential influential points. We saw similar patterns as we did with CpG site cg16277214 and concluded the 50% cutoff remains appropriate (Additional File 12). These sites were deemed unimportant by FDR and Bonferroni, but TT screening indicates they may warrant further investigation. As part of that further investigation, we examined the biologic pathways of the CpG sites selected at age 10, 18 and those only selected by the TT approach at age 18 (sample size too small, n = 1, to do at age 10). No CpG sites at any age were significant after adjustments. Although insignificant, it is difficult to draw conclusions to the broader population since our sample of CpG sites was randomly selected and candidate sites were selected on their associations with CpG site cg00002033.

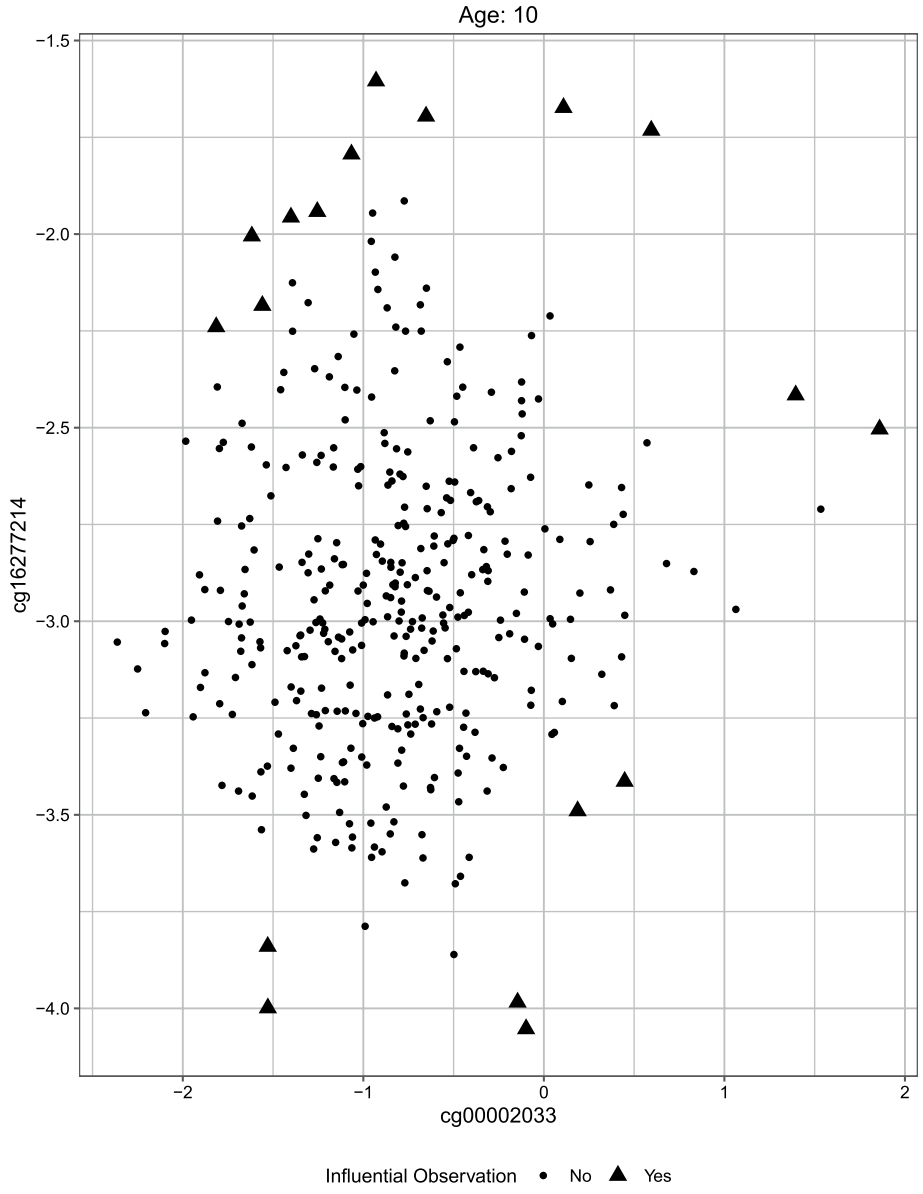

Fig. 3 Influential observations in regression of cg16277214 on cg00002033 at age 10. Legend: This scatterplot shows the association between CpG sites cg16277214 and cg00002033 at age 10. Influential points (Cook's $D > 4/n$) are marked as triangles; all other points are shown as dots

We used simulations assuming the random error follows a normal or a Cauchy distribution. Although, based on our simulation findings, the patterns observed are consistent across different situations, there is a possibility that the recommended 70% cutoff can be revised, and a systematic approach might be necessary. Furthermore, the impact of other regression flaws and biases, such as model misspecification, on the TT approach are not fully known at this time.

Conclusion

The screening approach implemented in the ttScreening R package has the ability to accurately screen high-dimensional data in the presence of influential data. However, it is recommended to increase the cutoff for the TT algorithm when this type of data is present.

Abbreviations

TT	ttScreening
CpGs	Cytosine-phosphate-guanine-dinucleotide site
FDR	False discovery rate

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12859-026-06369-4.

Supplementary file 1 (pdf 292 KB)

Supplementary file 2 (csv 541 KB)

Author Contributions

SS coded and conducted the simulation studies, created and synthesized all tables and figures, primary manuscript preparation; HZ collaborator of methodological theory and application, manuscript preparation; YJ collaborator of methodological theory and application, manuscript preparation; LX collaborator of methodological theory and application, manuscript preparation; MA assisted with the real data analysis and mauscript preparation; MR oversaw and assisted SS in analysis, originator of ttScreening and package author, and senior author in manuscript preparation. All authors read and approved the final manuscript.

Funding

No financial support was provided for this research.

Data Availability

Simulated datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request. The data analyzed in the real data section are from the Isle of Wight birth cohort study (Arshad et al., International Journal of Epidemiology. 2018 A 1; 47(4):1043-4i). Requests to the dataset used in the current study can be sent to Dr. Hongmei Zhang (hzhang6@memphis.edu) and Dr. Hasan Arshad (s.h.arshad@southampton.ac.uk).

Declarations

Competing interests

The authors declare no competing interests.

Ethical approval

Not applicable.

Consent for publication

Not applicable.

Received: 19 February 2025 / Accepted: 2 January 2026

Published online: 06 March 2026

References

1. Arshad SH, Holloway JW, Karmaus W, Zhang H, Ewart S, Mansfield L, et al. Cohort profile: the Isle of Wight whole population birth cohort (IOWBC). Int J Epidemiol. 2018;47(4):1043–1044i.
2. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc: Ser B (Methodol). 1995;57(1):289–300.
3. Cardenas A, Lutz SM, Everson TM, Perron P, Bouchard L, Hivert M-F. Mediation by placental DNA methylation of the association of prenatal maternal smoking and birth weight. Am J Epidemiol. 2019;188(11):1878–86.
4. Chen J, Just AC, Schwartz J, Hou L, Jafari N, Sun Z, et al. CpGFilter: model-based CpG probe filtering with replicates for epigenome-wide association studies. Bioinformatics. 2016;32(3):469–71.
5. Doherty T, Dempster E, Hannon E, Mill J, Poulton R, Corcoran D, et al. A comparison of feature selection methodologies and learning algorithms in the development of a DNA methylation-based telomere length estimator. BMC Bioinform. 2023;24(1):178.
6. Dunn OJ. Estimation of the medians for dependent variables. Ann Math Stat. 1959;30:192–7.
7. Fragou D, Pakkidi E, Aschner M, Samanidou V, Kovatsi L. Smoking and DNA methylation: correlation of methylation with smoking behavior and association with diseases and fetus development following prenatal exposure. Food Chem Toxicol. 2019;129:312–27.

8. Han L, Zhang H, Kaushal A, Rezwan FI, Kadalayil L, Karmaus W, et al. Changes in DNA methylation from pre-to post-adolescence are associated with pubertal exposures. Clin Epigenet. 2019;11:1–14.
9. Kaushal A, Zhang H, Karmaus WJ, Ray M, Torres MA, Smith AK, et al. Comparison of different cell type correction methods for genome-scale epigenetics studies. BMC Bioinform. 2017;18:1–12.
10. Luo R, Zhang H, Mukherjee N, Karmaus W, Patil V, Arshad H, et al. Association of grandmaternal smoking during pregnancy with DNA methylation of grandchildren: the Isle of Wight study. Epigenomics. 2021;13(18):1473–83.
11. Peluso A, Glen R, Ebbels TM. Multiple-testing correction in metabolome-wide association studies. BMC Bioinform. 2021;22:1–18.
12. Ray M, Tong X, Zhang H. R package version 1.6. 2018.https://CRAN.R-project.org/package=ttScreening.
13. Ray MA, Tong X, Lockett GA, Zhang H, Karmaus WJ. An efficient approach to screening epigenome-wide data. Biomed Res Int. 2016;2016(1):2615348.
14. Seeboth A, McCartney DL, Wang Y, Hillary RF, Stevenson AJ, Walker RM, et al. DNA methylation outlier burden, health, and ageing in generation Scotland and the Lothian birth cohorts of 1921 and 1936. Clin Epigenet. 2020;12:1–13.
15. Team RC. RA language and environment for statistical computing, R foundation for statistical. Computing. 2020.
16. Ye Z, Song G, Liang J, Yi S, Gao Y, Jiang H. Optimized screening of DNA methylation sites combined with gene expression analysis to identify diagnostic markers of colorectal cancer. BMC Cancer. 2023;23(1):617.
17. Zhang R-K, Liu J-L. Screening the genome for HCC-specific CpG methylation signatures as biomarkers for diagnosis and prognosis evaluation. BMC Med Genom. 2021;14(1):163.
18. Zhao J, Leng C, Li L, Wang H. High dimensional influence measure. ERN: other econometrics: econometric & statistical methods (topic). 2013. https://api.semanticscholar.org/CorpusID:88513171.

Publisher's Note