EPIGENOMICS
2025, VOL. 17, NO. 6, 389–396
https://doi.org/10.1080/17501911.2025.2479423

Taylor & Francis
Taylor & Francis Group

RESEARCH ARTICLE

Development and validation of a novel cell type estimation method for targeted bisulfite sequencing data

F. Berg[a], E. Köper[a], A. S. Limberg[a], K. Mattonet[a], B. Budeus[b], R. Kumsta[c], E. M. Hummel[a] and D. A. Moser[a]

[a]Department of Genetic Psychology, Faculty of Psychology, Ruhr-University Bochum, Bochum, Germany; [b]Institute of Cell Biology, Medical Faculty, University of Duisburg-Essen, Essen, Germany; [c]Department of Behavioural and Cognitive Sciences, Laboratory for Stress and Gene-Environment Interplay, University of Luxemburg, Esch-sur-Alzette, Luxemburg

ABSTRACT

Aims: Epidemiological studies of DNA methylation often use buccal swabs, which contain mixtures of cell types, but no low-cost methods exist for statistical correction in candidate gene studies using targeted bisulfite sequencing. This study aims to address this gap by estimating the proportion of buccal epithelial cells in swab and mouthwash samples.

Materials & methods: We applied a recently described and smoothly to implement method for estimating the proportion of buccal epithelial cells in buccal swab and mouthwash samples using targeted bisulfite sequencing. Additionally, we investigated the methylation of *actinin alpha 3* (***ACTN3***), a marker for cell type-specific methylation, following psychosocial and physical stress.

Results: The proposed estimation method showed strong correlation with the EpiDISH algorithm and effectively controlled for cellular heterogeneity. Over 90% of the variance in ***ACTN3*** methylation was explained by including the epithelial cell proportion in the model.

Conclusion: Our findings provide a solution for controlling cellular heterogeneity in buccal swab and mouthwash DNA methylation studies. This method is particularly relevant for candidate gene studies in clinical settings, but future work should refine it for more detailed analyses of cell type proportions to improve precision.

ARTICLE HISTORY
Received 10 September 2024
Accepted 10 March 2025

KEYWORDS
DNA methylation; buccal swabs; cell type estimations; cellular heterogeneity; EWAS; targeted bisulfite sequencing; acute stress; epigenetics

1. Introduction

DNA methylation is an epigenetic mark [1] that results from the transfer of a single methyl group (–CH3) to a cytosine residue, converting it to 5-methylcytosine. Methylation typically occurs at cytosines followed by a guanine nucleotide (CpG site) throughout the genome [2]. In a single somatic cell, methylation can be present in both alleles (100% methylation), only in one allele (50%), or absent altogether (0%) at any given CpG site [3]. However, methylation levels are not uniformly distributed across the genome. For instance, regions with high CpG density, known as CpG islands, typically show little to no methylation [4]. In contrast, other regions exhibit cell type-specific methylation patterns, which are the result of normal phylogenetic and ontogenetic development [5,6]. In fact, during ontogenesis, cell differentiation is invariably accompanied by extensive epigenetic reorganization of the entire (epi)genome, including DNA methylation. Regarding somatic tissues, this epigenetic reorganization occurs in a systematic manner, allowing somatic cells to be identified based on their cell type-specific DNA methylation patterns (e.g [2]). While studies on cancer cells benefit from cell type-specific DNA methylation as a measure of comparability (e.g [6]), studies correlating environmental exposures or mental states with DNA methylation levels are often complicated by cellular heterogeneity. This cellular heterogeneity is one of the most significant confounding factors in DNA methylation research [7]. For example, in psychological research, buccal swabs are commonly used to extract DNA [8]. However, buccal swabs contain not only buccal epithelial cells but also leukocytes [9], and these cell types display distinct DNA methylation patterns [10]. Additionally, the ratio of leukocytes to epithelial cells varies between samples due to intrinsic and extrinsic factors [9]. The blood cell composition also fluctuates based on factors such as time of day, menstrual cycle, and stress [11]. Consequently, observed changes in DNA methylation may reflect shifts in cellular composition rather than providing accurate insights into the trait being investigated [8,12]. This issue extends to studies of complex tissues like blood or oral mucosa. Notably, DNA methylation levels in complex tissues that contain a mixture of cell types reflect a weighted average of potentially distinct methylation levels [13]. Therefore, DNA methylation studies must account for this confounding factor. While epigenome-wide association studies (EWAS) and cell isolation methods can straightforward correct for cellular heterogeneity [14], next-generation sequencing techniques, such as targeted bisulfite sequencing, are only able to quantify methylation levels in individual cell types with extensive effort. This is unfortunate, as targeted bisulfite sequencing is particularly useful for

CONTACT F. Berg fabian.berg@rub.de; D. A. Moser dirk.moser@rub.de Department of Genetic Psychology, Faculty of Psychology, Ruhr-University Bochum, Universitätsstraße 150, Bochum 44801, Germany

Article highlights

- In targeted bisulfite sequencing data, two CpGs can be used to estimate the proportion of buccal epithelial cells in buccal swab and mouthwash samples
- Cellular heterogeneity, or the proportion of buccal epithelial cells, significantly impacts the measured *ACTN3* methylation of each sample
- Acute psychosocial and physical stress do not appear to influence the *ACTN3* methylation in the samples

validating epigenome-wide data and conducting candidate gene analyses [15]. To address the current gap in low-cost statistical correction methods, we extended the approach proposed by Eipel et al. [9] to calculate the proportion of buccal epithelial cells and applied it to targeted bisulfite sequencing data. Unlike estimation methods from epigenome-wide studies, Eipel and colleagues [9] enhanced their analysis by incorporating just two differentially methylated CpGs. Adapting their approach to targeted bisulfite sequencing requires only the additional quantification of two amplicons. To illustrate the impact of altered cell composition on targeted bisulfite sequencing data, we selected *actinin alpha 3* (*ACTN3*) for investigation due to its cell type-specific differential methylation pattern [16,17]. DNA methylation of *ACTN3* was analyzed in buccal swab and mouthwash samples within the context of a study on acute psychosocial and physical stress. We hypothesize that measurement errors, as well as stress-related shifts in cellular composition, lead to changes in the composition of buccal swab and mouthwash samples, resulting in alterations in the detected *ACTN3* methylation. In summary, this study had two main objectives. First, we aimed to replicate and extend the buccal epithelial cell estimation from [9] using two specific CpGs. Second, we applied this estimation to targeted bisulfite sequencing data of *ACTN3* to demonstrate how changes in cell composition following stress exposure could potentially confound the results.

2. Materials and methods

2.1. Datasets

Three different datasets were utilized in the analysis:

(1) Independent EWAS data (GSE154566): Adapted from Eipel et al. [9], epigenome-wide data from oral cavity and blood samples were used to estimate the proportion of buccal epithelial cells. We utilized a subsample from the Environmental Risk Longitudinal Twin Study [18]. Participants were 18-year-old twins who provided both buccal swab and blood samples and were healthy controls, with no history of stress-related diseases. Due to the requirement for both buccal swab and blood samples, $n = 37$ participants (40.54% female; 12 complete twin pairs), corresponding to a total of 74 samples, were included. Data were examined using the Illumina Infinium HumanMethylation850 BeadChip.

(2) Lab-internal EWAS data: Data from buccal swab, mouthwash, and blood samples collected after acute psychosocial and physical stress were used to replicate the buccal epithelial cell estimation of the GSE154566 dataset. Acute psychosocial stress was induced using the "Trier Social Stress Test" (TSST [19];), while acute physical stress was induced by a modified version of the "Physical Working Capacity" examination (PWC [20];). A total of 98 samples were collected before the stressor (−2 min), immediately after the stressor (+2 min), and 15 minutes after the stressor (+15 min). Data from $n = 12$ participants (50% female) were used, with four individuals participating in both the acute physical and psychosocial stress paradigms. As in the GSE154566 dataset, DNA methylation was analyzed using the Illumina Infinium HumanMethylation850 BeadChip.

(3) Lab-internal targeted bisulfite sequencing (TBS) data: Forty-six participants (67% female) were tested in the main acute stress study. Participants ($n = 46$) were healthy male ($n = 15$) and female ($n = 31$) volunteers of mostly European ancestry, age 18 to 29 years (mean = 21.8 ± 2.5 (SD)), with a normal body mass index (mean = 22.4 ± 2.3) and did not report regular alcohol intake or smoking behavior. Moreover, participants reported normal weight, no current pregnancy, no rotating shift work, and no stay abroad with time lag in the past 4 weeks, no history of or current mental health problems as well as no chronic or acute physical illnesses, and no current intake of medication. All participants gave written informed consent, and the study was approved by the local ethics committee (759/2022). Testing sessions were offered at least 4 weeks after getting vaccinated or donating blood. Females using oral contraceptive medication were tested at least 3 days after onset of a blister pack and before pausing oral application. To ensure hypothalamic – pituitary – adrenal axis reactivity comparable to the group of men, females who did not take oral contraceptives were tested during the luteal phase between day 17 and the last day of the menstrual cycle [21]. Data from none of these participants were included in the lab-internal EWAS dataset. However, as with the EWAS dataset, psychosocial and physical stress were induced using modified versions of the TSST and PWC, respectively. Sample characteristics are depicted in Table 1 and elsewhere [22]. The study followed a randomized crossover design, providing data from both stress conditions for each participant. Buccal swab, mouthwash, and blood samples were taken before the stressor, immediately after the stressor, and 45 minutes after the stressor (−2, +2, and +45 min, respectively). Blood samples were also collected 15 minutes after each stressor (+15 min). In total, 726 samples were available for statistical analysis of the TBS data.

Table 1. Sample characteristics.

Age, mean (SD)	21.8 (2.54)
BMI, mean (SD)	22.4 (2.34)
Sex, *n* (%)	
Female	15 (33%)
Female using oral contraceptive medication	16 (34%)
Male	15 (33%)
Smoking, *n* (%)	
Non-smoker	36 (80%)
Occasional smoker	9 (20%)

A total of n = 46 participants were included in the TBS acute stress study.

2.2. Sampling procedures

InnuPREP Swabs (Analytik Jena) and mouthwash samples (Listerine, JOHNSON & JOHNSON GmbH) were used to collect oral cavity tissue. Participants were instructed to swipe the tip of the swab along the inner surface of their cheek for 30 seconds. To increase the amount of tissue, the procedure was repeated with a second swab before the swabs were stored at −20°C. Ten ml of mouthwash solution, aliquoted into 50 ml reaction tubes, were swished in the mouth for 30 seconds, expectorated back into the tube, and stored at −20°C until DNA extraction. Blood samples were collected from the finger pulp following the manufacturer's instructions (Microvette® APT 250 EDTA K2E 250 µl, Sarstedt; Nümbrecht, Germany) and were immediately processed in the Genetic Psychology Laboratory as outlined below.

2.3. DNA extraction and targeted bisulfite sequencing

DNA was extracted from buccal swabs, mouthwash, and blood samples following a protocol originally described by Miller et al. [23]. DNA yield and purity were assessed using a Synergy2 plate reader (Biotek, Agilent; USA), and 300 ng of DNA from each sample was bisulfite-treated using the EZ DNA Methylation Gold Kit (Zymo Research, Freiburg, Germany). Targeted bisulfite sequencing [15] was performed on these samples for the candidate genes listed in Table 2. Prepared libraries were sequenced on an MiSeq at the Genomics & Transcriptomics Facility, University Hospital Essen, using a 600bp Flowcell. Samples were subsequently demultiplexed with bcl2fastq and quality-checked using FastQC.

2.4. Bioinformatic and statistical analysis

2.4.1. EWAS data

Illumina Infinium HumanMethylation850 BeadChip files from the Environmental Risk Longitudinal Twin Study were pre-processed and downloaded from the GEO database (GSE154566) in the form of normalized beta-values. Data from our lab, which utilized the same array, underwent standard quality control, preprocessing, and beta-value extraction using RnBeads [24]. Samples with quality issues, cross-reactive probes, probes on sex chromosomes, and probes containing SNPs in the last three bases of the target sequence were excluded.

2.4.2. TBS data

For DNA methylation assessment of each sample, FASTQ files were analyzed using the amplikyzer2 software [25]. Statistical analyses were performed based on the average *ACTN3* methylation provided by amplikyzer2. The methylation of buccal epithelial cell-specific CpGs was extracted from the respective amplicon's amplikyzer2 output. Samples with fewer than 1,000 reads were excluded from the analysis.

2.4.3. Buccal epithelial cell estimation

Following the procedure proposed by Eipel et al. [9], we screened the epigenome-wide for CpGs exhibiting maximum DNA methylation differences between swab and blood samples. Two targets (cg08141395 and cg12389346) were selected based on their suitability for amplification in targeted bisulfite sequencing and their correlation with the EWAS-predicted percentage of epithelial cells. The RPC method provided by the EpiDISH algorithm [26] was used to estimate the proportion of epithelial cells in all EWAS samples.

A multilevel model, adapted for the hierarchical structure of the GSE154566 data, was created to estimate the proportion of epithelial cells based on the methylation of both buccal epithelial cell-specific CpGs. The intercept and coefficients of this model were extracted to formulate a general-purpose estimation of the proportion of epithelial cells:

Estimated % of Epithelial Cells = 95.45113 + 0.01371 × cg08141395 methylation − 1.01601 × cg12389346 methylation.

The accuracy of our estimation was evaluated by calculating correlations with, and mean absolute deviations (MAD) from, the EpiDISH-predicted % of epithelial cells. Furthermore, this procedure was repeated with the EWAS data from our lab to validate the results (see Figure 2). Additionally, the association between both buccal epithelial cell-specific CpGs and the multilevel estimation of cellular proportions was investigated in the TBS dataset by calculating correlations (see Figure 4). Correlations between the buccal epithelial cell-specific CpGs were also assessed in both EWAS datasets and the TBS sample to ensure data comparability and transferability (see Figure 3).

2.4.4. Sources of variation in DNA methylation

The buccal epithelial cell estimation, based on cg08141395 and cg12389346, was tested by applying it to TBS data of *ACTN3*. Specifically, a random intercept model was created for the TBS dataset to determine whether acute stress and/or the proportion of buccal epithelial cells contribute to variations in average *ACTN3* methylation. Only buccal swab and mouthwash samples were included in the analysis. Time, type of acute stress, and the buccal epithelial cell estimation were added as predictors, nested within each participant and tissue sample. The amount of variance explained by the model was derived by calculating Nakagawa's R2. Additionally, we "normalized" *ACTN3* methylation by calculating the residuals of a linear model predicting average *ACTN3* methylation based on the estimated % of buccal epithelial cells. As a result, we present the average methylation of *ACTN3* irrespective of the

Table 2. Gene localization and primer sequences.

Name	Chromosomal position	Primer sequences
ACTN3 (178bp)	chr11:66,314,330–66,314,507	Fw: CTTGCTTCCTGGCACGAGgggttgggttgggttttatttaatg Rv: CAGGAAACAGCTATGACccaatcctcctcctattccatat
CTBP1 (220 bp)	chr4:1,214,984–1,215,203	Fw: CTTGCTTCCTGGCACGAGtttgggtttgtatttaggagtgga Rv: CAGGAAACAGCTATGACaaaaatactcaacctcctcactt
MAML2 (327 bp)	chr11:95,987,271–95,987,597	Fw: CTTGCTTCCTGGCACGAGtaggggaaagtgatttgatgttaagt Rv: CAGGAAACAGCTATGACaaaaccacaacaaacttacatattca

The gene-specific forward (Fw) and reverse (Rv) primers, as well as the genomic location of the amplicons used, are shown. Gene-specific, bisulfite-specific DNA sequences are indicated in lowercase letters; uppercase letters denote tag sequences used for amplification during the second-round PCR.

proportion of epithelial cells, thereby controlling for cellular heterogeneity.

All analyses were performed using R-Studio (version 2023.12.1 + 402).

3. Results

3.1. Two buccal epithelial cell specific CpGs enable estimation of cellular heterogeneity in swab and mouthwash samples

Epigenome-wide screening of the environmental risk longitudinal twin study dataset

Epigenome-wide screening of the Environmental Risk Longitudinal Twin Study (ER-LTS) dataset for CpGs exhibiting maximal differences in average methylation between buccal swab and blood samples identified several potential candidates for amplification. We ultimately selected cg08141395, located within the *Mastermind Like Transcriptional Coactivator 2 gene* (*MAML2*), which exhibited high average methylation in buccal swab samples and low average methylation in blood samples. Additionally, we chose cg12389346, located within the *C-Terminal Binding Protein 1* gene (*CTBP1*), as it demonstrated the opposite methylation pattern (Figure 1(a)).

Both CpGs displayed nearly perfect linear correlations with the predicted percentage of epithelial cells derived from the EpiDISH algorithm ($r = 0.99$, 95% CI [0.979, 0.992], for cg08141395 and $r = -0.99$, 95% CI [−0.994, −0.987], for cg12389346; see Figure 1(b,c)). Furthermore, our general-purpose formula for estimating the percentage of epithelial cells, based on a multilevel model, revealed a high correlation with the EpiDISH prediction ($r = 0.99$, 95% CI [0.987, 0.995]) and a mean absolute deviation (MAD) of 3.47% (see Figure 1(d)). These results were replicated in the lab-internal EWAS dataset (see Figure 2(a,d)), and both buccal epithelial cell-specific CpGs demonstrated high correlations across all three datasets (see Figure 3(a,c)). Additionally, the methylation of both CpGs was highly correlated with the multilevel estimation of epithelial cell proportions (see Figure 4(a,b)). Moreover, additional analysis using the GSE40279 dataset did not reveal any significant correlations between age and methylation levels of cg08141395 and cg12389346 in 656 blood samples (age 19–101). It can therefore be assumed that these two target CpGs are not affected by biological aging and show stable methylation patterns across the lifespan.

3.2. Variation in ACTN3 methylation is mainly due to cellular heterogeneity

Predicting differences in *ACTN3* methylation across sample time points, stress type, and the estimated proportion of buccal epithelial cells (adjusted for the multilevel data structure) revealed that

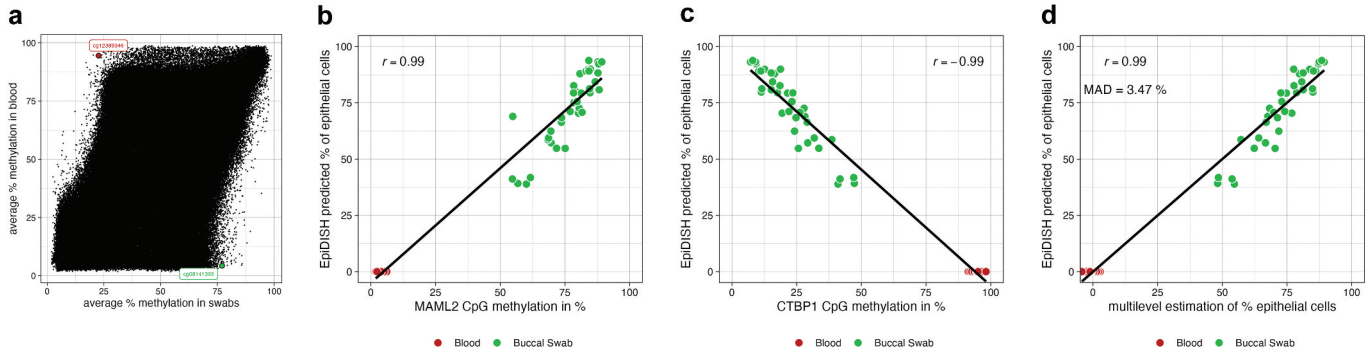

Figure 1. Comparison of DNA methylation at individual CpG sites between different tissues. (a) Average blood and swab methylation of 695.834 CpG sites. (b) Correlation between cg08141395 and the EpiDISH prediction of % epithelial cells. (c) Correlation between cg12389346 methylation and the EpiDISH prediction of % epithelial cells. (d) Correlation between the multilevel and the EpiDISH predicted proportion of epithelial cells. The multilevel model displays a MAD of 3.47% from the EpiDISH prediction.

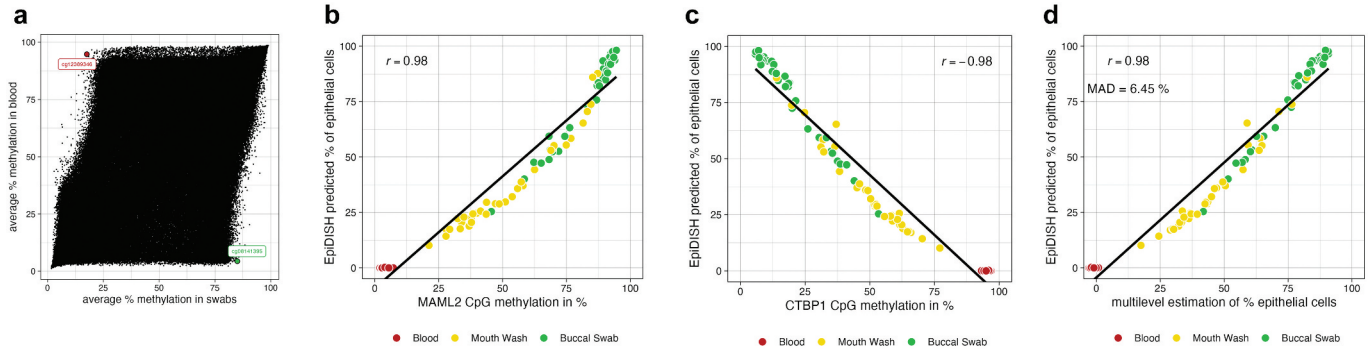

Figure 2. Replication of DNA methylation comparison at individual CpG sites between different tissues. (a) Average methylation of 770.621 CpGs in blood and buccal swab samples. (b) Correlation between cg08141395 and the proportion of epithelial cells predicted by the EpiDISH algorithm. (c) Correlation between cg12389346 and the EpiDISH predicted proportion of buccal epithelial cells. (d) Correlation between the multilevel and the EpiDISH estimation of % buccal epithelial cells. MAD in the lab internal EWAS dataset is 6.45%.

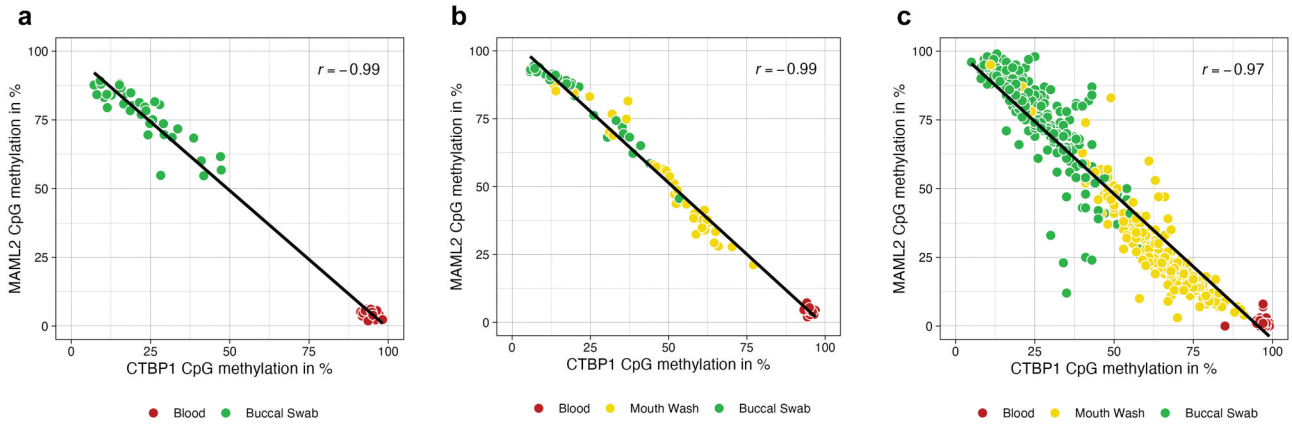

Figure 3. Linear relationship between DNA methylation of cg08141395 and cg12389346. (a) Correlation between cg08141395 and cg12389346 in the environmental risk longitudinal twin study EWAS dataset. (b) Correlation between cg08141395 and cg12389346 in the lab internal EWAS dataset. (c) Correlation between cg08141395 and cg12389346 in the lab internal targeted bisulfite dataset. Similar correlation values indicate adequate data comparability and transferability.

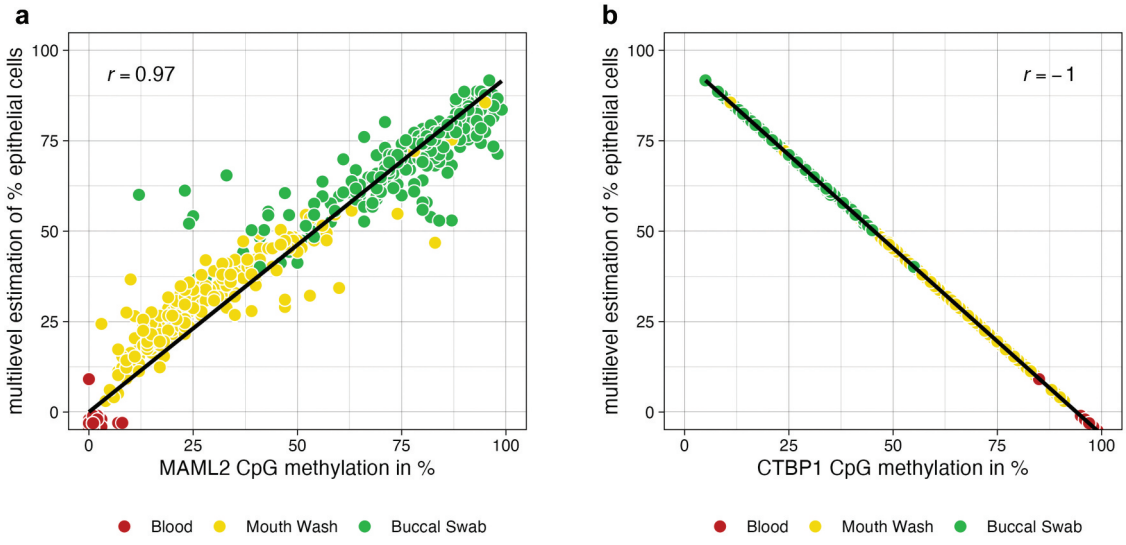

Figure 4. DNA methylation of cg08141395 and cg12389346 in relation to epithelial cell proportions. (a) cg08141395 correlation with the multilevel estimation of % buccal epithelial cells. (b) cg12389346 correlation with the proportion of epithelial cells based on the multilevel model prediction. The perfect linear correlation reflects the massive influence that cg12389346 has on the multilevel model estimation. Potentially, the proportion of buccal epithelial cells can be estimated by only one buccal cell specific CpG site.

only the estimated cell fraction significantly contributed to the variation in DNA methylation (see Table 3). Nakagawa's R2 calculation showed that the fixed effects explained 90.8% of the variance (95% CI [0.874, 0.936]). Additionally, methylation of the *ACTN3* amplicon exhibited a high correlation with the multilevel estimation of the epithelial cell proportion ($r = -0.96$, 95% CI [−0.964,

Table 3. Multilevel analysis of *ACTN3* methylation levels.

Fixed Effects	Coefficient	SE	T-ratio	p
Intercept	36.09	0.31	118.25	<.001***
Time	−0.04	0.03	−1.18	.24
Type of stress	−0.01	0.08	−0.15	.88
Multilevel estimation of % epithelial cells	−0.25	0.004	−70.22	< .001***
Random Effects	*Variance*	*SD*		
Intercept (Tissue/Participant)	0.84	0.92		
Intercept (Participant)	1.93	1.39		
Model fit	*REML*	*Nakagawa's R^2*		
	1285.6	Conditional	0.983 95% CI [0.978, 0.985]	
		Marginal	0.908 95% CI [0.874, 0.936]	

Fixed and random effects together with the model fit of the multilevel analysis are presented. Nakagawa's R2 shows that 90.8% of the variance are explained by fixed effects.

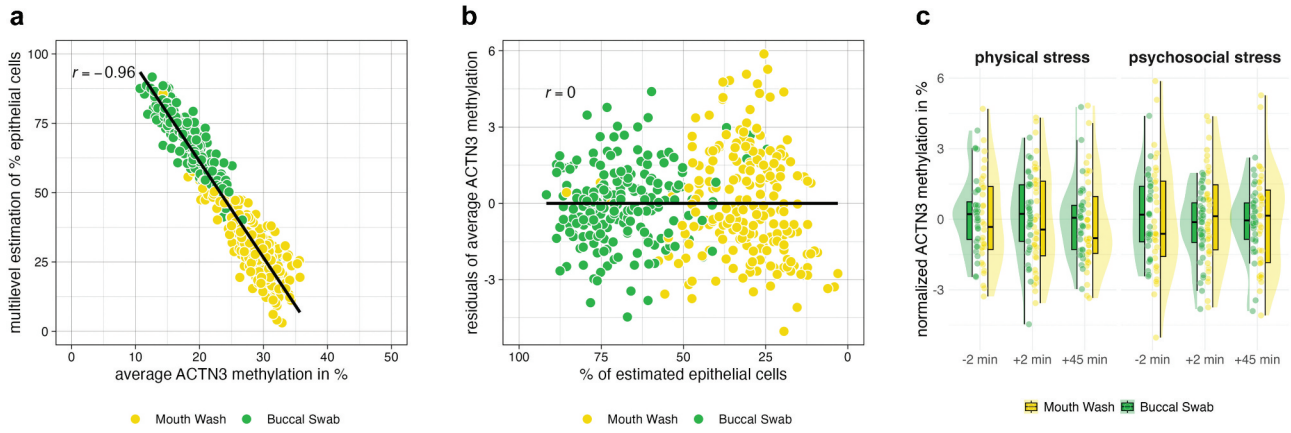

Figure 5. ACTN3 DNA methylation and the influence of buccal epithelial cell type proportions. (a) Correlation between the average *ACTN3* methylation and the multilevel estimation of the proportion of buccal epithelial cells. (b) Residuals/"Normalized" *ACTN3* methylation values of a linear model predicting *ACTN3* methylation based on the estimated proportion of buccal epithelial cells. (c) Raincloud plot of "normalized" *ACTN3* methylation values/residuals for each stressor and measurement time point combination.

−0.947]; see Figure 5(a)). In contrast, the "normalized" *ACTN3* methylation showed some residual variance that was unrelated to the cellular proportion ($r = 0$, 95% CI [−0.1, 0.1]; see Figure 5(b)). Moreover, the distribution of "normalized" *ACTN3* methylation values displayed substantial overlap across all time points and both acute stressors (see Figure 5(c)).

4. Discussion

4.1. Addressing cellular heterogeneity in targeted bisulfite sequencing

True DNA methylation levels of a specific cell type can only be assessed in isolated cells [14]. However, various statistical correction methods have been developed to estimate the proportion of cell types in blood and buccal samples using epigenome-wide data [26]. To date, no such statistical approaches have been available for targeted bisulfite sequencing data. To fill this gap, we introduce a general formula for estimating the proportion of buccal epithelial cells. Specifically, we followed the approach proposed by Eipel et al. [9] and performed an epigenome-wide screen to identify two CpGs with maximal differences in average methylation between blood and buccal swab samples. Using the Illumina Infinium HumanMethylation850 BeadChip, we screened nearly 700,000 CpGs and identified cg08141395, located within ***MAML2***, and cg12389346, located within ***CTBP1***, as suitable candidates for amplification.

Additionally, we validated our findings in a secondary EWAS sample and applied our formula to a targeted bisulfite sequencing dataset consisting of buccal swab and mouthwash samples. CpG site cg08141395 exhibited an average methylation of > 73% in buccal swabs and < 5% in blood samples, while cg12389346 showed > 94% methylation in blood and < 28% in buccal swab samples across all datasets. Both CpGs demonstrated high correlations ($|r \geq 0.98|$) with the EpiDISH predicted proportion of epithelial cells. The formula derived from a multilevel model, which used both CpGs to estimate the proportion of buccal epithelial cells, showed a high correlation ($|r \geq 0.98|$) with the EpiDISH prediction, with mean absolute deviations (MADs) ranging from 3.47% to 6.45%. While our multilevel estimation does not match the accuracy of the EpiDISH prediction, it enables control over the influence of cellular heterogeneity in DNA methylation studies of buccal swab and mouthwash samples.

To further assess this, we compared the effects of two acute stress paradigms and the estimated epithelial cell proportion on the DNA methylation of a cell type-specific methylated gene, *ACTN3*, as measured by targeted bisulfite sequencing. Raw TBS data revealed > 20% variation in *ACTN3* methylation. However, the multilevel model showed that only the estimated proportion of epithelial cells significantly predicted these differences, explaining more than 90% of the variance. In contrast, when *ACTN3* methylation was "normalized" (i.e., adjusted for epithelial cell proportion), the remaining variance showed no correlation with the cellular proportion. Furthermore, we observed substantial overlap in *ACTN3* methylation values across all measurement time points and both acute stressors, indicating that the observed variation was likely due to measurement errors rather than actual effects of the stressors.

This finding aligns with recent evidence demonstrating that methylation patterns in isolated cell types are highly reproducible [27], underscoring concerns raised in psychological research that differences in DNA methylation may reflect the cellular composition of buccal swabs [8]. Therefore, we recommend that future targeted bisulfite sequencing studies using buccal swab and mouthwash samples incorporate the quantification of two CpGs with maximal differences in methylation between oral and blood tissue. In addition to the CpGs proposed by Eipel et al. [9], we suggest adding cg08141395 and cg12389346 as field-tested candidates for estimating the proportion of buccal epithelial cells. This approach will allow researchers to control for cellular heterogeneity and any shifts in cellular composition that may arise.

4.2. Challenges and future directions

Despite the advances presented, there remains an unresolved need for statistical correction when analyzing blood samples

using targeted bisulfite sequencing. Blood samples also contain heterogeneous cell mixtures [14], which may similarly confound accurate methylation estimates. Additionally, our approach would benefit from further exploration into the number of CpGs required for accurate yet cost-effective estimation. Following the recommendation by Eipel et al. [9], we selected two CpGs for estimating the buccal epithelial cell proportion. However, our formula does not treat both CpGs equally but relies more heavily on the methylation of cg12389346. In contrast, the EpiDISH algorithm uses more than 500 CpGs for its estimation [26]. We acknowledge that using a small number of CpGs may limit precision. Therefore, future research should consider the inclusion of additional CpGs to refine the accuracy of cellular proportion estimates, particularly in cases of complex tissue compositions. Further studies comparing the performance of this approach with cytology-based methods, such as direct cell counts or tissue-specific markers, could also provide important insights into the validity and utility of this formula. This would not only improve the precision of cell type estimation but also help assess the generalizability of our formula across different populations and tissue types.

5. Conclusion

Cellular heterogeneity remains a common challenge in many epigenetic studies, often leading to misattribution of observed deviations in DNA methylation to interventions or phenotypes of interest. To mitigate this source of error, we have introduced a low-cost method for estimating the proportion of buccal epithelial cells in targeted bisulfite sequencing data. We demonstrated that accounting for cell type proportions can explain most of the variation in DNA methylation of a cell type-specific gene (*ACTN3*), while the actual effect of an acute stress paradigm on DNA methylation was negligible. However, future work should explore the inclusion of additional CpGs for more precise estimates and validate the method through comparisons with cytology-based cellular composition analysis. We hope our findings will emphasize the importance of routinely correcting for cellular heterogeneity in DNA methylation studies, ultimately improving the reliability and quality of epigenetic research.

Acknowledgments

We thank the master's students (Charlotte Gevers and Chiara Lindgraf), technical assistants (Kim Walusiacki and Annika Mühlenkamp), and student (Maximilian Niggemeier) from the Department of Genetic Psychology for their indispensable help in recruiting participants, providing technical assistance, and collecting data.

ChatGPT was used for the linguistic revision of the manuscript as well as for revising the code written in R-Studio. OpenAI. (2025). *ChatGPT* (Version January 2025). Retrieved [2025 01 09], from https://openai.com.

Author contributions

The study was conceptualized by Dirk A. Moser, Robert Kumsta, Elisabeth M. Hummel, Fabian Berg, Elisabeth Köper and Alicia S. Limberg, Fabian Berg and Elisabeth Köper collected the data. Samples were processed by Dirk A. Moser, Alicia S. Limberg, and Bettina Budeus. Data analysis was performed by Fabian Berg and Katharina Mattonet. Dirk A. Moser contributed to project administration and, along with Robert Kumsta, to supervising the project. The first version of the manuscript was written by Fabian Berg. Dirk A. Moser, Robert Kumsta and Fabian Berg revised and edited the manuscript for publication. All authors contributed to the final publication and approved the submitted version.

Disclosure statement

The authors have no relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript. This includes employment, consultancies, honoraria, stock ownership or options, expert testimony, grants or patents received or pending, or royalties.

No writing assistance was utilized in the production of this manuscript.

Ethical declaration

This study was conducted in accordance with ethical guidelines and principles for research involving human participants and approved by the local ethics committee (759/2022). Also, all participants gave written consent prior to participation.

Funding

The costs of the study were fully funded by the Department of Genetic Psychology. Moreover, we acknowledge support by the Open Access Publication Funds of the Ruhr-University Bochum.The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Data availability statement

Lab internal EWAS and TBS data is available from the corresponding authors upon request. The GSE154566 dataset is publicly available and can be downloaded from the NIH GEO datasets.

References

Papers of special note have been highlighted as either of interest (•) or of considerable interest (••) to readers.

1. Zhang G, Pradhan S. Mammalian epigenetic mechanisms. IUBMB Life. 2014;66(4):240–256. doi: 10.1002/iub.1264
2. Lee HJ, Hore TA, Reik W. Reprogramming the methylome: erasing memory and creating diversity. Cell Stem Cell. 2014;14(6):710–719. doi: 10.1016/j.stem.2014.05.008
3. Jones MJ, Moore SR, Kobor MS. Principles and challenges of applying epigenetic epidemiology to psychology. Annu Rev Psychol. 2018;69(1):459–485. doi: 10.1146/annurev-psych-122414-033653
4. Larsen F, Gundersen G, Lopez R, et al. Cpg Islands as gene markers in the human genome. Genomics. 1992;13(4):1095–1107. doi: 10.1016/0888-7543(92)90024-M
5. Lokk K, Modhukur V, Rajashekar B, et al. DNA methylome profiling of human tissues identifies global and tissue-specific methylation patterns. Genome Biol. 2014;15(4). doi: 10.1186/gb-2014-15-4-r54
6. Ziller MJ, Gu HC, Müller F, et al. Charting a dynamic DNA methylation landscape of the human genome. Nature. 2013;500(7463):477–481. doi: 10.1038/nature12433
7. Horsthemke B. A critical appraisal of clinical epigenetics. Clin Epigenetics. 2022;14(1). doi: 10.1186/s13148-022-01315-6
8. Kumsta R. The role of epigenetics for understanding mental health difficulties and its implications for psychotherapy research. Psychol Psychother-T. 2019;92(2):190–207. doi: 10.1111/papt.12227
   • **Depicts the issue of cellular heterogeneity confounding epigenetic studies in the field of clinical psychology.**
9. Eipel M, Mayer F, Arent T, et al. Epigenetic age predictions based on buccal swabs are more precise in combination with cell type-specific DNA methylation signatures. Aging-Us. 2016;8(5):1034–1048. doi: 10.18632/aging.100972

•• **Demonstrated that two CpGs are sufficient to estimate cell type proportions in buccal swab samples.**

10. Hannon E, Mansell G, Walker E, et al. Assessing the co-variability of DNA methylation across peripheral cells and tissues: implications for the interpretation of findings in epigenetic epidemiology. PLOS Genet. 2021;17(3):e1009443. doi: 10.1371/journal.pgen.1009443
11. Dhabhar FS, Malarkey WB, Neri E, et al. Stress-induced redistribution of immune cells—from barracks to boulevards to battlefields: a tale of three hormones – Curt Richter award winner. Psychoneuroendocrino. 2012;37(9):1345–1368. doi: 10.1016/j.psyneuen.2012.05.008
12. Koestler DC, Christensen BC, Karagas MR, et al. Blood-based profiles of DNA methylation predict the underlying distribution of cell types a validation analysis. Epigenetics-Us. 2013;8(8):816–826. doi: 10.4161/epi.25430
13. McGregor K, Bernatsky S, Colmegna I, et al. An evaluation of methods correcting for cell-type heterogeneity in DNA methylation studies. Genome Biol. 2016;17(1). doi: 10.1186/s13059-016-0935-y
14. Houseman EA, Accomando WP, Koestler DC, et al. DNA methylation arrays as surrogate measures of cell mixture distribution. BMC Bioinformatics. 2012;13(1):13. doi: 10.1186/1471-2105-13-86

•• **Introduced the workflow of using differentially methylated DNA regions to identify different cell types.**

15. Moser DA, Müller S, Hummel EM, et al. Targeted bisulfite sequencing: a novel tool for the assessment of DNA methylation with high sensitivity and increased coverage. Psychoneuroendocrino. 2020;120:120. doi: 10.1016/j.psyneuen.2020.104784
16. Beiter T, Niess AM, Moser D. Transcriptional memory in skeletal muscle. Don't forget (to) exercise. J Cell Physiol. 2020;235(7–8):5476–5489. doi: 10.1002/jcp.29535
17. Ehlert T, Simon P, Moser DA. Epigenetics in sports. Sports Med. 2013;43(2):93–110. doi: 10.1007/s40279-012-0012-y
18. Kandaswamy R, Hannon E, Arseneault L, et al. DNA methylation signatures of adolescent victimization: analysis of a longitudinal monozygotic twin sample. Epigenetics-Us. 2021;16(11):1169–1186. doi: 10.1080/15592294.2020.1853317
19. Kirschbaum C, Pirke KM, Hellhammer DH. The 'trier social stress test' – a tool for investigating psychobiological stress responses in a laboratory setting. Neuropsychobiology. 1993;28(1–2):76–81. doi: 10.1159/000119004
20. Andersen KLS, J R, Denolin H, et al. Fundamentals of exercise testing [book]. Ann Intern Med. 1971;76(3):54–55.
21. Kirschbaum C, Kudielka BM, Gaab J, et al. Impact of gender, menstrual cycle phase, and oral contraceptives on the activity of the hypothalamus-pituitary-adrenal axis. Psychosom Med. 1999;61(2):154–162. doi: 10.1097/00006842-199903000-00006
22. Limberg AS, Berg F, Koper E, et al. Cell-free DNA release following psychosocial and physical stress in women and men. Transl Psychiatry. 2025;15(1):26. doi: 10.1038/s41398-025-03242-5
23. Miller SA, Dykes DD, Polesky HF. A simple salting out procedure for extracting DNA from human nucleated cells. Nucl Acids Res. 1988;16(3):1215–1215. doi: 10.1093/nar/16.3.1215
24. Assenov Y, Müller F, Lutsik P, et al. Comprehensive analysis of DNA methylation data with RnBeads. Nat Methods. 2014;11(11):1138–1140. doi: 10.1038/nmeth.3115
25. Rahmann S, Beygo J, Kanber D, et al. Amplikyzer: automated methylation analysis of amplicons from bisulfite flowgram sequencing. PeerJ Preprints. 2013;1:e122v2. doi: 10.7287/peerj.preprints.122v2
26. Teschendorff AE, Breeze CE, Zheng SC, et al. A comparison of reference-based algorithms for correcting cell-type heterogeneity in epigenome-wide association studies. BMC Bioinformatics. 2017;18(1):105. doi: 10.1186/s12859-017-1511-5
27. Loyfer N, Magenheim J, Peretz A, et al. A DNA methylation atlas of normal human cell types. Nature. 2023;613(7943):355–364. doi: 10.1038/s41586-022-05580-6