


RESEARCH

Open Access



# A correlation map of genome-wide DNA methylation patterns between paired human brain and buccal samples

Yasmine Sommerer<sup>1</sup>, Olena Ohlei<sup>1</sup>, Valerija Dobricic<sup>1</sup>, Derek H. Oakley<sup>2</sup>, Tanja Wesse<sup>3</sup>, Sanaz Sedghpour Sabet<sup>3</sup>, Ilja Demuth<sup>4,5,6</sup>, Andre Franke<sup>3</sup>, Bradley T. Hyman<sup>7,8,9</sup>, Christina M. Lill<sup>1,10,12</sup> and Lars Bertram<sup>1,11\*</sup> 

## Abstract

Epigenome-wide association studies (EWAS) assessing the link between DNA methylation (DNAm) and phenotypes related to structural brain measures, cognitive function, and neurodegenerative diseases are becoming increasingly more popular. Due to the inaccessibility of brain tissue in humans, several studies use peripheral tissues such as blood, buccal swabs, and saliva as surrogates. To aid the functional interpretation of EWAS findings in such settings, there is a need to assess the correlation of DNAm variability across tissues in the same individuals. In this study, we performed a correlation analysis between DNAm data of a total of  $n = 120$  matched *post-mortem* buccal and prefrontal cortex samples. We identified nearly 25,000 (3% of approximately 730,000) cytosine-phosphate-guanine (CpG) sites showing significant (false discovery rate  $q < 0.05$ ) correlations between buccal and PFC samples. Correlated CpG sites showed a preponderance to being located in promoter regions and showed a significant enrichment of being determined by genetic factors, i.e. methylation quantitative trait loci (mQTL), based on buccal and dorsolateral prefrontal cortex mQTL databases. Our novel buccal–brain DNAm correlation map will provide a valuable resource for future EWAS using buccal samples for studying DNAm effects on phenotypes relating to the brain. All correlation results are made freely available to the public online.

**Keywords:** DNAm, Brain, Buccal, Inter-tissue correlation

## Introduction

DNA methylation (DNAm) is an epigenetic mechanism in vertebrate genomes that most often refers to the addition of a methyl-group to cytosine nucleotides within the DNA sequence. Most DNAm in somatic cells occurs in stretches of cytosine-phosphate-guanine (CpG) sites, where they typically, but not always, represent an epigenetic mark of translational repression [1]. Owing to the relative technical ease to generate high-resolution

DNAm data for thousands of CpG sites simultaneously, it currently represents one of the most frequently studied epigenetic marks. Accordingly, study designs exploiting DNAm profiles on a genome-wide scale—often referred to as epigenome-wide association studies (EWAS)—are becoming increasingly popular. Many EWAS aim to assess the relationship between DNAm patterns and certain brain-related phenotypes [2, 3], such as neuropsychiatric traits [4, 5], cognitive functions [6, 7], and risk for neurodegenerative diseases [8–10], with the goal to better understand the biology and pathophysiology of the traits of interest. However, given that the primary organ of interest, the brain, is typically inaccessible in living individuals, many studies use tissues that are more readily available. While most DNAm studies use blood

\*Correspondence: lars.bertram@uni-luebeck.de

<sup>1</sup> Lübeck Interdisciplinary Platform for Genome Analytics (LIGA), University of Lübeck, Ratzeburger Allee 160, Haus V50, 1st Floor, Room 319, 23562 Lübeck, Germany  
Full list of author information is available at the end of the article



© The Author(s) 2022. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

samples [3, 6, 7, 10], it has been hypothesized that buccal [11] or saliva [12] samples may be more informative for EWAS of psychiatric phenotypes. One key advantage of using peripheral tissues as surrogates is that samples can be obtained from living individuals and do not require *post-mortem* sampling. This also allows for study designs using longitudinal sampling and analysis. However, DNAm patterns are largely cell type- and tissue-dependent [13, 14], and therefore, it remains unclear how well peripheral DNAm patterns can be used to infer biological processes in the brain.

In the recent past, several attempts have been made to compare DNAm profiles between brain and peripheral tissues within the same individuals. However most studies focused on the comparison of blood and brain tissues [15, 16]. Only two reports recently compared buccal and brain samples, but sample sizes were very small with 12 and 27 matched sample pairs, respectively [17, 18]. One additional study used brain, thyroid, and heart tissue samples, each representing one developmental germ layer lineage, from ten individuals to identify so-called correlated regions of systemic interindividual variation (CoRSIV) [19]. Interestingly and perhaps not unexpectedly, most of the cited studies report a significant enrichment of methylation quantitative trait loci (mQTL), i.e. DNAm variations that are associated with genetic variants, among CpG sites correlated between tissues or cell-types [13, 15–17, 19].

Despite this recent progress, there currently remains a significant lack of studies systematically assessing DNAm patterns in paired buccal and brain specimen in sufficiently sized datasets. To close this gap, we here report the results of comprehensive and systematic correlation analyses of genome-wide DNAm patterns in 120 paired prefrontal cortex (PFC) and buccal swab samples. We identified 24,980 significantly correlated CpGs between both tissues and found a significant enrichment of both buccal and dorsolateral prefrontal cortex (DLPFC) mQTLs among the correlated CpG sites. All genome-wide DNAm correlation results are made freely available online ([http://www.liga.uni-luebeck.de/buccal\\_brain\\_correlation\\_results/](http://www.liga.uni-luebeck.de/buccal_brain_correlation_results/)), as we anticipate that the buccal–brain DNAm correlation map we generated in this study will provide a valuable resource for the interpretation of EWAS/DNAm studies for brain-related phenotypes.

## Material and methods

### Human samples

Matched prefrontal cortex (PFC) and buccal samples were obtained in two batches from the neuropathology unit at the Massachusetts Alzheimer's Disease Research Center (MADRC), Boston, MA, USA. Samples were shipped to our laboratory in two batches: "MADRC-1"

and "MADRC-2", encompassing 48 and 80 matched brain–buccal pairs, respectively. Buccal swabs were obtained from patients with neurodegenerative disease conditions and controls at the time of autopsy following an IRB-approved informed consent with specific inclusion of genetic studies. Consent forms were completed by next-of-kin or other legal representatives as specified by Massachusetts state law. Buccal-Prep Plus DNA Isolation Kit (Isohelix, UK) swabs were utilized to obtain buccal swabs; these were held at  $-80^{\circ}\text{C}$  without dehydration until DNA extraction (see below). One hemisphere of each harvested brain was coronally sectioned, flash-frozen on dry ice, cryopreserved at  $-80^{\circ}\text{C}$ , and used for subsequent PFC isolation and DNA methylation (DNAm) profiling (see below). The remaining hemisphere was fixed in 10% weight/volume formalin and subjected to detailed neuropathologic evaluation. Detailed descriptions of all MADRC buccal–brain samples used in this study can be found in Additional file 1: Table 1.

### DNA extraction and processing

DNA was extracted and processed in two laboratory batches according to their shipment charge (i.e. MADRC-1 and MADRC-2; Additional file 1: Table 1). Importantly, paired brain and buccal samples from the same shipment were processed simultaneously (incl. DNAm profiling, see below). For brain samples, genomic DNA was extracted from approximately 50 mg of frozen tissue using the DNeasy Blood & Tissue Kit (Qiagen, Hilden, Germany), while DNA from the buccal swabs was extracted using Buccal-Prep Plus DNA Isolation Kit (Isohelix, UK). All steps in the extraction procedure were conducted according to manufacturer's instructions. The quantity and the quality of obtained DNA were assessed using a NanoDrop ONE spectrophotometer (Thermo Fisher Scientific, USA).

### EPIC array profiling

DNAm profiling was performed using the "Infinium MethylationEPIC" array (Illumina, Inc., USA), as described previously [8]. In brief, experiments were performed on aliquots of DNA extracts diluted to  $\sim 50\text{ ng}/\mu\text{l}$  concentration. Bisulphite conversion of DNA samples was performed using the EZ DNA Methylation kit (Zymo Research, USA), following the alternative incubation conditions for the Illumina Infinium MethylationEPIC Array as recommended by the supplier. After hybridization to the EPIC array, scanning was performed on an iScan instrument (Illumina, Inc.) according to the manufacturer's instructions (Document#1000000077299v0). DNA samples from both shipment charges (MADRC-1 and MADRC-2) were processed in consecutive laboratory experiments to minimize batch effects. Raw DNAm

intensities were determined using the iScan control software (v2.3.0.0; Illumina, Inc.) and exported in .idat format for downstream processing and analysis.

#### DNA methylation data processing and quality control

DNAm data from each batch (MADRC-1 and MADRC-2) and tissue (PFC and buccal) were loaded into R and pre-processed separately. DNAm data pre-processing and quality control (QC) was performed using the same procedures as described previously [8] unless noted otherwise. In brief, this entailed using the R (v. 3.6.1) package *bigmelon* with default settings [20]. Samples were excluded when bisulphite conversion efficiency was below 80%. Outliers were removed using the *outlyx* function in *bigmelon* applying a threshold of 0.15. CpG sites on the X or Y chromosome, or those aligning to SNPs [21] or multiple locations in the genome [22] were removed from the analysis. The final analysis included a total of 120 matched PFC and buccal samples, with 44 samples from MADRC-1, and 76 samples from MADRC-2. Overall, a total of 730,157 QC'ed CpG sites were available in all four datasets and were used for the analyses.

To compare our results with those from the Braun et al. study evaluating the correlation of DNAm between buccal and brain samples [17], we downloaded the publicly available .idat files of that study (GEO accession number GSE111165), and loaded and pre-processed them with the R-package *ChAMP* [23] using default settings unless otherwise noted. Brain and buccal samples were loaded and pre-processed separately. Briefly, DNAm values with a detection *p*-value above 0.01 were set to N/A and CpG sites were completely removed if there were less than 3 beads in more than 5% of the samples, if they were on an X or Y chromosome, or if they aligned to SNPs [21] or multiple locations in the genome [22]. Normalization was performed with the *BMIQ* method. The analysis of this dataset comprised 21 pairs of matched buccal and brain samples, with 740,507 CpG sites. We note that 1,513 (65%) of the 2,367 significantly correlated CpGs according to the Braun et al. study [17] were not included in our reanalysis of the dataset as they were removed during QC (Additional file 1: Table 2). Almost 50% of excluded CpGs (*n*=931) were removed from our re-analysis of the data due to aligning to or being influenced by SNPs according to Zhou et al. [21], while ~350 CpG sites (23%) were removed due to their location on the X- or Y-chromosome. Despite these differences, we note that the vast majority (713; 83%) of the 854 remaining correlated CpG sites from Braun et al. [17] were also significantly correlated in our re-analysis of the Braun et al. data after multiple testing adjustment using a false discovery rate (FDR) *q*-value threshold of 0.05 (Additional file 1: Table 2).

#### Determination of and correction for DNAm covariates

First, cell-type composition estimates were obtained with the R package *EpiDISH* [14] for buccal samples and the *estimateCellCounts* function in the R package *Minfi* [24] for brain samples. Next, to assess the effects of potential confounders on the DNAm data, we used an adaptation of the singular value decomposition (SVD) approach described previously [25]. In short, SVD attempts to identify and correct for relevant variables that have a significant impact on genome-wide DNAm patterns and could act as confounders in subsequent analyses. Accordingly, we tested whether variation in cell type composition, bisulphite conversion efficiency, EPIC array ID, diagnosis, extraction date, and position on the EPIC array significantly associated with the variance in DNAm data in our data (as determined by principal component analysis [PCA], see below). These analyses were performed separately for buccal and brain datasets. To this end, we performed a PCA on the DNAm beta values after QC using the R base function *prcomp*. For this PCA, we first generated a subset of uncorrelated CpG sites by dividing the genome into 100 kb bins and using one random CpG site from each bin, resulting in 25,746 CpG sites included in each PCA. For the determination of relevant covariates for subsequent analyses, PCs explaining a substantial amount of variance in the DNAm data, as determined by scree plots (MADRC: Additional file 2: Fig. 1; Braun et al. data: Additional file 2: Fig. 2) were used. For numerical variables (bisulphite conversion efficiency and cell type composition estimates), a Pearson correlation test between the centred variables and the centred DNAm PCs was calculated with the R base function *cor.test*. For categorical variables (extraction date, EPIC array ID, position on the EPIC array, and diagnosis), a one-way ANOVA between the covariates and the DNAm PCs was performed with the R base function *aov*. Effects of variables explaining variance of at least one included DNAm PC with a *p*-value < 0.01 were removed from the DNAm beta values using the *removeBatchEffect* function in the R package *limma* [26]. The results of these analyses, as well as the number of DNAm PC eigenvalues (PCs) included for each dataset, can be found in Additional file 1: Table 3. The covariate-adjusted DNAm beta values of the two batches of PFC samples and buccal samples were combined in a “brain” and “buccal” data matrix, respectively. Lastly, the batch-defining variable (i.e. indicating either dataset MADRC-1 or MADRC-2) was removed from these combined matrices with the *removeBatchEffect* function in the R package *limma* [26]. All subsequent analyses were performed on these combined DNAm values adjusted for both covariates and batch.

To ensure that our results were not impacted by confounders not included in our adaptation of the SVD

method described above, we repeated our analyses by correcting the DNAm-values directly for the first three DNAm PCs. We used the DNAm PCs as described above and corrected the DNAm-values using the *removeBatchEffect* function in the R package *limma* [26].

#### Identification of CpG sites with correlated DNAm values between paired brain and buccal samples

Spearman rank correlations were calculated for each CpG site in a pair-wise manner across buccal and PFC samples using the R base function *cor.test*. The resulting *p*-values were adjusted for multiple testing using the FDR approach with the R base function *p.adjust*. FDR *q*-values < 0.05 were considered genome-wide significant in the context of this study. These analyses were performed for the matrices with the two batches (MADRC-1 and MADRC-2) combined, and for each batch separately. Our choice of using nonparametric Spearman rank (as opposed to parametric Pearson's correlation analyses) was motivated by the fact that neither the buccal nor the brain derived DNAm data were normally distributed, and computation of Pearson's *r* assumes linearity of the correlation, an assumption that is not necessarily fulfilled here in all instances.

#### Identification of mQTLs in buccal and blood samples from an independent dataset

To check for buccal-specific mQTLs, we used an independent in-house dataset with 837 buccal samples and 1,058 blood samples ascertained from the Berlin Aging Study II (BASE-II) [27, 28]. These data are referenced here as “unpublished data”, since a manuscript from our group with more details on this analysis is in preparation. In brief, for the buccal dataset both genome-wide QC'ed DNAm profiles (761,034 CpG-sites) and SNP genotyping data (7,663,257 SNPs) were available for mQTL analysis. DNAm data were derived from buccal-swab samples and generated and processed using the same procedures described above. For the blood dataset, genome-wide QC'ed DNAm profiles (763,828 CpG-sites) and SNP genotyping data (7,663,257 SNPs) were available for mQTL analysis. DNAm data was derived from blood samples which were extracted using commercial kits (Plus XL manual kit, LGC, UK). Genome-wide SNP genotyping data were generated from the same samples using the “Global Screening Array” (GSA) with shared custom content (Illumina, Inc.) using procedures outlined in Hong et al. [29]. To compute *cis* mQTLs (defined as within  $\pm 1$  Mb of the CpG site) in this dataset we used the matrix eQTL software [30], which performed an additive linear model with sex, genetic PC 1 to 5, DNAm PCs 1 to 10, and genotyping batch as covariates. Before association analysis, genome-wide DNAm profiles were adjusted

for cell type composition estimates. Only the DNAm and SNP effects that were below a *p*-value of 5.00E-02 were reported. *Cis* mQTLs with FDR *q* < 0.05 were defined as genome-wide significant for this arm of our analyses. Enrichment analyses for mQTLs within CpG sites correlated between PFC and buccal-swab samples were performed with the R base function *chisq.test*, using a subset of uncorrelated CpG sites according to 100 kb bins, as described above for the PCA.

#### Annotation of genomic regions to CpG sites

To assess whether there was a significant enrichment or depletion of CpG sites located in specific genomic regions, we used the annotation from the R package *IlluminaHumanMethylationEPICmanifest* to assign CpG sites to one of the following genomic regions: 1st exon, 3' untranslated region (UTR), 5'-UTR, gene body, exon boundary, intergenic region (IGR), the region from transcription start site (TSS) to 200 nucleotides upstream (TSS200), and the region from 200 nucleotides upstream of the TSS to 1500 nucleotides upstream (TSS1500). Enrichment analyses were performed with the R base function *chisq.test*.

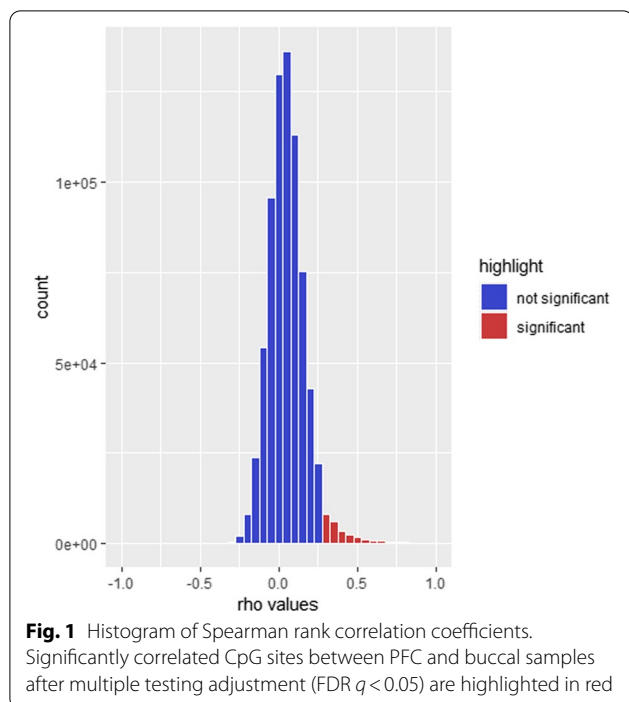
#### Gene ontology (GO) analysis

To further characterize the correlated CpG sites, a Gene Ontology (GO) enrichment analysis was performed with the *gometh* function in the R package *missMethyl* [31] using the significantly correlated (FDR *q* < 0.05) CpGs between PFC and buccal samples. We hypothesized that correlated CpG sites between buccal and brain might show an enrichment for “housekeeping” functions, which would explain the correlated DNAm-values. Nominally significant GO terms were subsequently submitted to the REVIGO tool (<http://revigo.irb.hr/>) [32] to identify and remove redundancy using Resnik's measure while allowing a terms similarity of 0.7.

## Results

### Spearman rank correlation analysis highlights 24,980 CpG sites showing significant correlation between PFC and buccal samples

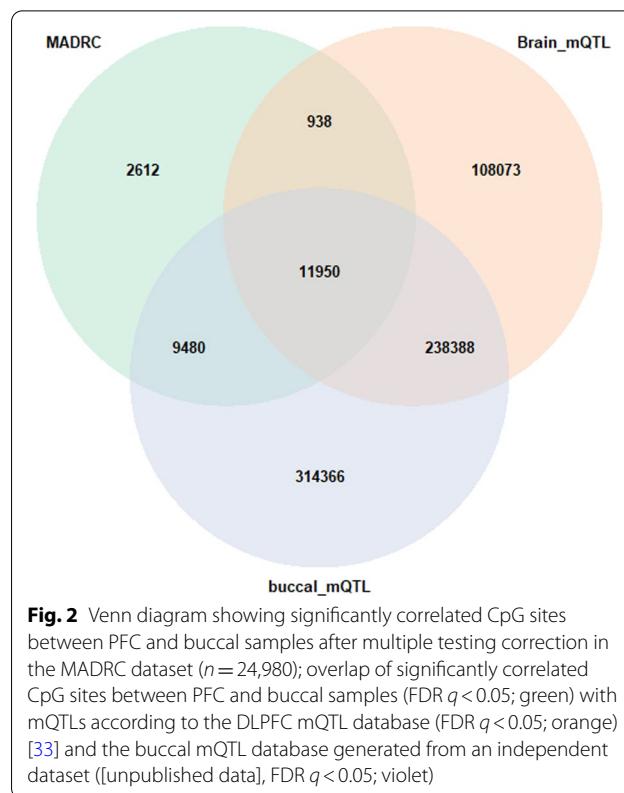
Out of the 730,157 CpG sites that were tested, 3% (*n* = 24,980) showed significant Spearman rank correlations of DNAm beta values between paired PFC and buccal samples after adjustment for multiple testing (FDR *q* < 0.05). Most of the significantly correlated CpG sites (*n* = 24,636; 99%) had a positive Spearman rank correlation coefficient (Fig. 1), which means that DNAm patterns were consistent in both tissues. The remainder (*n* = 344; 1%) showed negative correlations, meaning that the effect direction was opposite in buccal and brain samples. Furthermore, correlated CpG sites were



evenly distributed across the genome and showed no obvious preponderance for any significant genomic location (Additional file 2: Fig. 3). Lastly, the majority of significantly correlated CpG sites was also correlated between PFC and buccal samples at  $p < 0.05$  in analyses that were performed for each batch (“MADRC-1” and “MADRC-2”) separately (Additional file 2: Fig. 4). These latter computations were performed to check for spurious correlations that may have resulted from undetected confounding after merging both laboratory batches. To ensure that our analyses were not influenced by unknown confounders, we repeated the pair-wise Spearman rank correlation with DNAm data which were directly adjusted for the first three DNAm PCs. This analysis resulted in 34,392 significantly correlated CpGs ( $q < 0.05$ ). This included 20,914 of 24,980 (~84%) of the CpGs that were also identified in the main analysis. These findings suggest that the results of our study are not substantially impacted by confounding due to unknown factors.

#### Correlated CpG sites are enriched for both buccal and brain mQTLs

We next looked up our ~25 K correlated brain–buccal CpGs in two mQTL databases: Firstly, in an mQTL database ([https://eqtl.brainseq.org/WGBS\\_meQTL/](https://eqtl.brainseq.org/WGBS_meQTL/)) for dorsolateral prefrontal cortex (DLPFC) [33] mQTLs based on 165 samples, and secondly, in an in-house mQTL database for buccal samples that we generated in 839 individuals analysed as part of the Berlin Ageing Study



II (BASE-II; unpublished data). The look-up resulted in an enrichment of *cis* mQTLs in the fraction of CpG sites that were correlated between PFC and buccal samples when compared to all analysed CpGs ( $p = 4.34E-225$  for DLPFC mQTLs and  $p = 1.16E-300$  for buccal mQTLs using a Chi-squared test, Fig. 2). Overall, 22,368 CpG sites (90%) that were significantly correlated between PFC and buccal samples in our dataset were identified in mQTL analyses in either buccals ( $n = 21,430$ ) or DLPFC ( $n = 12,888$ ). Of these, 11,950 CpG sites were identified in mQTL analyses in both tissues. This also resulted in the identification of 2612 CpGs (10.46% of all correlated CpGs) which were neither buccal nor brain mQTLs, therefore providing potentially valuable information about the DNAm status in both tissues not determined by the DNA sequence.

#### Correlated CpG sites show good concordance with results from published datasets using matched brain and peripheral tissues

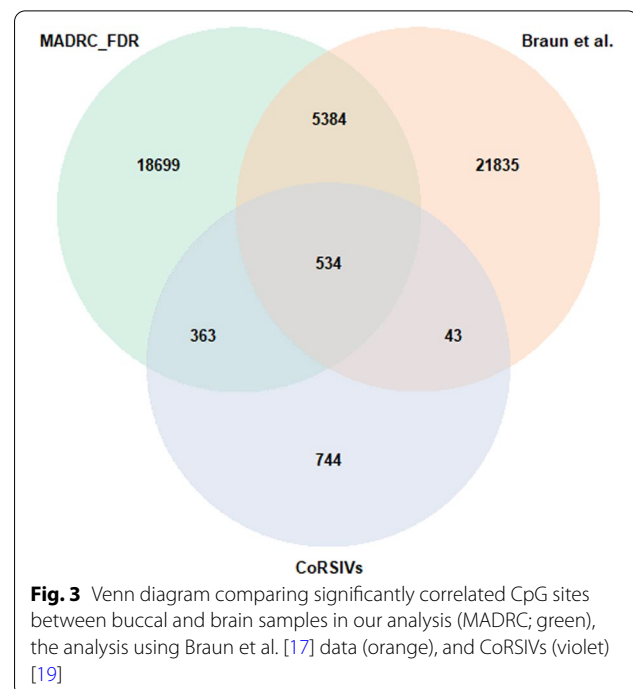
We compared our results to those of two previous publications evaluating the correspondence of DNAm between tissues [17, 19]. The first study by Braun et al. [17] correlated DNAm-values between brain tissues obtained during a surgery of epilepsy patients and buccal samples from the same individuals. Overall, they

identified 2367 CpGs (i.e. 0.29% of all tested) as significantly correlated [17]. To increase comparability between our and the Braun et al. dataset [17], we reprocessed the DNAm raw data from that study including covariate adjustment and Spearman rank correlations as applied to the MADRC datasets. Our re-analysis of the Braun et al. data (for details of the analysis see methods section) resulted in 27,796 CpGs showing significant (FDR  $q < 0.05$ ) correlation between buccal and brain tissue. While this number is comparable to the 24,980 correlated probes identified in the analyses of the MADRC data, it still represents a nearly tenfold difference as compared to the numbers originally published by Braun et al. [17]. This (stark) difference can likely be attributed to a more stringent multiple testing correction procedure (i.e. Bonferroni [Braun et al.] vs. FDR [here]) and differences in data processing and analysis strategies (Methods). Overall, a total of 5918 (24%) of the 24,980 correlated CpGs in our data also represented correlated CpGs in the Braun et al. data.

The second dataset used for comparison was recently published by Gunasekara et al. [19] who identified 9926 significant CoRSIVs across three different tissues (brain, thyroid, and heart) from 10 matched individuals. A total of 1311 (13%) of all CoRSIVs also had at least one CpG probe on the EPIC array, with some regions being represented by more than one CpG site. This resulted in a total of 1,684 individual CpG sites in the MADRC analysis that were located in a CoRSIV and could be used for comparison. A total of 897 CpG sites (53%) showed a significant correlation at FDR  $q < 0.05$  in our data, too. For a depiction of the three-way comparisons of the Braun et al., Gunasekara et al., and our MADRC data, see the Venn diagram in Fig. 3.

#### Genomic location of correlated CpG sites shows an enrichment in gene promoters and depletion in gene bodies

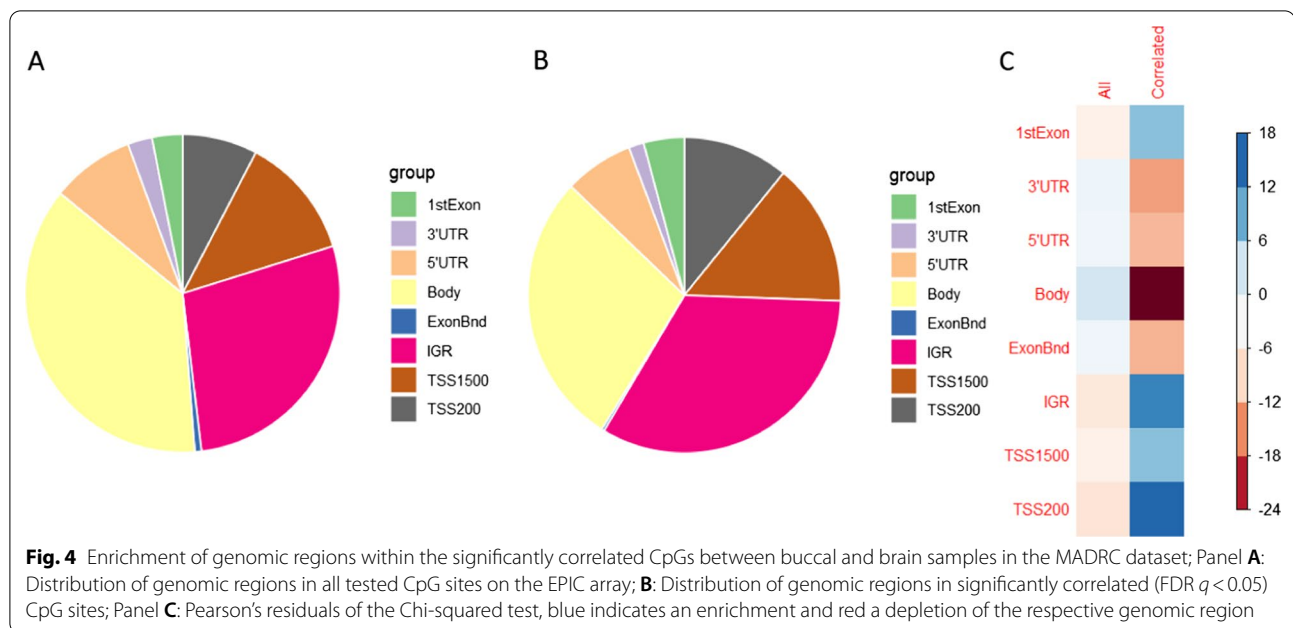
Next, we assessed whether there was an enrichment or depletion of correlated CpG sites in specific genomic regions in comparison to all CpGs. We noted a statistically significant change of the genomic region distribution within the CpG sites that were correlated between PFC and buccal sample DNAm in the MADRC dataset compared to the expected distribution in our data (Chi-squared test with  $p < 2.20E-16$ , Fig. 4). Specifically, we observed an enrichment of CpG sites located in the 1st exon, intergenic regions, regions 1,500 nucleotides upstream of transcription start sites, and regions 200 nucleotides upstream of transcription start sites, and a depletion in the 3'-UTR, 5'-UTR, gene bodies, and exon boundaries within the correlated CpG sites (Fig. 4). The



enrichment in IGRs and depletion in gene bodies is in line with previous observations made for CoRSIVs [19].

#### GO analysis of genes annotated to correlated CpG sites highlights cellular functions relating to major histocompatibility complex (MHC)

Next, we aimed to assess whether the correlated buccal–brain CpG sites fell into specific functional pathways and tested for an enrichment of specific gene ontology (GO) terms. Upon including all 24,980 significantly correlated (FDR  $q < 0.05$ ) CpG sites, this analysis revealed only one statistically significant GO term (GO:0007156, “homophilic cell adhesion via plasma membrane adhesion molecules”, FDR  $q = 0.04$ ). In order to remove redundant terms, all nominally significant GO terms ( $p < 0.05$ ) were submitted to REVIGO, which resulted in 246 nominally significant GO terms (Additional file 1: Table 4). Among the top GO terms showing at least a nominally significant enrichment were many terms related to “house-keeping” functions, such as peptide antigen binding, MHC protein complex, ion binding, cell–cell adhesion via plasma-membrane adhesion molecules, transferase activity, manganese ion binding, and nucleoside-triphosphatase regulator activity (Additional file 1: Table 4). In general, this is in line with the GO results presented in the Gunasekara et al. [19] publication (using GO terms associated with the CoRSIVs). One noteworthy overlapping annotation was observed with the “MHC protein



complex” likely highlighting the central role of MHC-mediated immune response in both tissues.

#### Look-up of Alzheimer’s disease EWAS results as an example of application for the buccal–brain correlation map

As a first “practical” application of our buccal–brain correlation results, we looked up top-hits from the hitherto largest EWAS across several brain regions [9] and for the entorhinal cortex (EC) for Alzheimer’s disease (AD) [8]. This look-up revealed that five of the CpGs that were previously highlighted in the context of AD also showed significant correlations at FDR  $q < 0.05$  and all correlations were positive (cg05030077 [MLST8/chr16:2255199], cg05923197 [BMP4/chr14:54418804], cg23950714 [DOK3/chr5:176935364], cg04252044 [chr3:188664747], cg24569831 [RGMA/chr15:93617168], Additional file 1: Table 5). Probe cg05923197 is also an mQTL according to the DLPFC mQTL database [33] and all five of them were identified as mQTLs in our buccal and blood mQTL databases (unpublished results). Furthermore, lowering the significance threshold to correlations significant at a nominal  $p$ -value of 0.05 increased this number to 29 CpGs showing association in one of the AD brain EWAS. Among these 29, 8 CpG sites were reported to be mQTLs in the DLPFC and 28 in both blood and buccal mQTL databases. Interestingly, all but two showed positive correlations in DNAm beta values between brain and buccal tissues (Additional file 1: Table 5). Based on these data, we predict that all 29 overlapping CpGs should show at least some degree of association evidence in AD EWAS performed in buccal samples. Even though there was a

large overlap between mQTLs in brain, buccal, or blood tissue (Additional file 1: Table 5), we note that assessing DNAm at these CpGs may still yield additional information beyond the variance explained by the genetic variants. Therefore, their potential for representing early biomarkers of AD should be the focus of future work.

#### Discussion

In this study, we generated high-resolution genome-wide DNAm profiles from brain and buccal samples collected *post-mortem* from the same individuals at the same timepoint. Comparing CpG sites across both tissues revealed ~25 K sites showing significant correlations in DNAm levels. This is in line with a previous study assessing the correlation of DNAm between brain and blood samples, which reported moderate to strong correlations for 1% to 6% of CpG sites [16]. Correlated CpGs showed an enrichment for being regulated by mQTLs using both buccal and brain databases, which is in line with results from previous publications [13, 15–17, 19]. In terms of physical location, correlated CpGs showed a significant enrichment in promoter regions and a significant depletion in gene bodies. A GO enrichment analysis highlighted terms related to molecular “house-keeping” functions, including several significant GO terms linked to the MHC protein complex, confirming previous findings [19]. To our knowledge, our study has generated the largest buccal–brain DNAm correlation map available to date and will hopefully prove to be a valuable resource for the interpretation of EWAS for brain-related phenotypes. Furthermore, other applications of our resource

relate to using it to aid the design of EWAS analyses, e.g. by focusing only on correlated CpGs highlighted in this work thereby increasing statistical power by reducing the multiple testing burden. To this end, we made all of our genome-wide DNAm correlation results freely available online ([http://www.liga.uni-luebeck.de/buccal\\_brain\\_correlation\\_results/](http://www.liga.uni-luebeck.de/buccal_brain_correlation_results/)).

Comparing our correlation results with those from a previous publication from Braun et al. [17] using matched brain and buccal samples resulted in an overall good correspondence in results (Additional file 1: Table 2). Differences in findings across both studies can likely be attributed to differences in clinical diagnoses of included individuals, as well as differences in technical aspects such as study design and QC procedures. For instance, the brain samples from Braun et al. [17] were obtained from living individuals during epilepsy surgery, whereas the MADRC samples were collected *post-mortem* and included individuals with different neurodegenerative diseases (Additional file 1: Table 1; see below). As a result, the brain samples from the Braun et al. study [17] were from many different brain regions, whereas the MADRC brain samples were all obtained from the PFC. In addition, in the MADRC dataset the extraction of brain and buccal samples was performed at the same time, i.e. at the time of autopsy, whereas brain and buccal samples in the Braun et al. study [17] were not always collected at the same time point (time range 0–638 days). The second comparison with published data was performed with the CoRSIVS from Gunasekara et al. [19]. Although the CoRSIV assessments from that study and buccal–brain correlation analyses performed here differed in many important aspects, the overlap between both analyses was more than 50% of all CpGs on the EPIC array that are located in CoRSIVs. We suggest that this number, rather than the 21% overlap with Braun et al., be considered as the lower bound of “true” correlations in DNAm patterns in human buccal vs. brain samples.

The strengths of our study are its pair-wise design (i.e. simultaneous collection and analysis of all paired buccal and brain samples), the use of the most current DNAm microarray (i.e. the EPIC array with ~730 K CpG probes available for analyses as opposed to the predecessor array with 450 K CpGs), our stringent QC and data processing procedures (e.g. to eliminate bias due to undetected confounding by certain biological or technical variables), and the comparatively large size of our sample (i.e.  $n=120$  here versus  $n=21$  in Braun et al. [17] and  $n=10$  in Gunasekara et al. [19]). Furthermore, we make use of a novel and hitherto unpublished buccal and blood tissue mQTL database from our group, allowing to determine the impact of genetics in the greatest detail possible. Despite these strengths, our study is also

subject to several limitations. First and probably most importantly, all DNAm profiles obtained from this study are from bulk tissue samples. While all DNAm data were corrected for cell type composition estimated from current reference panels, it cannot be excluded that undetected difference in cell type composition across samples has created a bias in results. Conceptionally, however, this bias (if it existed) should have increased the number of false-negative findings but would not invalidate our findings, i.e. it would result in a bias towards the null. Only studies applying single-cell sequencing could shed more light on the impact of cell-type specific differences in DNAm profiles. Second, we note that the majority of individuals used in this study were not “healthy controls” but had received some type of neuropathological diagnosis, mostly due to the presence of some neurodegenerative disorder (Additional file 1: Table 1). Since these underlying disease conditions likely had an impact on DNAm patterns in the brain and treatment regimens may have affected methylation in the brain and elsewhere, it cannot be excluded that the buccal–brain correlation results reported here were at least partially influenced by diagnosis status. In a similar vein, all subjects included here were relatively old, with a mean age of ~72 years. It is well known that DNAm patterns change over time and are different in aged vs. non-aged individuals [34–36], so use of our buccal–brain correlation map may be less informative for EWAS of younger individuals. Third, as described above, both types of sampled biospecimen (i.e. brain and buccal swabs) were collected *post-mortem*, i.e. with a specific and varying time interval (*post-mortem* interval; PMI) between death and sampling (Additional file 1: Table 1). It is difficult to predict whether and how DNAm patterns were affected by differing PMIs across individuals and tissues. However, evidence from previous work suggests that DNAm appears to be a rather stable epigenetic marker under varying conditions in brain samples collected *post-mortem* [37]. Fourth, we note the large overlap of correlated CpGs across brain and buccal tissue with mQTLs in buccal, brain, and blood samples. In theory, these could also be assessed in conventional GWAS designs which typically include much larger sample sizes. However, we emphasize that at least 10% of the correlated CpGs were neither identified as buccal or brain mQTLs. Especially for these CpGs our correlation map may be a useful tool for the interpretation of analysis results. Furthermore, even though specific genetic variants show association with DNAm levels at specific CpGs, for many sites there may still be components of variance in DNAm that are not explained by genetic variation, making a direct assessment of DNAm at these positions potentially useful and informative. Fifth, most of the CpGs highlighted by this study as significantly



correlated between buccal and brain samples are characterized by low Spearman rank correlation coefficients, starting at 0.28. To investigate CpGs showing stronger correlations, we re-evaluated our results using a more stringent threshold, i.e. using correlation coefficients smaller than  $-0.5$  or larger than  $0.5$ . This resulted in 4070 CpGs showing significant correlations of DNAm-values between brain and buccal samples. These CpGs displayed similar distributions regarding mQTLs (Additional file 2: Fig. 5), comparison with previous studies (Additional file 2: Fig. 6), and gene region enrichment (Additional file 2: Fig. 7), as compared to the full set of 24,980 CpGs. Lastly, despite stringent QC of the DNAm, including adjustment for covariates that may have had a substantial influence on DNAm (Methods) we cannot exclude that some correlations reported in this study are the result of some unknown and undetected confounding. However, given the pair-wise design of our study where we used matched buccal and brain samples from the same individuals, it appears unlikely that undetected confounding has led to a substantial and systematic inflation of our results. Notwithstanding, future studies are needed to verify and replicate the findings we present here, ideally using single-cell DNAm assessments in sufficiently sized samples.

In summary, our study on genome-wide DNAm patterns in paired buccal and brain samples highlighted  $\sim 25$  K sites showing significant correlations in DNAm levels across both tissues. To our knowledge, our study is the largest buccal–brain DNAm correlation map available to date and will hopefully prove to be a valuable resource for the interpretation of EWAS/DNAm studies for brain-related phenotypes. To this end, we made all of our genome-wide DNAm correlation results freely available online ([https://www.liga.uni-luebeck.de/buccal\\_brain\\_correlation\\_results/](https://www.liga.uni-luebeck.de/buccal_brain_correlation_results/)).

## Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s13148-022-01357-w>.

**Additional file 1.** Supplementary Tables.

**Additional file 2.** Supplementary Figures.

## Acknowledgements

We acknowledge the high-performance compute environment (“OmicsCluster”) at University of Lübeck where most data processing and analysis steps of this study were run.

## Author contributions

LB contributed to the design of the study, supervision, and acquisition of funding. DHO, BTH, and ID were involved in the ascertainment of buccal tissue. DHO and BTH contributed to the ascertainment of brain tissue and neuropathological examinations. VD, TW, SSS, and AF helped in handling of tissue samples and generation of molecular data. YS, OO, and CML contributed to data processing and statistical analyses. YS and LB were involved in the first

draft of the manuscript. All authors contributed to the critical revision and final version of manuscript. All authors read and approved the final manuscript.

## Funding

Open Access funding enabled and organized by Projekt DEAL. This work was supported by the Cure Alzheimer’s Fund (as part of the “CIRCUITS” consortium) to LB, Deutsche Forschungsgemeinschaft (DFG) and the National Science Foundation China (NSFC) as part of the Joint Sino-German research project (“MiRNet-AD”; #391523883) to LB, and by the EU Horizon 2020 Fund (as part of the “Lifebrain” consortium, #732592) to LB. The BASE-II research project (Co-PIs: Lars Bertram, Ilja Demuth, Denis Gerstorff, Ulman Lindenberger, Graham Pawelec, Elisabeth Steinhagen-Thiessen, and Gert G. Wagner) is supported by the German Federal Ministry of Education and Research (Bundesministerium für Bildung und Forschung, BMBF) under grant numbers #16SV5536K, #16SV5537, #16SV5538, #16SV5837, #01UW0808, 01GL1716A, and 01GL1716B. The Massachusetts Alzheimer’s Disease Research Center is supported by the National Institute on Aging NIA (Grant P30AG062421). CML is supported by the Heisenberg Program of the DFG (LI 2654/4-1).

## Availability of data and materials

All results pertaining to the buccal–brain DNAm correlation analyses are freely available via [http://www.liga.uni-luebeck.de/buccal\\_brain\\_correlation\\_results/](http://www.liga.uni-luebeck.de/buccal_brain_correlation_results/). The DNAm raw data generated and used for the analyses described in this manuscript are available to qualified researchers and qualified research projects.

## Declarations

### Ethics approval and consent to participate

MADRC: Sample extraction at the time of autopsy was done according to an IRB-approved informed consent with specific inclusion of genetic studies. Consent forms were completed by next-of-kin or other legal representatives as specified by Massachusetts state law.

BASE-II: All participants gave written informed consent. The medical assessments at baseline and follow-up were conducted in accordance with the Declaration of Helsinki and approved by the Ethics Committee of the Charité–Universitätsmedizin Berlin (approval numbers EA2/029/09 and EA2/144/16).

### Consent for publication

Not applicable.

### Competing interest

BTH has a family member who works at Novartis, and owns stock in Novartis; he serves on the SAB of Dewpoint and owns stock. He serves on a scientific advisory board or is a consultant for AbbVie, Avrobio, Axon, Biogen, BMS Cell Signaling, Genentech, Ionis, PPF, Novartis, Seer, Takeda, the US Dept of Justice, Vigil, Voyager. His laboratory is supported by Sponsored research agreements with Abbvie, F Prime, and research grants from the National Institutes of Health, Cure Alzheimer’s Fund, Tau Consortium, and the JPB Foundation. The other authors declare no conflict of interest.

### Author details

<sup>1</sup>Lübeck Interdisciplinary Platform for Genome Analytics (LIGA), University of Lübeck, Ratzeburger Allee 160, Haus V50, 1st Floor, Room 319, 23562 Lübeck, Germany. <sup>2</sup>Department of Pathology, Massachusetts General Hospital, Harvard Medical School, Charlestown, MA, USA. <sup>3</sup>Institute of Clinical Molecular Biology, Christian-Albrechts-University of Kiel, Kiel, Germany. <sup>4</sup>Charité – Universitätsmedizin Berlin, Freie Universität Berlin, Humboldt-Universität zu Berlin, Berlin, Germany. <sup>5</sup>Division of Lipid Metabolism, Department of Endocrinology and Metabolic Diseases, Berlin Institute of Health, Berlin, Germany. <sup>6</sup>BCRT - Berlin Institute of Health Center for Regenerative Therapies, Charité – Universitätsmedizin Berlin, Berlin, Germany. <sup>7</sup>Department of Neurology, Massachusetts General Hospital, Harvard Medical School, Charlestown, MA, USA. <sup>8</sup>Massachusetts Alzheimer’s Disease Research Center, Charlestown, MA, USA. <sup>9</sup>Harvard Medical School, Boston, MA, USA. <sup>10</sup>Ageing Epidemiology Unit (AGE), School of Public Health, Imperial College London, London, UK. <sup>11</sup>Department of Psychology, Center for Lifespan Changes in Brain and Cognition (LCBC), University of Oslo, Oslo, Norway. <sup>12</sup>Institute of Epidemiology and Social Medicine, University of Münster, Münster, Germany.

Received: 9 December 2021 Accepted: 17 October 2022  
Published online: 01 November 2022

## References

- Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, Tonti-Filippini J, et al. Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature*. 2009;462(7271):315–22.
- Sadeh N, Spielberg JM, Logue MW, Wolf EJ, Smith AK, Lusk J, et al. SKA2 methylation is associated with decreased prefrontal cortical thickness and greater PTSD severity among trauma-exposed veterans. *Mol Psychiatry*. 2015;21(3):357–63.
- Jia T, Chu C, Liu Y, van Dongen J, Papastergios E, Armstrong NJ, et al. Epigenome-wide meta-analysis of blood DNA methylation and its association with subcortical volumes: findings from the ENIGMA Epigenetics Working Group. *Mol Psychiatry*. 2019;2019(6):1–12.
- van Dongen J, Zilhão NR, Sugden K, Heijmans BT, t' Hoen PAC, van Meurs J, et al. Epigenome-wide association study of attention-deficit/hyperactivity disorder symptoms in adults. *Biol Psychiatry*. 2019;86(8):599–607.
- Hannon E, Dempster EL, Mansell G, Burrage J, Bass N, Bohlken MM, et al. Dna methylation meta-analysis reveals cellular alterations in psychosis and markers of treatment-resistant schizophrenia. *Elife*. 2021;1(10):1–53.
- McCartney DL, Hillary RF, Banos DT, Gadd DA, Walker RM, Nangle C, et al. Blood-based epigenome-wide analyses of cognitive abilities. *medRxiv*. 2021;23:1–16.
- Marioni RE, McRae AF, Bressler J, Colicino E, Hannon E, Li S, et al. Meta-analysis of epigenome-wide association studies of cognitive abilities. *Mol Psychiatry*. 2018;23(11):2133–44.
- Sommerer Y, Dobricic V, Schilling M, Ohlei O, Sabet SS, Wesse T, et al. Entorhinal cortex EWAS meta-analysis highlights four novel loci showing differential methylation in Alzheimer's disease. *bioRxiv*. 2021;02:450878.
- Smith RG, Pishva E, Shireby G, Smith AR, Roubroeks JAY, Hannon E, et al. A meta-analysis of epigenome-wide association studies in Alzheimer's disease highlights novel differentially methylated loci across cortex. *Nat Commun*. 2021;12(1):1–13.
- Nabais MF, Laws SM, Lin T, Vallerga CL, Armstrong NJ, Blair IP, et al. Meta-analysis of genome-wide DNA methylation identifies shared associations across neurodegenerative disorders. *Genome Biol*. 2021;22(1):1–30.
- Lowe R, Gemma C, Beyan H, Hawa MI, Bazeos A, David Leslie R, et al. Buccals are likely to be a more informative surrogate tissue than blood for epigenome-wide association studies. *Epigenetics*. 2013;8(4):445–54.
- Smith AK, Kilaru V, Klengel T, Mercer KB, Bradley B, Conneely KN, et al. DNA extracted from saliva for methylation studies of psychiatric traits: Evidence tissue specificity and relatedness to brain. *Am J Med Genet Part B Neuropsychiatr Genet*. 2015;168(1):36–44.
- Hannon E, Mansell G, Walker E, Nabais MF, Burrage J, Kepa A, 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.
- Teschendorff AE, Breeze CE, Zheng SC, Beck S. A comparison of reference-based algorithms for correcting cell-type heterogeneity in Epigenome-Wide Association Studies. *BMC Bioinforma*. 2017;18(1):1–14.
- Edgar RD, Jones MJ, Meaney MJ, Turecki G, Kobor MS. BECon: a tool for interpreting DNA methylation findings from blood in the context of brain. *Transl Psychiatry*. 2017;7(8):e1187–e1187.
- Hannon E, Lunnon K, Schalkwyk L, Mill J. Interindividual methylomic variation across blood, cortex, and cerebellum: Implications for epigenetic studies of neurological and neuropsychiatric phenotypes. *Epigenetics*. 2015;10(11):1024–32.
- Braun PR, Han S, Hing B, Nagahama Y, Gaul LN, Heinzman JT, et al. Genome-wide DNA methylation comparison between live human brain and peripheral tissues within individuals. *Transl Psychiatry*. 2019;9(1):1–10.
- Braun P, Hafner M, Nagahama Y, Hing B, McKane M, Grossbach A, et al. Genome-wide Dna methylation comparison between live human brain and peripheral tissues within individuals. *Eur Neuropsychopharmacol*. 2017;1(27):S506.
- Gunasekara CJ, Scott CA, Laritsky E, Baker MS, MacKay H, Duryea JD, et al. A genomic atlas of systemic interindividual epigenetic variation in humans. *Genome Biol*. 2019;20(1):1–12.
- Gorrie-Stone TJ, Smart MC, Saffari A, Malki K, Hannon E, Burrage J, et al. Bigmelon: tools for analysing large DNA methylation datasets. *Bioinformatics*. 2019;35(6):981–6.
- Zhou W, Laird PW, Shen H. Comprehensive characterization, annotation and innovative use of Infinium DNA methylation BeadChip probes. *Nucleic Acids Res*. 2017;45(4): e22.
- Nordlund J, Bäcklin CL, Wahlberg P, Busche S, Berglund EC, Eloranta ML, et al. Genome-wide signatures of differential DNA methylation in pediatric acute lymphoblastic leukemia. *Genome Biol*. 2013;14(9):1–15.
- Tian Y, Morris TJ, Webster AP, Yang Z, Beck S, Feber A, et al. ChAMP: updated methylation analysis pipeline for Illumina BeadChips. *Bioinformatics*. 2017;33(24):3982–4.
- Aryee MJ, Jaffe AE, Corrada-Bravo H, Ladd-Acosta C, Feinberg AP, Hansen KD, et al. Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics*. 2014;30(10):1363–9.
- Teschendorff AE, Menon U, Gentry-Maharaj A, Ramus SJ, Gayther SA, Apostolidou S, et al. An epigenetic signature in peripheral blood predicts active ovarian cancer. *PLoS ONE*. 2009;4(12): e8274.
- Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. Limma powers differential expression analyses for RNA-seq and microarray studies. *Nucleic Acids Res*. 2015;43(7):e47–e47.
- Demuth I, Banszerus V, Drewelies J, Düzel S, Seeland U, Spira D, et al. Cohort profile: follow-up of a Berlin Aging Study II (BASE-II) subsample as part of the GendAge study. *BMJ Open*. 2021;11(6): e045576.
- Bertram L, Böckenhoff A, Demuth I, Düzel S, Eckardt R, Li SC, et al. Cohort profile: the berlin aging study II (BASE-II). *Int J Epidemiol*. 2014;43(3):703–12.
- Hong S, Dobricic V, Ohlei O, Bos I, Vos SJB, Prokopenko D, et al. TMEM106B and CPOX are genetic determinants of cerebrospinal fluid Alzheimer's disease biomarker levels. *Alzheimer's Dement*. 2021;17(10):1628–40.
- Shabalin AA. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics*. 2012;28(10):1353–8.
- Phipson B, Maksimovic J, Oshlack A. missMethyl: an R package for analyzing data from Illumina's HumanMethylation450 platform. *Bioinformatics*. 2016;32(2):286–8.
- Supek F, Bošnjak M, Škunca N, Šmuc T. REVIGO summarizes and visualizes long lists of gene ontology terms. *PLoS ONE*. 2011;6(7): e21800.
- Perzel Mandell KA, Eagles NJ, Wilton R, Price AJ, Semick SA, Collado-Torres L, et al. Genome-wide sequencing-based identification of methylation quantitative trait loci and their role in schizophrenia risk. *Nat Commun*. 2021;12(1):1–12.
- Goel N, Karir P, Garg VK. Role of DNA methylation in human age prediction. *Mech Ageing Dev*. 2017;1(166):33–41.
- Davies G, Armstrong N, Bis JC, Bressler J, Chouraki V, Giddaluru S, et al. Genetic contributions to variation in general cognitive function: a meta-analysis of genome-wide association studies in the CHARGE consortium (N = 53 949). *Mol Psychiatry*. 2015;20(2):183–92.
- Horvath S. DNA methylation age of human tissues and cell types. *Genome Biol*. 2015;16(1):96.
- Ernst C, McGowan PO, Deleval V, Meaney MJ, Szyf M, Turecki G. The effects of pH on DNA methylation state: in vitro and post-mortem brain studies. *J Neurosci Methods*. 2008;174(1):123–5.

## Publisher's Note

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