Briefings in Bioinformatics, 2025, **26(4)**, bbaf427
https://doi.org/10.1093/bib/bbaf427
Problem Solving Protocol

Guidance for the design and analysis of cell-type-specific DNA methylation epidemiology studies

Emma M. Walker[1], Emma L. Dempster[1], Alice Franklin[1], Anthony Klokkaris[1], Barry Chioza[1], Jonathan P. Davies[1], Georgina E.T. Blake[1], Joe Burrage[1], Stefania Policicchio[2], Rosemary A. Bamford[1], Leonard C. Schalkwyk[3], Jonathan Mill[1], Eilis Hannon[1,*]

[1]Department of Clinical and Biomedical Sciences, University of Exeter Medical School, University of Exeter, Barrack Road, Exeter, Devon, EX2 5DW, United Kingdom
[2]Italian Institute of Technology Center for Human Technologies (CHT), Via Enrico Melen, 83, 16152 Genova GE, Italy
[3]School of Life Sciences University of Essex, Wivenhoe Park, Colchester, Essex, CO4 3SQ, United Kingdom
[*]Corresponding author. Department of Clinical and Biomedical Sciences, University of Exeter Medical School, RILD building, Royal Devon & Exeter Hospital, Barrack Road, Exeter EX2 5DW, UK. E-mail: E.J.Hannon@exeter.ac.uk

Abstract

Recent studies on the role of epigenetics in disease have focused on DNA methylation (DNAm) profiled in bulk tissues limiting the detection of the cell type affected by disease-related changes. Advances in isolating homogeneous populations of cells now make it possible to identify DNAm differences associated with disease in specific cell types. Critically, these datasets will require a bespoke analytical framework that can characterize whether the difference affects multiple or is specific to a particular cell type. We take advantage of a large set of DNAm profiles ($n = 751$) obtained from five different purified cell populations isolated from human prefrontal cortex samples and evaluate the effects on study design, data preprocessing, and statistical analysis for cell-specific studies, particularly for scenarios where multiple cell types are included. We describe novel quality control metrics that confirm successful isolation of purified cell populations, which when included in standard preprocessing pipelines provide confidence in the dataset. Our power calculations show substantial gains in detecting differentially methylated positions for some purified cell populations compared to bulk tissue analyses, countering concerns regarding the feasibility of generating large enough sample sizes for informative epidemiological studies. In a simulation study, we evaluated different regression models finding that this choice impacts on the robustness of the results. These findings informed our proposed two-stage framework for association analyses. Overall, our results provide guidance for cell-specific epigenome-wide association studies, establishing standards for study design and analysis, while showcasing the potential of cell-specific DNAm analyses to reveal links between epigenetic dysregulation and disease.

Keywords: epigenetic; epidemiology; cell-specific; DNA methylation

Introduction

Recent years have seen increased attention to the role of genetics and gene regulation in development and complex disease [1, 2] facilitated by sequencing and array-based technologies. This includes studies of the epigenome, which encompasses a diverse number of chemical modifications to DNA and nucleosomal histone proteins that directly influence gene expression. In contrast to the genome, the epigenome is highly dynamic, varying across development, between cell types, and in response to the environment. Consequently, this means that careful consideration of study design is required when investigating relationships between epigenetic variation and complex traits [3, 4]. The most studied epigenetic modification in the context of human health and disease is DNA methylation (DNAm) [5, 6], which involves the addition of a methyl group to a cytosine. Most existing association studies leverage the high-throughput nature of microarrays to profile DNAm at hundreds of thousands of positions across the genome, meaning it is feasible both practically and financially to identify disease-associated variation for large sample numbers.

It is well-established that a DNAm profile is primarily defined by tissue or cell type [7–11]. Therefore, the choice of tissue for profiling influences both the analytical results obtained and the nature of conclusions that can be drawn from these. Profiling the primary affected tissue, for example, blood for immune-related traits and the prefrontal cortex (PFC) for neuropsychiatric and neurodegenerative diseases, is pertinent to making mechanistic inferences about how DNAm variation contributes to a particular trait. However, even these 'bulk' tissues represent a heterogeneous mix of cell types. Each of these cell types has its own DNAm profile, with the resulting profile of the bulk tissue being an aggregate of those from the constituent cell types. As the proportion of each cell type within a sample can vary across individuals, systematic differences in cellular proportions that correlate with the phenotype of interest may manifest as differences in the overall DNAm profile [12]. To minimize potential false positive associations, quantitative covariates that capture the cellular composition of each sample are typically included in statistical analyses [12]. The major caveat of analyzing DNAm in bulk tissue is that it does not enable the identification of

Received: November 06, 2024. **Revised:** May 19, 2025. **Accepted:** June 18, 2025

which cell types are affected by the detected differences. In addition, subtle changes or differences in rarer cell types may be missed as they compete against the background signal from more abundant cell types. Elucidating which cell type(s) are affected by DNAm differences is critical for determining the genes and biological processes associated with specific complex traits and ultimately identifying novel targets for preventing and treating disease.

The case for generating cell-specific DNAm profiles to facilitate cell-specific analyses of epigenetic variation in disease is compelling. Although methods for single-cell DNAm profiling have been developed [13–15], these approaches are not currently amenable to large-scale analyses of human disease. Instead, techniques such as fluorescence activated nuclei sorting (FANS) and laser capture microdissection can be used to isolate purified cell populations from bulk tissue prior to genome-wide assays and have been applied to tissues such as whole blood [9, 16] and cortex [17–19]. Many datasets generated from these methods are small, aimed primarily at generating reference profiles for characterization of those cell types or as input for reference-based deconvolution algorithms that estimate the cellular composition from bulk tissue profiles [20]. There are, however, cases where these data have been used for epigenetic epidemiology to identify variation in DNAm associated with disease [19, 21–25].

Although the primary tissue for a particular disease may be obvious, the specific cell type involved is often less clear with cell-specific epigenome-wide association studies (EWAS) typically including multiple isolated cell types. In these scenarios, the objective differs from the traditional bulk tissue EWAS. Not only is the goal to identify loci in the genome where differences in DNAm correlate with the outcome of interest, but also to characterize whether the difference manifests across multiple cell types or is specific to a particular cell type. This will require a change in analytical approach because existing statistical approaches based on standard linear regression are potentially inadequate for analyzing these data. First, unlike most traditional bulk tissue analyses, multiple samples will be profiled for each individual. This threatens a major assumption of linear regression, that the observations are independent. Second, the statistical framework must capture differences between cases and controls that could be present in all cell types or are only present in a subset of cell types, by estimating case-control differences per cell type and assess whether these are statistically consistent.

In this manuscript, we evaluate the effects on study design, data preprocessing, and statistical analysis for cell-specific studies of DNAm, particularly where multiple cell types are considered, by taking advantage of a large DNAm dataset including five different purified cell populations isolated from PFC. First, we describe necessary extensions to established quality control pipelines that ensure that the isolation of purified cell populations has been successful. Second, we investigate the effect on statistical power of an association study that uses cell-specific DNAm data. Third, we assess the impact of how the data are normalized on the variance in DNAm. Finally, we evaluate multiple statistical frameworks for analyzing cell-specific DNAm data, using two simulation scenarios: a null association study and an association study with differentially methylation positions introduced. With these results, we provide guidance for the field in anticipation of future cell-specific EWAS with the objective of establishing standards for how these studies are designed, analyzed, and interpreted.

Methods

Isolation of neural nuclei from post-mortem brain tissue

Post-mortem tissue from 287 adult donors (aged 18–108 years old) was provided from multiple international brain banks (Cambridge, Edinburgh, Stanley, King's College London, Harvard, UCLA, Oxford, Miami, Douglas Bell, Pittsburgh and Mount Sinai Brain Banks). Tissue was collected under approved ethical regulation at each centre and transferred through Materials Transfer Agreements. Post-mortem human PFC samples were processed using a FANS protocol developed by our group [26]. The gating strategies we implemented are shown in Supplementary Fig. S1.

Methylomic profiling

DNA was extracted from frozen nuclei aliquots using a modified proteinase K-based extraction method developed specifically for these sample types [27]. 500 ng of genomic DNA from each sample was treated with sodium bisulfite using the Zymo EZ-96 DNA Methylation-Gold™ Kit (Cambridge Bioscience, UK) according to the manufacturer's standard protocol. All samples were then processed using the EPIC 850K array (Illumina Inc, CA, USA) according to the manufacturer's instructions, with minor amendments and quantified using an Illumina iScan System (Illumina, CA, USA). All sorted fractions from two randomly selected individuals were assigned to each BeadChip, within which the location of each fraction was randomized. Altogether, 293 Total (unsorted), 290 NeuNPos (NeuN+), 286 SOX10Pos (NeuN-/SOX10+), 27 IRF8Pos (NeuN-/SOX10-/IRF8+), 274 DoubleNeg (NeuN-/SOX10-), and 15 TripleNeg (NeuN-/SOX10-/IRF8-) nuclei samples were processed.

DNAm data preprocessing and quality control

DNAm data was loaded into R (version 3.6.3) from IDAT files using the package bigmelon [28]. These data were processed through a bespoke quality control pipeline developed for cell-specific DNAm data. The pipeline is structured into three stages:

Stage 1—confirming the quality of the DNAm data.
Stage 2—confirming the correct individual.
Stage 3—confirming the correctly labelled cell type.

The first two stages are common to most existing preprocessing pipelines. The final stage of the pipeline confirms that FANS isolation was successful and fractions representing distinct cell types were obtained. It leverages the fact that as cell-type identity is the primary source of variation in DNAm profiles [8, 10, 29], the major principal components (PCs) should cluster the samples by cell type. Using the first two PCs, the distance (in SD units) between each sample and the mean profile of its labelled cell type is used to identify instances where either the FANs isolation was not successful (and heterogeneous samples were collected) or where samples have potentially been mislabelled. Full details on the steps within each stage can be found in Supplementary Text S1.

After stringent quality control, 751 samples were retained: 218 Total, 164 DoubleNeg, 182 NeuNPos, 168 Sox10Pos, 12 IRF8Pos, and 7 TripleNeg samples.

Comparison of normalization strategies

This analysis was limited to cell types that had more than 100 samples (NeuNPos, Sox10Pos, and DoubleNeg), which were normalized using the *dasen*() function in the wateRmelon R package [30] in two ways:

(i) All samples from all cell types were normalized as a single dataset; and
(ii) Normalization was performed for each cell type separately.

These strategies were compared using three quantitative metrics proposed in the original wateRmelon manuscript [30] where, lower scores indicate a higher signal-to-noise ratio and a more effective normalization strategy (further details in Supplementary Text S2). In addition, we quantified the magnitude of transformation per sample, using the *qual*() function from the bigmelon package [28].

Power calculations

We conducted power calculations to assess the impact of isolating purified nuclei populations on statistical power, using the function *pwr.t.test()* from the R package pwr [31]. We consider the scenario with a binary outcome (i.e. case control study), using a two-sample *t*-test to compare the means of the two groups. We profiled the effect of varying sample size or mean difference between groups on statistical power for each cell type separately, using cell-specific SDs for each site. The significance level was set to an experiment-wide threshold of $P < 9 \times 10^{-8}$ [32].

To get an overall power estimate for a specific scenario for study with 846 232 autosomal sites, we calculated the cumulative percentage of sites that had a minimum level of statistical power. To reduce the computational burden of performing >800 000 calculations per cell type, we leveraged the fact that many sites have similar SDs. Sites were assigned to 500 bins of equal size (i.e. same number of sites per bin) and containing sites with similar SDs. The mean of the SDs across sites within each bin is used to calculate Cohen's d for the power calculation and the resulting power statistic is applied to all sites in that bin. When analyzing how the mean difference affected statistical power, we fixed the sample size at either 100 or 200 per group and evaluated a range of increasing mean differences (0.001, 0.002, 0.003, …, 0.1). When analyzing how sample size affected statistical power, we fixed the mean difference between groups (0.02 or 0.05) and evaluated 100 equally spaced sample sizes ranging from 0 up to the sample size needed to detect a minimum of 85% of sites with at least 80% power, rounded to the nearest 100. Note that in these results, we report the mean difference for DNAm measured as percentage points (i.e. bounded between 0 and 100) rather than as a proportion (i.e. bounded between 0 and 1).

Simulation study to assess statistical frameworks for cell-specific EWAS

To assess the different analytical frameworks, we implemented two simulation scenarios. First, we generated 100 null association studies by randomly assigning samples as either 'cases' or 'controls'. Second, taking each simulated null association study, we introduced a fixed number of differentially methylated positions (DMPs) at a random subset of sites, predetermined to be either common (i.e. affect all cell types) or specific to a single cell type. Further details can be found in Supplementary Text S3.

For each simulation, we compared four regression frameworks:

(i) Linear regression model for each cell type separately (within cell-type linear regression 'ctLR'). As there is only one sample per individual per cell type, the observations can be considered to be independent.
(ii) Linear regression model using all samples from all cell types ('allLR'). As there are multiple samples per individual, this approach is potentially biased.
(iii) Mixed effects linear regression model using all samples from all cell types (mixed effects regression 'MER') where a random intercept accounts for multiple samples per individual.
(iv) Clustered robust regression (CRR) model using all samples from all cell types ('CRR'), where samples from each individual are modelled as a cluster to account for the lack of independence.

All models included age, sex, and brain bank as covariates. Further details can be found in Supplementary Text S4. For each simulation and regression model, we recorded the number of significant associations at epigenome-wide significance (9×10^{-8}) [32] as well as three discovery thresholds (1×10^{-7}, 1×10^{-6}, and 1×10^{-5}), classifying these as either false positives or true positives.

Results

Novel quality control pipeline for processing cell-specific DNA methylation data

To ensure confidence and reliability in cell-sorted DNAm data, we developed a three-stage custom pipeline (Fig. 1; see Supplementary Text S1). Stage 1 confirms that the assay has generated high quality DNAm data. Stage 2 confirms that the sample matches their labelled individual by checking concordance with sex and genotype information. Stage 3 is tailored to the cell-sorted dimension of the dataset to confirm that FANS isolation was successful. For this, two novel metrics based on PCs were developed that quantify how closely each sample matches the average profile for its labelled cell type, using PCs. First, an individual level 'isolation efficiency score' was defined to verify that distinct fractions of nuclei were isolated for each FANs sort. Visual inspection of this score applied to an exemplar dataset determined a threshold of 5 was appropriate to identify individuals where sorting was unsuccessful and heterogeneous samples were collected (Supplementary Fig. S2). Second, we calculated the distance between each sample and its labelled cell type, retaining only those that were within two SDs of the cell type mean for the first two PCs.

Increased sensitivity for detecting differentially methylated positions following normalization within cell types

Normalization is used to minimize experimental technical variation by transforming samples to a common standard, making them quantitatively comparable and maximizing the statistical sensitivity. While studies report little impact of normalization strategy on DMP detection in EWAS of single sample types [33, 34], dramatic differences in DNAm profiles between cell types may lead to unpredictable behaviour if these data are forced to become more similar to each other. We compared the effect of normalizing across all samples from all cell types together to normalizing separately within each cell type using an established framework [30]. Across three metrics, where lower scores indicate better performance, normalizing the data separately for each cell type was the optimal approach (Table 1). Of note, NeuNPos samples are subject to a much larger transformation (mean across sites = 0.039, SD = 0.016), when normalized with the other cell types, compared to when they are normalized only with NeuNPos samples (mean = 0.027; SD = 0.011; Supplementary Fig. S3). In contrast, DoubleNeg and Sox10Pos samples were associated with similar levels of transformation when normalized separately by cell

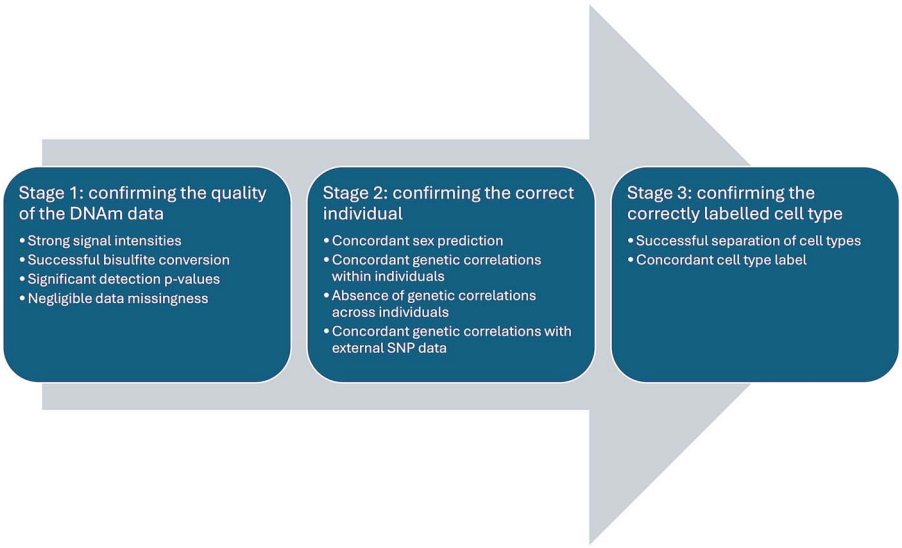

Figure 1. Overview of stages in quality control pipeline for cell-specific epigenetic data.

Table 1. Summary metrics to compare normalization strategies for cell-sorted DNAm data, where each column presents the results of a different metric proposed by Pidsley *et al.* to quantitative compare different normalization algorithms.

Normalization strategy	Metric		
	DMRSE Imprinted regions	GCOSE SNP probes	Seabird X-chromosome inactivation
Raw unnormalized data	1.578E-03	5.724E-05	0.053
Normalize all samples from all cell types together	1.600E-03	5.716E-05	0.037
Normalize each cell type separately	1.243E-03	5.107E-05	0.036

type (DoubleNeg mean = 0.033; SD = 0.013; Sox10Pos mean = 0.027; SD = 0.012) compared to when normalized altogether (DoubleNeg mean = 0.032; SD = 0.014; Sox10Pos mean = 0.026; SD = 0.011). This is consistent with the knowledge that the primary axis of variation in sorted PFC samples captures differences between NeuN positive and NeuN negative samples [29] meaning these samples require the largest manipulation to make them comparable.

Isolating homogeneous populations of cells reduces the number of samples required to reliably detect differentially methylated positions

It is often assumed that cell sorting is impractical for EWAS due to the large sample sizes required but extrapolating from power calculations derived from heterogeneous bulk tissues may be misleading. To test the hypothesis that purified cell populations have lower variance translating to increased power to detect cell-type-specific differences, we quantified probe level variation by cell type. Comparing the distribution of SDs across autosomal sites by cell type (Supplementary Fig. S4), lower variation across samples was observed as expected for the NeuNPos (mean = 0.036; SD = 0.023) and Sox10Pos (mean = 0.038; SD = 0.024) fractions relative to the Total samples (mean = 0.046; SD = 0.026). Total contains nuclei from all fractions and therefore can be considered a proxy for bulk PFC tissue. Of note the DoubleNeg (mean = 0.057; SD = 0.033) and IRF8Pos (mean = 0.055; SD = 0.036) fractions exhibit on average higher levels of variation with the TripleNeg (mean = 0.044; SD = 0.029) fraction showing similar levels of variation to Total.

The reduced variation naturally has a positive effect on power and sample size requirements (Fig. 2). For example, to detect at mean difference of 5% points between groups with power of 80% at 80% of probes on the EPIC array, you need 84 samples per group for NeuNPos or 88 samples per group for Sox10Pos (Supplementary Table S1). In contrast, 132 Total, 128 TripleNeg, 204 IRF8Pos, or 212 DoubleNeg samples per group are required. Instead to detect a 2% difference, you need 476 samples per group for NeuNPos compared to 799 samples for Total or 1224 samples per group for IRF8Pos. Alternatively keeping the sample size constant at 100 samples per group, 80% power is obtained at 80% of sites to detect a mean difference of 4.5% for NeuNPos, 4.7% for Sox10Pos, 5.7% for TripleNeg, 5.8% for Total, 7.3% for IRF8Pos, and 7.5% for DoubleNeg between groups. This decreases to a mean difference of 3.1% for NeuNPos, 3.3% for Sox10Pos, 4.0% for TripleNeg, 4.1% for Total, 5.1% for IRF8Pos, and 5.2% for DoubleNeg between groups if the sample size is increased to 200 samples per group.

Cell-specific EWAS are prone to false positives if within individual study design is not adequately controlled for

Cell-specific EWAS require an analytical framework that can account for multiple samples per individual and model potentially different magnitudes of disease associated variation across cell types. To evaluate the impact of different regression models on statistical sensitivity, we designed a simulation framework to compare four approaches. The first approach ('ctLR') is linear regression within each cell type, where there is no correlation

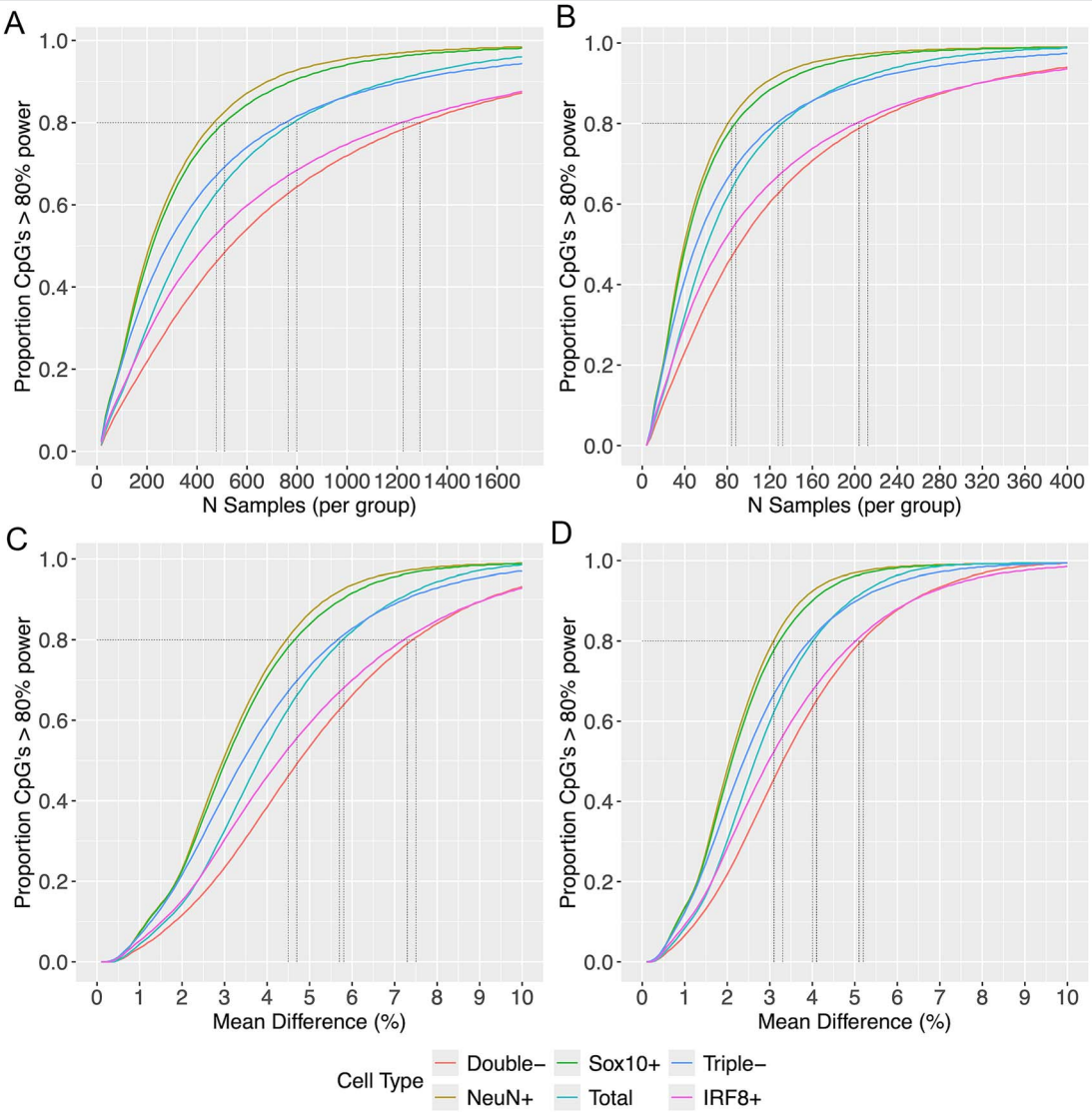

Figure 2. Cell-specific epigenetic association analyses increase statistical power over bulk tissue studies, as illustrated by power curves for purified brain cell types and bulk brain tissue across different experimental parameters, with Panels A and B showing power (y-axis) versus sample size for mean group differences of 5% and 2%, respectively, and Panels C and D showing power versus mean difference for fixed sample sizes of 100 and 200 cases versus controls, respectively, where each line represents a brain cell type or bulk tissue.

between samples to control for. It is a computationally cheap analysis per DNAm site but involves fitting one model per cell type and is unable to determine the cell-specificity of the effect. The second approach ('allLR') uses all the data available from all cell types in a linear regression model. This method violates the assumption of independent observations but by aggregating the data together could increase power to detect DMPs shared across cell types. The third approach ('MER') uses a mixed effects regression model to control for the structure within the data. The final approach ('CRR') uses a clustered robust regression model, an alternative approach that can account for related samples. The allLR, MER, and CRR methods all include both a main effect and interaction terms to capture DMPs with different impacts across cell types.

When deciding upon an efficient method, there are two desirable properties: minimizing the false positive rate and maximizing the true positive rate. To quantify the false positive rate for each approach, we simulated 100 null EWAS (see Supplementary Text S2). For regression models that include all data from all samples (allLR, MER, and CRR), depending on which cell type(s) a DMP affects, both or either of the main effect and interaction term may be significant, so both terms must be considered when assessing these methods. We observed that the distribution of log *P*-values for both the allLR and MER models is wider with longer tails indicative of smaller mean *P*-values (Supplementary Fig. S5). Applying the standard threshold for epigenome-wide significance (9×10^{-8} [32]), we observed that the ctLR and CRR are well calibrated with all models identifying a mean of <0.1 false positive DMPs (Fig. 3). In contrast, both the allLR and MER models had increased rates. These were dramatically inflated for the main effect term (allLR mean = 42 (SD = 136), MER mean = 53 [SD = 185]) and subtly higher also for the interaction term (allLR mean = 1 (SD = 2); MER mean = 2 (SD = 3)). This highlights that the choice of analysis method is critical and has dramatic effects on the robustness of the results. One mechanism to counteract this is to derive a significance threshold calibrated specifically for each method. Calculating the mean 5% family-wise error rate for each method across the null simulations (Supplementary Fig. S6), the significance thresholds for the ctLR and CRR models were comparable to the standard threshold

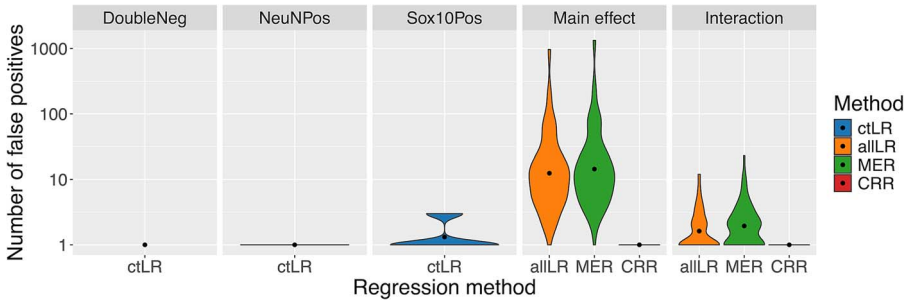

Figure 3. Choice of regression method is critical to minimize false positives in cell-specific association analyses, as shown by violin plots (y-axis: number of false positive DMPs on a log scale, $P < 9 \times 10^{-8}$) from 100 simulated null EWAS, where each violin represents a different statistical analysis, colored by regression method—ctLR (within cell-type linear regression), allLR (linear regression across all cell types), MER (mixed effects regression), and CRR (clustered robust regression).

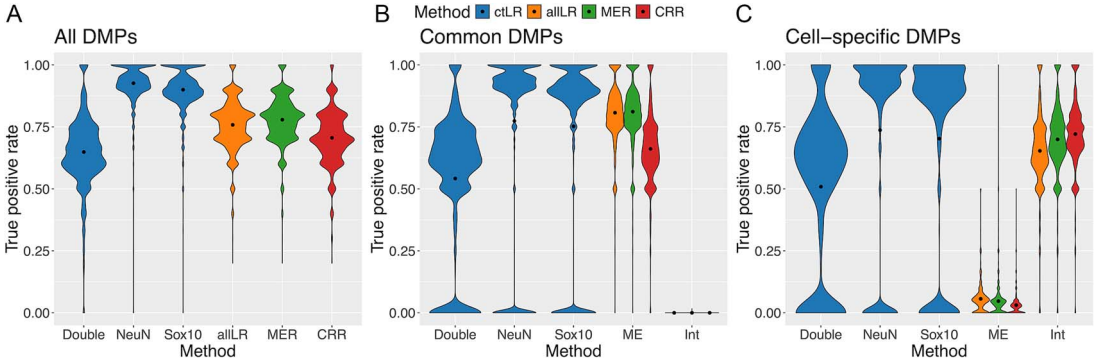

Figure 4. Violin plots compare true positive rates ($P < 9 \times 10^{-8}$) across cell-specific regression methods using results from 1,800 simulated EWAS in which 10, 100, or 1,000 differentially methylated positions (DMPs) with a 5% mean group difference were introduced, with varying proportions (0–1) of DMPs designated as cell-specific versus common across all cell types, and true positive rates are shown for (A) all DMPs, (B) cell-specific DMPs, and (C) common DMPs, with each violin representing a regression method—ctLR (within cell-type linear regression), allLR (linear regression across all cell types), MER (mixed effects regression), and CRR (clustered robust regression).

(range: $P = 9.7 \times 10^{-8}$–2.8×10^{-7}), whereas the thresholds for the allLR and MER models needed to be 2–7 orders of magnitude smaller (range: $P = 9.7 \times 10^{-16}$–5.3×10^{-9}).

To assess each method's accuracy at detecting true positive associations, we implemented a second simulation scenario where a fixed number of DMPs were introduced with effects that were either common to all three cell types or specific to just one. Considering DMPs with a mean difference of 5% between cases and controls, the ctLR in NeuNPos samples had the highest true positive rate (mean = 0.76, SD = 0.35; Fig. 4), slightly higher than the MER (mean = 0.74, SD = 0.09), ctLR in Sox10Pos samples (mean = 0.73, SD = 0.34), and allLR (mean = 0.72, SD = 0.09). There was then a drop in performance to the CRR method (mean = 0.66, SD = 0.10) followed by another drop in performance to the ctLR in DoubleNeg samples (mean = 0.49, SD = 0.26). The disparity in performance with the ctLR method for different cell types fits with the results of the power calculations, highlighting how the heterogeneity of DoubleNeg samples impacts the ability to detect genuine associations.

Generally, performance did not depend on whether a DMP was common to all cell types or specific to just one, although it is notable that the CRR model had a higher true positive rate for detecting cell-specific DMPs (mean = 0.68, SD = 0.13) than either the MER (mean = 0.65, SD = 0.14), or allLR (mean = 0.60, SD = 0.15). Regardless, these methods all performed slightly worse than the ctLR in either NeuNPos (mean = 0.72, SD = 0.36) or Sox10Pos (mean = 0.68, SD = 0.38) samples for these DMPs. The pattern of results was maintained when the mean difference between cases and controls was reduced to 2%, albeit with reduced true positive rates (Supplementary Fig. S7; range of mean true positive rates across all DMPs = 0.12–0.27). There was no meaningful effect of the number of DMPs on the performance of the regression method (Supplementary Fig. S8). Interestingly, as the proportion of cell-type-specific DMPs increased the true positive rate for the allLR and MER methods decreased and increased for the CRR method.

When EWAS fail to identify DMPs at the standard epigenome-wide significance threshold this is commonly attributed to a lack of statistical power. 'Discovery thresholds' are often used instead to uncover potential associations. However, since the validity of this approach is untested, we used our simulations to assess the potential benefits and identify a suitable discovery threshold. First, considering the number of false positives at a range of discovery thresholds (1×10^{-7}, 1×10^{-6}, and 1×10^{-5}), the methods that had minimal false positives at epigenome-wide significance, ctLR and CRR, maintained a mean of less than 1 false positive up to a threshold of $<10^{-6}$. This increased to between 4 and 10 at the most permissive P-value threshold of $<10^{-5}$ (Supplementary Table S2). Meanwhile, the methods that already had elevated mean false positive counts continued to accumulate additional false positives, increasing from 43 to 656 for allLR and from 55 to 785 for MER at the most relaxed threshold (P-value $<10^{-5}$). Secondly, considering the true positive rate, this increased as the significance threshold is relaxed for all methods, albeit to different degrees (Supplementary Table S3). There are smaller gains for the methods that were already associated with high rates, for example, ctLR in NeuNPos samples increases from 0.91 (SD = 0.08) at 9×10^{-8} to 0.95 (SD = 0.06) at 10^{-5} with bigger gains for methods that had lower rates at the most stringent threshold, e.g.,

CRR increases from 0.66 (SD = 0.10) at 9×10^{-8} to 0.78 (SD = 0.09) at 10^{-5}.

Discussion

Determining the cell types affected by differences in DNAm is the next critical step in enhancing our understanding of the role of epigenetics in health and disease. As cell-specific EWAS gain traction, now is the critical time to establish a framework that promotes statistically robust analyses facilitating the generation of meaningful and accurate biological insights. We present the first quantitative assessment of study design, quality control, and statistical analysis for cell-specific DNAm association studies. Using an exemplar dataset with five purified populations of nuclei obtained by FANS from post-mortem cortical tissue, we use both empirical power calculations and simulations to determine a statistically robust approach that identifies and characterizes the cell-specificity of positions in the genome associated with an outcome.

Balancing the need to minimize the false positive rate and maximize the true positive rate, we propose the following analytical strategy for EWAS of multiple cell types. First, we recommend performing a within cell type association analysis using linear regression. From each regression model (one per cell type), significant DMPs can be identified with low risk of false positive associations, subject to appropriate experimental design and adjustment for confounders. The caveat with this approach is that it cannot determine whether the identified DMP is cell type-specific or affects multiple cell types. Cross-referencing DMP lists across cell types may reveal overlaps, but a lack of overlap does not confirm absence of effect, as it could result from sampling variation or limited statistical power. Therefore, a second stage of the analysis is required to confirm whether the effect is common to all cell types by testing for differences or heterogeneity across cell types. We propose that a mixed effects model is used in this scenario with an interaction term and random effect to control for multiple samples per individual. While this method was associated with a high number of false positives, we would discourage using it to discover DMPs but instead to characterize the cell-type specificity of the effect. We additionally investigated the impact of using a 'discovery threshold' but this offered minimal benefit when identifying additional true positives. If required, we recommend not exceeding 10^{-6}, as this kept the number of false positives below 1.

Our motivation is informed by challenges observed in single-cell transcriptomics, where multiple replicates per sample (on a substantially greater scale) and the identification of cell-specific differences pose significant analytic challenges. Analyses of both empirical and synthetic datasets in that domain have shown that some proposed methods, including as we did mixed effects models, fail to account adequately for the variability between replicates and consequently are associated with inflated counts of false positives [35]. While single-cell sequencing is a complementary strategy for identifying cell-specific effects, it is fundamentally a different approach and currently lacks a robust way for quantifying DNAm. Profiling a pool of molecules from the same cell type is instead, more accurate and cost effective than profiling a single molecule. By positively selecting specific populations of nuclei, we ensure that sufficient material is collected for efficient quantification. Our experimental approach is therefore not only technically more robust but also more scalable and economical, making it well-suited for large cohort epigenetic epidemiology studies.

One of our key take home messages is that isolating populations of cells may not need to be performed in as many samples as would be needed for a bulk tissue EWAS. Our power calculations found that for some purified populations, there were substantial gains in statistical power for detecting DMPs. Consequently, in studies with limited samples (e.g. post-mortem brain tissue), cell-sorting could be a valuable strategy to boost statistical power and maximize the impact of a limited resource. Within specific cell types, it is anticipated that effect sizes will be greater, enhancing power further. However, for some cell-sorted fractions, such as IRF8Pos fractions (microglia enriched) and TripleNeg fractions (astrocyte enriched), there was an increase in variation and therefore a decrease in power. This suggests that within these fractions, further isolation may be required to not only increase the probability of finding genuine associations, but also to clarify which cell types show increased variation in DNAm.

The findings we present should be mindful of the following limitations. First, we only considered a case-control study design, the most commonly used study design for disease associated epigenetic differences. As the regression methods we used are easily adaptable for a broad range of phenotypes, including both continuous and categorical outcomes we are confident that our results would also hold for these other EWAS scenarios, which will also have to account for the structure in the data. Second, while our data only consists of neural purified cell populations, we believe the overarching conclusions of our study are generalizable to cell types from other tissues. For those interested in DNAm EWAS of brain cell types, we have developed an R package, CellPower, for the community to perform power calculations for brain cell-specific EWAS. Third, we only considered a limited number of parameter combinations for DMPs and sample size. We believe this is sufficient to address the questions we posed, but caution should be applied when extrapolating from the specific true positive and false positive rates we report to other EWAS scenarios. Fourth, we used a microarray to quantify DNAm with limited coverage of CpGs. There are other sequencing based technologies that can be used to profile DNAm more extensively across the genome that generate estimates with different statistical properties [36] and it may be that our results do not extend to these other technologies. All of these limitations could be addressed with follow up studies that apply our simulation framework to other study designs, or cell-specific data generated from different tissues or technologies to confirm that our findings translate beyond the scenario we explored.

Conclusion

In conclusion, our analyses provide a valuable insight into the potential of cell-specific DNAm association analyses and provide a benchmark against which future studies can be evaluated in terms of their study design, quality control pipelines and statistical analysis.

Key Points

- We propose an analytical framework for cell-specific DNAm data that incorporates bespoke quality control metrics and not only performs an association analysis to identify differentially methylation positions but that also characterizes whether it affects multiple cell types or is specific to a particular cell type.

- Consideration must be given to the most appropriate analytical model for these data, as even methods that theoretically adjust for correlations between samples, may be associated with inflated rates of false positives.
- Isolating purified cell types can lead to increases in statistical power to identify differentially methylated positions and therefore this might be an effective way of maximize the value of a limited resource such as post-mortem brain tissue samples.

Acknowledgements

We acknowledge the supply of samples from; The Cambridge Brain Bank, covered by current REC approval NRES 10/HO308/56; the Quebec Suicide Brain Bank at Douglas Mental Health University Institute, Canada; the MRC funded University of Edinburgh Brain & Tissue Bank; the Harvard Brain Tissue Resource Centre, which is supported by HHSN-271-2013-00030C; Brain Endowment Bank, at Miller School of Medicine, University of Miami; The Mount Sinai NBTR (NIH Brain and Tissue Repository), JJ Peters VA Medical Center; the Oxford Brain Bank, supported by the Medical Research Council (MRC), the NIHR Oxford Biomedical Research Centre and the Brains for Dementia Research programme, jointly funded by Alzheimer's Research UK and Alzheimer's Society; The Neuropathology Brain Bank at the University of Pittsburgh, School of Medicine Department of Psychiatry; The Stanley Medical Research Institute Brain Collection courtesy of Drs. Michael B. Knable, E. Fuller Torrey, Maree J. Webster, and Robert H. Yolken; The Human Brain and Spinal Fluid Resource Center (HBSFRC) NIH Neurobiobank. This study was supported by the National Institute for Health and Care Research Exeter Biomedical Research Centre. The views expressed are those of the author(s) and not necessarily those of the NIHR or the Department of Health and Social Care. For the purpose of open access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

Supplementary data

Supplementary data is available at *Briefings in Bioinformatics* online.

Conflict of interest: The authors declare no conflicts of interest.

Funding

These data were generated as part of Medical Research Council grant K013807 to J.M. and Alzheimer's Research UK (ARUK) grant ARUK-PPG2018A-010 to E.L.D. E.H., J.M., E.L.D., and L.C.S. were supported by Medical Research Council (MRC) grants K013807 and W004984 (awarded to J.M.). E.H. is supported by an Engineering and Physical Sciences Research Council Fellowship EP/V052527/1. Data analysis was undertaken using high-performance computing supported by a Medical Research Council (MRC) Clinical Infrastructure award (M008924) to J.M.

Data availability

Raw and processed DNAm data are available from GEO under accession number GSE279509.

Code availability

All code for the results presented in this manuscript are available at https://github.com/ejh243/BrainFANS. The bespoke quality control pipeline we describe can be found at https://github.com/ejh243/BrainFANS/tree/master/array/DNAm/preprocessing, codes for the analyses can be found at https://github.com/ejh243/BrainFANS/tree/master/array/DNAm/analysis/methodsDevelopment. We have also made our data and method for cell-specific EWAS power calculations available as a standalone R package which is available at https://github.com/ew367/CellPower/tree/main.

Ethics approval and consent to participate

Ethical approval for the study was granted by the College of Medicine & Health Research Ethics Committee under application reference number 6524714.

References

1. Battram T, Yousefi P, Crawford G. *et al.* The EWAS Catalog: a database of epigenome-wide association studies. *Wellcome Open Res* 2022;**7**:41. https://doi.org/10.12688/wellcomeopenres.17598.2
2. Li M, Zou D, Li Z. *et al.* EWAS atlas: a curated knowledgebase of epigenome-wide association studies. *Nucleic Acids Res* 2019;**47**:D983–8. https://doi.org/10.1093/nar/gky1027
3. Mill J, Heijmans BT. From promises to practical strategies in epigenetic epidemiology. *Nat Rev Genet* 2013;**14**:585–94. https://doi.org/10.1038/nrg3405
4. Relton CL, Davey SG. Epigenetic epidemiology of common complex disease: prospects for prediction, prevention, and treatment. *PLoS Med* 2010;**7**:e1000356. https://doi.org/10.1371/journal.pmed.1000356
5. Murphy TM, Mill J. Epigenetics in health and disease: heralding the EWAS era. *Lancet.* 2014;**383**:1952–4. https://doi.org/10.1016/S0140-6736(14)60269-5
6. Campagna MP, Xavier A, Lechner-Scott J. *et al.* Epigenome-wide association studies: current knowledge, strategies and recommendations. *Clin Epigenetics* 2021;**13**:214. https://doi.org/10.1186/s13148-021-01200-8
7. Hannon E, Lunnon K, Schalkwyk L. *et al.* Interindividual methylomic variation across blood, cortex, and cerebellum: implications for epigenetic studies of neurological and neuropsychiatric phenotypes. *Epigenetics* 2015;**10**:1024–32. https://doi.org/10.1080/15592294.2015.1100786
8. 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**:e1009443. https://doi.org/10.1371/journal.pgen.1009443
9. Salas LA, Zhang Z, Koestler DC. *et al.* Enhanced cell deconvolution of peripheral blood using DNA methylation for high-resolution immune profiling. *Nat Commun* 2022;**13**:761. https://doi.org/10.1038/s41467-021-27864-7
10. Roadmap Epigenomics Consortium, Kundaje A, Meuleman W. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* 2015;**518**:317–30. https://doi.org/10.1038/nature14248
11. Davies MN, Volta M, Pidsley R. *et al.* Functional annotation of the human brain methylome identifies tissue-specific

epigenetic variation across brain and blood. *Genome Biol* 2012;**13**:R43. https://doi.org/10.1186/gb-2012-13-6-r43
12. Jaffe AE, Irizarry RA. Accounting for cellular heterogeneity is critical in epigenome-wide association studies. *Genome Biol* 2014;**15**:R31. https://doi.org/10.1186/gb-2014-15-2-r31
13. Smallwood SA, Lee HJ, Angermueller C. *et al.* Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. *Nat Methods* 2014;**11**:817–20. https://doi.org/10.1038/nmeth.3035
14. Luo C, Rivkin A, Zhou J. *et al.* Robust single-cell DNA methylome profiling with snmC-seq2. *Nat Commun* 2018;**9**:3824. https://doi.org/10.1038/s41467-018-06355-2
15. Nichols RV, O'Connell BL, Mulqueen RM. *et al.* High-throughput robust single-cell DNA methylation profiling with sciMETv2. *Nat Commun* 2022;**13**:7627. https://doi.org/10.1038/s41467-022-35374-3
16. Accomando WP, Wiencke JK, Houseman EA. *et al.* Quantitative reconstruction of leukocyte subsets using DNA methylation. *Genome Biol* 2014;**15**:R50. https://doi.org/10.1186/gb-2014-15-3-r50
17. Hannon E, Dempster EL, Chioza B. *et al.* Quantifying the proportion of different cell types in the human cortex using DNA methylation profiles. *bioRxiv* 2023;2023.06.23.545974. https://doi.org/10.1101/2023.06.23.545974
18. Guintivano J, Aryee MJ, Kaminsky ZA. A cell epigenotype specific model for the correction of brain cellular heterogeneity bias and its application to age, brain region and major depression. *Epigenetics* 2013;**8**:290–302. https://doi.org/10.4161/epi.23924
19. Shireby G, Dempster EL, Policicchio S. *et al.* DNA methylation signatures of Alzheimer's disease neuropathology in the cortex are primarily driven by variation in non-neuronal cell-types. *Nat Commun* 2022;**13**:5620. https://doi.org/10.1038/s41467-022-33394-7
20. Houseman EA, Accomando WP, Koestler DC. *et al.* DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics* 2012;**13**:86. https://doi.org/10.1186/1471-2105-13-86
21. Gasparoni G, Bultmann S, Lutsik P. *et al.* DNA methylation analysis on purified neurons and glia dissects age and Alzheimer's disease-specific changes in the human cortex. *Epigenetics Chromatin* 2018;**11**:41. https://doi.org/10.1186/s13072-018-0211-3
22. Su D, Wang X, Campbell MR. *et al.* Distinct epigenetic effects of tobacco smoking in whole blood and among leukocyte subtypes. *PloS One* 2016;**11**:e0166486. https://doi.org/10.1371/journal.pone.0166486
23. Rhead B, Brorson IS, Berge T. *et al.* Increased DNA methylation of SLFN12 in CD4+ and CD8+ T cells from multiple sclerosis patients. *PloS One* 2018;**13**:e0206511. https://doi.org/10.1371/journal.pone.0206511
24. Natoli V, Charras A, Hofmann SR. *et al.* DNA methylation patterns in CD4. *Front Immunol* 2023;**14**:1245876. https://doi.org/10.3389/fimmu.2023.1245876
25. Rakyan VK, Beyan H, Down TA. *et al.* Identification of type 1 diabetes-associated DNA methylation variable positions that precede disease diagnosis. *PLoS Genet* 2011;**7**:e1002300. https://doi.org/10.1371/journal.pgen.1002300
26. Policicchio SSS, Davies JP, Chioza B. *et al.* Fluorescence-activated nuclei sorting (FANS) on human post-mortem cortex tissue enabling the isolation of distinct neural cell populations for multiple omic profiling. *Protocols.io* 2020. https://doi.org/10.17504/protocols.io.bmh2k38e
27. Policicchio SSS, Davies JP, Chioza B. *et al.* DNA Extraction from FANS sorted nuclei. *Protocols.io* 2020. https://dx.doi.org/10.17504/protocols.io.bmpmk5k6
28. Gorrie-Stone TJ, Smart MC, Saffari A. *et al.* Bigmelon: tools for analysing large DNA methylation datasets. *Bioinformatics* 2019;**35**:981–6. https://doi.org/10.1093/bioinformatics/bty713
29. Hannon E, Dempster EL, Davies JP. *et al.* Quantifying the proportion of different cell types in the human cortex using DNA methylation profiles. *BMC Biol* 2024;**22**:17. https://doi.org/10.1186/s12915-024-01827-y
30. Pidsley R, Wong YCC, Volta M. *et al.* A data-driven approach to preprocessing Illumina 450K methylation array data. *BMC Genomics* 2013;**14**:293. https://doi.org/10.1186/1471-2164-14-293
31. Champely S. pwr: Basic Functions for Power Analysis. 1.2–2 ed. https://CRAN.R-project.org/package=pwr2018.
32. Mansell G, Gorrie-Stone TJ, Bao Y. *et al.* Guidance for DNA methylation studies: Statistical insights from the Illumina EPIC array. *BMC Genomics* 2019;**20**:366. https://doi.org/10.1186/s12864-019-5761-7
33. van Rooij J, Mandaviya PR, Claringbould A. *et al.* Evaluation of commonly used analysis strategies for epigenome- and transcriptome-wide association studies through replication of large-scale population studies. *Genome Biol* 2019;**20**:235. https://doi.org/10.1186/s13059-019-1878-x
34. Wu MC, Joubert BR, Kuan PF. *et al.* A systematic assessment of normalization approaches for the Infinium 450K methylation platform. *Epigenetics* 2014;**9**:318–29. https://doi.org/10.4161/epi.27119
35. Squair JW, Gautier M, Kathe C. *et al.* Confronting false discoveries in single-cell differential expression. *Nat Commun* 2021;**12**:5692. https://doi.org/10.1038/s41467-021-25960-2
36. Seiler Vellame D, Castanho I, Dahir A. *et al.* Characterizing the properties of bisulfite sequencing data: maximizing power and sensitivity to identify between-group differences in DNA methylation. *BMC Genomics* 2021;**22**:446. https://doi.org/10.1186/s12864-021-07721-z

Briefings in Bioinformatics, 2025, **26(4)**, bbaf427
https://doi.org/10.1093/bib/bbaf427
Problem Solving Protocol