Accounting for cell lineage and sex effects in the identification of cell-specific DNA methylation using a Bayesian model selection algorithm

Nicole White[1☯], Miles Benton[2☯], Daniel Kennedy[1], Andrew Fox[3], Lyn Griffiths[2], Rodney Lea[2], Kerrie Mengersen[1]*

1 School of Mathematical Sciences, Queensland University of Technology, Brisbane, Queensland, Australia, **2** Genomics Research Center, Institute of Health and Biomedical Innovation, Queensland University of Technology, Brisbane, Queensland, Australia, **3** Florey Department of Neuroscience and Mental Health, Melbourne, Victoria, Australia

☯ These authors contributed equally to this work.
* k.mengersen@qut.edu.au

Abstract

Cell- and sex-specific differences in DNA methylation are major sources of epigenetic variation in whole blood. Heterogeneity attributable to cell type has motivated the identification of cell-specific methylation at the CpG level, however statistical methods for this purpose have been limited to pairwise comparisons between cell types or between the cell type of interest and whole blood. We developed a Bayesian model selection algorithm for the identification of cell-specific methylation profiles that incorporates knowledge of shared cell lineage and allows for the identification of differential methylation profiles in one or more cell types simultaneously. Under the proposed methodology, sex-specific differences in methylation by cell type are also assessed. Using publicly available, cell-sorted methylation data, we show that 51.3% of female CpG markers and 61.4% of male CpG markers identified were associated with differential methylation in more than one cell type. The impact of cell lineage on differential methylation was also highlighted. An evaluation of sex-specific differences revealed differences in CD56+NK methylation, within both single and multi- cell dependent methylation patterns. Our findings demonstrate the need to account for cell lineage in studies of differential methylation and associated sex effects.

Introduction

DNA methylation is a widely studied epigenetic modification that plays an essential role in the regulation of gene expression [1], cell differentiation [2] and the maintenance of chromatin structure [3]. Advances in high-throughput technologies [4–6] have made possible the collection of DNA methylation at the genome scale, allowing its relationship with biological processes to be interrogated. Studies of DNA methylation levels in humans, at both the global and

individual CpG levels, have revealed associations between aberrant methylation profiles and disease susceptibility [7], including carcinogenesis [8]. These studies have largely been conducted using samples of whole blood, in light of its convenience as a source of DNA and viability as a surrogate for target tissue. Its use, however, presents challenges for analysis and interpretation [9].

The role of DNA methylation in haematopoiesis has motivated the search for cell-specific methylation profiles at the individual CpG level and their association with lineage-specific gene expression [10–12]. To date, the discovery of cell-specific CpG markers has been based on the comparison of methylation levels between select purified cell subtypes [12] or against methylation levels observed in whole blood [13]. In [13], DNA methylation levels for seven purified cell subpopulations were compared and significant differences were identified between Lymphocyte and Myeloid cell populations. The definition of cell-specific methylation in this study and others was restricted the identification of differential methylation in a single cell type. The impact of accounting for more complex methylation profiles based on shared hematopoietic lineage remains unexplored.

Evidence of sex effects in DNA methylation is mixed and studies to date have focused primarily on whole blood. Previous studies of global and autosomal CpG methylation have reported a tendency for higher methylation in males [14] and sex-specific differences at varying numbers of CpG probes, across different chromosomes [15, 16]. A meta-analysis of published findings [17] identified 184 sex-specific, autosomal CpG probes, although average methylation differences were small and included studies did not account for cellular heterogeneity. Evidence for sex-specific methylation patterns for purified cell subtypes remains relatively unexplored. A recent study [18] examined sex-specific methylation levels for four purified immune cell subtypes (B cells, Monocytes, CD4+Foxp3-,CD8+T), and their role in immune-mediated diseases. Cell-specific methylation patterns in females versus males were compared and sex-specific differences were identified for each cell type. Similar to previous studies of cell-specific methylation, the identification of these differences was restricted to a single cell type, therefore precluding the exploration of sex-specific differences as a function of cell lineage.

In light of current limitations, this paper proposes new methodology for the identification of cell-specific methylation profiles at the individual CpG level and associated sex effects. Our methodology incorporates knowledge about cell lineage to accommodate the identification of differential methylation in one or multiple cell types, therefore expanding upon the range of methylation patterns currently studied. The main component of the methodology is a model choice algorithm, formulated within a Bayesian framework, which allows evidence for multiple cell-specific methylation patterns to be assessed simultaneously. Using cell-sorted methylation data from female and male samples, panels of cell-specific CpG markers are identified for each sex and common marker panels are derived. Subsequent analysis of identified CpG markers demonstrate the need to account for lineage in the discovery of cell-specific methylation patterns.

Materials and methods

Ethics statement

Female training data were collected as part of a study of epigenetics and multiple sclerosis at the Flory Institute in Melbourne, Australia. All patients involved provided informed consent and the study was granted ethics approval by the Eastern Health and Melbourne Health Human Research Ethics Committee. Remaining data used in this study were available from published studies and accessed from public data repositories.

Data

Publicly available Illumina 450K methylation data were obtained from six healthy males subjects [13] (GEO Accession number: GSE35069). The data contained methylation β-values on seven isolated cell populations: CD19B+ cells, CD14+ Monocytes, CD4$^+$T cells, CD8$^+$T cells, CD56$^+$ Natural Killers (NK), CD16$^+$ Neutrophils (Neu) and Eosinophils (Eos). Samples of the same cell types excluding Eosinophils were also obtained for five healthy female subjects, as part of a larger study (GEO Accession number: GSE88824). The methodology described in this paper was therefore applied to the six cell types common to both datasets. For both sexes, 445,603 CpG probes were available for analysis.

Consistent with previous findings [13], differences in DNA methylation levels across CpG sites were primarily driven by cell type differences, as opposed to being an artefact of between-subject variation (Fig 1). Cell lineage for each sex was subsequently inferred using hierarchical agglomerative clustering, applied to methylation β−values for each cell type and CpG probe, averaged over samples. Complete-linkage based on euclidean distance was assumed for the successive merging of cell types, however the choice of clustering metric did not impact the resulting hierarchy.

Additional cell sorted 450K methylation data on female and male subjects [18], aged 32-50 years, were obtained from public databases, for validation of model results. Female data consisted of cell-sorted samples on six healthy subjects downloaded from ArrayExpress (Accession number: E-ERAD-179) for CD19B$^+$B cells, CD14$^+$ Monocytes, CD4$^+$T and CD8$^+$ T cells. Methylation data on six healthy male subjects from the same study were downloaded from Gene Expression Ominibus (GEO Accession number: GSE71245). Following filtering and normalisation, 436,067 and 445,307 CpG sites were available for validation for females and males respectively, based on their correspondence with available sites in the training data.

Model formulation

The methodology presented in this paper was motivated by the identification of cell-specific methylation at the individual CpG level. A cell-specific CpG was defined as any CpG probe where the expected methylation level was different for one (single cell-specific) or more (multi- cell-specific) cell types, relative to others observed. The true cell-specific methylation pattern at each CpG probe was assumed unknown and treated as a model selection problem, with key details outlined in this Section.

Fig 1. Left to right: Principal components plot over samples (Circle = Female, Triangle = Male); Dendrogram of cell-sorted DNA methylation samples over 445,603 CpG probes for female and male samples; Comparison of inferred cell type lineage by clustering metric (Complete linkage, Single linkage, UPGMA, MPGMA). The response variable was equal to the mean methylation β−value by each cell type, at each CpG probe. UPGMA = Unweighted Pair Group Method with Arithmetic Mean, WPGMA = Weighted Pair Group Method with Arithmetic Mean.

https://doi.org/10.1371/journal.pone.0182455.g001

Table 1. Description of cell-specific models based on cell lineage from Fig 1. Each candidate model corresponds to a partition of cell types into non-overlapping groups. Cell types within each set of parentheses, {}, belong to the same partition and were assumed to have the same level of methylation. For each cell-specific model excluding 'All', the partition annotated by an asterisk (*) denotes the reference partition.

Candidate Model	Differentially methylated cell type(s)	Partition
Lymphocyte-I	All Lymphocytes	{CD19+},{CD4+},{CD8+},{CD56+},{CD14+,CD16+}*
Lymphocyte-II	Lymphocytes excl. CD19+ B cells	{CD4+},{CD8+},{CD56+},{CD19+,CD14+,CD16+}*
Myeloid	All Myeloids	{CD14+},{CD16+},{CD19+,CD4+,CD8+,CD56+}*
Pan T	T cells	{CD4+},{CD8+},{CD19+,CD56+,CD14+,CD16+}*
CD19+ B	CD19+ B cells	{CD19+},{CD4+,CD8+,CD56+,CD14+,CD16+}*
CD4+ T	CD4+ T cells	{CD4+},{CD19+,CD8+,CD56+,CD14+,CD16+}*
CD8+ T	CD8+ T cells	{CD8+},{CD19+,CD4+,CD56+,CD14+,CD16+}*
CD56+ NK	CD56+ Natural Killers	{CD56+},{CD19+,CD4+,CD8+,CD14+,CD16+}*
CD14+ Mono	CD14+ Monocytes	{CD14+},{CD19+,CD4+,CD8+,CD56+,CD16+}*
CD16+ Neu	CD16+ Neutrophils	{CD16+},{CD19+,CD4+,CD8+,CD56+,CD14+}*
All	All available cell types	{CD19+},{CD4+},{CD8+},{CD56+},{CD14+},{CD16+}
Null	None	{CD19+,CD4+,CD8+,CD56+,CD14+,CD16+}

https://doi.org/10.1371/journal.pone.0182455.t001

Using the results of hierarchical clustering (Fig 1), candidate models sought to characterise methylation levels with respect to the cell type(s) associated with each dendrogram split or node. This proposed approach to model specification was motivated by interest in identifying differentially methylated CpGs associated with each stage of whole blood lineage. Shared lineage was defined by the cell types present in a single split or node. In the corresponding candidate model, cell types with shared lineage were represented by unique, cell-specific means. Non-member cell types were assumed to share the same level of methylation to define a reference level, as associated levels of methylation were not of direct interest. This approach resulted in the specification of twelve candidate models, including the null and saturated cases (Table 1).

Each candidate model was translated into a linear model, where each cell type grouping was represented by a different mean parameter. For sex s, observed methylation β–values for each cell type at CpG k were represented by the vector, $\mathbf{y}_{iks} = (y_{iks1}, \ldots, y_{iksJ})$, for samples $i = 1, \ldots, n_s$. The partition defined under each candidate model was encoded into the design matrix, $\mathbf{X}_m$, of size $J \times P^{(m)}$, where $P^{(m)}$ was equal to the number of partitions for candidate model m. Mean methylation levels for each partition were represented by the vector $\boldsymbol{\mu}_{ks}^{(m)}$ of length $P^{(m)}$. The likelihood for a single CpG probe $k = 1, \ldots, K$ under candidate model m was given by,

$$p(\mathbf{y}_{ks}|\boldsymbol{\mu}_{ks}, \sigma_{ks}^2, \text{Model } m) = \prod_{i=1}^{n_s} \mathcal{N}_J(\mathbf{X}_m\boldsymbol{\mu}_{ks}^{(m)}, \sigma_{ks}^{2(m)}\mathbf{I}_{J\times J})$$

where $\mathcal{N}_J(\mathbf{a}, \mathbf{A})$ defines a J–dimensional Normal distribution with mean vector $\mathbf{a}$ and variance-covariance matrix $\mathbf{A}$. The unknown variance was assumed to be common across all cell types, denoted by $\sigma_{ks}^{2(m)}\mathbf{I}_{J\times J}$ where $\mathbf{I}$ is the identity matrix. This variance-covariance structure was chosen for identifiability reasons given sample sizes available.

A Bayesian approach to model selection was adopted, which allowed for probabilistic statements to be made about the relative fit of each candidate model. Under this approach, the posterior probability of each model conditional on the observed data was calculated for all candidate models. The posterior probability of model m compared to other candidates m' was

calculated as,

$$Pr(\text{Model } m|\mathbf{y}_{ks}) = \frac{p(\mathbf{y}_{ks}|\text{Model } m)Pr(\text{Model } m)}{\sum_{m'} p(\mathbf{y}_{ks}|\text{Model } m')Pr(\text{Model } m')} \qquad (1)$$

where the sum of probabilities over all candidate models was equal to 1. The term $p(\mathbf{y}_{ks}|\text{Model } m)$ was obtained by integrating over all unknown parameters from the likelihood and prior distributions. Prior probabilities of each model, $Pr(\text{Model } m)$, were assumed to be equal to reflect a lack of model preference *a priori*.

Prior distributions for remaining parameters were selected such that Eq 1 could be derived analytically. A g-prior distribution [19] was adopted for each $\boldsymbol{\mu}_{ks}^{(m)}$,

$$\boldsymbol{\mu}_{ks}^{(m)}|g_s, \sigma_{ks}^2 \sim \mathcal{N}_{p^{(m)}}\left(b_0, g_s \frac{\sigma_{ks}^{2(m)}}{n_s}(\mathbf{X}_m^T \mathbf{X}_m)^{-1}\right).$$

The g-prior is a popular choice in linear model selection settings, as it allows the experimenter to introduce information on the scale of $\mathbf{X}_m$. Expected methylation levels were centered around the overall methylation level, b_0, with variance proportional the standard error of each partition. The scaling factor $g_s > 0$ is interpreted as a relative weighting of the prior versus the observed data. Here, g_s was assumed common over all CpG probes and estimated by maximising the marginal likelihood averaged over candidate models. In this paper, b_0 was set to the global methylation mean, averaged over CpGs and cell types. A conjugate prior distribution for $\sigma_{ks}^{2(m)}$ of form $p(\sigma_{ks}^{2(m)}) \propto 1/\sigma_{ks}^{2(m)}$ completed model specification [19].

The expectation-maximisation (EM) algorithm was used to obtain a global Empirical Bayes estimate for g_s [20]. This computational approach provided significant computational benefits over sampling-based approaches, namely Markov chain Monte Carlo (MCMC), and was possible given closed-form solutions for $p(\mathbf{y}_{ks}|\text{Model } m)$ conditional on g_s. In addition, the desired posterior model probabilities from Eq 1 were available upon convergence of the EM algorithm. The proposed methodology was implemented in R, with code available as Supporting Information (S1 File).

Making the most of the Bayesian approach for model based inference

The accommodation of model and parameter uncertainty under the Bayesian approach formed the basis of subsequent inference, namely the identification of cell-specific CpGs, the estimation of differential methylation by cell type and the assessment of sex-specific differences by cell type. Brief details of each inference are provided in this Section.

CpG marker identification. CpG probes or 'markers' associated with each cell-specific pattern were identified by comparing posterior model probabilities at each CpG probe. For each candidate model and sex, markers were identified by reviewing the set of K model probabilities: $Pr(\text{Model } m|\mathbf{y}_{1s}), \ldots, Pr(\text{Model } m|\mathbf{y}_{Ks})$. A 5% Bayes' False Discovery Rate (FDR) [21] was applied to each set of probabilities to control the expected number of false discoveries. Common CpG markers were defined as CpGs that were identified in both female and male samples, for the same candidate model. The compilation of common marker panels allowed for the assessment of sex-specific differences by cell type, within each candidate model.

Estimation of differential methylation. Posterior inference about each marker panel focused on the estimation of differential methylation by cell type relative to its corresponding reference partition (Table 1). Under model m, the posterior distribution of differential

methylation for cell-specific partition p and sex s is,

$$\mu_{kps}^{(m)} - \mu_{k1s}^{(m)} \sim \mathcal{N}\left(\hat{\mu}_{ksp}^{(m)} - \hat{\mu}_{ks1}^{(m)}, \frac{\hat{g}_s}{1+\hat{g}_s} \frac{\hat{\sigma}_{ks}^{2(m)}}{n_s}\left(\frac{1}{c_p^{(m)}} + \frac{1}{c_1^{(m)}}\right)\right) \quad (2)$$

where c_p is the number of cell types in partition p and $(\hat{\boldsymbol{\mu}}_{ks}^{(m)}, \hat{\sigma}_{ks}^{2(m)})$ are posterior estimates of each mean and residual variance, respectively, for which direct estimates were available. For CpGs identified under each marker panel, Eq 2 was summarised in terms of a posterior mean and 95% credible interval (CI). The posterior probabilities of differential methylation being within selected ranges (<0.10, $0.10-0.20$, $0.20-0.30$, 0.30, 0.40, $0.40-0.50$, >0.5), were also calculated.

The posterior distribution in Eq 2 was also used to validate selected CpG markers. Validation of common CpG markers was limited to CD19$^+$ B, CD4$^+$ T, CD8$^+$ T and Pan T models, in light of validation samples available. Using all validation samples for each sex, the average β −value for each cell type and CpG was computed. The difference between each average β −value and reference partition was then compared with the corresponding 95% CI from Eq 2 for the appropriate candidate model. Concordance rates with respect to the predicted direction of methylation (hypomethylated, hypermethylated) for validation samples were also calculated. This joint approach was motivated by the limitation that validation based on coverage of 95% CIs relied on a posterior estimate of $\sigma_{ks}^{2(m)}$. When this estimate is small and/or underestimated, proportions of validated markers based on CI coverage only were likely to be low, even if differences in methylation between training and validation samples were small. Finally, it is noted that not all cell types observed in the training data were available in the validation samples. Given the properties of the multivariate Normal distribution, this discrepancy was addressed by using the appropriate marginal distributions for the available cell types.

Evaluation of sex effects. A similar expression to Eq 2 was derived to evaluate sex-specific methylation differences within common CpG marker panels. In this case, focus was on the comparison of female and male methylation estimates for differentially methylated cell types, as defined by the corresponding candidate model. The posterior distribution of this difference across model partitions was Multivariate Normal,

$$\boldsymbol{\mu}_{k1}^{(m)} - \boldsymbol{\mu}_{k2}^{(m)} \sim \mathcal{N}_{p^{(m)}}\left(\hat{\boldsymbol{\mu}}_{k1}^{(m)} - \hat{\boldsymbol{\mu}}_{k2}^{(m)}, \hat{\sigma}_{k1}^{2(m)}\boldsymbol{\Sigma}_{m1} + \hat{\sigma}_{k2}^{2(m)}\boldsymbol{\Sigma}_{m2}\right) \quad (3)$$

where $\hat{\boldsymbol{\mu}}_{ks}^{(m)} = \frac{\hat{g}_s}{1+\hat{g}_s}(\mathbf{X}_m^T\mathbf{X}_m)^{-1}\mathbf{X}_m^T\bar{\mathbf{y}}_{ks}$ for $s = 1, 2$. To assess evidence for sex effects, the posterior probability that the difference in methylation between sexes using Eq 3 was at least 0.10 was calculated for each cell-specific partition. This calculation again relied on estimates of the residual variance, $\hat{\sigma}_{ks}^{2(m)}$ for each sex which were set to their posterior mean estimate. A sex-specific difference for a given partition was declared if the posterior probability of a difference greater than 0.10 exceeded 0.95.

CpG markers to genes: Assessment of SNP effects, genomic features and pathway enrichment

To provide additional evidence to support our method of marker identification, a pathways enrichment analysis was performed to explore the underlying biology of gene sets derived from common CpG marker panels. Common CpG markers were mapped to genes using Illumina Human Methylation 450K annotation data available from Bioconductor [22]. SNP information was also collated to infer the percentage of SNP associated markers with associations limited to SNPs located directly on the CpG loci. Using KEGG functional analysis in

WebGestalt [23], a hypergeometric test was applied to each marker panel. A 5% FDR [24] was applied to resulting p-values to identify significant pathway enrichment for derived gene lists.

Results

Cell lineage impacts the identification of differentially methylated CpGs

Across all cell-specific models, 83,449 and 97,747 CpG markers were identified for females and males, respectively (Table 2). Among female samples, 42,834 markers (51.3%) were associated with differential methylation in multiple cell types, which included a three-fold increase in Pan T markers compared with males. A larger proportion of multi- cell-specific markers among males was observed (64.8%), due to larger numbers identified under Lymphocyte-II and Myeloid models. Among single cell-specific models, CD19+B was the most frequently observed marker type for both sexes, with 25,611 in females and 18,271 in males.

A total of 42,452 CpGs were associated with the same cell-specific methylation pattern for both sexes, corresponding to 9.5% of the observed methylome. Within this subset, 23,551 (55.5%) were defined by differential methylation in more than one cell type. Over all candidate models, smaller frequencies of common markers were associated with T cell-specific markers (CD4+, CD8+, Pan T).

Differential methylation is affected by cell lineage among common CpG markers

Common CpG markers associated with differential methylation in Myeloid cell types (CD14+ Mono, CD16+ Neu) were consistently hypomethylated across all relevant marker panels (Table 3). Among Lymphocytes, CD8+T markers were the least likely to be hypomethylated (35.99%). Smaller proportions of hypomethylation among Lymphocyte-I and II panels were indicative of greater levels of methylation among Lymphocyte versus Myeloid cell subtypes.

The impact of cell lineage on differential methylation was greatest among marker panels related to Lymphocytes (Fig 2). Lower levels of differential methylation (<0.10) were concentrated within single cell dependent markers (CD19+B, CD4+T, CD8+T, CD56+NK). In contrast, the comparison of distributions indicated a wider range of posterior estimates observed for Pan T, Lymphocyte-I and Lymphocyte-II panels. The comparison of distributions associated with CD14+ Monocytes showed little evidence of being affected by cell lineage.

Table 2. Number of CpG markers identified by candidate model for females versus males, based on a 5% Bayes False Discovery Rate (FDR). For each model, the number of common markers identified in both sexes is also listed. The number of CpGs identified for a single sex only are given in brackets for each candidate model.

Candidate model	Female	Male	Common
All	3873 (1699)	3553 (1379)	2174
Myeloid	12541 (3637)	29534 (20630)	8904
Lymphocyte-I	15596 (7369)	14105 (5878)	8227
Lymphocyte-II	4241 (1032)	14015 (10806)	3209
Pan T	6583 (5546)	2126 (1089)	1037
CD19+ B	25611 (11987)	18271 (4647)	13624
CD4+ T	1277 (821)	2050 (1594)	456
CD8+ T	2299 (1985)	2742 (2428)	314
CD56+ NK	4893 (3302)	4346 (2755)	1591
CD14+ Mono	1383 (627)	1681 (925)	756
CD16+ Neu	5152 (2992)	5324 (3164)	2160

Table 3. Percentage of hypomethylation among common markers by cell type and sex. Results are summarised for single cell-specific (CD19+B, CD4+T,CD8+T,CD56+NK,CD14+Mono,CD16+Neu) and multi- cell-specific (Myeloid, Pan T, Lymphocyte-I, Lymphocyte-II) marker panels.

	Sex	Single cell-specific markers					
		CD19+B	CD4+T	CD8+T	CD56+NK	CD14+Mono	CD16+Neu
	Female	58.24	65.57	35.99	90.19	94.71	96.39
	Male	58.24	65.57	35.99	90.19	94.71	96.39
		Multi- cell-specific markers					
Myeloid	Female	–	–	–	–	93.53	93.41
	Male	–	–	–	–	93.53	93.41
Pan T	Female	–	43.30	43.30	–	–	–
	Male	–	43.30	43.30	–	–	–
Lymphocyte-I	Female	29.52	26.70	27.26	29.11	–	–
	Male	29.52	26.70	27.46	28.86	–	–
Lymphocyte-II	Female	–	44.13	45.56	45.72	–	–
	Male	–	44.13	45.65	45.81	–	–

https://doi.org/10.1371/journal.pone.0182455.t003

The distribution of differential methylation by sex among common Pan T markers revealed considerable variation for both hypermethylated and hypomethylated states compared to CD4+T and CD8+ T panels (Fig 3). Differential methylation levels among Pan T markers tended to be greater for CD4+ T cells; 25 markers in this subset showed strong evidence of differential methylation greater than 0.5 for both sexes (S1 Fig).

Validation of common CpG markers for B- and T- lymphocytes

The validation of common markers with respect to mean differences was generally higher in males than females, across immune cell subtypes (Table 4). Whilst validation based on coverage of credible intervals was low, differences between training and validation estimates were relatively small across all markers tested (Fig 4); for CD4+T markers, approximately 80% of absolute differences were less than 10% (S2 Fig). Furthermore, comparisons with respect to inferred methylation state showed very high concordance rates for all marker panels and these findings were consistent for both sexes.

Genomic feature distributions and enrichment analysis

The effects of cell lineage were evident in the comparison of genomic features, with higher proportions of markers residing in Transcription Start Sites (TSS) for Lymphocyte cell subtypes compared with Myeloids (Table 5). Common markers were concentrated in the gene body and intergenic regions, with 54.73% of CD56+NK markers located in the gene body. Among Lymphocyte cell subtypes, TSS proportions were highest for T cells, with 18.28% and 16.48% for CD4+T and CD8+T cells, respectively.

Moderate proportions of markers were directly associated with SNPs and these levels were maintained between sex-specific and common marker panels, averaging 30.7% in common markers (S1 Table). Higher SNP proportions were observed in marker panels related to the lineage of Myeloid cells, except for CD56+ NK markers of which 36.02% were SNP associated. For CpG probes not assigned to any common marker panel, a similar degree of SNP association was observed (26.61%).

Significant pathways enrichment for common markers were associated with immune cell subtypes (Table 6) all including biologically relevant pathways.
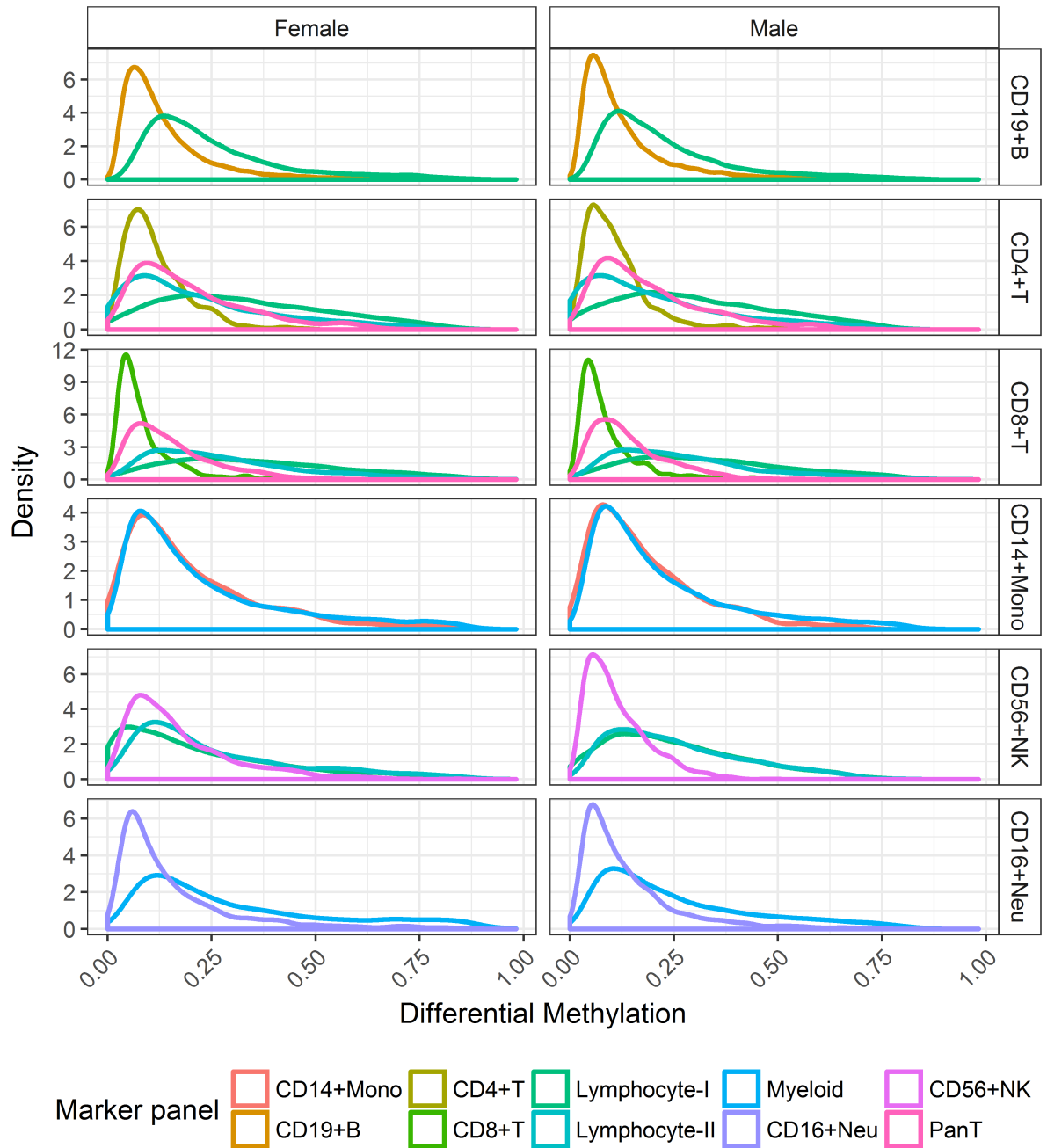
Fig 2. Distribution of posterior mean estimates of differential methylation for each purified cell type across corresponding marker panels. Posterior estimates are summarised for females (first column) and males (second column) across common CpG markers.

https://doi.org/10.1371/journal.pone.0182455.g002

Assessment of common marker profiles shows sex effects in CD56$^+$NK, CD16$^+$Neutrophils

The majority of common CpG markers did not exhibit sex-specific differences in methylation profile (Table 7). Differences within CD56$^+$NK and Lymphocyte-I/II panels indicated 12.32, 14.15 and 20.64% of respective markers exhibited sex effects, with the majority corresponding
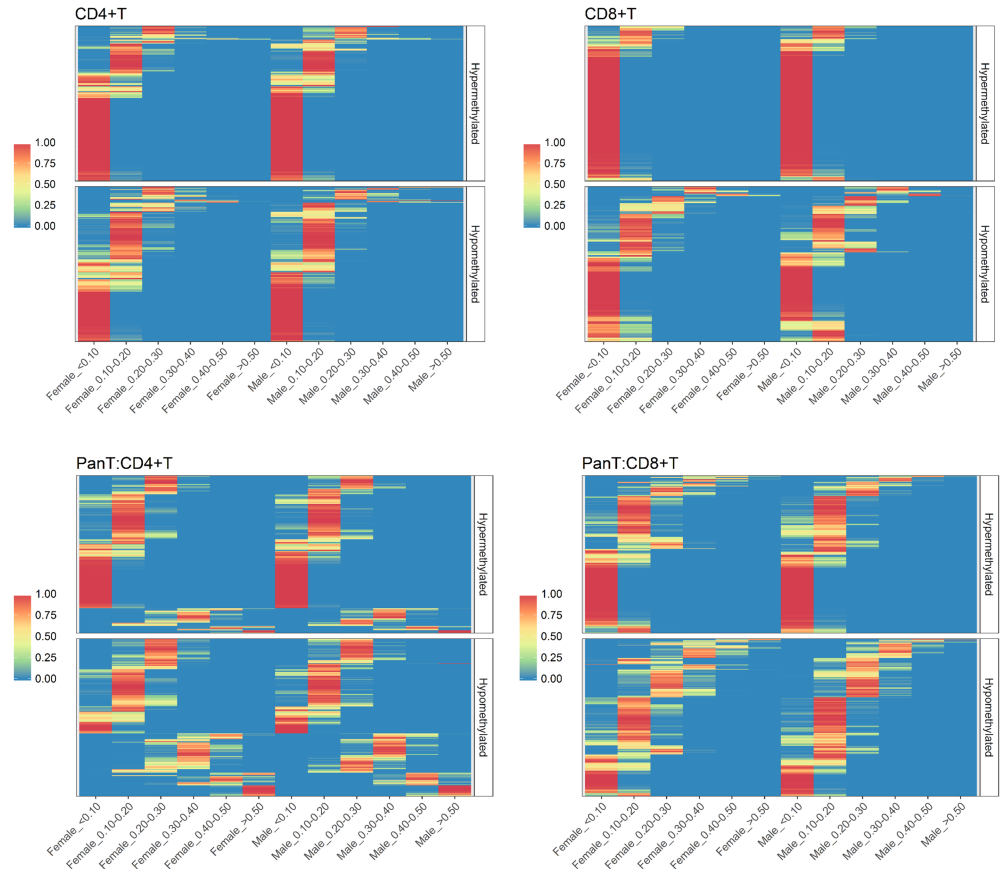
Fig 3. Posterior probability heatmaps for varying levels of differential methylation for common CD4⁺T, CD8⁺T and Pan T markers. For each cell type, CpG markers (rows) are ordered the same for males and females, to enable sex-specific comparisons. Markers are further classified by methylation state (Hypomethylated, Hypermethylated), based on their posterior mean estimate of differential methylation. First row (L-R): CD4⁺T, CD8⁺T; Second row: Pan T.

https://doi.org/10.1371/journal.pone.0182455.g003

to autosomal CpGs. No sex-specific differences were identified among common CD8$^+$ T markers.

Sex-specific differences identified among single cell-specific markers were uniquely mapped to 215 genes (S2 Table). The majority of identified cases were concentrated in the

Table 4. Percentage of validated, common CpG markers based on the coverage of 95% credible intervals inferred from the training data, by sex and marker type. The outcome of interest for validation was the estimated difference in methylation. For Pan T markers, results are summarised by individual cell type and jointly, the latter corresponding to both CD4⁺T and CD8⁺T differences being validated for the same marker.

Marker panel	Differential methylation		Methylation state	
	Female	**Male**	**Female**	**Male**
CD19⁺B	36.9	25.8	99.9	99.7
CD4⁺T	37.7	53.9	97.6	100
CD8⁺T	38.1	63.4	98.6	99.7
PanT	21.4	31.6	99.7	99.5
Pan T: CD4⁺T only	46.7	54.5	100	100
Pan T: CD8⁺T only	42.7	41.1	99.7	99.5

https://doi.org/10.1371/journal.pone.0182455.t004
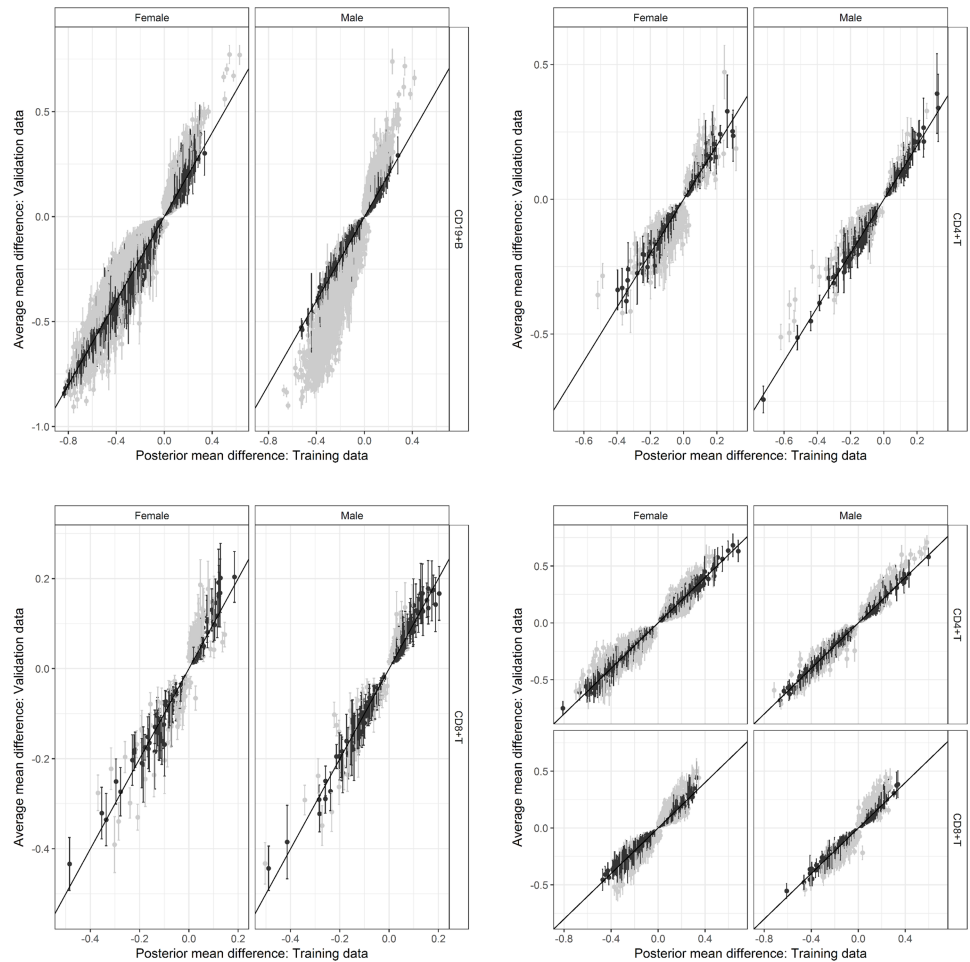
Fig 4. Validation of common marker panels for immune cell subtypes, based on the comparison of mean differences for each cell type relative to its corresponding reference partition. For each marker type, posterior mean differences inferred from the training data (x-axis) are compared with the average mean difference calculated from the validation data (y-axis). Estimates of each posterior mean difference are accompanied by a respective 95% CI. Validated markers based on the coverage of each 95%CI are indicated in black. First row (L-R): CD19+, CD4+ T; Second row (L-R): CD8+ T, Pan T. CI: Credible interval.

https://doi.org/10.1371/journal.pone.0182455.g004

Table 5. Distribution of genomic features among common CpG markers by cell-specific methylation pattern.

Marker panel	1stExon	3'UTR	5'UTR	Body	Intergenic	TSS1500	TSS200
Myeloid	2.88	5.90	11.17	46.85	15.41	13.65	4.14
Lymphocyte-I	4.10	4.20	14.26	40.84	13.26	17.05	6.28
Lymphocyte-II	3.58	4.31	14.48	40.68	12.92	18.15	5.88
anT	4.32	2.19	14.29	40.75	13.26	17.98	7.20
CD19+B	7.07	4.32	13.55	38.71	13.81	14.20	8.35
CD4+T	3.79	3.52	15.14	41.64	12.40	18.28	5.22
CD8+T	6.67	2.78	12.59	38.52	13.15	16.48	9.81
CD56+NK	2.21	5.07	12.17	54.73	12.43	9.80	3.59
CD14+Mono	1.72	4.38	10.47	47.30	20.60	11.33	4.21
CD16+Neu	1.72	8.12	10.63	53.68	12.21	10.89	2.75

https://doi.org/10.1371/journal.pone.0182455.t005

Table 6. Summary of biologically relevant pathways identified in enrichment analysis of common marker panels. All pathways were identified based on a Benjamini-Hochberg (BH) adjusted p-value < 0.05.

| Marker panel	Pathway	p-value
CD14+Monocytes	Fc gamma R-mediated phagocytosis	0.01
	Cytokine-cytokine receptor interaction	0.04
CD19+B	B cell receptor signaling pathway	2.84×10^{-13}
	T cell receptor signaling pathway	3.07×10^{-11}
	Fc epsilon RI signaling pathway	4.58×10^{-9}
CD4+T	T cell receptor signaling	2.73×10^{-7}
	Antigen processing and presentation	0.0002
	Cytokine-cytokine receptor interaction	0.0007
CD8+T	Antigen processing and presentation	0.02
CD56+NK	Natural killer cell mediated cytotoxicity	6.35×10^{-5}
	Chemokine signaling pathway	4.85×10^{-5}
	T cell receptor signaling pathway	0.0003
CD16+Neu	Endocytosis	1.07×10^{-8}
	Chemokine signaling pathway	0.0002
	Fc gamma R-mediated phagocytosis	0.0003
	Phagosome	0.04
Pan T	Chemokine signaling pathway	3.33×10^{-5}
	T cell receptor signaling pathway	0.0005
	B cell receptor signaling pathway	0.004
	Natural killer cell mediated cytotoxicity	0.005

CD56+ NK common marker panel. Four CD4+ T markers were associated with greater methylation observed in females and were mapped to a single gene, *CD40LG*, located on the X chromosome. Sex-specific differences and annotation information for all common markers identified is provided in the Supplementary Material (S2 File).

Sex effects in CD56+ NK methylation were prominent in Lymphocyte-I and Lymphocyte-II common marker panels, in addition to the CD56+ NK panel (Fig 5). Within the Lymphocyte-I panel, 1698 common CpG markers were identified as sex-specific, of which 1627 showed differences in CD56+ NK methylation only. Similarly, 442 common Lymphocyte-II markers

Table 7. Summary of sex-specific differences by common CpG marker panel. A difference between male and female methylation estimates ≥0.10 was the outcome of interest. A marker was declared sex-specific if the posterior probability for this outcome exceeded 0.95, for one or more model-based partitions. For each marker panel, total numbers of sex-specific markers and autosomal sex-specific markers are given.

	Non sex-specific	Sex-specific differences	
Marker panel	(%)	Total	Total Autosomal
Myeloid	91.69	740	685
Lymphocyte-I	79.36	1698	1635
Lymphocyte-II	85.85	454	444
Pan T	99.42	6	2
CD19+ B	99.75	34	11
CD4+ T	98.90	5	1
CD8+ T	100	0	0
CD56+ NK	87.68	196	194
CD14+ Mono	99.07	7	4
CD16+ Neu	96.67	72	66
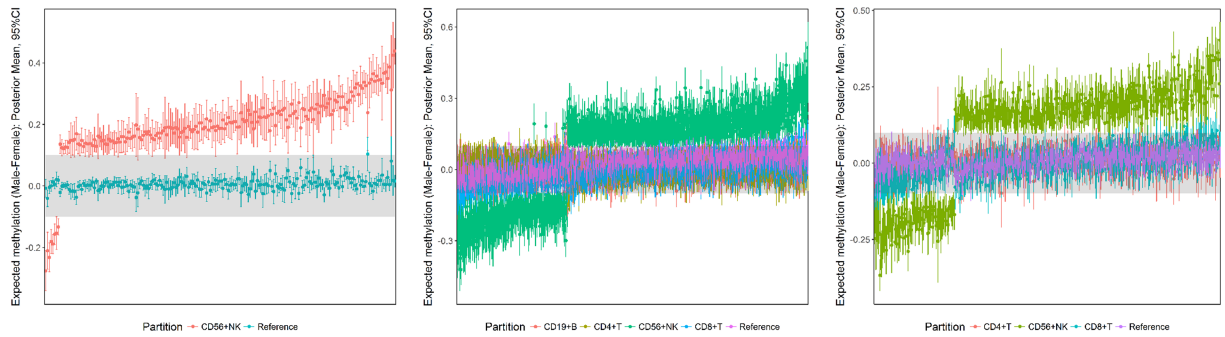
Fig 5. Posterior summaries of sex-specific differences in selected common marker panels defined by differences in CD56+ NK methylation only. Each difference is summarised by the posterior mean estimate and corresponding 95% CI. The shaded region corresponds to a difference of ±0.1. L-R: CD56+ NK, Lymphocyte-I, Lymphocyte-II.

showed differences in CD56+ NK methylation between sexes. The tendency for CD56+ NK methylation to be higher in males was common across all three panels.

684 common Myeloid markers were associated with differences in CD14+ Monocytes or CD16+ Neutrophils only. Differences with respect to CD14+ Monocytes tended to be higher in females, compared with higher male methylation in CD16+ Neutrophils (Fig 6).

Discussion

This paper has proposed new statistical methodology for the discovery of cell-specific methylation profiles in whole blood, applying principles of Bayesian model selection. The characterisation of CpGs by differential methylation in one or more cell types builds significantly upon existing work that has been restricted to univariate analyses of differential methlyation by cell type. Sex-specific differences in both the prevalence of cell-specific profiles and methylation signal for select cell types were also demonstrated. For immune cell subtypes, validation of common markers using external cell-sorted samples produced favourable results. Enrichment analyses of common marker panels provided additional support for the proposed methodology, where it was demonstrated that the detection of cell-specific methylation at the individual CpG level was also biologically meaningful.
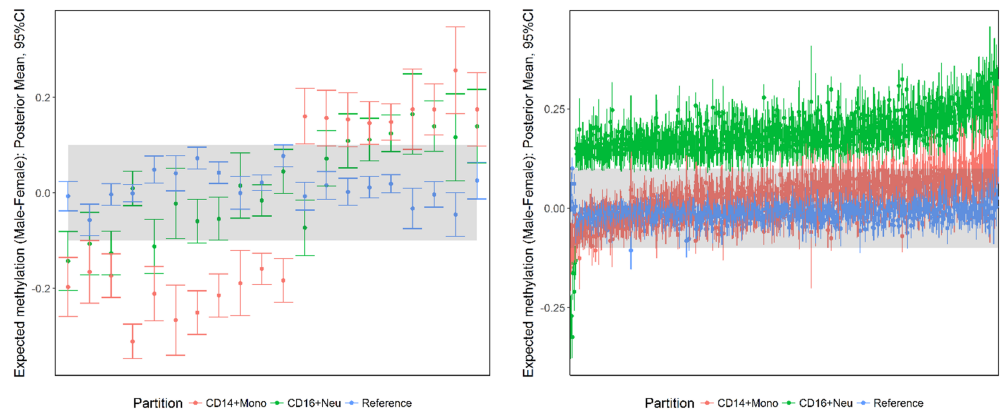
Fig 6. Posterior summaries of sex-specific differences in common Myeloid markers defined by differences in CD14+Monocytes (Left) or CD16+ Neutrophils (Right) only. The difference at each marker is summarised by the posterior mean and corresponding 95% CI. The shaded region corresponds to a methylation difference of ± 0.10.

The adoption of a Bayesian approach was motivated by the benefits it offered over traditional approaches to model selection. The calculation of posterior model probabilities provided an informative measure of model evidence that accounted for selection uncertainty between multiple competing hypotheses [25]. The availability of these probabilities therefore enabled multiple candidate models to be compared simultaneously. Resulting probabilities were also used directly to identify CpG markers associated with each cell-specific pattern, whilst controlling for a defined false discovery rate. Whilst equal prior weights were assumed for all candidate models, other specifications are possible [26]. Future research in this area should consider alternative prior specifications and their impact on marker identification. For example, spatial structure among neighbouring CpGs could be incorporated, shifting the emphasis of identification to cell-specific differentially methylated regions (DMRs), offering an alternative to current statistical approaches [27].

The incorporation of knowledge about hematopoietic lineage into model specification represents a semi-supervised approach that reflects relevant cell biology. Consistent with other model selection strategies, our approach assumes that the defined set of candidate models is exhaustive and true patterns beyond this set are not of primary interest. In the event that the true methylation pattern does not align with any of the candidate models specified, two outcomes are likely. In the first instance, corresponding probes are assigned to the saturated model and may be examined further as part of a larger marker panel. Alternatively, the vector of model probabilities for each CpG may be diffuse over multiple possibilities and therefore not be identified for any cell-specific pattern under the chosen criteria. The feasibility of unsupervised approaches to model selection, where the full list of methylation patterns is determined by the observed data is appealing, but their routine application in high dimensional settings is prohibitive. Furthermore, there is potential for patterns identified to be an artefact of random noise present in cell sorted samples as opposed to true biological signal. Integrated methods of model selection represent an avenue for future work, where new cell-specific patterns are proposed based on the combination of cell lineage and their prevalence in the observed methylome.

It is well documented that there are sex-specific differences in the proportions of circulating white blood cells [28, 29]. The application of the proposed methodology to female and male samples has highlighted the importance of accounting for sex effects in DNA methylation analyses. Greater numbers of CD19$^+$B and T cell-specific markers in females are consistent with previous findings and are possibly indicative of higher levels of cell activation [30]. The association between sex-specific differences in select CD4$^+$T markers and the *CD40LG* gene have also been identified previously. Previous studies have pointed to allele specific methylation for this gene [31, 32] where CD4$^+$T hypomethylation is observed in healthy males compared with healthy women who carry one methylated and one hypomethylated allele. One of our major findings was large differences of methylation between males and females in markers defined as CD56$^+$NK specific. This is interesting when considered alongside the observation that males show an increase in circulatory NK cells compared to females [30], which adds further support for the accuracy of the approach. Additionally there is some evidence of sex-specific methylation differences in CD56$^+$ NK, as well as CD8$^+$ T cells [33]. Under the proposed approach, we have provided a potential solution to accurately account for potential bias introduced by sex effects at the marker level.

The presence of cell-specific CpG markers highlights the need to account for cellular composition prior to conducting Epigenome Wide Association Studies (EWAS), in whole blood. Methods for this purpose have been developed [34, 35] based on the assembly of methylation 'signatures' from cell-sorted data which are then projected onto heterogeneous samples to predict cell type proportions. The inclusion of other marker panels identified in our study may

lead to further improvement in cell mixture estimation, in particular for immune cell subtypes that may be present in low proportions. Furthermore, the performance of these algorithms rely on consistent cell-type effects across cohorts [36]. Given the sex-specific methylation differences we have identified in this study, failure to account for sex effects may also impact upon the quality of cell mixture estimation and should therefore be given due consideration.

It is common practice in array-based methylation studies to exclude CpG sites which contain SNPs both within the probe and on the CpG site. While this is a valid approach to filtering before analysis, it will often lead to dramatic reduction of overall data. As a result, it is likely that sites of potential interest may be lost before any association can be made. By mapping hg38 annotated SNPs to all 450K CpG loci, we were able to ascertain the overall proportion of cell marker sites which have a SNP present; on average, across the common set of markers, this was approximately 30.7% of markers. In light of these results, we suggest that deconvolution studies and methods should account for SNP events at cell marker sites, noting the proportion that are present. For the filtering stage, we recommend that the overall rarity of the SNP variant be taken into account, for example, retaining CpGs which also have a rare (MAF < 0.01) variant mapping. This approach is likely to be beneficial to the overall study design and outcome.

Supporting information

S1 Table. Percentage of markers associated with SNPs, for sex-specific and common markers by candidate model. The category 'Unassigned' refers to all CpG probes that were not assigned to any marker panel, based on a 5% Bayes' FDR.
(PDF)

S2 Table. Distribution of sex-specific common markers over chromosomes, for single cell-dependent makers only. A sex-specific marker was declared if the posterior probability of a methylation difference ≥ 0.10 was greater than 0.95 for at least one cell type.
(PDF)

S1 Fig. Distribution of observed methylation β−values by cell type and sex for selected common Pan T markers. Markers were identified as having high levels of differential methylation (>0.5) in CD4$^+$ T cells. Markers were identified if the corresponding posterior probability of differential methylation >0.5 exceeded 0.95.
(TIFF)

S2 Fig. Validation analysis: Empirical cumulative distribution function (CDF) of the absolute difference between posterior and validation estimates by validation status. Validation status was defined by the coverage of each 95% CI, estimated from the training data. First row (L-R): CD19$^+$B, CD4$^+$T; Second row (L-R): CD8$^+$T, Pan T.
(TIFF)

S1 File. R code for Bayesian model selection algorithm. Core R functions required to prepare data and compute posterior model probabilities via the EM algorithm, across all listed candidate models.
(ZIP)

S2 File. List of common CpG markers associated with a sex-specific difference of ≥ 0.10. Illumina Human Methylation 450K annotation data are also included for each CpG marker.
(XLSX)

Author Contributions

Conceptualization: Nicole White, Miles Benton, Daniel Kennedy, Rodney Lea.

Data curation: Miles Benton, Andrew Fox.

Formal analysis: Nicole White, Miles Benton.

Funding acquisition: Lyn Griffiths, Kerrie Mengersen.

Investigation: Andrew Fox.

Methodology: Nicole White.

Supervision: Lyn Griffiths, Rodney Lea, Kerrie Mengersen.

Writing – original draft: Nicole White, Miles Benton.

Writing – review & editing: Nicole White, Miles Benton, Daniel Kennedy, Andrew Fox, Lyn Griffiths, Rodney Lea, Kerrie Mengersen.

References

1. Muers M. Gene expression: Disentangling DNA methylation. Nat Rev Genet. 2013; 14(8):519–519. https://doi.org/10.1038/nrg3535 PMID: 23797851

2. Meissner A, Mikkelsen TS, Gu H, Wernig M, Hanna J, Sivachenko A, et al. Genome-scale DNA methylation maps of pluripotent and differentiated cells. Nature. 2008; 454(7205) 766–770. https://doi.org/10.1038/nature07107 PMID: 18600261

3. Hashimshony T, Zhang J, Keshet I, Bustin M, Cedar H. The role of DNA methylation in setting up chromatin structure during development. Nature Genetics. 2003; 34(2):187–192. https://doi.org/10.1038/ng1158 PMID: 12740577

4. Dedeurwaerder S, Defrance M, Bizet M, Calonne E, Bontempi G, Fuks F. A comprehensive overview of Infinium HumanMethylation450 data processing. Briefings in bioinformatics. 2014; 15(6): 929–941. https://doi.org/10.1093/bib/bbt054 PMID: 23990268

5. Dedeurwaerder S, Defrance M, Calonne E, Denis H, Sotiriou C, Fuks F. Evaluation of the Infinium Methylation 450K technology. Epigenomics. 2011; 3(6): 771–784. https://doi.org/10.2217/epi.11.105 PMID: 22126295

6. Sandoval J, Heyn H, Moran S, Serra-Musach J, Pujana MA, Bibikova M, et al. Validation of a DNA methylation microarray for 450,000 CpG sites in the human genome. Epigenetics. 2011; 6(6):692–702. https://doi.org/10.4161/epi.6.6.16196 PMID: 21593595

7. Rakyan VK, Beyan H, Down TA, Hawa MI, Maslau S, Aden D, et al. Identification of type 1 diabetes—associated DNA methylation variable positions that precede disease diagnosis. PLoS Genetics 2011; 7 (9):e1002300. https://doi.org/10.1371/journal.pgen.1002300 PMID: 21980303

8. Franco R, Schoneveld O, Georgakilas AG, Panayiotidis MI. Oxidative stress DNA methylation and carcinogenesis. Cancer Letters. 2008; 266(1):6–11. https://doi.org/10.1016/j.canlet.2008.02.026 PMID: 18372104

9. Houseman EA, Kim S, Kelsey KT, Wiencke JK. DNA methylation in whole blood: uses and challenges. Current Environmental Health Reports. 2015; 2(2):145–154. https://doi.org/10.1007/s40572-015-0050-3 PMID: 26231364

10. Adalsteinsson BT, Gudnason H, Aspelund T, Harris TB, Launer LJ, Eiriksdottir G, et al. Heterogeneity in white blood cells has potential to confound DNA methylation measurements. PloS One. 2012; 7(10): e46705. https://doi.org/10.1371/journal.pone.0046705 PMID: 23071618

11. Bocker MT, Hellwig I, Breiling A, Eckstein V, Ho AD, Lyko F. Genome-wide promoter DNA methylation dynamics of human hematopoietic progenitor cells during differentiation and aging. Blood. 2011; 117 (19):e182–e189. https://doi.org/10.1182/blood-2011-01-331926 PMID: 21427290

12. Glossop JR, Nixon NB, Emes RD, Haworth KE, Packham JC, Dawes PT, et al. Epigenome-wide profiling identifies significant differences in DNA methylation between matched-pairs of T- and B-lymphocytes from healthy individuals. Epigenetics. 2013; 8(11):1188–1197. https://doi.org/10.4161/epi.26265 PMID: 24005183

13. Reinius LE, Acevedo N, Joerink M, Pershagen G, Dahlén SE, Greco D, et al. Differential DNA methylation in purified human blood cells: implications for cell lineage and studies on disease susceptibility. PloS One. 2012; 7(7):e41361. https://doi.org/10.1371/journal.pone.0041361 PMID: 22848472

14. El-Maarri O, Becker T, Junen J, Manzoor SS, Diaz-Lacava A, Schwaab R, et al. Gender specific differences in levels of DNA methylation at selected loci from human total blood: a tendency toward higher methylation levels in males. Human Genetics. 2007; 122(5):505–514. https://doi.org/10.1007/s00439-007-0430-3 PMID: 17851693

15. Liu J, Morgan M, Hutchison K, Calhoun VD. A study of the influence of sex on genome wide methylation. PloS One. 2010; 5(4):e10028. https://doi.org/10.1371/journal.pone.0010028 PMID: 20386599

16. Eckhardt F, Lewin J, Cortese R, Rakyan V, Attwood J, Burger M, et al. DNA methylation profiling of human chromosomes 6, 20 and 22. Nature Genetics. 2006; 38(12):1378–1385. https://doi.org/10.1038/ng1909 PMID: 17072317

17. McCarthy NS, Melton PE, Cadby G, Yazar S, Franchina M, Moses EK, et al. Meta-analysis of human methylation data for evidence of sex-specific autosomal patterns. BMC Genomics. 2014; 15:981. https://doi.org/10.1186/1471-2164-15-981 PMID: 25406947

18. Mamrut S, Avidan N, Staun-Ram E, Ginzburg E, Truffault F, Berrih-Aknin S, et al. Integrative analysis of methylome and transcriptome in human blood identifies extensive sex-and immune cell-specific differentially methylated regions. Epigenetics. 2015; 10(10):943–957. https://doi.org/10.1080/15592294.2015.1084462 PMID: 26291385

19. Zellner A. On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In: Goel PK, Zellner A, editors. Bayesian inference and decision techniques: Essays in Honor of Bruno De Finetti. Amsterdam: North-Holland; 1986. pp. 233–243.

20. Liang F, Paulo R, Molina G, Clyde MA, Berger JO. Mixtures of g priors for Bayesian variable selection. Journal of the American Statistical Association. 2012; 103(481):410–423. https://doi.org/10.1198/016214507000001337

21. Newton MA, Noueiry A, Sarkar D, Ahlquist P. Detecting differential gene expression with a semiparametric hierarchical mixture method. Biostatistics. 2004; 5(2):155–176. https://doi.org/10.1093/biostatistics/5.2.155 PMID: 15054023

22. Triche Jr T. IlluminaHumanMethylation450k.db: Illumina Human Methylation 450K annotation data. R package version 2.0.9. 2014.

23. Wang J, Duncan D, Shi Z, Zhang B. WEB-based GEne SeT AnaLysis Toolkit (WebGestalt): Update 2013. Nucleic Acids Research. 2013; 41:W77–W83. https://doi.org/10.1093/nar/gkt439 PMID: 23703215

24. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society Series B (Methodological). 1995; 57(1):289–300.

25. Friel N, Wyse J. Estimating the evidence–a review. Statistica Neerlandica. 2012; 66(3):288–308. https://doi.org/10.1111/j.1467-9574.2011.00515.x

26. Fernandez C, Ley E, Steel MF. Benchmark priors for Bayesian model averaging. Journal of Econometrics. 2001; 100(2):381–427. https://doi.org/10.1016/S0304-4076(00)00076-2

27. Chen DP, Lin YC, Fann CS. Methods for identifying differentially methylated regions for sequence- and array-based data. Briefings in Functional Genomics. 2016; 15(6): 485–490. PMID: 27323952

28. Fish EN. The X-files in immunity: sex-based differences predispose immune responses. Nature Reviews Immunology. 2008; 8(9):737–744. https://doi.org/10.1038/nri2394 PMID: 18728636

29. Klein SL, Flanagan KL. Sex differences in immune responses. Nature Reviews Immunology. 2016; 16 (10):626–638. https://doi.org/10.1038/nri.2016.90 PMID: 27546235

30. Abdullah M, Chai P, Chong M, Tohit E, Ramasamy R, Pei C, et al. Gender effect on in vitro lymphocyte subset levels of healthy individuals. Cellular Immunology. 2012; 272(2):214–219. https://doi.org/10.1016/j.cellimm.2011.10.009 PMID: 22078320

31. Schmidl C, Klug M, Boeld TJ, Andreesen R, Hoffmann P, Edinger M, et al. Lineage-specific DNA methylation in T cells correlates with histone methylation and enhancer activity. Genome Research. 2009; 19 (7):1165–1174. https://doi.org/10.1101/gr.091470.109 PMID: 19494038

32. Jeffries M, Dozmorov M, Tang Y, Merrill JT, Wren JD, Sawalha AH. Genome-wide DNA methylation patterns in CD4+ T cells from patients with systemic lupus erythematosus. Epigenetics. 2011; 6 (5):593–601. https://doi.org/10.4161/epi.6.5.15374 PMID: 21436623

33. Inoshita M, Numata S, Tajima A, Kinoshita M, Umehara H, Yamamori H, et al. Sex differences of leukocytes DNA methylation adjusted for estimated cellular proportions. Biology of Sex Differences. 2015; 6:11. https://doi.org/10.1186/s13293-015-0029-7 PMID: 26113971

34. Houseman EA, Accomando WP, Koestler DC, Christensen BC, Marsit CJ, Nelson HH, 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 PMID: 22568884

35. Koestler DC, Jones MJ, Usset J, Christensen BC, Butler RA, Kobor MS, et al. Improving cell mixture deconvolution by identifying optimal DNA methylation libraries (IDOL). BMC Bioinformatics. 2016; 17:120. https://doi.org/10.1186/s12859-016-0943-7 PMID: 26956433

36. Yousefi P, Huen K, Quach H, Motwani G, Hubbard A, Eskenazi B, et al. Estimation of blood cellular heterogeneity in newborns and children for epigenome-wide association studies. Environmental and Molecular Mutagenesis. 2015; 56(9):751–758. https://doi.org/10.1002/em.21966 PMID: 26332589