Reference-free deconvolution of DNA methylation data and mediation by cell composition effects

E. Andres Houseman[1], Molly L. Kile[1], David C. Christiani[2], Tan A. Ince[3], Karl T. Kelsey[4], Carmen J. Marsit[5]

1. School of Biological and Population Health Sciences, College of Public Health and Human Sciences, Oregon State University; Corvallis, OR, USA. Email: andres.houseman@oregonstate.edu

2. Department of Environmental Health, Harvard T. H. Chan School of Public Health; Boston, MA, USA.

3. Department of Pathology, University of Miami, Miller School of Medicine; Miami, FL, USA.

4. Department of Epidemiology, Department of Pathology and Laboratory Medicine, Brown University

5. Department of Community and Family Medicine, Dartmouth Medical School; Hanover, NH, USA.

Abstract

We propose a simple method for reference-free deconvolution that provides both proportions of putative cell types defined by their underlying methylomes, the number of these constituent cell types, as well as a method for evaluating the extent to which the underlying methylomes reflect specific types of cells. We have demonstrated these methods in an analysis of 23 Infinium data sets from 13 distinct data collection efforts; these empirical evaluations show that our algorithm can reasonably estimate the number of constituent types, return cell proportion estimates that demonstrate anticipated associations with underlying phenotypic data; and methylomes that reflect the underlying biology of constituent cell types. Thus the methodology permits an explicit quantitation of the mediation of phenotypic associations with DNA methylation by cell composition effects. Although more work is needed to investigate functional information related to estimated methylomes, our proposed method provides a novel and useful foundation for conducting DNA methylation studies on heterogeneous tissues lacking reference data.

Introduction

In the last decade, there has been an increasing interest in epigenome-wide association studies (EWAS), which aim to investigate associations between DNA methylation and health or exposure phenotypes across the genome. Numerous publications have reported associations between DNA methylation profiled in a single tissue and disease states or exposure phenotypes. Most of these studies have used whole blood[1] or cord blood[2-4], but some have used other media such as buccal swabs[5], adipose tissue[6, 7], and placenta[8-11].

However, most tissues are complex mosaic of cells derived from at least two and sometimes three different germ layers; endoderm, mesoderm and ectoderm that give rise to both epithelial and stromal compartments. Just the epithelial component of an organ can be composed of many cell types; for example we found that breast epithelium is composed of at least 10-12 cell types[12] with potentially distinct DNA methylation profiles[13]. Added to this complexity are the cells in the stromal component with distinct functions, including vascular and lymphoid endothelial cells and pericytes, immune cells such as macrophages, leukocytes and lymphocytes, stromal fibroblasts, myofibroblasts, myoepithelial cells, as well as adipose cells, endocrine cells, nerve cells and other cellular and tissue elements that have different but systematically varying developmental origins. The complexity of the epigenome in normal tissues has been described in a recent analysis of 111 reference human epigenomes of human tissues[14]. Thus, because normal tissue development, individual cellular differentiation and cellular lineage determination are regulated by epigenetic mechanisms, which include chromatin alterations as well as DNA methylation[15-18], many phenotypic associations with DNA methylation may be explained in whole or in part by systematic associations with the distribution of underlying cell types. This has been demonstrated statistically in numerous papers[19-22] and in one notable recently published manuscript which identified and confirmed the specific cell subtype responsible for the highly replicated relationship between tobacco smoking exposure and DNA methylation of the GPR15 locus[23]. This phenomenon has led to an interest in methods for adjusting EWAS studies for cell-type heterogeneity. In *referenced-based* deconvolution methods, the distribution of cell types is obtained by projecting whole-tissue DNA methylation data onto linear spaces spanned by cell-type-specific methylation profiles for a specific set of CpGs that distinguish the cell types, so-called *differentially methylated positions* (DMPs)[19]; these methods require the existence of a reference set consisting of the cell-type specific methylation profiles, such as those that exist for blood[19, 24, 25]. However, no such reference sets exist for solid tissues of interest, such as adipose and placenta, or even tumors, thus motivating *reference-free* methods[13, 26, 27] that seek to adjust DNA methylation associations for cell-type distribution.

Numerous cell-type deconvolution methods are currently available, many of them based on mRNA or protein expression[28]; all of them are essentially either reference-based, i.e. supervised by the pre-selection of loci known to differentiate cell types, or else reference-free, i.e. essentially unsupervised. While reference-based deconvolution methods allow for direct inference of the relationship between phenotypic variation and altered cell composition of characterized cell subtypes, reference-free approaches can provide only limited, if any, information on the types of cells contributing to the phenotypic association. In this article we

propose a simple method for reference-free deconvolution that addresses this challenge and that provides both interpretable outputs – proportions of putative cell types defined by their underlying DNA methylation profiles – as well as a means for evaluating the extent to which the underlying profiles reflect specific types of cells.

Our fundamental approach is as follows: we assume an $m \times n$ matrix $\mathbf{Y}$ representing DNA methylation data collected for n subjects or specimens, each measured on an array of m CpG loci, and that the measured values are constrained to the unit interval $[0,1]$, each roughly representing the fraction of methylated cytosine molecules in the given sample at a specific genomic position. This conforms to the typical *average beta* output of popular platforms such as the Infinium arrays by Illumina, Inc. (San Diego, CA), i.e. the older HumanMethylation27 (27K) platform, which interrogates 27,578 CpG loci, and the newer HumanMethylation450 (450K) platform, which interrogates 485,412 CpG loci; however, it also conforms to the results of sequencing-based platforms such as whole genome bisulfite sequencing (WGBS). In reference-based methods, the following relation is assumed to hold: $\mathbf{Y} = \mathbf{M}\mathbf{\Omega}^{\mathrm{T}}$, where $\mathbf{M}$ is a *known* $m \times K$ matrix representing m CpG-specific methylation states for K cell types and $\mathbf{\Omega}$ is an $n \times K$ matrix representing subject-specific cell-type distributions (each row representing the cell-type proportions for a given subject, i.e. the entries of $\mathbf{\Omega}$ lie within $[0,1]$ and the rows of $\mathbf{\Omega}$ sum to values less than one). Reference-free methods attempt to circumvent lack of knowledge about $\mathbf{M}$ either by using a two-stage regression analysis (e.g. the Houseman approach[27]) or else fitting a high-dimensional mixed-effects model and equating the resulting random coefficients with cell-mixture effects (i.e. the Zou approach[26]); both methods rely on a predetermined model positing associations between DNA methylation $\mathbf{Y}$ and phenotypes $\mathbf{X}$. For example, the Houseman method posits the model $\mathbf{Y} = \mathbf{A}\mathbf{X}^{\mathrm{T}} + \mathbf{R}$, where $\mathbf{X}$ is an $n \times d$ design matrix of phenotype variables and potential confounders; the $m \times d$ regression coefficient matrix $\mathbf{A}$ and the $m \times n$ error matrix $\mathbf{R}$ are both assumed to have further linear structure involving $\mathbf{M}$, and the common variation between $\mathbf{A}$ and $\mathbf{R}$ is assumed to represent systematic association with cell type distribution. However, results of this approach are somewhat influenced by the choice of the dimension of the linear subspace of $[\mathbf{A}, \mathbf{R}]$ representing the common variance induced by $\mathbf{M}$ [20]; consequently there has been recent concern that the method may over-adjust for cell distribution. A similar problem exists with the Zou approach, which models the phenotype as a linear function of DNA methylation, and in which the choice of a tuning parameter can influence the extent to which phenotypic associations are putatively explained by heterogeneity in underlying cell types. Here, we propose that a variant of non-negative matrix factorization be used to decompose $\mathbf{Y}$ as $\mathbf{Y} = \mathbf{M}\mathbf{\Omega}^{\mathrm{T}}$, where the entries of the unknown matrices $\mathbf{M}$ and $\mathbf{\Omega}$ are constrained to lie in the unit interval and the rows of $\mathbf{\Omega}$ are constrained to sum to a value less than or equal to one. This approach is similar to existing approaches for estimating the proportion of normal tissue cells in a tumor sample or otherwise deconvolving mixtures of cells[29-33]. Additionally, this factorization conforms to the biological assumption that DNA methylation measurements $\mathbf{Y}$, regardless of associated metadata $\mathbf{X}$, ultimately arise as linear combinations of constituent methylomes, as we have previously argued[20]. However, such constrained factorizations can be computationally intensive, and it is

still necessary to specify the number K of assumed cell types, so in Supplementary Section S1 we propose a fast approximation that facilitates resampling, which is the basis of our method for determining K, described in Section S2. Note that $K=1$ corresponds to the case where there are no relevant constituent cell types, which should be true for relatively pure media. If associations remain between $\mathbf{\Omega}$ and $\mathbf{X}$, i.e. if the associations between $\mathbf{X}$ and $\mathbf{Y}$ factor through the decomposition $\mathbf{Y}=\mathbf{M}\mathbf{\Omega}^{\mathrm{T}}$, then these associations are potentially explained by systematic changes in cell composition. Evidence for mediation of associations by cell type is substantially strengthened if the methylomes represented by $\mathbf{M}$ map to biological processes that correspond to distinct populations of cells. To that end, we propose a simple companion analytical procedure for the interpretation of the methylomes represented by $\mathbf{M}$. Denote each row of $\mathbf{M}$ (corresponding to one CpG) as the $K\times 1$ vector $\boldsymbol{\mu}_j$, $j\in\{1,...,m\}$. CpG loci that most differentiate the K putative cell types will tend to have distinct values within $\boldsymbol{\mu}_j$; thus high values of the row-variance $s_j^2=\mathrm{var}\{\mu_{j1},...,\mu_{jK}\}$ should correspond to CpGs that are most relevant to the biological distinctions among the K cell types, and this can be tested with auxiliary annotation data. Figure 1 illustrates our approach.

We demonstrate these methods in analyses of 23 genome-scale DNA methylation data sets from 13 distinct data collection efforts, including four blood data sets, several breast tumor data sets (including data from The Cancer Genome Atlas, *TCGA*), vascular and liver tissues, sperm, and four separate media collected on the same population, Bangladeshi neonates, including placenta. In addition, we leverage data derived from The Roadmap Epigenomics Project, demonstrating their utility in addressing the biological relevance of fitted methylomes $\mathbf{M}$.

Results

To test our proposed approach, we analyzed 23 DNA methylation datasets from 13 distinct studies, each set of DNA methylation measurements obtained via the Infinium 27K or 450K platform. Four blood data sets[3, 22, 34, 35] were included as positive controls (given the existing reference data and known heterogeneity), each collected in the context of an epidemiologic study, and each assumed to exhibit heterogeneity in cell type as previously described[3, 19, 22]. Sperm[36] and isolated vascular tissues[37] were included as negative controls, presumed to represent relative homogeneity in terms of constituent cell types. Note that four datasets arose from one study on arsenic exposure in Bangladeshi neonates[3, 9], in which four separate tissues were obtained from the same individuals. Also included were arterial tissue[38], liver tissue[39], and data from cancer data sets[40-43], including breast tissues from *TCGA*[44]. Table 1 lists the data sets, their sources and their short descriptions. Figure S3.1 shows the results of hierarchical clustering applied to 26,476 CpG sites common across the datasets (Manhattan distances based on mean methylation for the data set, clustering based on Ward's method implemented as *Ward.D* in R version 3.2.2). Figure S3.2 summarizes the number of CpGs analyzed for each data set, by fraction of samples observed for each CpG. Note the strong clustering of data sets by type of media. The ordering of data sets in Table 1 and many subsequent figures are based on the clusters shown in Figure S3.1.

Estimated Numbers of Cell Types

Using the method described in Supplementary Section S2 with 25 iterations, for each data set we found the decomposition $\mathbf{Y} = \mathbf{M}\mathbf{\Omega}^{\mathrm{T}}$, for values of K varying from 2 to either $K_{\max} = 10$ or the maximum possible given the sample size ($K_{\max} = 8$ for *BV+LV*, 2 for *BV* and *LV*, 7 for *AR[n]*). We then used our bootstrap approach (Supplementary Section S3) for determining the number $\hat{K}$ of classes for each data set, displayed in Figure 2a, which demonstrates heterogeneity in the number of classes $\hat{K}$ estimated. $\hat{K} \geq 3$ for blood data sets ($\hat{K} = 3$ for the cord blood data set *BL-as*, larger for the other three peripheral blood datasets). For breast tissues (both tumor and normal) $\hat{K}$ was typically large. $\hat{K} \geq 3$ for the artery and liver data sets having three distinct sources each (*AR[np]* and *L[np]*). $\hat{K} = 1$ for the pure blood and lymphatic vessel data sets (*BV* and *LV*); $\hat{K} = 2$ for other vessel data sets consisting of normal tissue (*AR[n]* , *AR-as*, *BV+LV*, *UV-as*), for the normal liver data set *L[n]*, for sperm (*SP*) and for placenta (*PL-as*). We remark that $\hat{K}$ was typically lower for datasets that were more likely to be comprised of homogeneous tissues. We also remark that our proposed method of selecting $\hat{K}$ is based on minimizing a bootstrapped deviance statistic, and that the variation of this statistic across values of K can be informative. For example, with the *BL-ra* dataset, the deviance dropped precipitously from $K = 1$ to $K = 3$, while for the sperm data set the deviance remained flat from $K = 1$ to $K = 6$ before rapidly increasing (Figure 2b).

Associations with Phenotypic Metadata

To examine the associations between $\mathbf{\Omega}$ and various metadata associated with the subjects/specimens in the corresponding study, we fit a quasi-binomial model for each row of $\mathbf{\Omega}$; Table 1 provides the covariate model $\mathbf{X}$ used for each data set. As described below and in detail in Supplementary Section S4, to circumvent dependence of results on the choice of K, we examined associations over the range $K \in \{1,...,K_{\max}\}$, using a permutation test (1000 permutations) for inference on each covariate. Table 2 provides a summary of permutation test results. As shown in Table 2, cell mixture proportions $\mathbf{\Omega}$ were typically significantly associated with major phenotypes of interest and occasionally with age (e.g. *bl-hn* and *BR-tcga[t]*); the exception was the sperm dataset, for which $\mathbf{\Omega}$ was not significantly associated with fraction. Note in particular that for breast tumors, ER status variables (or histology variables incorporating ER status) were significantly associated with $\mathbf{\Omega}$. Sex was typically not significantly associated with $\mathbf{\Omega}$. As shown in Figure 2, the associations between $\mathbf{\Omega}$ and phenotype could be quite striking (e.g. rheumatoid arthritis status) or completely lacking (e.g. sperm fraction). Figure 3a shows clustering heatmap of $\mathbf{\Omega}$ ($K = \hat{K} = 10$) for *BL-ra*, one of the positive control blood data sets, with annotation track showing the associated phenotype rheumatoid arthritis case/control status. Figure 3b shows a similar clustering heatmap for the negative control *SP* (sperm, $K = \hat{K} = 2$), along with the associated phenotype, specimen fraction. Other clustering heatmaps are provided supplementary files.

We also considered the effect of $\mathbf{\Omega}$ on CpG-specific associations of DNA methylation with $\mathbf{X}$. As described below and in detail in Section S5, we computed regression coefficients for logit-methylation (i.e. M-values) upon $[\mathbf{X}, \mathbf{\Omega}]$ for $K \in \{1,...,K_{\max}\}$ (for $K = 1$ the covariate model was simply $\mathbf{X}$); for $\mathbf{\Omega}$ and for each covariate, we then used the resulting nominal p-values to estimate the proportion π_0 of null associations. For demographic variables (age, sex, race), Figure S5.1 illustrates the value of π_0 for the overall association of $\mathbf{\Omega}$ with DNA methylation ($K = K^* = \max(2, \hat{K})$). For demographic variables, Figure S5.2 provides a comparison of π_0 from the $K = 1$ model with π_0 from the $K = K^*$ model. Figure 4 displays a similar comparison for other variables. These figures demonstrate that adjustment by $\mathbf{\Omega}$ very often resulted in higher values of π_0, the estimated proportion of null associations. Exceptions were age in *bl-hn* and sex in *AR[np]* and *AR-as*, where adjustment by $\mathbf{\Omega}$ reduced π_0 (Figure S5.2). On the other hand, the proportion of null associations with $\mathbf{\Omega}$ was typically low: except for the homogeneous tissue datasets *SP* and *BL+LV*, π_0 was less than 0.2; of the others, except for the arsenic-exposure data sets *UV-as*, *AR-as*, and *PL-as*, π_0 was extremely close to zero.

Interpretation of Putative Cell Types

We examined the biological relevance of resulting matrices $\mathbf{M}$ in several different ways. First, for each data set, we computed row-variances s_j^2 (as described above) both for $K = 2$ and for $K^* = \max(2, \hat{K})$. For each of these two values of K, we classified each CpG $j \in \{1,...,m\}$ by whether its row-variance s_j^2 lay above the 75th percentile $q_{0.75}(s^2)$, reasoning that these CpGs could be considered as important distinguishers of cell type. Next, we obtained a list of DMPs for differentiating distinct major types of leukocytes (*Blood DMPs*), and another list of CpGs mapped to genes considered Polycomb Group proteins (*PcG loci*), the construction of both lists described in detail in Supplementary Section S6. For each data set we computed the odds ratio for the association of high row-variance ($s_j^2 > q_{0.75}(s^2)$) with DMP set membership (*Blood DMPs* or *PcG loci*), using Fisher's exact test to compute the corresponding p-values. Odds ratios are depicted in Figure 5, with $\log_{10}$ p-values given in Table S6.1. *Blood DMP* status showed the highest associations in blood data sets, although also somewhat high associations in *L[np]* and *AR[np]* (data sets having tissues with potentially inflammatory components to pathology). Data sets with tumors (*BR-tcga[t]*, *br-1[t]*, *br-2[t]*, *br-3[t]*, *g[nt]*, and *g[t]*) showed high association of *PcG loci* with cell-type distinguishing CpGs, but so did the data set with normal gastric tissue, *g[n]*. As shown in Figure S6.1, the *Bilenky* DMPs based on breast tissue showed the highest association with cell-type distinguishing CpGs in the data sets with breast tissue, although associations were also high in *L[np]*, *AR[n]*, and *AR[np]*. As shown in Figure S6.2, the *REMC* DMPs, based on comparison of ectodermal/mesodermal/endodermal distinctions among embryonic stem cells, showed relatively weak (or negative) associations with cell-type distinguishing CpGs for all datasets.

We also developed a novel approach based on WGBS data from the Roadmap Epigenomics Project for 24 primary tissues. For each sample, we obtained the 470,909 CpGs overlapping with CpGs from either Infinium array (and having fewer than 3 missing values), clustering the tissue samples based on the 15,000 most variable of these CpGs (Manhattan distance metric with Ward's method of clustering). The resulting dendrogram, shown in Figure S7.1, demonstrates substantial clustering along general tissue type. We also applied our deconvolution algorithm to these 24 tissue samples ($K = 6$), with results shown in Figure S7.2; note that the deconvolution of these tissues resulted in constituent cell types that roughly aligned with anticipated anatomical associations, e.g. tissues with substantial smooth or skeletal muscle mapped to one cell type, tissues with a substantial lymphoid/immune component mapped to another, and central nervous tissues map to yet another. We reasoned that *similar* tissue types would differ principally in the proportion of underlying normal constituent cell types, and thus provide information on cell-type heterogeneity underlying other tissues of similar type. Consequently, we selected the tissue pairs corresponding to the 25 smallest Manhattan distances (as calculated for the clustering in Figure S7.1), with pairs illustrated as network edges in Figure S7.3. Due to small numbers of DMPs (10 or fewer) we excluded two pairs; for each of the remaining 23 pairs, we identified, among the 15,000 CpGs most variable across all 24 tissue types, those CpGs that differed in methylation fraction by greater than 0.70 between the two samples; we considered these CpGs to be Infinium-specific DMPs for tissue-specific heterogeneity. Using these 23 sets of DMPs, we conducted a gene-set analysis as described in the previous paragraph. The clustering heatmap in Figure 6 presents the odds ratios for the 450K data with $K^* = \max(2, \hat{K})$; the heatmap in Figure S6.4 presents the odds ratios for the 27K data with $K^* = \max(2, \hat{K})$, and the odds ratios for $K = 2$ are given in Figures S7.5 and S7.6. Corresponding p-values are given in Tables S7.1, S7.2 and S7.3. Note that we excluded additional pairs from the 27K array analysis due to small DMP overlap with the 27K array. As shown in Figure 6, positions that distinguished immune-related tissues (CD34+ hematopoietic stem cells vs. thymus or spleen) were highly associated with CpGs distinguishing cell types in the two 450K blood data sets, as well as in the mixed liver tissue dataset *L[np]* and the mixed arterial dataset *AR[np]*, consistent with the findings demonstrated in Figure 5a. In the arterial data sets *AR[n]* and *AR[np]*, the normal breast data set *BR-tcga[n]* and to some extent the normal mixed vessel data set *BV+LV*, high associations were found for CpGs that distinguished smooth muscle content (aorta vs. psoas muscle, heart atrium vs. ventricle, heart atrium vs. esophagus). Interestingly, *AR[n]*, *AR[np]*, and *BR-tcga[n]* displayed associations with CpGs distinguishing lung and esophagus, potentially an epithelial cell comparison (although potentially also representing a distinction in smooth muscle content). All other positive associations were relatively weak. Strong negative associations with CpGs that distinguished right atrium from left ventricle were observed for *L[n]*, *SP*, *UV-as*, and *AR-as*, although these results may be driven by small numbers of CpGs (see p-values in Table S7.2). Patterns were similar for $K = 2$ (Figure S7.5). Patterns were similar in 27K blood data sets; additionally, the normal gastric data set g[n] displayed high association with DMPs distinguishing Roadmap stomach tissues (Figure S7.4). Interestingly, *L[n]* was the only dataset displaying mostly negative (though weak) associations.

<u>Additional Analyses on 450K Blood Datasets</u>

We analyzed four blood data sets as positive controls, with the expectation that the resulting cell-type proportions $\mathbf{\Omega}$ would show substantial associations with $\mathbf{X}$. Practically, considering that reference data sets exist for blood, estimation of associations between phenotypic metadata and major types of leukocytes would typically employ the reference-based estimation of $\mathbf{\Omega}$ rather than the essentially unsupervised approach we have proposed. Two additional avenues of investigation emerge: (1) the extent to which the reference-based and reference-free approaches are consistent in their results; and (2) the extent to which the unsupervised approach may provide additional information on immune response and inflammation (as represented by distributions of leukocytes including their various activation states) beyond associations with simply the major types of leukocytes, i.e. those existing in currently available reference sets. To this end, we further analyzed the two 450K blood data sets, *BL-ra* and *BL-as*, estimating for each data set two sets of cell-type proportion matrices ($K = 7$): $\mathbf{\Omega}_0$ (reference-based) and $\mathbf{\Omega}_1$ (reference-free). We used a common set of DMPs for each estimation procedure, with details provided in Supplementary Section S8. Note that for the reference-based approach, we fit $\mathbf{Y} = \mathbf{M}_0\mathbf{\Omega}_0^{\mathrm{T}}$ with essentially known $\mathbf{M}_0$, while for the reference-free approach, we estimated $\mathbf{M}_1$ in the context of fitting $\mathbf{Y} = \mathbf{M}_1\mathbf{\Omega}_1^{\mathrm{T}}$. We note that, in general, we did not anticipate $\mathbf{\Omega}_0$ and $\mathbf{\Omega}_1$ to be equal. The reason is that the unsupervised, reference-free approach will find only the major axes of variation within a given data set, not necessarily all relevant distinctions of major cell types. For example, if a data set consists of only two distinct immune profiles (with very little variation among the subjects sharing a profile), then the reference-free approach will typically find only two cell types, those corresponding to each profile. However, $\mathbf{M}_0$ and $\mathbf{M}_1$ should be related to by a mixing matrix $\mathbf{\Psi}$ that reassigns the "correct" cell types to the unsupervised decomposition, i.e. $\mathbf{M}_1 = \mathbf{M}_0\mathbf{\Psi}^{\mathrm{T}}$. It follows that $\mathbf{\Omega}_0 \approx \mathbf{\Omega}_1\mathbf{\Psi}$, thus phenotypic associations with $\mathbf{\Omega}_0$ should match those with $\mathbf{\Omega}_1\mathbf{\Psi}$. Figures S8.1 and S8.2 depict the mixing matrices $\mathbf{\Psi}$, obtained by constrained projection, for *BL-ra* and *BL-as*, respectively. Though our essentially unsupervised approach resulted in cell proportion estimates $\mathbf{\Omega}_1$ quite distinct from $\mathbf{\Omega}_0$, for both *BL-ra* and *BL-as* the reference-based solution $\mathbf{\Omega}_0$ was nevertheless reasonably similar to a linear combination $\mathbf{\Omega}_1\mathbf{\Psi}$ of the unsupervised solution $\mathbf{\Omega}_1$ (Figures S8.3 and S8.4). Phenotypic associations with the "re-mixed" $\mathbf{\Omega}_1\mathbf{\Psi}$ cell proportion estimates were remarkably similar to associations with the reference-based solution $\mathbf{\Omega}_0$ (Figures S8.5 and S8.6), with only one notable reversal: in the *BL-ra* dataset, the relative magnitudes of CD4+ and CD8+ coefficients were reversed, but all were still significantly and negatively associated with rheumatoid arthritis status.

It follows that if $\mathbf{M}_1$ contains information on immune function not readily apparent from $\mathbf{M}_0$, then the information should be evident in the residual matrix $\mathbf{M}_1 - \mathbf{M}_0\mathbf{\Psi}^{\mathrm{T}}$. In fact, a number of CpGs still showed substantial heterogeneity among the residual methylomes $\mathbf{M}_1$ in comparison with the residual methylomes $\mathbf{M}_1 - \mathbf{M}_0\mathbf{\Psi}^{\mathrm{T}}$ (Figures S8.7 and S8.8). The residual methylomes

$\mathbf{M}_1 - \mathbf{M}_0 \mathbf{\Psi}^{\mathrm{T}}$, which reflect residual epigenetic information suggestive of cell-type heterogeneity but unaccounted for by the reference methylomes $\mathbf{M}_0$, displayed substantially diminished association with the DMPs based on Roadmap WGBS data, compared with the unadjusted methylomes $\mathbf{M}_1$ (Figure S8.9 vs Figure S8.10). However, with loci mapped to genes known to reflect immune activation or regulation, both $\mathbf{M}_1$ and $\mathbf{M}_1 - \mathbf{M}_0 \mathbf{\Psi}^{\mathrm{T}}$ typically displayed similar heterogeneity across constituent methylomes (Figures S8.11 and S8.12). In particular, they identified two strongly significant processes in the rheumatoid arthritis data set, *Th1 & Th2 differentiation* and *T-Cell Polarization* (Figure S8.13), while for arsenic exposure in Bangladeshi neonates, they identified one strongly significant process, *Regulators of T-Cell Activation* (Figure S8.14). These results were somewhat dissimilar from results obtained using the reference-based cell proportions $\mathbf{\Omega}_0$ in *limma* to adjust for cell type (Figures S8.15 through S8.17). In particular, the limma-based methods found significant associations for the less specific gene set *T-Cell Differentiation*, but not for *Th1 & Th2 differentiation*. Thus, subtle immune effects may be more readily apparent from the row variances of $\mathbf{M}_1$ than from the methylation associations obtained by adjusting for the reference-based cell proportions.

Additional Analyses on Datasets with Normal and Pathological Tissue

Finally, in an analysis of cell proportions $\mathbf{\Omega}$ obtained using the Roadmap-derived methylomes as a reference, notable distinctions between normal and pathological tissues were revealed (Figures S9.1 through S9.3). In particular, gastric tumors differed from normal gastric tissues in having greater immunological/inflammation content but lesser gastrointestinal content, atherosclerotic carotid (and to some extent atherosclerotic aorta) differed from normal aorta in having greater immunological/inflammation content but lesser muscular content, and cirrhotic tissues differed from normal liver tissues in having greater immunological/inflammation content but lesser gastrointestinal content (with the pattern more striking for cirrhotic tissues related to viral infection than for cirrhotic tissues related to alcohol abuse).

Discussion

We have proposed a simple method for reference-free deconvolution that provides both interpretable outputs, i.e. proportions of putative cell types defined by their underlying methylomes, as well as a method for evaluating the extent to which the underlying reflect specific types of cells. We have demonstrated these methods in a wide array of methylation datasets in various tissues and focused on differing exposures or outcomes.

Overall our deconvolution approach is similar to many others that have been proposed[29-33]. In particular, it is very similar to a recent publication that applied a convex-mixtures approach to deconvolve RNA expression[33]. Our approach differs from this one in that it deconvolves DNA methylation, with a corresponding constraint on the values of $\mathbf{M}$; in addition, importantly, we have provided a more comprehensive approach for interpreting the resulting columns of $\mathbf{M}$.

We have provided a novel approach for estimating the number $\hat{K}$ of cell types, which we have shown to reflect the level of cellular heterogeneity anticipated from each tissue we analyzed.

More heterogeneous tissues (blood, breast, and gastric tissues) resulted in higher values of $\hat{K}$, while the more homogeneous tissues had lower values: $\hat{K}=1$ for the (admittedly small) isolated lymphatic and blood vessel endothelium data sets, while $\hat{K}=2$ for sperm and umbilical cord endothelial tissues. Note, however, that Figure 2b reflects ambiguity in the choice of $\hat{K}$ for sperm and equally supports the choice $\hat{K}=1$; similar plots shown in Figure 2b unambiguously suggest $\hat{K}>1$ for two blood data sets and for an artery data set. A similar plot for UV-as (not shown) displayed an unambiguous preference for the choice of $\hat{K}=2$, but the two putative types of cells did not associate with any metadata (Table 2). Taken together, these results demonstrate that our algorithm returns reasonably reliable values of $\hat{K}$ reflecting cellular heterogeneity.

Cell mixture proportions $\mathbf{\Omega}$ were typically significantly associated with major phenotypes of interest, with the notable exception of sperm, umbilical vein endothelium, and placental artery (the former two assumed to be homogenous tissues). Thus, the radically dimension-reduced DNA methylation data in $\mathbf{\Omega}$ can still retain strongly significant associations with major phenotypes of interest. However, for other covariates, especially demographic confounders, there was considerable variation in significance, demonstrating that $\mathbf{\Omega}$ can also show null associations with some covariates. Taken together, the results show that $\mathbf{\Omega}$ can distinguish signal from noise. As the limma analysis demonstrated, residual signal can still exist in $\mathbf{Y}$ even after adjusting for $\mathbf{\Omega}$, although often in a more diminished capacity. In a few rare cases, the signal increased after adjusting for $\mathbf{\Omega}$. Taken together, these results suggest that a substantial proportion of the association between $\mathbf{Y}$ and phenotypic metadata $\mathbf{X}$ can be factored through the decomposition $\mathbf{Y}=\mathbf{M}\mathbf{\Omega}^{\mathrm{T}}$, occasionally clarifying the residual signal, but more often diminishing it. This finding is significant, as it would strongly suggest that the results of the vast majority of EWAS studies are driven by physiologic changes of the underlying composition of cells within the samples obtained. This is nicely highlighted by a recent report identifying a specific cell type driving the associations between smoking and changes in DNA methylation in peripheral blood[23]. This is in contrast to the prevailing current interpretation of most findings, which has aligned more strongly with the concept of metastable epialleles. These alleles represent loci where environmental conditions during development dictate 'setpoints' for the levels of methylation at particular gene sequences that are consistent across tissues within any person, also yielding differences in concordant gene expression[45-48]. The methods described in our work may have some utility for future discovery of these alleles in that within-person, cross-tissue comparisons of methylation profiles would be expected to be enriched for metastable alleles when the loci that are reflective of subsets of cell types are described and removed from comparisons.

On the other hand, we demonstrated that columns of $\mathbf{M}$ correlate with external biological annotation data in a manner concordant with their interpretation as methylomes specific to constituent cell types. *Blood DMP* status showed the highest associations in blood data sets, although also somewhat high associations in *L[np]* and *AR[np]*, data sets having tissues with potentially inflammatory components to pathology. Data sets with tumors (*BR-tcga[t]*, *br-1[t]*, *br-2[t]*, *br-3[t]*, *g[nt]*, and *g[t]*) demonstrated high associations with *PcG Loci* , reflecting the

mitotic activity of tumors; that normal gastric tissue, *g[n]* also showed a high association with *PcG Loci* is consistent with the high level of cellular turnover in gastric tissues. The Bilenky DMRs showed the strongest associations for breast tissues, consistent with the fact that the Bilenky DMRs were obtained from breast tissue, but also demonstrated strong associations for liver pathologies, and in arterial tissues *AR[n]* and *AR[np]*. Breast and arterial tissues have a mix of epithelial and smooth muscle tissues, which may explain the arterial results. The association in *L[np]* may reflect the fibrous character of pathological liver tissue. REMC DMRs demonstrated only weak correlation or strong negative correlation with all tissues, perhaps reflecting the embryonic/developmental nature of the REMC DMRs.

Comparisons with DMPs constructed from Roadmap WGBS data also demonstrated that the columns of $\mathbf{M}$ reflect epigenetic content concordant with anatomical expectations; in particular, blood datasets displayed associations with DMPs distinguishing CD34+ hematopoietic stem cells vs. thymus or spleen, as were datasets *L[np]* and *AR[np]*, which both included tissue pathologies involving inflammation and immune response. Arterial data sets displayed associations with DMPs distinguishing smooth muscle from endothelium. Associations with Roadmap-based DMPs were typically weak for homogeneous tissues, in particular sperm. Interestingly, the normal liver tissue data set *L[n]* had mostly weak negative associations; one possible explanation is that the primary tissues available from the Roadmap were too dissimilar from normal liver tissue to distinguish subtle anatomical features. Using the Roadmap data as a pseudo-reference, normal and pathological tissues were revealed to differ anatomically along anticipated lines, specifically in that pathological tissues had greater cellular content reflective of immune or inflammation processes, and lesser gastrointestinal content (gastric and liver tissue) or muscular content (arterial tissue). Taken together, these results suggest that the columns of $\mathbf{M}$ reflect methylomes of constituent cell types.

We remark that the unsupervised deconvolution approach we have proposed cannot be guaranteed to recover the methylomes of all constituent cell types; instead, it recovers the major axes of cellular variation. This was evident in the comparison of reference-based and reference-free deconvolution of blood datasets *BL-ra* and *BL-as*; for these datasets, the reference-free approach recovered the linear combination of reference methylomes most relevant to characterizing the underlying variation. However, when “re-mixed” back to proportions of known cell types using a reference methylome, associations with phenotypic metadata were consistent with those obtained from reference-based deconvolution. While on its surface this suggests that the reference-free approach has no value when a reference methylome is known (as is the case with blood), further analysis of the residual information in the unsupervised deconvolution demonstrated that reference-free deconvolution may reveal distinctions in cell type relevant to characterizing the underlying variation in the dataset but more subtle than the potential distinctions fixed in advance by the reference set. For example, the unsupervised approach identified two strongly significant processes in the rheumatoid arthritis data set, *Th1 & Th2 differentiation* and *T-Cell Polarization*; this finding is consistent with known Th1/Th2 differentiation processes[49, 50] and T-Cell polarization process[51, 52] involved in the etiology of rheumatoid arthritis. Similarly, the unsupervised approach identified *Regulators of T-Cell Activation* as a significant process in the dataset investigating arsenic exposure in Bangladesh. In fact, the impact of arsenic exposure on regulation of T-cells has been noted[53],

and even observed in another study conducted in Bangladesh[54]. We remark that we obtained somewhat different results using reference-based cell proportions $\mathbf{\Omega}_0$ in *limma* to adjust for cell type. Thus, the reference-free approach may provide important information that complements a reference-based approach.

As a general point, we have demonstrated the links suggested in Figure 1; thus, we have shown that it is possible to use a reference-free approach to characterize the extent to which phenotypic associations with DNA methylation data are explained by differences in constituent cell types. We remark that such distinctions may be subtle, such as variation in smooth muscle content or the presence of leukocytes with specialized immunological states. There may still exist associations residual to those with variations in putative underlying cell types, although they will often be diminished after adjusting for cell type in the manner we have proposed. Other reference-free approaches can also distinguish between associations driven by variation in cell type and those that are more focal to individual CpG sites, but our proposed method has several advantages over these existing methods. The first is that it is not particularly intensive computationally; the second is that it provides an easy and interpretable way to estimate the underlying number of constituent cell types; the third is that it provides estimates of cell-proportions that are directly interpretable and comparable with estimates obtained from reference data sets. In particular, our method provides a means for extracting information that is more subtle than that available from reference data sets but may nevertheless reflect additional variation in constituent cell types. While similar insights may be obtained simply by examining the CpG-specific associations, we note that there is ongoing controversy on what "adjustment for cell-type" means in the context of EWAS analysis. We have previously argued that all epigenetic variation is ultimately mediated by cell-type, if the meaning of "cell-type" is conceived of broadly enough[20]; a more useful framing of the question is how to identify types of cells that are relevant to the biological variation being studied. Our proposed approach helps in partitioning the underlying variation into units that resemble cell-specific methylomes, so that these methylomes or the overt functional characteristics of these cells may be further analyzed using additional biological characterization data.

We remark on a few current limitations of our approach. One is that we have used a crude gene-set procedure based on variance, which removes "signed" information and thus precludes the use of algorithms based on expression signature, such as CTen[55]. Another related limitation is a lack of relevant annotation data. Further work is necessary to adapt the method we have proposed here to "signed" comparisons, thus enabling a wider array of annotation tools, and to develop other relevant annotation datasets relevant to identifying subtle cell types.

Methods

Empirical Examination of Proposed Methods

We removed chromosome Y data from all datasets; and we also removed chromosome X data from all but the breast datasets. For the 450K data sets downloaded from TCGA, and for the 450K data collected to investigate associations with arsenic exposure in Bangladeshi neonates[3, 9], we used the *FunNorm* algorithm (*Bioconductor* package *minfi*) to process the raw *idat* files;

we obtained all other data sets as processed average beta values from Gene Expression Omnibus (GEO). For 450K data sets, we excluded CpGs with cross-hybridizing probes or probes with SNPs[56], and used the *BMIQ* algorithm[57] (*Bioconductor* package *wateRmelon*) to align the scales of Type I and Type II probes. Finally, for each data set, we excluded CpGs having missing measurements for over half the specimens.

Associations with Phenotypic Metadata

As described in Supplementary Section S4, permutation tests were used to assess omnibus significance of covariates $\mathbf{X}$ with fitted cell proportions $\mathbf{\Omega}$. As described in Supplementary Section S5, we further compared associations of $\mathbf{Y}$ with $\mathbf{X}$ before and after including terms from $\mathbf{\Omega}$ in the regression model for $\mathbf{Y}$, using the *limma* procedure[58] (via the R package *limma*) to compute regression coefficients, using the R package *qvalue* to estimate both q-values and the overall proportion π_0 of null associations.

Interpretation of Putative Cell Types

We obtained a list of DMPs for differentiating distinct major types of leukocytes (*Blood DMPs*) from the Reinius reference set[25], and constructed a set of CpGs mapped to genes considered Polycomb Group proteins (*PcG loci*), compiled from four references[59-62] as in our previous articles[20, 27]. We also constructed a set of CpGs based on differentially methylated regions (DMRs) obtained from WGBS data collected by the Epigenomics Roadmap Project. Supplementary Section S6 describes the details of the construction of these DMP sets. In addition, we developed a novel approach based on WGBS data from the Roadmap Epigenomics Project for 24 primary tissues, described in detail in Supplementary Section S7.

Additional Analyses on 450K Blood Datasets

To compare reference-based analysis with our proposed approach, we analyzed the two 450K blood data sets, *BL-ra* and *BL-as*, estimating for each data set two sets of cell-type proportion matrices ($K = 7$): $\mathbf{\Omega}_0$ (reference-based) and $\mathbf{\Omega}_1$ (reference-free). Details appear in Supplementary Section S8. Briefly, to obtain the mixing matrix $\mathbf{\Psi}$ that relates matrices $\mathbf{M}_0$ and $\mathbf{M}_1$, we used a constrained projection similar to that used to obtain the reference-based cell proportion matrix $\mathbf{\Omega}_0$, and compared $\mathbf{M}_1$ with $\mathbf{M}_1 - \mathbf{M}_0\mathbf{\Psi}^{\mathrm{T}}$ by identifying CpGs with high variation across their constituent methylomes. In addition, we compared these highly varying CpGs with with immune activation and immune regulation pathways compiled from six sources[63-69].

Additional Analyses on Datasets with Normal and Pathological Tissue

We projected Infinium data from each of the three datasets sets *g[nt]*, *AR[np]*, and *L[np]* onto the profile matrix $\mathbf{M}$ obtained by decomposing the Roadmap WGBS data ($\mathbf{Y} = \mathbf{M}\mathbf{\Omega}^{\mathrm{T}}$); we then averaged the resulting specimen-specific cell proportions $\mathbf{\Omega}$ over tissue status (normal gastric tissue vs. gastric tumor, normal aorta vs. atherosclerotic aorta and atherosclerotic carotid, and

normal liver vs. alcohol-related cirrhotic liver and cirrhotic liver due to viral infection). Details and results appear in Section S9.

Acknowledgements

Funding for this work was provided by grants from the National Institutes for Health: NIMH R01MH094609 (EAH and CJM), NIEHS P01 ES022832 (CJM), K01 ES017800 (MLK), R01ES024991 (EAH, TAI), R01ES015533 (DC). Funding was also provided by EPA grant RD83544201 (CJM).

References

1. Houseman, E.A., Kim, S., Kelsey, K.T. & Wiencke, J.K. DNA Methylation in Whole Blood: Uses and Challenges. *Current environmental health reports* **2**, 145-154 (2015).
2. Herbstman, J.B. et al. Predictors and consequences of global DNA methylation in cord blood and at three years. *PloS one* **8**, e72824 (2013).
3. Kile, M.L. et al. Effect of prenatal arsenic exposure on DNA methylation and leukocyte subpopulations in cord blood. *Epigenetics : official journal of the DNA Methylation Society* **9**, 774-782 (2014).
4. Koestler, D.C., Avissar-Whiting, M., Houseman, E.A., Karagas, M.R. & Marsit, C.J. Differential DNA methylation in umbilical cord blood of infants exposed to low levels of arsenic in utero. *Environ Health Perspect* **121**, 971-977 (2013).
5. Smith, A.K. et al. DNA extracted from saliva for methylation studies of psychiatric traits: evidence tissue specificity and relatedness to brain. *American journal of medical genetics. Part B, Neuropsychiatric genetics : the official publication of the International Society of Psychiatric Genetics* **168B**, 36-44 (2015).
6. Agha, G. et al. Adiposity is associated with DNA methylation profile in adipose tissue. *Int J Epidemiol* **44**, 1277-1287 (2015).
7. Busche, S. et al. Population whole-genome bisulfite sequencing across two tissues highlights the environment as the principal source of human methylome variation. *Genome biology* **16**, 290 (2015).
8. Banister, C.E. et al. Infant growth restriction is associated with distinct patterns of DNA methylation in human placentas. *Epigenetics : official journal of the DNA Methylation Society* **6**, 920-927 (2011).
9. Cardenas, A. et al. In utero arsenic exposure and epigenome-wide associations in placenta, umbilical artery, and human umbilical vein endothelial cells. *Epigenetics : official journal of the DNA Methylation Society* **10**, 1054-1063 (2015).
10. Maccani, J.Z.J. et al. DNA methylation changes in the placenta are associated with fetal manganese exposure. *Reprod Toxicol* **57**, 43-49 (2015).
11. Maccani, J.Z.J. et al. Placental DNA Methylation Related to Both Infant Toenail Mercury and Adverse Neurobehavioral Outcomes. *Environ Health Persp* **123**, 723-729 (2015).
12. Santagata, S. et al. Taxonomy of breast cancer based on normal cell phenotype predicts outcome. *The Journal of clinical investigation* **124**, 859-870 (2014).

13. Houseman, E.A. & Ince, T.A. Normal cell-type epigenetics and breast cancer classification: a case study of cell mixture-adjusted analysis of DNA methylation data from tumors. *Cancer informatics* **13**, 53-64 (2014).
14. Roadmap Epigenomics, C. et al. Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317-330 (2015).
15. Christensen, B.C. et al. Aging and environmental exposures alter tissue-specific DNA methylation dependent upon CpG island context. *PLoS Genet* **5**, e1000602 (2009).
16. Ji, H. et al. Comprehensive methylome map of lineage commitment from haematopoietic progenitors. *Nature* **467**, 338-342 (2010).
17. Khavari, D.A., Sen, G.L. & Rinn, J.L. DNA methylation and epigenetic control of cellular differentiation. *Cell Cycle* **9**, 3880-3883 (2010).
18. Natoli, G. Maintaining cell identity through global control of genomic organization. *Immunity* **33**, 12-24 (2010).
19. Houseman, E.A. et al. DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC bioinformatics* **13**, 86 (2012).
20. Houseman, E.A., Kelsey, K.T., Wiencke, J.K. & Marsit, C.J. Cell-composition effects in the analysis of DNA methylation array data: a mathematical perspective. *BMC bioinformatics* **16** (2015).
21. Jaffe, A.E. & Irizarry, R.A. Accounting for cellular heterogeneity is critical in epigenome-wide association studies. *Genome biology* **15**, R31 (2014).
22. Liu, Y. et al. Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis. *Nat Biotechnol* **31**, 142-147 (2013).
23. Bauer, M. et al. A varying T cell subtype explains apparent tobacco smoking induced single CpG hypomethylation in whole blood. *Clinical epigenetics* **7**, 81 (2015).
24. Accomando, W.P. et al. Decreased NK cells in patients with head and neck cancer determined in archival DNA. *Clin Cancer Res* **18**, 6147-6154 (2012).
25. Reinius, L.E. et al. Differential DNA methylation in purified human blood cells: implications for cell lineage and studies on disease susceptibility. *PloS one* **7**, e41361 (2012).
26. Zou, J., Lippert, C., Heckerman, D., Aryee, M. & Listgarten, J. Epigenome-wide association studies without the need for cell-type composition. *Nat Methods* **11**, 309-311 (2014).
27. Houseman, E.A., Molitor, J. & Marsit, C.J. Reference-Free Cell Mixture Adjustments in Analysis of DNA Methylation Data. *Bioinformatics* (2014).
28. Shen-Orr, S.S. & Gaujoux, R. Computational deconvolution: extracting cell type-specific information from heterogeneous samples. *Current opinion in immunology* **25**, 571-578 (2013).
29. Erkkila, T. et al. Probabilistic analysis of gene expression measurements from heterogeneous tissues. *Bioinformatics* **26**, 2571-2577 (2010).
30. Gaujoux, R. & Seoighe, C. Semi-supervised Nonnegative Matrix Factorization for gene expression deconvolution: a case study. *Infection, genetics and evolution : journal of molecular epidemiology and evolutionary genetics in infectious diseases* **12**, 913-921 (2012).
31. Quon, G. et al. Computational purification of individual tumor gene expression profiles leads to significant improvements in prognostic prediction. *Genome medicine* **5**, 29 (2013).
32. Teschendorff, A.E., Zhuang, J. & Widschwendter, M. Independent surrogate variable analysis to deconvolve confounding factors in large-scale microarray profiling studies. *Bioinformatics* **27**, 1496-1505 (2011).
33. Wang, N. et al. Mathematical modelling of transcriptional heterogeneity identifies novel markers and subpopulations in complex tissues. *Scientific reports* **6**, 18909 (2016).
34. Langevin, S.M. et al. Peripheral blood DNA methylation profiles are indicative of head and neck squamous cell carcinoma: an epigenome-wide association study. *Epigenetics : official journal of the DNA Methylation Society* **7**, 291-299 (2012).

35. Teschendorff, A.E. et al. An epigenetic signature in peripheral blood predicts active ovarian cancer. *PloS one* **4**, e8274 (2009).
36. Krausz, C. et al. Novel insights into DNA methylation features in spermatozoa: stability and peculiarities. *PloS one* **7**, e44479 (2012).
37. Bronneke, S. et al. DNA methylation regulates lineage-specifying genes in primary lymphatic and blood endothelial cells. *Angiogenesis* **15**, 317-329 (2012).
38. Zaina, S. et al. A DNA Methylation Map of Human Atherosclerosis. *Circulation: Cardiovascular Genetics*, CIRCGENETICS. 113.000441 (2014).
39. Hlady, R.A. et al. Epigenetic signatures of alcohol abuse and hepatitis infection during human hepatocarcinogenesis. *Oncotarget* **5** (2014).
40. Fackler, M.J. et al. Genome-wide methylation analysis identifies genes specific to breast cancer hormone receptor status and risk of recurrence. *Cancer Res* **71**, 6195-6207 (2011).
41. Zhuang, J. et al. The dynamics and prognostic potential of DNA methylation changes at stem cell gene loci in women's cancer. *PLoS genetics* **8**, e1002517 (2012).
42. Zouridis, H. et al. Methylation subtypes and large-scale epigenetic alterations in gastric cancer. *Science translational medicine* **4**, 156ra140-156ra140 (2012).
43. Dedeurwaerder, S. et al. DNA methylation profiling reveals a predominant immune component in breast cancers. *EMBO Mol Med* **3**, 726-741 (2011).
44. Cancer Genome Atlas, N. Comprehensive molecular portraits of human breast tumours. *Nature* **490**, 61-70 (2012).
45. Dolinoy, D.C., Das, R., Weidman, J.R. & Jirtle, R.L. Metastable epialleles, imprinting, and the fetal origins of adult diseases. *Pediatric research* **61**, 30R-37R (2007).
46. Harris, R.A., Nagy-Szakal, D. & Kellermayer, R. Human metastable epiallele candidates link to common disorders. *Epigenetics : official journal of the DNA Methylation Society* **8**, 157-163 (2013).
47. Rakyan, V.K., Blewitt, M.E., Druker, R., Preis, J.I. & Whitelaw, E. Metastable epialleles in mammals. *Trends in genetics : TIG* **18**, 348-351 (2002).
48. Silver, M.J. et al. Independent genomewide screens identify the tumor suppressor VTRNA2-1 as a human epiallele responsive to periconceptional environment. *Genome biology* **16**, 118 (2015).
49. Diehl, S. & Rincon, M. The two faces of IL-6 on Th1/Th2 differentiation. *Molecular immunology* **39**, 531-536 (2002).
50. Schulze-Koops, H. & Kalden, J.R. The balance of Th1/Th2 cytokines in rheumatoid arthritis. *Best practice & research. Clinical rheumatology* **15**, 677-691 (2001).
51. Gutcher, I. & Becher, B. APC-derived cytokines and T cell polarization in autoimmune inflammation. *The Journal of clinical investigation* **117**, 1119-1127 (2007).
52. Kim, C.H. et al. Rules of chemokine receptor association with T cell polarization in vivo. *The Journal of clinical investigation* **108**, 1331-1339 (2001).
53. Hernandez-Castro, B. et al. Effect of arsenic on regulatory T cells. *Journal of clinical immunology* **29**, 461-469 (2009).
54. Ahmed, S. et al. Arsenic-associated oxidative stress, inflammation, and immune disruption in human placenta and cord blood. *Environ Health Perspect* **119**, 258-264 (2011).
55. Shoemaker, J.E. et al. CTen: a web-based platform for identifying enriched cell types from heterogeneous microarray data. *BMC genomics* **13**, 460 (2012).
56. Chen, Y.-a. et al. Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray. *Epigenetics : official journal of the DNA Methylation Society* **8**, 203 (2013).
57. Teschendorff, A.E. et al. A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data. *Bioinformatics* **29**, 189-196 (2013).

58. Smyth, G.K. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol* **3**, 3 (2004).
59. Bracken, A.P., Dietrich, N., Pasini, D., Hansen, K.H. & Helin, K. Genome-wide mapping of Polycomb target genes unravels their roles in cell fate transitions. *Genes & development* **20**, 1123-1136 (2006).
60. Lee, T.I. et al. Control of developmental regulators by Polycomb in human embryonic stem cells. *Cell* **125**, 301-313 (2006).
61. Schlesinger, Y. et al. Polycomb-mediated methylation on Lys27 of histone H3 pre-marks genes for de novo methylation in cancer. *Nature genetics* **39**, 232-236 (2006).
62. Squazzo, S.L. et al. Suz12 binds to silenced regions of the genome in a cell-type-specific manner. *Genome research* **16**, 890-900 (2006).
63. Criscione, L.G. & Pisetsky, D.S. B lymphocytes and systemic lupus erythematosus. *Current rheumatology reports* **5**, 264-269 (2003).
64. Jaeckel, E. Animal models of autoimmune hepatitis. *Seminars in liver disease* **22**, 325-338 (2002).
65. Poindexter, N.J., Sahin, A., Hunt, K.K. & Grimm, E.A. Analysis of dendritic cells in tumor-free and tumor-containing sentinel lymph nodes from patients with breast cancer. *Breast cancer research : BCR* **6**, R408-415 (2004).
66. Ragde, H., Cavanagh, W.A. & Tjoa, B.A. Dendritic cell based vaccines: progress in immunotherapy studies for prostate cancer. *The Journal of urology* **172**, 2532-2538 (2004).
67. Tseng, S.Y. & Dustin, M.L. T-cell activation: a multidimensional signaling network. *Current opinion in cell biology* **14**, 575-580 (2002).
68. Wang, E., Panelli, M.C. & Marincola, F.M. Understanding the response to immunotherapy in humans. *Springer seminars in immunopathology* **27**, 105-117 (2005).
69. Tedder, T.F., Poe, J.C., Fujimoto, M., Haas, K.M. & Sato, S. The CD19-CD21 signal transduction complex of B lymphocytes regulates the balance between health and autoimmune disease: systemic sclerosis as a model system. *Current directions in autoimmunity* **8**, 55-90 (2005).

Figures

Figure 1. If associations between DNA methylation data $\mathbf{Y}$ and phenotypic metadata $\mathbf{X}$ factor through the decomposition $\mathbf{Y} = \mathbf{M}\mathbf{\Omega}^{\mathrm{T}}$, and the data in $\mathbf{M}$ serve to distinguish cell types by their associations with relevant annotation data, then associations between $\mathbf{X}$ and $\mathbf{Y}$ are explained in whole or in part by differences in the distribution of constituent cell types.

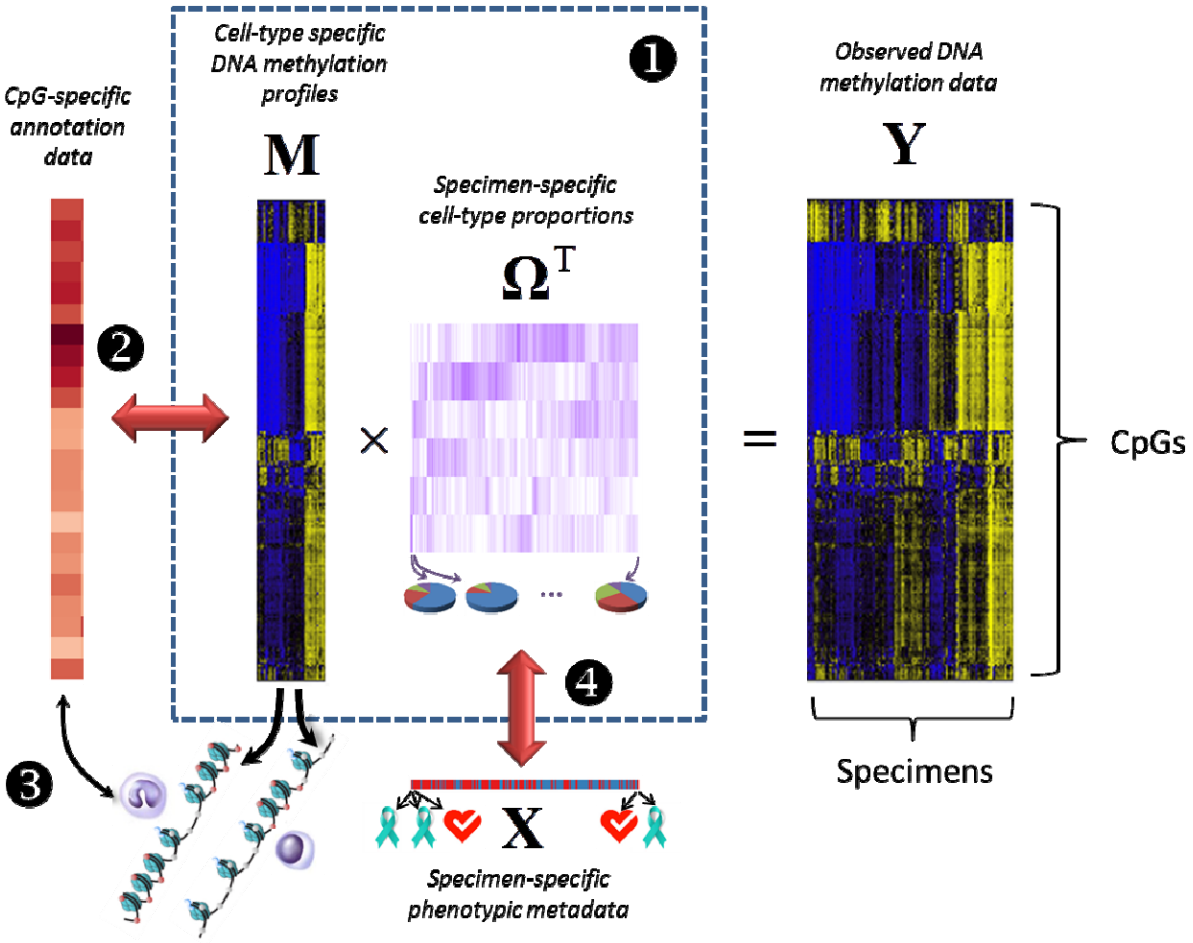


Figure 2. (A) Estimated number $\hat{K}$ of classes for each data set. (B) Bootstrapped deviance profiles for four selected data sets, along with mean deviance, median deviance, and quartiles for each value of K.

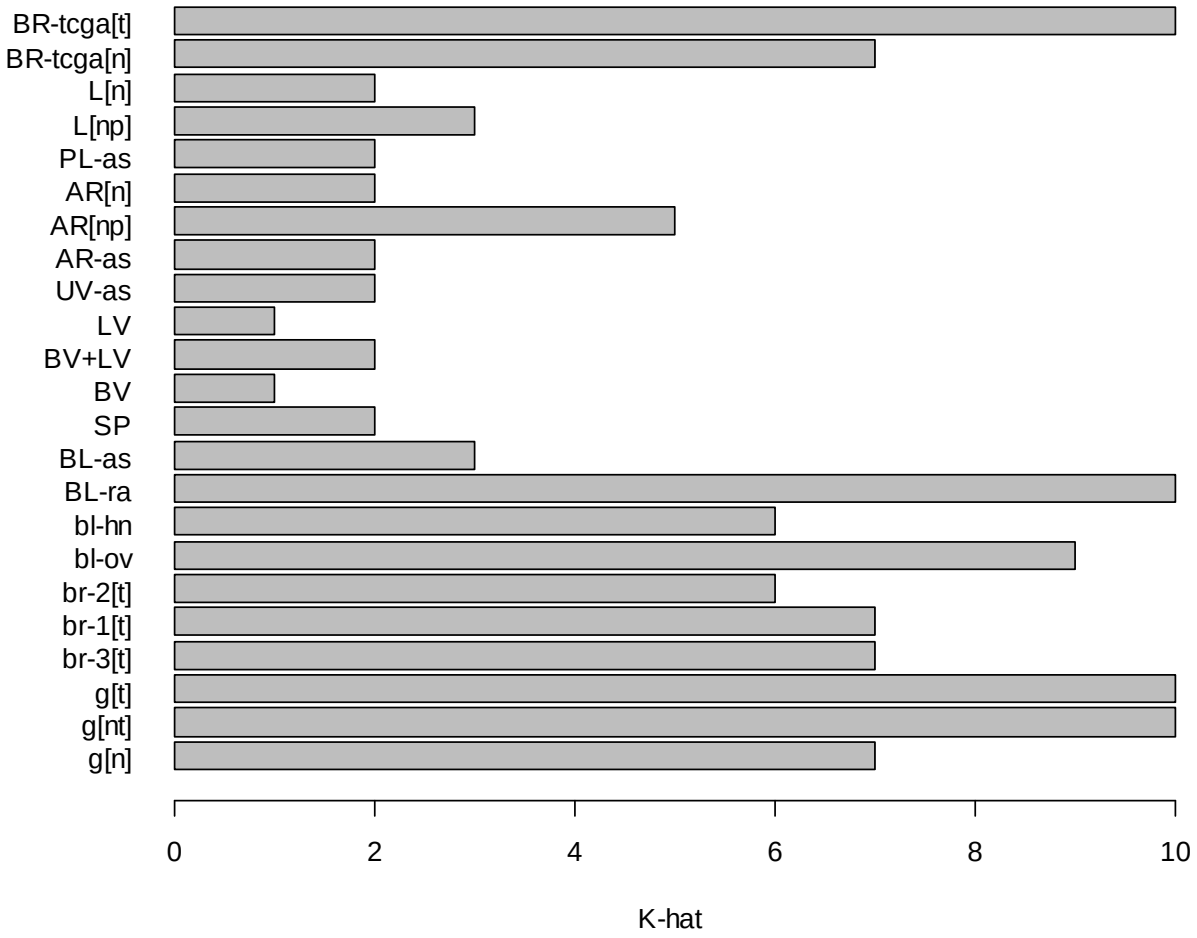


Figure 3. Clustering heatmaps of cell proportion matrix $\mathbf{\Omega}$ for two data sets; purple intensity indicates cell proportion. (A) Blood from rheumatoid arthritis cases and controls (*BL-ra,* $K = \hat{K} = 10$). (B) Sperm (*SP*, $K = \hat{K} = 2$).

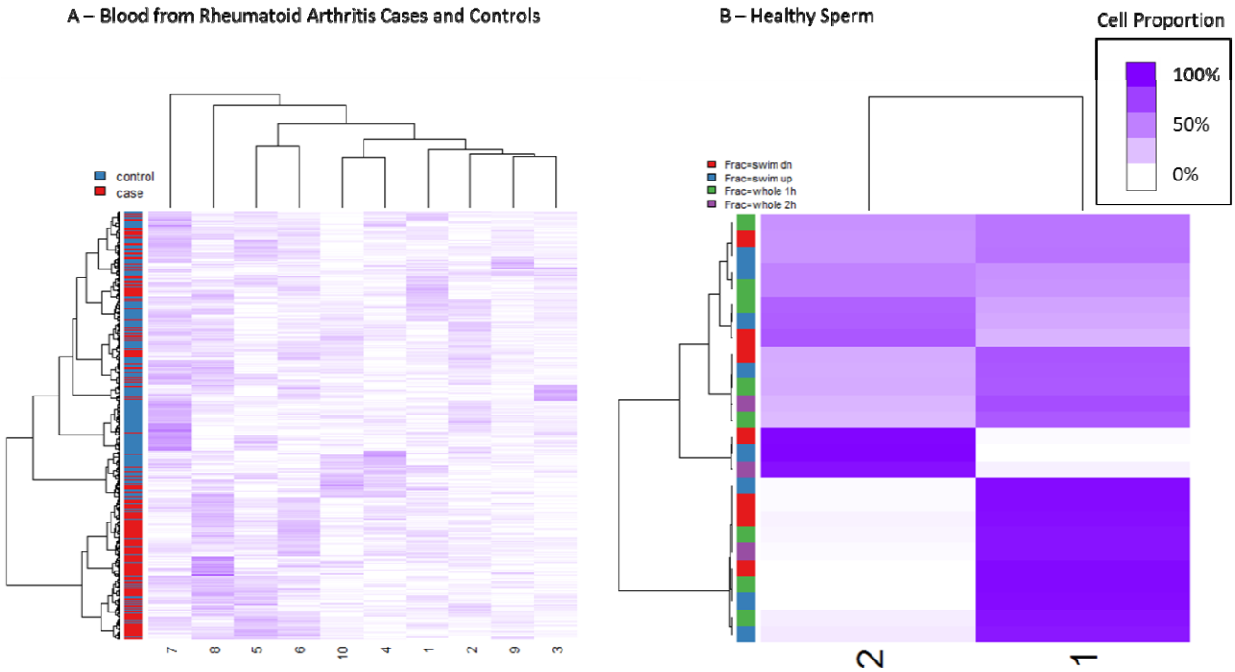


Figure 4. Comparison of π_0 (proportion of null association CpGs) from the $K = 1$ model with π_0 from the $K = K^*$ model; only non-demographic variables are shown.

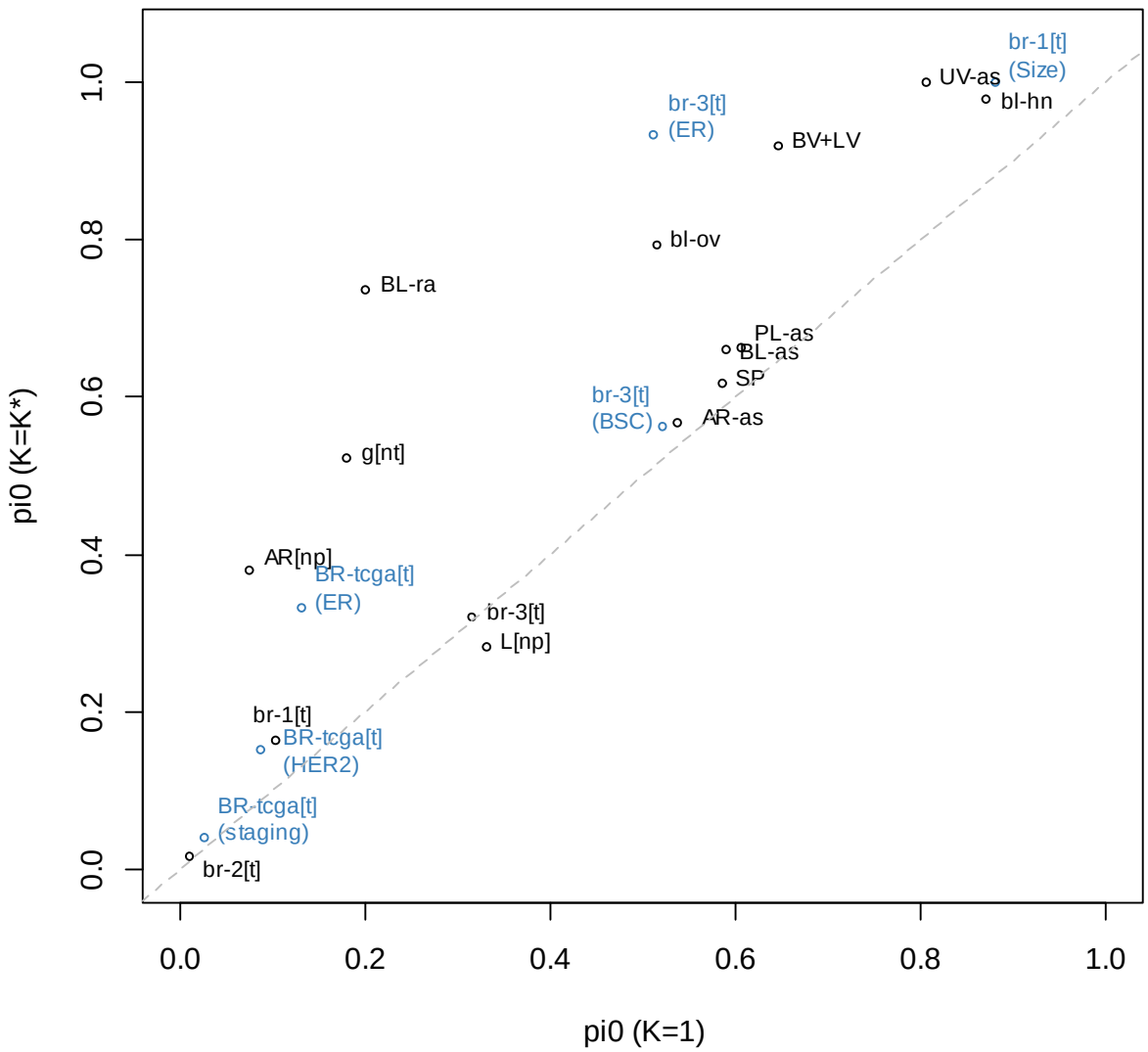


Figure 5. Gene-set odds ratios, showing the association of gene set membership with the set of CpGs whose values are highly variable across fitted methylomes ($s_j^2 > q_{0.75}(s^2)$). (A) Blood DMRs. (B) CpGs mapped to polycomb group protein genes.

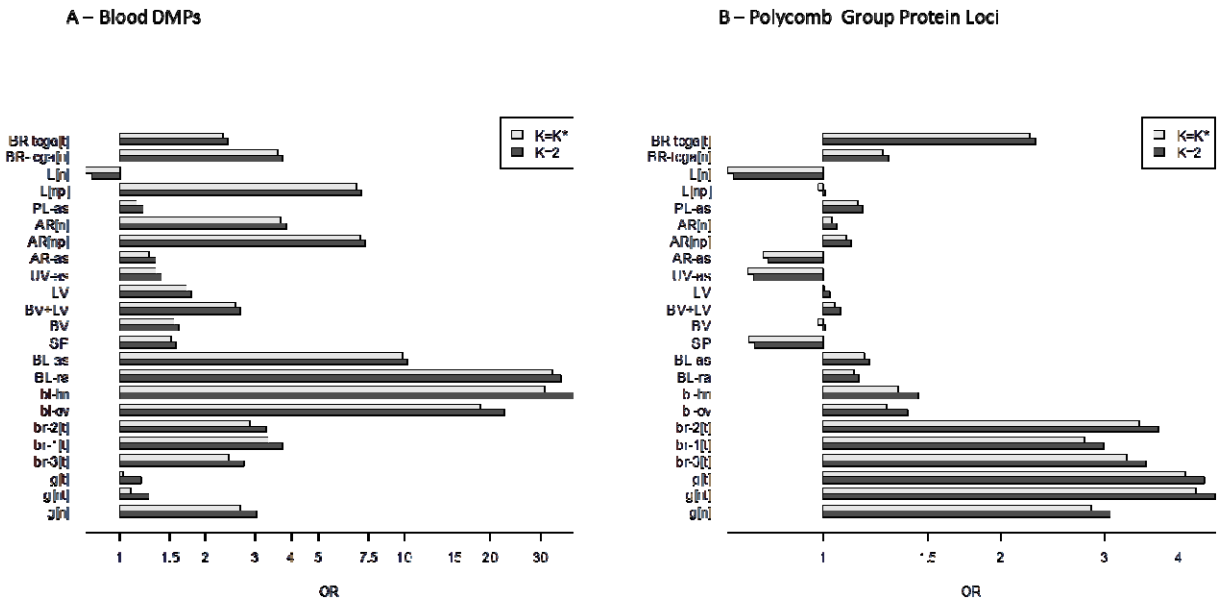
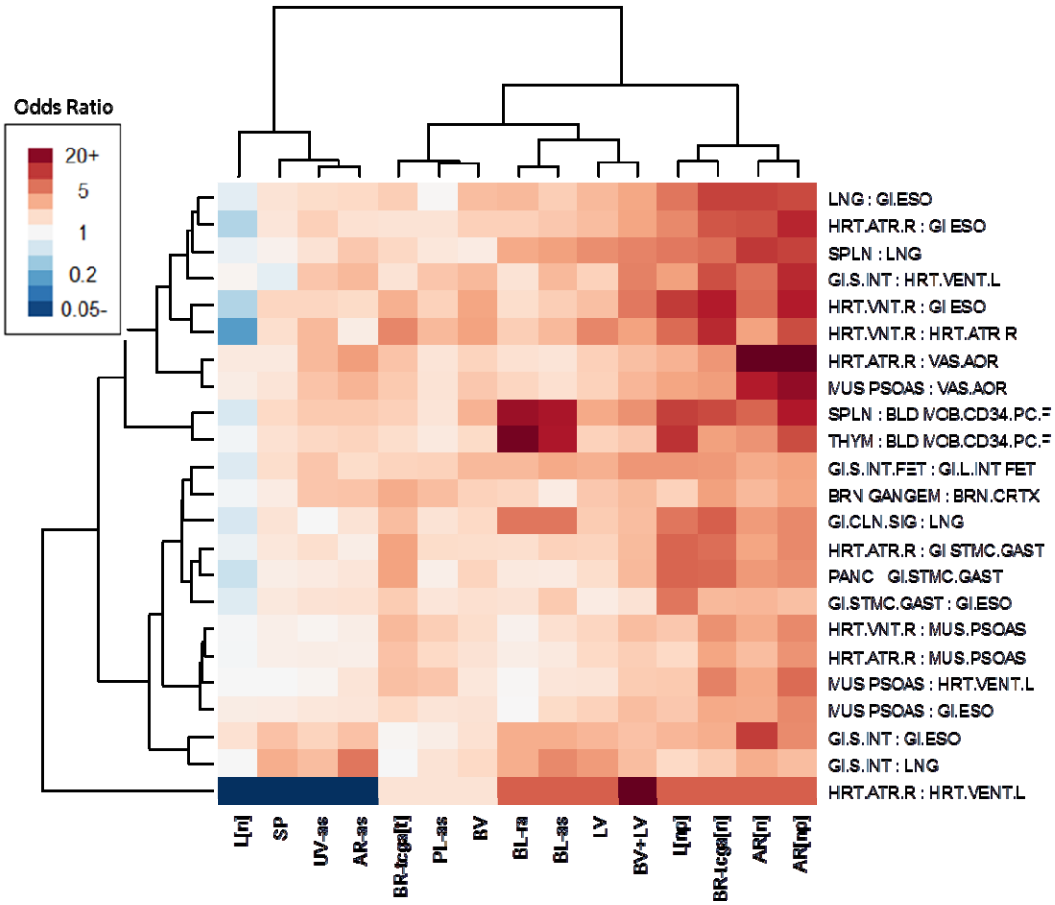


Figure 6. Gene-set odds ratios for 450K data sets, showing association of sets of DMPs distinguishing various Roadmap Epigenomics WGBS specimens with the set of CpGs whose values are highly variable across fitted methylomes ($s_j^2 > q_{0.75}(s^2)$).



Tables

Table 1 – Summary of Datasets

Code	Tissue	Source	Ref	Platform	Source Description	N	Covariate model
g[nt]	gastric tissue: tumor+normal	GEO: GSE30601	42	27K	203 gastric tumors and 94 matched gastric non-malignant samples.	297	Tumor[normal|tumor]
g[n]	gastric tissue: normal					94	-
g[t]	gastric tissue: tumor					203	-
br-1[t]	breast: tumor	GEO: GSE20712	43	27K	119 breast tumor samples with histological information. Removed 29 samples with ambiguous or missing histology.	119	Histology[basal|HER2|LumA|LumB] + Age[young|old] + Size[small|large]
br-2[t]	breast: tumor	GEO: GSE31979	40	27K	103 primary invasive breast tumors.	90	Histology[basal|ER-|ER+|HER2|LumA|LumB] + Age
br-3[t]	breast: tumor	GEO: GSE32393	41	27K	Breast tumor samples: 91 invasive ductal, 13 invasive lobular, 10 mucinous or medullary; 76 were ER+.	114	ER[ER-|ER+] + Histology[duct|lob|muc or med] + Age
bl-ov	peripheral blood	GEO: GSE19711	35	27K	Whole blood from 131 ovarian cancer cases (drawn pre-treatment) and 274 controls.	402	Case[control|ovarian cancer case] + Age
bl-hn	peripheral blood	GEO: GSE30229*	34	27K	Peripheral blood from 92 head and neck squamous cell carcinoma (HNSCC) patients and 92 controls. Removed 2 outlier cases.	182	Case[control|HNSCC case] + Age
BL-ra	peripheral blood	GEO: GSE42861	22	450K	Peripheral blood from 354 rheumatoid arthritis patients and 335 controls.	689	Case[control|arthritis case]
BL-as	cord blood	(not public)	3	450K**	Cord blood from 45 Bangladeshi neonates, with corresponding drinking water arsenic concentrations.	45	Log-arsenic + Sex[female|male]
SP	sperm	GEO: GSE47627	36	450K	26 normal sperm samples.	26	Fraction[swim down|swim up|whole 1h|whole 2h]
BV+LV	endothelial tissue	GEO: GSE34487	37	450K	16 vascular samples: 6 primary blood vessel endothelial cell samples and 10 primary lymphatic endothelial cell samples.	16	Source[BV|LV]
BV	endothelial tissue: blood vessel					6	-
LV	endothelial tissue: lymphatic vessel					10	-
UV-as	umbilical vein endothelial tissue	(not public)	9	450K**	Umbilical vein endothelial tissues from 51 Bangladeshi neonates, with corresponding drinking water arsenic concentrations.	51	Log-arsenic + Sex[female|male]
AR-as	placental artery	(not public)	9	450K**	Placental arteries from 46 Bangladeshi neonates, with corresponding drinking water arsenic concentrations.	46	Log-arsenic + Sex[female|male]
AR[np]	arterial tissue: atherosclerotic + normal	GEO: GSE46394	38	450K	15 normal aortic tissues, 15 atherosclerotic aortic lesions, 19 carotid atherosclerotic samples.	49	Source[normal|ath|carotid ath] + Sex[female|male] + Age
AR[n]	arterial tissue: normal aorta					15	-
PL-as	placenta	(not public)	9	450K**	Placentas from 45 Bangladeshi neonates, with corresponding drinking water arsenic concentrations.	45	Log-arsenic + Sex[female|male]
L[np]	liver tissue: cirrhotic + normal	GEO: GSE60753	39	450K	34 normal liver tissues, 21 cirrhotic tissues (due to alcoholism), 45 cirrhotic tissues [due to chronic hepatitis B (HBV) or C (HCV) viral	100	Source[normal|CirrEtOH|CirrV]
L[n]	liver tissue: normal					34	-
BR-tcga[n]	breast: normal	TCGA (11/2014)	44	450K*	96 normal breast tissues (matched to tumor) from The Cancer Genome Atlas, downloaded Nov. 2014	96	Age + Race[white|other]
BR-tcga[t]	breast: tumor			450K*	725 breast tumors from The Cancer Genome Atlas, downloaded Nov. 2014	725	Age + Race[white|other] + Staging[II+|III+|IV/X|?] + ER[ER+|ER-] + HER2[HER2+|HER2-|HER2?]

Table 2 – Inference with Phenotypic Metadata

Data Set	Permutaton P-values
g[nt]	Tumor<0.001
br-1[t]	Histology<0.001; Age=0.059; Size=0.016
br-2[t]	Histology<0.001; Age=0.06
br-3[t]	ER<0.001; Histology=0.295; Age=0.008; BSC=0.297
bl-ov	Case<0.001; Age=0.999
bl-hn	Case<0.001; Age<0.001
BL-ra	Case<0.001
BL-as	Log-arsenic<0.001; Sex=0.263
SP	Fraction=0.994
BV+LV	Source=0.013
UV-as	Log-arsenic=0.515; Sex=0.962
AR-as	Log-arsenic=0.285; Sex=0.505
AR[np]	Source<0.001; Sex=0.043; Age=0.377
PL-as	Log-arsenic=0.006; Sex=0.451
L[np]	Source<0.001
BR-tcga[n]	Age=0.089; Race=0.153
BR-tcga[t]	Age<0.001; Race<0.001; Staging=0.013; ER<0.001; HER2<0.001