Cell-type-specific resolution epigenetics without the need for cell sorting or single-cell biology

Elior Rahmani[1], Regev Schweiger[2], Brooke Rhead[3], Lindsey A. Criswell[4], Lisa F. Barcellos[3], Eleazar Eskin[1,5], Saharon Rosset[6], Sriram Sankararaman[1], Eran Halperin[1,5,7]

[1]Department of Computer Science, University of California, Los Angeles, Los Angeles, CA, USA

[2]Blavatnik School of Computer Science, Tel Aviv University, Tel Aviv, Israel

[3]School of Public Health, University of California, Berkeley, Berkeley, CA, USA

[4]Russell / Engleman Rheumatology Research Center, Department of Medicine, University of California, San Francisco, San Francisco, CA, USA

[5]Department of Human Genetics, University of California, Los Angeles, Los Angeles, CA, USA

[6]Department of Statistics, Tel Aviv University, Tel Aviv, Israel

[7]Department of Anesthesiology and Perioperative Medicine, University of California, Los Angeles, Los Angeles, CA, USA

High costs and technical limitations of cell sorting and single-cell techniques currently restrict the collection of large-scale, cell-type-specific DNA methylation data for a large number of individuals. This, in turn, impedes our ability to tackle key biological questions that pertain to variation within a population, such as identification of disease-associated genes at a cell-type-specific resolution. Here, we show mathematically and experimentally that cell-type-specific methylation levels of an individual can be learned from its tissue-level bulk data, as if the sample has been profiled with a single-cell resolution and then signals were aggregated in each cell population separately. Thus, our proposed approach provides an

unprecedented way to perform powerful large-scale epigenetic studies with cell-type-specific resolution using relatively easily obtainable large tissue-level data. We revisit previous studies with methylation and reveal novel associations with leukocyte composition in blood and multiple novel cell-type-specific associations with rheumatoid arthritis (RA). For the latter, further evidence demonstrates correlation of the associated CpGs with cell-type-specific expression of known RA risk genes, thus rendering our results consistent with the possibility that contributors to RA pathogenesis are regulated by cell-type-specific changes in methylation.

1 Introduction

Each cell type in the body of an organism performs a unique repertoire of required functions. Hence, disruption of cellular processes in particular cell types may lead to phenotypic alterations or development of disease. This presumption in conjunction with the complexity of tissue-level (“bulk”) data has led to many cell-type-specific genomic studies, in which genomic features, such as gene expression levels, are assayed from isolated cell types in a group of individuals and studied in the context of a phenotype or condition of interest (e.g., [1–4]).

In fact, in order to reveal cellular mechanisms affecting disease it is critical to study cell-type-specific effects. For example, it has been shown that cell-type-specific effects can contribute to our understanding of the principles of regulatory variation [5] and the underlying transcriptional landscape of heterogeneous tissues such as the human brain [6], it can provide a finer characterization of tumor heterogeneity [7,8], and it may reveal disease-related pathways and mechanisms of genes that were detected in genetic association studies [9,10]. Moreover, these findings are typically not revealed when a heterogeneous tissue is studied. For example, in [9] it has been shown that the FTO allele associated with obesity represses mitochondrial thermogenesis in adipocyte precursor cells. Particularly, in that study it is shown that the developmental regulators IRX3 and IRX5 had genotype-associated expression in primary preadipocytes, while genotype-associated expression was not observed in whole-adipose tissue, indicating that the effect was cell-type specific and restricted to preadipocytes.

In spite of the clear motivation to conduct studies with a cell-type-specific resolution, while developments in genomic profiling technologies have led to the availability of many large bulk data sets with hundreds or thousands of individuals (e.g., [11–13]), cell-type-specific data sets with a large number of individuals are still relatively scarce. Particularly, cell-type-specific studies are typically drastically restricted in their sample sizes owing to high costs and technical limitations imposed by both cell sorting and single-cell approaches. This restriction is especially profound for epigenetic studies with single-cell DNA methylation - while pioneering works on single-cell methylation have demonstrated significant advances (e.g. [14–17]), profiling methylation with single-cell resolution is still limited in coverage and throughput and currently cannot be practically used to routinely obtain large-scale data for population studies (the most eminent recent studies included data from only a few individuals). This, in turn, substantially limits our ability to tackle questions such as identification of disease-related altered regulation of genes in specific cell types and mapping of diseases to specific manifesting cell types.

Technologies for profiling single-cell methylation are currently still under development, and some of these attempts will potentially allow sometime in the future for the analysis of cell-type-specific methylation across or within populations. However, even if such technologies emerge in the near future, the large number of existing bulk methylation samples that have been collected by now are still an extremely valuable resource for genomic research (e.g., more than 100,000 bulk profiles to date in the Gene Expression Omnibus (GEO) alone [18]). These data reflect years of substantial community-wide effort of data collection from multiple organisms, tissues, and under different conditions, and it is therefore of great importance to develop new statistical approaches that can provide cell-type-specific insights from bulk data.

Here, we introduce Tensor Composition Analysis (TCA), a novel computational approach for learning cell-type-specific DNA methylation signals (a tensor of samples by methylation sites by cell-types) from a typical two-dimensional bulk data (samples by methylation sites). Conceptually, TCA emulates the scenario in which each sample in the bulk data has been profiled with a single-cell resolution and then signals were aggregated in each cell population separately.

We demonstrate the utility of TCA by applying it to data from previously published epigenome-wide association studies (EWAS). Particularly, we apply TCA to a previous large methylation study with rheumatoid arthritis (RA), in which DNA methylation profiles (CpG sites) were collected from cases and controls and tested for association with RA status [19]. Our analysis reveals novel cell-type-specific associations of methylation with RA without the need to collect cost prohibitive cell-type-specific data for a large number of individuals. Finally, we used independent data sets of cell-sorted methylation data to test the replicability of our results, and we provide additional independent evidence suggesting that some of the associated CpGs act as cell-type-specific regulators of expression in RA risk genes, thus shedding light on the cell-type specificity of RA pathogenesis.

2 Results

Different cell types are known to differ in their methylation patterns. Therefore, an individual bulk sample collected from a heterogeneous tissue represents a combination of different signals coming from the different cell types in the tissue. Since cell-type composition varies across individuals, testing for correlation between bulk methylation levels and a phenotype of interest may lead to spurious associations in case the phenotype is correlated with the cell-type composition [20]. A widely acceptable solution to this problem is to incorporate the cell-type composition information into the analysis of the phenotype by introducing it as covariates in a regression analysis. This approach results in an adjusted analysis which is conceptually similar to a study in which the cases and controls are matched on cell-type distribution. Even though this procedure is useful in order to eliminate spurious findings, it does not leverage the cell-type-specific signal, and thus results in a sever power loss as explained below.

Given no statistical relation between the phenotype and the cell-type composition, association studies typically assume a model with the following structure:

$$y = x\beta + \epsilon \tag{1}$$

Here, y represents the phenotype, x and β represent the bulk methylation level at a particular site under test and its corresponding effect size, and ϵ represents noise. This standard formulation assumes that a single parameter (β) describes the statistical relation between the phenotype and the bulk methylation level. We argue that this formulation is a major oversimplification of the nature of the underlying biology. In general, different cell types may have different statistical relations with the phenotype. Thus, a more realistic formulation would be:

$$y = \sum_{h=1}^{k} x_h \beta_h + \epsilon \tag{2}$$

Here, $x_1, ..., x_k$ are the methylation levels in each of the k cell types composing the studied tissue and $\beta_1, ..., \beta_k$ are their corresponding cell-type-specific effects.

Applying a standard analysis to bulk data may fail to detect even strong cell-type-specific associations with a phenotype. For instance, consider the scenario of a case/control study, where the methylation of one particular cell type is associated with the disease. In this scenario, due to the signals arising from other cell types, the observed bulk levels may obscure the real association and not demonstrate a difference between the cases and controls; importantly, in general, merely taking into account the variation in cell-type composition between individuals does not allow the detection of the association (Figure 1). Thus, allowing analysis with a cell-type-specific resolution (i.e. obtaining $x_1, ..., x_k$) - beyond its importance for revealing disease-manifesting cell types - is also crucial for the detection of true signals.

We consider a new model for DNA methylation. We attribute some of the methylation variation to factors which are known to alter methylation status (e.g., age [21] and sex [22]), and we regard the rest of the variability as individual-specific intrinsic variability, which we assume to come

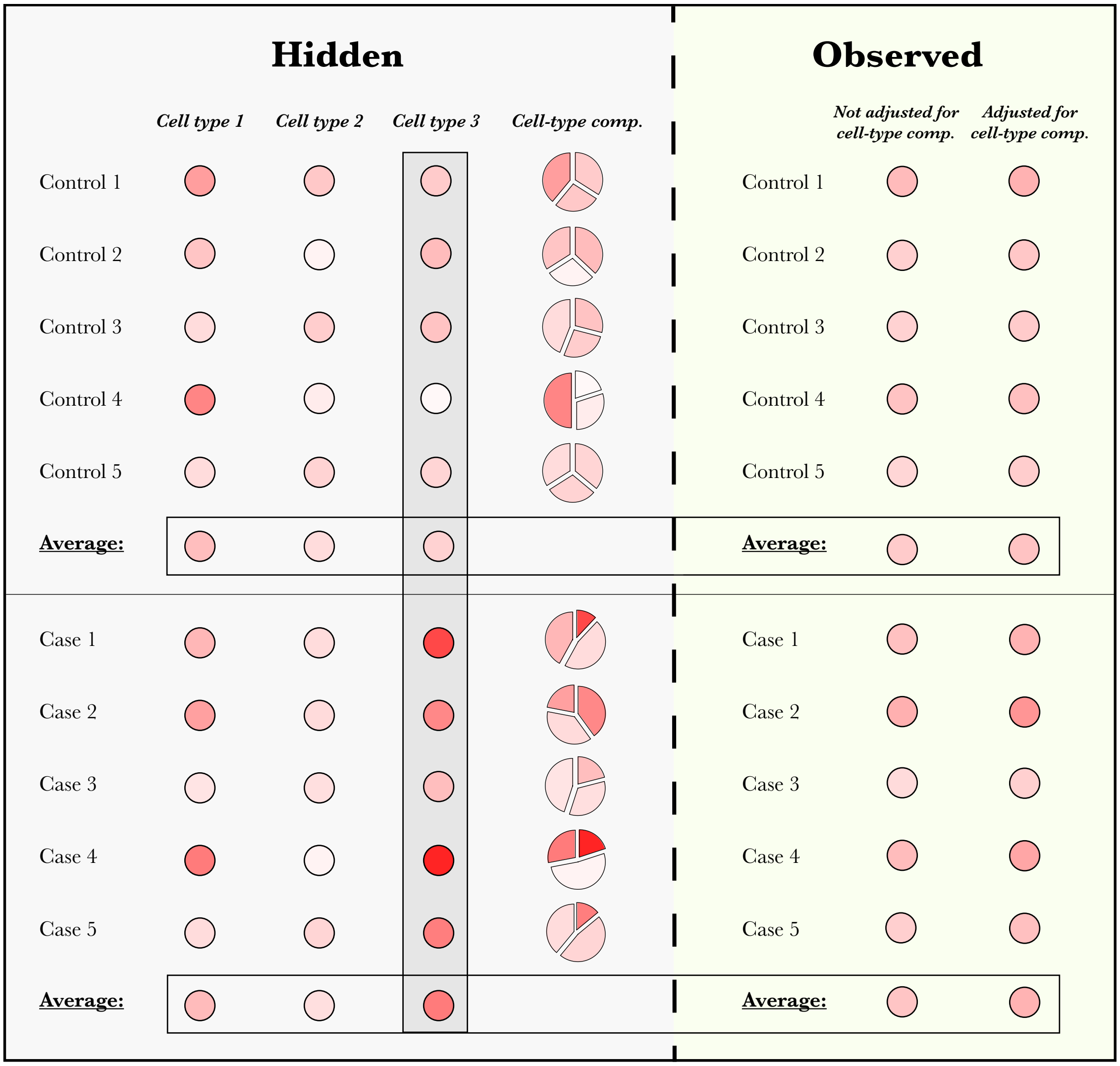

Figure 1: Observed bulk methylation levels may obscure cell-type-specific signals. Neither the observed methylation levels nor the observed levels after adjusting for the variability in cell-type composition can demonstrate a clear difference between cases and controls, in spite of a clear (hidden) difference in cell type 3. Methylation levels are represented by a gradient of red color, and adjusted observed levels were calculated for each sample by removing the cell-type-specific mean levels, weighted by its cell-type composition.

from a distribution. We summarize and illustrate the model in Figure 2. Based on this model, we developed Tensor Composition Analysis (TCA), a method for learning the unique cell-type-specific methylomes for each individual sample from its bulk data. TCA requires knowledge of the cell-type composition of the individuals in the data. In cases where the cell-type composition is unknown, it can be computationally estimated using standard methods [23–27]. As we later show, TCA performs well even in cases where only noisy estimates of the cell-type composition are available.

Applying TCA for detecting cell-type-specific associations in epigenetic studies In order to empirically verify that TCA can learn cell-type-specific methylation levels, we first leveraged whole-blood methylation data collected from sorted leukocytes [28] to simulate heterogeneous bulk methylation data. While the bulk data captured the cell-type-specific signals to some extent, as expected, TCA performed substantially better (Supplementary Figures S1 and S2). We further observed that TCA effectively captures effects of methylation altering covariates (Supplementary Figure S3 and **??**).

We next evaluated the performance of TCA in detecting cell-type-specific associations by simulating bulk methylation and corresponding phenotypes with cell-type-specific effects. Our experiments verify that TCA yields a substantial increase in power under different scenarios when compared to a standard regression analysis of the bulk levels. Particularly, in its worst performing scenario, TCA achieved a median of 2.4 fold increase in power (across all tested effect sizes) over the standard approach and a median of 12.1 fold increase in power in the best performing scenario

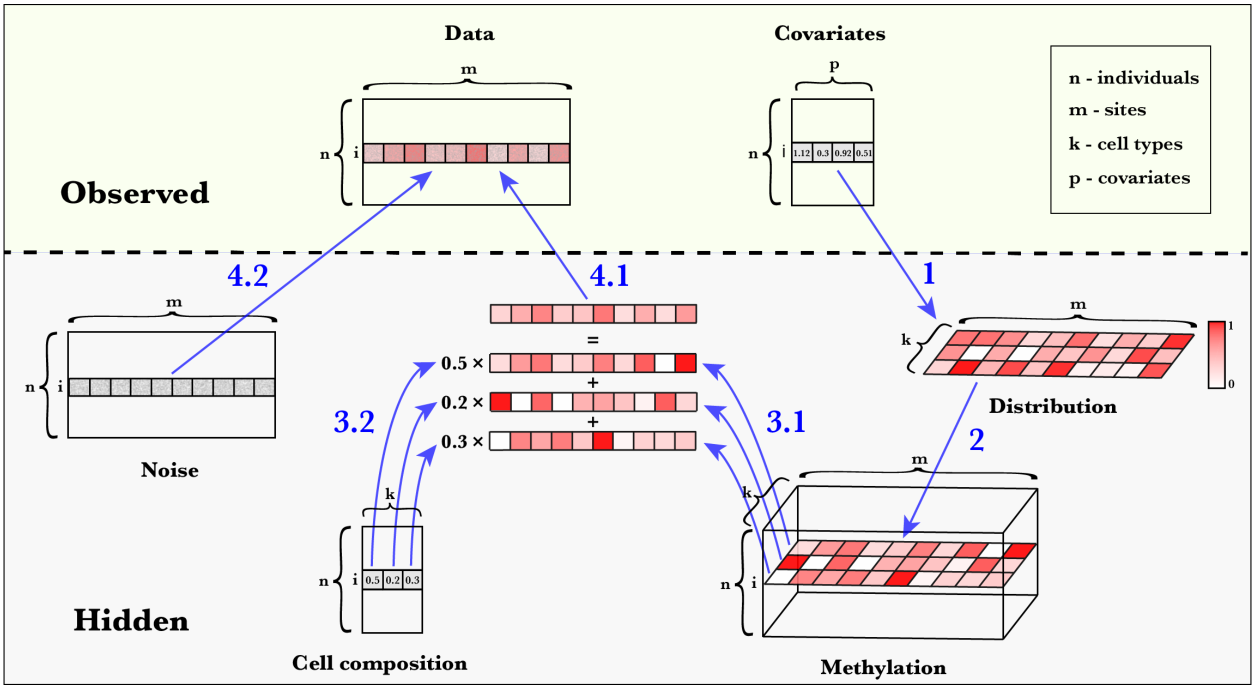

Figure 2: A summary of the TCA model for bulk DNA methylation data, presented as a four-steps generative model. Step 1: methylation altering covariates (e.g., age and sex) of a particular individual i can affect the methylation distribution of individual i. Step 2: the cell-type-specific methylomes of individual i are generated for each of the k cell types in the studied tissue. Step 3: the cell-type-specific methylomes of individual i (3.1) are combined according to the cell-type composition of the individual (3.2). Step 4: the true signal of the heterogeneous mixture (4.1) is distorted due to additional variation introduced by different sources of noise such as batch effects and other experiment-specific artificial variability (4.2); this results in the observed data. Methylation levels are represented by a gradient of red color

(Figure 3). Remarkably, TCA improved the most upon the power of the standard approach in a scenario where all cell types have the exact same effect size, although the standard analysis conceptually assumes all cell types to have the same effect size (Figure 3).

Surprisingly, in spite of the high power given by TCA, we found it to be conservative (i.e. less false positives than expected; Supplementary Figure S5). This results from the optimization of the model (Supplementary Note). Finally, we performed an additional power analysis stratified by cell types, which, once again, showed that TCA robustly outperforms the alternative standard regression approach (Supplementary Figures S6 and S7).

Cell-type-specific differential methylation in immune activity In general, the methylation levels in a particular cell type are not expected to be related to the tissue cell-type composition. Therefore, in the analysis of sorted-cell or single-cell methylation, there is no need to account for cell-type composition. In contrast, it is now widely acknowledged that in analysis of bulk methylation one has to account for cell-type composition [20]. Thus, for a phenotype that is highly correlated with the cell-type composition, the correction for cell-type composition on bulk methylation data will inevitably mask the signal, potentially resulting in no findings (i.e. false negatives). As opposed to bulk, cell-type specific analysis would not mask the signal in this case. To demonstrate this, we consider an extreme case where the phenotype is the cell-type composition. Specifically, we defined the level of immune activity of an individual as its total lymphocyte proportion in whole-blood, and aimed at finding methylation sites that are associated with regulation of immune activity.

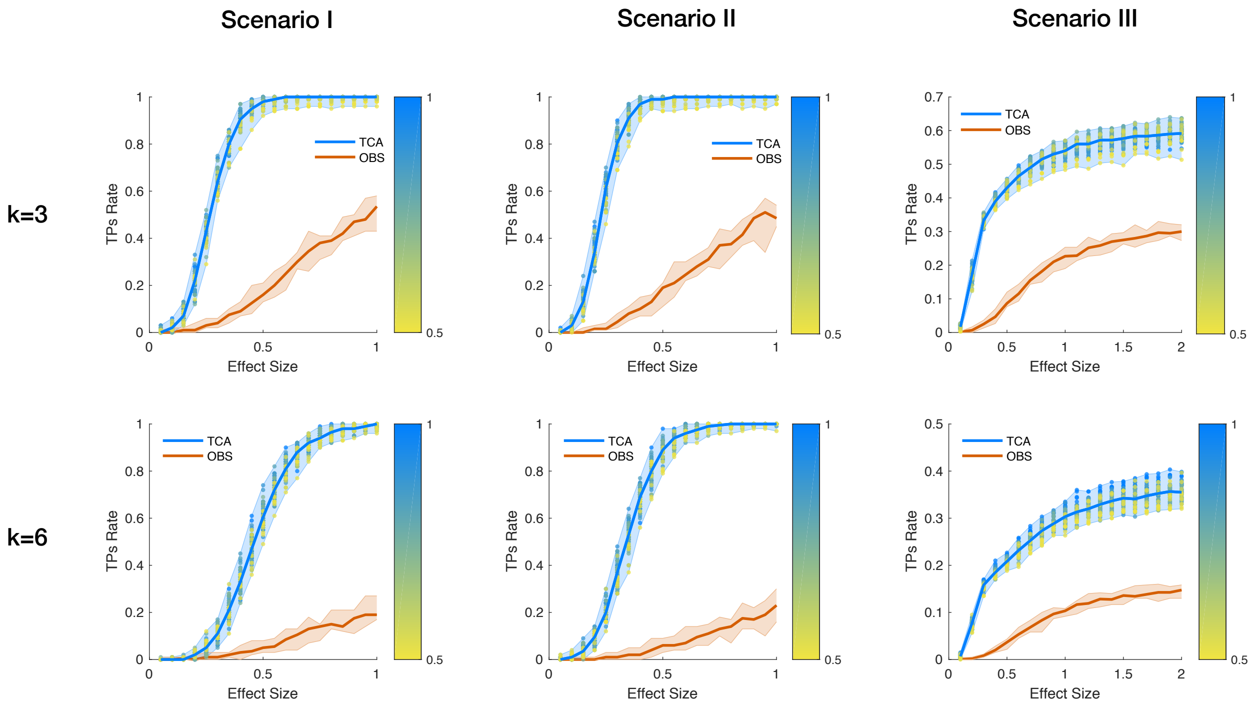

Figure 3: An evaluation of power for detecting cell-type-specific associations with DNA methylation. Performance was evaluated using two approaches: TCA and a standard linear regression with the observed bulk data (OBS). The numbers of true positives (TPs) were measured under three scenarios using a range of effect sizes: different effect sizes for different cell types (Scenario I), the same effect size for all cell types (Scenario II), and only a single effect size for a single cell type (Scenario III); each of the scenarios was evaluated under the assumption of three constituting cell types (k=3) and six constituting cell types (k=6). The colored areas reflect results across multiple simulations, and the colored dots reflect the results of TCA under different initializations of the cell-type composition estimates, where the color gradients represent the mean absolute correlation of the initial estimates with the true values (across all cell types).

Since bulk methylation data is a composition of signals that depend on to the cell-type proportions, a standard regression approach with whole-blood methylation is expected to fail to distinguish between false and true associations with immune activity. We verified this using whole-blood methylation data from a previous study by Liu et al. ($n = 658$) [19] (Figure 4a). Importantly, accounting for the cell-type composition in this case would eliminate any true signal in the data, as the immune response phenotype is perfectly defined by the cell-type composition.

We next performed cell-type-specific analysis using TCA, which resulted in 8 experiment-wide significant associations (p-value<9.87e-07; Figure 4b and Supplementary File 1). Importantly, 6 of the associated CpGs reside in 5 genes that were either linked in GWAS to leukocyte composition in blood or that are known to play a direct role in regulation of leukocytes: CD247, CLEC2D, PDCD1, PTPRCAP, and DOK2 (Supplementary File 1). The remaining associated CpGs reside in the genes SDF4 and SEMA6B, which were not previously reported as related to leukocyte composition. Using a second large whole-blood methylation data set (n=650) [29], we could replicate the associations with 4 out of the 7 genes (PTPRCAP, DOK2, SDF4 and SEMA6B; p-value<0.0063; Supplementary File 1). Our results are therefore consistent with the possibility that methylation modifications in these genes are involved in regulation of immune activity.

Cell-type-specific differential methylation in rheumatoid arthritis RA is an autoimmune chronic inflammatory disease which has been previously related to changes in DNA methylation [30,31]. In order to further demonstrate the utility of TCA, we revisited the largest previous whole-blood methylation study with RA by Liu et al. ($n = 650$) [19]. As a first attempt to detect associations

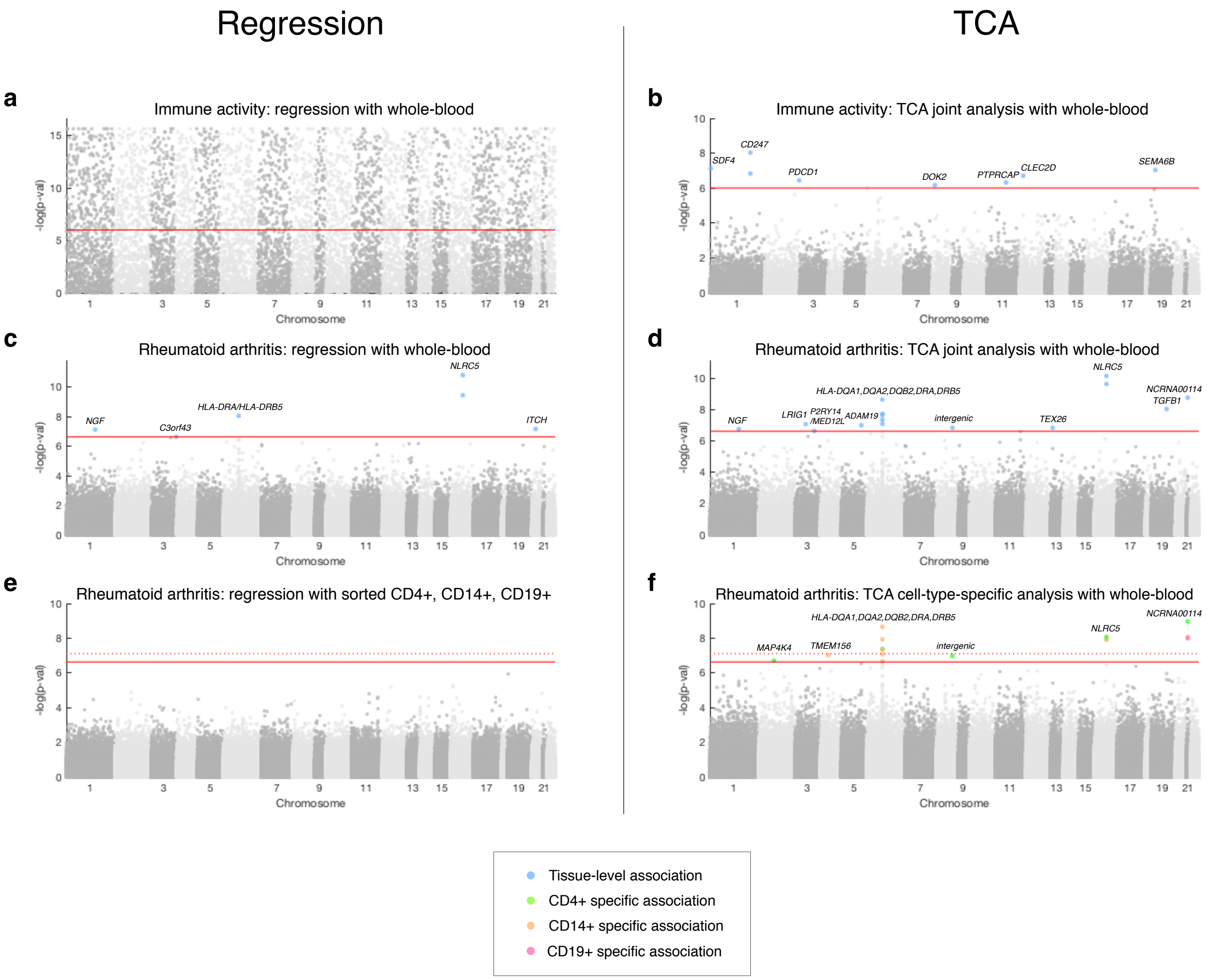

Figure 4: Results of the association analysis with level of immune activity and with rheumatoid arthritis, presented by Manhattan plots of the -log10 P-values for the association tests. Horizontal red lines represent the experiment-wide significance threshold. (a-b) Shown are results with immune activity using standard regression analysis and TCA analysis. (c-d) Shown are results with RA using regression analysis and TCA analysis under the assumption of a single effect size for all cell types. (e-f) Shown are results with RA using regression analysis with cell-sorted methylation and TCA cell-type-specific analysis with whole-blood methylation. Horizontal red dotted lines represent the significance threshold adjusted for three experiments corresponding to the three cell types.

between methylation and RA status, we applied a standard regression analysis, which yielded 6 experiment-wide significant associations (p-value<2.33e-7 ;Figure 4c and Supplementary File 2), overall in line with previous studies that analyzed this data set [24,32]. Since the standard analysis conceptually assumes a single effect size for all cell types, we next applied TCA under the assumption of a single effect size for all cell types. Remarkably, TCA found 15 experiment-wide significant CpGs, which altogether highlighted RA as an enriched pathway (p-value=1.45e-07; Figure 4d and Supplementary File 2).

The presumption that only some particular immune cell-types are related to the pathogenesis of RA, have led to studies with methylation collected from sorted populations of leukocytes (e.g., [33–35]). In a recent study by Rhead et al., some of us investigated differences in methylation patterns between RA cases and controls using data collected from sorted cells [35]. Particularly, methylation levels were collected from two sub-populations of CD4+ T cells (memory cells and naive cells; n=90, n=88), CD14+ monocytes (n=90), and CD19+ B cells (n=87). Although this study involved a considerable data collection effort in attempt to provide insights into the methylome of RA patients at a cell-type-specific resolution, it does not allow the detection of experiment-wide significant associations (Figure 4e), possibly owing to the limited sample size.

In order to overcome the sample size limitation, we applied TCA on the larger whole-blood data by Liu et al. Unlike the previous analysis, where we assumed that all cell types have the same effect size, in this analysis we tested for associations specifically with methylation levels in CD4+, CD14+, and CD19+ cells, without the restriction of a single effect size. Overall, this analysis reported 15 novel cell-type-specific associations with 11 CpGs: 4 associations in CD4+, 5 in CD14+, and one association in CD19+cells (p-value<2.33e-07; Figure 4f and Supplementary File 2). Considering a more stringent significance threshold in order to account for the three separate experiments we conducted for the three cell types resulted in 10 cell-type-specific associations with 7 CpGs (p-value<7.78e-08). Importantly, we found these CpGs to be enriched for involvement in the RA pathway (p-value=9.47e-07); particularly, 4 of these CpGs reside in HLA genes (or in an intergenic HLA region) that were previously reported in GWAS as RA genetic risk loci: HLA-DRA, DRB5, DQA1, and DQA2 (Supplementary File2).

Using the sorted-cell methylation data by Rhead et al. together with another data set with CD4+ methylation from an RA study by Guo et al. (n=24), we were able to validate two of the CD4+ associations and two of the CD14+ associations (Supplementary File 2). The lack of replication evidence for the rest of the associated CpGs may be explained in part by the small sample size available for replication (n≤90), as the p-values of many of them tended to be small (Supplementary File 2), or by the fact that each data set was collected from a different population; specifically, Liu et al. studied a Swedish population, Rhead et al. studied a heterogeneous European population, and Guo et al. studied a Han Chinese population.

In order to shed light on potential mechanisms related to these associations, we leveraged data from a previous study in a multi-ethnic cohort of unaffected individuals with both methylation and gene expression levels collected from sorted CD14+ (n=1,202) and sorted T cells (n=214) [36]. For each of the 5 CpGs reported by TCA as CD14+ specific associations with RA, we evaluated its correlation in CD14+ with CD14+ expression levels. Similarly, for each of the 4 CpGs reported by TCA as CD4+ specific associations with RA, we evaluated its correlation in T cells with T cell expression levels. In 5 of 9 of the cases, we found the methylation levels to be significantly correlated with the expression of groups of genes that are enriched for the RA pathway (p-value<2e-04 ; Supplementary File 3). Of particular interest is cg13081526, which was validated in the sorted data as a CD14+ specific association. We found this CpG to be highly correlated (or highly negatively correlated) with the CD14+ expression of 23 genes, 16 of which reside in the HLA region (Supplementary File 3).

Finally, we further investigated the potential relation of gene expression with the combined effect of cg13081526 and two additional CpGs (cg13778567 and cg18816397) that were reported by TCA as CD14+ specific associations and were found to be enriched for correlation with genes in the RA pathway. Interestingly, we found these 3 CpGs to be strongly associated with the CD14+ specific expression of 35 genes; particularly, these 3 CpGs could explain most of the variation in the CD14+ expression levels of three known RA risk genes: HLA-DRB1, DRB4, and DRB6 ($R^2 > 0.5$, p-value<1.64e-192 for all 3 genes; Supplementary File 3). Altogether, our evidence from multiple data sets is consistent with the possibility that cell-type-specific variation in the methylation of the associated CpGs play a role in cell-type-specific regulation of the expression of genes that are known to be related to RA pathogenesis.

3 Discussion

We proposed a methodology that can reveal novel cell-type-specific associations from bulk methylation data, i.e., without the need to collect cost prohibitive cell-type-specific data. This methodology is particularly useful in light of the large number of bulk samples that have been collected by now, and due to the fact that currently single-cell methylation technologies are not practically scalable to large population studies. Importantly, we found that TCA is substantially superior to a standard regression analysis of bulk data, even in the case where all cell types share the same effect size. We therefore suggest that TCA should always be preferred in analysis of bulk methylation data.

Notably, a recent attempt to provide cell-type-specific context in genetic studies aims at identifying trait-relevant tissues or cell types by leveraging genetic data and known tissue or cell-type-specific functional annotations [37,38]. This approach yielded some promising results in relating trait-associated genetic loci to relevant tissues and cell types. However, it is limited to only one particular task and it is bounded by design to consider only genetic signals, whereas non-genetic signals are often also of interest in genomic studies. Moreover, this approach can only suggest an implicit cell-type-specific context by binding known annotations with heritability. In contrast, the approach taken in TCA allows the extraction of explicit cell-type-specific signals, which can potentially allow many opportunities and applications in biological research.

A potential limitation of TCA is the need for rarely available cell-type proportions as an input. We alleviate this issue by allowing TCA to get estimates of the cell-type proportions using standard methods [23,27] and then re-estimating them following the TCA model. As we showed, this allows TCA to provide good results even when just moderately reasonable initial estimates of the cell-type proportions are available. In practice, obtaining such estimates can be done using either a reference-based approach [23] or a semi-supervised approach [27], in case a methylation reference is not available for the studied tissue.

Our experiments and mathematical results show that TCA can extract cell-type-specific signals from abundant cell types better compared with lowly abundant cell types. Another potential limitation is expected to be in the case where the proportion of one cell type strongly covary with the proportion of a second cell type. In case of a true association in just one of the two cell types, performing a marginal association test on each cell type separately might fail to effectively distinguish between the signals of the two cell types and report an association in both cell types. In light of these limitations, future studies are likely to benefit from including small replication data sets from sorted or single cells.

Finally, in this paper we focus on the application of TCA to epigenetic association studies. However, TCA can be formulated as a general statistical framework for obtaining underlying three-dimensional information from two-dimensional convolved signals, a capability which can benefit various domains in biology and beyond.

4 Methods

Here we summarize the model and mathematical methods. Further details are provided in the Supplementary Note. Since TCA can most naturally be described as a generalization of matrix factorization, we further provide a brief technical overview of matrix factorization (Supplementary Note).

The model Let Z^i_{hj} denote the value coming from cell type $h \in 1, ..., k$ at methylation site $j \in 1, ...m$ in sample $i \in 1, ...n$, we assume:

$$Z^i_{hj} | \mu_{hj}, \sigma_{hj} \sim N(\mu_{hj}, \sigma^2_{hj}) \tag{3}$$

In theory, the methylation status of a given site within a particular cell is a binary condition. However, unlike in the case of genotypes, methylation status may be different between different cells (even within the same individual, site and, cell type). We therefore consider a fraction of methylation rather than a fixed binary value. In array methylation data, possibly owing to the large number of cells used to construct each individual signal, we empirically observe that a normal assumption is reasonable. Admittedly, normality may not hold for values near the boundaries, however, in practice, we typically ignore sites with mean levels that are near the boundaries (i.e. sites whose values are consistently methylated or consistently unmethylated). This, in conjunction with the relatively low variability demonstrated by the vast majority of methylation sites, makes the normality assumption reasonable and therefore widely accepted in the context of statistical analysis of DNA metylation.

Let $W \in \mathbb{R}^{k \times n}$ be a non-negative constant weights matrix of k cell types for each of the n samples (i.e. cell-type proportions; each column sums up to 1), we assume the following model for site j of sample i in the observed heterogeneous methylation data matrix X:

$$X_{ij} = \sum_{h=1}^{k} w_{hi} Z_{hj}^{i} + \epsilon_{ij}, \; \epsilon_{ij} \sim N(0, \tau^2) \tag{4}$$

where w_{hi} is the proportion of the h-th cell type of sample i in W, and ϵ_{ij} represents an additional component of measurement noise which is independent across all samples. We therefore get that X_{ij} follows a normal distribution with parameters that are unique for each individual i and site j. Put differently, we assume that the entries of X are independent but also different in their means and variances.

Tensor Composition Analysis (TCA) Following the assumptions in (3) and in (4), the conditional probability of $Z_j^i = \left(Z_{1j}^i, ..., Z_{kj}^i\right)^T$ given X_{ij} can be shown (Supplementary Note) to satisfy

$$Pr(Z_j^i = z_j^i | X_{ij} = x_{ij}, w_i, \mu_j, \sigma_j, \tau) \propto exp\left(-\frac{1}{2}(a_{ij} - z_j^i)^T S_{ij}^{-1}(a_{ij} - z_j^i)\right) \tag{5}$$

where

$$\Sigma_j = diag(\sigma_{1j}^2, ..., \sigma_{kj}^2) \tag{6}$$

$$S_{ij} = \left(\frac{w_i w_i^T}{\tau^2} + \Sigma_j^{-1}\right)^{-1} \tag{7}$$

$$a_{ij} = S_{ij}\left(\frac{x_{ij}}{\tau^2} w_i + \Sigma_j^{-1}\mu_j\right) \tag{8}$$

Essentially, our suggested method, TCA, leverages the information given by the observed values $\{x_{ij}\}$ for learning a three-dimensional tensor consisted of estimates of the underlying values

$\{z^i_{hj}\}$. This is done by setting the estimator $\hat{z}^i_j$ to be the mode of the conditional distribution in (5):

$$\hat{z}^i_j = a_{ij} = \left(\frac{w_i w_i^T}{\tau^2} + \Sigma_j^{-1}\right)^{-1}\left(\frac{x_{ij}}{\tau^2} w_i + \Sigma_j^{-1}\mu_j\right) \tag{9}$$

TCA requires the cell-type proportions W as an input. Given W, the parameters τ, $\{\mu_j\}$, $\{\sigma_j\}$ can be estimated from the observed data under the assumption in (4). In practice, the cell-type proportions are typically unknown. In such cases, W can be estimated computationally using standard methods (e.g., [23,27]) and then re-estimated under the TCA model in an alternating optimization procedure with the rest of the parameters in the model. The TCA model can further account for covariates, which may either directly affect Z^i_j (e.g., age and sex) or affect the mixture X_{ij} (e.g., batch effects). For more details and a full derivation of the conditional distribution of Z^i_j, while accounting for covariates, and for information about parameters inference see Supplementary Note.

In order to see why TCA can learn non-trivial information about the $\{z^i_{hj}\}$ values, consider a simplified case where $\tau = 0, \mu_{hj} = 0, \sigma_{hj} = 1$ for each h and a specific given j. In this case, it can be shown (Supplementary Note) that

$$Z^i_{hj} | X_{ij} = x_{ij} \sim N\left(\frac{w_{hi} x_{ij}}{\sum_{l=1}^k w_{li}^2}, 1 - \frac{w_{hi}^2}{\sum_{l=1}^k w_{li}^2}\right) \tag{10}$$

That is, given the observed value x_{ij}, the conditional distribution of Z^i_{hj} has a lower variance compared with that of the marginal distribution of Z^i_{hj} ($\sigma^2_{hj} = 1$), thus reducing the uncertainty and allowing us to provide non-trivial estimates of the $\{z^i_{hj}\}$ values. This result further implies that in the context of DNA methylation, where the weights matrix W corresponds to a matrix of cell-type proportions, we should expect to gain better estimates for the $\{z^i_{hj}\}$ levels in more abundant cell types compared with cell types with typically lower abundance. For more details see Supplementary Note.

Applying TCA to epigenetic association studies We next consider the problem of detecting statistical associations between DNA methylation levels and biological phenotypes. Let $X \in \mathbb{R}^{n \times m}$ be an individuals by sites matrix of methylation levels, and let Y denote an n-length vector of phenotypic levels measured from the same n individuals, typical association studies usually consider the following model for testing a particular site j for association with Y:

$$Y_i \quad = \quad X_{ij}\beta_j + e_i, \; e_i \sim N(0, \sigma^2) \tag{11}$$

where Y_i is the phenotypic level of individual i, β_j is the effect size of the j-th site, and e_i is a component of i.i.d. noise. For convenience of presentation, we omit potential covariates which can be incorporated into the model. In a typical EWAS, we fit the above model for each feature, and we look for all features j for which we have a sufficient statistical evidence of non-zero effect size (i.e. $\beta_j \neq 0$).

In principle, one can use TCA for estimating cell-type-specific levels, and then look for cell-type-specific associations by fitting the model in (11) with the estimated cell-type-specific levels (instead of directly using X). However, an alternative one-step approach can be also used. This approach leverages the information we gain about z_{hj}^i given that $X_{ij} = x_{ij}$ for directly modeling the phenotype as having cell-type-specific effects. Specifically, consider the following model:

$$Y_i = Z_{lj}^i \beta_{lj} + e_i, e_i \sim N(0, \phi^2) \tag{12}$$

where β_{lj} denotes the cell-type-specific effect size of some cell type of interest l. Provided with the observed information x_{ij}, while keeping the assumptions in (3) and in (4), it can be shown (Supplementary Note) that:

$$Y_i|X_{ij} = x_{ij} \sim N\left(\beta_{lj}\left(\mu_{lj} + \frac{w_{li}\sigma_{lj}^2\tilde{x}_{ij}}{\tau^2 + \sum_{h=1}^k w_{hi}^2\sigma_{hj}^2}\right), \phi^2 + \beta_{lj}^2\left(\sigma_{lj}^2 - \frac{w_{li}^2\sigma_{lj}^4}{\tau^2 + \sum_{h=1}^k w_{hi}^2\sigma_{hj}^2}\right)\right) \tag{13}$$

$$\tilde{x}_{ij} = x_{ij} - \sum_{h=1}^{k} w_{hi}\mu_{hj} \tag{14}$$

This shows that directly modeling $Y_i|X_{ij}$ effectively integrates the information over all possible values of Z_{lj}^i. Given W, μ_j, σ_j, τ (typically estimated from X; Supplementary Note), we can estimate ϕ and the effect size β_{lj} using maximum likelihood. The estimate $\hat{\beta}_{lj}$ can be then tested for significance using a generalized likelihood ratio test. Similarly, we can consider a joint test for the combined effects of more than one cell type. A full derivation of the statistical test is described in the Supplementary Note. In this paper, whenever association testing was conducted, we used this direct modeling of the phenotype given the observed methylation levels.

Finally, we note that in principle one can also use the model in equation (4) for testing for cell-type-specific associations by treating the phenotype of interest as a covariate and estimating its effect size. However, TCA provides a way to deconvolve the data into cell-type-specific levels, which is of independent interest beyond the specific application for association studies. Moreover, model directionality often matters, and the TCA framework allows us to directly model the phenotype rather than merely treat it as another covariate. Particularly, in the context of this paper, it is known that methylation levels are actively involved in many cellular processes such as regulation of gene expression [39], thus, making DNA methylation a potential contributing determinant in disease (which further justifies the modeling of the phenotype as an outcome).

Implementation of TCA TCA was implemented in Matlab and is available from github at http://github.com/cozygene/TCA. TCA requires for its execution a heterogeneous DNA methylation data matrix and corresponding cell-type proportions for the samples in the data. In case where cell counts are not available, TCA can take estimates of the cell-type proportions, which are then optimized with the rest of the parameters in the model.

For the real data experiments, we used GLINT [40] for generating initial estimates of the cell-type proportions for the whole-blood data sets. GLINT provides estimates according to the Houseman et al. model [23], using a panel of 300 highly informative methylation sites in blood [41] and a reference data collected from sorted blood cells [28]. Given these estimates, we used the TCA model to re-estimate the cell-type proportions using the top 500 sites selected by the feature selection procedure of ReFACTor [24].

Data simulation We simulated data following our model and similarly to an approach that we previously described in details elsewhere [27]. Briefly, we estimated cell-type-specific means and standard deviations in each site using reference data of methylation levels collected from sorted blood cells [28]. Since we expected cell-type-specific associations to be mostly present in CpG sites that are highly differentially methylated across different cell types, we considered cell-type-specific means and standard deviations from sites which demonstrated the highest variability in cell-type-specific mean levels across the different cell types.

Using the estimated parameters of a given site, we generated cell-type-specific DNA methylation levels using normal distributions, conditional on the range $[0, 1]$. In cases where covariates were simulated to have an effect on the cell-type-specific methylation levels, the means of the normal distributions were tuned for each sample to account for its covariates and the corresponding effect sizes (shared across samples; Supplementary Note). We generated cell-type proportions for each sample using a Dirichlet distribution with parameters that were estimated from blood cell counts elsewhere [27]. Specifically, the Dirichlet distribution modeled the distribution of 6 cell types: granulocytes, monocytes and 4 sub-types of lymphocytes (CD4+, CD8+, NK and B cells). In the case of three constituting cell types (granulocytes, monocytes, and lymphocytes), we set the Dirichlet parameter of lymphocytes to be the sum of the parameters of all the lymphocyte subtypes. Eventually, for each sample, we composed its methylation level at each site by taking a linear combination of the simulated cell-type-specific levels of that site, weighted by the cell composition of that sample, and added an additional i.i.d normal noise conditional on the range $[0, 1]$ to simulate technical noise ($\tau = 0.01$). In cases where covariates were simulated to have a global effect on the methylation levels (i.e. non-cell-type-specific effect, such as batch effects), we further added an additional component of variation for each sample according to its global covariates and their corresponding effect sizes.

Data sets We used 3 methylation data sets that were previously collected in RA studies with the Illumina 450K human DNA methylation array: a whole-blood data set by Liu et al. of 354 RA cases and 332 controls (GEO accession GSE42861) [19], a CD4+ methylation data set of 12 RA cases and 12 controls with matching age and sex (for each RA case a control sample with matching age and sex was collected) by Guo et al. (GEO accession GSE71841) [34], and cell-sorted methylation data collected from 63 female RA patients and 31 female control subjects in CD4+ memory cells, CD4+ naive cells, CD14+ monocytes, and CD19+ B cells; these sorted-cell data were originally described by Rhead et al. [35].

We further used data from a previous study by Reynolds et al. with both 450K methylation array data (GEO accessions GSE56581 and GSE56046) and Illumina HumanHT-12 expression array data (GEO accessions GSE56580 and GSE56045) collected from CD14+ monocytes (n=1,202) and from T cells (n=214) [36]. In addition, for replicating the association results with immune activity, we used another 450K methylation array data set that was previously studied by Hannum et al. in the context of aging rates (n=656; GEO accession GSE40279) [29]. Finally, for the simulation experiments we used methylation reference of sorted leukocyte cell types collected in 6 individuals from the Gene Omnibus Database (GEO accession GSE35069) [28].

We preprocessed the Liu et al. data and the Hannum et al. data according to a recently suggested normalization pipeline [42]. The full preprocessing details for these two data sets were previously described by us elsewhere [27]. Since IDAT files were not available for the Guo et al. data set, we used the methylation intensity levels published by the authors. Following recommendations by Lenhe et al., we performed a quantile normalisation of the methylation intensity values, subdivided by probe type, probe sub-type and color channel. The normalized levels were then used to calculate beta normalized methylation levels (according to the recommendation by Illumina). The full preprocessing details for the the Rhead et al. data are described elsewhere [35]; here, we further excluded a small batch consisted of only 4 individuals. Finally, for the association experiments with methylation, we further discarded consistently methylated probes and consistently unmethylated probes from the data (mean value higher than 0.9 or lower than 0.1, respectively).

Power simulations We simulated data and sampled for each site under test a normally distributed phenotype with additional effects of the cell-type-specific methylation levels of the site. We set the variance of each phenotype to the variance of the site under test, in order to eliminate the dependency of the power in the variance of the tested site (and therefore allow a clear quantification of the true positives rate under a given effect size). Particularly, when simulating an effect coming from a single cell type, we randomly generated a phenotype from a normal distribution with the variance set to the variance of the site under test in the specific cell type under test. Similarly, when simulating effects coming from all cell types, we randomly generated a phenotype from a normal distribution with the variance set to the total variance of the site under test (i.e. across all cell types).

We performed the power evaluation using simulated data with 3 constituting cell types (k=3) and using simulated data with 6 constituting cell types (k=6). We considered three scenarios across a range of effect sizes as follows: different effect sizes for different cell types (using s joint test), the same effect size for all cell types (using a joint test, under the assumption of the same effect for all cell types), and a scenario with only a single associated cell type (a marginal test). In the first scenario, effect sizes for the different cell types were drawn from a normal distribution with the particular effect size under test set to be the mean (with standard deviation $\sigma = 0.05$), and in the third scenario we evaluated the aggregated performance of all the marginal tests across all constituting cell types in the simulation. We further repeated the marginal test while stratifying the evaluation by cell type (i.e. the marginal test was performed under the third scenario for each cell type separately). In each of these experiment, we calculated the true positives rate of the associations that were reported as significant while adjusting for the number of sites in the simulated data.

For each scenario and for each number of constituting cell types, we simulated 10 data sets, each included 500 samples and 100 sites. Importantly, throughout the simulation study, we considered for each simulated data set the case where only noisy estimates of the cell-type proportions are available (and therefore need to be re-estimated together with the rest of the parameters in the TCA model). Specifically, for each sample in the data we replaced its cell-type proportions with randomly sampled proportions coming from a Dirichlet distribution with the original cell-type proportions of the individuals as the parameters. For each level of noise, these parameters were multiplied by a factor that controlled the level of similarity of the sampled proportions to the original proportions. Finally, for evaluating false positives rates, we followed the above procedure, however, without adding additional effects coming from methylation levels. We evaluated the false positives rate by considering the fraction of sites with p-value$<$0.05.

Analysis of immune activity We used the Liu et al. data [19] as the discovery data (n=658) and the Hannum et al. data [29] as the replication data (n=650). Since we expected to observe associations with regulation of cell-type composition in CpGs that demonstrate differential methylation between different cell types, we considered for this analysis only CpGs that were reported as differentially methylated between different whole-blood cell types [20]. Specifically, we considered the sites in the intersection between the set of Bonferroni-significant CpGs that were reported as differentially methylated in whole-blood and the available CpGs in both the discovery and replication data sets; this resulted in a set of 50,123 CpGs that were available for this analysis.

We performed a standard linear regression analysis using GLINT [40] and a TCA analysis under the assumption of the same effect size in all cell types. In the analysis of the Liu et al. data we controlled for RA status, gender, age, smoking status, and known batch information, and in the analysis of the Hannum et al. data we controlled for gender, age, ethnicity and the first two EPISTRUCTURE principal components [43] in order to account for the population structure in this data set. In both data sets, in order to take into account potentially unknown technical confounding effects, we further included the first ten principal components calculated from the intensity levels of a set of 220 control probes in the Illumina methylation array, as suggested by Lenhe et al. [42] in an approach similar to the remove unwanted variation method (RUV) [44]. These probes are expected to demonstrate no true biological signal and therefore allow to capture global technical variation in the data.

In the replication analysis, we applied a Bonferroni threshold in reporting significance, controlling for the number of genome-wide significant associations that were reported in the discovery data. The results are summarized in Supplementary File 1, where additional description for the associated genes is provided from GeneCards [45], the GWAS catalog [46], and GeneHancer [47].

Analysis of rheumatoid arthritis We used the Liu et al. data [19] as the discovery data (n=658, 214,096 Cpgs). We applied a standard logistic regression analysis with the RA status as an outcome using GLINT [40] and TCA analysis: under the assumption of a single effect for all cell types (joint test), and for each of CD4+, CD14+, and CD19+, under the assumption of a single associated cell type (marginal test). In every analysis, we accounted for the same variables described in the immune activity analysis with this data set. In order to test the associations reported by TCA for enrichment for the RA pathway, we used missMethyl [48], an R package that allows to run enrichment analysis for disease directly on CpGs (while accounting for gene length bias).

In the replication analysis with the Rhead et al. data, we applied a standard logistic regression analysis using GLINT [40] on each of the CD14+ (n=90) and CD19+ (n=87) data sets, while accounting for age, smoking status, and batch information. Since the Rhead et al. data included sorted-cell methylation from two sub-types of CD4+, for the replication analysis of CD4+ (n=81) we performed for each site a logistic regression analysis using both its CD4+ naive cells methylation levels and CD4+ memory cells methylation.

Taking a standard approach in the analysis of the Guo et al. CD4+ sorted methylation data resulted in a severe inflation in test statistic. Since the cases and controls in the sample were matched for age and sex, we suspected that technical variation might have led to this inflation. In order to test that, we calculated the first principal component of control probes, similarly to the approach taken in the analysis of the Liu et al. data. However, since IDAT files were not available for the Guo et al .data, and therefore the same set of 220 control probes that were used in the Liu et al. data were not available, we used the methylation intensity levels of the 220 sites with the least variation in the data as control probes. Indeed, we found that the first PC of the control probes corresponds to the case/control status in the data almost perfectly (r=0.91, p-value=6.29e-10). As a result, p-values obtained using a standard analysis of the Guo et al. data set are not reliable. We therefore considered the following non-parametric procedure. We ranked the sites according to their absolute difference in mean methylation levels between cases and controls, and considered a simple enrichment test, wherein the p-value of a site was determined as its rank divided by the total number of sites in the ranking.

We considered a Bonferroni correction for reporting significance in the replication analysis, controlling for the number of genome-wide significant associations that were reported by the cell-type-specific analysis of TCA in the discovery data. Since two independent data sets were available for testing the replicability of the CD4+ specific associations (Rhead et al. and Guo et al.), we considered sites with replication p-value$<$0.05 in both data sets as successfully replicated. The results are summarized in Supplementary File 2, where additional description for the associated genes is provided from GeneCards [45], the GWAS catalog [46], and GeneHancer [47].

Finally, in the analysis of the Reynolds et al. data with both methylation and expression levels, we first looked for significant correlations between methylation and the log-transformed expression levels, while accounting for the total number of hypotheses (the number of genes times the number of CpGs that were reported by TCA for CD+4 and CD14+). Enrichment test for the RA pathway was performned for the set of significantly correlated genes (for each of the tested CpGs separately) using clusterProfiler [49]. In order to find the genes whose expression can be well explained by the 3 CD14+ specific associations that were reported by TCA and were found to be enriched for correlation with RA pathway genes (cg13081526, cg13778567 and cg18816397), we fitted a linear model for the log-transformed expression levels of each gene in the CD14+ expression data using the 3 CpGs and the pairwise interactions between these 3 CpGs. The results with the Reynolds are summarized in Supplementary File 3.

Acknowledgments. EH and ER were partially supported by NSF grant 1705197. ER and RS were supported in part by the Israel Science Foundation (Grant 1425/13) and by the Edmond J. Safra Center for Bioinformatics at Tel-Aviv University. SS was supported in part by is supported in part by NIH grants R00GM111744, R35GM125055), NSF Grant III-1705121), an Alfred P. Sloan Research Fellowship, and a gift from the Okawa Foundation.

1. Fukazawa, Y. *et al.* Lymph node t cell responses predict the efficacy of live attenuated siv vaccines. *Nature medicine* **18**, 1673 (2012).

2. Becker, A. M. *et al.* Sle peripheral blood b cell, t cell and myeloid cell transcriptomes display unique profiles and each subset contributes to the interferon signature. *PloS one* **8**, e67003

(2013).

3. Plitas, G. *et al.* Regulatory t cells exhibit distinct features in human breast cancer. *Immunity* **45**, 1122–1134 (2016).

4. Schwarzer, A. *et al.* The non-coding rna landscape of human hematopoiesis and leukemia. *Nature Communications* **8** (2017).

5. Buenrostro, J. D. *et al.* Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* **523**, 486 (2015).

6. Lake, B. B. *et al.* Neuronal subtypes and diversity revealed by single-nucleus rna sequencing of the human brain. *Science* **352**, 1586–1590 (2016).

7. Tirosh, I. *et al.* Single-cell rna-seq supports a developmental hierarchy in human oligodendroglioma. *Nature* **539**, 309 (2016).

8. Tirosh, I. *et al.* Dissecting the multicellular ecosystem of metastatic melanoma by single-cell rna-seq. *Science* **352**, 189–196 (2016).

9. Claussnitzer, M. *et al.* Fto obesity variant circuitry and adipocyte browning in humans. *New England Journal of Medicine* **373**, 895–907 (2015).

10. Mostafavi, S. *et al.* Parsing the interferon transcriptional network and its disease associations. *Cell* **164**, 564–578 (2016).

11. Battle, A. *et al.* Characterizing the genetic basis of transcriptome diversity through rna-sequencing of 922 individuals. *Genome research* **24**, 14–24 (2014).

12. Wright, F. A. *et al.* Heritability and genomics of gene expression in peripheral blood. *Nature genetics* **46**, 430–437 (2014).

13. Pfeifferm, L. *et al.* Dna methylation of lipid-related genes affects blood lipid levels. *Circulation: Genomic and Precision Medicine* CIRCGENETICS–114 (2015).

14. Smallwood, S. A. *et al.* Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. *Nature methods* **11**, 817 (2014).

15. Schwartzman, O. & Tanay, A. Single-cell epigenomics: techniques and emerging applications. *Nature Reviews Genetics* **16**, 716 (2015).

16. Clark, S. J., Lee, H. J., Smallwood, S. A., Kelsey, G. & Reik, W. Single-cell epigenomics: powerful new methods for understanding gene regulation and cell identity. *Genome biology* **17**, 72 (2016).

17. Angermueller, C. *et al.* Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity. *Nature methods* **13**, 229 (2016).

18. Edgar, R., Domrachev, M. & Lash, A. E. Gene expression omnibus: Ncbi gene expression and hybridization array data repository. *Nucleic acids research* **30**, 207–210 (2002).

19. Liu, Y. *et al.* Epigenome-wide association data implicate dna methylation as an intermediary of genetic risk in rheumatoid arthritis. *Nature biotechnology* **31**, 142–147 (2013).

20. Jaffe, A. E. & Irizarry, R. A. Accounting for cellular heterogeneity is critical in epigenome-wide association studies. *Genome Biol* **15**, R31 (2014).

21. Horvath, S. Dna methylation age of human tissues and cell types. *Genome biology* **14**, R115 (2013).

22. Singmann, P. *et al.* Characterization of whole-genome autosomal differences of dna methylation between men and women. *Epigenetics & chromatin* **8**, 1–13 (2015).

23. Houseman, E. A. *et al.* DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC bioinformatics* (2012).

24. Rahmani, E. *et al.* Sparse pca corrects for cell type heterogeneity in epigenome-wide association studies. *Nature methods* **13**, 443–445 (2016).

25. Houseman, E. A. *et al.* Reference-free deconvolution of dna methylation data and mediation by cell composition effects. *BMC bioinformatics* **17**, 259 (2016).

26. Lutsik, P. *et al.* Medecom: discovery and quantification of latent components of heterogeneous methylomes. *Genome biology* **18**, 55 (2017).

27. Rahmani, E., Schweiger, R., Shenhav, L., Eskin, E. & Halperin, E. A bayesian framework for estimating cell type composition from dna methylation without the need for methylation reference. *bioRxiv* 112417 (2017).

28. 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).

29. Hannum, G. *et al.* Genome-wide methylation profiles reveal quantitative views of human aging rates. *Molecular cell* **49**, 359–367 (2013).

30. Glant, T. T., Mikecz, K. & Rauch, T. A. Epigenetics in the pathogenesis of rheumatoid arthritis. *BMC medicine* **12**, 35 (2014).

31. Cribbs, A., Feldmann, M. & Oppermann, U. Towards an understanding of the role of dna methylation in rheumatoid arthritis: therapeutic and diagnostic implications. *Therapeutic advances in musculoskeletal disease* **7**, 206–219 (2015).

32. Zou, J., Lippert, C., Heckerman, D., Aryee, M. & Listgarten, J. Epigenome-wide association studies without the need for cell-type composition. *Nature methods* **11**, 309–11 (2014).

33. de Andres, M. C. *et al.* Assessment of global dna methylation in peripheral blood cell subpopulations of early rheumatoid arthritis before and after methotrexate. *Arthritis research & therapy* **17**, 233 (2015).

34. Guo, S. *et al.* Genome-wide dna methylation patterns in cd4+ t cells from chinese han patients with rheumatoid arthritis. *Modern rheumatology* **27**, 441–447 (2017).

35. Rhead, B. *et al.* Rheumatoid arthritis naive t cells share hypermethylation sites with synoviocytes. *Arthritis & Rheumatology* **69**, 550–559 (2017).

36. Reynolds, L. M. *et al.* Age-related variations in the methylome associated with gene expression in human monocytes and t cells. *Nature communications* **5**, 5366 (2014).

37. Finucane, H. K. *et al.* Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nature genetics* **47**, 1228 (2015).

38. Hao, X., Zeng, P., Zhang, S. & Zhou, X. Identifying and exploiting trait-relevant tissues with multiple functional annotations in genome-wide association studies. *PLoS genetics* **14**, e1007186 (2018).

39. Jaenisch, R. & Bird, A. Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nature genetics* **33**, 245 (2003).

40. Rahmani, E. *et al.* Glint: a user-friendly toolset for the analysis of high-throughput dna-methylation array data. *Bioinformatics* btx059 (2017).

41. Koestler, D. C. *et al.* Improving cell mixture deconvolution by id entifying optimal dna methylation libraries (idol). *BMC bioinformatics* **17**, 1 (2016).

42. Lehne, B. *et al.* A coherent approach for analysis of the illumina humanmethylation450 beadchip improves data quality and performance in epigenome-wide association studies. *Genome biology* **16**, 37 (2015).

43. Rahmani, E. *et al.* Genome-wide methylation data mirror ancestry information. *Epigenetics & chromatin* **10**, 1 (2017).

44. Gagnon-Bartsch, J. A. & Speed, T. P. Using control genes to correct for unwanted variation in microarray data. *Biostatistics* **13**, 539–552 (2012).

45. Stelzer, G. *et al.* The genecards suite: from gene data mining to disease genome sequence analyses. *Current protocols in bioinformatics* **54**, 1–30 (2016).

46. MacArthur, J. *et al.* The new nhgri-ebi catalog of published genome-wide association studies (gwas catalog). *Nucleic acids research* **45**, D896–D901 (2016).

47. Fishilevich, S. *et al.* Genehancer: genome-wide integration of enhancers and target genes in genecards. *Database* **2017** (2017).

48. Phipson, B., Maksimovic, J. & Oshlack, A. missmethyl: an r package for analyzing data from illuminas humanmethylation450 platform. *Bioinformatics* **32**, 286–288 (2015).

49. Yu, G., Wang, L.-G., Han, Y. & He, Q.-Y. clusterprofiler: an r package for comparing biological themes among gene clusters. *Omics: a journal of integrative biology* **16**, 284–287 (2012).