bioRxiv preprint doi: https://doi.org/10.1101/2020.06.18.158758; this version posted June 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Nonlinear ridge regression improves robustness of cell-type-specific differential expression studies

Fumihiko Takeuchi and Norihiro Kato

Department of Gene Diagnostics and Therapeutics, Research Institute, National Center for Global Health and Medicine (NCGM), Tokyo, Japan

Correspondence

Fumihiko Takeuchi

Department of Gene Diagnostics and Therapeutics, Research Institute, National Center for Global Health and Medicine (NCGM)

1-21-1 Toyama, Shinjuku-ku, Tokyo, 162-8655, Japan

Email: fumihiko@takeuchi.name

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.18.158758; this version posted June 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Abstract

Background: Epigenome-wide association studies (EWAS) and differential gene expression analyses are generally performed on tissue samples, which consist of multiple cell types. Cell-type-specific effects of a trait, such as disease, on the omics expression are of interest but difficult or costly to measure experimentally. By measuring omics data for the bulk tissue, cell type composition of a sample can be inferred statistically. Subsequently, cell-type-specific effects are estimated by linear regression that includes terms representing the interaction between the cell type proportions and the trait. This approach involves two issues, scaling and multicollinearity.

Results: First, although cell composition is analyzed in linear scale, differential methylation/expression is analyzed suitably in the logit/log scale. To simultaneously analyze two scales, we developed nonlinear regression. Second, we show that the interaction terms are highly collinear, which is obstructive to ordinary regression. To cope with the multicollinearity, we applied ridge regularization. In simulated and real data, the improvement was modest by nonlinear regression and substantial by ridge regularization.

Conclusion: Nonlinear ridge regression performed cell-type-specific association test on bulk omics data more robustly than previous methods. The omicwas package for R implements nonlinear ridge regression for cell-type-specific EWAS, differential gene expression and QTL analyses. The software is freely available from https://github.com/fumi-github/omicwas

Keywords

Epigenome-wide association study, Differential gene expression analysis, Cell type, Ridge regression, Nonlinear regression, mQTL, eQTL

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.18.158758; this version posted June 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Background

Epigenome-wide association studies (EWAS) and differential gene expression analyses elucidate the association of disease traits (or conditions) with the level of omics expression, namely DNA methylation and gene expression. Thus far, tissue samples, which consist of heterogeneous cell types, have mainly been examined, because cell sorting is not feasible in most tissues and single-cell assay is still expensive. Nevertheless, the cell type composition of a sample can be quantified statistically by comparing omics measurement of the target sample with reference data obtained from sorted or single cells [1,2]. By utilizing the composition, the disease association specific to a cell type was statistically inferred for gene expression [3-10] and DNA methylation [11-14].

For the imputation of cell type composition, omics markers are usually analyzed in the original linear scale, which measures the proportion of mRNA molecules from a specific gene or the proportion of methylated cytosine molecules among all cytosines at a specific CpG site [15]. The proportion can differ between cell types, and the weighted average of cell-type-specific proportions becomes the proportion in a bulk tissue sample. Using the fact that the weight equals the cell type composition, the cell type composition of a sample is imputed. In contrast, gene expression analyses are performed in the log-transformed scale because the signal and noise are normally distributed after log-transformation [16]. In DNA methylation analysis, the logit-transformed scale, which is called the M-value, is statistically valid [17]. Consequently, the optimal scales for analyzing differential gene expression or methylation can differ from the optimal scale for analyzing cell type composition.

Aiming to perform cell-type-specific EWAS or differential gene expression analyses by using unsorted tissue samples, we study two issues that have been overlooked. Whereas previous studies were performed in linear scale, we develop a nonlinear regression, which simultaneously analyzes cell type composition in linear scale and differential expression/methylation in log/logit scale. The second issue is multicollinearity. Cell-type-specific effects of a trait, such as disease, on omics expression are usually estimated by linear

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.18.158758; this version posted June 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

regression that includes terms representing the interaction between the cell type proportions and the trait. We show that the interaction terms can mutually be highly correlated, which obstructs ordinary regression. To cope with the multicollinearity, we implement ridge regularization. Our methods and previous ones are compared in simulated and real data.

Results

Multicollinearity of interaction terms

Typically, cell-type-specific effects of a trait on omics marker expression is analyzed by the linear regression in equation (2). The goal is to estimate $\beta_{h,k}$, the effect of trait *k* on the expression level in cell type *h*. This is estimated based on the relation between the bulk expression level Y_i of a sample and the regressor $W_{h,i}X_{i,k}$, which is an interaction term defined as the product of the cell type proportion $W_{h,i}$ and the trait value $X_{i,k}$ of the sample. The variable W_h for cell type composition cannot be mean-centered, and interaction terms involving uncentered variables cause multicollinearity [18]. We first survey the extent of multicollinearity in real data for cell-type-specific association.

In peripheral blood leukocyte data from a rheumatoid arthritis study (GSE42861), the proportion of cell types ranged from 0.59 for neutrophils to 0.01 for eosinophils (Table 1A). The proportion of neutrophils was negatively correlated with the proportion of other cell types (apart from monocytes) with correlation coefficient of −0.68 to −0.46, whereas the correlation was weaker for other pairs (Table 1B). Rheumatoid arthritis status was modestly correlated with proportions of cell types. The product of the disease status X_k, centered to have zero mean, and the proportion of a cell type becomes an interaction term. The correlation coefficients between the interaction terms are mostly >0.8, apart from eosinophils (Table 1C). The ratio of mean to SD of the proportion is high for all cell types apart from eosinophils (Table 1A). The interaction terms for high-ratio cell types are strongly correlated

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.18.158758; this version posted June 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

with X_k, which in turn causes strong correlation between the relevant interaction terms.

The situation was the same for the interaction with age in GTEx data. The granulocytes (which include neutrophils and eosinophils) were the most abundant (Table 2A). The proportion of granulocytes was negatively correlated with other cell types (apart from monocytes) with correlation coefficient of −0.89 to −0.41, and the correlation between other pairs was generally weaker (Table 2B). Age was modestly correlated with proportions of cell types. In this dataset, the ratio of mean to SD of the proportion was high in all cell types (Table 2A), which caused strong mutual correlation between interaction terms (Table 2C).

In the above empirical data, multicollinearity between interaction terms seemed to arise not due to the correlation between cell type proportions or X_k, but due to the high ratio of mean to SD in the cell type proportions. Subsequently, this property was derived mathematically. As we derived in equation (17), the correlation between interaction terms $W_h X_k$ and $W_{h'} X_k$ approaches to one, when the ratios $\mathrm{E}[W_h]/\mathrm{SD}[W_h]$ and $\mathrm{E}[W_{h'}]/\mathrm{SD}[W_{h'}]$ are high, irrespective of $\mathrm{Cor}[W_h, W_{h'}]$ (Figure 1). The ratio was 1.6 to 5.3 (apart from eosinophils) in the rheumatoid arthritis dataset and ≥4.3 in the GTEx dataset. We looked up datasets of several ethnicities and found the ratio to be ≥1.5 in majority of cell types (Additional file 1: Table S1). Thus, multicollinearity can be a common problem for cell-type-specific association analyses.

Evaluation in simulated data

By using simulated data, we evaluated previous methods and new approaches of the omicwas package. In order to simultaneously analyze two scales, the linear scale for heterogeneous cell mixing and the log/logit scale for trait effects, we applied nonlinear regression in omicwas (equations (4) and (5)). To cope with the multicollinearity of interaction terms, we applied ridge regularization (equations (9) and (10)).

Previous regression type methods are based either on the full model of linear regression (equation (2)) or the marginal model (equation (3)). The

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.18.158758; this version posted June 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

full model fits and tests cell-type-specific effects for all cell types simultaneously, and its derivatives include TOAST, csSAM.lm, CellDMC.unfiltered and CellDMC.filtered. The marginal model fits and tests cell-type-specific effect for one cell type at a time, and its derivatives include csSAM.monovariate and TCA.

The simulation data was generated from real datasets of DNA methylation and gene expression. The original cell type composition was retained for all samples, and the case-control status was randomly assigned. In each sample, expression level in each cell type was randomly determined according to a scenario, and then averaged according to the sample's cell type composition. Under each statistical algorithm, the disease association in the target cell type was assessed by a Z-score, comparing cases vs controls.

In scenario A for DNA methylation, expression of all cell types had identical distribution, irrespective of the case/control status (Figure 2A). The type I error rate was controlled (≤0.05) in all algorithms. In scenario B, cases had higher expression level in one randomly selected cell type, and that cell type was tested (Figure 2B). Here, the most appropriate algorithm is the marginal test applied to the perturbed cell type, which indeed attained the highest power. For the most abundant neutrophils, the Z-score was in the high range of 9.9 to 14.9 for the marginal test. With regards to the power, the ridge regression methods (omicwas.identity.ridge and omicwas.logit.ridge) came next. The algorithms based on full model, without ridge regularization, (Full, TOAST, CellDMC.unfiltered, omicwas.identity and omicwas.logit) gained modest power. TCA, which is similar to the marginal test, detected neutrophil-specific association with high Z-score, but the power over all cell types was modest. In scenario C, the expression level of cases was lower in one cell type, which was not the tested cell type (Figure 2C). Since the expression of the tested cell type is identical between cases and controls, a correct algorithm should detect no signal. The type I error rate was inflated, being highest for the marginal test, followed by the ridge regression methods and TCA. Extremely strong spurious signals of Z-score < −6 were detected in marginal and TCA. Scenario D combined scenarios B and C, where the tested cell type had higher expression in cases, and one non-

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.18.158758; this version posted June 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

tested cell type had lower expression in cases (Figure 2D). The distribution of neutrophil Z-score was similar to scenario B, and the spurious signals with low Z-scores were similar to scenario C. Over all scenarios, the similarity in performance of omicwas.identity vs omicwas.logit, as well as omicwas.identity.ridge vs omicwas.logit.ridge, indicates that the scaling was not influential in DNA methylation data.

The results for simulated gene expression data were similar. In scenario A with no true signal, type I error rate was controlled (≤0.05) in all algorithms (Figure 3A). In scenario B, where true signal exists only for the tested cell type, the power was the highest in marginal and relatively high in csSAM.monovariate (Figure 3B). The power was in decreasing order, omicwas.log.ridge > omicwas.identity.ridge > omicwas.log > omicwas.identity; proper scaling modestly improved performance. In scenario C, where cases have lower expression in one non-target cell type, the type I error was inflated in the negative direction, with the largest inflation in marginal, and moderate inflation in ridge regression methods and csSAM.monovariate (Figure 3C). Extremely strong false signals of Z-score < −6 occurred in marginal and csSAM.monovariate. In scenario D, where the tested cell type has higher expression in cases, while one non-tested cell type has lower expression, we could observe the overlay of power gain of scenario B and type I error inflation of scenario C (Figure 3D).

Although we roughly grouped previous algorithms into derivatives of full or derivatives of marginal, some implement treatments beyond simple linear models. For example, the TCA algorithm tends to detect neutrophil signals similarly as the marginal test (Fig. 2B), yet had smaller type I error rate (Fig. 2C).

Cell-type-specific association with rheumatoid arthritis and age

The cell-type-specific association of DNA methylation with rheumatoid arthritis was predicted using bulk peripheral blood leukocyte data and was evaluated in sorted monocytes (Figure 4A) and B cells (Figure 4B). Whereas the full model (and its derivatives) performed the best and the marginal model (and its derivatives) performed the worst in monocytes, the

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.18.158758; this version posted June 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

performance ranking was opposite in B cells. A robust algorithm would consistently achieve high performance relative to the best algorithm in each instance. Nonlinear ridge regression (omicwas.logit.ridge) was the most robust, performing 65% to 93% relative to the best method.

The cell-type-specific association of gene expression with age was predicted using whole blood data and was evaluated in sorted $CD4^+$ T cells (Figure 4C) and monocytes (Figure 4D). All algorithms performed poorly in $CD4^+$ T cells, and the marginal model performed the best in monocytes. Overall, nonlinear ridge regression (omicwas.log.ridge) was next to the marginal model, performing 21% to 47% to the marginal.

For dataset GSE42861 and for GTEx whole blood, the omicwas.logit.ridge and omicwas.log.ridge models of the omicwas package was computed in 8.1 and 0.7 hours respectively, using 8 cores of a 2.5 GHz Xeon CPU Linux server.

Discussion

Aiming to elucidate cell-type-specific trait association in DNA methylation and gene expression, this article explored two aspects, multicollinearity and scale. We observed multicollinearity in real data and derived mathematically how it emerges. To cope with the multicollinearity, we proposed ridge regression. To properly handle multiple scales simultaneously, we developed nonlinear regression. By testing in simulated and real data, we found proper scaling to modestly improve performance. In contrast, ridge regression achieved performance that was more robust than previous methods.

The statistical methods discussed in this article are applicable, in principle, to any tissue. For validation of the methods, we need datasets for bulk tissue as well as sorted cells, ideally of >100 samples. Currently, the publicly available data is limited to peripheral blood. By no means, we claim the rheumatoid arthritis EWAS datasets [19-21] or the datasets for age association of gene expression [22,23] to be representative. Nevertheless, we think verification in real data is important, which has not been performed previously in large sample size.

By the performance in simulated and real data, we can roughly divide

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.18.158758; this version posted June 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

algorithms into three groups: full (and its derivatives), marginal (and its derivatives) and ridge models. In marginal models, we test one cell type at a time. If we knew in advance that one particular cell type is associated with the trait, which would be a rare situation, testing that cell type in the marginal model is the most simple and correct approach. Indeed, under such a simulated scenario, the marginal test attained highest power (Figs. 2B, 3B). However, when the test target cell type is not associated, but instead another cell type is associated, the marginal tests can pick up false signals due to the collinearity between regressor variables (Figs. 2C, 3C). The high power and high error rate of the marginal tests can lead to unstable performance; in real data, the marginal tests were the most powerful for detecting B cell specific association with rheumatoid arthritis (Fig. 4B) but were the least powerful for monocytes (Fig. 4A). The full model tests all cell types together, and its performance was the opposite of the marginal. By fitting all cell types simultaneously, the full model adjusts for the effects of other cell types. The full models did not detect false association coming indirectly from non-target cell types (Figs. 2C, 3C), yet their power was relatively low (Figs. 2B, 3B). The ridge tests (omicwas.identity.ridge, omicwas.logit.ridge and omicwas.log.ridge) were in the middle between full and marginal tests with regards to the power (Figs. 2B, 3B, 4). The false positives of ridge tests were modest compared to the marginal tests (Figs. 2C, 3C).

We mathematically modeled and implemented the logit scale for DNA methylation and log scale for gene expression. It turns out that the improvement by formulating the nonlinear scale was negligible for DNA methylation (Fig. 2B) and modest for gene expression (Fig. 3B; omicwas.identity vs omicwas.log, and omicwas.identity.ridge vs omicwas.log.ridge). This implies that previous works, which were almost exclusively in linear scale, were not losing much power due to scaling.

Conclusions

For cell-type-specific differential expression analysis by using unsorted tissue samples, we recommend trying ridge regression as a first choice because it

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.18.158758; this version posted June 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

balances power and type I error. Although marginal tests can be powerful when the tested cell type actually is the only one associated with the trait, caution is needed due to its high type I error rate. For a signal detected by the marginal test, reanalysis in full model could be valuable. Ridge regression is preferable compared to the full model without ridge regularization because ridge estimator of the effect size has smaller MSE (equation (13)). Nonlinear regression, which models scales properly, is recommended more than the linear regression, yet the difference can be modest. We do not claim the ridge model to substitute previous models. Indeed, we think none of the current algorithms is superior to others in all aspects, indicating possibility for future improvement.

Methods

Linear regression

We begin by describing the linear regressions used in previous studies. Let the indexes be *h* for a cell type, *i* for a sample, *j* for an omics marker (CpG site or gene), *k* for a trait that has cell-type-specific effects on marker expression, and *l* for a trait that has a uniform effect across cell types. The input data is given in four matrices. The matrix $W_{h,i}$ represents cell type composition. The matrices $X_{i,k}$ and $C_{i,l}$ represent the values of the traits that have cell-type-specific and uniform effects, respectively. We assume the two matrices are centered: $\sum_i X_{i,k} = \sum_i C_{i,l} = 0$. The matrix $Y_{i,j}$ represents the omics marker expression level in tissue samples.

The parameters we estimate are the cell-type-specific trait effect $\beta_{h,j,k}$, tissue-uniform trait effect $\gamma_{j,l}$, and basal marker level $\alpha_{h,j}$ in each cell type. For the remaining of the first five sections (up to "Multicollinearity of interaction terms"), we focus on one marker *j*, and omit the index for readability. For cell type *h*, the marker level of sample *i* is

$$\alpha_h + \sum_k \beta_{h,k} X_{i,k}. \quad (1)$$

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.18.158758; this version posted June 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

This is a representative value rather than a mean because we do not model a probability distribution for cell-type-specific expression. By averaging the value over cell types with weight $W_{h,i}$, and combining with the tissue-uniform trait effects, we obtain the mean marker level in bulk tissue of sample *i*,

$$\mu_i = \sum_h \alpha_h W_{h,i} + \sum_{h,k} \beta_{h,k} W_{h,i} X_{i,k} + \sum_l \gamma_l C_{i,l}.$$

With regards to the statistical model, we assume the error of the marker level to be normally distributed with variance σ^2, independently among samples, as

$$Y_i = \mu_i + \varepsilon_i,$$
$$\varepsilon_i \sim N(0, \sigma^2).$$

The statistical significance of all parameters is tested under the *full* model of linear regression,

$$Y_i = \sum_h \alpha_h W_{h,i} + \sum_{h,k} \beta_{h,k} W_{h,i} X_{i,k} + \sum_l \gamma_l C_{i,l} + \varepsilon_i, \quad (2)$$

or its derivatives [5,10,13]. Alternatively, the cell-type-specific effects of traits can be fitted and tested for one cell type *h* at a time by the *marginal* model,

$$Y_i = \sum_{h'} \alpha_{h'} W_{h',i} + \sum_k \beta_{h,k} W_{h,i} X_{i,k} + \sum_l \gamma_l C_{i,l} + \varepsilon_i, \quad (3)$$

or its derivatives [7-9,11,14].

Nonlinear regression

Aiming to simultaneously analyze cell type composition in linear scale and differential expression/methylation in log/logit scale, we develop a nonlinear regression model. The differential analyses are performed after applying normalizing transformation. The normalizing function is the natural logarithm *f* = log for gene expression, and *f* = logit for methylation (see Background). Conventional linear regression can be formulated by defining *f* as the identity function. We denote the inverse function of *f* by *g*; *g* = exp for gene expression, and *g* = logistic for methylation. Thus, *f* converts from the linear scale to the normalized scale, and *g* does the opposite.

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.18.158758; this version posted June 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

The marker level in a specific cell type (formula (1)) is modeled in the normalized scale. The level is linearized by applying function *g*, then averaged over cell types with weight $W_{h,i}$, and normalized by applying function *f*. Combined with the tissue-uniform trait effects, the mean normalized marker level in bulk tissue of sample *i* becomes

$$\mu_i = f\left(\sum_h W_{h,i}\, g\left(\alpha_h + \sum_k \beta_{h,k} X_{i,k}\right)\right) + \sum_l \gamma_l C_{i,l}. \quad (4)$$

We assume the normalized marker level to have an error that is normally distributed with variance σ^2, independently among samples, as

$$f(Y_i) = \mu_i + \varepsilon_i, \quad (5)$$

$$\varepsilon_i \sim N(0, \sigma^2).$$

We obtain the ordinary least squares (OLS) estimator of the parameters by minimizing the residual sum of squares,

$$\mathrm{RSS} = \sum_i (f(Y_i) - \mu_i)^2, \quad (6)$$

and then estimate the error variance as

$$\widehat{\sigma^2} = \frac{1}{n-p}\mathrm{RSS}, \quad (7)$$

where *n* is the number of samples and *p* is the number of parameters [[24], section 6.3.1].

Ridge regression

The parameters $\beta_{h,k}$ for cell-type-specific effect cannot be estimated accurately by ordinary linear regression because the regressors $W_{h,i}X_{i,k}$ in equation (2) are highly correlated between cell types (see below). Multicollinearity also occurs to the nonlinear case in formula (4) because of local linearity. To cope with the multicollinearity, we apply ridge regression with a regularization parameter $\lambda \geq 0$, and obtain the ridge estimator of the parameters that minimizes

$$\mathrm{RSS} + \lambda \sum_{h,k} {\beta_{h,k}}^2, \quad (8)$$

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.18.158758; this version posted June 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

where the second term penalizes $\beta_{h,k}$ for taking large absolute values. The ridge estimator $\widehat{\boldsymbol{\theta}}(\lambda)$ is asymptotically normally distributed (see Additional file 2: Supplementary note) with

$$\mathrm{Mean}\left[\widehat{\boldsymbol{\theta}}(\lambda)\right] = Q(\lambda)^{-1}\, Q(0)\, \boldsymbol{\theta}, \quad (9)$$

$$\mathrm{Var}\left[\widehat{\boldsymbol{\theta}}(\lambda)\right] = \sigma^2\, Q(\lambda)^{-1} \left(\frac{\partial \boldsymbol{\mu}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)^T \left(\frac{\partial \boldsymbol{\mu}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right) Q(\lambda)^{-1}, \qquad (10)$$

$$Q(\lambda) = \left(\frac{\partial \boldsymbol{\mu}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)^T \left(\frac{\partial \boldsymbol{\mu}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right) + \lambda \begin{pmatrix} O & O & O \\ O & I & O \\ O & O & O \end{pmatrix} - (f(Y) - \boldsymbol{\mu}(\boldsymbol{\theta})) \cdot \left(\frac{\partial^2 \boldsymbol{\mu}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}\right),$$

where $\boldsymbol{\mu}$ is the vector form of μ_i, $\boldsymbol{\theta}$ is the vector form of the parameters α_h, $\beta_{h,k}$ and γ_l combined, $(\partial \boldsymbol{\mu}/\partial \boldsymbol{\theta})$ is the Jacobian matrix, $(\partial^2 \boldsymbol{\mu}/\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T)$ is the array of Hessian matrices for μ_i taken over samples, and T indicates matrix transposition. The product of $f(Y) - \boldsymbol{\mu}(\boldsymbol{\theta})$ and the Hessian is taken by multiplying for each sample and then summing up over samples. The matrix after λ has one only in the diagonal corresponding to $\beta_{h,k}$. The assigned value $\boldsymbol{\theta}$ is the true parameter value. By taking the expectation of Q, we obtain a rougher approximation [25] as

$$\mathrm{Mean}\left[\widehat{\boldsymbol{\theta}}(\lambda)\right] = Q^*(\lambda)^{-1}\, Q^*(0)\, \boldsymbol{\theta}, \qquad (11)$$

$$\mathrm{Var}\left[\widehat{\boldsymbol{\theta}}(\lambda)\right] = \sigma^2\, Q^*(\lambda)^{-1} \left(\frac{\partial \boldsymbol{\mu}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)^T \left(\frac{\partial \boldsymbol{\mu}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right) Q^*(\lambda)^{-1}, \qquad (12)$$

$$Q^*(\lambda) = \mathrm{E}[Q(\lambda)] = \left(\frac{\partial \boldsymbol{\mu}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)^T \left(\frac{\partial \boldsymbol{\mu}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right) + \lambda \begin{pmatrix} O & O & O \\ O & I & O \\ O & O & O \end{pmatrix}.$$

The matrices Q and Q^* are the observed and expected Fisher matrices multiplied by σ^2 and adapted to ridge regression, respectively.

Since our objective is to predict the cell-type-specific trait effects, we choose the regularization parameter λ that can minimize the mean squared error (MSE) of $\beta_{h,k}$. Our methodology is based on [26]. To simplify the explanation, we assume the Jacobian matrices $(\partial \boldsymbol{\mu}(\boldsymbol{\theta})/\partial \boldsymbol{\alpha})$, $(\partial \boldsymbol{\mu}(\boldsymbol{\theta})/\partial \boldsymbol{\beta})$ and $(\partial \boldsymbol{\mu}(\boldsymbol{\theta})/\partial \boldsymbol{\gamma})$ to be mutually orthogonal, where $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are the vector forms of α_h, $\beta_{h,k}$ and γ_l, respectively. Then, from formulae (11) and (12), the ridge estimator $\widehat{\boldsymbol{\beta}}(\lambda)$ is asymptotically normally distributed with

$$\mathrm{Mean}\left[\widehat{\boldsymbol{\beta}}(\lambda)\right] = \left[\left(\frac{\partial \boldsymbol{\mu}(\boldsymbol{\theta})}{\partial \boldsymbol{\beta}}\right)^T \left(\frac{\partial \boldsymbol{\mu}(\boldsymbol{\theta})}{\partial \boldsymbol{\beta}}\right) + \lambda I\right]^{-1} \left(\frac{\partial \boldsymbol{\mu}(\boldsymbol{\theta})}{\partial \boldsymbol{\beta}}\right)^T \left(\frac{\partial \boldsymbol{\mu}(\boldsymbol{\theta})}{\partial \boldsymbol{\beta}}\right) \boldsymbol{\beta},$$

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.18.158758; this version posted June 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

$$\mathrm{Var}[\widehat{\boldsymbol{\beta}}(\lambda)] = \sigma^2 \left[\left(\frac{\partial \boldsymbol{\mu}(\boldsymbol{\theta})}{\partial \boldsymbol{\beta}}\right)^T \left(\frac{\partial \boldsymbol{\mu}(\boldsymbol{\theta})}{\partial \boldsymbol{\beta}}\right) + \lambda I\right]^{-1} \left(\frac{\partial \boldsymbol{\mu}(\boldsymbol{\theta})}{\partial \boldsymbol{\beta}}\right)^T \left(\frac{\partial \boldsymbol{\mu}(\boldsymbol{\theta})}{\partial \boldsymbol{\beta}}\right) \left[\left(\frac{\partial \boldsymbol{\mu}(\boldsymbol{\theta})}{\partial \boldsymbol{\beta}}\right)^T \left(\frac{\partial \boldsymbol{\mu}(\boldsymbol{\theta})}{\partial \boldsymbol{\beta}}\right) + \lambda I\right]^{-1},$$

where the assigned values $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ are the true parameter values. We apply singular value decomposition

$$\left(\frac{\partial \boldsymbol{\mu}(\boldsymbol{\theta})}{\partial \boldsymbol{\beta}}\right) = UDV^T,$$

where U and V are orthogonal matrices, the columns of V are $\boldsymbol{v}_1, \cdots, \boldsymbol{v}_M$, and the diagonals of diagonal matrix D are sorted $d_1 \geq \cdots \geq d_M \geq 0$. The bias, variance and MSE of the ridge estimator are decomposed as

$$\begin{aligned}
\mathrm{Bias}[\widehat{\boldsymbol{\beta}}(\lambda)] &= \mathrm{E}[\widehat{\boldsymbol{\beta}}(\lambda) - \boldsymbol{\beta}] \\
&= -\lambda \left[\left(\frac{\partial \boldsymbol{\mu}(\boldsymbol{\theta})}{\partial \boldsymbol{\beta}}\right)^T \left(\frac{\partial \boldsymbol{\mu}(\boldsymbol{\theta})}{\partial \boldsymbol{\beta}}\right) + \lambda I\right]^{-1} \boldsymbol{\beta} \\
&= \left\{\sum_{m=1}^{M} \boldsymbol{v}_m \frac{-\lambda}{{d_m}^2 + \lambda} {\boldsymbol{v}_m}^T\right\} \boldsymbol{\beta}, \\
\mathrm{Var}[\widehat{\boldsymbol{\beta}}(\lambda)] &= \sigma^2 \sum_{m=1}^{M} \boldsymbol{v}_m \frac{{d_m}^2}{\left({d_m}^2 + \lambda\right)^2} {\boldsymbol{v}_m}^T, \\
\mathrm{MSE}[\widehat{\boldsymbol{\beta}}(\lambda)] &= \mathrm{E}\left[\left\|\widehat{\boldsymbol{\beta}}(\lambda) - \boldsymbol{\beta}\right\|^2\right] \\
&= \left\|\mathrm{Bias}[\widehat{\boldsymbol{\beta}}(\lambda)]\right\|^2 + \mathrm{tr}\left(\mathrm{Var}[\widehat{\boldsymbol{\beta}}(\lambda)]\right) \\
&= \sum_{m=1}^{M} \left(\frac{\lambda}{{d_m}^2 + \lambda}\right)^2 ({\boldsymbol{v}_m}^T \boldsymbol{\beta})^2 + \left(\frac{{d_m}^2}{{d_m}^2 + \lambda}\right)^2 \left(\frac{\sigma^2}{{d_m}^2}\right). \quad (13)
\end{aligned}$$

For each m in the summation of (13), the minimum of the summand is attained at $\lambda_m = \sigma^2/({\boldsymbol{v}_m}^T \boldsymbol{\beta})^2$. To minimize MSE, we need to find some "average" of the optimal λ_m over the range of m. Hoerl et al. [27] proposed to take the harmonic mean $\lambda = M\sigma^2/\|\boldsymbol{\beta}\|^2$. However, if an OLS estimator $\widehat{\boldsymbol{\beta}}(0)$ is plugged in, $\|\boldsymbol{\beta}\|^2$ is biased upwards, and λ is biased downwards. Indeed, with regards to the estimator of $1/\sqrt{\lambda_m}$, we notice that

$$\frac{1}{\sigma} {\boldsymbol{v}_m}^T \widehat{\boldsymbol{\beta}}(0) \sim N\left(\frac{1}{\sigma} {\boldsymbol{v}_m}^T \boldsymbol{\beta}, \frac{1}{{d_m}^2}\right),$$

where the terms with larger m have larger variance. Thus, we take the average of $({\boldsymbol{v}_m}^T \widehat{\boldsymbol{\beta}}(0))^2/\sigma^2$, weighted by ${d_m}^2/\sum_{m=1}^{M} {d_m}^2$, and also subtract the upward bias as,

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.18.158758; this version posted June 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

$$\kappa = \frac{1}{\sum_{m=1}^{M} {d_m}^2} \sum_{m=1}^{M} \left\{ \frac{{d_m}^2 \left({\boldsymbol{v}_m}^T \widehat{\boldsymbol{\beta}}(0) \right)^2}{\sigma^2} - 1 \right\}. \quad (14)$$

The weighting and subtraction were mentioned in [26], where the subtraction term was dismissed, under the assumption of large effect-size $\boldsymbol{\beta}$. Since the effect-size could be small in our application, we keep the subtraction term. The statistic κ can be nonpositive, and is unbiased in the sense that

$$\mathrm{E}[\kappa] = \frac{1}{\sum_{m=1}^{M} {d_m}^2} \sum_{m=1}^{M} \frac{{d_m}^2 ({\boldsymbol{v}_m}^T \boldsymbol{\beta})^2}{\sigma^2} = \frac{1}{\sum_{m=1}^{M} {d_m}^2} \sum_{m=1}^{M} \frac{{d_m}^2}{\lambda_m}.$$

Our choice of regularization parameter is

$$\lambda = \begin{cases} 1/\kappa & \text{if } \kappa > 0, \\ {d_1}^2 & \text{otherwise,} \end{cases} \quad (15)$$

where ${d_1}^2$ is taken instead of positive infinity.

Implementation of omicwas package

For each omics marker, the parameters $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ (denoted in combination by $\boldsymbol{\theta}$) are estimated and tested by nonlinear ridge regression in the following steps. As we assume the magnitude of trait effects $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ to be much smaller than that of basal marker level $\boldsymbol{\alpha}$, we first fit $\boldsymbol{\alpha}$ alone for numerical stability.

1. Compute OLS estimator $\widehat{\boldsymbol{\alpha}}(0)$ by minimizing formula (6) under $\boldsymbol{\beta} = \boldsymbol{\gamma} = \mathbf{0}$. Apply Wald test.
2. Calculate $\widehat{\sigma^2}$ by formula (7). Use it as a substitute for σ^2. The residual degrees of freedom $n - p$ is the number of samples minus the number of parameters in $\boldsymbol{\alpha}$.
3. Compute OLS estimators $\widehat{\boldsymbol{\beta}}(0)$ and $\widehat{\boldsymbol{\gamma}}(0)$ by minimizing formula (6) under $\boldsymbol{\alpha} = \widehat{\boldsymbol{\alpha}}(0)$. Let $\widehat{\boldsymbol{\theta}}(0) = (\widehat{\boldsymbol{\alpha}}(0)^T, \widehat{\boldsymbol{\beta}}(0)^T, \widehat{\boldsymbol{\gamma}}(0)^T)^T$.
4. Apply singular value decomposition $\left(\partial \boldsymbol{\mu}(\widehat{\boldsymbol{\theta}}(0)) / \partial \boldsymbol{\beta}\right) = UDV^T$.
5. Calculate κ and then the regularization parameter λ by formulae (14) and (15).
6. Compute ridge estimators $\widehat{\boldsymbol{\beta}}(\lambda)$ and $\widehat{\boldsymbol{\gamma}}(\lambda)$ by minimizing formula (8) under $\boldsymbol{\alpha} = \widehat{\boldsymbol{\alpha}}(0)$. Let $\widehat{\boldsymbol{\theta}}(\lambda) = (\widehat{\boldsymbol{\alpha}}(0)^T, \widehat{\boldsymbol{\beta}}(\lambda)^T, \widehat{\boldsymbol{\gamma}}(\lambda)^T)^T$.
7. Approximate the variance of ridge estimator, according to formula (10),

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.18.158758; this version posted June 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

by

$$\mathrm{Var}\left[\begin{pmatrix}\widehat{\boldsymbol{\beta}}(\lambda)\\ \widehat{\boldsymbol{\gamma}}(\lambda)\end{pmatrix}\right] = \widehat{\sigma^2}\, Q(\lambda)^{-1} \left(\frac{\partial \boldsymbol{\mu}\left(\widehat{\boldsymbol{\theta}}(\lambda)\right)}{\partial \begin{pmatrix}\boldsymbol{\beta}\\ \boldsymbol{\gamma}\end{pmatrix}}\right)^T \left(\frac{\partial \boldsymbol{\mu}\left(\widehat{\boldsymbol{\theta}}(\lambda)\right)}{\partial \begin{pmatrix}\boldsymbol{\beta}\\ \boldsymbol{\gamma}\end{pmatrix}}\right) Q(\lambda)^{-1},$$

$$Q(\lambda) = \left(\frac{\partial \boldsymbol{\mu}\left(\widehat{\boldsymbol{\theta}}(\lambda)\right)}{\partial \begin{pmatrix}\boldsymbol{\beta}\\ \boldsymbol{\gamma}\end{pmatrix}}\right)^T \left(\frac{\partial \boldsymbol{\mu}\left(\widehat{\boldsymbol{\theta}}(\lambda)\right)}{\partial \begin{pmatrix}\boldsymbol{\beta}\\ \boldsymbol{\gamma}\end{pmatrix}}\right) + \lambda \begin{pmatrix} I & O\\ O & O\end{pmatrix} - \left(f(Y) - \boldsymbol{\mu}\left(\widehat{\boldsymbol{\theta}}(\lambda)\right)\right) \cdot \left(\frac{\partial^2 \boldsymbol{\mu}\left(\widehat{\boldsymbol{\theta}}(\lambda)\right)}{\partial \begin{pmatrix}\boldsymbol{\beta}\\ \boldsymbol{\gamma}\end{pmatrix} \partial \begin{pmatrix}\boldsymbol{\beta}\\ \boldsymbol{\gamma}\end{pmatrix}^T}\right).$$

8. Apply the "non-exact" *t*-type test [28]. For the *s*-th coordinate,

$$\frac{\begin{pmatrix}\widehat{\boldsymbol{\beta}}(\lambda)\\ \widehat{\boldsymbol{\gamma}}(\lambda)\end{pmatrix}_s}{\sqrt{\mathrm{Var}\left[\begin{pmatrix}\widehat{\boldsymbol{\beta}}(\lambda)\\ \widehat{\boldsymbol{\gamma}}(\lambda)\end{pmatrix}\right]_{s,s}}} \sim t_{n-p}, \quad (16)$$

under the null hypothesis $\begin{pmatrix}\boldsymbol{\beta}\\ \boldsymbol{\gamma}\end{pmatrix}_s = 0$.

The formula (16) is the same as a Wald test, but the test differs, because the ridge estimators are not maximum-likelihood estimators. The algorithm was implemented as a package for the R statistical language. We used the NL2SOL algorithm of the PORT library [29] for minimization.

In analyses of quantitative trait locus (QTL), such as methylation QTL (mQTL) and expression QTL (eQTL), an association analysis that takes the genotypes of a single nucleotide polymorphism (SNP) as $X_{i,k}$ is repeated for many SNPs. In order to speed up the computation, we perform rounds of linear regression. First, the parameters $\widehat{\boldsymbol{\alpha}}(0)$ and $\widehat{\boldsymbol{\gamma}}(0)$ are fit by ordinary linear regression under $\boldsymbol{\beta} = \mathbf{0}$, which does not depend on $X_{i,k}$. By taking the residuals, we practically dispense with $\widehat{\boldsymbol{\alpha}}(0)$ and $\widehat{\boldsymbol{\gamma}}(0)$ in the remaining steps. Next, for $X_{i,k}$ of each SNP, $\widehat{\boldsymbol{\beta}}(0)$ is fit by ordinary linear regression under $\boldsymbol{\alpha} = \widehat{\boldsymbol{\alpha}}(0)$, $\boldsymbol{\gamma} = \widehat{\boldsymbol{\gamma}}(0)$. The regularization parameter λ is computed according to steps 4 and 5 above. Finally, $\widehat{\boldsymbol{\beta}}(\lambda)$ is fitted and tested by linear ridge regression under $\boldsymbol{\alpha} = \widehat{\boldsymbol{\alpha}}(0)$, $\boldsymbol{\gamma} = \widehat{\boldsymbol{\gamma}}(0)$.

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.18.158758; this version posted June 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Multicollinearity of interaction terms

The regressors for cell-type-specific trait effects in the full model (equation (2)) are the interaction terms $W_{h,i}X_{i,k}$. To assess multicollinearity, we mathematically derive the correlation coefficient between two interaction terms $W_{h,i}X_{i,k}$ and $W_{h',i}X_{i,k}$. In this section, we treat $W_{h,i}$, $W_{h',i}$ and $X_{i,k}$ as sampled instances of random variables W_h, $W_{h'}$ and X_k, respectively. For simplicity, we assume W_h and $W_{h'}$ are independent of X_k. Let E[•], Var[•], Cov[•] and Cor[•] denote the expectation, variance, covariance and correlation, respectively. Since X_k is centered, $\mathrm{E}[W_h X_k] = \mathrm{E}[W_{h'} X_k] = 0$. The correlation coefficient between interaction terms becomes

$$\mathrm{Cor}[W_h X_k, W_{h'} X_k] = \frac{\mathrm{E}[W_h X_k W_{h'} X_k]}{\sqrt{\mathrm{E}[{W_h}^2 {X_k}^2]}\sqrt{\mathrm{E}[{W_{h'}}^2 {X_k}^2]}}$$

$$= \frac{\mathrm{E}[W_h W_{h'}]}{\sqrt{\mathrm{E}[{W_h}^2]}\sqrt{\mathrm{E}[{W_{h'}}^2]}}$$

$$= \frac{\mathrm{Cov}[W_h, W_{h'}] + \mathrm{E}[W_h]\,\mathrm{E}[W_{h'}]}{\sqrt{\mathrm{Var}[W_h] + \mathrm{E}[W_h]^2}\sqrt{\mathrm{Var}[W_{h'}] + \mathrm{E}[W_{h'}]^2}}$$

$$= \frac{\mathrm{Cor}[W_h, W_{h'}] + \frac{\mathrm{E}[W_h]}{\sqrt{\mathrm{Var}[W_h]}}\frac{\mathrm{E}[W_{h'}]}{\sqrt{\mathrm{Var}[W_{h'}]}}}{\sqrt{1 + \frac{\mathrm{E}[W_h]^2}{\mathrm{Var}[W_h]}}\sqrt{1 + \frac{\mathrm{E}[W_{h'}]^2}{\mathrm{Var}[W_{h'}]}}}. \quad (17)$$

If the ratios $\mathrm{E}[W_h]/\sqrt{\mathrm{Var}[W_h]}$ and $\mathrm{E}[W_{h'}]/\sqrt{\mathrm{Var}[W_{h'}]}$ are high, the correlation of interaction terms approaches to one, irrespective of $\mathrm{Cor}[W_h, W_{h'}]$.

EWAS of rheumatoid arthritis

EWAS datasets for rheumatoid arthritis were downloaded from the Gene Expression Omnibus (GEO). Using the RnBeads package (version 2.2.0) [30] of R, IDAT files of HumanMethylation450 array were preprocessed by removing low quality samples and markers, by normalizing methylation level, and by removing markers on sex chromosomes and outlier samples. The association of methylation level with disease status was tested with

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.18.158758; this version posted June 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

adjustment for sex, age, smoking status and experiment batch; the covariates were assumed to have uniform effects across cell types. After quality control, dataset GSE42861 included bulk peripheral blood leukocyte data for 336 cases and 322 controls [20]. GSE131989 included sorted $CD14^+$ monocyte data for 63 cases and 31 controls [21]. By meta-analysis of GSE131989 and GSE87095 [19], we obtained sorted $CD19^+$ B cell data for 108 cases and 95 controls. The cell type composition of bulk samples was imputed using the Houseman algorithm [31] in the GLINT software (version 1.0.4) [32].

Differential gene expression by age

Whole blood RNA-seq data of GTEx v7 was downloaded from the GTEx website [22]. Genes of low quality or on sex chromosomes were removed, expression level was normalized, outlier samples were removed, and 389 samples were retained. The association of read count with age was tested with adjustment for sex. From GEO dataset GSE56047 [23], we obtained sorted $CD14^+$ monocyte data for 1202 samples and sorted $CD4^+$ T cell data for 214 samples. The cell type composition of bulk samples was imputed using the DeconCell package (version 0.1.0) [9] of R.

Simulation of cell-type-specific disease association

Bulk tissue sample data for case-control comparison were simulated based on real data. We generated four scenarios. Each omics marker was simulated independently. The mean expression level was defined for each cell type, separately in cases and controls. The standard deviation (SD) was set to be the same for each combination. We tested disease association specific to one cell type, which we call the target cell type. In each scenario, the mean expression level was set as follows.

A. The mean was equal for all cell types both in cases and controls (null scenario).
B. The mean in cases was higher by 1 SD for the target cell type. Other combinations had the same mean value.

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.18.158758; this version posted June 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

C. The mean in cases was lower by 1 SD for one non-target cell type. Other combinations had the same mean value.

D. The mean in cases was higher by 1 SD for the target cell type, and lower by 1 SD for one non-target cell type. Other combinations had the same middle mean value.

The target and non-target cell types were randomly chosen for each marker. For each sample, the cell-type-specific expression level was randomly sampled from a normal distribution that was specified in the scenario. The cell-type-specific expression levels were converted to the linear scale, and then averaged across cell types, according to the predefined cell type composition. The result becomes the bulk expression level of the sample in linear scale.

We used the above-mentioned bulk tissue data, namely DNA methylation data for 658 peripheral blood leukocyte samples (GSE42861) and gene expression data for 389 whole blood samples (GTEx). We applied the same simulation procedure to each dataset. The cell type composition in the original data was retained for all samples. Half of the samples were randomly assigned as cases, and the other half were assigned as controls. Normalizing transformation (i.e., logit or log) was applied to the bulk expression data, and 500 omics markers were randomly selected. For each marker, we measured the average μ and the standard deviation σ of the expression level. For control samples, the expression level in each cell type was sampled from $N(\mu, \sigma^2)$. For case samples, the expression level in each cell type was sampled from $N(\mu, \sigma^2)$, $N(\mu + \sigma, \sigma^2)$ or $N(\mu - \sigma, \sigma^2)$ according to the scenario.

Evaluation of statistical methods

Cell-type-specific effects of traits was statistically tested by using bulk tissue data as input. We applied the omicwas package with the normalizing function f = log, logit, identity without ridge regularization (omicwas.log, omicwas.logit, omicwas.identity) or under ridge regression (omicwas.log.ridge, omicwas.logit.ridge, omicwas.identity.ridge). The omicwas package was used also for conventional linear regression under the full and marginal models.

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.18.158758; this version posted June 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Among previous methods, we evaluated those that accept cell type composition as input and compute test statistics for cell-type-specific association. For DNA methylation data, we applied TOAST (version 1.2.0) [10], CellDMC (version 2.4.0) [13] and TCA (version 1.1.0) [14]. CellDMC first tests association for all combinations, and then filters out those not differentially methylated. We took all of the initial results as CellDMC.unfiltered; in CellDMC.filtered, Z-score was set to zero for those filtered out. For gene expression data, we applied TOAST and csSAM (version 1.4) [5]. For csSAM, we either fitted all cell types together or one cell type at a time, and denoted the results as csSAM.lm and csSAM.monovariate, respectively. The csSAM method is applicable to binomial traits but not to quantitative traits.

For simulated data, we adopted the nominal significance level $P < 0.05$ (two-sided). In scenario B, the power was defined as the frequency of Z-score > 1.96.

For the association with rheumatoid arthritis and age, "true" association was determined from the measurements in physically sorted blood cells, under the nominal significance level $P < 0.05$ (two-sided). The significant markers were "up-regulated" (in rheumatoid arthritis cases or elders) or "down-regulated." For a set of differentially expressed markers in a cell type (e.g., up-regulated in monocytes), the prediction performance of an algorithm was measured by the area under the curve (AUC) of receiver operating characteristic (ROC). Standard error of AUC was computed by the jackknife estimator by splitting the markers into 100 groups by chromosomal position. The relative performance of an algorithm was evaluated by its AUC – 0.5 divided by that for the best algorithm in each scenario.

Supplementary information

Additional file 1: Table S1. Blood cell type proportion in Tsimane Amerindians, Caucasians and Hispanics.

Additional file 2: Supplementary note. Asymptotic distribution of ridge estimator.

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.18.158758; this version posted June 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Abbreviations

AUC : area under the curve, eQTL: expression QTL, EWAS: epigenome-wide association study, GEO: Gene Expression Omnibus, mQTL: methylation QTL, MSE: mean squared error, OLS: ordinary least squares, QTL: quantitative trait locus, ROC: receiver operating characteristic, SD: standard deviation, SNP: single nucleotide polymorphism

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

The datasets generated and analyzed during the current study are available in the figshare repository, https://dx.doi.org/10.6084/m9.figshare.10718282

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by JSPS KAKENHI [grant number JP16K07218] and by the NCGM Intramural Research Fund [grant numbers 19A2004, 20A1013]. The funding body had no role in the design and collection of the study, experiments, analyses and interpretations of data, and in writing the manuscript.

Author's contributions

FT developed the methodology, wrote the software, implemented the study, and wrote the manuscript. NK revised the manuscript. All authors read and approved the final manuscript.

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.18.158758; this version posted June 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Acknowledgements

Not applicable.

References

1. Teschendorff AE, Zheng SC. Cell-type deconvolution in epigenome-wide association studies: a review and recommendations. Epigenomics. 2017;9:757–68.

2. Sturm G, Finotello F, Petitprez F, Zhang JD, Baumbach J, Fridman WH, et al. Comprehensive evaluation of transcriptome-based cell-type quantification methods for immuno-oncology. Bioinformatics. 2019;35:i436–45.

3. Ghosh D. Mixture models for assessing differential expression in complex tissues using microarray data. Bioinformatics. 2004;20:1663–9.

4. Stuart RO, Wachsman W, Berry CC, Wang-Rodriguez J, Wasserman L, Klacansky I, et al. In silico dissection of cell-type-associated patterns of gene expression in prostate cancer. Proc Natl Acad Sci USA. National Academy of Sciences; 2004;101:615–20.

5. Shen-Orr SS, Tibshirani R, Khatri P, Bodian DL, Staedtler F, Perry NM, et al. Cell type–specific gene expression differences in complex tissues. Nat Meth. Nature Publishing Group; 2010;7:287–9.

6. Erkkilä T, Lehmusvaara S, Ruusuvuori P, Visakorpi T, Shmulevich I, Lähdesmäki H. Probabilistic analysis of gene expression measurements from heterogeneous tissues. Bioinformatics. 2010;26:2571–7.

7. Kuhn A, Thu D, Waldvogel HJ, Faull RLM, Luthi-Carter R. Population-specific expression analysis (PSEA) reveals molecular changes in diseased brain. Nat Meth. Nature Publishing Group; 2011;8:945–7.

8. Westra H-J, Arends D, Esko T, Peters MJ, Schurmann C, Schramm K, et al. Cell Specific eQTL Analysis without Sorting Cells. Pastinen T, editor. PLoS

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.18.158758; this version posted June 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Genet. Public Library of Science; 2015;11:e1005223–17.

9. Aguirre-Gamboa R, de Klein N, di Tommaso J. Deconvolution of bulk blood eQTL effects into immune cell subpopulations. bioRxiv. 2019.

10. Li Z, Wu Z, Jin P, Wu H. Dissecting differential signals in high-throughput data from complex tissues. Hancock J, editor. Bioinformatics. 2019;35:3898–905.

11. Montaño CM, Irizarry RA, Kaufmann WE, Talbot K, Gur RE, Feinberg AP, et al. Measuring cell-type specific differential methylation in human brain tissue. Genome Biol. BioMed Central; 2013;14:R94–9.

12. White N, Benton M, Kennedy D, Fox A, Griffiths L, Lea R, et al. Accounting for cell lineage and sex effects in the identification of cell-specific DNA methylation using a Bayesian model selection algorithm. Sawalha AH, editor. PLoS ONE. 2017;12:e0182455–18.

13. Zheng SC, Breeze CE, Beck S, Teschendorff AE. Identification of differentially methylated cell types in epigenome-wide association studies. Nat Meth. Nature Publishing Group; 2018;15:1059–66.

14. Rahmani E, Schweiger R, Rhead B, Criswell LA, Barcellos LF, Eskin E, et al. Cell-type-specific resolution epigenetics without the need for cell sorting or single-cell biology. Nature Communications. Nature Publishing Group; 2019;10:3417–11.

15. Cobos FA, Vandesompele J, Mestdagh P. Computational deconvolution of transcriptomics data from mixed cell populations. Bioinformatics. 2018;34:1969–79.

16. Hoyle DC, Rattray M, Jupp R, Brass A. Making sense of microarray data distributions. Bioinformatics. 2002;18:576–84.

17. Du P, Zhang X, Huang C-C, Jafari N, Kibbe WA, Hou L, et al. Comparison of Beta-value and M-value methods for quantifying methylation

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.18.158758; this version posted June 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

levels by microarray analysis. BMC Bioinformatics. BioMed Central; 2010;11:1–9.

18. Aiken LS, West SG. Multiple Regression: Testing and Interpreting Interactions. Sage Publications; 1991.

19. Julià A, Absher D, López-Lasanta M, Palau N, Pluma A, Waite Jones L, et al. Epigenome-wide association study of rheumatoid arthritis identifies differentially methylated loci in B cells. Hum Mol Genet. 2017;26:2803–11.

20. Liu Y, Aryee MJ, Padyukov L, Fallin MD, Hesselberg E, Runarsson A, et al. Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis. Nat Biotechnol. Nature Publishing Group; 2013;31:142–7.

21. Rhead B, Holingue C, Cole M, Shao X, Quach HL, Quach D, et al. Rheumatoid Arthritis Naive T Cells Share Hypermethylation Sites With Synoviocytes. Arthritis & Rheumatology. John Wiley & Sons, Ltd; 2017;69:550–9.

22. GTEx Consortium. Genetic effects on gene expression across human tissues. Nature. Nature Publishing Group; 2017;550:204–13.

23. Reynolds LM, Taylor JR, Ding J, Lohman K, Johnson C, Siscovick D, et al. Age-related variations in the methylome associated with gene expression in human monocytes and T cells. Nature Communications. 2014;5:5366.

24. Riazoshams H, Midi H, Ghilagaber G. Robust Nonlinear Regression: with Applications using R. John Wiley & Sons; 2019.

25. Lim C. Robust ridge regression estimators for nonlinear models with applications to high throughput screening assay data. Statist. Med. 2014;34:1185–98.

26. Lawless JF, Wang P. A simulation study of ridge and other regression estimators. Communications in Statistics - Theory and Methods.

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.18.158758; this version posted June 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

1976;5:307–23.

27. Hoerl AE, Kannard RW, Baldwin KF. Ridge regression: some simulations. Communications in Statistics - Theory and Methods. 1975;4:105–23.

28. Halawa AM, Bassiouni El MY. Tests of regression coefficients under ridge regression models. Journal of Statistical Computation and Simulation. 2000;65:341–56.

29. Dennis JE, Gay DM, Welsch RE. An adaptive nonlinear least-squares algorithm. ACM Transactions on Mathematical Software. 1981;7:348–68.

30. Müller F, Scherer M, Assenov Y, Lutsik P, Walter J, Lengauer T, et al. RnBeads 2.0: comprehensive analysis of DNA methylation data. Genome Biol. BioMed Central; 2019;20:55–12.

31. Houseman EA, Accomando WP, Koestler DC, Christensen BC, Marsit CJ, Nelson HH, et al. DNA methylation arrays as surrogate measures of cell mixture distribution. BMC Bioinformatics. 2012;13:86.

32. Rahmani E, Yedidim R, Shenhav L, Schweiger R, Weissbrod O, Zaitlen N, et al. GLINT: a user-friendly toolset for the analysis of high-throughput DNA-methylation array data. Hancock JM, editor. Bioinformatics. 2017;33:1870–2.

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.18.158758; this version posted June 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

TABLES

Table 1A Blood cell type proportion in rheumatoid arthritis dataset

Cell type	Neu	CD4+T	CD8+T	NK	Mono	Bcells	Eos
Mean	0.59	0.10	0.08	0.08	0.07	0.07	0.01
SD	0.11	0.06	0.05	0.04	0.02	0.03	0.02
Mean/SD	5.3	1.8	1.6	1.9	3.2	2.6	0.4

Table 1B Correlation between blood cell type proportion and rheumatoid arthritis (X_k)

r	**Neu**	**CD4+T**	**CD8+T**	**NK**	**Mono**	**Bcells**	**Eos**	**X_k=Disease**
Neu	1	-0.68	-0.60	-0.46	-0.06	-0.49	-0.48	0.44
CD4+T	-0.68	1	0.14	0.05	-0.17	0.38	0.26	-0.33
CD8+T	-0.60	0.14	1	0.08	-0.05	0.19	0.13	-0.27
NK	-0.46	0.05	0.08	1	-0.04	0.01	0.11	-0.27
Mono	-0.06	-0.17	-0.05	-0.04	1	-0.17	0.05	0.10
Bcells	-0.49	0.38	0.19	0.01	-0.17	1	0.11	-0.22
Eos	-0.48	0.26	0.13	0.11	0.05	0.11	1	-0.10

Table 1C Correlation between interaction terms

r	**Neu*X_k**	**CD4+T*X_k**	**CD8+T*X_k**	**NK*X_k**	**Mono*X_k**	**Bcells*X_k**	**Eos*X_k**
Neu*X_k	1	0.83	0.80	0.85	0.93	0.90	0.27
CD4+T*X_k	0.83	1	0.78	0.78	0.83	0.88	0.42
CD8+T*X_k	0.80	0.78	1	0.77	0.82	0.83	0.35
NK*X_k	0.85	0.78	0.77	1	0.85	0.83	0.35
Mono*X_k	0.93	0.83	0.82	0.85	1	0.88	0.35
Bcells*X_k	0.90	0.88	0.83	0.83	0.88	1	0.36
Eos*X_k	0.27	0.42	0.35	0.35	0.35	0.36	1

Neu, neutrophils; Mono, monocytes; Eos, eosinophils.

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.18.158758; this version posted June 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

683

Table 2A Blood cell type proportion in GTEx dataset

Cell type	Gran	CD4+T	CD8+T	Mono	NK	Bcells
Mean	0.53	0.22	0.10	0.07	0.05	0.03
SD	0.037	0.020	0.013	0.004	0.012	0.003
Mean/SD	14.4	10.7	7.7	16.6	4.3	8.3

Table 2B Correlation between blood cell type proportion and age (X_k)

r	**Gran**	**CD4+T**	**CD8+T**	**Mono**	**NK**	**Bcells**	**X_k=Age**
Gran	1	-0.89	-0.83	0.56	-0.76	-0.41	-0.23
CD4+T	-0.89	1	0.59	-0.64	0.50	0.51	0.14
CD8+T	-0.83	0.59	1	-0.40	0.59	0.15	0.15
Mono	0.56	-0.64	-0.40	1	-0.44	-0.42	0.02
NK	-0.76	0.50	0.59	-0.44	1	0.13	0.31
Bcells	-0.41	0.51	0.15	-0.42	0.13	1	-0.03

Table 2C Correlation between interaction terms

r	**Gran*X_k**	**CD4+T*X_k**	**CD8+T*X_k**	**Mono*X_k**	**NK*X_k**	**Bcells*X_k**
Gran*X_k	1	0.99	0.98	1.00	0.96	0.99
CD4+T*X_k	0.99	1	1.00	0.99	0.98	1.00
CD8+T*X_k	0.98	1.00	1	0.99	0.98	0.99
Mono*X_k	1.00	0.99	0.99	1	0.96	0.99
NK*X_k	0.96	0.98	0.98	0.96	1	0.97
Bcells*X_k	0.99	1.00	0.99	0.99	0.97	1

Gra, granulocytes; Mono, monocytes.

684

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.18.158758; this version posted June 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

FIGURE LEGENDS

Figure 1

Contour plot of the correlation coefficient between interaction terms $W_h X_k$ and $W_{h'} X_k$. W_h and $W_{h'}$ represent proportions of cell types h and h', and X_k represents the value of trait *k*. For this plot, we assume the ratios $\mathrm{E}[W_h]/\mathrm{SD}[W_h]$ and $\mathrm{E}[W_{h'}]/\mathrm{SD}[W_{h'}]$ to be equal. As the ratio increases 1.5, 2 to 3, the correlation coefficient raises >0.5, >0.7 to >0.8, over most range of $\mathrm{Cor}[W_h, W_{h'}]$. SD stands for standard deviation.

Figure 2

Detection of cell-type-specific association in simulated data for DNA methylation. (A), (B), (C) and (D) correspond to the respective scenarios. Results from different algorithms are aligned horizontally. Vertical axis indicates the Z-score for the disease effect (cases vs controls) specific to the target cell type. Points are colored according to the target cell type. The middle bar of the box plot indicates the median, and the lower and upper hinges correspond to the first and third quartiles. The whiskers extend to the value no further than 1.5 * inter-quartile range from the hinges. Neu, neutrophils; Mono, monocytes; Eos, eosinophils.

Figure 3

Detection of cell-type-specific association in simulated data for gene expression. (A), (B), (C) and (D) correspond to the respective scenarios. Results from different algorithms are aligned horizontally. Vertical axis indicates the Z-score for the disease effect (cases vs controls) specific to the target cell type. Points are colored according to the target cell type. The middle bar of the box plot indicates the median, and the lower and upper hinges correspond to the first and third quartiles. The whiskers extend to the value no further than 1.5 * inter-quartile range from the hinges. Gra, granulocytes; Mono, monocytes.

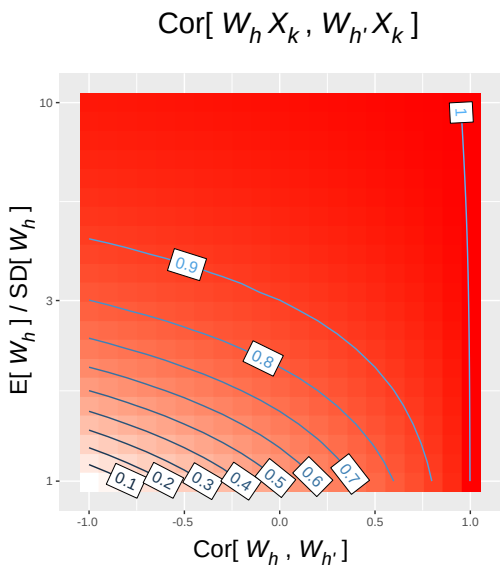
bioRxiv preprint doi: https://doi.org/10.1101/2020.06.18.158758; this version posted June 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Figure 4

Performance of cell-type-specific association prediction. For rheumatoid arthritis association of DNA methylation in monocytes (A) and B cells (B); Age association of gene expression in $CD4^+$ T cells (C) and monocytes (D). The prediction is evaluated separately for up-regulated and down-regulated markers. The AUC of ROC and its 95% confidence interval are plotted for each statistical algorithm.

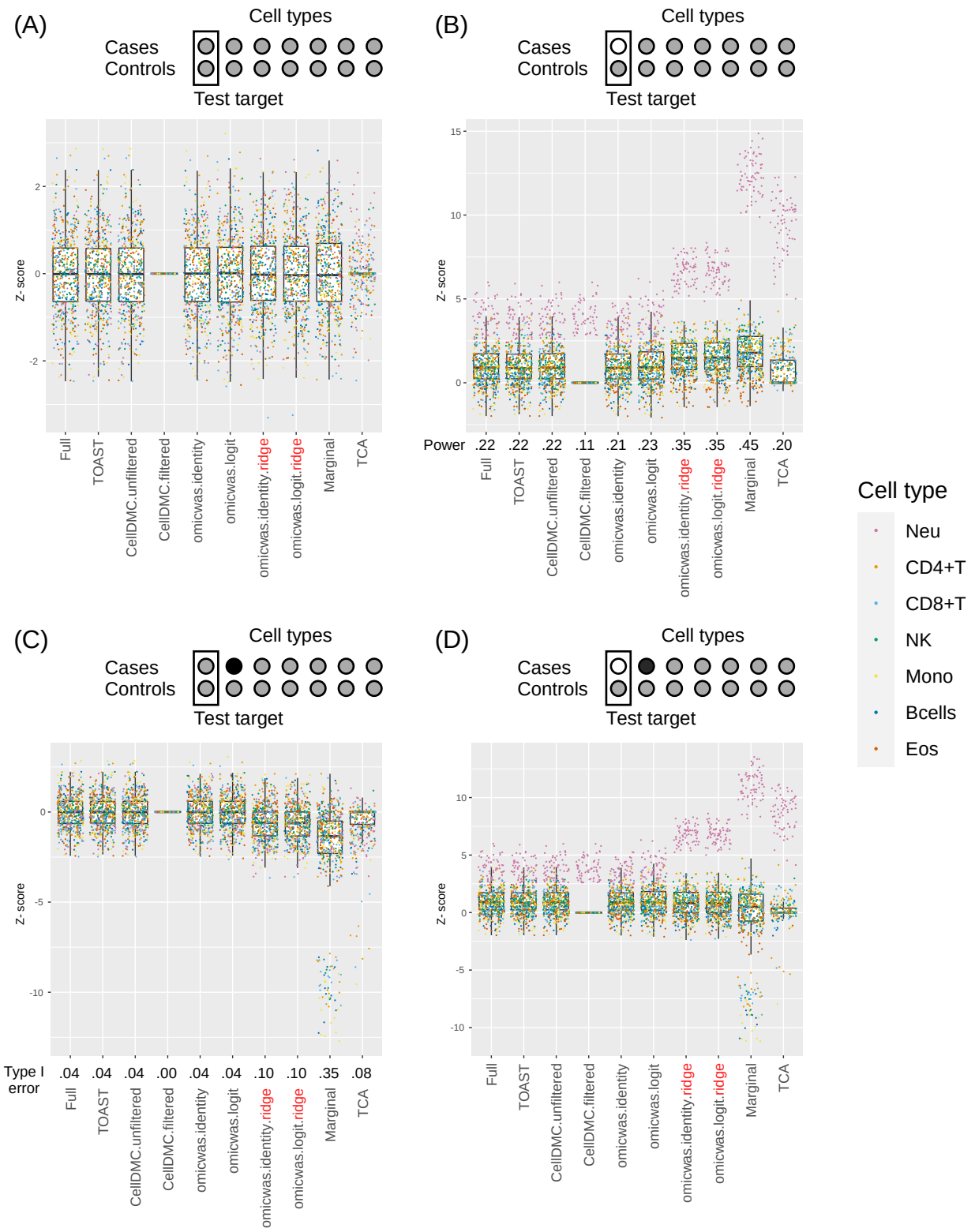
bioRxiv preprint doi: https://doi.org/10.1101/2020.06.18.158758; this version posted June 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Figure 1



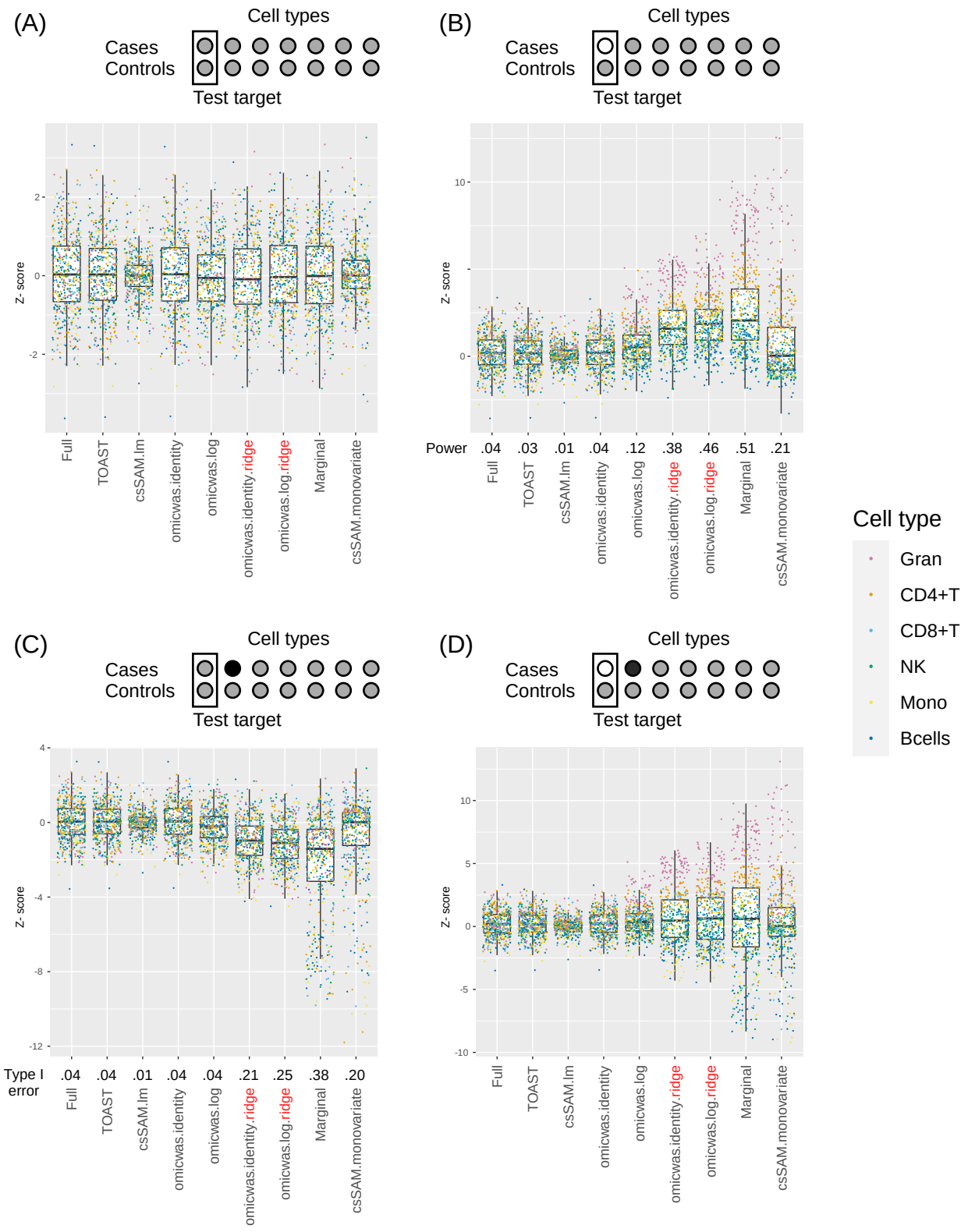
bioRxiv preprint doi: https://doi.org/10.1101/2020.06.18.158758; this version posted June 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Figure 2



bioRxiv preprint doi: https://doi.org/10.1101/2020.06.18.158758; this version posted June 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

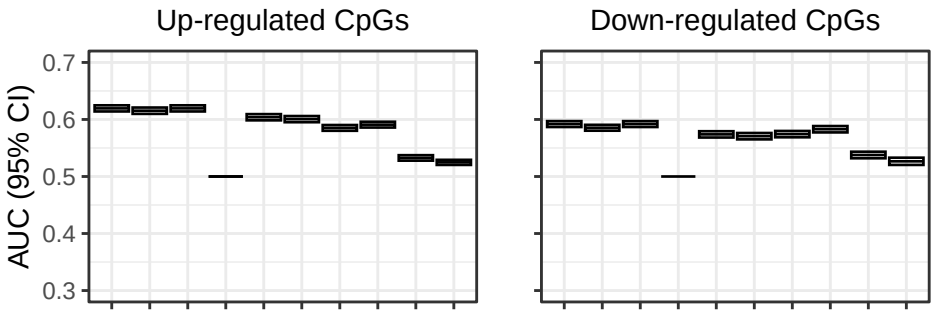
Figure 3



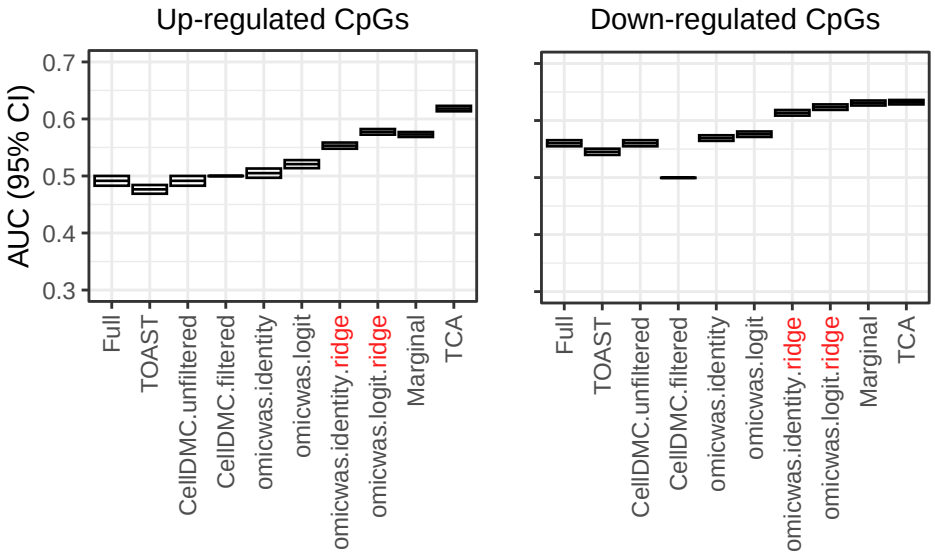
bioRxiv preprint doi: https://doi.org/10.1101/2020.06.18.158758; this version posted June 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Figure 4

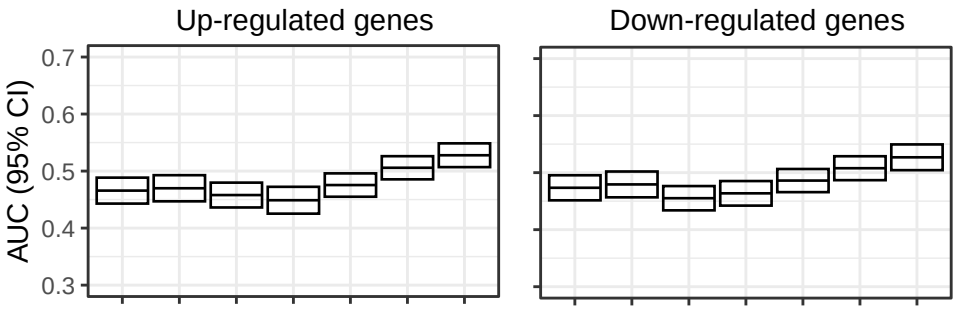
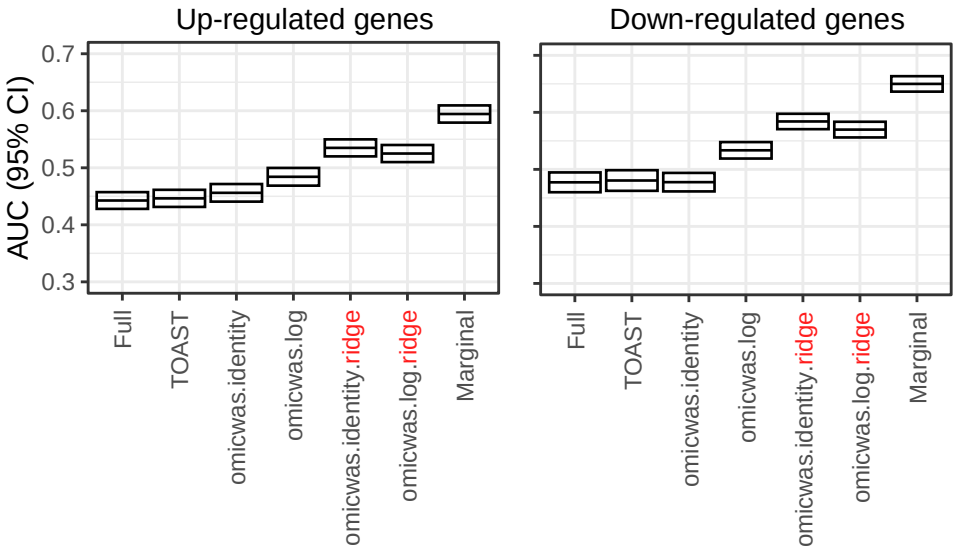
(A) Rheumatoid arthritis association in Monocytes



(B) Rheumatoid arthritis association in B cells



(C) Age association in CD4+ T cells



(D) Age association in Monocytes