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

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

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 applied 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 data, nonlinear ridge regression attained well-balanced sensitivity, specificity and precision. In real data, nonlinear ridge regression detected signals consistently over the examined cases.

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, Nonlinear regression, Ridge regression, mQTL, eQTL

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

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). For each omics marker, 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 sample *i* 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. We assume that Y_i, $W_{h,i}$ and $X_{i,k}$ are given as input data.

The variable $W_{h,i}$ for cell type composition cannot be mean-centered for our purpose. If $W_{h,i}$ were centered, we would obtain, instead of $\beta_{h,k}$, the deviation of $\beta_{h,k}$ from the average across cell types. In general, interaction terms involving uncentered variables can become collinear [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 were mostly >0.8, apart from eosinophils (Table 1C). The coefficient

of variation (CV), which is the ratio of standard deviation to mean, of the proportion was low for all cell types apart from eosinophils (Table 1A). The interaction terms for low-CV cell types were strongly correlated with X_k, which in turn caused 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 CV of the proportion was low 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 low CV 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 one when $\mathrm{CV}[W_h]$ and $\mathrm{CV}[W_{h'}]$ are low, irrespective of $\mathrm{Cor}[W_h, W_{h'}]$ (Fig. 1). The CV was 0.2 to 0.6 (apart from eosinophils) in the rheumatoid arthritis dataset and 0.1 to 0.2 in the GTEx dataset. We looked up datasets of several ethnicities and found the CV to be ≤0.6 in majority of blood cell types (Additional file 1: Table S1). Thus, multicollinearity can be a common problem for cell-type-specific association analyses. Biologically, in tissues where cell type composition is tightly controlled, the CV of cell type proportion becomes low and the multicollinearity is exacerbated.

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 (equations (4) and (5)). To

126 cope with the multicollinearity of interaction terms, we applied ridge
127 regularization (formula (8)).
128 Previous regression type methods are based either on the full model of
129 linear regression (equation (2)) or the marginal model (equation (3)). The
130 full model fits and tests cell-type-specific effects for all cell types
131 simultaneously, and its variations include TOAST, csSAM.lm and CellDMC. The
132 marginal model fits and tests cell-type-specific effect for one cell type at a
133 time, and its variations include csSAM.monovariate and TCA. We also
134 examined a hybrid of the two models (Marginal.Full005), which becomes
135 positive when the models agree.
136 The simulation data was generated from real datasets of DNA methylation
137 (658 samples; 451,725 CpG sites) and gene expression (389 samples;
138 14,038 genes). The original cell type composition was retained for all samples,
139 and the case-control status was randomly assigned. Ninety-five percent of
140 omics markers were set to be unassociated with disease status, 2.5% were
141 up-regulated in cases at one cell type, and 2.5% were similarly down-
142 regulated. The cell-type-specific effect-size was fixed in a simulation trial,
143 either to methylation odds ratio of 1.3, 1.6 or 1.9 or to gene expression fold
144 change of 1.7, 3.0 or 5.0. The significance level was set to $P < 2.4 \times 10^{-7}$ for
145 DNA methylation and false discovery rate <5% for gene expression. In each
146 simulation trial, the sensitivity, specificity and precision for detecting cell-
147 type-specific association was calculated for each cell type. To compare
148 algorithms, the performance measures for the same effect-size and cell type
149 were averaged over the simulation trials.
150 Overall, in the simulation for DNA methylation the sensitivity (Fig. 2) was
151 higher under large effect-size (bottom row of panels) and in abundant cell
152 types (left columns of panels). The average specificity (Fig. 3) was high
153 across the effect-size settings and across cell types, being >0.97 for the
154 Marginal model, >0.98 for TCA and >0.999 for other algorithms. The
155 precision (Fig. 4) was higher in abundant cell types within each effect-size
156 setting. As effect-size increased, the precision decreased in neutrophils but
157 increased in the other minor cell types. Excluding the cases where all
158 algorithms lacked sensitivity (monocytes and B cells under methOR=1.3 and

159 eosinophils), the average precision of omicwas.logit.ridge was >0.78 and was
160 the highest in 13/16 of the cases.
161 There was trade-off between sensitivity and precision. Among the
162 algorithms, the Marginal model attained the highest sensitivity and the lowest
163 precision. TCA, which is a variation of the marginal model, had relatively high
164 sensitivity and relatively low precision. Marginal.Full005 attained the second
165 highest sensitivity and moderate precision. The ridge regressions
166 (omicwas.logit.ridge and omicwas.identity.ridge) attained moderate
167 sensitivity and high precision. The full models without ridge regularization
168 (omicwas.logit, omicwas.identity, Full, TOAST and CellDMC) had the lowest
169 sensitivity.
170 The overall tendency was similar in the simulation for gene expression.
171 The sensitivity (Fig. 5) was higher under large effect-size and in abundant
172 cell types. The average specificity (Fig. 6) was high across the effect-size
173 settings and across cell types, being >0.96 for the marginal models (Marginal
174 and csSAM.monovariate), >0.98 for the nonlinear and ridge regressions
175 (omicwas.log.ridge, omicwas.identity.ridge, omicwas.log) and >0.996 for
176 other algorithms. The precision (Fig. 7) was higher in abundant cell types
177 within each effect-size setting. As effect-size increased, the precision
178 decreased in granulocytes but increased in the other minor cell types.
179 Excluding the full models (omicwas.identity, Full, TOAST, csSAM.lm) that
180 lacked sensitivity, the algorithms that were frequently top in average
181 precision were omicwas.identity.ridge (5 cases), omicwas.log (5 cases),
182 omicwas.log.ridge (3 cases) and Marginal.Full005 (3 cases).
183 There again was trade-off between sensitivity and precision. Among the
184 algorithms, the marginal models attained the highest sensitivity but relatively
185 low precision. The nonlinear and ridge regressions and Marginal.Full005
186 attained moderate sensitivity and highest precision. The full models had very
187 low average sensitivity of <0.01.
188 For gene expression, we also simulated a scenario where cell-type-specific
189 disease effect occurred in cell type "marker" genes (Additional files 2, 3, 4:
190 Figs. S1, S2, S3). In other words, the expression level in the target cell type
191 differed between cases and controls, and the expression level in other cell

192 types was zero (in linear scale). Thus, non-target cell types did not introduce
193 noise to bulk expression level. The Marginal model attained the highest
194 average sensitivity of >0.93 but relatively low average precision of ~0.16. As
195 equal number of differentially expressed genes were generated in the six cell
196 types, picking up signals for all such genes, including those not for the tested
197 cell type, would result in precision of 1/6 = 0.16. The full models had low
198 sensitivity. The nonlinear and ridge regressions and Marginal.Full005 attained
199 moderate sensitivity and moderate precision. With regards to the frequency
200 of being the top in average precision, the algorithms were ordered
201 Marginal.Full005 (9 cases), omicwas.log.ridge (6 cases),
202 omicwas.identity.ridge (2 cases) and omicwas.log (1 case), excluding the full
203 models that lacked sensitivity.

204 **Cell-type-specific association with rheumatoid arthritis and age**

205 The detection of cell-type-specific association in bulk tissue was evaluated by
206 using physically sorted cells. In principle, sorted cells should serve as genuine
207 verification, however, due to the relatively small sample size (94 or 203 for
208 rheumatoid arthritis and 214 or 1202 for age) the available datasets were
209 underpowered to generate a gold standard list of differentially expressed
210 omics markers [19]. Instead, we generated a benchmark set of differentially
211 expressed markers by imposing a relaxed significance level of $P < 0.05$; the
212 set would be enriched for true differentially expressed markers yet also
213 include unassociated markers. The benchmark set was cross-checked with
214 the prediction by each algorithm; the prediction cutoff value was varied and
215 the performance was measured by the area under the curve (AUC) of receiver
216 operating characteristic (ROC). In other words, we examined if the prediction
217 in bulk tissue by an algorithm could be replicated in sorted cells under $P <$
218 0.05.

219 The cell-type-specific association of DNA methylation with rheumatoid
220 arthritis was predicted using bulk peripheral blood leukocyte data and was
221 evaluated in sorted monocytes (Fig. 8A) and B cells (Fig. 8B). Whereas the
222 full model (and its variations) performed the best and the marginal model
223 (and its variations) performed the worst in monocytes, the performance

224 ranking was opposite in B cells. A robust algorithm would consistently achieve
225 high performance relative to the best algorithm in each instance. Nonlinear
226 ridge regression (omicwas.logit.ridge) was the most robust, performing 65%
227 to 93% relative to the best method. The hybrid Marginal.Full005 did not
228 perform well.

229 The cell-type-specific association of gene expression with age was
230 predicted using whole blood data and was evaluated in sorted $CD4^+$ T cells
231 (Fig. 8C) and monocytes (Fig. 8D). All algorithms performed poorly in $CD4^+$
232 T cells, and the marginal model performed the best in monocytes. Overall,
233 nonlinear ridge regression (omicwas.log.ridge) was next to the marginal
234 model, performing 21% to 47% to the marginal. The dataset of sorted $CD4^+$
235 T cells (214 samples) is smaller than the monocyte dataset (1202 samples)
236 thus could be underpowered to pick enough true differentially expressed
237 genes into the benchmark set.

238 For DNA methylation dataset GSE42861 and for GTEx gene expression
239 dataset, the omicwas.logit.ridge and omicwas.log.ridge models of the
240 omicwas package was computed in 8.1 and 0.7 hours respectively, using 8
241 cores of a 2.5 GHz Xeon CPU Linux server.
242

243 **Discussion**

244 Aiming to elucidate cell-type-specific trait association in DNA methylation and
245 gene expression, this article explored two aspects, multicollinearity and scale.
246 We observed multicollinearity in real data and derived mathematically how it
247 emerges. To cope with the multicollinearity, we applied ridge regularization.
248 To properly handle multiple scales simultaneously, we applied nonlinear
249 regression. Among the examined algorithms, nonlinear ridge regression
250 attained moderate sensitivity and highest precision in simulated data, and
251 achieved robust performance in real data. We also developed an algorithm
252 that combines full and marginal models, which attained balanced sensitivity
253 and precision in simulation.

254 The statistical methods discussed in this article are applicable, in principle,
255 to any tissue. For validation of the methods, we need datasets for bulk tissue

as well as sorted cells, ideally of several hundred samples. Currently, the publicly available data is limited to peripheral blood. By no means, we claim the rheumatoid arthritis EWAS datasets [20-22] or the datasets for age association of gene expression [23,24] to be representative. Nevertheless, we think verification in real data is valuable.

By the performance in simulated and real data, we can roughly divide algorithms into three groups: full (and its variations), marginal (and its variations) and the third group that includes ridge regressions and the hybrid Marginal.Full005. 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 with the marginal model is the most simple and correct approach. However, when the test target cell type is not associated, but instead another cell type is associated, the marginal models can pick up false signals due to the collinearity between regressor variables. Indeed, marginal models attained highest sensitivity (Figs. 2, 5) and relatively low precision (Figs. 4, 7), which could lead to unstable performance. In real data, the marginal models were the most powerful for detecting B cell specific association with rheumatoid arthritis (Fig. 8B) but were the least powerful for monocytes (Fig. 8A). The full model fits and tests all cell types simultaneously, by which it adjusts for the effects of other cell types. Due to the simultaneous inclusion of collinear predictors, the sensitivity was low (Figs. 2, 5). The ridge regressions (omicwas.identity.ridge, omicwas.logit.ridge and omicwas.log.ridge) were in the middle between full and marginal models with regards to the sensitivity (Figs. 2, 5), while attaining the highest specificity (Figs. 4, 7). The nonlinear ridge regressions consistently achieved high performance relative to the best algorithm in real data (Fig. 8). The hybrid Marginal.Full005 algorithm is intended to gain sensitivity by the marginal model while keeping precision by incorporating the full model. It attained moderate sensitivity (Figs. 2, 5) and moderate precision (Figs. 4, 7) in simulation but did not perform well in real data (Fig. 8).

Conclusions

For cell-type-specific differential expression analysis by using unsorted tissue samples, we recommend trying the nonlinear ridge regression as a first choice because it balances sensitivity and precision and was robust in experimental data. Although marginal models can be powerful when the tested cell type actually is the only one associated with the trait, caution is needed due to its low precision. Under the idea of first scanning by the marginal model and then reanalyzing in full model, we developed the hybrid Marginal.Full005 algorithm, which attained balanced sensitivity and precision but was not corroborated in experimental data. Ridge regression is preferable compared to the full model without ridge regularization because ridge estimator of the effect-size has smaller mean squared error (equation (13)). We do not claim the ridge regression to substitute previous algorithms. 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: $\sum_i X_{i,k} = \sum_i C_{i,l} = 0$. For example, $X_{i,k} = 0.5$ for disease cases and $X_{i,k} = -0.5$ for controls when the number of cases and controls are equal. 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

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

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 variations [5,9,13]. 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 variations [7,8,10,11,14].

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.

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 [[25], 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

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

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 5: Supplementary note) with

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

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

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

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 superscript T indicates matrix transposition. The dot product of $(f(Y) - \boldsymbol{\mu}(\boldsymbol{\theta}))^T$ and the array of Hessians 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 [26] 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 [27]. 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

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

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

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

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. [28] proposed to use the harmonic mean $\lambda = M\sigma^2/\|\boldsymbol{\beta}\|^2$. However, if an OLS estimator $\widehat{\boldsymbol{\beta}}(0)$ is actually plugged into $\|\boldsymbol{\beta}\|^2$, the denominator is biased upwards, and the computed mean 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,

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

The weighting and subtraction were mentioned in [27], 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

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

equals the weighted sum of $1/\lambda_m$. 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), by

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

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

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

8. Apply the "non-exact" *t*-type test [29]. 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 [30] 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 $\boldsymbol{\alpha} = \widehat{\boldsymbol{\alpha}}(0)$, $\boldsymbol{\gamma} = \widehat{\boldsymbol{\gamma}}(0)$.

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; note that the sample index *i* is omitted. For simplicity, we assume W_h and $W_{h'}$ are independent of X_k. Let E[•], Var[•], Cov[•], Cor[•] and CV[•] denote the expectation, variance, covariance, correlation and coefficient of variation, 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

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

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

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

$$= \frac{\dfrac{\mathrm{Cor}[W_h, W_{h'}]\sqrt{\mathrm{Var}[W_h]}\sqrt{\mathrm{Var}[W_{h'}]}}{\mathrm{E}[W_h]\mathrm{E}[W_{h'}]} + 1}{\sqrt{\dfrac{\mathrm{Var}[W_h]}{\mathrm{E}[W_h]^2} + 1}\sqrt{\dfrac{\mathrm{Var}[W_{h'}]}{\mathrm{E}[W_{h'}]^2} + 1}}$$

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

If $\mathrm{CV}[W_h]$ and $\mathrm{CV}[W_{h'}]$ approach zero, the correlation of interaction terms approaches one, irrespective of $\mathrm{Cor}[W_h, W_{h'}]$.

498 **EWAS of rheumatoid arthritis**

499 EWAS datasets for rheumatoid arthritis were downloaded from the Gene
500 Expression Omnibus. Using the RnBeads package (version 2.2.0) [31] of R,
501 IDAT files of HumanMethylation450 array were preprocessed by removing
502 low quality samples and markers, by normalizing methylation level, and by
503 removing markers on sex chromosomes and outlier samples. The association
504 of methylation level with disease status was tested with adjustment for sex,
505 age, smoking status and experiment batch; the covariates were assumed to
506 have uniform effects across cell types. After quality control, dataset
507 GSE42861 included bulk peripheral blood leukocyte data for 336 cases and
508 322 controls [21].

509 The cell type composition of bulk samples was imputed using the
510 Houseman algorithm [32] in the GLINT software (version 1.0.4) [33]. The
511 reference data of GLINT software characterizes seven cell types [34] by 300
512 CpG sites [35], of which 284 were measured in our data. We used prediction
513 results for the seven cell types (Table 1).

514 Dataset GSE131989 included sorted $CD14^+$ monocyte data for 63 cases
515 and 31 controls [22]. By meta-analysis of GSE131989 and GSE87095 [20],
516 we obtained sorted $CD19^+$ B cell data for 108 cases and 95 controls. Under
517 the nominal significance level $P < 0.05$ (two-sided), the number of CpG sites
518 up- or down-regulated in cases were 20,545 (5%) and 14,742 (3%),
519 respectively, in $CD14^+$ monocyte and 28,216 (6%) and 26,447 (6%) in $CD19^+$
520 B cell.

521 **Differential gene expression by age**

522 Whole blood RNA-seq data of GTEx v7 was downloaded from the GTEx
523 website [23]. Genes of low quality or on sex chromosomes were removed,
524 expression level was normalized, outlier samples were removed, and 389
525 samples were retained. The association of read count with age was tested
526 with adjustment for sex.

527 The cell type composition of bulk samples was imputed using the
528 DeconCell package (version 0.1.0) [10] of R. The reference data of DeconCell

529 characterizes 33 cell types by two to 217 signature genes. Our data measured
530 39% of the signature genes. We used prediction results for six main cell types
531 (Table 2), for which the prediction performance was 44.4 to 90.9.
532 From the Gene Expression Omnibus dataset GSE56047 [24], we obtained
533 sorted $CD14^+$ monocyte data for 1202 samples and sorted $CD4^+$ T cell data
534 for 214 samples. Under the nominal significance level $P < 0.05$ (two-sided),
535 the number of genes up- or down-regulated by higher age were 2081 (16%)
536 and 1984 (15%), respectively, in $CD14^+$ monocyte and 693 (5%) and 842
537 (6%) in $CD4^+$ T cell.

538 **Simulation of cell-type-specific disease association**

539 Bulk tissue sample data for case-control comparison were simulated based
540 on the above-mentioned EWAS dataset GSE42861 and GTEx gene expression
541 dataset. We randomly assigned the case-control status to the samples.
542 Among the omics markers, 2.5% were set to be up-regulated in cases in
543 single cell type, 2.5% were similarly down-regulated, and 95% were
544 unrelated to case-control status. The cell-type-specific effect-size of the
545 differentially expressed markers was fixed within a simulation trial, and was
546 chosen from methylation odds ratio (OR) of 1.3, 1.6 or 1.9 for EWAS [19]
547 and fold-change of 1.7, 3.0 or 5.0 for gene expression analysis; the effect-
548 sizes correspond to log(1.3), log(1.6) and log(1.9) or log(1.7), log(3.0) and
549 log(5.0) in normalized scale.
550 For each effect-size, we performed 20 simulation trials for the EWAS
551 dataset and 50 trials for the gene expression dataset. In each simulation trial,
552 we randomly assigned half of the samples as cases ($X_{i,k} = 0.5$) and the other
553 half as controls ($X_{i,k} = -0.5$). We retained the covariates matrix $C_{i,l}$ and the
554 cell type composition matrix $W_{h,i}$ from the original data. From the original
555 bulk expression level matrix $Y_{i,j}$, 95% of the markers were randomly chosen
556 and retained; these markers had no association with disease because the
557 case-control status was randomized. Cell-type-specific association was
558 introduced into the remaining 5% of markers, such that an equal number of
559 markers were up- or down-regulated in each cell type. For example, in the
560 EWAS dataset, $451{,}725 \times 0.05 \times 0.5 \div 7 = 1613$ CpG sites were up-regulated in

561 neutrophils of cases.

562 The bulk expression level of a marker *j* with normalized-scale effect-size
563 β specific to a cell type *h* was generated as follows. First, the average μ and
564 the variance σ^2 of the normalized bulk expression level $f(Y_{i,j})$ in the original
565 data was measured. Next, we generated normalized expression level in each
566 cell type. For cases, the expression level in cell type *h* was randomly sampled
567 from the normal distribution $N(\mu+\beta,\sigma^2)$ and the expression level in each of
568 the other cell types was sampled from $N(\mu,\sigma^2)$. For controls, the expression
569 level in each cell type was sampled from $N(\mu,\sigma^2)$. Finally, for each individual,
570 the expression levels in cell types were converted to the linear scale,
571 multiplied by the cell type composition and added, to obtain the bulk
572 expression level in linear scale.

573 For gene expression, we also simulated a scenario where cell-type-specific
574 disease effect occurs in cell type marker genes. The simulation procedure is
575 same as above except that the expression level was set to zero (in linear
576 scale) in all cell types other than the target cell type *h*, for both cases and
577 controls.

578 **Evaluation of statistical methods**

579 Cell-type-specific effects of traits was statistically tested by using bulk tissue
580 data as input. We applied the omicwas package with the normalizing function
581 *f* = log, logit, identity without ridge regularization (omicwas.log,
582 omicwas.logit, omicwas.identity) or under ridge regression
583 (omicwas.log.ridge, omicwas.logit.ridge, omicwas.identity.ridge). The
584 omicwas package was used also for conventional linear regression under the
585 full and marginal models. We also developed a hybrid of marginal and full
586 models (Marginal.Full005): if the effect direction agreed in two models and if
587 P < 0.05 in the full model, we adopted the Z-score of the marginal model;
588 otherwise, the Z-score was set to zero.

589 Among previous methods, we evaluated those that accept cell type
590 composition as input and compute test statistics for cell-type-specific
591 association. For DNA methylation data, we applied TOAST (version 1.2.0) [9],
592 CellDMC (version 2.0.2) [13] and TCA (version 1.0.0) [14]. 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 of EWAS dataset GSE42861, we adopted the significance level $P < 2.4 \times 10^{-7}$, which accounts for the correlation among the probes on HumanMethylation450 array [36]. For the GTEx gene expression dataset, multiple testing was controlled by the Benjamini-Hochberg procedure with the false discovery rate <5% in each cell type [37].

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 obtained benchmark set of differentially expressed 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 AUC of 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 setting.

Supplementary information

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

Additional file 2: Fig. S1. Sensitivity for detecting cell-type-specific association in simulated data for gene expression of marker genes.

Additional file 3: Fig. S2. Specificity for detecting cell-type-specific association in simulated data for gene expression of marker genes.

Additional file 4: Fig. S3. Precision (positive predictive value) for detecting cell-type-specific association in simulated data for gene expression of marker genes.

Additional file 5: Supplementary note. Asymptotic distribution of ridge

estimator.

Abbreviations

AUC: area under the curve, CV: coefficient of variation, eQTL: expression QTL, EWAS: epigenome-wide association study, mQTL: methylation QTL, MSE: mean squared error, OLS: ordinary least squares, OR: odds ratio, QTL: quantitative trait locus, ROC: receiver operating characteristic, 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,

650 and wrote the manuscript. NK revised the manuscript. All authors read and
651 approved the final manuscript.

652 **Acknowledgements**

653 Not applicable.

654

655 **References**

656 1. Teschendorff AE, Zheng SC. Cell-type deconvolution in epigenome-wide
657 association studies: a review and recommendations. Epigenomics.
658 2017;9:757–68.

659 2. Sturm G, Finotello F, Petitprez F, Zhang JD, Baumbach J, Fridman WH, et
660 al. Comprehensive evaluation of transcriptome-based cell-type
661 quantification methods for immuno-oncology. Bioinformatics.
662 2019;35:i436–45.

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

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

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

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

675 7. Kuhn A, Thu D, Waldvogel HJ, Faull RLM, Luthi-Carter R. Population-
676 specific expression analysis (PSEA) reveals molecular changes in diseased

677 brain. Nat Meth. Nature Publishing Group; 2011;8:945–7.

678 8. Westra H-J, Arends D, Esko T, Peters MJ, Schurmann C, Schramm K, et
679 al. Cell Specific eQTL Analysis without Sorting Cells. Pastinen T, editor. PLoS
680 Genet. Public Library of Science; 2015;11:e1005223–17.

681 9. Li Z, Wu Z, Jin P, Wu H. Dissecting differential signals in high-throughput
682 data from complex tissues. Hancock J, editor. Bioinformatics.
683 2019;35:3898–905.

684 10. Aguirre-Gamboa R, de Klein N, di Tommaso J, Claringbould A, van der
685 Wijst MG, de Vries D, et al. Deconvolution of bulk blood eQTL effects into
686 immune cell subpopulations. BMC Bioinformatics. 2020;21:243.

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

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

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

697 14. Rahmani E, Schweiger R, Rhead B, Criswell LA, Barcellos LF, Eskin E, et
698 al. Cell-type-specific resolution epigenetics without the need for cell sorting
699 or single-cell biology. Nature Communications. Nature Publishing Group;
700 2019;10:3417–11.

701 15. Cobos FA, Vandesompele J, Mestdagh P. Computational deconvolution
702 of transcriptomics data from mixed cell populations. Bioinformatics.
703 2018;34:1969–79.

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

706 17. Du P, Zhang X, Huang C-C, Jafari N, Kibbe WA, Hou L, et al.
707 Comparison of Beta-value and M-value methods for quantifying methylation
708 levels by microarray analysis. BMC Bioinformatics. BioMed Central;
709 2010;11:1–9.

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

712 19. Rakyan VK, Down TA, Balding DJ, Beck S. Epigenome-wide association
713 studies for common human diseases. Nat Rev Genet. 2011;12:529–41.

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

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

721 22. Rhead B, Holingue C, Cole M, Shao X, Quach HL, Quach D, et al.
722 Rheumatoid Arthritis Naive T Cells Share Hypermethylation Sites With
723 Synoviocytes. Arthritis & Rheumatology. John Wiley & Sons, Ltd;
724 2017;69:550–9.

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

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

730 25. Riazoshams H, Midi H, Ghilagaber G. Robust Nonlinear Regression: with

731 Applications using R. John Wiley & Sons; 2019.

732 26. Lim C. Robust ridge regression estimators for nonlinear models with
733 applications to high throughput screening assay data. Statist. Med.
734 2014;34:1185–98.

735 27. Lawless JF, Wang P. A simulation study of ridge and other regression
736 estimators. Communications in Statistics - Theory and Methods.
737 1976;5:307–23.

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

740 29. Halawa AM, Bassiouni El MY. Tests of regression coefficients under ridge
741 regression models. Journal of Statistical Computation and Simulation.
742 2000;65:341–56.

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

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

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

751 33. Rahmani E, Yedidim R, Shenhav L, Schweiger R, Weissbrod O, Zaitlen
752 N, et al. GLINT: a user-friendly toolset for the analysis of high-throughput
753 DNA-methylation array data. Hancock JM, editor. Bioinformatics.
754 2017;33:1870–2.

755 34. Reinius LE, Acevedo N, Joerink M, Pershagen G, Dahlén S-E, Greco D, et
756 al. Differential DNA Methylation in Purified Human Blood Cells: Implications
757 for Cell Lineage and Studies on Disease Susceptibility. Ting AH, editor. PLoS

758 ONE. Public Library of Science; 2012;7:e41361–13.

759 35. Koestler D. Improving Cell Mixture Deconvolution by Identifying Optimal
760 DNA methylation Libraries (IDOL). BMC Bioinformatics. BMC Bioinformatics;
761 2016;:1–21.

762 36. Saffari A, Silver MJ, Zavattari P, Moi L, Columbano A, Meaburn EL, et al.
763 Estimation of a significance threshold for epigenome-wide association
764 studies. Genet Epidemiol. John Wiley & Sons, Ltd; 2017;42:20–33.

765 37. Anders S, McCarthy DJ, Chen Y, Okoniewski M, Smyth GK, Huber W, et
766 al. Count-based differential expression analysis of RNA sequencing data using
767 R and Bioconductor. Nat Protoc. 2013;8:1765–86.
768

769 **TABLES**

770

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
CV	0.2	0.6	0.6	0.5	0.3	0.4	2.7

SD, standard deviation; CV, coefficient of variation

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.

771

772

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
CV	0.1	0.1	0.1	0.1	0.2	0.1

SD, standard deviation; CV, coefficient of variation

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.

773

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 coefficient of variation $\mathrm{CV}[W_h]$ and $\mathrm{CV}[W_{h'}]$ to be equal. As the CV decreases 0.6, 0.4 to 0.2, the correlation coefficient raises >0.5, >0.7 to >0.9, over most range of $\mathrm{Cor}[W_h, W_{h'}]$.

Figure 2

Sensitivity for detecting cell-type-specific association in simulated data for DNA methylation. Panels are aligned in rows according to the simulation settings with the methylation odds ratio of 1.3, 1.6 or 1.9. In each row, panels for different cell types are aligned in decreasing order of proportion. The vertical axis indicates sensitivity. In each panel, results from different algorithms are aligned horizontally in different colors. Results from 20 simulation trials are summarized in a box plot. 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. MethOR, methylation odds ratio; Neu, neutrophils; Mono, monocytes; Eos, eosinophils.

Figure 3

Specificity for detecting cell-type-specific association in simulated data for DNA methylation. The figure format is same as Fig. 3.

Figure 4

Precision (positive predictive value) for detecting cell-type-specific association in simulated data for DNA methylation. The figure format is same as Fig. 3.

801 **Figure 5**

802 Sensitivity for detecting cell-type-specific association in simulated data for
803 gene expression. Panels are aligned in rows according to the simulation
804 settings with the gene expression fold change of 1.7, 3.0 or 5.0. In each row,
805 panels for different cell types are aligned in decreasing order of proportion.
806 The vertical axis indicates sensitivity. In each panel, results from different
807 algorithms are aligned horizontally in different colors. Results from 50
808 simulation trials are summarized in a box plot. The middle bar of the box plot
809 indicates the median, and the lower and upper hinges correspond to the first
810 and third quartiles. The whiskers extend to the value no further than 1.5 ×
811 inter-quartile range from the hinges. FC, fold change; Gran, granulocytes;
812 Mono, monocytes.

813 **Figure 6**

814 Specificity for detecting cell-type-specific association in simulated data for
815 gene expression. The figure format is same as Fig. 5.

816 **Figure 7**

817 Precision (positive predictive value) for detecting cell-type-specific
818 association in simulated data for gene expression. The figure format is same
819 as Fig. 5.

820 **Figure 8**

821 Performance of cell-type-specific association prediction. For rheumatoid
822 arthritis association of DNA methylation in monocytes (A) and B cells (B);
823 Age association of gene expression in $CD4^+$ T cells (C) and monocytes (D).
824 The prediction is evaluated separately for up-regulated and down-regulated
825 markers. The AUC of ROC and its 95% confidence interval are plotted for each
826 statistical algorithm.
827

Figure 1

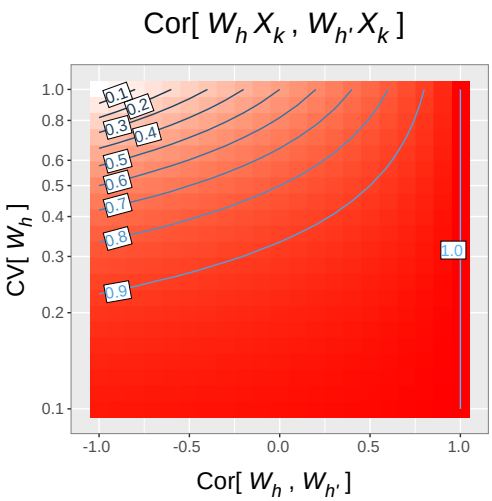

Cor[$W_h X_k$, $W_{h'} X_k$]
CV[W_h]
Cor[W_h , $W_{h'}$]
1.0
0.8
0.6
0.5
0.4
0.3
0.2
0.1
-1.0
-0.5
0.0
0.5
1.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0

Figure 2

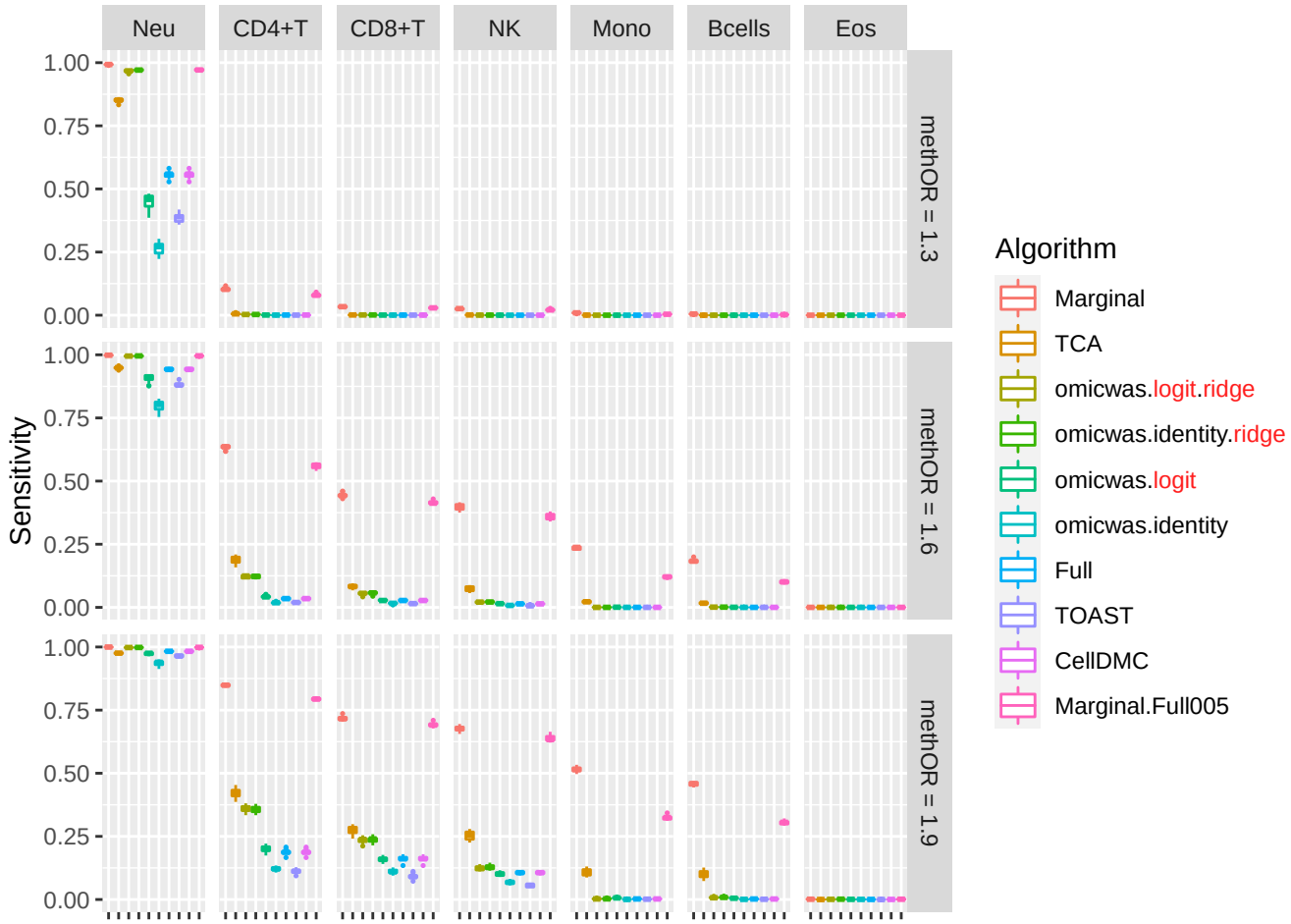

Neu
CD4+T
CD8+T
NK
Mono
Bcells
Eos
methOR = 1.3
methOR = 1.6
methOR = 1.9
Sensitivity
1.00
0.75
0.50
0.25
0.00
Algorithm
Marginal
TCA
omicwas.logit.ridge
omicwas.identity.ridge
omicwas.logit
omicwas.identity
Full
TOAST
CellDMC
Marginal.Full005

Figure 3

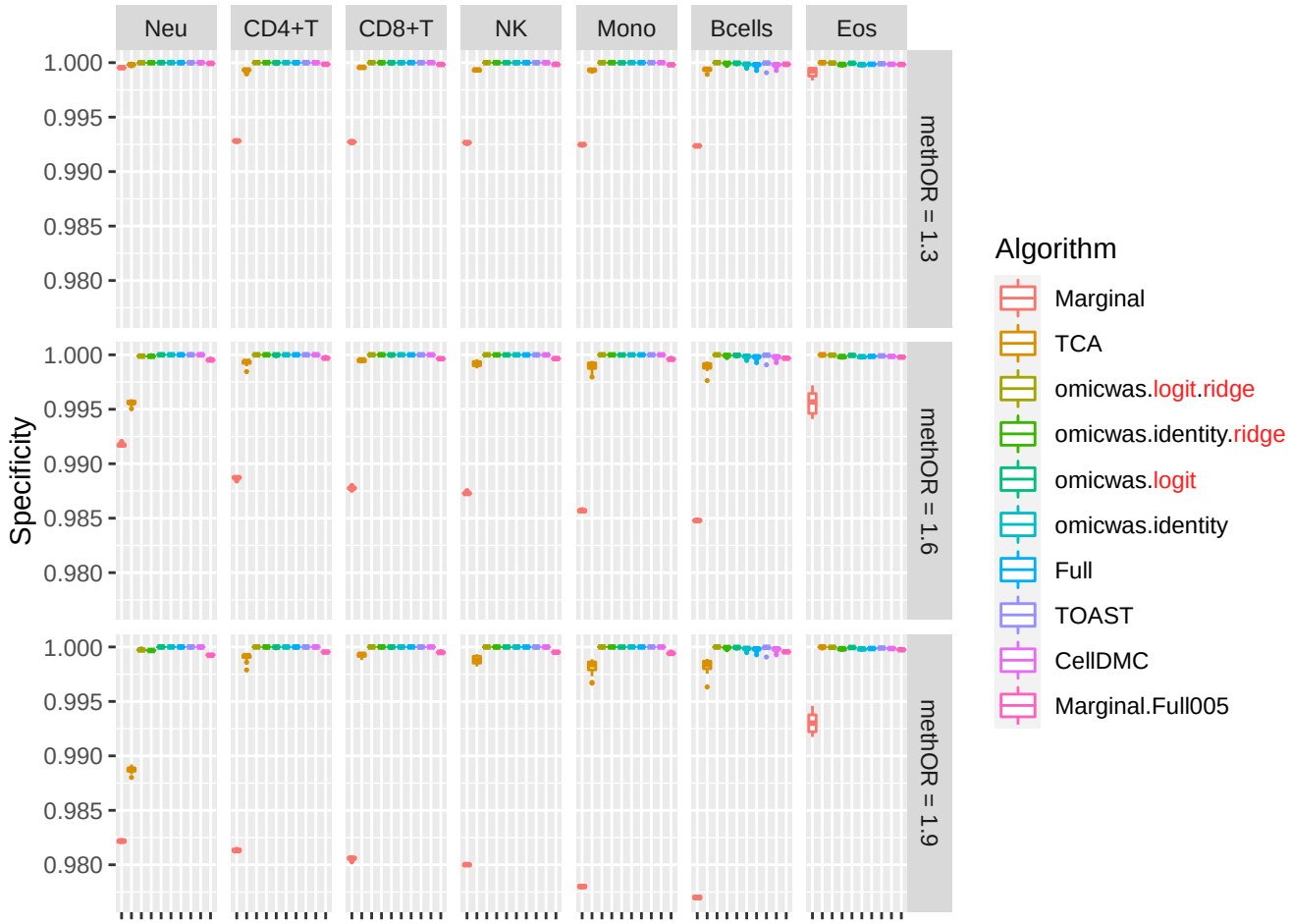

Neu
CD4+T
CD8+T
NK
Mono
Bcells
Eos
Specificity
1.000
0.995
0.990
0.985
0.980
methOR = 1.3
methOR = 1.6
methOR = 1.9
Algorithm
Marginal
TCA
omicwas.logit.ridge
omicwas.identity.ridge
omicwas.logit
omicwas.identity
Full
TOAST
CellDMC
Marginal.Full005

Figure 4

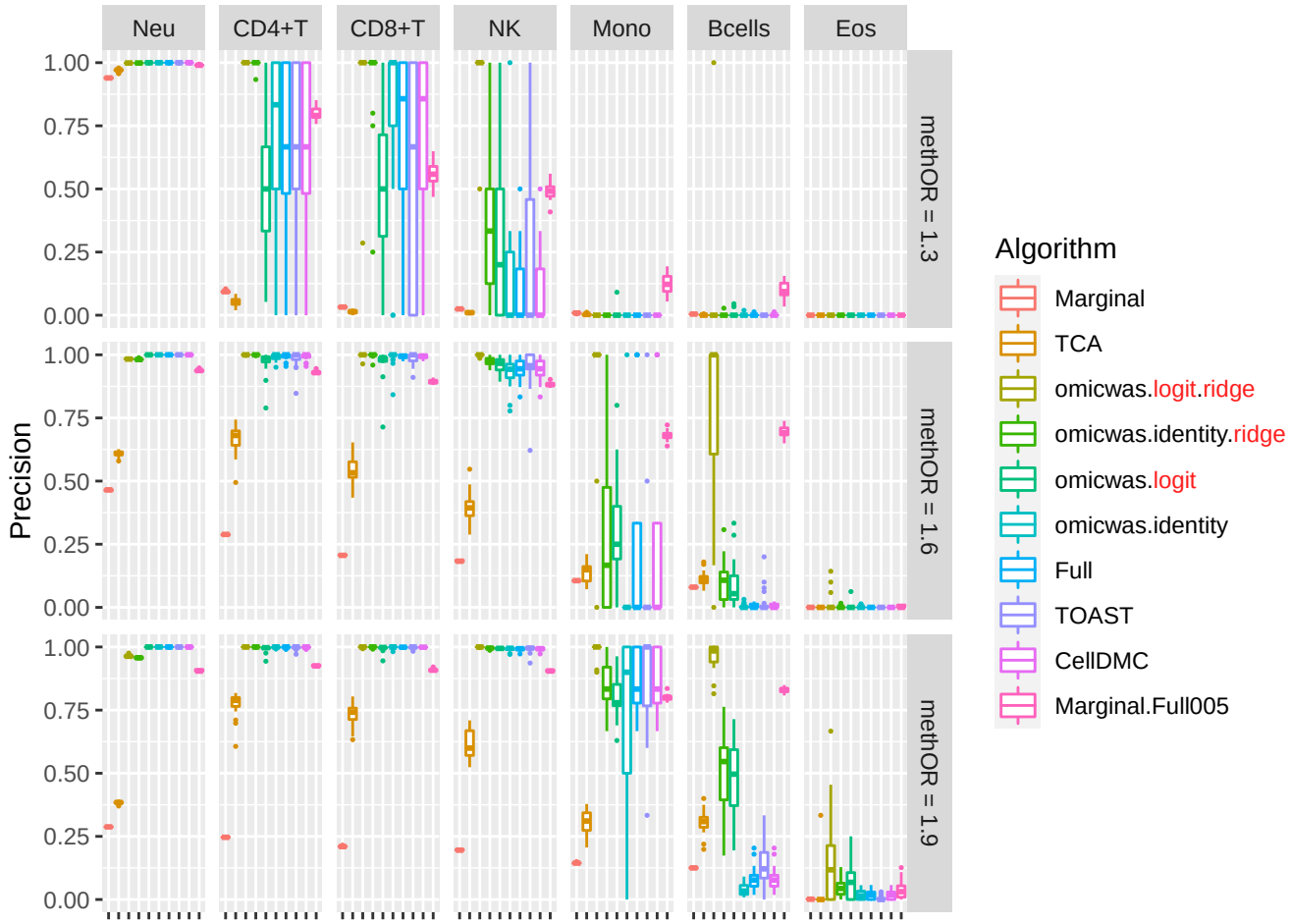

Neu
CD4+T
CD8+T
NK
Mono
Bcells
Eos
Precision
1.00
0.75
0.50
0.25
0.00
methOR = 1.3
methOR = 1.6
methOR = 1.9
Algorithm
Marginal
TCA
omicwas.logit.ridge
omicwas.identity.ridge
omicwas.logit
omicwas.identity
Full
TOAST
CellDMC
Marginal.Full005

Figure 5

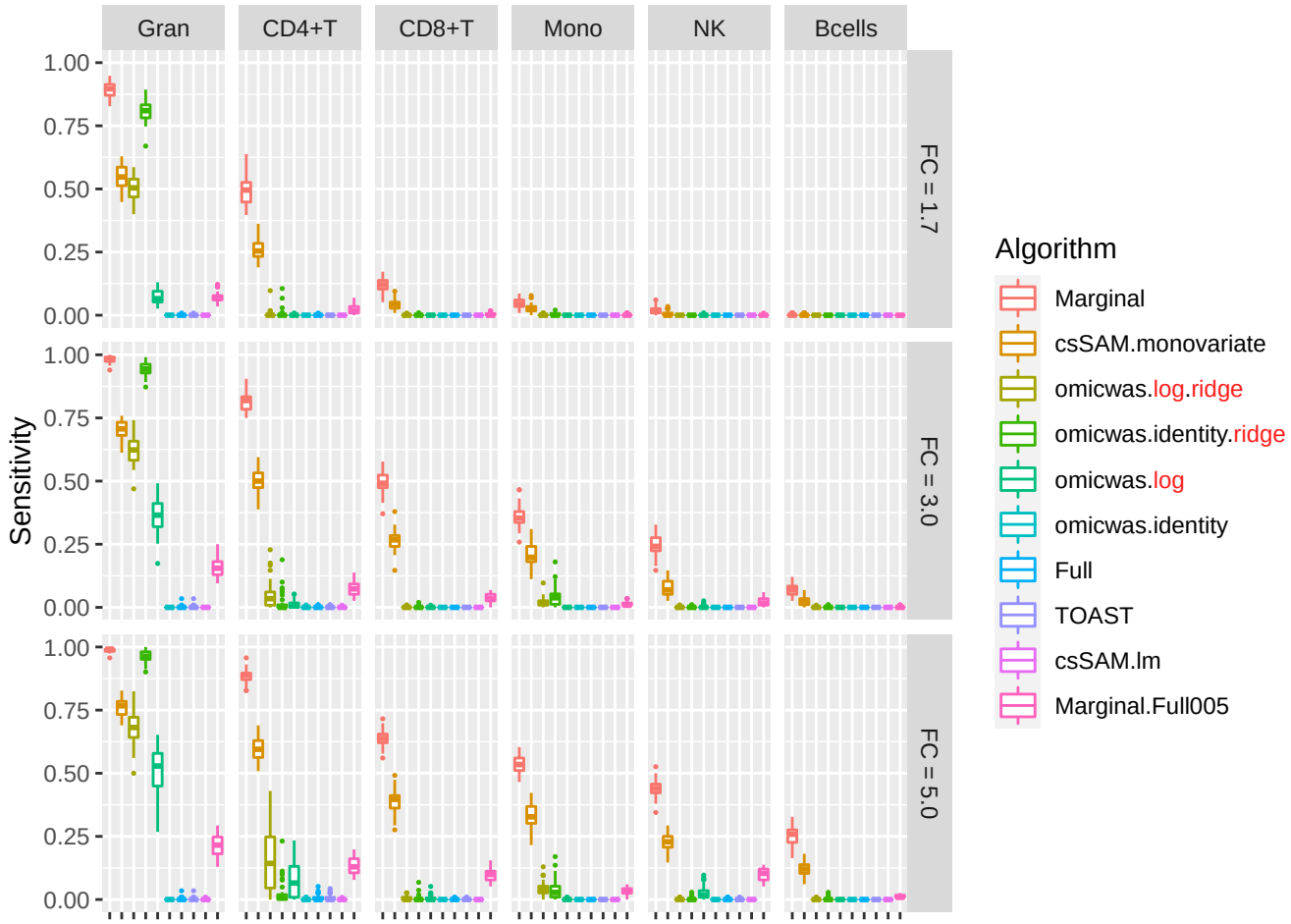

Gran
CD4+T
CD8+T
Mono
NK
Bcells
FC = 1.7
FC = 3.0
FC = 5.0
Sensitivity
1.00
0.75
0.50
0.25
0.00
Algorithm
Marginal
csSAM.monovariate
omicwas.log.ridge
omicwas.identity.ridge
omicwas.log
omicwas.identity
Full
TOAST
csSAM.lm
Marginal.Full005

Figure 6

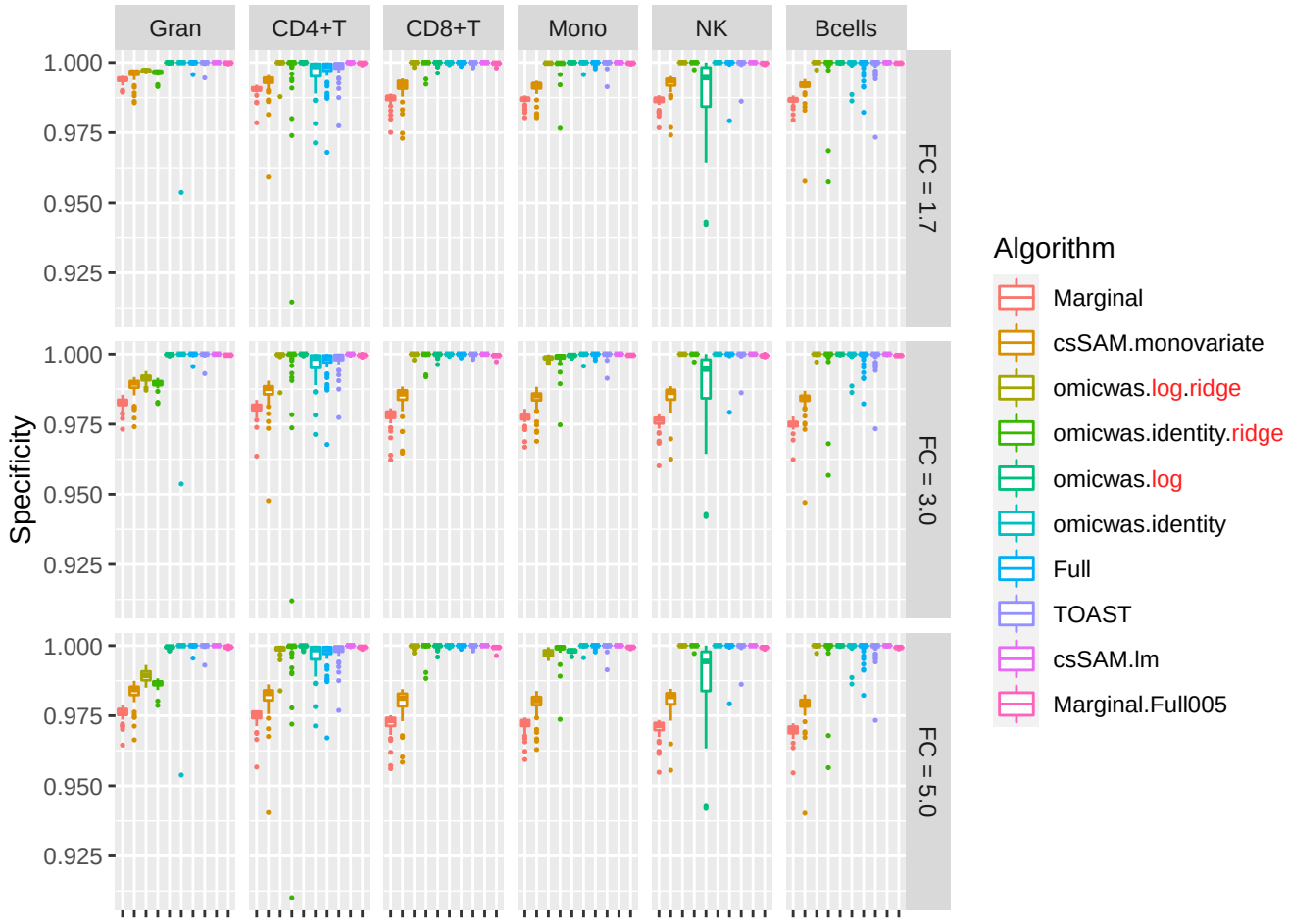

Gran
CD4+T
CD8+T
Mono
NK
Bcells
FC = 1.7
FC = 3.0
FC = 5.0
Specificity
1.000
0.975
0.950
0.925
Algorithm
Marginal
csSAM.monovariate
omicwas.log.ridge
omicwas.identity.ridge
omicwas.log
omicwas.identity
Full
TOAST
csSAM.lm
Marginal.Full005

Figure 7

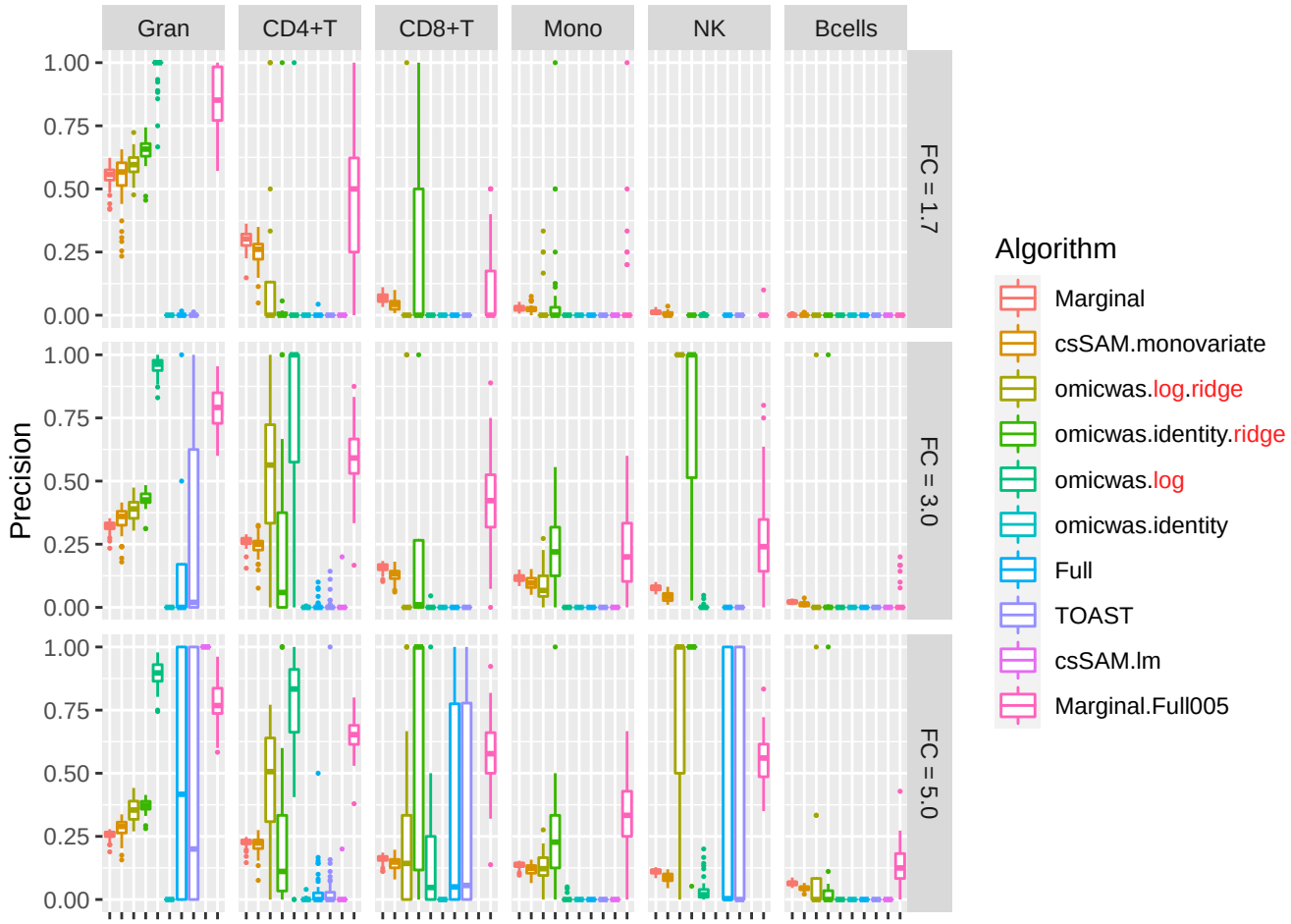
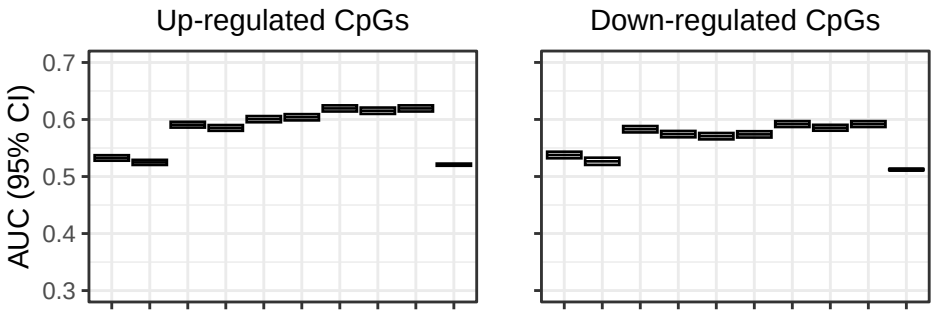
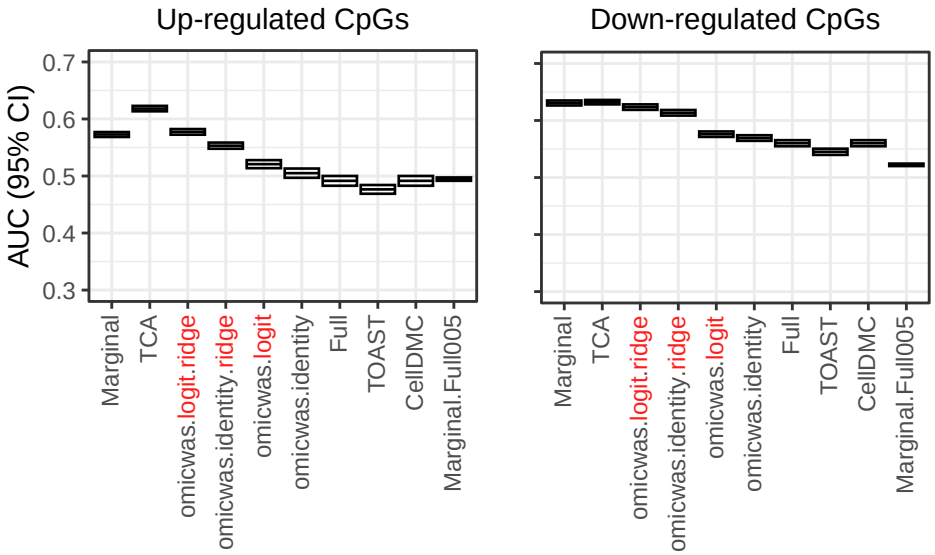

Gran
CD4+T
CD8+T
Mono
NK
Bcells
FC = 1.7
FC = 3.0
FC = 5.0
Precision
1.00
0.75
0.50
0.25
0.00
Algorithm
Marginal
csSAM.monovariate
omicwas.log.ridge
omicwas.identity.ridge
omicwas.log
omicwas.identity
Full
TOAST
csSAM.lm
Marginal.Full005

Figure 8

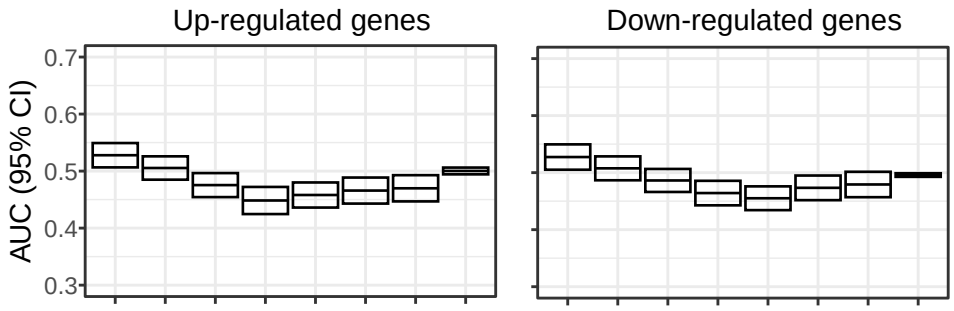
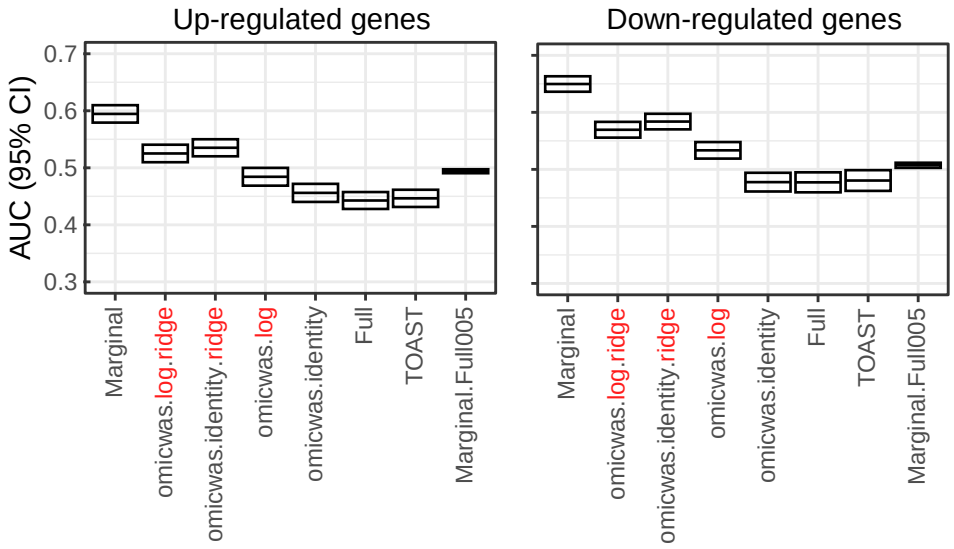
(A) Rheumatoid arthritis association in Monocytes



(B) Rheumatoid arthritis association in B cells



(C) Age association in CD4^{+} T cells



(D) Age association in Monocytes