Nonlinear ridge regression improves 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. Marginal model attained the lowest precision and highest sensitivity and was the only algorithm to detect weak signal in real data.

Conclusion: Nonlinear ridge regression performed cell-type-specific association test on bulk omics data with well-balanced performance. 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], although the linear scale could yield comparable performance under large sample size [18]. 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 [19]. 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 (19), 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

126 for trait effects, we applied nonlinear regression (equations (4) and (5)). To
127 cope with the multicollinearity of interaction terms, we applied ridge
128 regularization (formula (10)).

129 Previous regression type methods are based either on the full model of
130 linear regression (equation (2)) or the marginal model (equation (3)). The
131 full model fits and tests cell-type-specific effects for all cell types
132 simultaneously, and its variations include TOAST, csSAM.lm and CellDMC. The
133 marginal model fits and tests cell-type-specific effect for one cell type at a
134 time, and its variations include csSAM.monovariate and TCA. We also
135 examined a hybrid of the two models (Marginal.Full005), which becomes
136 positive when the models agree.

137 The simulation data was generated from real datasets of DNA methylation
138 (658 samples; 451,725 CpG sites) and gene expression (389 samples;
139 14,038 genes). The original cell type composition was retained for all samples,
140 and the case-control status was randomly assigned. Ninety-five percent of
141 omics markers were set to be unassociated with disease status, 2.5% were
142 up-regulated in cases at one cell type, and 2.5% were similarly down-
143 regulated. The cell-type-specific effect-size was fixed in a simulation trial,
144 either to methylation odds ratio (OR) of 1.3, 1.6 or 1.9 or to gene expression
145 fold change of 1.7, 3.0 or 5.0. The significance level was set to $P < 2.4 \times 10^{-}$
146 ^{7} for DNA methylation and false discovery rate <5% for gene expression. In
147 each simulation trial, the sensitivity, specificity and precision for detecting
148 cell-type-specific association was calculated for each cell type. To compare
149 algorithms, the performance measures for the same effect-size and cell type
150 were averaged over the simulation trials.

151 Overall, in the simulation for DNA methylation the sensitivity (Fig. 2) was
152 higher under large effect-size (bottom row of panels) and in abundant cell
153 types (left columns of panels). The average specificity (Fig. 3) was high
154 across the effect-size settings and across cell types, being >0.97 for the
155 Marginal model, >0.98 for TCA and >0.999 for other algorithms. The
156 precision (Fig. 4) was higher in abundant cell types within each effect-size
157 setting. As effect-size increased, the precision decreased in neutrophils but
158 increased in the other minor cell types. Excluding the cases where all

159 algorithms lacked sensitivity (monocytes and B cells under methOR=1.3 and
160 eosinophils), the average precision of omicwas.logit.ridge was >0.79 and was
161 the highest in 13/16 of the cases.

162 There was trade-off between sensitivity and precision. Among the
163 algorithms, the Marginal model attained the highest sensitivity and the lowest
164 precision. TCA, which is a variation of the marginal model, had relatively high
165 sensitivity and relatively low precision. Marginal.Full005 attained the second
166 highest sensitivity and moderate precision. The ridge regressions
167 (omicwas.logit.ridge and omicwas.identity.ridge) attained moderate
168 sensitivity and high precision. The full models without ridge regularization
169 (omicwas.logit, omicwas.identity, Full, TOAST and CellDMC) had the lowest
170 sensitivity.

171 The overall tendency was similar in the simulation for gene expression.
172 The sensitivity (Fig. 5) was higher under large effect-size and in abundant
173 cell types. The average specificity (Fig. 6) was high across the effect-size
174 settings and across cell types, being >0.96 for the marginal models (Marginal
175 and csSAM.monovariate), >0.98 for the nonlinear and ridge regressions
176 (omicwas.log.ridge, omicwas.identity.ridge, omicwas.log) and >0.996 for
177 other algorithms. The precision (Fig. 7) was higher in abundant cell types
178 within each effect-size setting. As effect-size increased, the precision
179 decreased in granulocytes but increased in the other minor cell types.
180 Excluding the full models (omicwas.identity, Full, TOAST, csSAM.lm) that
181 lacked sensitivity, the algorithms that were frequently top in average
182 precision were omicwas.identity.ridge (5 cases), omicwas.log (5 cases),
183 omicwas.log.ridge (3 cases) and Marginal.Full005 (3 cases).

184 There again was trade-off between sensitivity and precision. Among the
185 algorithms, the marginal models attained the highest sensitivity but relatively
186 low precision. The nonlinear and ridge regressions and Marginal.Full005
187 attained moderate sensitivity and highest precision. The full models had very
188 low average sensitivity of <0.01.

189 For gene expression, we also simulated a scenario where cell-type-specific
190 disease effect occurred in cell type "marker" genes (Additional files 2, 3, 4:
191 Figs. S1, S2, S3). In other words, the expression level in the target cell type

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

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

206 The detection of cell-type-specific association in bulk tissue was evaluated by
207 using physically sorted cells. In principle, sorted cells should serve as genuine
208 verification, however, due to the relatively small sample size (94 or 203 for
209 rheumatoid arthritis and 214 or 1202 for age) the available datasets were
210 underpowered to generate a gold standard list of differentially expressed
211 omics markers [20]. Instead, we generated a benchmark set of differentially
212 expressed markers by imposing a relaxed significance level of $P < 0.05$; the
213 set would be enriched for true differentially expressed markers yet also
214 include unassociated markers. The benchmark set was cross-checked with
215 the prediction by each algorithm; in the same manner as the simulation
216 analysis, we assessed the sensitivity, specificity and precision.

217 The cell-type-specific association of DNA methylation with rheumatoid
218 arthritis was predicted using bulk peripheral blood leukocyte data and was
219 evaluated in sorted monocytes and B cells (Fig. 8). The input bulk methylation
220 data was normalized by applying the logit-transformation for the
221 Marginal.logit algorithm, which otherwise was the same as Marginal. Although
222 the sensitivity was extremely low for all algorithms, it was positive in both
223 cell types for Marginal ($0.8–1.1 \times 10^{-4}$), Marginal.logit ($1.2–1.4 \times 10^{-4}$), TCA

224 (0.5–1.2 × 10^{-4}) and omicwas.logit (0.5 × 10^{-4}). The cell-type-specific
225 association of DNA methylation with age was predicted using the same bulk
226 dataset and was evaluated in sorted $CD4^+$T cells and monocytes (Fig. 8). The
227 Marginal and Marginal.logit models attained by far the highest sensitivity
228 (both 0.15–0.27) in both cell types, and moderate precision (0.59–0.68 and
229 0.60–0.68 respectively).
230 The cell-type-specific association of gene expression with age was
231 predicted using whole blood data and was evaluated in sorted $CD4^+$ T cells
232 and monocytes (Fig. 9). The input bulk gene expression data was normalized
233 by applying the log-transformation for the Marginal.log algorithm, which
234 otherwise was the same as Marginal. Although the sensitivity was low for all
235 algorithms, it was positive in both cell types for Marginal (0.02–0.07),
236 Marginal.log (0.07–0.11), omicwas.identity.ridge (0.01–0.22) and
237 omicwas.log (0.03–0.05). The precision was modest for Marginal (0.06–0.31),
238 Marginal.log (0.07–0.28), omicwas.identity.ridge (0.03–0.21) and
239 omicwas.log (0.04–0.17). The dataset of sorted $CD4^+$ T cells (214 samples)
240 is smaller than the monocyte dataset (1202 samples) thus could be
241 underpowered to pick enough true differentially expressed genes into the
242 benchmark set.
243 For DNA methylation dataset GSE42861 and for GTEx gene expression
244 dataset, the omicwas.logit.ridge and omicwas.log.ridge models of the
245 omicwas package was computed in 8.1 and 0.7 hours respectively, using 8
246 cores of a 2.5 GHz Xeon CPU Linux server.
247

248 **Discussion**

249 Aiming to elucidate cell-type-specific trait association in DNA methylation and
250 gene expression, this article explored two aspects, multicollinearity and scale.
251 We observed multicollinearity in real data and derived mathematically how it
252 emerges. To cope with the multicollinearity, we applied ridge regularization.
253 To properly handle multiple scales simultaneously, we applied nonlinear
254 regression. Among the examined algorithms, nonlinear ridge regression
255 attained moderate sensitivity and highest precision in simulated data. We also

256 developed an algorithm that combines full and marginal models, which
257 attained balanced sensitivity and precision in simulation. In real benchmark
258 data, all algorithms performed poorly yet the marginal models tended to
259 attain the highest sensitivity.

260 The statistical methods discussed in this article are applicable, in principle,
261 to any tissue. For validation of the methods, we need datasets for bulk tissue
262 as well as sorted cells, ideally of several hundred samples. Currently, the
263 publicly available data is limited to peripheral blood. By no means, the
264 rheumatoid arthritis EWAS datasets [21-23] or the datasets for age
265 association of gene expression [24,25] are representative. Nevertheless, we
266 think verification in real data is valuable.

267 By the performance in simulated and real data, we can roughly divide
268 algorithms into three groups: full (and its variations), marginal (and its
269 variations) and the third group that includes ridge regressions and the hybrid
270 Marginal.Full005. In marginal models, we test one cell type at a time. If we
271 knew in advance that one particular cell type is associated with the trait,
272 which would be a rare situation, testing that cell type with the marginal model
273 is the most simple and correct approach. However, when the test target cell
274 type is not associated, but instead another cell type is associated, the
275 marginal models can pick up false signals due to the collinearity between
276 regressor variables. Indeed, marginal models attained highest sensitivity
277 (Figs. 2, 5) and relatively low precision (Figs. 4, 7), which could lead to
278 unstable performance. The full model fits and tests all cell types
279 simultaneously, by which it adjusts for the effects of other cell types. Due to
280 the simultaneous inclusion of collinear predictors, the sensitivity was low (Figs.
281 2, 5). The ridge regressions (omicwas.identity.ridge, omicwas.logit.ridge and
282 omicwas.log.ridge) were in the middle between full and marginal models with
283 regards to the sensitivity (Figs. 2, 5), while attaining the highest specificity
284 (Figs. 4, 7). The hybrid Marginal.Full005 algorithm is intended to gain
285 sensitivity by the marginal model while keeping precision by incorporating
286 the full model. It attained moderate sensitivity (Figs. 2, 5) and moderate
287 precision (Figs. 4, 7) in simulation. In real data, all algorithms performed
288 poorly yet Marginal, Marginal.logit and Marginal.log tended to attain the

highest sensitivity. With regards to the performance measures of all algorithms, the association of DNA methylation with age (Fig. 8) was roughly similar to the simulation setting of methylation OR = 1.6 for B cells (Figs. 2–4), and the association of gene expression with age (Fig. 9) was roughly similar to the simulation setting of fold change = 1.7 for $CD8^+T$ cells (Figs. 5–7). For the respective simulation settings, the median coefficient of determination for the Marginal model was 0.020 and 0.007, indicating weak association.

A limitation of our simulation is that only one cell type was assumed to be associated with disease status at each marker. In reality, two or more cell types can be associated with disease under homogeneous or heterogeneous effect. In the physically sorted cells, the association of DNA methylation with rheumatoid arthritis tended to be consistent between monocytes and B cells; the association statistics across CpG sites were positively correlated with Spearman's rank correlation coefficient of 0.20 (P-value $< 2.2 \times 10^{-16}$). Similarly, the association with age tended to be consistent between $CD4^+T$ cells and monocytes with correlation coefficient of 0.27 and 0.07 (P-value $< 2.2 \times 10^{-16}$), respectively, for DNA methylation and gene expression. The consistency suggests that multiple cell types tend to be associated under homogeneous effect. If the association is completely consistent, the effect-size is uniform across cell types. As there is no cell-type-specific effect, a simple regression by disease (or relevant trait), ignoring the cell type composition, becomes the appropriate modeling (formula (8)). Moreover, when cell type composition has low CV (as observed in Tables 1 and 2), the marginal model with normalized input (formula (9)) becomes almost identical to the simple regression. In other words, the marginal model can pick up signal in cases where effect-size is homogeneous across cell types. Correspondingly, in real data of DNA methylation Marginal.logit performed the best and was slightly better than Marginal (Fig. 8), and in gene expression Marginal.log performed mostly the best and was better than Marginal (Fig. 9).

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. Although marginal models can be powerful when the tested cell type actually is the only one associated with the trait, caution is needed in 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 (15)). The number of cell types associated with disease at each marker was restricted to one in our simulation but could be two or more with homogeneous or heterogeneous effect. If the effect-size is uniform across all cell types, a simple regression by disease status is suitable, which can be substituted with the marginal model that takes normalized input. 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

352 equal. The matrix $Y_{i,j}$ represents the omics marker expression level in tissue
353 samples.

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

359 $$\alpha_h + \sum_k \beta_{h,k} X_{i,k}. \quad (1)$$

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

364 $$\mu_i = \sum_h \alpha_h W_{h,i} + \sum_{h,k} \beta_{h,k} W_{h,i} X_{i,k} + \sum_l \gamma_l C_{i,l}.$$

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

368 $$Y_i = \mu_i + \varepsilon_i,$$
369 $$\varepsilon_i \sim N(0, \sigma^2).$$

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

372 $$Y_i = \sum_h \alpha_h W_{h,i} + \sum_{h,k} \beta_{h,k} W_{h,i} X_{i,k} + \sum_l \gamma_l C_{i,l} + \varepsilon_i, \quad (2)$$

373 or its variations [5,9,13]. Alternatively, the cell-type-specific effects of traits
374 can be fitted and tested for one cell type *h* at a time by the *marginal* model,

375 $$Y_i = \sum_{h'} \alpha_{h'} W_{h',i} + \sum_k \beta_{h,k} W_{h,i} X_{i,k} + \sum_l \gamma_l C_{i,l} + \varepsilon_i, \quad (3)$$

376 or its variations [7,8,10,11,14].

377 **Nonlinear regression**

378 Aiming to simultaneously analyze cell type composition in linear scale and
379 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 [[26], section 6.3.1].

In the special case where the marker expression is homogeneous across cell types, the formulae become simple. Suppose that α_h regardless of cell type *h* equals α and that $\beta_{h,k}$ equals β_k. The regression formulae (4) and (5) of sample *i* reduces to

$$f(Y_i) = \alpha + \sum_k \beta_k X_{i,k} + \sum_l \gamma_l C_{i,l} + \varepsilon_i. \quad (8)$$

On the other hand, the marginal model for cell type *h* in formula (3) reduces to

$$Y_i = \alpha + \sum_k \beta_k W_{h,i} X_{i,k} + \sum_l \gamma_l C_{i,l} + \varepsilon_i.$$

Moreover, when the CV for cell type composition is low, the cell type proportion $W_{h,i}$ of sample *i* approximately equals the average $\overline{W_h}$ taken over samples. Thus, the formula reduces further to

$$Y_i = \alpha + \overline{W_h} \cdot \sum_k \beta_k X_{i,k} + \sum_l \gamma_l C_{i,l} + \varepsilon_i.$$

If we replace the input bulk expression level Y_i with the normalized value $f(Y_i)$, the model becomes

$$f(Y_i) = \alpha + \overline{W_h} \cdot \sum_k \beta_k X_{i,k} + \sum_l \gamma_l C_{i,l} + \varepsilon_i. \quad (9)$$

Under the special case of cell-type-homogeneous expression and low-CV cell type composition, formula (8) for nonlinear regression and formula (9) for the marginal model with normalized input become almost identical. The difference is the multiplication by constant $\overline{W_h}$, which does not change the test statistics for β_k.

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

$$\mathrm{RSS} + \lambda \sum_{h,k} {\beta_{h,k}}^2, \quad (10)$$

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

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

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

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

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 [27] as

$$\mathrm{Mean}\left[\widehat{\boldsymbol{\theta}}(\lambda)\right] = Q^*(\lambda)^{-1}\, Q^*(0)\, \boldsymbol{\theta}, \qquad (13)$$

$$\mathrm{Var}\left[\widehat{\boldsymbol{\theta}}(\lambda)\right] = \sigma^2\, Q^*(\lambda)^{-1} \left(\frac{\partial \boldsymbol{\mu}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)^T \left(\frac{\partial \boldsymbol{\mu}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right) Q^*(\lambda)^{-1}, \qquad (14)$$

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

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 [28]. 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 (13) and (14), the ridge estimator $\widehat{\boldsymbol{\beta}}(\lambda)$ is asymptotically normally distributed with

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

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

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

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

$$\begin{aligned}
\mathrm{Bias}\left[\widehat{\boldsymbol{\beta}}(\lambda)\right] &= \mathrm{E}\left[\widehat{\boldsymbol{\beta}}(\lambda) - \boldsymbol{\beta}\right] \\
&= -\lambda\left[\left(\frac{\partial \boldsymbol{\mu}(\boldsymbol{\theta})}{\partial \boldsymbol{\beta}}\right)^T\left(\frac{\partial \boldsymbol{\mu}(\boldsymbol{\theta})}{\partial \boldsymbol{\beta}}\right) + \lambda I\right]^{-1}\boldsymbol{\beta} \\
&= \left\{\sum_{m=1}^{M} \boldsymbol{v}_m \frac{-\lambda}{{d_m}^2 + \lambda}{\boldsymbol{v}_m}^T\right\}\boldsymbol{\beta}, \\
\mathrm{Var}\left[\widehat{\boldsymbol{\beta}}(\lambda)\right] &= \sigma^2 \sum_{m=1}^{M} \boldsymbol{v}_m \frac{{d_m}^2}{\left({d_m}^2 + \lambda\right)^2}{\boldsymbol{v}_m}^T, \\
\mathrm{MSE}\left[\widehat{\boldsymbol{\beta}}(\lambda)\right] &= \mathrm{E}\left[\left\|\widehat{\boldsymbol{\beta}}(\lambda) - \boldsymbol{\beta}\right\|^2\right] \\
&= \left\|\mathrm{Bias}\left[\widehat{\boldsymbol{\beta}}(\lambda)\right]\right\|^2 + \mathrm{tr}\left(\mathrm{Var}\left[\widehat{\boldsymbol{\beta}}(\lambda)\right]\right) \\
&= \sum_{m=1}^{M}\left(\frac{\lambda}{{d_m}^2 + \lambda}\right)^2 ({\boldsymbol{v}_m}^T\boldsymbol{\beta})^2 + \left(\frac{{d_m}^2}{{d_m}^2 + \lambda}\right)^2\left(\frac{\sigma^2}{{d_m}^2}\right). \quad (15)
\end{aligned}$$

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. [29] 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,

$$\kappa = \frac{1}{\sum_{m=1}^{M}{d_m}^2}\sum_{m=1}^{M}\left\{\frac{{d_m}^2\left({\boldsymbol{v}_m}^T\widehat{\boldsymbol{\beta}}(0)\right)^2}{\sigma^2} - 1\right\}. \quad (16)$$

The weighting and subtraction were mentioned in [28], 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

$$\lambda = \begin{cases} 1/\kappa & \text{if } \kappa > 0, \\ {d_1}^2 & \text{otherwise,} \end{cases} \quad (17)$$

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 (16) and (17).
6. Compute ridge estimators $\widehat{\boldsymbol{\beta}}(\lambda)$ and $\widehat{\boldsymbol{\gamma}}(\lambda)$ by minimizing formula (10) 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 (12), 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 [30]. For the *s*-th coordinate,

$$\frac{\begin{pmatrix}\widehat{\boldsymbol{\beta}}(\lambda)\\ \widehat{\boldsymbol{\gamma}}(\lambda)\end{pmatrix}_s}{\sqrt{\mathrm{Var}\left[\begin{pmatrix}\widehat{\boldsymbol{\beta}}(\lambda)\\ \widehat{\boldsymbol{\gamma}}(\lambda)\end{pmatrix}\right]_{s,s}}} \sim t_{n-p}, \quad (18)$$

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

The formula (18) 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 [31] 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 $\boldsymbol{\alpha} = \widehat{\boldsymbol{\alpha}}(0)$, $\boldsymbol{\gamma} = \widehat{\boldsymbol{\gamma}}(0)$.

536 **Multicollinearity of interaction terms**

537 The regressors for cell-type-specific trait effects in the full model (equation
538 (2)) are the interaction terms $W_{h,i}X_{i,k}$. To assess multicollinearity, we
539 mathematically derive the correlation coefficient between two interaction
540 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
541 sampled instances of random variables W_h, $W_{h'}$ and X_k, respectively; note
542 that the sample index *i* is omitted. For simplicity, we assume W_h and $W_{h'}$ are
543 independent of X_k. Let E[•], Var[•], Cov[•], Cor[•] and CV[•] denote the
544 expectation, variance, covariance, correlation and coefficient of variation,
545 respectively. Since X_k is centered, $\mathrm{E}[W_h X_k] = \mathrm{E}[W_{h'} X_k] = 0$. The correlation
546 coefficient between interaction terms becomes

547 $$\mathrm{Cor}[W_h X_k, W_{h'} X_k] = \frac{\mathrm{E}[W_h X_k W_{h'} X_k]}{\sqrt{\mathrm{E}\left[{W_h}^2 {X_k}^2\right]}\sqrt{\mathrm{E}\left[{W_{h'}}^2 {X_k}^2\right]}}$$

548 $$= \frac{\mathrm{E}[W_h W_{h'}]}{\sqrt{\mathrm{E}\left[{W_h}^2\right]}\sqrt{\mathrm{E}\left[{W_{h'}}^2\right]}}$$

549 $$= \frac{\mathrm{Cov}[W_h, W_{h'}] + \mathrm{E}[W_h]\,\mathrm{E}[W_{h'}]}{\sqrt{\mathrm{Var}[W_h] + \mathrm{E}[W_h]^2}\sqrt{\mathrm{Var}[W_{h'}] + \mathrm{E}[W_{h'}]^2}}$$

550 $$= \frac{\dfrac{\mathrm{Cor}[W_h, W_{h'}]\sqrt{\mathrm{Var}[W_h]}\sqrt{\mathrm{Var}[W_{h'}]}}{\mathrm{E}[W_h]\mathrm{E}[W_{h'}]} + 1}{\sqrt{\dfrac{\mathrm{Var}[W_h]}{\mathrm{E}[W_h]^2} + 1}\sqrt{\dfrac{\mathrm{Var}[W_{h'}]}{\mathrm{E}[W_{h'}]^2} + 1}}$$

551 $$= \frac{\mathrm{Cor}[W_h, W_{h'}]\,\mathrm{CV}[W_h]\,\mathrm{CV}[W_{h'}] + 1}{\sqrt{\mathrm{CV}[W_h]^2 + 1}\sqrt{\mathrm{CV}[W_{h'}]^2 + 1}}. \quad (19)$$

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

554 **EWAS of rheumatoid arthritis and age**

555 EWAS datasets for rheumatoid arthritis were downloaded from the Gene
556 Expression Omnibus. Using the RnBeads package (version 2.2.0) [32] of R,
557 IDAT files of HumanMethylation450 array were preprocessed by removing

558 low quality samples and markers, by normalizing methylation level, and by
559 removing markers on sex chromosomes and outlier samples. The association
560 of methylation level with disease status was tested with adjustment for sex,
561 age, smoking status and experiment batch; the covariates were assumed to
562 have uniform effects across cell types. Alternatively, the association of
563 methylation level with age was tested with adjustment for disease status ,
564 sex, smoking status and experiment batch. After quality control, dataset
565 GSE42861 included bulk peripheral blood leukocyte data for 336 cases and
566 322 controls [22].

567 The cell type composition of bulk samples was imputed using the
568 Houseman algorithm [33] in the GLINT software (version 1.0.4) [34]. The
569 reference data of GLINT software characterizes seven cell types [35] by 300
570 CpG sites [36], of which 284 were measured in our data. We used prediction
571 results for the seven cell types (Table 1).

572 Dataset GSE131989 included sorted $CD14^+$ monocyte data for 63 cases
573 and 31 controls [23]. By meta-analysis of GSE131989 and GSE87095 [21],
574 we obtained sorted $CD19^+$ B cell data for 108 cases and 95 controls. Under
575 the nominal significance level $P < 0.05$ (two-sided), the number of CpG sites
576 up- or down-regulated in cases were 20,869 (5%) and 14,911 (3%),
577 respectively, in $CD14^+$ monocyte and 28,004 (6%) and 26,582 (6%) in $CD19^+$
578 B cell.

579 From the Gene Expression Omnibus dataset GSE56047 [25], we obtained
580 sorted $CD14^+$ monocyte data for 1200 samples and sorted $CD4^+$ T cell data
581 for 214 samples. Under the nominal significance level $P < 0.05$ (two-sided),
582 the number of CpG sites up- or down-regulated by higher age were 45,283
583 (10%) and 80,871 (18%), respectively, in $CD14^+$ monocyte and 35,822 (8%)
584 and 25,020 (5%) in $CD4^+$ T cell.

585 **Differential gene expression by age**

586 Whole blood RNA-seq data of GTEx v7 was downloaded from the GTEx
587 website [24]. Genes of low quality or on sex chromosomes were removed,
588 expression level was normalized, outlier samples were removed, and 389

589 samples were retained. The association of read count with age was tested
590 with adjustment for sex.
591 The cell type composition of bulk samples was imputed using the
592 DeconCell package (version 0.1.0) [10] of R. The reference data of DeconCell
593 characterizes 33 cell types by two to 217 signature genes. Our data measured
594 39% of the signature genes. We used prediction results for six main cell types
595 (Table 2), for which the prediction performance was 44.4 to 90.9.
596 From the Gene Expression Omnibus dataset GSE56047 [25], we obtained
597 sorted $CD14^+$ monocyte data for 1202 samples and sorted $CD4^+$ T cell data
598 for 214 samples. Under the nominal significance level $P < 0.05$ (two-sided),
599 the number of genes up- or down-regulated by higher age were 2715 (11%)
600 and 3240 (13%), respectively, in $CD14^+$ monocyte and 1082 (4%) and 1246
601 (5%) in $CD4^+$ T cell.

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

603 Bulk tissue sample data for case-control comparison were simulated based
604 on the above-mentioned EWAS dataset GSE42861 and GTEx gene expression
605 dataset. We randomly assigned the case-control status to the samples.
606 Among the omics markers, 2.5% were set to be up-regulated in cases in
607 single cell type, 2.5% were similarly down-regulated, and 95% were
608 unrelated to case-control status. The cell-type-specific effect-size of the
609 differentially expressed markers was fixed within a simulation trial, and was
610 chosen from methylation OR of 1.3, 1.6 or 1.9 for EWAS [20] and fold-change
611 of 1.7, 3.0 or 5.0 for gene expression analysis; the effect-sizes correspond to
612 log(1.3), log(1.6) and log(1.9) or log(1.7), log(3.0) and log(5.0) in
613 normalized scale. If the mean methylation level of a CpG site in cases and
614 controls are μ_{case} and μ_{control}, respectively, the methylation odds become
615 $\mu_{\text{case}}/(1-\mu_{\text{case}})$ and $\mu_{\text{control}}/(1-\mu_{\text{control}})$. The methylation OR represents the
616 case-control contrast of methylation level by the ratio of odds,
617 $\{\mu_{\text{case}}/(1-\mu_{\text{case}})\}/\{\mu_{\text{control}}/(1-\mu_{\text{control}})\}$ (see [20]).
618 For each effect-size, we performed 50 simulation trials. In each simulation
619 trial, we randomly assigned half of the samples as cases ($X_{i,k} = 0.5$) and the

other half as controls ($X_{i,k} = -0.5$). We retained the covariates matrix $C_{i,l}$ and the cell type composition matrix $W_{h,i}$ from the original data. From the original bulk expression level matrix $Y_{i,j}$, 95% of the markers were randomly chosen and retained; these markers had no association with disease because the case-control status was randomized. Cell-type-specific association was introduced into the remaining 5% of markers, such that an equal number of markers were up- or down-regulated in each cell type. For example, in the EWAS dataset, $451{,}725 \times 0.05 \times 0.5 \div 7 = 1613$ CpG sites were up-regulated in neutrophils of cases.

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

In the truly disease-associated cell type *h*, we introduced signal β and noise σ^2. The signal level was fixed in a simulation trial, for example to methylation OR = 1.3. Since the noise level was taken from real data, the level varied between markers. In the process of obtaining bulk expression the expression of all cell types was mixed, which dilutes the signal. The signal dilution becomes stronger if *h* is a minor cell type. The mixing process adds noise from other cell types, which becomes stronger if *h* is a minor cell type. Consequently, minor cell types tend to manifest weaker association in bulk tissue. We empirically measured the strength of association by the coefficient of determination, R^2, for the marginal model. The coefficient of determination is defined as the proportion of variance explained by the model, and $R^2/(1 - R^2)$ equals the signal-to-noise ratio. Under methylation OR of 1.3, 1.6, 1.9 for EWAS simulation, the median R^2 was 0.322, 0.589, 0.712 for

653 neutrophils, 0.010, 0.033, 0.057 for NK cells, and 0.001, 0.003, 0.005 for
654 eosinophils. Under fold-change of 1.7, 3.0 or 5.0 for gene expression
655 simulation, the median R^2 was 0.135, 0.331, 0.434 for granulocytes, 0.007,
656 0.026, 0.049 for $CD8^+$ T cells, and 0.001, 0.003, 0.007 for B cells.
657 For gene expression, we also simulated a scenario where cell-type-specific
658 disease effect occurs in cell type marker genes. The simulation procedure is
659 same as above except that the expression level was set to zero (in linear
660 scale) in all cell types other than the target cell type *h*, for both cases and
661 controls.

662 **Evaluation of statistical methods**

663 Cell-type-specific effects of traits was statistically tested by using bulk tissue
664 data as input. We applied the omicwas package with the normalizing function
665 *f* = log, logit, identity without ridge regularization (omicwas.log,
666 omicwas.logit, omicwas.identity) or under ridge regression
667 (omicwas.log.ridge, omicwas.logit.ridge, omicwas.identity.ridge). The
668 omicwas package was used also for conventional linear regression under the
669 full and marginal models. We also developed a hybrid of marginal and full
670 models (Marginal.Full005): if the effect direction agreed in two models and if
671 $P < 0.05$ in the full model, we adopted the Z-score of the marginal model;
672 otherwise, the Z-score was set to zero.
673 Among previous methods, we evaluated those that accept cell type
674 composition as input and compute test statistics for cell-type-specific
675 association. For DNA methylation data, we applied TOAST (version 1.2.0) [9],
676 CellDMC (version 2.0.2) [13] and TCA (version 1.0.0) [14]. For gene
677 expression data, we applied TOAST and csSAM (version 1.4) [5]. For csSAM,
678 we either fitted all cell types together or one cell type at a time, and denoted
679 the results as csSAM.lm and csSAM.monovariate, respectively. The csSAM
680 method is applicable to binomial traits but not to quantitative traits.
681 For simulated data of EWAS dataset GSE42861, we adopted the
682 significance level $P < 2.4 \times 10^{-7}$, which accounts for the correlation among
683 the probes on HumanMethylation450 array [37]. 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 [38].

The performance of an algorithm for the simulated data was assessed by sensitivity, specificity and precision. The performance measures were obtained from each simulation trial. For a target cell type *h*, we counted the four possible outcomes, true positives (TP_h), true negatives (TN_h), false positives (FP_h) and false negatives (FN_h). The sum $\mathrm{TP}_h + \mathrm{TN}_h + \mathrm{FP}_h + \mathrm{FN}_h$ equals the total number of omics markers (which was 451,725 CpG sites for DNA methylation and 14,038 genes for gene expression). For 5% of the markers, one randomly selected cell type h^* was set to be truly associated with disease status at data generation. The remaining 95% of the markers were null cases with no truly associated cell types. The outcome counts can be subtotaled according to the truly associated cell type, which is denoted in superscript,

$$\mathrm{TP}_h = \mathrm{TP}_h^{h^*=h},$$

$$\mathrm{TN}_h = \sum\nolimits_{h^* \neq h} \mathrm{TN}_h^{h^*} + \mathrm{TN}_h^{\mathrm{Null}},$$

$$\mathrm{FP}_h = \sum\nolimits_{h^*} \mathrm{FP}_h^{h^*} + \mathrm{FP}_h^{\mathrm{Null}},$$

$$\mathrm{FN}_h = \mathrm{FN}_h^{h^*=h}.$$

Remark that FP_h can occur when in cell type *h* a marker is truly up-regulated in disease cases but an algorithm predicts the marker to be down-regulated in *h*. The performance measures can be represented as

$$\mathrm{sensitivity}_h = \frac{\mathrm{TP}_h}{\mathrm{TP}_h + \mathrm{FN}_h} = \frac{\mathrm{TP}_h^{h^*=h}}{\mathrm{TP}_h^{h^*=h} + \mathrm{FN}_h^{h^*=h}},$$

$$\mathrm{specificity}_h = \frac{\mathrm{TN}_h}{\mathrm{TN}_h + \mathrm{FP}_h} = \frac{\sum_{h^* \neq h} \mathrm{TN}_h^{h^*} + \mathrm{TN}_h^{\mathrm{Null}}}{\mathrm{FP}_h^{h^*=h} + \sum_{h^* \neq h}\left(\mathrm{TN}_h^{h^*} + \mathrm{FP}_h^{h^*}\right) + \left(\mathrm{TN}_h^{\mathrm{Null}} + \mathrm{FP}_h^{\mathrm{Null}}\right)},$$

$$\mathrm{precision}_h = \frac{\mathrm{TP}_h}{\mathrm{TP}_h + \mathrm{FP}_h} = \frac{\mathrm{TP}_h^{h^*=h}}{\left(\mathrm{TP}_h^{h^*=h} + \mathrm{FP}_h^{h^*=h}\right) + \sum_{h^* \neq h} \mathrm{FP}_h^{h^*} + \mathrm{FP}_h^{\mathrm{Null}}}.$$

Whereas sensitivity is obtained solely from markers that are truly associated in the target cell type *h*, the specificity and precision are obtained by aggregating with the markers associated in other cell types and the null markers.

712 For the association with rheumatoid arthritis and age, "true" association
713 was determined from the measurements in physically sorted blood cells,
714 under the nominal significance level $P < 0.05$ (two-sided). In the same
715 manner as the simulation analysis, we assessed the sensitivity, specificity
716 and precision.
717

718 **Supplementary information**

719 **Additional file 1: Table S1.** Blood cell type proportion in Tsimane
720 Amerindians, Caucasians and Hispanics.
721 **Additional file 2: Fig. S1.** Sensitivity for detecting cell-type-specific
722 association in simulated data for gene expression of marker genes.
723 **Additional file 3: Fig. S2.** Specificity for detecting cell-type-specific
724 association in simulated data for gene expression of marker genes.
725 **Additional file 4: Fig. S3.** Precision (positive predictive value) for detecting
726 cell-type-specific association in simulated data for gene expression of marker
727 genes.
728 **Additional file 5: Supplementary note.** Asymptotic distribution of ridge
729 estimator.

730 **Abbreviations**

731 CV: coefficient of variation, eQTL: expression QTL, EWAS: epigenome-wide
732 association study, mQTL: methylation QTL, MSE: mean squared error, OLS:
733 ordinary least squares, OR: odds ratio, QTL: quantitative trait locus, SNP:
734 single nucleotide polymorphism

735 **Declarations**

736 **Ethics approval and consent to participate**

737 Not applicable.

738 **Consent for publication**

739 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.

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

9. Li Z, Wu Z, Jin P, Wu H. Dissecting differential signals in high-throughput data from complex tissues. Hancock J, editor. Bioinformatics.

787 2019;35:3898–905.

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

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

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

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

801 14. Rahmani E, Schweiger R, Rhead B, Criswell LA, Barcellos LF, Eskin E, et
802 al. Cell-type-specific resolution epigenetics without the need for cell sorting
803 or single-cell biology. Nature Communications. Nature Publishing Group;
804 2019;10:3417–11.

805 15. Cobos FA, Vandesompele J, Mestdagh P. Computational deconvolution
806 of transcriptomics data from mixed cell populations. Bioinformatics.
807 2018;34:1969–79.

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

810 17. Du P, Zhang X, Huang C-C, Jafari N, Kibbe WA, Hou L, et al.
811 Comparison of Beta-value and M-value methods for quantifying methylation
812 levels by microarray analysis. BMC Bioinformatics. BioMed Central;
813 2010;11:1–9.

814 18. Zhuang J, Widschwendter M, Teschendorff AE. A comparison of feature
815 selection and classification methods in DNA methylation studies using the
816 Illumina Infinium platform. BMC Bioinformatics. BioMed Central;
817 2012;13:1–14.

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

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

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

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

829 23. Rhead B, Holingue C, Cole M, Shao X, Quach HL, Quach D, et al.
830 Rheumatoid Arthritis Naive T Cells Share Hypermethylation Sites With
831 Synoviocytes. Arthritis & Rheumatology. John Wiley & Sons, Ltd;
832 2017;69:550–9.

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

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

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

840 27. Lim C. Robust ridge regression estimators for nonlinear models with

841 applications to high throughput screening assay data. Statist. Med.
842 2014;34:1185–98.

843 28. Lawless JF, Wang P. A simulation study of ridge and other regression
844 estimators. Communications in Statistics - Theory and Methods.
845 1976;5:307–23.

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

848 30. Halawa AM, Bassiouni El MY. Tests of regression coefficients under ridge
849 regression models. Journal of Statistical Computation and Simulation.
850 2000;65:341–56.

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

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

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

859 34. Rahmani E, Yedidim R, Shenhav L, Schweiger R, Weissbrod O, Zaitlen
860 N, et al. GLINT: a user-friendly toolset for the analysis of high-throughput
861 DNA-methylation array data. Hancock JM, editor. Bioinformatics.
862 2017;33:1870–2.

863 35. Reinius LE, Acevedo N, Joerink M, Pershagen G, Dahlén S-E, Greco D, et
864 al. Differential DNA Methylation in Purified Human Blood Cells: Implications
865 for Cell Lineage and Studies on Disease Susceptibility. Ting AH, editor. PLoS
866 ONE. Public Library of Science; 2012;7:e41361–13.

867 36. Koestler D. Improving Cell Mixture Deconvolution by Identifying Optimal

868 DNA methylation Libraries (IDOL). BMC Bioinformatics. BMC Bioinformatics;
869 2016;:1–21.

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

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

876

877 **TABLES**

878

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.

879

880

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.

881

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.

909 **Figure 5**

910 Sensitivity for detecting cell-type-specific association in simulated data for
911 gene expression. Panels are aligned in rows according to the simulation
912 settings with the gene expression fold change of 1.7, 3.0 or 5.0. In each row,
913 panels for different cell types are aligned in decreasing order of proportion.
914 The vertical axis indicates sensitivity. In each panel, results from different
915 algorithms are aligned horizontally in different colors. Results from 50
916 simulation trials are summarized in a box plot. The middle bar of the box plot
917 indicates the median, and the lower and upper hinges correspond to the first
918 and third quartiles. The whiskers extend to the value no further than 1.5 ×
919 inter-quartile range from the hinges. FC, fold change; Gran, granulocytes;
920 Mono, monocytes.

921 **Figure 6**

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

924 **Figure 7**

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

928 **Figure 8**

929 Performance of the predictions for cell-type-specific association of DNA
930 methylation. For the association with rheumatoid arthritis in monocytes and
931 B cells and the association with age in $CD4^+$ T cells and monocytes, sensitivity
932 (top), specificity (middle) and precision (bottom) are plotted. In each panel,
933 results from different algorithms are aligned horizontally in different colors.
934 Precision is not plotted when there were no positive CpG sites. RA,
935 rheumatoid arthritis; Mono, monocytes.

936 **Figure 9**

937 Performance of the predictions for cell-type-specific association of gene
938 expression. For the association with age in $CD4^+$ T cells and monocytes,
939 sensitivity (top), specificity (middle) and precision (bottom) are plotted. In
940 each panel, results from different algorithms are aligned horizontally in
941 different colors. Precision is not plotted when there were no positive genes.
942 Mono, monocytes.
943

Figure 1

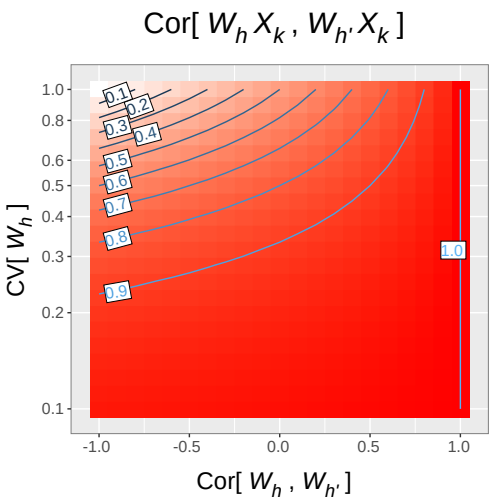

Cor[$W_h X_k$, $W_{h'} X_k$]
CV[W_h]
Cor[W_h , $W_{h'}$]
1.0
0.8
0.6
0.5
0.4
0.3
0.2
0.1
-1.0
-0.5
0.0
0.5
1.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0

Figure 2

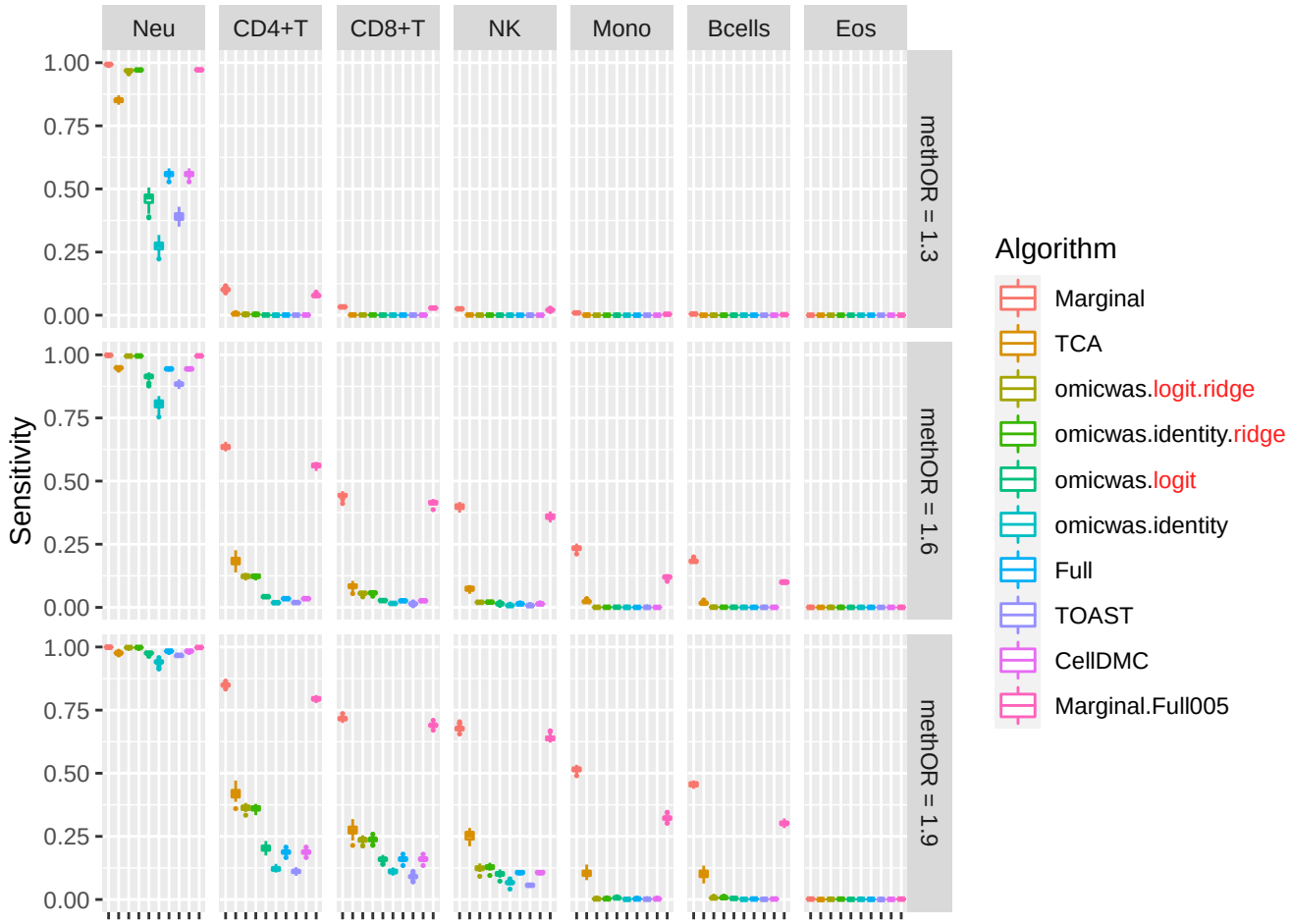

Neu
CD4+T
CD8+T
NK
Mono
Bcells
Eos
methOR = 1.3
methOR = 1.6
methOR = 1.9
Sensitivity
1.00
0.75
0.50
0.25
0.00
Algorithm
Marginal
TCA
omicwas.logit.ridge
omicwas.identity.ridge
omicwas.logit
omicwas.identity
Full
TOAST
CellDMC
Marginal.Full005

Figure 3

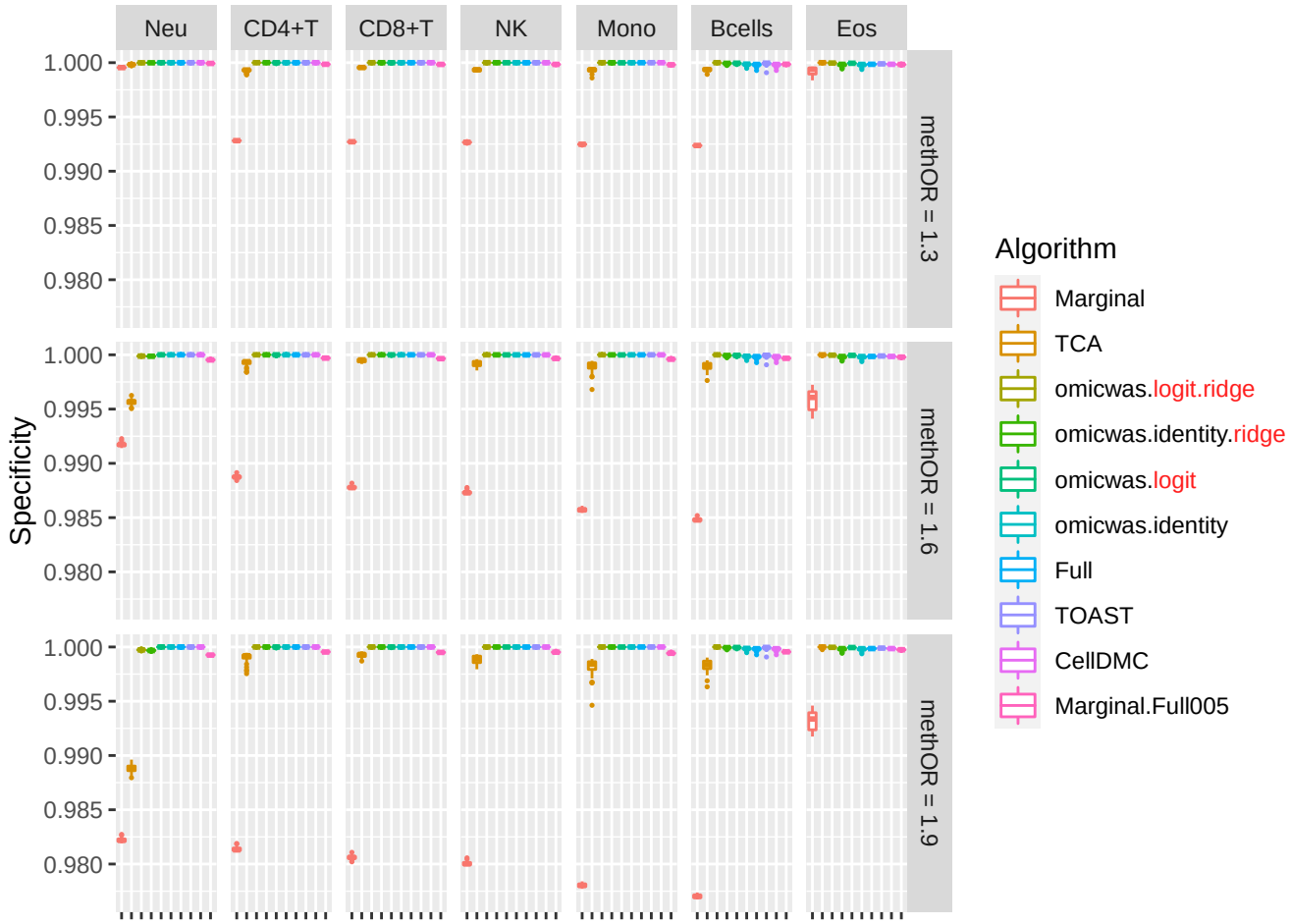

Neu
CD4+T
CD8+T
NK
Mono
Bcells
Eos
Specificity
1.000
0.995
0.990
0.985
0.980
methOR = 1.3
methOR = 1.6
methOR = 1.9
Algorithm
Marginal
TCA
omicwas.logit.ridge
omicwas.identity.ridge
omicwas.logit
omicwas.identity
Full
TOAST
CellDMC
Marginal.Full005

Figure 4

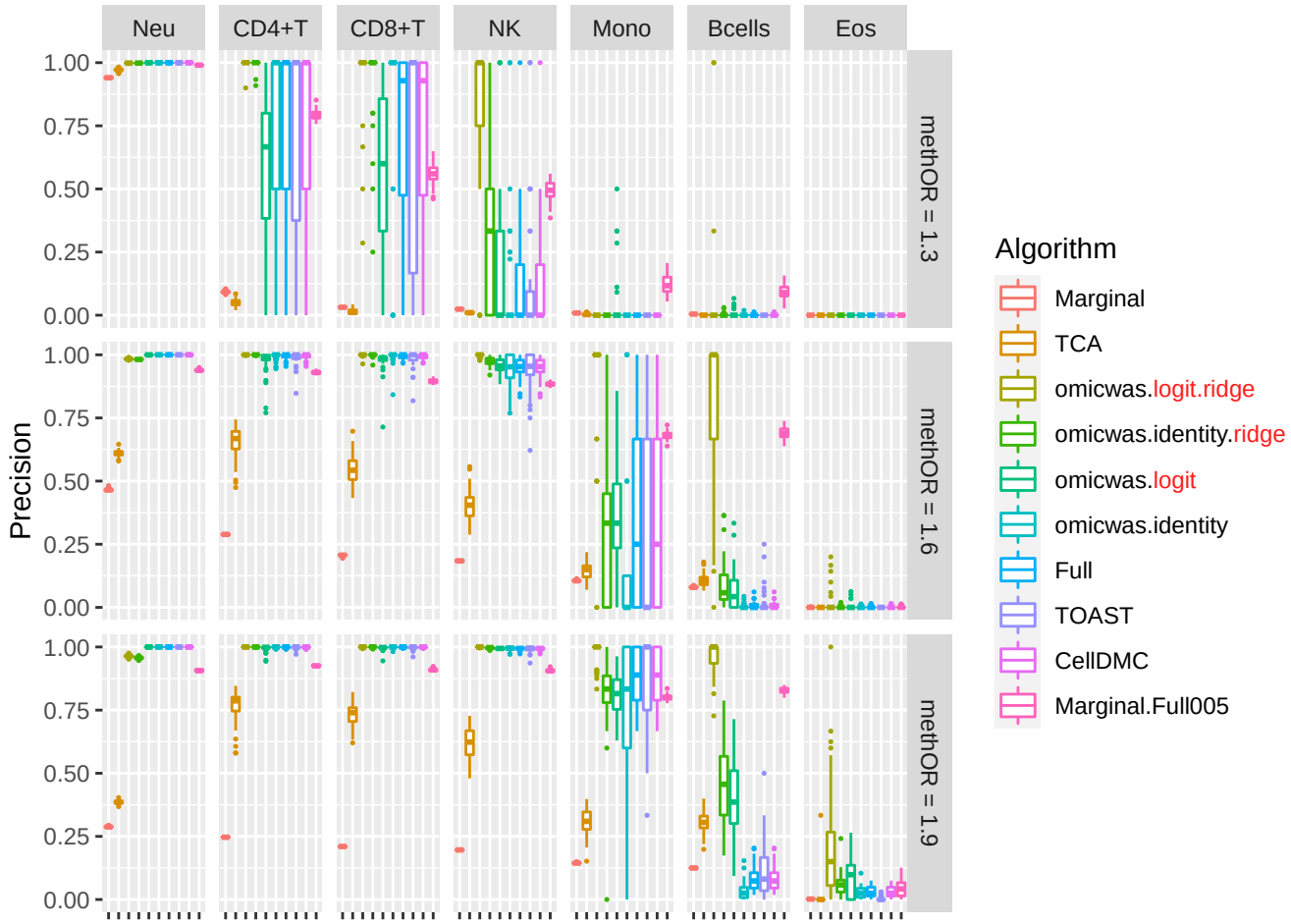

Neu
CD4+T
CD8+T
NK
Mono
Bcells
Eos
Precision
1.00
0.75
0.50
0.25
0.00
methOR = 1.3
methOR = 1.6
methOR = 1.9
Algorithm
Marginal
TCA
omicwas.logit.ridge
omicwas.identity.ridge
omicwas.logit
omicwas.identity
Full
TOAST
CellDMC
Marginal.Full005

Figure 5

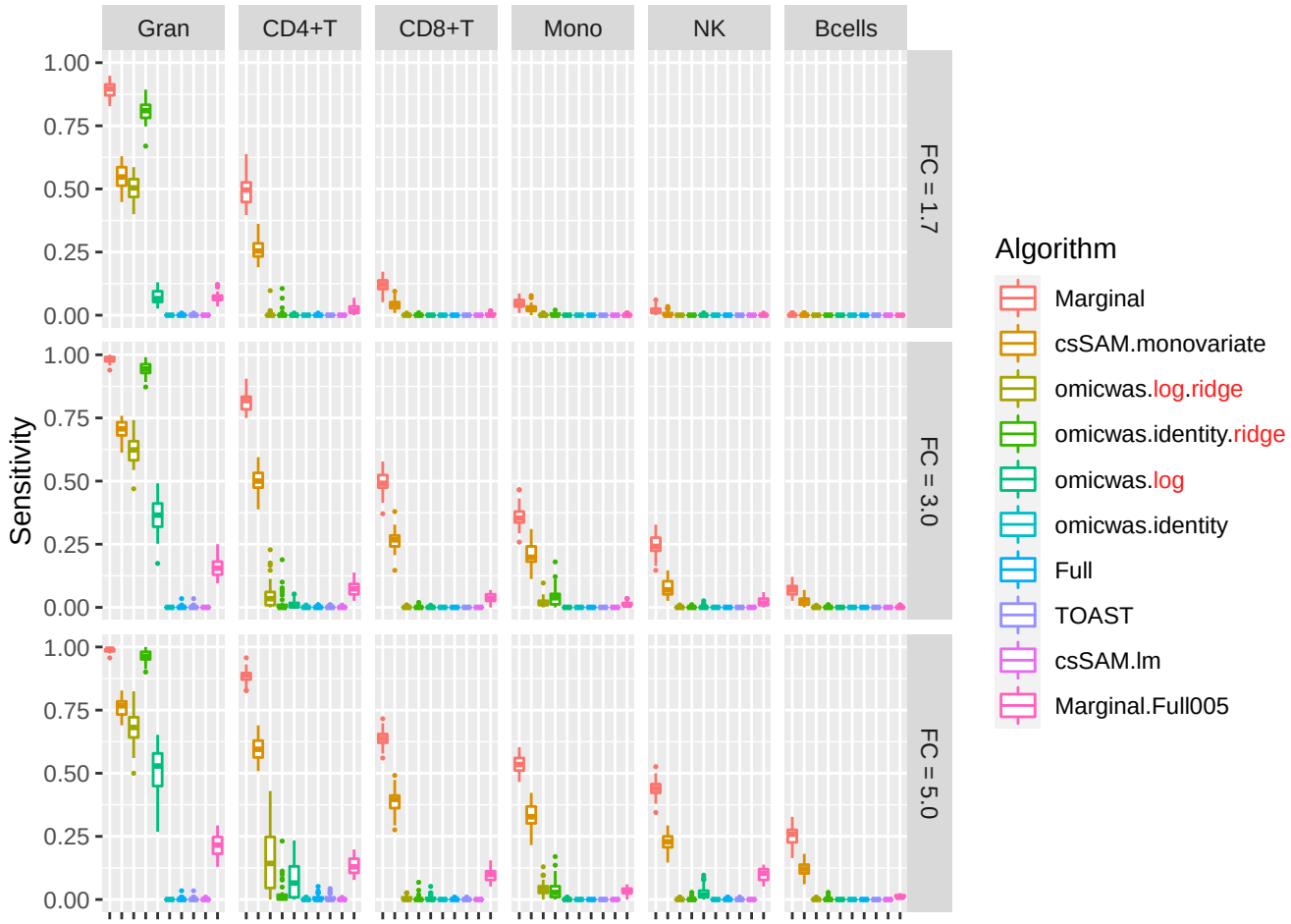

Gran
CD4+T
CD8+T
Mono
NK
Bcells
FC = 1.7
FC = 3.0
FC = 5.0
Sensitivity
1.00
0.75
0.50
0.25
0.00
Algorithm
Marginal
csSAM.monovariate
omicwas.log.ridge
omicwas.identity.ridge
omicwas.log
omicwas.identity
Full
TOAST
csSAM.lm
Marginal.Full005

Figure 6

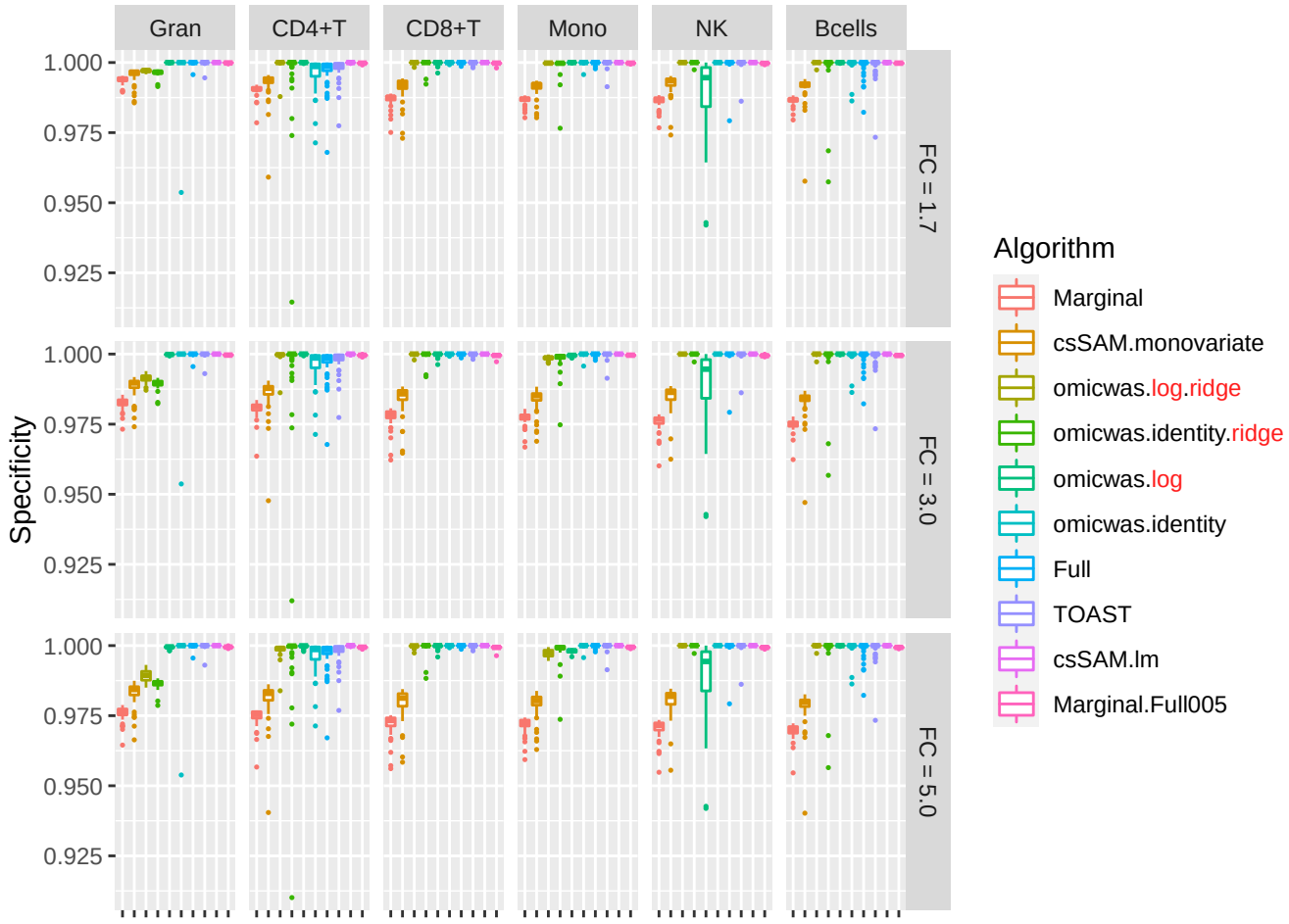

Gran
CD4+T
CD8+T
Mono
NK
Bcells
FC = 1.7
FC = 3.0
FC = 5.0
Specificity
1.000
0.975
0.950
0.925
Algorithm
Marginal
csSAM.monovariate
omicwas.log.ridge
omicwas.identity.ridge
omicwas.log
omicwas.identity
Full
TOAST
csSAM.lm
Marginal.Full005

Figure 7

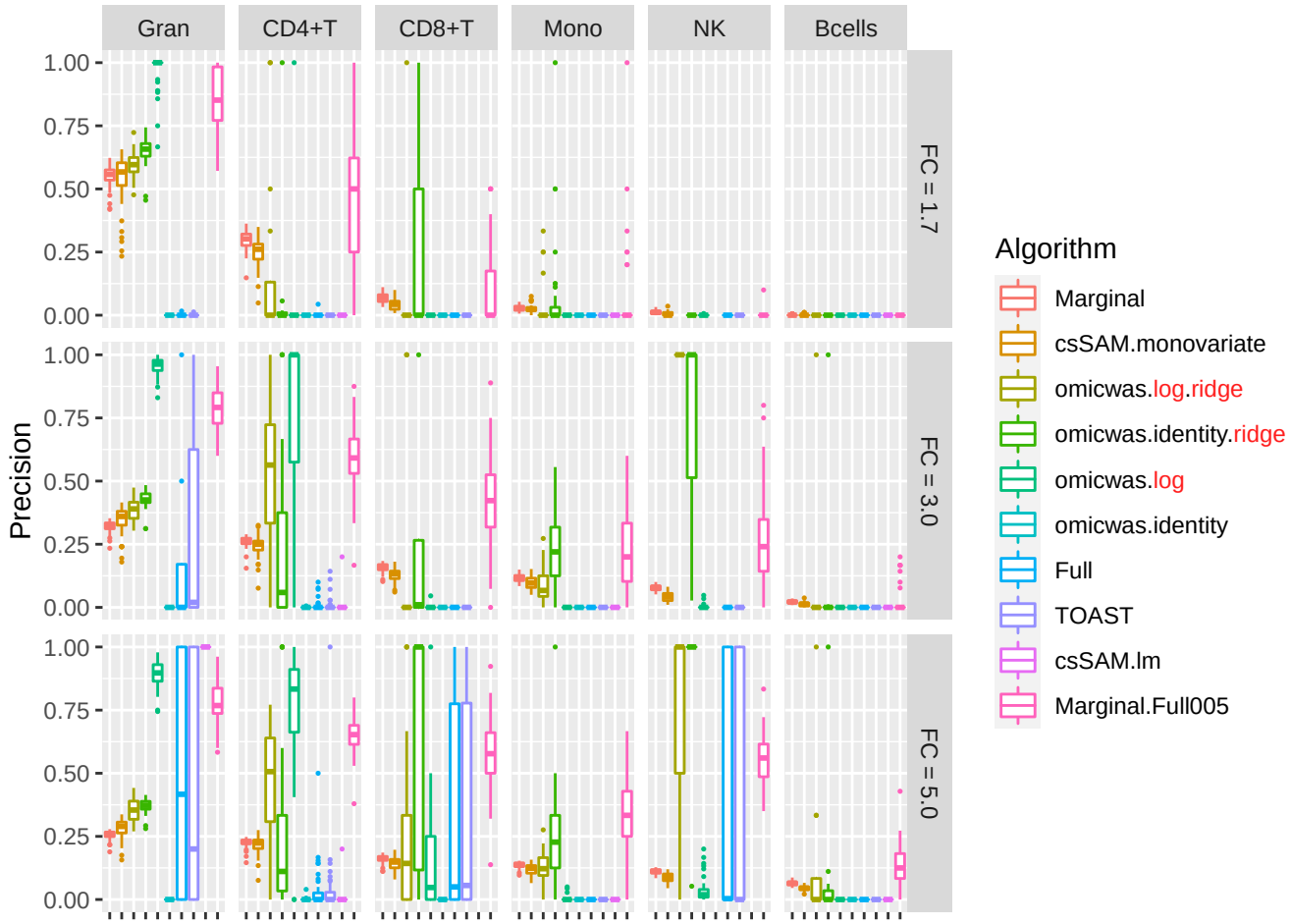

Gran
CD4+T
CD8+T
Mono
NK
Bcells
FC = 1.7
FC = 3.0
FC = 5.0
Precision
1.00
0.75
0.50
0.25
0.00
Algorithm
Marginal
csSAM.monovariate
omicwas.log.ridge
omicwas.identity.ridge
omicwas.log
omicwas.identity
Full
TOAST
csSAM.lm
Marginal.Full005

Figure 8

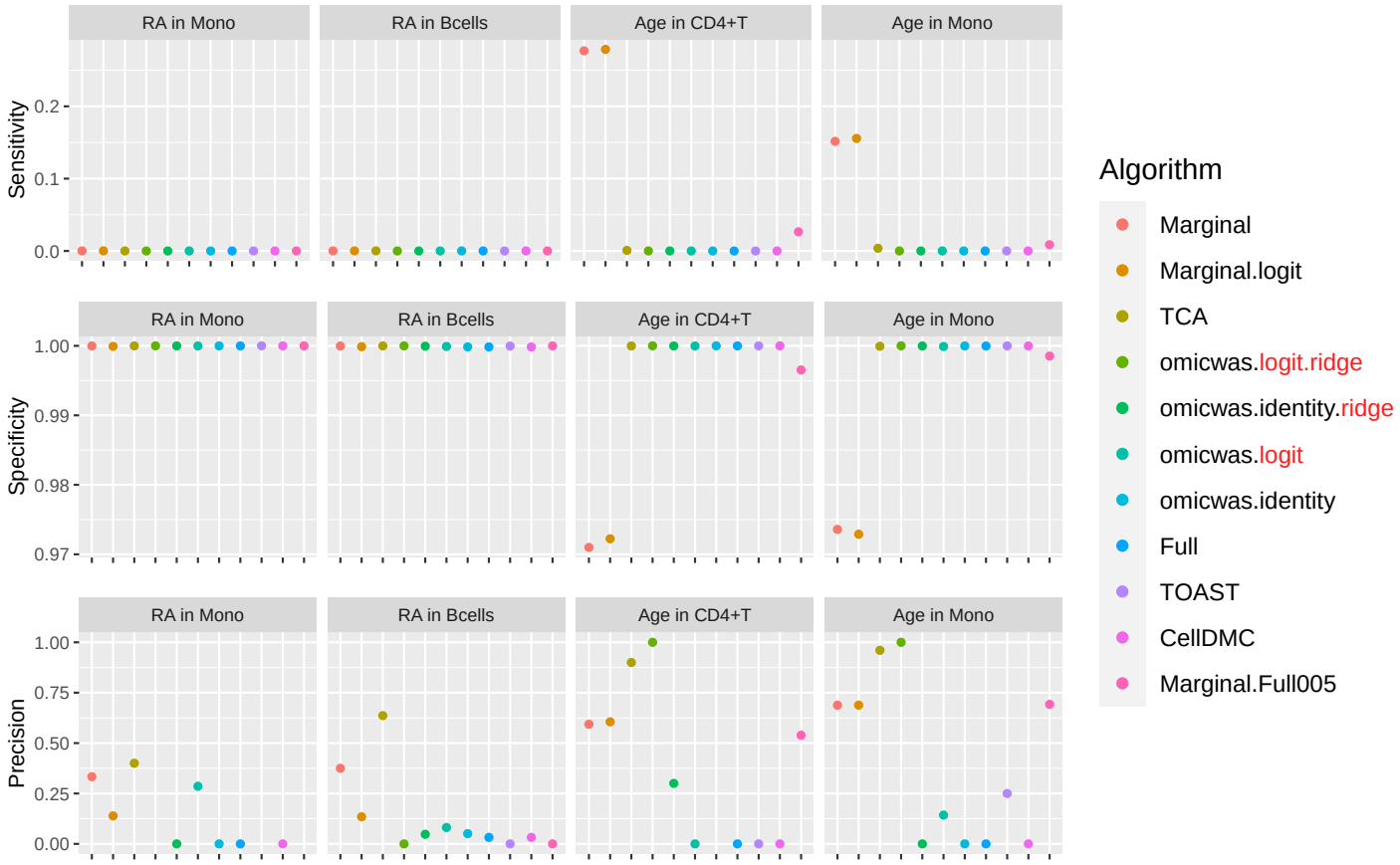
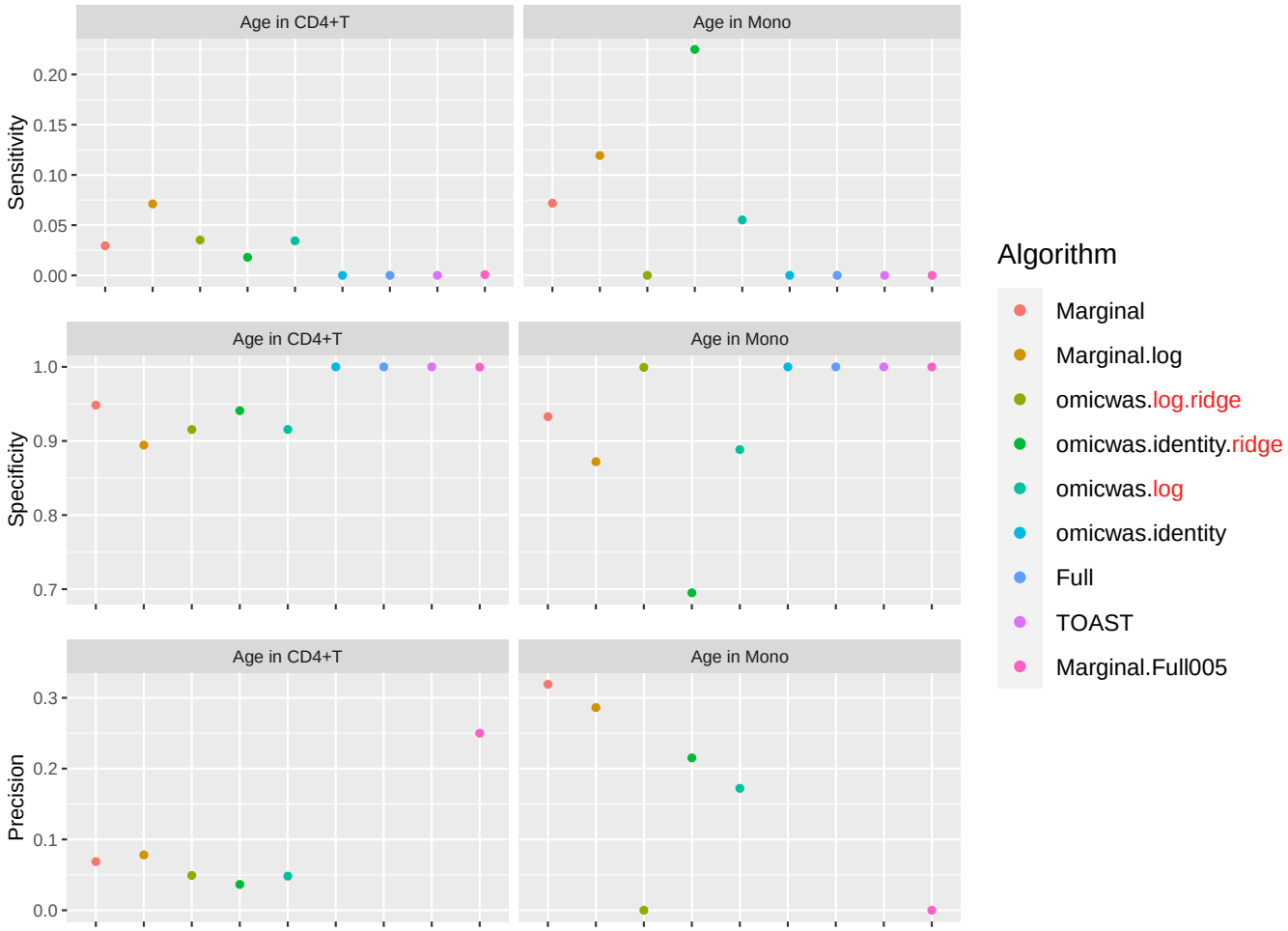

RA in Mono
RA in Bcells
Age in CD4+T
Age in Mono
Sensitivity
0.2
0.1
0.0
Specificity
1.00
0.99
0.98
0.97
Precision
1.00
0.75
0.50
0.25
0.00
Algorithm
Marginal
Marginal.logit
TCA
omicwas.logit.ridge
omicwas.identity.ridge
omicwas.logit
omicwas.identity
Full
TOAST
CellDMC
Marginal.Full005

Figure 9


Age in CD4+T
Age in Mono
Sensitivity
0.20
0.15
0.10
0.05
0.00
Specificity
1.0
0.9
0.8
0.7
Precision
0.3
0.2
0.1
0.0
Algorithm
Marginal
Marginal.log
omicwas.log.ridge
omicwas.identity.ridge
omicwas.log
omicwas.identity
Full
TOAST
Marginal.Full005