Sparse latent factor regression models for genome-wide and epigenome-wide association studies

Basile Jumentier[1,3] Kevin Caye[1] Barbara Heude[2]
Johanna Lepeule[3,⋆] Olivier François[1,⋆]

Authors' affiliations:

[1] Université Grenoble-Alpes, Centre National de la Recherche Scientifique, Grenoble INP, TIMC-IMAG CNRS UMR 5525, 38000 Grenoble, France.

[2] Université de Paris, CRESS, Inserm, INRAE, F-75004 Paris, France.

[3] Université Grenoble-Alpes, Centre National de la Recherche Scientifique, Institut National de la Santé et de la Recherche Médicale, Institute for Advanced Biosciences, INSERM U 1209, CNRS UMR 5309, 38000 Grenoble, France.

⋆ Corresponding authors: `johanna.lepeule@inserm.fr` , `olivier.francois@imag.fr`

Abstract

Association of phenotypes or exposures with genomic and epigenomic data faces important statistical challenges. One of these challenges is to remove variation due to unobserved confounding factors, such as individual ancestry or cell-type composition in tissues. This issue can be addressed with penalized latent factor regression models, where penalties are introduced to cope with high dimension in the data. If a relatively small proportion of genomic or epigenomic markers correlate with the variable of interest, sparsity penalties may help to capture the relevant associations, but the improvement over non-sparse approaches has not been fully evaluated yet. In this study, we introduced least-squares algorithms that jointly estimate effect sizes and confounding factors in sparse latent factor regression models. Computer simulations provided evidence that sparse latent factor regression models achieve higher statistical performance than other sparse methods, including the least absolute shrinkage and selection operator (LASSO) and a Bayesian sparse linear mixed model (BSLMM). Additional simulations based on real data showed that sparse latent factor regression models were more robust to departure from the generative model than non-sparse approaches, such as surrogate variable analysis (SVA) and other methods. We applied sparse latent factor regression models to a genome-wide association study of a flowering trait for the plant *Arabidopsis thaliana* and to an epigenome-wide association study of smoking status in pregnant women. For both applications, sparse latent factor regression models facilitated the estimation of non-null effect sizes while avoiding multiple testing problems. The results were not only consistent with previous discoveries, but they also pinpointed new genes with functional annotations relevant to each application.

1 Introduction

Association studies represent one of the most powerful tool to identify genomic variation correlating with disease states, exposure levels or phenotypes. Those studies are divided into several categories according to the nature of the genomic markers evaluated. For example, genome-wide association studies (GWAS) focus on single-nucleotide polymorphisms in different individuals to estimate disease allele effects (Balding, 2006), while epigenome-wide association studies (EWAS) measure epigenetic marks, such as DNA methylation levels to derive associations between epigenetic variation and exposure levels or to assess effects on phenotypic traits (Rakyan *et al.*, 2011). Despite their success in identifying the genetic architecture of phenotypic traits or genomic targets of exposure, association studies are plagued with the problem of confounding, which arises when unobserved variables correlate with the variable of interest and with the genomic markers simultaneously (Wang *et al.*, 2017).

Historical approaches to the confounding issue remove hidden confounders by considering corrections for inflation (Devlin and Roeder, 1999) and empirical null-hypothesis testing methods (Efron, 2004). Alternative approaches evaluate hidden confounders by using linear combinations of observed variables, often called factors. In GWAS, a frequently-used factor approach consists of computing the largest principal components of the genotype matrix, and includes them as covariates in linear regression models (Price *et al.*, 2006). The variable of interest may, however, be co-linear to largest principal components, and removing their effects can result in loss of statistical power. To increase power, methods based on latent factor regression models have been proposed (Leek and Storey, 2007; Carvalho *et al.*, 2008). Latent factor regression models employ deconvolution methods in which unobserved vari-

ables, including batch effects, individual ancestry or tissue cell-type composition are integrated in the regression model by using latent factors. In these models, effect sizes and latent factors are estimated jointly. The latent factor regression framework encompasses several methods which include surrogate variable analysis (SVA, Leek and Storey (2007)), latent factor mixed models (LFMM, Frichot *et al.* (2013)), residual principal component analysis (Kalaitzis and Lawrence, 2012), and confounder adjusted testing and estimation (CATE, Wang *et al.* (2017)). Each method has specific merits relative to some category of association study, and the performances of the methods have been extensively debated in recent surveys (for example, see Kaushal *et al.* (2017)).

A property of many latent factor regression models is to use regularization parameters inducing constraints on effect size estimates. Among those methods, sparse regression models suppose that a relatively small proportion of all genomic variables correlate with the variable of interest or affect the phenotype, and evaluate associations while avoiding multiple testing problems (Tibshirani, 1996; Hoggart *et al.*, 2008; Wu *et al.*, 2009). Sparse regression models have been coupled with linear mixed models to combine the benefits of both for polygenic trait studies with Bayesian sparse linear mixed model (BSLMM) (Zhou *et al.*, 2012, 2013). Sparse regression models can include confounding factors that are usually estimated separately of effect sizes. In this study, we introduce least-squares algorithms that jointly estimate effect sizes and confounding factors in sparse latent factor regression models. We estimate effect sizes based on regularized least-squares methods with L^1 and nuclear norm penalties. Thus our method allows identifying non-null effect sizes without the use of multiple statistical tests. We refer to our models as *sparse latent factor mixed models* or *sparse*

LFMM. We present estimation algorithms for sparse LFMM and theoretical results in the next section. Then we compare the performances of sparse LFMM with other sparse regression models (LASSO, BSLMM), and with non-sparse regression models (SVA, CATE, LFMM). To illustrate our approach, we used sparse LFMM to perform a GWAS of flowering time for the plant *Arabidopsis thaliana* and to perform an epigenome-wide association study (EWAS) of smoking status in pregnant women.

2 Latent factor regression models

2.1 Models

Latent factor regression models evaluate associations between the elements of a response matrix, $\mathbf{Y}$, and variables of interest, called *primary* variables, $\mathbf{X}$, measured for n individuals. The response matrix records p markers, which can represent any type of omic data (genotypes, DNA methylation, etc), collected for the individuals. The $\mathbf{X}$ matrix can also incorporate nuisance variables such as observed confounders (age, sex, etc), and its dimension is $n \times d$, where d represents the total number of primary and nuisance variables. Latent factor regression models are regression models combining fixed and latent effects as follows

$$\mathbf{Y} = \mathbf{X}\mathbf{B}^T + \mathbf{W} + \mathbf{E}. \tag{1}$$

Fixed effect sizes are recorded in the $\mathbf{B}$ matrix, which has dimension $p \times d$. The $\mathbf{E}$ matrix represents residual errors, and has the same dimension as the response matrix. The matrix $\mathbf{W}$ is a *latent matrix* of rank K, defined by K latent factors (Leek and Storey, 2007; Frichot *et al.*, 2013; Wang *et al.*, 2017). The value of K is unknown, and it is generally determined by model choice or cross-validation procedures. The

K latent factors, $\mathbf{U}$, are defined from the singular value decomposition of the latent matrix

$$\mathbf{W} = \mathbf{U}\mathbf{V}^T\,,$$

where $\mathbf{V}$ is a $K \times p$ matrix of loadings (Eckart and Young, 1936). The matrices $\mathbf{U}$ and $\mathbf{V}$ are unique up to a change of sign.

Naive statistical estimates for the $\mathbf{B}$ and $\mathbf{W}$ matrices in equation (1) could be obtained through the minimization of a classical least-squares loss function

$$\mathcal{L}(\mathbf{B}, \mathbf{W}) = \|\mathbf{Y} - \mathbf{W} - \mathbf{X}\mathbf{B}^T\|_F^2\,, \tag{2}$$

where $\|.\|_F$ is the Frobenius matrix norm. A minimum value of the loss function is attained when $\mathbf{W}$ is computed as the rank K singular value decomposition of $\mathbf{Y}$. In this case, the $\mathbf{B}$ matrix can be obtained from the estimates of a linear regression of the residual matrix $(\mathbf{Y} - \mathbf{W})$ on $\mathbf{X}$. To motivate the introduction of regularization terms in the loss function, we remark that the interpretation of latent factors obtained from this solution as confounder estimates may be incorrect, because it fails to include any information on the primary variable, $\mathbf{X}$. Assuming that latent factors are computed only from the response matrix contradicts the definition of confounding variables (Wang *et al.*, 2017). In addition, the definition is problematic, because it does not lead to a unique minimum of the loss function. To see it, consider any matrix $\mathbf{P}$ with dimensions $d \times p$ and check that

$$\|\mathbf{Y} - (\mathbf{U} - \mathbf{X}\mathbf{P})\mathbf{V}^T + \mathbf{X}(\mathbf{B}^T - \mathbf{P}\mathbf{V}^T))\|_F^2 = \|\mathbf{Y} - \mathbf{U}\mathbf{V}^T + \mathbf{X}\mathbf{B}^T\|_F^2.$$

As a consequence, $\mathbf{B}$ and $(\mathbf{B} - \mathbf{V}\mathbf{P}^T)$ correspond to valid minima, and there is

an infinite space of possible solutions. To conclude, the loss function needs to be modified in order to warrant dependency of $\mathbf{W}$ on both $\mathbf{Y}$ and $\mathbf{X}$, and to enable the computation of well-defined solutions.

2.2 Sparse estimation algorithms

L^1-regularized least-square problem. To solve the problems outlined in the above section, a sparse regularization approach is considered. This approach introduces penalties based on the L^1 norm of the regression coefficients and on the nuclear norm of the latent matrix

$$\mathcal{L}_{\text{sparse}}(\mathbf{W}, \mathbf{B}) = \left\|\mathbf{Y} - \mathbf{W} - \mathbf{X}\mathbf{B}^T\right\|_F^2 + \mu\|\mathbf{B}\|_1 + \gamma\|\mathbf{W}\|_* , \quad \mu, \gamma > 0, \tag{3}$$

where $\|\mathbf{B}\|_1$ denotes the L^1 norm of $\mathbf{B}$, μ is an L^1 regularization parameter, $\mathbf{W}$ is the latent matrix, $\|\mathbf{W}\|_*$ denotes its nuclear norm, and γ is a regularization parameter for the nuclear norm. The L^1 penalty induces sparsity on the fixed effects (Tibshirani, 1996), and corresponds to the prior information that not all response variables may be associated with the primary variables. More specifically, the prior implies that a restricted number of rows of the effect size matrix $\mathbf{B}$ are non-zero. The second regularization term is based on the nuclear norm, and it is introduced to penalize large numbers of latent factors. With these penalty terms, $\mathcal{L}_{\text{sparse}}(\mathbf{W}, \mathbf{B})$ is a convex function, and convex mimimization algorithms can be applied to obtain estimates of $\mathbf{B}$ and $\mathbf{W}$ (Mishra *et al.*, 2013).

Sparse latent factor mixed model algorithm. To simplify the description of the estimation algorithm, let us assume that the explanatory variables, $\mathbf{X}$, are scaled so that $\mathbf{X}^T\mathbf{X} = \mathbf{Id}_d$. Note that our program implementation is more general, and

does not make this restrictive assumption. Here, it is introduced to explain the sparse LFMM algorithm with simplified notations. We developed a block-coordinate descent method for minimizing the convex loss function $\mathcal{L}_{\text{sparse}}(\mathbf{W}, \mathbf{B})$ with respect to $\mathbf{B}$ and $\mathbf{W}$. The algorithm is initialized from the null matrix $\hat{\mathbf{W}}_0 = 0$, and iterates the following steps.

1. Find $\hat{\mathbf{B}}_t$ a minimum of the penalized loss function

$$\mathcal{L}^{(1)}_{\text{sparse}}(\mathbf{B}) = \|(\mathbf{Y} - \hat{\mathbf{W}}_{t-1}) - \mathbf{X}\mathbf{B}^T\|_F^2 + \mu\|\mathbf{B}\|_1 \,, \tag{4}$$

2. Find $\hat{\mathbf{W}}_t$ a minimum of the penalized loss function

$$\mathcal{L}^{(2)}_{\text{sparse}}(\mathbf{W}) = \|(\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}}_t^T) - \mathbf{W}\|_F^2 + \gamma\|\mathbf{W}\|_* . \tag{5}$$

The algorithm cycles through the two steps until a convergence criterion is met or the allocated computing resource is depleted. Each minimization step has a well-defined and unique solution. To see it, note that Step 1 corresponds to an L^1-regularized regression of the residual matrix $(\mathbf{Y} - \hat{\mathbf{W}}_{t-1})$ on the explanatory variables. To compute the regression coefficients, we used the Friedman block-coordinate descent method (Friedman *et al.*, 2007). According to Tibshirani (1996), we obtained

$$\hat{\mathbf{B}}_t = \text{sign}(\bar{\mathbf{B}}_t)(\bar{\mathbf{B}}_t - \mu)_+ \,, \tag{6}$$

where $s_+ = \max(0, s)$, $\text{sign}(s)$ is the sign of s, and $\bar{\mathbf{B}}_t$ is the linear regression estimate, $\bar{\mathbf{B}}_t = \mathbf{X}^T\mathbf{Y} - \hat{\mathbf{W}}_{t-1}$. Step 2 consists of finding a low rank approximation of the residual matrix $\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}}_t^T$ (Cai *et al.*, 2008). This approximation starts with a singular value decomposition (SVD) of the residual matrix, $\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}}_t^T = \mathbf{M}\mathbf{S}\mathbf{N}^T$, with $\mathbf{M}$ a unitary

matrix of dimension $n \times n$, $\mathbf{N}$ a unitary matrix of dimension $p \times p$, and $\mathbf{S}$ the matrix of singular values $(s_j)_{j=1,\ldots,n}$. Then, we obtain

$$\hat{\mathbf{W}}_t = \mathbf{M}\bar{\mathbf{S}}\mathbf{N}^T \tag{7}$$

where $\bar{\mathbf{S}}$ is the diagonal matrix with diagonal terms $\bar{s}_j = (s_j - \gamma)_+$, $j = 1, \ldots, n$. Building on results from Tseng (2001), the following statement holds.

Theorem 1. *Let $\mu > 0$ and $\gamma > 0$. Then the block-coordinate descent algorithm cycling through Step 1 and Step 2 converges to estimates of $\mathbf{W}$ and $\mathbf{B}$ defining a global minimum of the penalized loss function $\mathcal{L}_{\text{sparse}}(\mathbf{W}, \mathbf{B})$.*

Note that the algorithmic complexities of Step 1 and Step 2 are bounded by a term of order $O(pn + K(p + n))$. The computing time of sparse LFMM estimates is generally longer than for the CATE algorithm (Wang *et al.*, 2017) or the ridge LFMM algorithm detailed below (Caye *et al.*, 2019). Sparse LFMM needs to perform SVD and projections several times until convergence while CATE and ridge LFMM require a single iteration.

2.3 Ridge regression algorithms

Caye *et al.* (2019) considered a related approach, referred to as ridge LFMM, where the statistical estimates of the parameter matrices $\mathbf{B}$ and $\mathbf{W}$ are computed after minimizing the loss function with L^2 norm regularization defined as follows

$$\mathcal{L}_{\text{ridge}}(\mathbf{B}, \mathbf{W}) = \|\mathbf{Y} - \mathbf{W} - \mathbf{X}\mathbf{B}^T\|_F^2 + \lambda\|\mathbf{B}\|_2^2, \quad \lambda > 0, \tag{8}$$

where $\|.\|_F$ is the Frobenius norm, $\|.\|_2$ is the L^2 norm, and λ is a regularization pa-

rameter. The minimization algorithm starts with an SVD of the explanatory matrix, $\mathbf{X} = \mathbf{Q}\mathbf{\Sigma}\mathbf{R}^T$, where $\mathbf{Q}$ is an $n \times n$ unitary matrix, $\mathbf{R}$ is an $d \times d$ unitary matrix and $\mathbf{\Sigma}$ is an $n \times d$ matrix containing the singular values of $\mathbf{X}$, denoted by $(\sigma_j)_{j=1,\ldots,d}$. The ridge estimates are computed as follows

$$\hat{\mathbf{W}} = \mathbf{Q}\mathbf{D}_\lambda^{-1}\mathrm{svd}_K(\mathbf{D}_\lambda\mathbf{Q}^T\mathbf{Y}) \tag{9}$$

$$\hat{\mathbf{B}}^T = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{Id}_d)^{-1}\mathbf{X}^T(\mathbf{Y} - \hat{\mathbf{W}}), \tag{10}$$

where $\mathrm{svd}_K(\mathbf{A})$ is the SVD of rank K of $\mathbf{A}$, $\mathbf{Id}_d$ is the $d \times d$ identity matrix, and $\mathbf{D}_\lambda$ is the $n \times n$ diagonal matrix with coefficients defined as

$$\mathbf{d}_\lambda = \left(\sqrt{\frac{\lambda}{\lambda + \sigma_1^2}}, \ldots, \sqrt{\frac{\lambda}{\lambda + \sigma_d^2}}, 1, \ldots, 1\right).$$

For $\lambda > 0$, the solution of the regularized least-squares problem is unique (Caye *et al.*, 2019), and the corresponding matrices are called the *ridge estimates*. For completeness, we provide a short proof for this result, stated in (Caye *et al.*, 2019), in the appendix. Using random projections to compute low rank approximations, the complexity of the estimation ridge LFMM algorithm is of order $O(n^2p + np\log K)$ (Halko *et al.*, 2011). For studies in which the number of samples, n, is much smaller than the number of response variables, p, computing times of ridge estimates are therefore faster than those of sparse LFMM.

3 Results

Generative model experiments. In a first series of experiments, we compared sparse LFMM with LASSO and three non-sparse approaches (ridge LFMM, CATE,

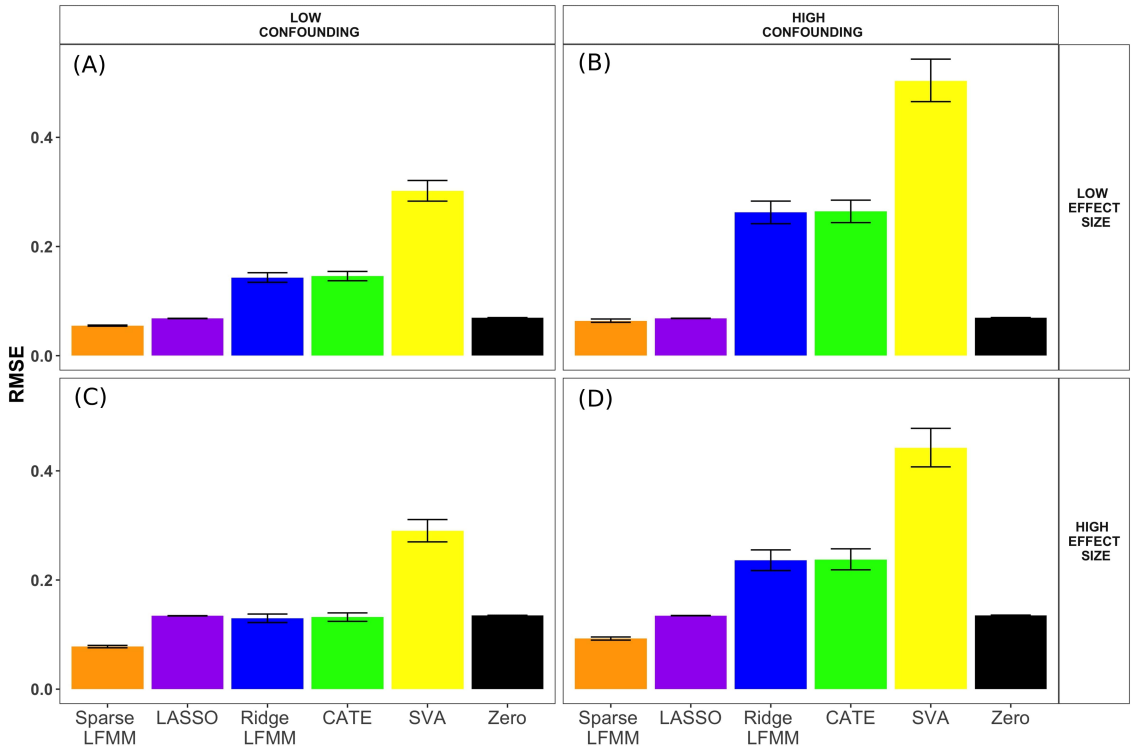

Figure 1. Root Mean Square Error (RMSE) as a function of the effect size of causal markers and confounding intensity. Two sparse methods (sparse LFMM, LASSO) and three non-sparse methods (ridge LFMM, CATE and SVA) were compared. The "Zero" value corresponds to an RMSE obtained with all effect sizes set to zero (null-model error). Generative model simulation parameters: (A) Lower effect sizes and confounding intensities (B) Lower effect sizes and higher confounding intensities. (C) Higher effect sizes and lower confounding intensities. (D) Higher effect sizes and confounding intensities.

SVA). The data were simulated from the generative model defined in equation (1), and the performance of each algorithm was measured in four scenarios showing higher or lower effect sizes and confounding intensities (Figure 1). For all experiments, we computed statistical errors (RMSE) for the effect size estimates of each method (Figure 1). To provide a reference value for the RMSE, we measured the error made when all effect sizes were estimated as being null ("Zero" value or null-model error). The null-model error was equal to 0.069 in low effect size scenarios and equal to 0.135 in high effect size scenarios. A powerful method was expected to reach error levels lower than the null-model error. The RMSEs of sparse LFMM ranged from 0.055 to 0.092, less than those of the null-model. The RMSEs of LASSO were close to the ones of sparse LFMM in the low effect size scenarios. In contrast, non-sparse methods led to RMSEs higher than the null-model error, ranging between 0.13 and 0.26 for ridge LFMM and CATE, and rising up to 0.50 for SVA. For the effect sizes associated with causal markers, non-sparse methods reached lower RMSE values than those of sparse methods, ranging between 0.12 and 0.26 for ridge LFMM and CATE, and between 0.60 and 1.03 for sparse LFMM (Figure S1). Regarding precision and F-score - which is a harmonic mean of power and precision, the performances of all methods were higher in scenarios with higher effect size and lower confounding intensity. Sparse LFMM performed similarly to or less than the LASSO when the size of the causal effects was small (Figure 2AB), but it reached higher F-scores for larger effect sizes (Figure 2CD). In those simulations, sparse LFMM obtained lower F-scores than ridge LFMM and CATE. The difference was substantial when the sizes of the causal effects were small ($F \approx 0.51$ versus $F \approx 0.76$, Figure 1AB), but the differences were small for the larger effect sizes ($F \approx 0.75$, Figure 2CD). In all scenarios, sparse

LFMM obtained better scores than SVA. In summary, sparse LFMM was associated with the smallest overall statistical error, but the estimates of effect size were biased more severely with this method than with non-sparse methods. Sparse LFMM was generally preferable to LASSO and SVA. Once non-null effect sizes are identified by sparse LFMM, a consensus strategy would use ridge LFMM or CATE for evaluating the effect sizes of the candidate markers.

Empirical simulation experiments. In a second series of experiments, we used realistic simulations to compare sparse LFMM to other sparse and non-sparse methods. Simulations were based on 162 ecotypes of the model plant *Arabidopsis thaliana* using 53,859 SNP genotypes in chromosome 5. The simulations considered lower and higher effect sizes and gene by environment ($G \times E$) interaction levels. Those simulations departed from generative model simulations, and they were introduced to evaluate the robustness of effect size estimates in each approach. In lower $G \times E$ interaction scenarios, sparse LFMM obtained the highest scores (F in (0.57,0.60), precision in (0.81,0.82), Figure 3AC) compared to BSLMM (F in (0.36,0.44)), and to non-sparse methods (F ranging between 0.25 and 0.28). In higher $G \times E$ interaction scenarios, all methods obtained very low performances for the low effect size scenario, but sparse LFMM obtained among the highest F-score and precision. When the effect size was higher, sparse LFMM reached higher performances ($F \approx 0.28$ and accuracy ≈ 0.33) than the other methods (Figure 3D). In those realistic simulations, sparse LFMM demonstrated greater robustness to departure from the generative model assumptions than the other sparse methods (BSLMM, LASSO), and also compared favorably with non-sparse methods (ridge LFMM, CATE, SVA).

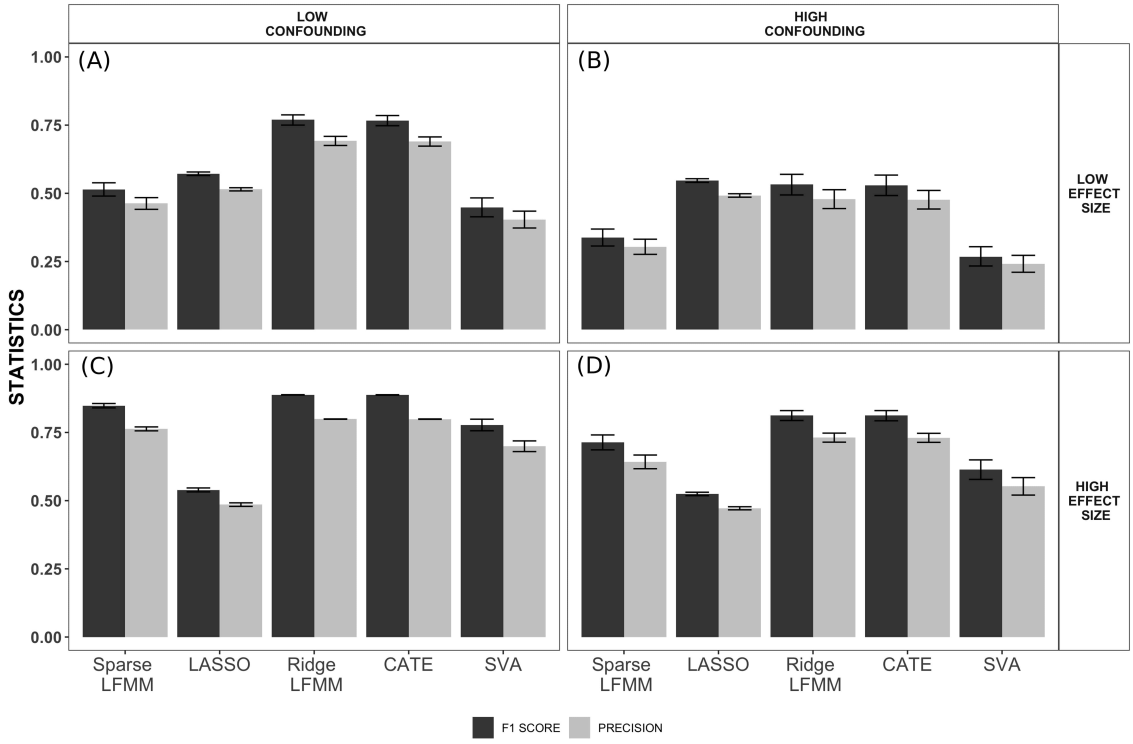

Figure 2. ***F*-score and precision as a function of effect size of the causal markers and confounding intensity.** Two sparse methods (sparse LFMM, LASSO) and three non-sparse methods (ridge LFMM, CATE and SVA) were compared. *F*-score is the harmonic mean of precision and recall. Generative model simulation parameters: (A) Lower effect sizes and confounding intensities (B) Lower effect sizes and higher confounding intensities. (C) Higher effect sizes and lower confounding intensities. (D) Higher effect sizes and confounding intensities.

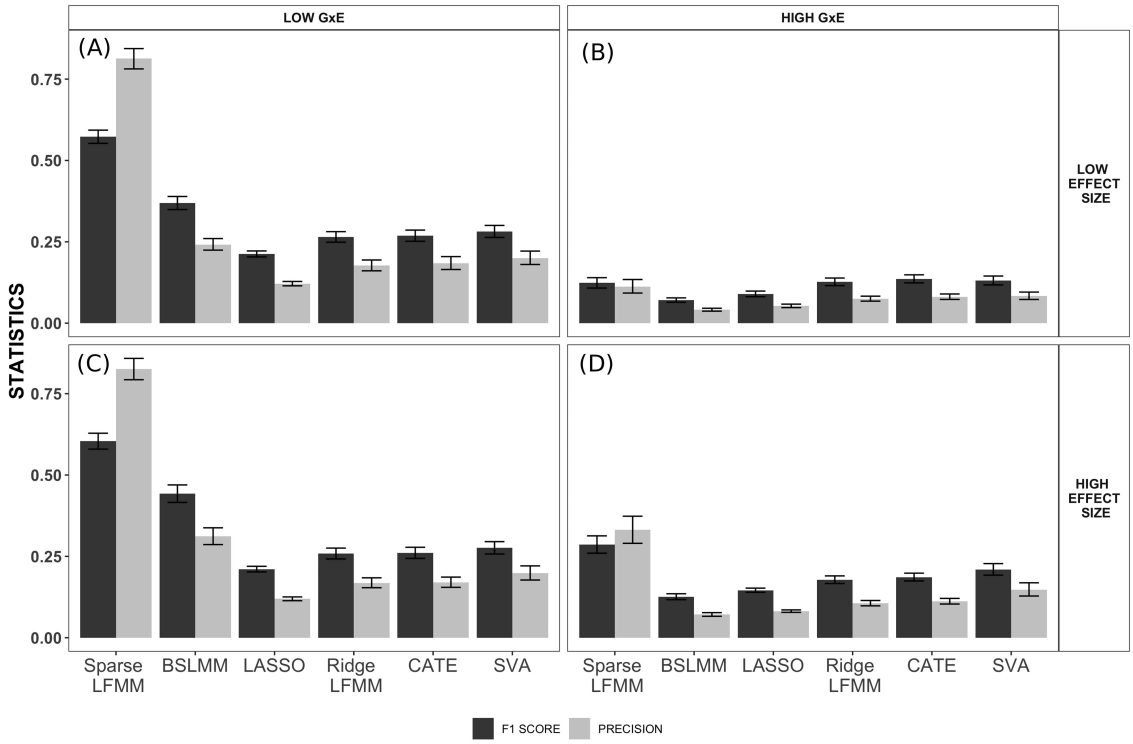

Figure 3. Empirical simulation data (F-score and precision). F-score and precision as a function of the effect size of the causal markers and of the strength of the interaction between genotype and environment ($G \times E$). Three sparse methods (sparse LFMM, BSLMM and LASSO) and three non-sparse methods (ridge LFMM, CATE and SVA) were compared. F-score is the harmonic mean of precision and recall. Simulation parameters: (A) Lower effect sizes and lower $G \times E$ (B) Lower effect sizes and higher $G \times E$. (C) Higher effect sizes and lower $G \times E$. (D) Higher effect sizes and higher $G \times E$.

Runtimes and number of factors. Next, we evaluated runtimes for sparse LFMM, and compared those runtimes with BSLMM and ridge LFMM (Figure S2). Whatever the number of individuals or markers, ridge LFMM was the fastest method, and sparse LFMM was the slowest method. Higher computation times for sparse LFMM were not surprising because the method iterates many cycles before convergence, whereas ridge LFMM is an exact approach. It took around 2,000 seconds for sparse LFMM to complete runs with $n = 1,000$ individuals and $p = 100,000$ markers. With default values for MCMC parameters, BSLMM runtimes were of the same order as those of sparse LFMM. To assess the choice of K by cross-validation, we varied the number of latent factors between 3 and 10, and compared the values estimated by cross validation with the true values. In 73% simulations, the number of latent factors was correctly estimated, and in the remaining 17% simulations, the true value of K was overestimated by one unit (Figure S3).

GWAS of flowering time in *A. thaliana*. To illustrate the use of latent factor models in a context where confounding is difficult to control for, we performed a GWAS of flowering time using $p = 53,859$ SNPs genotyped in chromosome 5 for $n = 162$ European accessions of the model plant *A. thaliana*. The sparse methods (sparse LFMM, LASSO, BSLMM) differed in their estimate of the number of null effect sizes (Figure 4ABC, Figure S4). The LASSO approach estimated 99.85% null effect sizes while the proportions were equal to 99.24% and 98.18% for BSLMM and sparse LFMM respectively. The LASSO was the most conservative approach, and sparse LFMM the most liberal one. Sparse LFMM shared 3.9% of hits with LASSO, and 5.5% with BSLMM (Figure S4). Less than 1% of all hits were common to the

three approaches. The (non-null) effect sizes for hits varied on distinct scales, with LASSO exhibiting the strongest biases. All sparse methods detected the same top hit at around 4 Mb, corresponding to a SNP located within the *FLC* gene, consistent with the results of Atwell *et al.* (2010). The second hit in (Atwell *et al.*, 2010), located in the gene *DOG1*, was also identified by sparse LFMM. BSLMM had more difficulties in identifying previously discovered genes. Given the high correlation – greater than 94 % – between effect sizes obtained with non-sparse methods, we grouped their results by averaging their estimates. Non-sparse methods exhibited effect sizes in a range of values closer to sparse LFMM than to LASSO and BSLMM, but higher statistical errors were observed for those approaches (Figure 4D). Overall, we found a significant correlation between the non-null effect sizes estimated by sparse LFMM and the corresponding effect sizes found by non-sparse methods ($\rho = 0.8065$, $P < 10^{-16}$). In addition, sparse LFMM and the non-sparse methods found new hits around 13.9 Mb and 6.5 Mb of chr 5, corresponding to the *SAP* and *ACL5* genes respectively.

EWAS of exposure to smoking during pregnancy. To evaluate association between smoking during pregnancy and placental DNA methylation, we performed an EWAS considering tobacco consumption as a primary variable. To this objective, we considered beta-normalized methylation levels at $p = 425,878$ probed CpG sites for $n = 668$ women (Heude *et al.*, 2016; Rousseaux *et al.*, 2019). The placentas were collected at delivery from women included in the EDEN mother-child cohort. Using sparse LFMM, the proportion of null effect sizes was equal to 99.698%, for a total number of 1,287 hits (Figure S5). To characterize the targeted CpGs, we evaluated

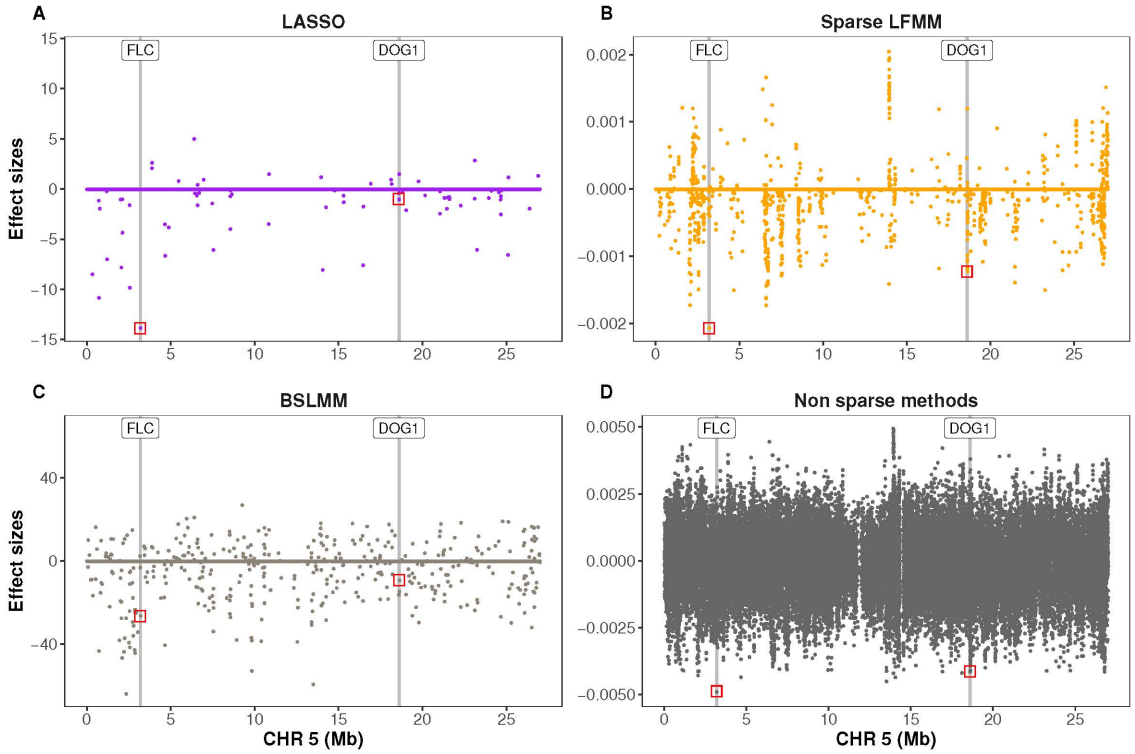

Figure 4. GWAS of flowering time in *A. thaliana* **(chromosome 5).** A) Effect size estimates for LASSO. B) Effect size estimates for sparse LFMM. C) Effect size estimates for sparse BSLMM. D) Average effect size estimates for non-sparse methods (ridge LFMM, CATE and SVA). Grey bars represent *Arabidopsis* SNPs associated with the FT16 phenotype in (Atwell *et al.*, 2010), and correspond to the *FLC* and *DOG1* genes.

whether there was an enrichment of enhancer and promoter regions in candidate regions compared to the methylome (Figure S6 and Figure S7). For the 1,287 CpGs with non-null effect sizes, 25.48% were found in enhancer regions, compared to 22.73% for the whole methylome, and 6.83% were found in promoter regions, compared to 19.94% for the whole methylome. We compared the CpGs having the highest effect sizes in each method (Figure S8). Sparse LFMM shared 45.3 % of hits with non-sparse models (represented by ridge LFMM), and 2.8 % of hits with LASSO (Table S1). Among the 51 top hits shared by sparse LFMM and ridge LFMM, 25 were found in the body of a gene, 11 were not associated with a gene, 20 were in enhancer regions and 2 in promoter regions. Note that in this analysis, we averaged the effect sizes of non-sparse methods because their correlation was greater than 99%. The results of sparse LFMM agreed with the results of non-sparse methods better than with those of LASSO. The Pearson correlation between the non-null effect sizes estimated by sparse LFMM and the corresponding effect sizes estimated non-sparse methods was equal to $\rho = 80.38\%$ ($P < 10^{-16}$), whereas the Pearson correlation between non-null effect sizes of sparse LFMM and LASSO was equal to $\rho = 61.86\%$ ($P < 10^{-16}$). To focus on a specific chromosome, we detailed the outputs of all approaches for chromosome 3, which contained the epigenome-wide top hit for sparse LFMM and for non-sparse methods (cg27402634, located on an enhancer, Figure 5). This CpG was also detected with LASSO (Figure S9). The sparse LFMM hits shared three additional CpGs with non-sparse methods: cg09627057, cg18557837 and cg12662091. Overall, sparse LFMM detected 61 CpGs with non-null effect sizes: 43 were located in genes, 22 in enhancer regions and 6 in promoter regions.

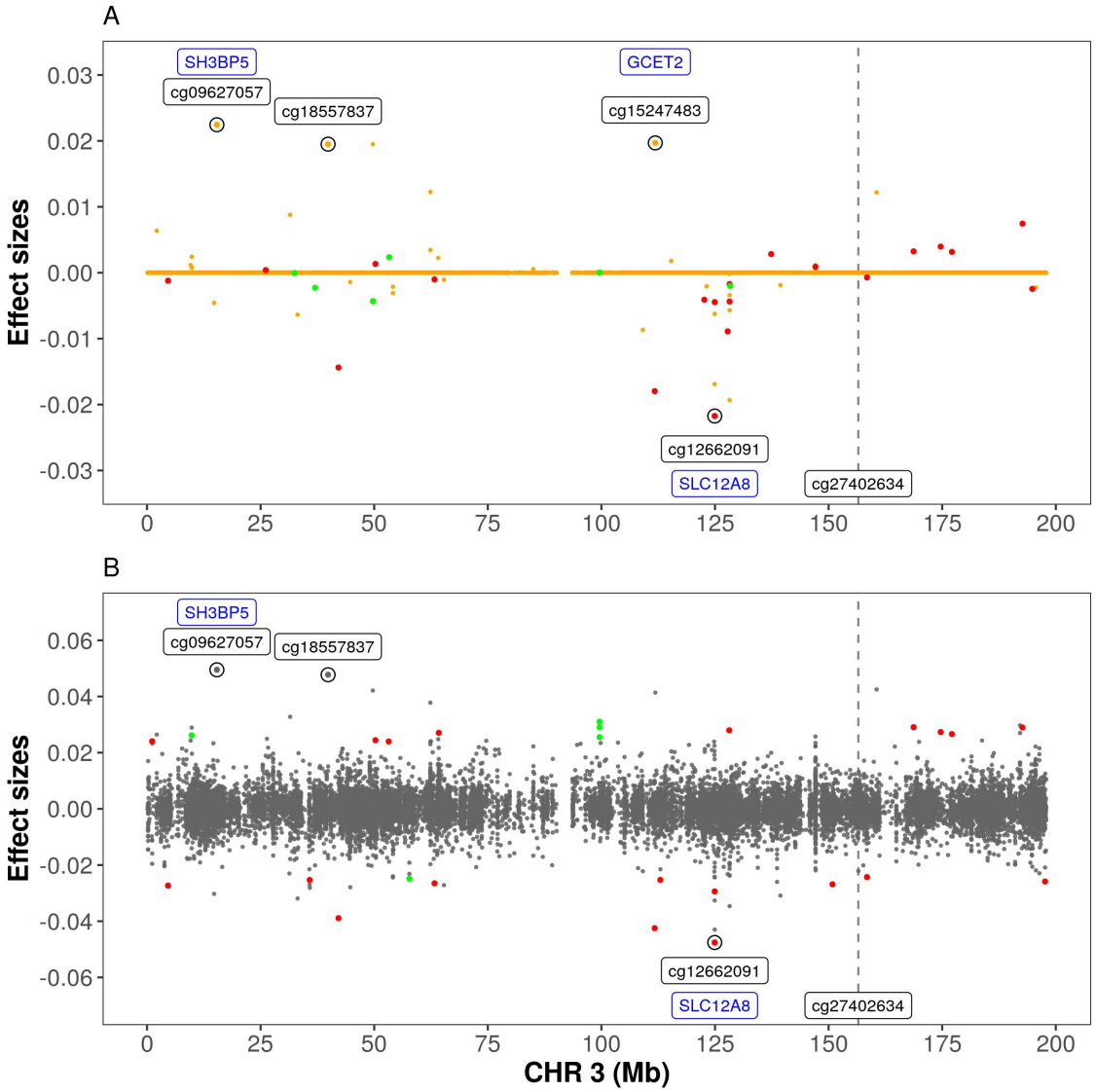

Figure 5. DNA methylation EWAS of smoking status in pregnant women (chromosome 3). A) Estimated effect size for sparse LFMM. The effect size at cg27402634 is equal to $\beta = -0.117$ (out of range). B) Estimated effect size for non-sparse methods (ridge LFMM, CATE and SVA). The effect size at cg27402634 is equal to $\beta = -0.141$ (out of range). CpGs with the highest effects are circled (genes in blue color). Red dots represent CpGs located in enhancer regions. Green dots represent CpGs located in promoter regions (Illumina annotations).

4 Discussion

We introduced sparse latent factor regression methods for the joint estimation of effect sizes and latent factors in genomic and epigenomic association studies. In generative and in empirical simulations, sparse LFMM obtained higher F-score and precision than previously introduced sparse methods, BSLMM and LASSO. Compared to three non-sparse methods (ridge LFMM, CATE and SVA), statistical errors of effect size estimates were reduced. In simulations based on a real data set, sparse LFMM reached the highest precision and F-score, showing that the method was more robust to departure from model assumptions than the other methods. For the causal markers, the effect sizes estimated by sparse LFMM and the corresponding effect sizes estimated by non-sparse methods were strongly correlated. Effect size estimates had a lower bias in non-sparse methods compared to sparse methods. These results suggest to combine sparse LFMM with a non-sparse method in the following way. At a first stage, sparse LFMM can be used to estimate the support of causal markers (non-null effect sizes). Then ridge LFMM and CATE can be used to estimate the effect sizes of the selected markers.

In a GWAS of flowering time using 53,859 SNPs in the fifth chromosome of 162 European accessions of the plant *A. thaliana*, sparse LFMM identified the *FLC* and *DOG1* genes to be associated with the FT16 phenotype. The two genes were previously reported as being associated with this phenotype in (Atwell *et al.*, 2010). The *FLC* gene plays a central role in flowering induced by vernalization (Sheldon *et al.*, 2000), and *DOG1* is involved in the control of dormancy and seed germination (Nishimura *et al.*, 2018). The second hit of sparse LFMM corresponded to SNPs linked to *SAP*, which is a transcriptional regulator involved in the specification of

floral identity (Byzova *et al.*, 1999). This association was also significant for non-sparse methods. In addition, the new method detected SNPs located in the *ACL5* gene, which plays a role in internodal growth and organ size (Hanzawa *et al.*, 1997). In summary, sparse methods facilitated the selection of non-null effect sizes. The results for sparse LFMM were not only consistent with previous discoveries, but they also identified new candidate genes with interesting functional annotations.

Next, we applied sparse LFMM in an EWAS of placental DNA methylation for women exposed to smoking during pregnancy, which is considered an important risk factor for child health (Lumley *et al.*, 2009). The CpG with the highest effect size in sparse LFMM and non-sparse methods (cg27402634) is located in an enhancer region, close to the *LEKR1* gene which was associated with birth weight in a GWAS from the Early Growth Genetics (EGG) consortium (`http://egg-consortium.org/birth-weight.html`). This association was detected as a top hit in an independent study of placental methylation and smoking (Morales *et al.*, 2019). (Rousseaux *et al.*, 2019) also detected the association with cg27402634 in an EWAS based on a slightly different study population, and with other measures of the level of tobacco consumption. (Morales *et al.*, 2019) carried out a Sobel analysis of mediation between smoking and birth weight, found the test significant for cg27402634. In the list of 51 CpGs with high effect sizes, several additional statistical associations between placental methylation and maternal smoking have been reported in previous studies, including cg21992501 in the gene *TTC27* (Cardenas *et al.*, 2019), cg25585967 and cg17823829, respectively in the *TRIO* and *KDM5B* genes (Morales *et al.*, 2019; Everson *et al.*, 2019). Turning to the rest of the methylome, we found additional associations that may adversely affect mother-child health. To

better characterize the CpGs in those associations, we evaluated whether there was an enrichment in enhancer and in promoter regions. We found there was an enrichment of enhancer regions and a depletion of promoter regions for CpGs with non-null effects, consistent with the findings of (Rousseaux *et al.*, 2019). Overall our new method allowed us to confirm some previously discovered associations, and also detected new associations including genes for which methylation changes have detrimental effects on the health of the child.

Conclusion. Removing variation due to unobserved confounding factors is extremely difficult in any type of association study. Assuming that a small proportion of all markers correlate with the exposure or phenotype, we addressed the confounding issue by using sparse latent factor regression models, providing mathematical guarantees that global solutions of least squares estimation problems are proposed. Sparsity constraints in our algorithm allowed the selection of markers without any need for statistical testing. The application of our method to real data sets highlighted new associations with relevant biological meaning. The methods are reproducible and are implemented in the `R` package `lfmm`.

Materials and Methods

Cross-validation method. Choosing regularization parameters of sparse LFMM or ridge LFMM and the number of latent factors can be done by using cross-validation methods. The cross-validation approach partitions the data into a training set and a test set. The training set is used to fit model parameters, and prediction errors are measured on the test set. In our approach, the response and explanatory variables

are partitioned according to their rows (individuals). We denote by I the subset of individual labels on which prediction errors are computed. Estimates of effect sizes, $\hat{\mathbf{B}}_{-I}$, and loading values, $\hat{\mathbf{V}}_{-I}$, are computed on the training set. Next, we partition the set of columns of the response matrix, and denote by J the subset of columns on which prediction errors are computed. A factor matrix, $\hat{\mathbf{U}}_{-J}$, is estimated from the complementary subset as follows

$$\hat{\mathbf{U}}_{-J} = (\mathbf{Y}[I, -J] - \mathbf{X}[I,]\hat{\mathbf{B}}_{-I}^T[-J,])\hat{\mathbf{V}}_{-I}[-J,]. \tag{11}$$

In these notations, the brackets indicate which subsets of rows and columns are selected. A prediction error is then computed as follows

$$\text{Error} = \left\| \mathbf{Y}[I, J] - \hat{\mathbf{U}}_{-J}\hat{\mathbf{V}}_{-I}^T[J,] - \mathbf{X}[I,]\hat{\mathbf{B}}_{-I}^T[J,] \right\|_F. \tag{12}$$

Regularization parameters and the number of factors leading to the lowest prediction error were retained in data analysis.

Heuristics for regularization parameters and number of factors. Additional heuristics were used to determine the number of latent factors and the regularization parameter of the nuclear norm of the latent matrix. In order to choose the number of latent factors, K, we considered the matrix $\mathbf{D}_\lambda$, defined for the ridge algorithm, and the unitary matrix $\mathbf{Q}$, obtained from an SVD of $\mathbf{X}$. The number of latent factors, K, can be estimated by using a spectral analysis of the matrix $\mathbf{D}_0\mathbf{Q}^T\mathbf{Y}$. In our experiments, we used the "elbow" method based on the scree plot of eigenvalues of the matrix $\mathbf{D}_0\mathbf{Q}^T\mathbf{Y}$. Values for K were confirmed by prediction errors computed by cross-validation. The L^1-regularization parameter, μ, was determined by inspection of the proportion of non-zero effect sizes in the $\mathbf{B}$ matrix, which was estimated by

cross-validation. Having set the proportion of non-null effect sizes, μ was computed by using the regularization path approach proposed by Friedman *et al.* (2010) as follows. The regularization path algorithm was initialized with the smallest values of μ such that

$$\hat{\mathbf{B}}_1 = \text{sign}(\bar{\mathbf{B}}_1)(\bar{\mathbf{B}}_1 - \mu)_+ = 0, \tag{13}$$

where $\hat{\mathbf{B}}_1$ resulted from Step 1 in the sparse LFMM algorithm, and $\bar{\mathbf{B}}_1$ is the linear regression estimate. Then, we built a sequence of μ values that decreased from the inferred value of the parameter $\mu^{\max}$ to $\mu^{\min} = \epsilon\mu^{\max}$. We eventually measured the number of non-null elements in $\hat{\mathbf{B}}_t$, and stopped when the target proportion was reached. The nuclear norm parameter (γ) determines the rank of the latent matrix $\mathbf{W}$. We used a heuristic approach to evaluate γ from the number of latent factors K. Based on the singular values $(\lambda_1, \ldots, \lambda_n)$ of the response matrix $\mathbf{Y}$, we set

$$\gamma = \frac{(\lambda_K + \lambda_{K+1})}{2}. \tag{14}$$

With this value of γ, sparse LFMM always converged to a latent matrix estimate having rank K in our experiments.

Estimation algorithms. Sparse LFMM was compared to two other sparse methods. As a baseline, we used Least Absolute Shrinkage and Selection Operator (LASSO) regression models (Tibshirani, 1996; Friedman *et al.*, 2010). LASSO regression models did not include any correction for confounding, and strong biases were expected in effect size estimates. The LASSO models were implemented in the `R` package `glmnet`, and the regularization parameter was selected by using a 5-fold cross validation approach (Zeng *et al.*, 2017). We also used Bayesian Sparse Linear Mixed Models

(BSLMM) implemented in the GEMMA software (Zhou *et al.*, 2013). BSLMM is a hybrid method that combines sparse regression models with linear mixed models. BSLMM uses a Markov chain Monte Carlo (MCMC) method to estimate effect sizes. The MCMC burn-in period and sampling sizes were set to 10,000 (Zeng *et al.*, 2017). To determine the proportion of non-zero effect sizes, two parameters were tuned (pmin and pmax). Those parameters correspond to the logarithm of the maximum and minimum expected proportions of non-zero effect size. We also compared sparse LFMM to three non-sparse algorithms, all based on the generative model defined in equation (1). First we implemented Surrogate Variable Analysis (SVA, Leek and Storey (2007)). SVA was introduced to overcome the problems caused by heterogeneity in gene expression studies. The algorithm starts with estimating the loading values of a principal component analysis for the residuals of the regression of the response matrix $\mathbf{Y}$ on $\mathbf{X}$. In a second step, SVA determines a subset of response variables exhibiting low correlation with $\mathbf{X}$, and uses this subset of variables to estimate the latent factors. SVA was implemented in the `R` package `sva`. Next, we implemented the Confounder Adjusted Testing and Estimation (CATE) method (Wang *et al.*, 2017). CATE uses a linear transformation of the response matrix such that the first axis of this transformation is colinear to $\mathbf{X}$ and the other axes are orthogonal to $\mathbf{X}$. CATE was used without negative controls, and it was implemented in the `R` package `cate`. We eventually used the ridge version of LFMM implemented in the `R` package `lfmm` (Caye *et al.*, 2019).

Generative model simulations. We defined the *confounding intensity* as the percentage of variance of the primary variable $\mathbf{X}$ explained by the latent factors $\mathbf{U}$.

Following Caye *et al.* (2019), we performed simulations of a primary variable, $\mathbf{X}$, with $d = 1$, and $K = 6$ independent latent factors, $\mathbf{U}$, for two values of confounding intensity, $R^2 = 0.1$ (lower) and $R^2 = 0.5$ (higher). The joint distribution of ($\mathbf{X}$, $\mathbf{U}$) was a multivariate Gaussian distribution. Having defined primary variables and latent factors, we used the generative model defined in equation (1) to simulate a response matrix, $\mathbf{Y}$. To create sparse models, only a small proportion of effect sizes, around 0.8%, were allowed to be different from zero. Non-null effect sizes were sampled according to a Gaussian distribution, $\mathrm{N}(B, 0.2)$, where B could take two values, $B = 0.75$ (lower value) and $B = 1.5$ (higher value). Residual errors and loadings, $\mathbf{V}$, were sampled according to a standard Gaussian distribution. The dimensions of the response matrix were set to $n = 400$ individuals and $p = 10,000$ variables. Two hundred simulations were performed for each combination of parameters (800 simulations).

Empirical simulations. We used the R package `naturalgwas` to simulate associations of phenotypes based on a matrix of sampled genotypes (François and Caye, 2018). With this program, phenotypic simulations incorporate realistic features such as geographic population genetic structure and gene-by-environment interactions where environmental variables are derived from a bioclimatic database. When estimating effect sizes, population genetic structure and gene-by-environment interactions are considered to be the main sources of confounding. Phenotypes were simulated for $n = 162$ publicly available Single Nucleotide Polymorphisms (SNPs) genotyped from the fifth chromosome of the model plant *Arabidopsis thaliana* (Atwell *et al.*, 2010). The response matrix contained $p = 53,859$ SNPs, with minor allele fre-

quency greater than 5%. The number of confounding factors was set to $K = 6$, and the phenotypes were generated from a combination of five causal SNPs with identical effect sizes. Two values of effect size were implemented, $B = 6$ (lower effect size) and $B = 9$ (higher effect size). Additionnally, two values of gene-by-environment interaction were implemented, $G \times E = 0.1$ (lower $G \times E$) and $G \times E = 0.9$ (higher $G \times E$). For each parameter combination, two hundred simulations were performed.

Evaluation metrics. In the simulation study, all methods were used with their default parameters, and the number of latent factors was set to $K = 6$ in all latent factor models. To evaluate the capabilities of methods to identify true positives, we used precision, which corresponds to the proportion of true positives in a list of positive markers, the recall, which is the number of true positives divided by the number of causal markers, and the F-score, which is the harmonic mean of precision and recall. To compute precision and F-score in generative model experiments, a list of 100 markers with the largest absolute estimated effect sizes was considered for each data set and method. In empirical simulations, the measures were modified to account for linkage disequilibrium (LD) in the data. Candidate markers within a window of size 10kb around a causal marker were considered to be true discoveries (LD-$r^2 < 0.2$, François and Caye (2018)). In addition to the F-score, we used the root mean squared error (RMSE) to evaluate the statistical errors of effect size estimates. We also used simulations from the generative model to assess the capability of the cross validation algorithm to estimate the number of latent factors in sparse LFMM. In program runs, the number of latent factors varied between $K = 3$ and $K = 10$ and the value estimated by the cross validation algorithm was compared with the

true value ($K = 6$).

GWAS of plant phenotype. Sparse LFMM and a set of other methods were used to perform association studies for two distinct types of genomic data including genotypic and epigenetic markers. For *Arabidopsis thaliana*, we considered $n = 162$ European accessions and $p = 53,859$ SNPs from the fifth chromosome ot the plant genome to investigate associations with the flowering time phenotype FT16-TO: 0000344 (Atwell *et al.*, 2010). FT16 corresponds to the number of days required for an individual plant to reach the flowering stage. In the sparse LFMM algorithm, the percentage of non-null effect size was set to 0.01. The parameters pmin and pmax defining sparsity in the BSLMM algorithm were fixed to pmin$= -5$ and pmax$= -4$ respectively. These values correspond to the logarithm of expected proportions of non-null effect sizes in BSLMM. For all factor methods, the number of latent factors was determined by cross-validation and set to $K = 10$.

EWAS of exposure to tobacco consumption. Our second application to real data concerned an EWAS based on the EDEN mother-child cohort (Heude *et al.*, 2016). Beta-normalized methylation levels at $p = 425,878$ probed CpG sites were measured for $n = 668$ women. We tested the association between smoking status (219 current smokers women and 449 non-current smokers women) and DNA methylation (mDNA) levels in the mother's placenta. Detailed information on the study population and protocols for placental DNA methylation assessment processing could be found in (Abraham *et al.*, 2018; Rousseaux *et al.*, 2019). The proportion of null effect sizes in sparse LFMM was equal to 0.999. For latent factor models, the number of latent factors was estimated by cross-validation, and was equal to $K = 7$.

Acknowledgements. This article was developed in the framework of the Grenoble Alpes Data Institute, supported by the French National Research Agency under the Investissements d'Avenir program (ANR-15-IDEX-02). It received support from LabEx PERSYVAL Lab, ANR-11-LABX-0025-01, and from the French National Research Agency (Agence Nationale pour la Recherche) ETAPE, ANR-18-CE36-0005. We thank the participants of the EDEN cohort. We thank the midwife research assistants for data collection, the psychologists and the data entry operators. We also thank the EDEN mother-child cohort study group which includes I Annesi-Maesano, JY Bernard, J Botton, M-A Charles, P Dargent- Molina, B de Lauzon- Guillain, P Ducimetière, M de Agostini, B Foliguet, A Forhan, X Fritel, A Germa, V Goua, R Hankard, B Heude, M Kaminski, B Larroque, N Lelong, J Lepeule, G Magnin, L Marchand, C Nabet, F Pierre, R Slama, MJ Saurel-Cubizolles, M Schweitzer, O Thiebaugeorges.

Program Availability. All codes are publicly available. Sparse LFMM was implemented in the `R` package `lfmm` available from Github (`https://bcm-uga.github.io/lfmm/`) and submitted to the Comprehensive R Archive Network (`https://cran.r-project.org/`).

Data Availability. The *Arabidopsis thaliana* data are publicly available from the 1,001 genomes database (`https://1001genomes.org/`). The EDEN individual-level data have restricted access owing to ethical and legal conditions in France. They are available upon request from the EDEN steering committee at `etude.eden@inserm.fr` and through collaborations with the principal investigators of EDEN.

Fundings. The EDEN study was supported by Foundation for medical research (FRM), National Agency for Research (ANR), National Institute for Research in Public health (IRESP: TGIR cohorte santé 2008 program), French Ministry of Health (DGS), French Ministry of Research, INSERM Bone and Joint Diseases National Research (PRO-A), and Human Nutrition National Research Programs, Paris-Sud University, Nestlé, French National Institute for Population Health Surveillance (InVS), French National Institute for Health Education (INPES), the European Union FP7 programmes (FP7/2007-2013, HELIX, ESCAPE, ENRIECO, Medall projects), Diabetes National Research Program (through a collaboration with the French Association of Diabetic Patients (AFD)), French Agency for Environmental Health Safety (now ANSES), Mutuelle Générale de l'Education Nationale a complementary health insurance (MGEN), French national agency for food security, French-speaking association for the study of diabetes and metabolism (ALFEDIAM).

References

Abraham, E., Rousseaux, S., Agier, L., Giorgis-Allemand, L., Tost, J., Galineau, J., Hulin, A., Siroux, V., Vaiman, D., Charles, M.-A., Heude, B., Forhan, A., Schwartz, J., Chuffart, F., Bourova-Flin, E., Khochbin, S., Slama, R., and Lepeule, J., (2018). Pregnancy exposure to atmospheric pollution and meteorological conditions and placental DNA methylation. *Environ. Int.*, **118**, 334-347.

Akama, T.O., Misra, A.K., Hindsgaul, O., and Fukuda, M.N. (2002). Enzymatic synthesis in vitro of the disulfated disaccharide unit of corneal keratan sulfate. *J. Biol. Chem.*, **277**, 42505-42513.

Atwell, S., Huang, Y.S., Vilhjàlmsson, B.J., Willems, G., Horton, M., Li, Y., Meng, D., Platt, A., Tarone, A.M., Hu, T.T., Jiang, R., Muliyati, N.W., Zhang, X., Amer, M.A., Baxter, I., Brachi, B., Chory, J., Dean, C., Debieu, M., de Meaux, J., Ecker, J.R., Faure, N., Kniskern, J.M., Jones, J.D.G., Michael, T., Nemri, A., Roux, F., Salt, D.E., Tang, C., Todesco, M., Traw, M.B., Weigel, D., Marjoram, P., Borevitz, J.O., Bergelson, J., and Nordborg, M. (2010). Genome-wide association study of 107 phenotypes in *Arabidopsis thaliana* inbred lines. *Nature*, **465**, 627-631.

Balding, D.J. (2006) A tutorial on statistical methods for population association studies. *Nat. Rev. Genet.*, **7**, 781-781.

Bertsekas, D. P. (1999) *Nonlinear Programming*. Belmont: Athena Scientific.

Byzova, M.V., Franken, J., Aarts, M.G.M., de Almeida-Engler, J., Engler, G., Mariani, C., Van Lookeren Campagne, M.M., Angenent, G.C. (1999). Arabidopsis STERILE APETALA, a multifunctional gene regulating inflorescence, flower, and ovule development. *Genes Dev.*, **13**, 1002-1014.

Cai, J-F., Candès, E.J. and Shen, Z. (2010) A singular value thresholding algorithm for matrix completion. *SIAM J. Optim.*, **20** 1956-1982.

Carvalho, C. M. *et al.* (2008) High-dimensional sparse factor modeling: applications in gene expression genomics. *J. Am. Stat. Assoc.*, **103**, 1438-1456.

Cardenas, A., Lutz, S.M., Everson, T.M., Perron, P., Bouchard, L., Hivert, M.-F., (2019). Mediation by placental DNA methylation of the association of prenatal maternal smoking and birth weight. *Am. J. Epidemiol.*, **188**, 1878-1886.

Caye, K., Jumentier, B., Lepeule, J., François, O. (2019) LFMM 2: Fast and accurate inference of gene-environment associations in genome-wide studies. *Mol. Biol. Evol.*, **36**, 852-860.

Devlin, B. and Roeder K. (1999) Genomic control for association studies. *Biometrics*, **55**, 997-1004.

Eckart, C. and Young, G. (1936) The approximation of one matrix by another of lower rank. *Psychometrika*, **1**, 211-218.

Efron, B. (2004) Large-scale simultaneous hypothesis testing: The choice of a null hypothesis. *J. Am. Stat. Assoc.*, **99**, 96-104.

Everson, T.M., Vives-Usano, M., Seyve, E., Cardenas, A., Lacasaña, M., Craig, J.M., Lesseur, C., Baker, E.R., Fernandez-Jimenez, N., Heude, B., Perron, P., Gonzalez-Alzaga, B., Halliday, J., Deyssenroth, M.A., Karagas, M.R., Iñiguez, C., Bouchard, L., Carmona-Saez, P., Loke, Y.J., Hao, K., Belmonte, T., Charles, M.A., Martorell-Marugan, J., Muggli, E., Chen, J., Fernandez, M.F., Tost, J., Gomez-Martin, A., London, S.J., Sunyer, J., Marsit, C.J., Lepeule, J., Hivert, M.-F., Bustamante, M., (2019). Placental DNA methylation signatures of maternal smoking during pregnancy and potential impacts on fetal growth. *BioRxiv*, 663567.

François, O., Caye, K. (2018) Naturalgwas: An R package for evaluating genome-wide association methods with empirical data. *Mol. Ecol. Resour.*, **18**(4), 789-797.

Friedman, J., Hastie, T., Höfling, H., Tibshirani, R. (2007) Pathwise coordinate optimization. *Ann. Appl. Stat.*, **1**, 302-332.

Friedman, J., Hastie, T., Tibshirani, T. (2010) Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.*, **33**.

Frichot, E., Schoville, S. D., Bouchard, G. and François, O. (2013) Testing for associations between loci and environmental gradients using latent factor mixed models. *Mol. Biol. Evol.*, **30**, 1687-1699.

Frichot, E. and François, O. (2015) LEA: an R package for landscape and ecological association studies. *Methods Ecol. Evol.*, **6**, 925-929.

Gautier, M. (2015) Genome-wide scan for adaptive divergence and association with population-specific covariates. *Genetics*, **201**, 1555-1579.

Halko, N., Martinsson, P. G. and Tropp, J. A. (2011) Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Rev.*, **53**, 217-288.

Hanzawa, Y., Takahashi, T., and Komeda, Y. (1997). ACL5: an Arabidopsis gene required for internodal elongation after flowering. *Plant J.*, **12**, 863-874.

Hastie, T., Tibshirani, R., and Friedman, J. (2009) *The Elements of Statistical Learning*. Springer Series in Statistics, Springer, NY, USA.

Heude, B., Forhan, A., Slama, R., Douhaud, L., Bedel, S., Saurel-Cubizolles, M.-J., Hankard, R., Thiebaugeorges, O., De Agostini, M., Annesi-Maesano, I., Kaminski, M., and Charles, M.-A. (2016). Cohort Profile: The EDEN mother-child cohort on the prenatal and early postnatal determinants of child health and development. *Int. J. Epidemiol.*, **45**, 353-363.

Hoggart, C.J., Whittaker, J.C., Iorio, M.D., and Balding, D.J.(2008) Simultaneous analysis of all SNPs in genome-wide and re-sequencing association studies. *PLoS Genet.*, **4**, e1000130.

Houseman, E.A., Kile, M.L., Christiani, D.C., Ince, T.A., Kelsey, K.T., and Marsit, C.J. (2016). Reference-free deconvolution of DNA methylation data and mediation by cell composition effects. *BMC Bioinformatics*, **17**.

Jaffe, A. E. and Irizarry, R. A. (2014) Accounting for cellular heterogeneity is critical in epigenome-wide association studies. *Genome Biol.*, **15**, R3.

Kalaitzis, A.A., and Lawrence, N.D. (2012) Residual component analysis: Generalising PCA for more flexible inference in linear-Gaussian models. Proceedings of the 29th International Conference on Machine Learning, *ICML 2012*, **1**, 209-216.

Kaushal, A. *et al.* (2017) Comparison of different cell type correction methods for genome-scale epigenetics studies. *BMC Bioinformatics*, **18**, 216.

Leek, J. T., and Storey, J. D. (2007) Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet.*, **3**, e161.

Lumley, J., Chamberlain, C., Dowswell, T., Oliver, S., Oakley, L., and Watson, L., (2009). Interventions for promoting smoking cessation during pregnancy. *Cochrane Database of Systematic Reviews 2009*, **Issue 3**. Art. No.: CD001055.

Mishra, B., Meyer, G., Bach, F., and Sepulchre, R. (2013) Low-rank optimization with trace norm penalty. *SIAM J. Optim.*, **23**, 2124-2149.

Morales, E., Vilahur, N., Salas, L.A., Motta, V., Fernandez, M.F., Murcia, M., Llop, S., Tardon, A., Fernandez-Tardon, G., Santa-Marina, L., Gallastegui, M., Bollati, V., Estivill, X., Olea, N., Sunyer, J., Bustamante, M., (2016). Genome-wide DNA methylation study in human placenta identifies novel loci associated with maternal smoking during pregnancy. *Int. J. Epidemiol.*, **45**, 1644-1655.

Nishimura, N., Tsuchiya, W., Moresco, J.J., Hayashi, Y., Satoh, K., Kaiwa, N., Irisa, T., Kinoshita, T., Schroeder, J.I., Yates, J.R., Hirayama, T., Yamazaki, T. (2018). Control of seed dormancy and germination by DOG1-AHG1 PP2C phosphatase complex via binding to heme. *Nat. Commun.*, **9**.

Price, A.L. *et al.* (2006) Principal component analysis corrects for stratification in genome-wide association studies. *Nat. Genet.*, **38**, 904-909.

Rakyan, V. K., Down, T. A., Balding, D. J. and Beck, S. (2011) Epigenome-wide association studies for common human diseases. *Nat. Rev. Genet.*, **12**, 529-541.

Rellstab, C., Gugerli, F., Eckert, A. J., Hancock, A. M., and Holderegger, R. (2015) A practical guide to environmental association analysis in landscape genomics. *Mol. Ecol.*, **24**, 4348-4370.

Rousseaux, S., Seyve, E., Chuffart, F., Bourova-Flin, E., Benmerad, M., *et al.* (2019). Maternal exposure to cigarette smoking induces immediate and durable changes in placental DNA methylation affecting enhancer and imprinting control regions. *BioRxiv*, 852186.

Sayin, N., Kara, N., Pekel, G., and Altinkaynak, H. (2014). Effects of chronic smoking

on central corneal thickness, endothelial cell, and dry eye parameters. *Cutan. Ocul. Toxicol.*, **33**, 201-205.

Sheldon, C.C., Rouse, D.T., Finnegan, E.J., Peacock, W.J., Dennis, E.S. (2000). The molecular basis of vernalization: The central role of Flowering Locus C (FLC). *Plant Biol.*, **97**, 6.

Tibshirani, R. (1996) Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B*, **58** 267-288.

The BIOS Consortium, van Iterson, M., van Zwet, E.W., Heijmans (2017) Controlling bias and inflation in epigenome-and transcriptome-wide association studies using the empirical null distribution. *Genome Biol.*, **18**, 19.

Tseng, P. (2001) Convergence of a block coordinate descent method for nondifferentiable minimization. *J. Optim. Theor. Appl.*, **109**, 475-494.

Wang, J., Zhao, Q., Hastie, T., Owen, A.B. (2017) Confounder adjustment in multiple hypothesis testing. *Ann. Statist.*, **45**, 1863-1894.

Witten, D. M., Tibshirani, R., and Hastie, T. (2009) A penalized matrix decomposition with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, **10**(3), 515-534.

Wu, T.T., Chen, Y.F., Hastie, T., Sobel, E., and Lange, K. (2009) Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics*, 25, 714-721.

Yu, J., Pressoir, G., Briggs, W.H., Bi, I.V., Yamasaki, M., *et al.* (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature Genetics*, 38, 203-208.

Zeng, P., Zhou, X., Huang, S., (2017). Prediction of gene expression with cis-SNPs using mixed models and regularization methods. *BMC Genomics*, **18**.

Zhou, X., Stephens, M. (2012) Genome-wide efficient mixed-model analysis for association studies. *Nature Genetics*, **44**, 821.

Zhou, X., Carbonetto, P., and Stephens, M. (2013) Polygenic modeling with Bayesian sparse linear mixed models. *PLoS Genetics*, **9**(2), e1003264.

Zou, H., Hastie, T., and Tibshirani, R. (2006) Sparse principal component analysis. *Journal of Computational and Graphical Statistics*, **15**, 265-286.

Appendix: Proofs of theorems

This section provides mathematical proofs for the theorems stated in section 2.

Theorem 1. *Let $\mu > 0$ and $\gamma > 0$. Then the block-coordinate descent algorithm cycling through Step 1 and Step 2 converges to estimates of* $\mathbf{W}$ *and* $\mathbf{B}$ *defining a global minimum of the penalized loss function* $\mathcal{L}_{\text{sparse}}(\mathbf{W}, \mathbf{B})$.

Proof. The proof arguments are based on a result of Tseng (2001). Consider the Cartesian product of closed convex sets $A = A_1 \times A_2 \times ... \times A_m$, and let $f(\mathbf{z})$ be a continuous convex function defined on A and such that

$$f(z_1, \cdots, z_m) = g(z_1, \cdots, z_m) + \sum_{i=1}^{m} f_i(z_i)\,,$$

where $g(\mathbf{z})$ is a differentiable convex function, and for each $i = 1, \ldots, m$, $f_i(z_i)$ is a continuous convex function. Let $(\mathbf{z}^t)$ be the sequence of values defined by the following block-coordinate descent algorithm

$$z_i^{t+1} \in \arg\min_{\zeta \in A_i} f(z_1^t, \ldots, z_{i-1}^t, \zeta, z_{i+1}^t, \ldots, z_m^t)\,, \quad i = 1, \ldots, m. \tag{15}$$

Then a limit point of the sequence $(\mathbf{z}^t)$ defines a global minimum of the function $f(\mathbf{z})$. The theorem's proof is a consequence of the convexity of the penalized loss function $\mathcal{L}_{\text{sparse}}(\mathbf{W}, \mathbf{B})$, and the fact that we can write

$$\mathcal{L}_{\text{sparse}}(\mathbf{B}, \mathbf{W}) = g(\mathbf{B}, \mathbf{W})/2 + f_1(\mathbf{B}) + f_2(\mathbf{W})$$

where $g(\mathbf{B}, \mathbf{W}) = \|\mathbf{Y} - \mathbf{W} - \mathbf{X}\mathbf{B}^T\|_F^2$ is a differentiable convex function, and $f_1(\mathbf{B}) = \|\mathbf{B}\|_1^2$, $f_2(\mathbf{W}) = \|\mathbf{W}\|_*^2$ are continuous convex functions. Tseng's result can be applied with the function $f(\mathbf{B}, \mathbf{W}) = \mathcal{L}_{\text{sparse}}(\mathbf{B}, \mathbf{W})$ to conclude the proof (see also (Bertsekas, 1999)).

□

Theorem 2. *Let $\lambda > 0$ and assume $\sigma_i^2 > 0$ for all $i = 1, \ldots, d$. The estimates $\hat{\mathbf{W}}$ and $\hat{\mathbf{B}}$ computed as follows*

$$\hat{\mathbf{W}} = \mathbf{Q}\mathbf{D}_\lambda^{-1}\text{svd}_K(\mathbf{D}_\lambda \mathbf{Q}^T \mathbf{Y}) \tag{16}$$

$$\hat{\mathbf{B}}^T = (\mathbf{X}^T\mathbf{X} + \lambda \mathbf{Id}_d)^{-1}\mathbf{X}^T(\mathbf{Y} - \hat{\mathbf{W}}), \tag{17}$$

where $\mathrm{svd}_K(\mathbf{A})$ *is the rank* K *SVD of the matrix* $\mathbf{A}$*,* $\mathbf{Id}_d$ *is the* $d \times d$ *identity matrix, and* $\mathbf{D}_\lambda$ *is the* $n \times n$ *diagonal matrix with coefficients defined as*

$$\mathbf{d}_\lambda = \left(\sqrt{\frac{\lambda}{\lambda + \sigma_1^2}}, \ldots, \sqrt{\frac{\lambda}{\lambda + \sigma_d^2}}, 1, \ldots, 1 \right).$$

define a global mimimum of the penalized loss function $\mathcal{L}_{\mathrm{ridge}}(\mathbf{B}, \mathbf{W})$*.*

Proof. Given $\mathbf{W}$, a global minimum for $\mathcal{L}_{\mathrm{ridge}}(\mathbf{B}, \mathbf{W})$ is obtained with the ridge estimates for a linear regression of the response matrix $\mathbf{Y} - \mathbf{W}$ on $\mathbf{X}$.

$$\hat{\mathbf{B}}^T = (\mathbf{X}^T\mathbf{X} + \lambda \mathbf{Id}_d)^{-1}\mathbf{X}^T(\mathbf{Y} - \mathbf{W}). \tag{18}$$

Thus, the problem amounts to minimizing the function $\mathcal{L}(\mathbf{W}) = \mathcal{L}_{\mathrm{ridge}}(\hat{\mathbf{B}}, \mathbf{W})$ with respect to $\mathbf{W}$. By definition of the $\mathbf{D}_\lambda$ and $\mathbf{Q}$ matrices, the loss function rewrites as

$$\mathcal{L}(\mathbf{W}) = \left\| \mathbf{D}_\lambda \mathbf{Q}^T(\mathbf{Y} - \mathbf{W}) \right\|_F^2 . \tag{19}$$

Minimizing the above loss function is equivalent to finding the best approximation of rank K for the matrix $\mathbf{D}_\lambda \mathbf{Q}^T \mathbf{Y}$. According to Eckart and Young (1936), this approximation is given by the rank K singular value decomposition of $\mathbf{D}_\lambda \mathbf{Q}^T \mathbf{Y}$. Eventually we obtain that

$$\hat{\mathbf{W}} = \mathbf{Q}\mathbf{D}_\lambda^{-1}\mathrm{svd}_K(\mathbf{D}_\lambda \mathbf{Q}^T \mathbf{Y}) \tag{20}$$

defines the unique global minimum of the $\mathcal{L}(\mathbf{W})$ function. □

Supplementary materials

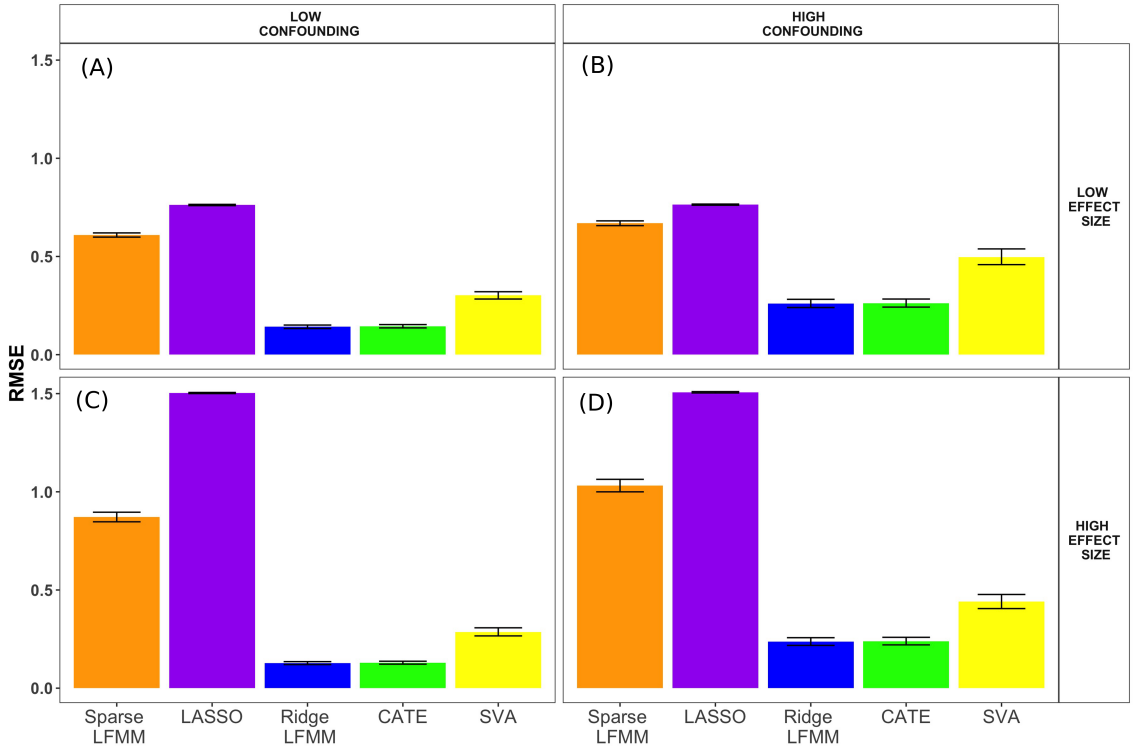

Figure S1. Generative model simulations (RMSE for causal markers only). Root Mean Square Error (RMSE) of causal effect sizes as a function of the effect size of the causal markers and of the confounding intensity. Two sparse methods (sparse LFMM, LASSO) and three non-sparse methods (ridge LFMM, CATE and SVA) were compared. Simulation parameters: (A) Lower effect sizes and confounding intensities (B) Lower effect sizes and higher confounding intensities. (C) Higher effect sizes and lower confounding intensities. (D) Higher effect sizes and confounding intensities.

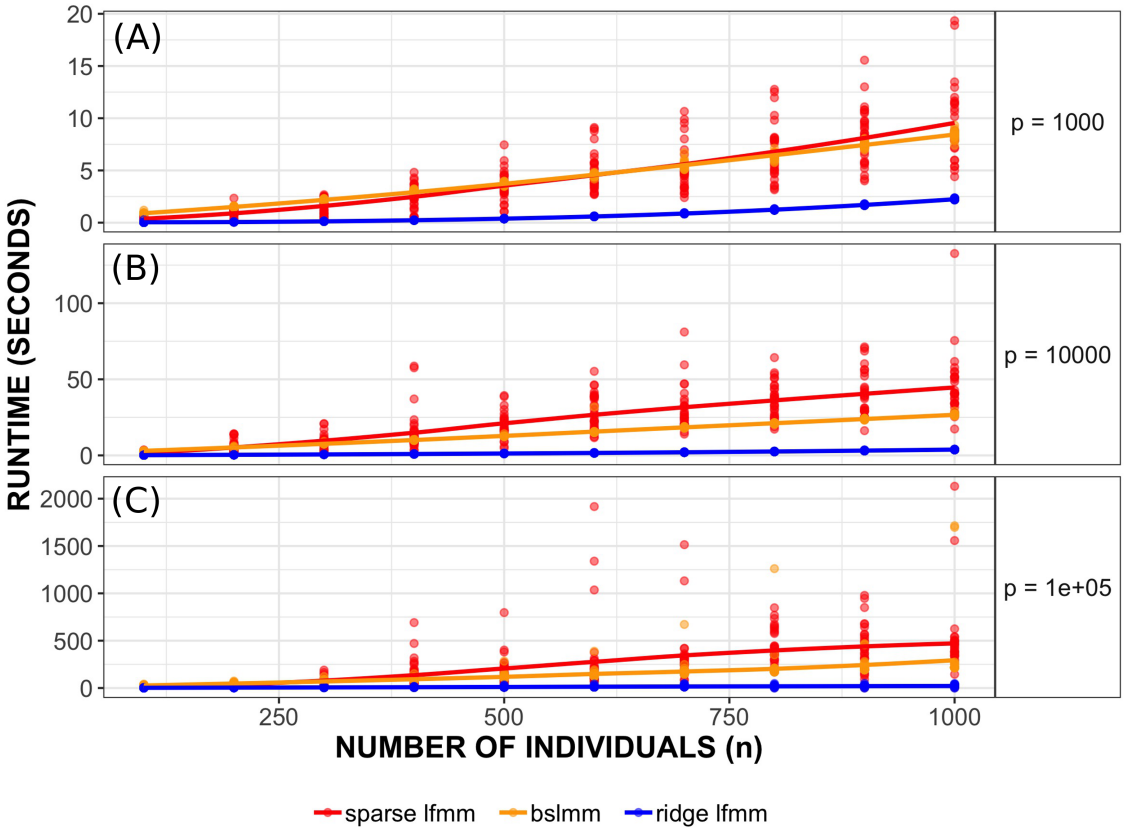

Figure S2. Comparison of the runtimes of three methods. Runtimes as a function of the number of markers (p) and the number of individuals (n). (A) $p = 1000$. (B) $p = 10,000$. (C) $p = 100,000$.

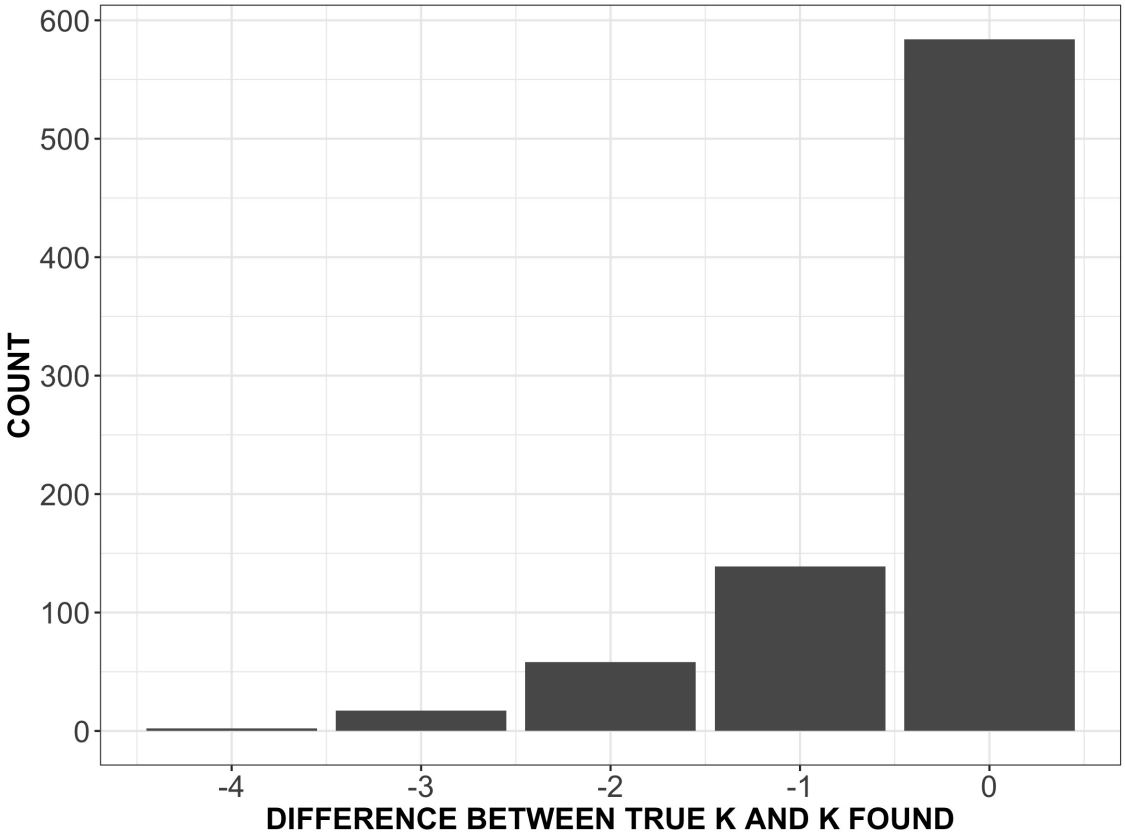

Figure S3. Estimation of the number of latent factors (K) in the generative simulations. Difference between the true K of the simulations and the K estimated by our cross validation algorithm.

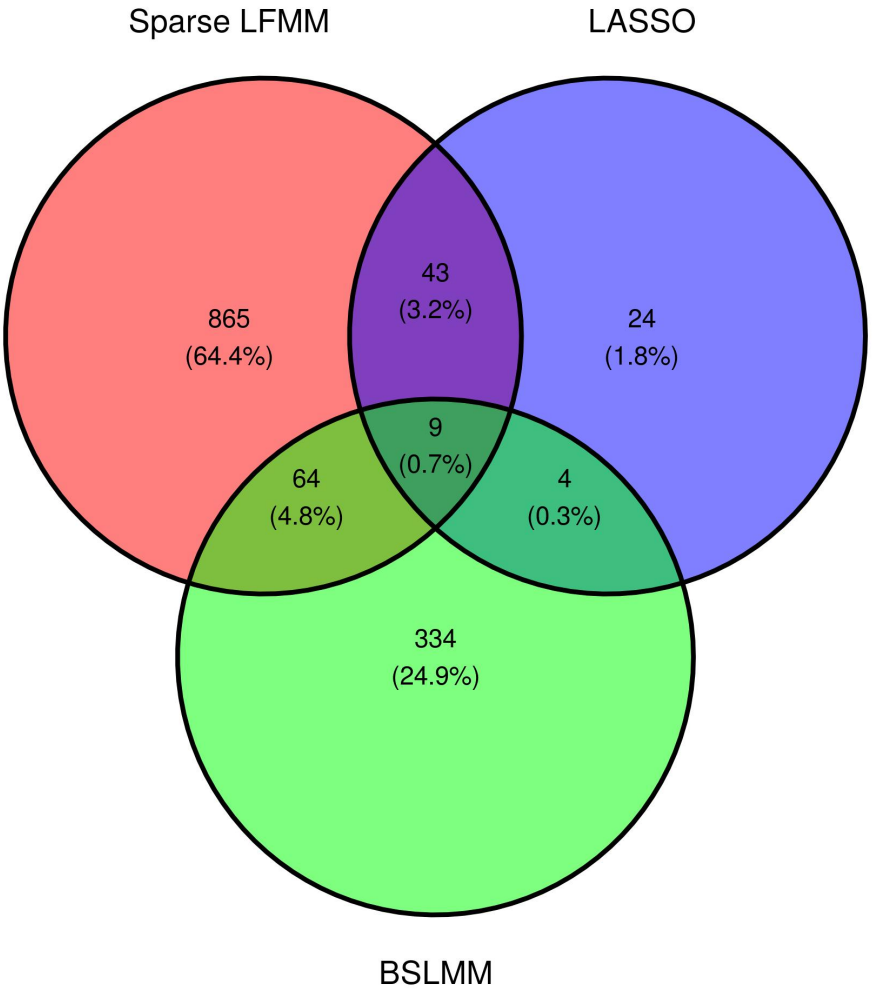

Figure S4. GWAS of a flowering trait with sparse LFMM, LASSO and BSLMM. Venn diagram of SNPs associated with the FT16 phenotype in each approach. The hits correspond to SNPs having non-null effect size estimates.

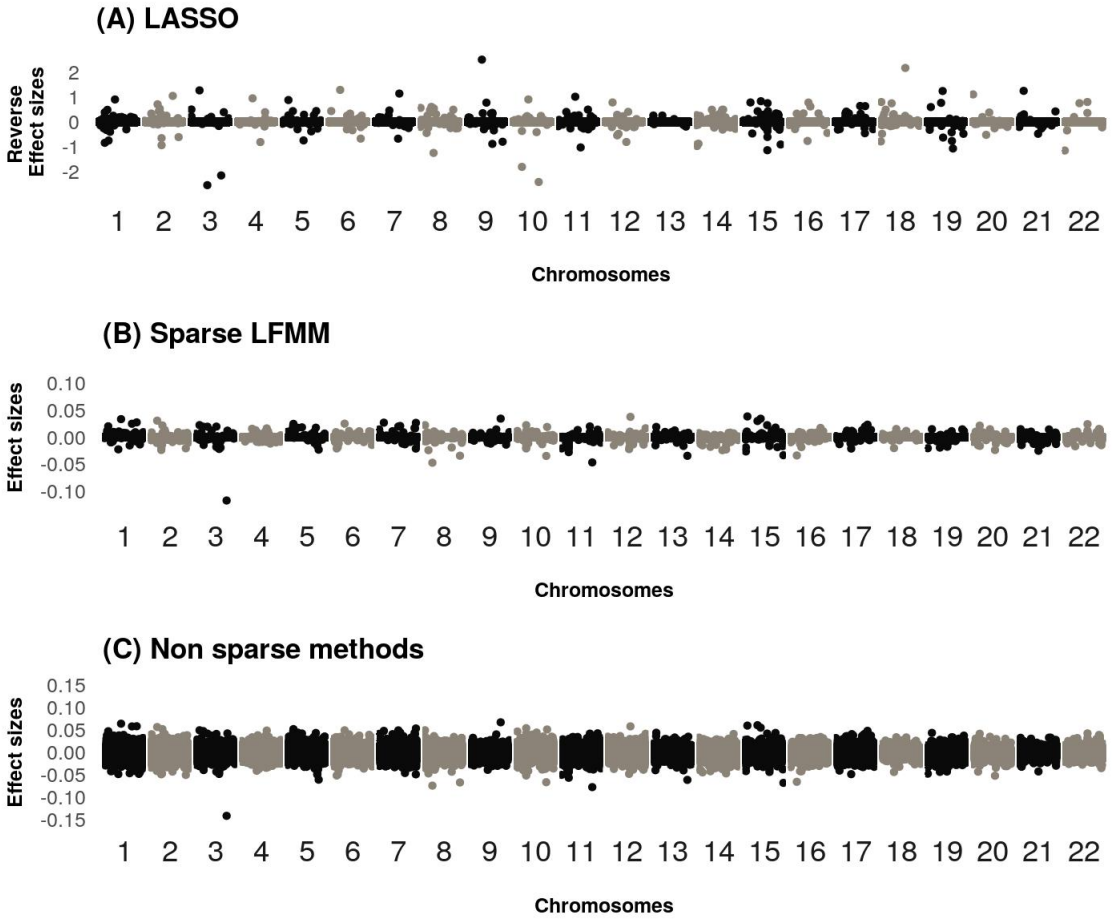

Figure S5. DNA methylation EWAS of smoking status in pregnant women (all chromosomes). A) Estimated reverse effect size for LASSO. B) Estimated effect size for sparse LFMM. C) Estimated effect size for non-sparse methods (ridge LFMM, CATE and SVA).

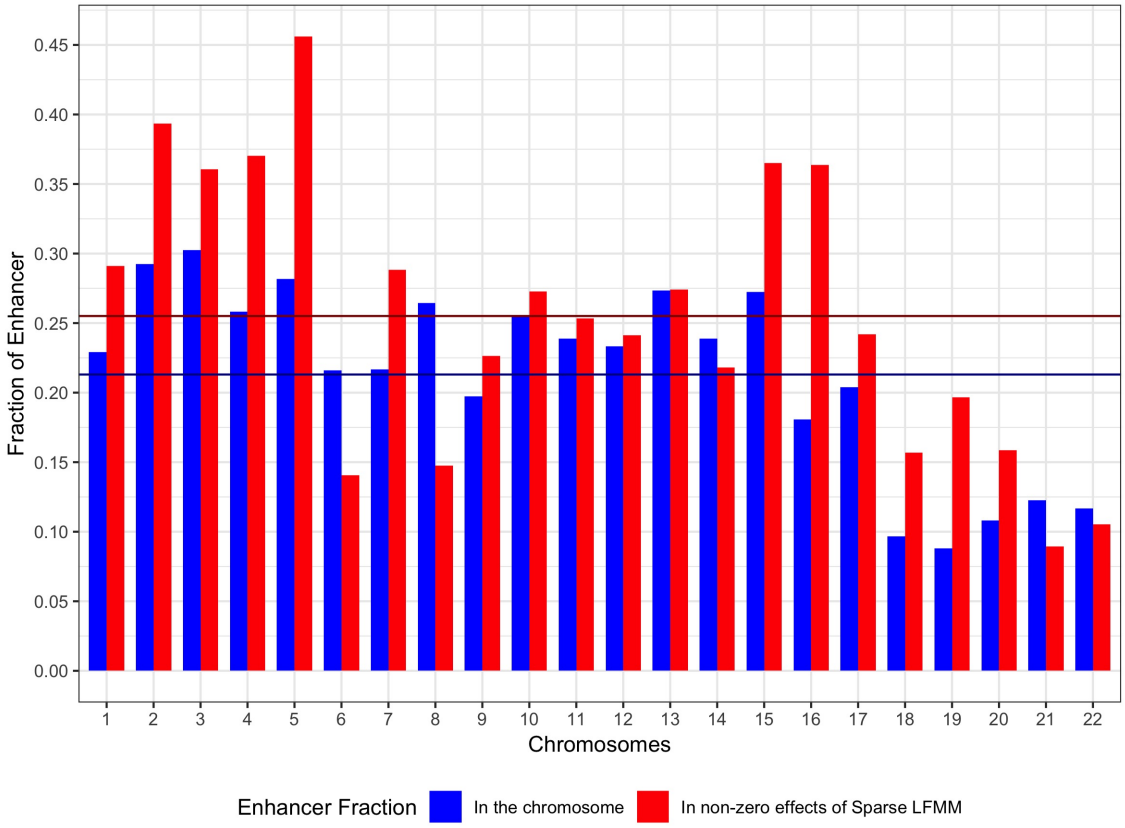

Figure S6. EWAS of smoking status in women. Over-representation of enhancer regions in sparse LFMM candidate regions compared to the methylome. Blue bars correspond to the fraction of enhancer regions in each chromosome. Red bars correspond to the fraction of enhancer regions detected by sparse LFMM. The horizontal blue line represent the average number of enhancer regions per chromosome for the methylome. The red line represents the average number for sparse LFMM.

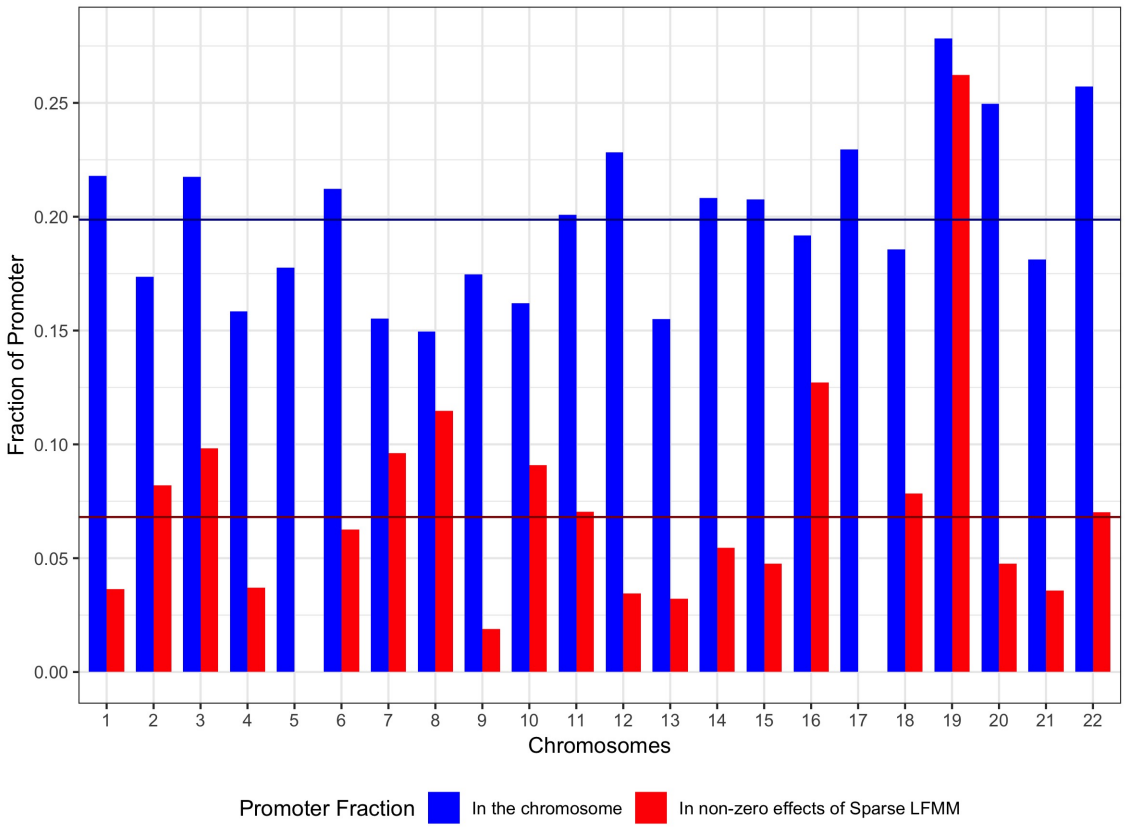

Figure S7. EWAS of smoking status in women. Under-representation of promoter regions in sparse LFMM candidate regions compared to the methylome. Blue bars correspond to the fraction of promoter regions in each chromosome. Red bars correspond to the fraction of promoter regions detected by sparse LFMM. The horizontal blue line represent the average number of promoter regions per chromosome for the methylome. The red line represents the average number for sparse LFMM.

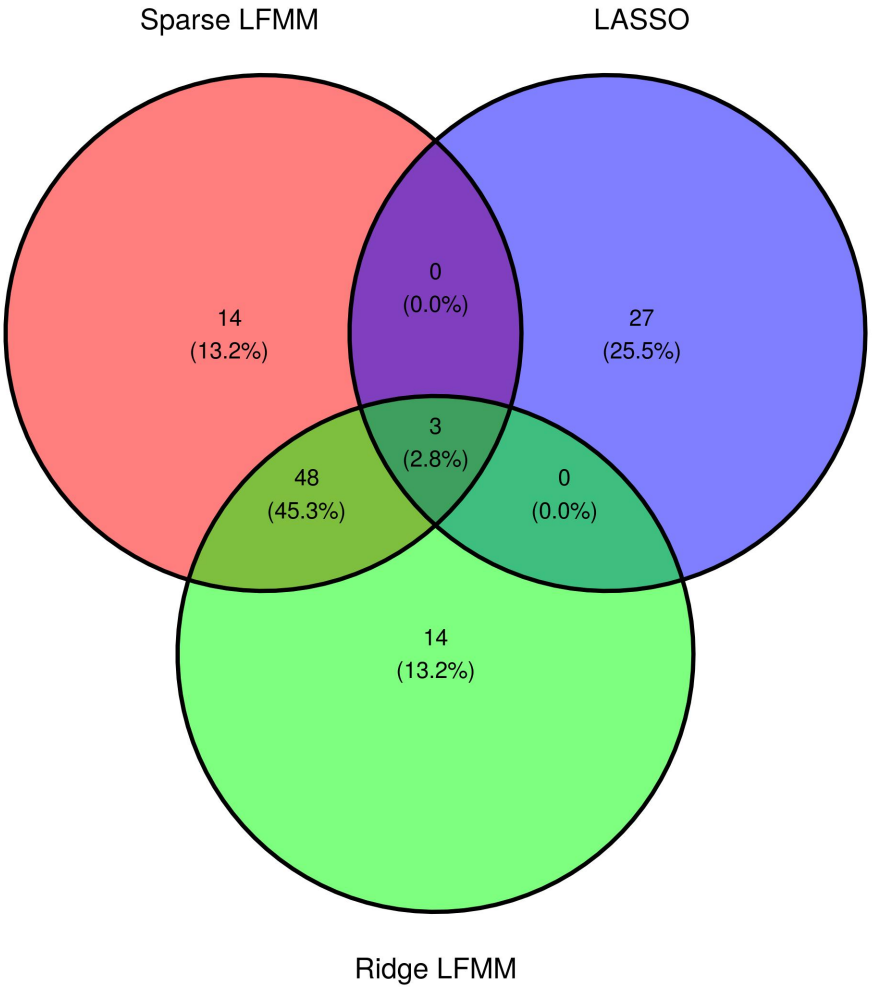

Figure S8. EWAS of smoking status in women with sparse LFMM, LASSO and Ridge LFMM. Venn diagram of CpGs associated with tobacco consumption in each approach. For sparse LFMM, the hits correspond to the 5% of the highest non-null effect sizes (65 hits). For LASSO, the hits correspond to the 5% of the highest non-null effect sizes (30 hits). For ridge LFMM, the hits correspond to the 65 CpGs having the highest effect sizes.

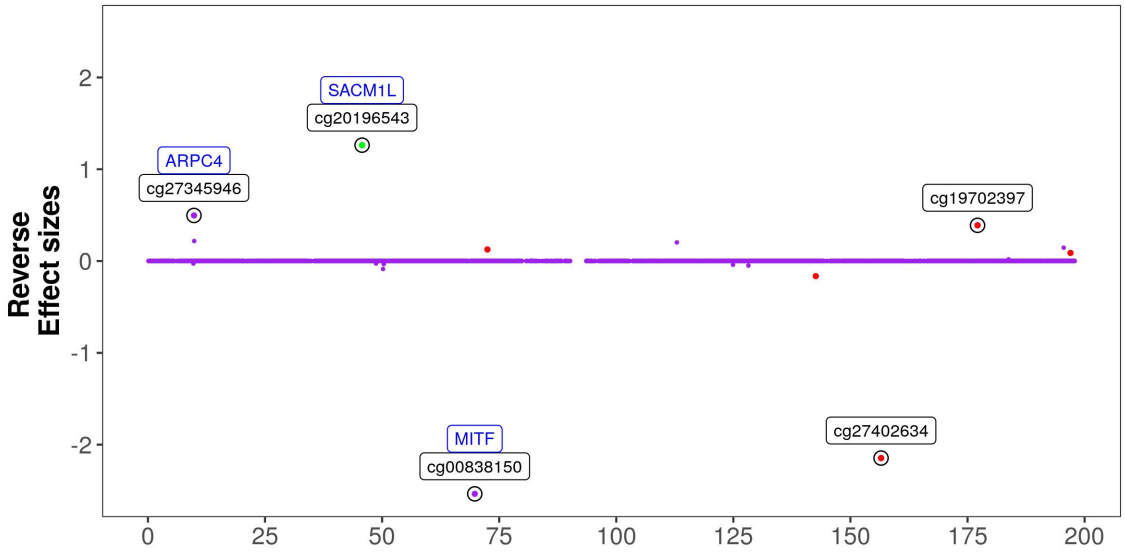

Figure S9. Estimated reverse effect size for LASSO in an EWAS of smoking status in pregnant women (chromosome 3). The 5 CpGs having the highest reverse effect size are circled (the genes associated with these CpGs are in blue color). Red dots represent CpGs located in enhancer regions. Green dots represent CpGs located in promoter regions (Illumina annotations).