BIOINFORMATICS

Vol. 00 no. 00 2005
Pages 1–6

Supplementary Material for Assessing multivariate gene-metabolome associations with rare variants using Bayesian reduced rank regression

Pekka Marttinen[1,2], Matti Pirinen[3], Antti-Pekka Sarin[3,4], Jussi Gillberg[1], Johannes Kettunen[3,4], Ida Surakka[3,4], Antti J. Kangas[5], Pasi Soininen[5,6], Paul O'Reilly[7], Marika Kaakinen[8,9], Mika Kähönen[10], Terho Lehtimäki[11], Mika Ala-Korpela[5,6,12], Olli T. Raitakari[13,14], Veikko Salomaa[15], Marjo-Riitta Järvelin[7,8,9,16,17], Samuli Ripatti[3,4,18,19,*], Samuel Kaski[1,20,*]

1 MODEL DETAILS

Detailed structure of our model is shown as a Bayesian network in Supplementary Figure 7, see, e.g., Neapolitan (2004). The priors for the the low-rank covariance inducing part of the model, $H\Lambda + E$, are set exactly as in Bhattacharya and Dunson (2011), as follows:

$$\lambda_{jh}|\phi^{\Lambda}_{jh}, \tau_h \sim N\left(0, (\phi_{jh}\tau_h)^{-1}\right), \quad \phi^{\Lambda}_{jh} \sim \text{Ga}(\nu/2, \nu/2),$$
$$\tau_h = \prod_{l=1}^{h} \delta_l \quad \delta_1 \sim \text{Ga}(a_1, 1), \quad \delta_l \sim \text{Ga}(a_2, 1), \quad l \geq 2,$$
$$\sigma_j^{-2} \sim \text{Ga}(a_\sigma, b_\sigma), \quad (j = 1, \ldots, P), \tag{1}$$

where δ_l ($l = 1, \ldots, \infty$) are independent, τ_h is a global shrinkage parameter for the hth column of Λ and ϕ^{Λ}_{ph}s are local shrinkage parameters for the elements of the hth column to increase flexibility of the prior.

For the SNP-to-phenotype regression coefficient matrix, $\Theta = \Psi\Gamma$, we introduce a prior similar to (1). For the matrix $\Psi = [\psi_{jh}]$:

$$\psi_{jh}|\phi^{\Psi}_{jh}, \tau^*_h \sim N\left(0, \left(\phi^{\Psi}_{jh}\tau^*_h\right)^{-1}\right),$$
$$\phi^{\Psi}_{jh} \sim \text{Ga}(\nu/2, \nu/2), \quad \tau^*_h = \prod_{l=1}^{h} \delta^*_l,$$
$$\delta^*_1 \sim \text{Ga}(a_3, 1), \quad \delta^*_l \sim \text{Ga}(a_4, 1), \quad l \geq 2, \tag{2}$$

where the parameter τ^*_h acts as a global shrinkage parameter for the hth column of Ψ. The priors for the matrix $\Gamma = [\gamma_{jh}]$ are set as follows:

$$\gamma_{jh}|\tau^*_j \sim N\left(0, \left(\tau^*_j\right)^{-1}\right). \tag{3}$$

Here we include the elementwise shrinkage parameters ϕ^{Ψ}_{jh} only for matrix Ψ and not for Γ to represent our prior expectation of associations to be sparse on the SNP side but dense on the metabolite side. Note that in (3), the parameters τ^*_j represent global shrinkage parameters for the *rows* of Γ, as opposed to (1) and (2) in which the *columns* were shrunk. Furthermore, note that the parameters τ^*_h and δ^*_l and the corresponding hyperparameters a_3 and a_4 are shared between Ψ and Γ, because the scales of Ψ and Γ are not identifiable separately. The hyperparameter ν is common also with the prior for Λ given in (1), and we use a fixed value $\nu = 3$ similarly to Bhattacharya and Dunson (2011).

The prior distributions for hyperparameters a_3 and a_4 are set using results from the next section. Equation (16) specifies the contribution of the first component to the overall explained variation. This formula is used to select the parameters of a Gamma distribution for a_4 which imposes the distribution of the contribution of the first component to have quantiles $q_{0.01} = 0.3$ and $q_{0.99} = 0.999$, i.e. the contribution of the first component is with probability 0.98 in the interval 30%-99.9%. However, we note that with $K_1 = 1$, the first component explains the full effect and a_4 is not actually needed. Second, using Equation (15), a Gamma distribution is specified for a_3 which imposes the mean PTVE, μ_{PTVE} to satisfy the properties of the informative prior distribution, as specified in the main document. To solve for a_3 using Equation (15), we use Monte Carlo simulation to integrate out a_4.

2 RELATION TO OTHER METHODS

2.1 Canonical correlation analysis

Canonical correlation analysis (CCA) is a classical tool that can be used for measuring the strength of association between multivariate data sets (Hotelling, 1936). CCA has recently been used to investigate associations between multiple SNPs and multiple phenotypes (Tang and Ferreira, 2012; Marttinen *et al.*, 2013). However, with CCA, the *a priori* expected strength of association is stronger when testing with a greater number of SNPs. This results in reduced power when a set comprises only a few SNPs, and overfitting when testing with larger sets. Although recently introduced sparse variants of CCA (Waaijenborg *et al.*, 2008; Witten *et al.*, 2009; Parkhomenko *et al.*, 2009) can be used to alleviate the problem of overfitting, there does not seem to exist a fully satisfactory solution for correcting for this bias. Furthermore, canonical correlation analysis disregards the directionality of the association studies following from the causal understanding of the problem, namely that the SNPs are affecting the phenotypes, and not the other way around. Due to this, the interpretation of the outcome from canonical correlation analysis may be cumbersome and incorporating external knowledge in terms of intuitive prior distributions may not be straightforward, although model-based formulations exist (Klami and Kaski, 2007; Wang, 2007).

*to whom correspondence should be addressed

2.2 Regression models with latent confounding factors

The importance of accounting for latent non-genetic factors in association studies is well acknowledged (Stegle *et al.*, 2010; Fusi *et al.*, 2012). Let $\mathbf{Y} = (\mathbf{y}_1, \ldots, \mathbf{y}_G)$ here denote a collection of G phenotypes from N individuals, i.e. $\mathbf{Y}$ has N rows. Joint modelling of confounding factors and causal SNPs is based on estimating the model

$$\mathbf{Y} = \mu + \mathbf{SV} + \mathbf{XW} + \epsilon, \tag{4}$$

where $\mathbf{S} = (\mathbf{s}_1, \ldots, \mathbf{s}_K)$ is the collection of K vectors $\mathbf{s}_k$ of length N comprising the observed SNPs from N individuals. Matrix $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_Q)$ is a collection of Q hidden factors. Noise matrix ϵ contains i.i.d. univariate noise terms $\epsilon_{ng} \sim N(0, \sigma_e^2)$. The difference to our model is that in our model several SNPs are affecting the phenotypes through the reduced rank regression formulation, whereas here a standard regression is utilized, typically for just one SNP at a time.

2.3 Reduced rank regression

Baysian reduced rank regression, originally presented by Geweke (1996), is based on the model

$$Y = X\Psi\Phi + ZA + E, \tag{5}$$

where the matrices are interpreted similarly to our model. The main difference is that in (5) the rows of E are assumed to follow a multivariate normal distribution with covariance matrix Σ, which is given Inverse-Wishart prior distribution. In our approach a low-rank noise covariance is assumed, see the model description in the main document. Another difference is that in (5) the columns of Ψ and rows of Γ are given symmetric priors, but identifiability is ensured by fixing certain elements in the matrices to unity. In our model we shrink the columns of Ψ and rows of Γ increasingly, encouraging the largest effects to correspond to the first columns/rows.

A comparison carried out by (Carriero *et al.*, 2011) showed that rank reduction combined with shrinkage from Bayesian priors improves substantially the accuracy of predictions in a multivariate time series (52 variables), compared to some alternatives (classical reduced rank regression, factor models, Bayesian VAR, multivariate boosting). Another recent article by Vounou *et al.* (2010) presents a regularized reduced rank regression model, and uses this for association studies where the phenotype is a high-dimensional vector of brain-features. They showed that the model outperforms the standard exhaustive pairwise testing approach. However, the correlations between the phenotypes were not taken into account in their method.

2.4 Bayesian infinite sparse factor analysis

Bayesian infinite sparse factor analysis model presented in Bhattacharya and Dunson (2011) is defined by:

$$y_i = \Lambda h_i + \epsilon_i, \quad \epsilon_i \sim N_p(0, \Sigma),$$

where Λ is a factor loadings matrix, h_i is a vector of hidden factors, and $\Sigma = \text{diag}(\sigma_1^2, \ldots, \sigma_p^2)$. Bhattacharya and Dunson (2011) assume that Λ has infinitely many columns, however, such that the induced covariance matrix, $\Lambda\Lambda^T + \Sigma$ stays finite. This is achieved by introducing a prior that shrinks the columns of Λ increasingly as the column index increases. Note that although Λ is not rotation invariant, the prior induced on the covariance does not depend on the rotation. Furthermore, the prior encourages that the first columns of Λ to correspond to the most influential latent confounders. In our method, we use this factor analysis model to represent the multivariate noise.

3 PTVE-RARE SCORE

Let x denote here the S-dimensional SNP-vector and y the P-dimensional phenotype. Suppose we are interested in the prediction $\hat{y}_p$ for the pth phenotype. By letting θ denote here the pth column from the regression coefficient matrix $\Theta = \Psi\Gamma$, we can write:

$$\begin{aligned}\hat{y}_p &= \theta^T x \\ &= \theta_R^T x_R + \theta_C^T x_C,\end{aligned}$$

where x_R and x_C are subvectors of x comprising the rare and the common variants with the corresponding division of the regression coefficients into θ_R and θ_C. The variance of the prediction can be divided similarly:

$$\begin{aligned}\text{Var}(\hat{y}_p) = \theta_R^T \text{Var}(x_R)\theta_R + \theta_C^T \text{Var}(x_C)\theta_C \\ +2\theta_C^T \text{Cov}(x_C, x_R)\theta_R.\end{aligned}$$

We define the part of the variation of $\hat{y}_p$ attributed to the rare variants as:

$$\text{Var-rare}(p) \equiv \theta_R^T \text{Var}(x_R)\theta_R + \theta_C^T \text{Cov}(x_C, x_R)\theta_R. \tag{6}$$

The variation due to the cross-correlation term $2\theta_C^T \text{Cov}(x_C, x_R)\theta_R$ is in (6) divided evenly between the rare and common variants. This prevents counterintuitive situations possibly caused by negative correlation between the rare and the common variants where the variation due to the rare variants might be larger than the total variance. The PTVE-rare score is now defined as

$$\text{PTVE-rare} = \frac{\sum_{p=1}^{P} \text{Var-rare}(p)}{\text{Tr}(\text{Cov}(Y))}.$$

It is straightforward to estimate the posterior distribution of PTVE-rare similarly to PTVE using samples from the posterior distribution of the parameters.

4 COMPUTATION

For estimating the parameters of the model, we use the Gibbs sampling Markov chain Monte Carlo (MCMC) algorithm for reduced rank ragression (Geman and Geman, 1984; Geweke, 1996; Karlsson, 2012). The algorithm iteratively updates the parameters of the model one-by-one by sampling them from their conditional posterior probability distributions, given the current values of other parameters, and it is theoretically guaranteed to produce samples from the posterior distribution facilitating inferences concerning both qualitative aspects of the model (which SNPs affect which factors, which factors affect which phenotypes) and quantitative aspects of the model (the sizes of the effects).

However, when we experimented with the standard formulation of the Gibbs sampling algorithm, we found it to exhibit unstable

behavior when trying to learn the very small effects present in the genomics data: sometimes the algorithm got stuck into a single mode for an arbitrary number of iterations, and suddenly switched into a different mode, rendering the results useless in practice (running long enough to reach convergence over the different modes was not feasible in practice). The unsatisfactory behaviour has been discussed by others as well, and a solution based on restricting the lengths of the row-vectors of Γ has been suggested (Koop *et al.*, 2006). Recently, an informative prior formulation utilizing the correlations between the responses has been proposed (Gillberg et al., personal communication).

Here, we approximate the posterior distribution by using a point estimate for one of the parameters, Γ, leading to sensible results in practice. An efficient two-stage strategy is used for computing the approximation: first, the algorithm starts with an informative initialization based on the singular value decomposition, followed by a short MCMC run, and a point estimate $\bar{\Gamma}$ is computed using samples from the second half of the simulation. In the second stage the MCMC is run until convergence to learn the rest of the parameters, keeping $\bar{\Gamma}$ fixed in order to prevent the unwanted mode switching. The interpretation of the procedure is that, in the first stage, a combination of phenotypes putatively affected by the SNPs is learned. In the second stage, the SNPs, if any, affecting the combination and the effect sizes are thoroughly estimated. Details of the Gibbs sampling algorithm are provided in the next Section.

5 GIBBS SAMPLING UPDATES FOR THE REDUCED RANK REGRESSION

The Gibbs sampling algorithm proceeds by alternatingly updating the reduced rank regression part of the model $X\Psi\Gamma$, and the noise model $H\Lambda^T + E$. The standard regression ZA could be updated similarly; however, we regress out the known factors as a preprocessing step, hence these updates are not shown here. For updating the parameters related to the reduced rank regression part of the model, $\{\Psi, \Gamma, \Phi^{\Psi}, a_3, a_4, \delta_l^*\}$, we first compute the residuals

$$Y^* = Y - H^{(i)}\Lambda^{T(i)},$$

where $H^{(i)}$ and $\Lambda^{T(i)}$ are the current values of the corresponding variables. The parameter updates, given the residuals, are straightforward modifications of the updates for the standard Bayesian reduced rank regression model (Geweke, 1996; Karlsson, 2012) and they are provided below. Then, we calculate residuals

$$Y^{**} = Y - X\Psi^{(i)}\Gamma^{(i)},$$

where $\Psi^{(i)}$ and $\Gamma^{(i)}$ are now the current values of the Ψ and Γ parameters. Given the residuals Y^{**} the noise model including parameters $\left\{\Lambda, H, \Phi^{\Lambda}, a_1, a_2, \delta_l\right\}$ can be updated using Gibbs sampling steps derived for a Bayesian factor analysis model (Bhattacharya and Dunson, 2011). As we use exactly the same formulation for our noise model, these steps will not be repeated here.

Gibbs sampling MCMC updates for the original Bayesian reduced rank regression were derived by Geweke (1996) and Karlsson (2012). Here we derive updates for the reduced rank regression parameters Γ and Ψ (and related hyperparameters) which are accommodated to our specific model formulation. The derivations follow straightforwardly as most distributions are of semi-conjugate form. To keep the notation simple, let Y here represent the residuals, $Y - H\Lambda^T - ZA$, where effects of the latent factors have been removed, such that Y can be written as

$$Y = X\Psi\Gamma + E, \tag{7}$$

where the columns of E are independent.

In deriving the update equations, we exploit a standard result for Bayesian linear regression, see e.g. Bishop *et al.* (2006), which states that if

$$\beta \sim N(0, \Sigma_\beta)$$

and

$$y|X \sim N\left(X\beta, \Sigma_y\right), \tag{8}$$

then

$$\beta|y, X \sim N(\Sigma_{\beta|y}(X^T\Sigma_y^{-1}y), \Sigma_{\beta|y}), \tag{9}$$

where

$$\Sigma_{\beta|y} \sim \left(\Sigma_\beta^{-1} + X^T\Sigma_y^{-1}X\right)^{-1}.$$

This result can be used in order to derive the conditional distributions for Ψ and Γ, by transforming the reduced rank regression (7) appropriately into the form of the standard linear regression model (8).

5.1 Update for Γ

Because the columns of the residual matrix E are independent, we can update the columns Γ_i one-by-one by observing that

$$Y_i \sim N\left(X\Psi\Gamma_i, \sigma_i^2 I_N\right), \quad i = 1, \ldots, P,$$

which immediately leads to update equations

$$\pi(\Gamma_i|Y, \Theta_{-\Gamma}) = N\left(\mu_{\Gamma_i|Y,\Theta_{-\Gamma}}, \Sigma_{\Gamma_i|Y,\Theta_{-\Gamma}}\right), \quad i = 1, \ldots, P,$$

where

$$\mu_{\Gamma_i|Y,\Theta_{-\Gamma}} = \Sigma_{\Gamma_i|Y,\Theta_{-\Gamma}}\left(\Psi^T X^T Y_i\right)\sigma_i^{-2}$$
$$\Sigma_{\Gamma_i|Y,\Theta_{-\Gamma}} = \left(\Sigma_{\Gamma_i}^{-1} + \sigma_i^{-2}\Psi^T X^T X\Psi\right)^{-1},$$

where Σ_{Γ_i} is a diagonal matrix with prior variances of the elements of Γ_i, see (3), collected of the diagonal.

5.2 Update for Ψ

In order to derive the update equations for Ψ, the model (7) is written in a vectorized form as follows:

$$\text{vec}(Y) \sim N\left(\left(\Gamma^T \otimes X\right)\text{vec}(\Psi), \Sigma_E \otimes I_N\right),$$

where Σ_E is a diagonal matrix with variances of the columns of E on the diagonal. Again, the conditional distribution follows by applying the results for the linear model (9) and applying straightforward algebra for Kronecker products:

$$\pi(\text{vec}(\Psi)|Y,\Theta_{-\Psi}) = N(\mu_{\Psi|Y,\Theta_{-\Psi}}, \Sigma_{\Psi|Y,\Theta_{-\Psi}}),$$

with

$$\mu_{\Psi|Y,\Theta_{-\Psi}} = \Sigma_{\Psi|Y,\Theta_{-\Psi}} \text{vec}\left(X^T Y \Sigma_E^{-1} \Gamma^T\right)$$
$$\Sigma_{\Psi|Y,\Theta_{-\Psi}} = \left(\Sigma_\Psi^{-1} + \left(\Gamma \Sigma_E^{-1} \Gamma^T\right) \otimes \left(X^T X\right)\right)^{-1},$$

where Σ_Ψ is a diagonal matrix, with prior variances, see (2) of the elements of Ψ, read columnwise, on the diagonal.

5.3 Update for Φ^Ψ

Recall from (2) that the element ϕ_{jh}^Ψ appears as a factor in the precision of ψ_{jh}, the (j,h)th element of the Ψ matrix and has the conjugate prior distribution Ga$(\nu/2, \nu/2)$. Thus, it follows that Φ^Ψcan be updated element-by-element using the conditional distrbution:

$$\phi_{jh}^\Psi \sim \text{Ga}\left(\text{shape} = \alpha_{\phi_{jh}^\Psi|Y,\Theta_{-\phi_{jh}^\Psi}}, \text{rate} = \beta_{\phi_{jh}^\Psi|Y,\Theta_{-\phi_{jh}^\Psi}}\right),$$

where

$$\alpha_{\phi_{jh}^\Psi|Y,\Theta_{-\phi_{jh}^\Psi}} = \frac{\nu+1}{2}$$
$$\beta_{\phi_{jh}^\Psi|Y,\Theta_{-\phi_{jh}^\Psi}} = \frac{1}{2}\left(\nu + \tau_h \psi_{jh}^2\right).$$

5.4 Update for $\delta_l^*, l = 1, \ldots, K_1$

First, notice that δ_1^* appears as a factor in the precision of normally distributed variables (the elements of first the column of Ψ and the first row of Γ) and has the conjugate Ga$(a_3, 1)$ prior, see Equations (2) and (3). Therefore, the conditional posterior distribution of δ_1^* is also a Gamma distribution, with parameters as follows:

$$\pi\left(\delta_1^*|Y,\Theta_{-\delta_1^*}\right) = \text{Ga}\left(\text{shape} = \alpha_{\delta_1^*|Y,\Theta_{-\delta_1^*}}, \text{rate} = \beta_{\delta_1^*|Y,\Theta_{-\delta_1^*}}\right),$$

where

$$\alpha_{\delta_1^*|Y,\Theta_{-\delta_1^*}} = a_3 + \frac{(P+S)K_1}{2}$$
$$\beta_{\delta_1^*|Y,\Theta_{-\delta_1^*}} = 1 + \frac{1}{2}\left(\sum_{h=1}^{K_1} \tau_h^{(1)} \sum_{j=1}^{S} \phi_{jh}^\Psi \psi_{jh}^2 + \sum_{j=1}^{K_1} \tau_j^{(1)} \sum_{h=1}^{P} \gamma_{jh}^2\right),$$

where we have used notation $\tau_h^{(l)} = \prod_{t=1,\ldots,h,t\neq l} \delta_t^*$, $h, l = 1, \ldots, K_1$.

Similarly, the δ_l^* appears as a factor in the precisions of columns from l to K_1 of Ψ and the corresponding rows of Γ. The conditional distribution of δ_l^* is therefore

$$\pi\left(\delta_l^*|Y,\Theta_{-\delta_l^*}\right) = \text{Ga}\left(\text{shape} = \alpha_{\delta_l^*|Y,\Theta_{-\delta_l^*}}, \text{rate} = \beta_{\delta_l^*|Y,\Theta_{-\delta_l^*}}\right),$$

where

$$\alpha_{\delta_l^*|Y,\Theta_{-\delta_l^*}} = a_4 + \frac{(P+S)(K_1-l+1)}{2}$$
$$\beta_{\delta_l^*|Y,\Theta_{-\delta_l^*}} = 1 + \frac{1}{2}\left(\sum_{j=l}^{K_1} \tau_j^{(l)} \sum_{h=1}^{P} \gamma_{jh}^2 + \sum_{h=l}^{K_1} \tau_h^{(l)} \sum_{j=1}^{S} \phi_{jh}^\Psi \psi_{jh}^2\right)$$

5.5 Update for a_3 and a_4

The conditional distribution of these parameters, given data and all other parameters, is not of simple form, therefore we update the parameters jointly using a Metropolis-Hastings step within the Gibbs sampler. Let a_3^* and a_4^* denote the proposed values. The sampler switches to the new values with probability

$$\min\left\{1, \alpha\left[(a_3,a_4) \longrightarrow (a_3^*, a_4^*)\right]\right\}$$

where the acceptance ratio is the following:

$$\alpha\left[(a_3,a_4) \longrightarrow (a_3^*, a_4^*)\right] = \frac{\pi\left(a_3^*, a_4^*|Y,\Theta_{-(a_3^*,a_4^*)}\right) J\left[(a_3^*,a_4^*) \to (a_3,a_4)\right]}{\pi\left(a_3,a_4|Y,\Theta_{-(a_3,a_4)}\right) J\left[(a_3,a_4) \to (a_3^*,a_4^*)\right]}. \quad (10)$$

Here, J denotes the proposal distribution that consists of proposing the new values for the two parameters independently of each other:

$$J\left[(a_3,a_4) \to (a_3^*,a_4^*)\right] = J'\left(a_3 \to a_3^*\right) J'\left(a_4 \to a_4^*\right).$$

The new values are proposed on a logarithmic scale from a normal distribution centered on the current value, as follows:

$$J'\left(a \to a^*\right) = N(\log(a^*)|\text{mean} = \log(a), \text{sd}= \log(a)/10).$$

The posterior distribution $\pi\left(a_3,a_4|Y,\Theta_{-(a_3,a_4)}\right)$ appearing in the acceptance ratio (10) has the following form

$$\pi\left(a_3,a_4|Y,\Theta_{-(a_3,a_4)}\right) \propto \text{Ga}(a_3|\alpha_{a_3},\beta_{a_3})\text{Ga}(a_4|\alpha_{a_4},\beta_{a_4}) \times \text{Ga}(\delta_1^*|a_3)\prod_{l=2}^{K_2}\text{Ga}(\delta_l^*|a_4). \quad (11)$$

The proportionality in (11) includes the truncation of the gamma prior distributions to legitimate values $a_3 > 2$ and $a_4 > 3$, such that the results from Supplementary Section 6 can be applied. However, the proportionality constant is the same in the numerator and denominator of (10) and, thus, cancels in the computations.

6 PROOFS

In this section we derive results which characterize the dependency of the prior distribution of PTVE on the model hyperparameter a_3 and a_4. The following lemma shows that the amount of variance of the ith phenotype explained by the hth component, given a fixed value of the variance parameter τ_h^*, is proportional to the total variation of the SNPs that are used as predictors.

Lemma 1:

$$Var(x^T \Psi_h \gamma_{hi}|\tau_h^*) = 3\left(\tau_h^*\right)^{-2} \sum_{i=1}^{S} Var(x_i)$$

Proof of Lemma 1:

$$
\begin{aligned}
Var(x^T\Psi_h\gamma_{hi}|\tau_h^*) &= \text{Var}(x^T\Psi_h|\tau_h^*)Var(\gamma_{hi}|\tau_h^*) \\
&= (\tau_h^*)^{-1}\, E\left(x^t\Psi_h\Psi_h^T x|\tau_h^*\right) \\
&= (\tau_h^*)^{-1}\, E\left(\text{Tr}\left\{x^t\Psi_h\Psi_h^T x\right\}|\tau_h^*\right) \\
&= (\tau_h^*)^{-1}\, E\left(\text{Tr}\left\{xx^t\Psi_h\Psi_h^T\right\}|\tau_h^*\right) \\
&= (\tau_h^*)^{-1}\,\text{Tr}\left\{E\left(xx^t\Psi_h\Psi_h^T|\tau_h^*\right)\right\} \\
&= (\tau_h^*)^{-1}\,\text{Tr}\left\{E\left(xx^t|\tau_h^*\right)E(\Psi_h\Psi_h^T|\tau_h^*)\right\} \\
&= (\tau_h^*)^{-1}\,\text{Tr}\left\{E(xx^t)E[(\phi_{1h}^{\Psi})^{-1}]\,(\tau_h^*)^{-1}\,I\right\} \\
&= (\tau_h^*)^{-2}\,E[(\phi_{1h}^{\Psi})^{-1}]\text{Tr}\left\{E(xx^T)\right\} \\
&= (\tau_h^*)^{-2}\,\frac{\nu}{\nu-2}\text{Tr}\left\{E(xx^T)\right\} \\
&= (\tau_h^*)^{-2}\,\frac{\nu}{\nu-2}\sum_{i=1}^{S}\text{Var}(x_i).
\end{aligned}
$$

The lemma follows when $\nu = 3$.

Recall that $\tau_h^* = \prod_{l=1}^{h}\delta_l^*$. Thus, if the parameters δ_l^* are given, the variance explained by all components is obtained by summing over the components:

$$
\sum_{h=1}^{\infty}\text{Var}(x^T\Psi_h\gamma_{hi}|\delta_l^*, l=1,\ldots) \tag{12}
$$

$$
= \left[\frac{\nu}{\nu-2}\sum_{i=1}^{S}\text{Var}(x_i)\right]\sum_{h=1}^{\infty}(\delta_1^*)^{-2}\prod_{l=2}^{h}(\delta_l^*)^{-2} \tag{13}
$$

We can take an expectation of (13) over the distribution of δ_l^*, $l = 1, 2, \ldots$

$$
\begin{aligned}
&E\left[\sum_{h=1}^{\infty}\text{Var}(x^T\Psi_h\gamma_{hi}|\delta_l^*, l=1,\ldots)\right] \\
&= \left[\frac{\nu}{\nu-2}\sum_{i=1}^{S}\text{Var}(x_i)\right]\sum_{h=1}^{\infty}E\left[(\delta_1^*)^{-2}\right]\prod_{l=2}^{h}E\left[(\delta_l^*)^{-2}\right] \\
&= \left[\frac{\nu}{\nu-2}\sum_{i=1}^{S}\text{Var}(x_i)\right]\sum_{h=1}^{\infty}E\left[(\delta_1^*)^{-2}\right]E\left[(\delta_2^*)^{-2}\right]^{h-1} \\
&= \frac{\nu}{\nu-2}\sum_{i=1}^{S}\text{Var}(x_i)\frac{E\left[(\delta_1^*)^{-2}\right]}{1-E\left[(\delta_2^*)^{-2}\right]}
\end{aligned} \tag{14}
$$

The convergence of the infinite sum is guaranteed as long as $E[(\delta_1^*)^{-2}] < \infty$ and $E[(\delta_2^*)^{-2}] < 1$, for which sufficient conditions are $a_3 > 2$ and $a_4 > 3$, see below. The expectation for the proportion of total variation explained is obtained by multiplying (14) with the number of phenotypes and dividing by the total variation in the phenotypes, $\sum_{i=1}^{P}\text{Var}(Y_i)$, yielding

$$
\frac{\nu P}{\nu-2}\frac{E\left[(\delta_1^*)^{-2}\right]}{1-E\left[(\delta_2^*)^{-2}\right]}\frac{\sum_{i=1}^{S}\text{Var}(x_i)}{\sum_{i=1}^{P}\text{Var}(Y_i)}
$$

By plugging in the expectations

$$
E\left[(\delta_1^*)^{-2}\right] = \frac{\Gamma(a_3-2)}{\Gamma(a_3)} \qquad \text{and} \qquad E\left[(\delta_2^*)^{-2}\right] = \frac{\Gamma(a_4-2)}{\Gamma(a_4)}
$$

we get our first corollary:

Corollary 1: The expected proportion of the total variation explained by all SNPs, given fixed values of the hyperparameters a_3 and a_4 is equal to

$$
\frac{\nu P}{\nu-2}\frac{\Gamma(a_3-2)/\Gamma(a_3)}{1-\Gamma(a_4-2)/\Gamma(a_4)}\frac{\sum_{i=1}^{S}\text{Var}(x_i)}{\sum_{i=1}^{P}\text{Var}(Y_i)}. \tag{15}
$$

The second corollary specifies the share of the total variation explained that can be attributed to the first component, and is obtained straightforwardly from the calculations above.

Corollary 2: Suppose we are given fixed values of hyperparameters a_3 and a_4. Of the proportion of total variation that is explained by all components, given in (15), the share of the first component is equal to

$$
E(\delta_1^{-2})/\frac{E(\delta_1^{-2})}{1-E(\delta_2^{-2})} = 1-E(\delta_2^{-2}) = 1-\frac{\Gamma(a_4-2)}{\Gamma(a_4)}. \tag{16}
$$

7 REPLICATION P-VALUE

In order to compute the p-value in the replication data sets for the multivariate association detected in the NFBC1966 data set, we reduced the multivariate association into an association between two univariate variables: the genotype combination and the phenotype combination. The univariate genotype combination was obtained simply by scaling the genotypes in the replication data in the same way as with the NFBC1966 data and multiplying the resulting genotype matrix with $\bar{\Psi}$ learned with NFBC1966 data. On the phenotype side, we estimated a univariate variable q_i for each individual i representing the phenotypes by finding the maximum likelihood value for

$$
y_i \sim N_P(q_i\bar{\Gamma}, \bar{\Sigma}),
$$

where y_i is the P-dimensional phenotype of individual i, $\bar{\Gamma}$ is the value of the Γ parameter estimated with NFBC1966 data and $\bar{\Sigma}$ is the estimated covariance of the phenotypes. Thus, q_i represents the value of the latent factor which is the most likely to have generated the observed phenotypes through the coefficient matrix $\bar{\Gamma}$. Simple linear regression model was then used to check whether the genotype combination and the phenotype combination were positively correlated.

SUPPLEMENTARY FIGURES AND TABLES

Supplementary Figures 1-7 and Tables 1-2 are located in the end of this document. Supplementary Tables 3-8 can be found in another file *supplementary_Tables_3_to_8.pdf*.

ACKNOWLEDGEMENT

Funding: This work was financially supported by the Academy of Finland (grant number 251170 to the Finnish Centre of Excellence in Computational Inference Research COIN; grant number 259272 to PM; grant number 257654 to MP; grant number 137870 to PS). VS was supported by the Academy of Finland, grant number 139635, and the Finnish Foundation for Cardiovascular Research. MAK was supported by the Sigrid Juselius Foundation, the Finnish Funding Agency for Technology TEKES, and the Strategic Research Funding from the University of Oulu.

NFBC1966 has been financially supported the Academy of Finland (project grants 104781, 120315, 129269, 1114194, 24300796, Center of Excellence in Complex Disease Genetics and SALVE), University Hospital Oulu, Biocenter, University of Oulu, Finland (75617), NHLBI grant 5R01HL087679-02 through the STAMPEED program (1RL1MH083268-01), NIH/NIMH (5R01MH63706:02), ENGAGE project and grant agreement HEALTH-F4-2007-201413, EU FP7 EurHEALTHAgeing -277849 and the Medical Research Council, UK (G0500539, G0600705, G1002319, PrevMetSyn/SALVE).

The Young Finns Study has been financially supported by the Academy of Finland: grants 126925, 121584, 124282, 129378 (Salve), 117787 (Gendi), and 41071 (Skidi), the Social Insurance Institution of Finland, Kuopio, Tampere and Turku University Hospital Medical Funds (grant 9N035 for TeLeht), Juho Vainio Foundation, Paavo Nurmi Foundation, Finnish Foundation of Cardiovascular Research and Finnish Cultural Foundation, Tampere Tuberculosis Foundation and Emil Aaltonen Foundation (T.L).

REFERENCES

Bhattacharya, A. and Dunson, D. (2011). Sparse Bayesian infinite factor models. *Biometrika*, **98**(2), 291–306.

Bishop, C. M. *et al.* (2006). *Pattern recognition and machine learning*. Springer, New York.

Carriero, A., Kapetanios, G., and Marcellino, M. (2011). Forecasting large datasets with bayesian reduced rank multivariate models. *Journal of Applied Econometrics*, **26**(5), 735–761.

Fusi, N., Stegle, O., and Lawrence, N. (2012). Joint modelling of confounding factors and prominent genetic regulators provides increased accuracy in genetical genomics studies. *PLoS Computational Biology*, **8**(1), e1002330.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, (6), 721–741.

Geweke, J. (1996). Bayesian reduced rank regression in econometrics. *Journal of Econometrics*, **75**(1), 121–146.

Hotelling, H. (1936). Relations between two sets of variates. *Biometrika*, **28**(3/4), 321–377.

Karlsson, S. (2012). Conditional posteriors for the reduced rank regression model. Technical report, Orebro University, School of Business.

Klami, A. and Kaski, S. (2007). Local dependent components. In *Proceedings of the 24th International Conference on Machine Learning*, pages 425–432. ACM.

Koop, G., Strachan, R. W., Van Dijk, H., and Villani, M. (2006). Bayesian approaches to cointegration. In *The Palgrave Handbook of Theoretical Econometrics*, pages 871–898. Palgrave Macmillan.

Marttinen, P., Gillberg, J., Havulinna, A., Corander, J., and Kaski, S. (2013). Genome-wide association studies with high-dimensional phenotypes. *Statistical Applications in Genetics and Molecular Biology*. in press, pre-print from http://arxiv.org/abs/1211.1144.

Neapolitan, R. E. (2004). *Learning Bayesian networks*. Pearson Prentice Hall Upper Saddle River.

Parkhomenko, E., Tritchler, D., and Beyene, J. (2009). Sparse canonical correlation analysis with application to genomic data integration. *Statistical Applications in Genetics and Molecular Biology*, **8**(1), 1–34.

Stegle, O., Parts, L., Durbin, R., and Winn, J. (2010). A Bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies. *PLoS Computational Biology*, **6**(5), e1000770.

Tang, C. S. and Ferreira, M. A. (2012). A gene-based test of association using canonical correlation analysis. *Bioinformatics*, **28**(6), 845–850.

Vounou, M., Nichols, T., and Montana, G. (2010). Discovering genetic associations with high-dimensional neuroimaging phenotypes: a sparse reduced-rank regression approach. *Neuroimage*, **53**(3), 1147–1159.

Waaijenborg, S., Verselewel de Witt Hamer, P. C., and Zwinderman, A. H. (2008). Quantifying the association between gene expressions and dna-markers by penalized canonical correlation analysis. *Statistical Applications in Genetics and Molecular Biology*, **7**, 1–29.

Wang, C. (2007). Variational Bayesian approach to canonical correlation analysis. *IEEE Transactions on Neural Networks*, **18**(3), 905–910.

Wheeler, D. L., Barrett, T., Benson, D. A., Bryant, S. H., Canese, K., Chetvernin, V., Church, D. M., DiCuccio, M., Edgar, R., Federhen, S., *et al.* (2007). Database resources of the national center for biotechnology information. *Nucleic Acids Research*, **35**(suppl 1), D5–D12.

Witten, D. M., Tibshirani, R., *et al.* (2009). Extensions of sparse canonical correlation analysis with applications to genomic data. *Statistical Applications in Genetics and Molecular Biology*, **8**(1), 1–27.

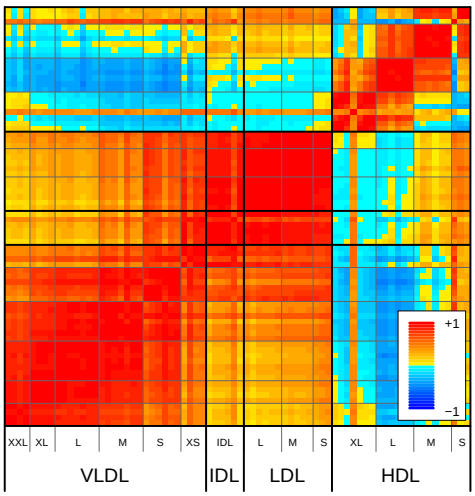

Fig. 1. Empirical correlation matrix of the lipoprotein traits. A classification into the different lipoprotein classes is shown below the x-axis.

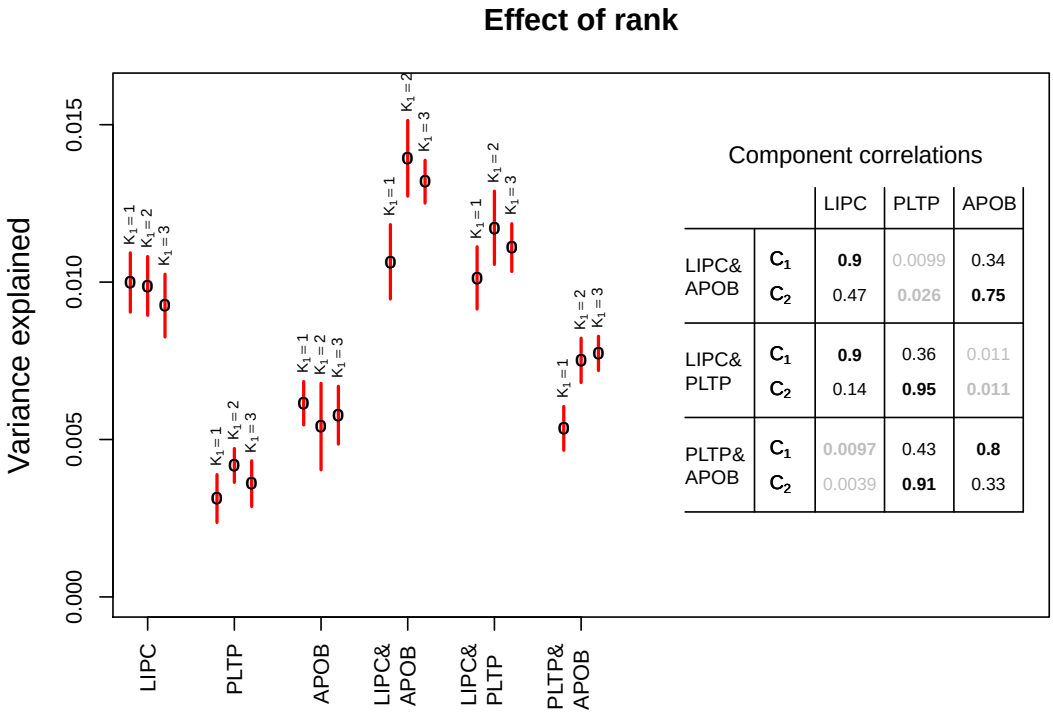

Fig. 2. Impact of model rank on the results. Estimated proportion of total variance explained is shown for three genes, *LIPC*, *PLTP*, *APOB*, and for all two-gene joint models. The black circle marks the mean and the red lines show +/- 2 standard deviations. The table on the right shows correlations between the two components of the joint models from the analyses with $K_1 = 2$ and the single-gene 1-dimensional model components.

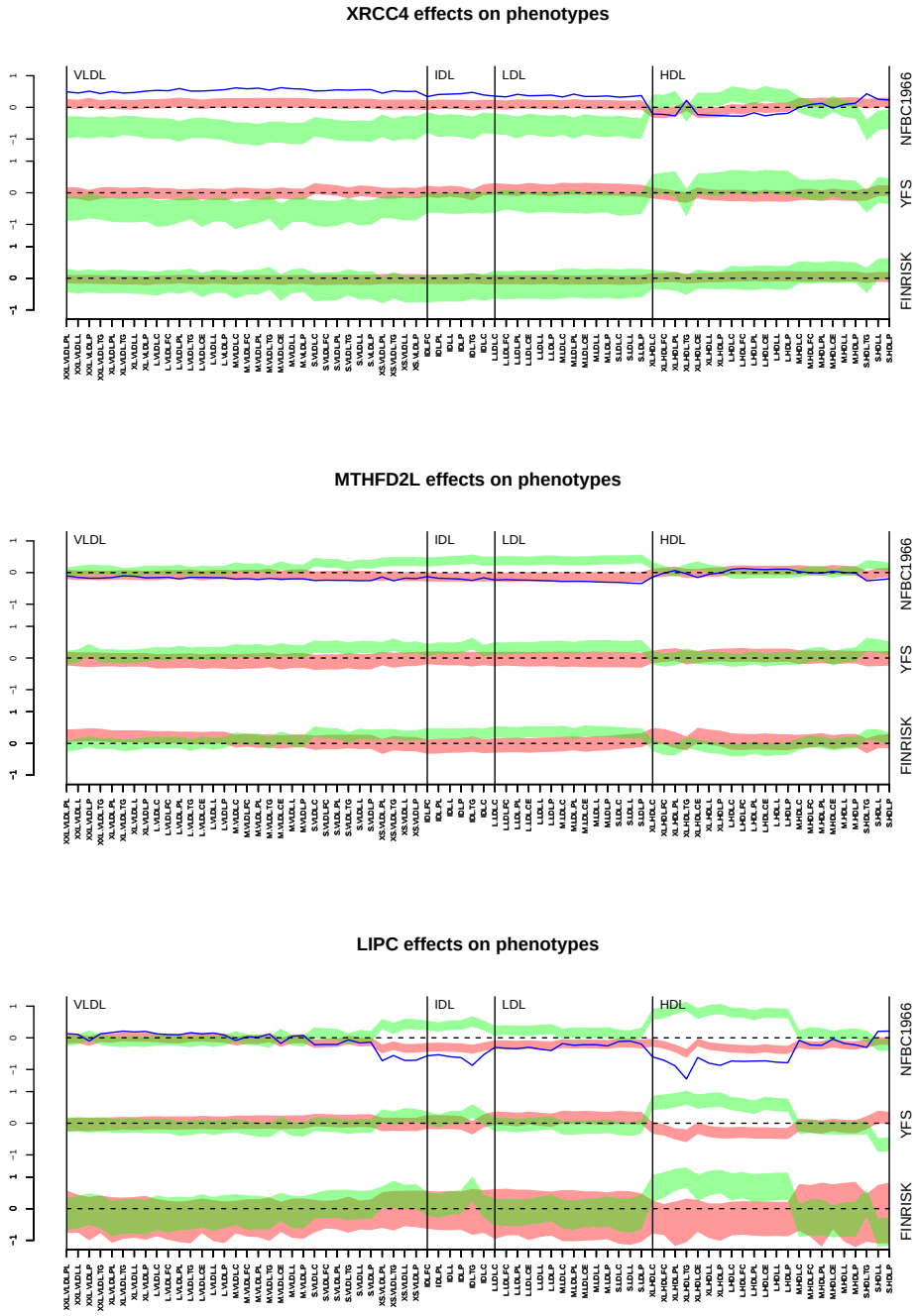

Fig. 3. The effects of *XRCC4, MTHFD2L* and *LIPC* on the lipoprotein traits. The traits are shown on the x-axis. The green/red area shows the 95 per cent confidence interval for the mean trait value in the group annotated with green/red background in Figure 3 of the main document. For example, individuals in the green group have lower VLDL trait values than individuals in the red group. One unit on the y-axis denotes one standard deviation. The results are shown for all three data sets. The solid blue line in the NFBC1966 plot shows the effect as estimated by the model, and it gives the predicted average difference between the red and the green groups. The values defining the boundaries of the high and low genotype combination groups have been optimized to minimize the overlap between the confidence intervals using the NFBC1966 data.

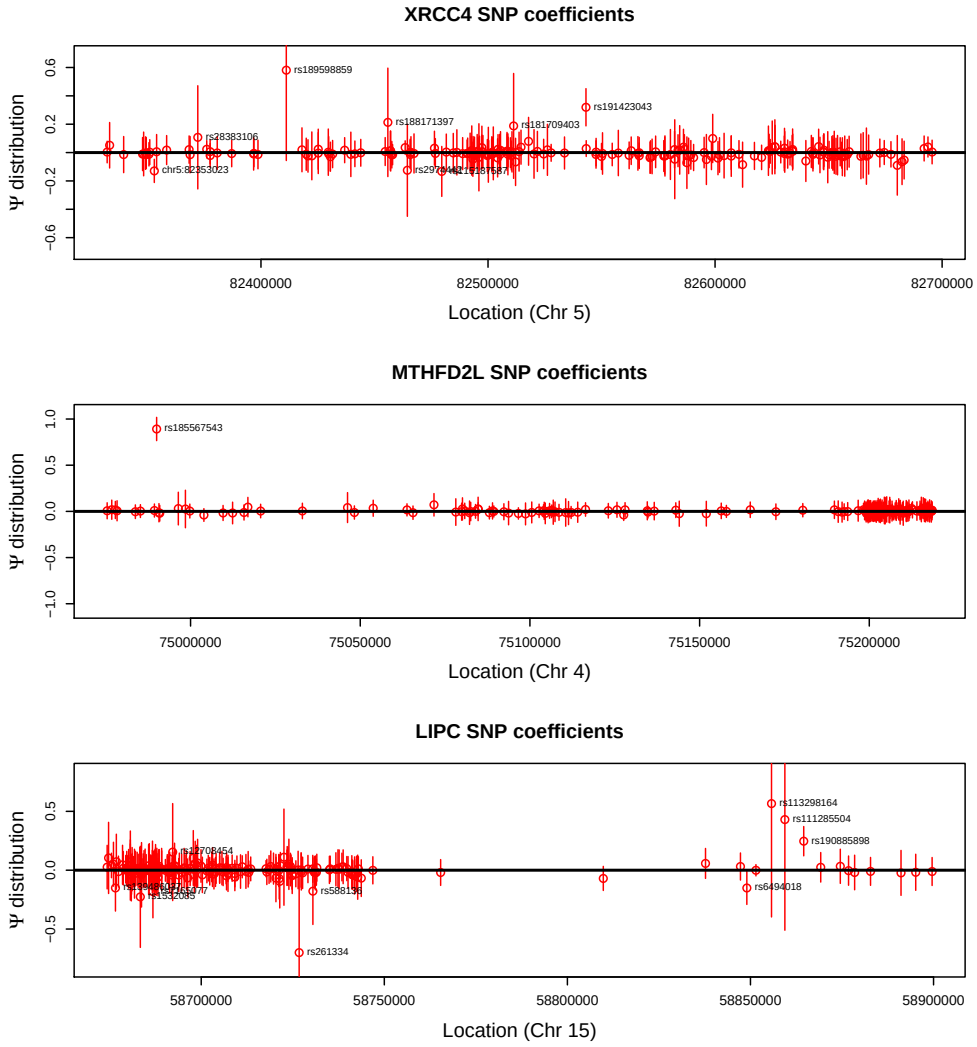

Fig. 4. Marginal confidence intervals for the SNP coefficients for three genes: *XRCC4, MTHFD2L* and *LIPC*, derived from posterior samples for the Ψ matrix. Notice how, when the effect is concentrated on a single SNP, as in *MTHFDL*, other SNPs, although linked with the lead SNP, get very small weights. The large standard deviations for rs113298164 and rs111285504 in the results for *LIPC* are due to the fact that either one but not both of these SNPs is needed to explain the effect; hence, these two coefficients are strongly negatively correlated with each other. The coefficients shown here differ from those presented in Supplementary Table 1, because the Ψ matrix represents effects of the SNPs scaled to have unit variance, whereas in Supplementary Table 1 the coefficients have been scaled back to correspond to the original 0,1,2 SNP encoding.

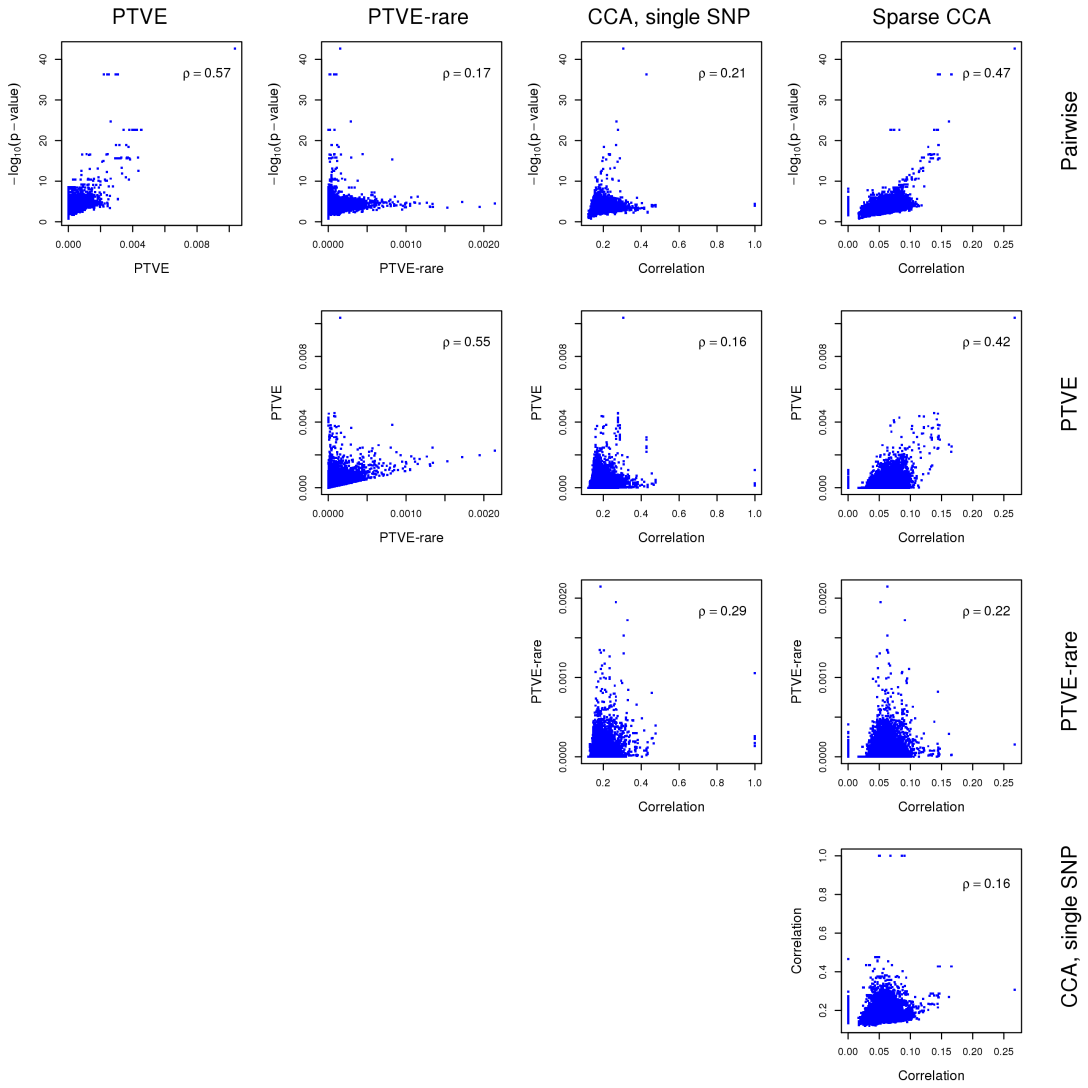

Fig. 5. Correlations between different methods. The panels show pairwise comparisons between scores from different methods included in the study. The names of the methods are shown on top of the columns and on the right of the rows. The correlation between the scores is shown within each panel.

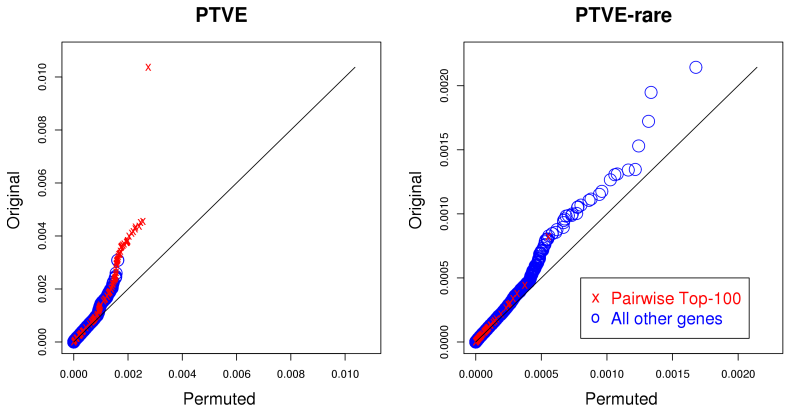

Fig. 6. A Q-Q plot for PTVE and PTVE-rare scores obtained by the Bayesian reduced rank regression. The panel on the left shows the PTVE scores for all human genes, plotted against their expected values obtained using analyses of permuted data sets. The panel on the right shows a Q-Q plot for the PTVE-rare scores. As additional annotation, a hundred genes with the highest scores in the pairwise analysis are drawn using a red cross, whereas all other genes are drawn using a circle. Notice how genes with high PTVE values can also be seen in the pairwise analysis. On the other hand, genes with the largest PTVE-rare scores are in general not detectable by the pairwise analysis.

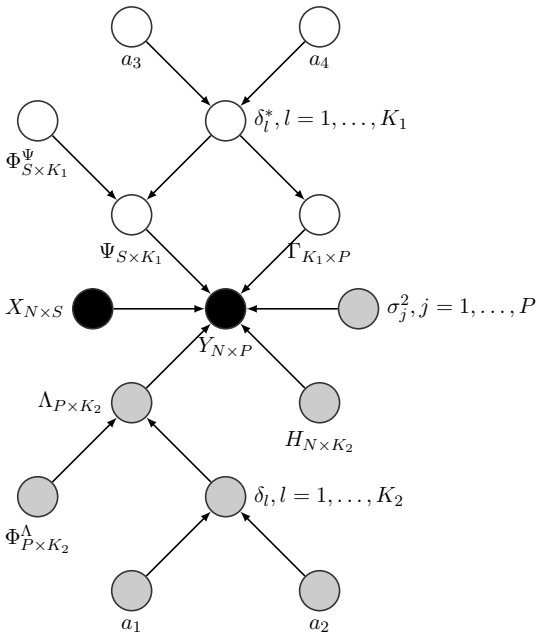

Fig. 7. Bayesian network representing the structure of the model. Black filled circles denote the observed data. Grey-colored circles represent the variables related to the noise model. White circles represent the parameters related to the reduced rank regression part of the model. For clarity, the variables related to the standard regression part of the model are not shown.

Table 1. Detailed information on the associated SNPs in the significant genes from Table 1 of the main text. The columns are interpreted as follows: **MAF**, the minor allele frequency; **Effect**, the coefficient of the SNP in the genotype combination (to be interpreted as relative to one another); **contribution**, the cumulative contribution of the SNPs on the total amount of variation explained by the gene up to 80 per cent. *significant replication in only one of the two replication data sets.

a)

Chr	Locus	SNP	MAF	Effect	Contribution
5	*XRCC4*	rs189598859	0.0034	9.3	0.33
		rs28360178	0.082	1.5	0.58
		rs181709403	0.0011	8.2	0.67
		rs191423043	0.00023	16	0.75
		rs115640857	0.062	-0.88	0.81
16	*SPIRE2**	rs1110400	0.0027	-18	0.065
		chr16:89971087:I	0.0029	36	0.22
		rs149408054	0.00034	-17	0.48
		rs60958597	0.0032	-470	0.61
		rs8059075	0.0032	450	0.92

b)

Chr	Locus	SNP	MAF	Effect	Contribution
2	*DTNB**	rs744976	0.49	-0.46	0.27
		rs75591229	0.052	1.8	0.45
		rs142390827	0.0013	-8.2	0.6
		chr2:25625422:I	0.37	-0.61	0.72
		rs17745484	0.34	-0.56	0.8
		rs150371006	0.028	-1.1	0.85
4	*MTHFD2L*	rs185567543	0.026	5.1	0.89

Table 2. Background information about the genes with significant replication. The table reports genome-wide significant associations located within 1Mb of the gene, as reported in the *Gene* database (Wheeler *et al.*, 2007).

Chr	Locus	Trait	P-value (-log10)	Distance
5	*XRCC4*	Mortality	16.7	35 kb
		Spondylitis, Ankylosing	9.0	520 kb
4	*MTHFD2L*	Leukocyte Count	16.7	46 kb