Supplementary materials for

HEFT: eQTL analysis of many thousands of expressed genes while simultaneously controlling for hidden factors

Chuan Gao [1*], Nicole L. Tignor[2], Jacqueline Salit[2], Yael Strulovici-Barel[2], Neil R. Hackett[2], Ronald G. Crystal[2] and Jason G. Mezey [1,2*]

In this material, we provide details and derivations for the Hidden Expression Factor (HEFT) analysis model that combines a multivariate ridge regression and factor analysis to simultaneously correct for the hidden factors in expression Quantitative Trait Loci (eQTL) analysis (**S.1**). We also present detailed methods for the simulation and lung SAE data analyses described in the main text (**S.2**), as well as additional figures and discussion of results for these analyses (**S.3**). For the theory work, we show that the ridge penalty for the multivariate regression component of the model is not just imposed blindly, but instead arises naturally as a necessary component for the model to work. We note that for the derivations, we assume that we are working with a factor model where the factors are correlated, but since assuming uncorrelated factors has little influence on the results (Rubin and Thayer, 1982), we only implemented the uncorrelated version. We also note that the algorithm for the sole purpose of factor analysis, without including the multivariate regression, has been presented by Rubin (Rubin and Thayer, 1982).

S.1 The model and inference

S.1.1 The HEFT model

As described in the main text, our full eQTL model with hidden structures for a single genotype with m expression variables and sample of size n can be written as

$$\mathbf{Y} = \mu \mathbf{1}_{\mathbf{m}}^{'} + \mathbf{X}\beta + \mathbf{\Lambda F} + \mathbf{W} \tag{S.1.1.1}$$

where $\mathbf{Y}$ is an $n \times m$ matrix of measured expression variables, $\mathbf{1_m}$ is vector of 1s of length m, μ is an $n \times 1$ vector of row means, $\mathbf{X}$ is a $n \times 2$ matrix with the first column set to 1 and second column set to the genotype, β is the $2 \times m$ matrix of column means and genotypic effects, $\mathbf{\Lambda}$ and $\mathbf{F}$ are the $n \times p$ loading matrix and $p \times m$ matrix of values for p factors, and $\mathbf{W}$ is the the $n \times m$ error matrix, where we make the standard assumption that covariance among samples can be well modeled by non-error terms such that each column of matrix $\mathbf{W}$ has a normal distribution $W_j \sim \mathbf{N}(\mathbf{0}, \mathbf{\Psi_j})$ with diagonal $n \times n$ matrix $\mathbf{\Psi_j}$. To avoid the potential problems caused by biased estimates of unconstrained error variances, we assume that expression variables have been scaled to a common variance and we constrain each of the $\mathbf{\Psi}$ to be $\mathbf{I}\sigma^2$.

Generally, when estimating the fixed effect β with confounding in a linear mixed model framework, the hidden factors are integrated out (Fusi *et al.*, 2012; Listgarten *et al.*, 2010). That is, assuming $\mathbf{F} \sim N(\mathbf{0}, \mathbf{\Sigma})$ with covariance matrix $\mathbf{\Sigma}$, an incomplete likelihood written in the following form can be used to obtain the Maximum Likelihood Estimator (MLE).

$$L(\theta|\mathbf{D}) = \frac{1}{(2\pi)^{nm/2}|\mathbf{\Psi} + \mathbf{\Lambda\Sigma\Lambda^T}|^{m/2}} \mathbf{exp}\left(\mathbf{tr}(-\frac{1}{2}(\mathbf{H} - \mathbf{X}\beta)^T(\mathbf{\Psi} + \mathbf{\Lambda\Sigma\Lambda^T})^{-1}(\mathbf{H} - \mathbf{X}\beta))\right) \tag{S.1.1.2}$$

where $\mathbf{H} = \mathbf{Y} - \mu\mathbf{1}_{\mathbf{m}}^{'}$. In the linear mixed model context (Kang *et al.*, 2008), a covariance matrix $\mathbf{R}\sigma^{\mathbf{2}}$, which is equivalent to $\mathbf{\Lambda\Sigma\Lambda^T}$ is used to capture the covariance structure of the hidden factor, where $\mathbf{R}$ is a pre-fixed similarity matrix obtained in advance, and the MLE of β is obtained in the following form

$$\hat{\beta} = (\mathbf{X^T}(\mathbf{R}\sigma^\mathbf{2} + \mathbf{\Psi})^\mathbf{-1}\mathbf{X})^\mathbf{-1}\mathbf{X^T}(\mathbf{R}\sigma^\mathbf{2} + \mathbf{\Psi})^\mathbf{-1}\mathbf{H} \qquad \text{(S.1.1.3)}$$

This linear mixed model approach helps correct the spurious associations caused by non-orthogonal structure (structures that create effects that are non-orthogonal to the genotypic effects), however, at a price of reducing power because the approach pools the error term and the covariance structure of the hidden factor instead of partitioning them.

We approach the problem from the factor analysis angle by explicitly modeling the hidden factors with the goal of partitioning factor variance from the error term, while simultaneously estimating the fixed effect. We accomplish this by placing a ridge penalty $||\mathbf{\Xi}^T\beta||^2$ on the fixed effects β and by considering the complete log likelihood, which takes the following form

$$\begin{aligned} l_c &= -\frac{\mathbf{1}}{\mathbf{2}}\mathbf{tr}(\mathbf{FF^T}) - \frac{\mathbf{m}}{\mathbf{2}}\mathbf{log}|\mathbf{\Psi}| - ||\mathbf{\Xi}^T\beta||^2 \\ &\quad -\frac{\mathbf{1}}{\mathbf{2}}\mathbf{tr}((\mathbf{H} - \mathbf{X}\beta - \mathbf{\Lambda F})(\mathbf{H} - \mathbf{X}\beta - \mathbf{\Lambda F})^\mathbf{T}\mathbf{\Psi}^\mathbf{-1}) \qquad \text{(S.1.1.4)}\end{aligned}$$

Next, we lay out the necessary pieces of the EM algorithm based on this likelihood.

S.1.2 The Expectation-Maximization algorithm

S.1.2.1 The Expectation step

In the expectation step, we transform the incomplete likelihood in equation S.1.1.2 to the complete likelihood in equation S.1.1.4. This transformation is made possible by noticing that the hidden factor $\mathbf{F}$ can be substituted by its expected value conditional on $\mathbf{Y}$. To get the $\mathbf{E}(\mathbf{F}|\mathbf{Y})$, we note that the joint distribution of $\mathbf{F}$ and $\mathbf{Y}$, the latter in terms of $\mathbf{H} = \mathbf{Y} - \mu\mathbf{1}'_\mathbf{m}$, can be written as

$$\begin{pmatrix} \mathbf{F} \\ \mathbf{H} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ \mathbf{X}\beta \end{pmatrix}\begin{pmatrix} \mathbf{\Sigma} & \mathbf{\Sigma\Lambda^T} \\ \mathbf{\Lambda\Sigma} & \mathbf{\Lambda\Sigma\Lambda^T} + \mathbf{\Psi} \end{pmatrix}\right) \qquad \text{(S.1.2.1)}$$

from which the conditional variance of $\mathbf{F}$ can be written as

$$\mathbf{V}(\mathbf{F}|\mathbf{Y}) = \mathbf{\Sigma} - \mathbf{\Sigma\Lambda^T}(\mathbf{\Lambda\Sigma\Lambda^T} + \mathbf{\Psi})^\mathbf{-1}\mathbf{\Lambda\Sigma^T} \qquad \text{(S.1.2.2)}$$

and the conditional expected value of $\mathbf{F}$ takes the the following form

$$\mathbf{E}(\mathbf{F}|\mathbf{Y}) = \mathbf{\Sigma\Lambda^T}(\mathbf{\Lambda\Sigma\Lambda^T} + \mathbf{\Psi})^\mathbf{-1}(\mathbf{H} - \mathbf{X}\beta) \qquad \text{(S.1.2.3)}$$

Notice that in the two equations above, the inversion of $(\mathbf{\Lambda\Sigma\Lambda^T} + \mathbf{\Psi})^\mathbf{-1}$ has computation complexity of n^3, where n is the sample size. We can simplify the computations by converting the complexity from n^3 to p^3, where p is the number of factors by using the Woodbury matrix identity

$$\mathbf{AB^T}(\mathbf{BAB^T} + \mathbf{R})^\mathbf{-1} = (\mathbf{A}^\mathbf{-1} + \mathbf{B^T}\mathbf{R}^\mathbf{-1}\mathbf{B})^\mathbf{-1}\mathbf{B^T}\mathbf{R}^\mathbf{-1} \qquad \text{(S.1.2.4)}$$

and after some algebra, we get

$$\mathbf{V}(\mathbf{F}|\mathbf{Y}) = \mathbf{\Sigma} - (\mathbf{\Sigma}^\mathbf{-1} + \mathbf{\Lambda^T}\mathbf{\Psi}^\mathbf{-1}\mathbf{\Lambda})^\mathbf{-1}\mathbf{\Lambda^T}\mathbf{\Psi}^\mathbf{-1}\mathbf{\Lambda\Sigma^T} \qquad \text{(S.1.2.5)}$$

$$\mathbf{E}(\mathbf{F}|\mathbf{Y}) \;=\; (\mathbf{\Sigma^{-1}} + \mathbf{\Lambda^T}\mathbf{\Psi^{-1}}\mathbf{\Lambda})^{\mathbf{-1}}\mathbf{\Lambda^T}\mathbf{\Psi^{-1}}(\mathbf{H} - \mathbf{X}\beta) \tag{S.1.2.6}$$

Importantly, for the special case with $\mathbf{\Sigma} = \mathbf{I}$, we have

$$\mathbf{V}(\mathbf{F}|\mathbf{Y}) \;=\; (\mathbf{I} + \mathbf{\Lambda^T}\mathbf{\Psi^{-1}}\mathbf{\Lambda})^{\mathbf{-1}} \tag{S.1.2.7}$$

$$\mathbf{E}(\mathbf{F}|\mathbf{Y}) = (\mathbf{I} + \mathbf{\Lambda^T}\mathbf{\Psi^{-1}}\mathbf{\Lambda})^{\mathbf{-1}}\mathbf{\Lambda^T}\mathbf{\Psi^{-1}}(\mathbf{H} - \mathbf{X}\beta) \tag{S.1.2.8}$$

which is the form of a ridge estimator. Thus, the $\mathbf{E}(\mathbf{F}|\mathbf{Y})$ incorporates the same penalty that is imposed on the fixed effects, a result that arises naturally in the form of the expectation. We discuss the importance of this critical result in the next section.

S.1.2.2 The Maximization step

Finding the Maximum Likelihood Estimator (MLE) for the parameters β, $\mathbf{\Lambda}$, $\mathbf{\Psi}$, and μ involves taking the first derivative of the likelihood shown in equation S.1.1.4 with respect to each parameter, setting to 0 and solving the equation. We provide the derivation of β and the rest of the parameters follow using the same approach.

The log-likelihood of the ridge model can be written as:

$$l_c \;=\; -\mathbf{log}|\mathbf{\Psi}| - \mathbf{tr}((\mathbf{H} - \mathbf{\Lambda F} - \mathbf{X}\beta)(\mathbf{H} - \mathbf{\Lambda F} - \mathbf{X}\beta)^{\mathbf{T}}\mathbf{\Psi^{-1}}) - \mathbf{tr}(\mathbf{\Xi^T}\beta\beta^{\mathbf{T}}\mathbf{\Xi}) \tag{S.1.2.9}$$

we take the derivative of l_c with respect to β, which gives

$$\frac{\partial l_c}{\partial \beta} \;=\; -\mathbf{2}\mathbf{X^T}\mathbf{\Psi^{-1}}(\mathbf{H} - \mathbf{\Lambda F}) + \mathbf{2}\mathbf{X^T}\mathbf{\Psi^{-1}}\mathbf{X}\beta + \mathbf{2}\mathbf{\Xi}\mathbf{\Xi^T}\beta \tag{S.1.2.10}$$

set to 0 and solve for β, which gives

$$\hat{\beta} = (\mathbf{X^T}\mathbf{\Psi^{-1}}\mathbf{X} + \mathbf{\Xi}\mathbf{\Xi^T})^{\mathbf{-1}}\mathbf{X^T}\mathbf{\Psi^{-1}}(\mathbf{H} - \mathbf{\Lambda F}) \tag{S.1.2.11}$$

Note that if we assume $\beta \sim N(\mathbf{0}, \mathbf{\Theta})$ and treat $\mathbf{\Lambda F}$ as an observed variable, then the distribution of β and $\mathbf{Y}$ can be written as

$$\begin{pmatrix} \beta \\ \mathbf{H} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ \mathbf{\Lambda F} \end{pmatrix}\begin{pmatrix} \mathbf{\Theta} & \mathbf{\Theta X^T} \\ \mathbf{X\Theta} & \mathbf{X\Theta X^T} + \mathbf{\Psi} \end{pmatrix}\right) \tag{S.1.2.12}$$

Then using the property of the joint normal distribution

$$\begin{aligned} \mathbf{E}(\beta|\mathbf{Y}) \;&=\; \mathbf{\Theta X^T}(\mathbf{X\Theta X^T} + \mathbf{\Psi})^{\mathbf{-1}}(\mathbf{H} - \mathbf{\Lambda F}) \\ &=\; (\mathbf{X^T}\mathbf{\Psi^{-1}}\mathbf{X} + \mathbf{\Theta^{-1}})^{\mathbf{-1}}\mathbf{X^T}\mathbf{\Psi^{-1}}(\mathbf{H} - \mathbf{\Lambda F}) \end{aligned} \tag{S.1.2.13}$$

comparing equation S.1.2.13 to equation S.1.2.11, we see that if we set $\mathbf{\Theta^{-1}} = \mathbf{\Xi\Xi^T}$, the two equations take the same form, such that assuming a prior of $\beta \sim N(\mathbf{0}, \mathbf{\Theta})$ is equivalent to imposing a ridge penalty. Thus, from the result above, both $\mathbf{E}(\mathbf{F}|\mathbf{Y})$ and $\mathbf{E}(\beta|\mathbf{Y})$ take the form of a ridge estimator.

Recognizing that $\mathbf{E}(\mathbf{F}|\mathbf{Y})$ of the random effect is a shrinkage parameter of the same form as $\mathbf{E}(\beta|\mathbf{Y})$ (an L2 norm) is the key to making HEFT work. It suggests that to correct for hidden factors that are correlated with the fixed covariate $\mathbf{X}$, a penalty has to be imposed on β. Otherwise, the shrinkage property of the random effect will push the non-orthogonal hidden factor away towards β, which would defeat the purpose of including $\mathbf{\Lambda F}$ to correct for the false positives.

This also suggests that by controlling the variance term of the prior, we can effectively control how much shrinkage we can impose on these parameters. To prevent over-shrinkage of either the β or factors, here, we used the same amount of shrinkage for β and $\mathbf{F}$ by applying $\mathbf{\Theta^{-1}} = \mathbf{I}$ and $\mathbf{\Sigma} = \mathbf{I}$, respectively.

The MLE for the rest of the parameters can be derived similarly using the same principle, where $\mathbf{\Lambda}$ takes the following form

$$\hat{\mathbf{\Lambda}} = \mathbf{Y}(\mathbf{E}(\mathbf{F}|\mathbf{Y}))^{\mathbf{T}}(\mathbf{E}(\mathbf{F}\mathbf{F}^{\mathbf{T}}|\mathbf{Y}))^{\mathbf{-1}} \tag{S.1.2.14}$$

where $\mathbf{E}(\mathbf{F}\mathbf{F}^{\mathbf{T}}|\mathbf{Y}) = \mathbf{E}(\mathbf{F}|\mathbf{Y})\mathbf{E}(\mathbf{F}|\mathbf{Y})^{\mathbf{T}} + m\mathbf{Var}(\mathbf{F}|\mathbf{Y})$, and for $\mathbf{\Psi}$ the MLE is

$$\hat{\mathbf{\Psi}} = \frac{1}{m}\mathbf{diag}((\mathbf{H} - \mathbf{X}\beta)(\mathbf{H} - \mathbf{X}\beta)^{\mathbf{T}} - \mathbf{\Lambda}\mathbf{E}(\mathbf{F}|\mathbf{Y})(\mathbf{H} - \mathbf{X}\beta)^{\mathbf{T}}) \tag{S.1.2.15}$$

We note that the diagonal matrix $\mathbf{\Psi}$ acts as a weight for each sample, which can be used to produce more accurate parameter estimates when they are drawn from distributions with heterogenous variance. However, if improperly learned, these variances can lead to an ill-conditioned system. To avoid this, we further restraint $\mathbf{\Psi} = \mathbf{I}\sigma^{\mathbf{2}}$. To get $\mathbf{\Psi}$, we simply set each element to the average of all elements across the diagonal.

Finally, the global mean μ is simply set to

$$\hat{\mu} = \frac{1}{m}\sum_{i=1}^{m}(Y_i - X\beta_i) \tag{S.1.2.16}$$

S.1.2.3 Unifying the fixed effects and the factor effects

Now that we know the expected value of the factor effects takes the same form as the penalized fixed effects, we can unify the ridge regression model and the factor model into a combined framework by making the following substitutions. Letting $\mathbf{\Omega} = [\mathbf{X\Lambda}]$ and $\mathbf{\Gamma} = \begin{bmatrix} \beta \\ \mathbf{F} \end{bmatrix}$, we can simplify the model as

$$\mathbf{Y} = \mu\mathbf{1}_{\mathbf{m}}^{'} + \mathbf{\Omega\Gamma} + \mathbf{W} \tag{S.1.2.17}$$

Notice that this treatment transforms the model into a simple factor analysis model, which has the same form of maximum likelihood estimator for $\mathbf{\Omega}$ as for $\mathbf{\Lambda}$, except that the first few columns of $\mathbf{\Omega}$ are fixed covariates, and the rest of the columns are for the hidden factors. We can use the same type of EM algorithm for the factor analysis model for inference of the parameters, with the iteration steps listed below.

1. Initialize parameter values
2. Find μ by using equation S.1.2.16
3. Set $\mathbf{H} = \mathbf{Y} - \mu\mathbf{1}_{\mathbf{m}}'$
4. $\mathbf{Var}(\mathbf{\Gamma}|\mathbf{Y})_{\mathbf{t+1}} = (\mathbf{I} + \mathbf{\Omega}_{\mathbf{t}}^{\mathbf{T}}\mathbf{\Psi}_{\mathbf{t}}^{-\mathbf{1}}\mathbf{\Omega}_{\mathbf{t}})^{-\mathbf{1}}$
5. $\mathbf{E}(\mathbf{\Gamma}|\mathbf{Y})_{\mathbf{t+1}} = \mathbf{Var}(\mathbf{\Gamma}|\mathbf{Y})_{\mathbf{t+1}}\mathbf{\Omega}_{\mathbf{t}}^{\mathbf{T}}\mathbf{\Psi}_{\mathbf{t}}^{-\mathbf{1}}\mathbf{H}$
6. Set the corresponding row of $\mathbf{E}(\mathbf{\Gamma}|\mathbf{Y})$ to β

7. Keep the fixed effects and the known covariates in the $\boldsymbol{\Omega}$ matrix fixed, update the rest using the following formula $\boldsymbol{\Omega_{t+1}} = \mathbf{HE(\Gamma|Y)_t^T E(\Gamma\Gamma^T|Y)_t^{-1}}$
8. $\mathbf{S_{t+1}} = \frac{\mathbf{1}}{\mathbf{m}}(\mathbf{HH^T} - \boldsymbol{\Omega}_{\mathbf{t+1}}\mathbf{E(\Gamma|Y)_{t+1}H^T})$
9. $\boldsymbol{\Psi}_{\mathbf{ii}}^{\mathbf{(t+1)}} = \frac{\mathbf{tr(S)}}{\mathbf{n}}$
10. Iterate until convergence

The initial values of the parameters are randomly chosen, but special care was taken to guard against ill conditioned values. Specifically, all β's were initialized to be 0, $\boldsymbol{\Lambda}$ were randomly generated from $\mathbf{N(0,I)}$, and $\boldsymbol{\Psi}$ were set to 0.5 across the diagonal. The convergence of the algorithm can be diagnosed by checking whether the update of the likelihood or parameters approach a specified tolerance threshold. We prefer checking the tolerance of the likelihood, which can be calculated as in equation S.1.1.4.

From these steps, we note that the EM algorithm has time complexity scaling $max(O(p^3), O(nmp))$ by noting that for the E step, the bottleneck is in the inversion step $(\mathbf{I} + \boldsymbol{\Omega}_\mathbf{t}^\mathbf{T}\boldsymbol{\Psi}_\mathbf{t}^{-\mathbf{1}}\boldsymbol{\Omega}_\mathbf{t})^{-\mathbf{1}}$, which scales as $O(p^3)$, and for the maximization step, the most expensive component is $\mathbf{HE(\Gamma|Y)_t^T E(\Gamma\Gamma^T|Y)_t^{-1}}$, where the scaling of the matrix multiplication component is $O(nmp)$. We also note that when the factor number p is small, the contribution of the matrix inversion is minimal.

S.1.3 Convexity of the objective function

We next show that the objective function for our model is strictly convex, such that the EM algorithm monotonically climbs the likelihood surface to the MLE.

By observing that log-likelihoood in it's quadratic form

$$l_1 = \sum_{i=1}^{n}(y_i - \mu - X\beta_i - \Lambda F_i)^T \boldsymbol{\Psi}^{-1}(y_i - \mu - X\beta_i - \Lambda F_i) \tag{S.1.3.1}$$

and noting the following proposition

Proposition: If $F(x_1, x_2, ..., x_n) = \mathbf{x}^T C\mathbf{x}$ is a quadratic form for n variables, and if matrix C is symmetric, then F is convex $\Leftrightarrow$ C is semi positive definite

Since we know that $\boldsymbol{\Psi}$ is semi positive definite, the above objective with respect to each i is convex. Note also that while both $\boldsymbol{\Lambda}$ and $\mathbf{F}$ are unknown this is not a concern because we are only interested in the nuisance parameter $\boldsymbol{\Lambda}\mathbf{F}$ as a whole, so the two unknowns effectively combine as one variable. Now, using the following Theorem

Theorem: If f is a function in n variables defined on a convex subset $S \subseteq R^n$, then if $f = \sum_i^n a_i f_i$, where each $a_i \geq 0$ and each f_i is a convex function defined on S, then f is convex.

and given the form of S.1.3.1 the likelihood is convex.

Finally, we note that the EM algorithm, with the latent variable transformed into observed variable, guarantees convergence to the mode of the above likelihood, where a proof can be found in (Bishop, 2007).

S.1.4 Selecting the factor numbers

Various techniques can be used to select the number of factor used by the model. For example, the Akaike information criterion (AIC)

$$AIC = 2k - 2ln(L) \tag{S.1.4.1}$$

or the Bayesian information criterion (BIC)

$$BIC = -2ln(L) + Kln(n) \tag{S.1.4.2}$$

where K is the parameter number, n is the sample size and L is the likelihood. Although we found these two criteria work well for a relatively simple data, they demonstrated less satisfactory performance for more complicated data that are affected by multiple hidden factors. In contrast, we found that selecting the factor number by manually examining the eigen spectrum of the data was a reasonable strategy, where performance of HEFT was robust to cases where this approach resulted in the inclusion of larger than the true number of factors (see main text and supplement simulation analysis results below). The eigen spectrum is obtained by performing a Principal Component Analysis on the data, then the variance proportion that is explained by each component is calculated. The factor number is then selected based on how well a top set of eigenvalues can can be visually separated from the rest.

S.1.5 The test statistics

One approach to developing a test statistic is calculating a Likelihood Ratio Test (LRT) to calculate approximate p-values of the fixed effects. The LRT is performed by calculating the following value,

$$LR = -2 * (l_0 - l_1) \tag{S.1.5.1}$$

where l_0 and l_1 correspond respectively to the log likelihood of the null model and the full model. For the purposes of eQTL analysis, to calculate these two log likelihoods, we use the complete model in equation S.1.1.4, which is a function of all $\mathbf{Y}$'s and a single genotype, then we delete one column of $\mathbf{Y}$ at a time, where for each deletion, we calculate a null likelihood for the deleted gene as a function of all other genes and the same genotype, such that m null likelihoods will be generated for a genotype, each corresponding to the deleted gene and the genotype pair.

We note that since the LRT does not have a well-behaved distribution for the HEFT model and since a large number of runs of the EM are required to apply a LRT test ($m \times l$, where l is the total number of genotypes), which can be a computational burden when very small p-values need to be obtained for the ranking of the tests, requiring the tolerance of the EM to be set higher, such that the corresponding runs take longer. We therefore prefer a one step Wald test approach, where this test is performed by constructing a t-type test statistic

$$\frac{\hat{\beta} - 0}{\sqrt{Var(\hat{\beta})}} \tag{S.1.5.2}$$

where we are testing the significance of only one X_i, Y_j pair at a time, and $Var(\hat{\beta})$ is the corresponding diagonal element of the covariance matrix of $\hat{\beta}$. With the matrix of predictors and the loadings (as well as other covariates) is $\mathbf{X}$, we can calculate $Var(\hat{\beta})$ as follows

$$Cov(\hat{\beta}) = \mathbf{(I + X^T \Psi^{-1} X)^{-1} X^T \Psi^{-1} Var(y_i) \Psi^{-1} X (I + X^T \Psi^{-1} X)^{-1}} \tag{S.1.5.3}$$

where $Var(y_i)$ can be calculated as the variance of $y_i - \mathbf{X}\hat{\beta}$.

S.2 Supplementary methods

S.2.1 Additional details for simulated data and analyses

As described in the main text, we simulated data for each of the following scenarios (Table S1): a) no eQTLs and no hidden factors (null scenario 1), b) no eQTLs with hidden factors (null scenario 2), c) eQTL where each affects one expressed gene (no pleiotropy) and no hidden factors, d) a combination of pleiotropic and non-pleiotropic eQTL and no hidden factors, e) non-pleiotropic eQTLs with hidden factors, f) a combination of pleiotropic and non-pleiotropic eQTL with hidden factors. For each of the scenarios with hidden factors (b, e, f), we simulated 10 datasets where the hidden factors effects were orthogonal to the entire set of markers and 10 datasets with hidden factors that were non-orthogonal to a non-trivial subset of the markers. For the scenarios with no hidden factors (a, c, d), we also simulated 10 datasets each. The sample size for each dataset was fixed at n=200. To generate the genetic markers of each dataset, SNP genotypes were generated using the coalescent simulator MaCS (Chen *et al.*, 2009) using the default approximation for tree width. We simulated 5 Mb of marker data for a single diploid populations of size $N_e = 10,000$, with a population mutation rate of $\theta = 4N\mu = 0.001$ and the recombination rate of $\rho = 4N\kappa = 0.00045$, values taken from Voight et al. (Voight *et al.*, 2005). For a dataset, we randomly selected 1000 SNPs from those with a derived Minor Allele Frequency (MAF) greater than 0.1, producing average linkage disequilibrium of 0.45 ± 0.01 for all ten datasets for pairwise markers measured by r^2.

Table S1: The parameter set up for all simulations showing the combinations of hidden factors, eQTLs, and pleiotropic eQTLs, used for each scenario, where × means the parameter is absent and √ indicate the parameter is present. The bottom rows also show the heritability range for eQTL (when present)

	Scenarios					
	a	b	c	d	e	f
eQTLs	×	×	√	√	√	√
Pleiotropic eQTLs	×	×	×	√	×	√
Factors	×	√	×	×	√	√
Heritability(min/max)-non-orth	×	×	3.6e-06/0.81	8.6e-07/0.91	2.1e-06/0.81	5.5e-07/0.87
Heritability(min/max)-orth	×	×			1.1e-06/0.84	5.0e-07/0.83

To generate the gene expression values of each dataset, we simulated 500 gene expression variables with standard normal error. For the eQTL scenarios with no pleiotropy (c and e) we randomly selected 50 uncorrelated markers to be eQTL, where the additive effect of each on a randomly selected gene was drawn from a standard normal. For the cases with pleiotropy (d and f) we included 50 eQTL with individual gene effects and selected an additional 20 uncorrelated SNPs each influencing 20 expression variables each, where again, the effect on each gene was selected from a standard normal. Overall, the total variation explained by the eQTLs for a given gene ranged from 5.0e-07 to 0.92, with the vast majority in the range of 0-0.025. For each dataset with hidden factors (b, e, f), we additionally incorporated the effects of four factors. To simulate a non-orthogonal factor, the scores of individuals on the first principal component of the correlation matrix of 100 randomly selected markers was used to assign individuals into five total groups, where the individuals with the largest 40 scores were assigned to group 1, individuals with the next largest 40 were assigned to group 2, etc. (i.e. factor effects were orthogonal to each other although non-orthogonal to the 100 SNPs). For each group, a single effect was then assigned drawing from $N(0, 1)$ or from $N(0, 3)$. For orthogonal factors, we applied the same procedure but randomly assigned each individual to one of the five groups. While for the latter, this does not prevent a factor from being non-orthogonal to some markers, we found that each factor was approximately orthogonal to almost all of the markers in the dataset in practice.

We analyzed each simulated dataset with the eight methods mentioned in the main text: a linear regression (LR), a two-step version of our method (HEFT-TS), the mixed model approach with full rank similarity matrix, LMM (Listgarten *et al.*, 2010) and low rank similarity matrix, PANAMA (Fusi *et al.*, 2012), the variational Bayes method PEER (Stegle *et al.*, 2010), the surrogate variable method SVA (Leek and Storey, 2007), the low rank sparse representation method LORS (Yang *et al.*, 2013), and HEFT. For the two-step analysis, we estimated factor structure using a factor analysis of the expressed genes and then used the residuals $\mathbf{Y} - \mathbf{\Lambda F}$ to do a secondary analysis, i.e. we applied a two-step approach within the HEFT framework (HEFT-TS). For LMM (Listgarten *et al.*, 2010), we could not get the software provided by the authors to work, so we re-implemented the algorithm. We note that we used the same convergence and other implementation criteria as described by the authors (Listgarten *et al.*, 2010) and that our implementation performed as they described. We also note that we did not make use of their populations structure component to allow for an appropriate comparison because there is no population structure in our simulated data (i.e. we applied LMM-EH and not LMM-PS-EH). While PEER (Stegle *et al.*, 2010) can in theory perform simultaneous analysis of eQTL and hidden factors, there are no simultaneous inference components implemented in the available R software package. We therefore applied PEER using their two-step option. For SVA and LORS, we applied these methods using the default setting.

For each analysis method, the association of each SNP-gene expression pair was assessed. For LR and HEFT we used the resulting p-values. For HEFT-TS and PEER, we followed the same procedure as applied in the PEER paper (Stegle *et al.*, 2010), where we extracted a p-value-like statistic from a linear regression model applied to the residuals after fitting the factor model. For LMM, we calculated the p-value statistic as described in their paper. As LORS does not generate p-values, we evaluated it by ranking its regression coefficients.

S.2.2 Lung Airway Dataset

We used HEFT, PEER, PANAMA, and linear regression to identify eQTL affecting gene expression in the lung Small Airway Epithelium (SAE) using a dataset that included 79 smokers and 37 nonsmokers recruited from the New York City area. The individuals in the sample were of different genders, different ancestry groups, and were characterized as non-smokers or smokers and were further labeled as healthy or having a lung disease phenotype (see Table S2). Details concerning data collection for these samples have been provided elsewhere (Harvey *et al.*, 2008). Briefly, SAE cell populations were collected by bronchial brushing of the small airway (Raman *et al.*, 2009) and RNA was hybridized to the HG-U133 Plus 2.0 microarray (Affymetrix, Santa Clara, CA) using standard protocols. To avoid the problem of probe sets mapping to wrong genes, we used the custom mapping provided by (Dai *et al.*, 2005) and the Robust Multi-array Average (RMA) (Irizarry *et al.*, 2003a,b) normalization method to convert array probe expression measurements into a single expression measurements for genes with unique Entrez gene IDs. We further removed genes with individual expression values beyond 3 standard deviation of the mean, which appeared likely to be outliers. This provided data on ∼7,575 protein-coding genes, an unknown subset of which are operating in the regulation and response behaviors of the pulmonary environment.

Blood was also collected from each individual and Affymetrix 500k microarrays were used to provide SNP genotypes. After filtering SNPs with a MAF below 0.1, significant deviations from Hardy-Weinberg equilibrium as assessed by a p-value < 0.05 for an efficient exact test (Wigginton *et al.*, 2005), and those genotypes with any missing observations using PLINK (Purcell *et al.*, 2007), this left 191,959 genotypes for analysis. The complete expression and genotype dataset analyzed in this study have been deposited in NCBI's Gene Expression Omnibus (Edgar *et al.*, 2002) and are accessible through GEO Series accession number GSE32030 (`http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?token=hvuxloaoaekwoza&acc=GSE40364`).

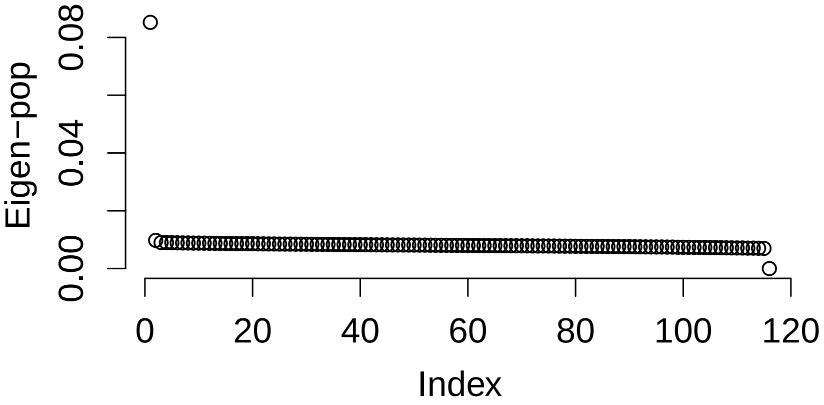
We initially applied the factor component of the HEFT model to just the expression data treating the known smoking covariate as missing information to assess the recovery ability of the factor component of the model. We then applied a complete HEFT analysis to the entire dataset. For this latter analysis, we selected the hidden factor number by visually examine the eigen spectrum of the gene expression correlation matrix and selected 5 factors that are clearly separable from the rest (Figure S1). We note that we also tried different factor numbers (3, 7, 12), where the p-value distributions of these analyses were not qualitatively different. To account for the obvious population structure in these data (Figure S1), we applied a factor analysis to the genotype covariance matrix (Engelhardt and Stephens, 2010) and incorporated the loadings of the first factors as fixed covariates, where this factor number was selected from the genotype covariance matrix eigen spectrum. We additionally included fixed covariates including gender, disease status and the smoking status. For binary covariates such as gender and smoking, we encoded them as 0 and 1, while three level disease status was encoded as a $n \times 2$ binary design matrix of either 0 or 1. For a baseline comparison, we also applied a multiple regression model including all of the same fixed covariates. Two thresholds for assessing significance of each SNP-expression pair were applied: a Bonferroni corrected threshold of $0.05 \ / \ (7,575 \times 191,959) = 3.438578\text{e-}11$ and a Benjamini-Hochberg control of the false discovery rate at $q = 0.05$.

S.3 Supplementary results

S.3.1 Comparison of HEFT to hidden factor methods

S.3.1.1 Performance for null and standard eQTL scenarios.

For datasets simulated under scenario a), where there are no eQTL and no hidden factors (null scenario 1), all eQTL analysis methods for which p-values were produced, all methods except PANAMA returned a uniform distribution of p-values for the set of all SNP-gene tests as measured by genomic inflation factor in a range of 1.00-1.04 (Aulchenko *et al.*, 2007; Devlin *et al.*, 2004), indicating they all performed appropriately for this null scenario (Table S3 and Figure S2). This outcome was observed regardless of the number of factors that were provided to HEFT-TS, PEER, and HEFT, indicating that these methods are also robust to incorporating the wrong number of factors (>0) for this null scenario. We note that PANAMA also returns a uniform p-value distribution but as the software automatically takes into account the multiple test scenario by calculating a q values, all values are close to one (as expected).We also note the LORS does not produce p-values so we could not check whether this method performed as expected under this null scenario.



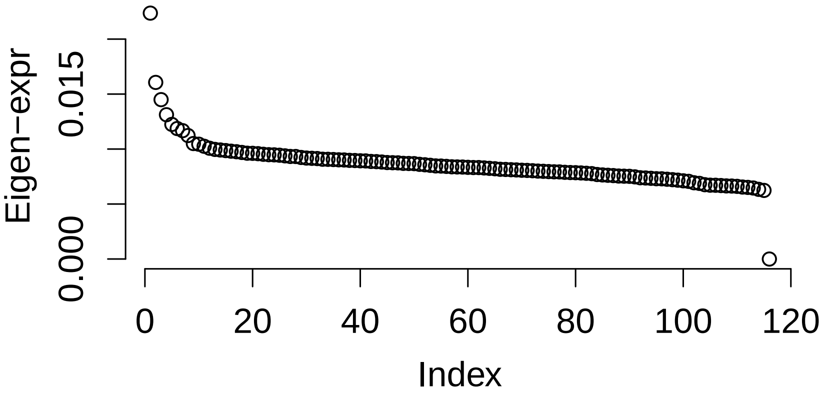


Figure S1: Eigen spectrum plot for separate Principal Component Analyses (PCA) applied to the SAE sample genotypes (top) and gene expression measurements (bottom), where the x-axis shows the index of the eigenvalues (eigenvectors) and the y-axis shows the variance proportion explained by each eigenvector.

Table S2: Population demographics of the human lung airway epithelium study

	Small airway epithelium[1]		
	Healthy non-smk[2]	Healthy smk[3]	Disease smk[4]
# of Samples	38	37	41
Gender (M/F)	28/10	22/15	31/10
Ethnicity (B/W)[6]	22/16	27/10	24/17
Age[7]	43±12	43±7	51±10
Pack-year history[8]	0	28±17	37±24

1. Data are presented as mean ± standard deviation where appropriate

2. Life-long nonsmokers with normal lung functions as measured by spirometry and diffusion capacity of carbon monoxide (Harvey *et al.*, 2008)

3. Current or ex-smokers with normal lung functions as measured by normal spirometry and diffusion capacity of carbon monoxide (Harvey *et al.*, 2008)

4. Current or ex-smokers with pulmonary disease as defined by their lung functions: either with Chronic Obstructive Pulmonary Disease (COPD) as defined by the GOLD criteria (Harvey *et al.*, 2008) or early emphysema as defined by normal spirometry and reduced diffusion capacity of carbon monoxide (<80%) (Harvey *et al.*, 2008)

5. African American (B=Black) or Caucasian (W=white)

6. Presented as mean ± standard deviation

7. Calculated for each individual as the number of packs of cigarettes smoked per day times the number of years of self-reported smoking history presented as mean ± standard deviation

For datasets simulated under scenario b, where there are no eQTL and hidden factors (null scenario 2), we considered performance for cases where the effects of the four hidden factors were (approximately) orthogonal to all SNPs and cases where the effects of the four hidden factors were non-orthogonal to 10% of the SNPs. For the orthogonal case, with the correct factor number ($p = 4$) all methods including LR produced an almost uniform distribution of p-values (λ=1-1.04) as expected (Table S3 and Figure S3). For the case of non-orthogonal hidden factors under this same null scenario b), the performance for LR diverged far from the null expectation where far too many small p-values were returned, a result that in practice would result in a large number of false positives (Table S3 and Figure S4). This result is expected given that linear regression is unable to distinguish an eQTL signal from the effects of hidden factors. Again, we note that we could not check LORS under this null scenario.

For the standard eQTL scenario c (50 non-pleiotropic eQTL with no hidden factors) and scenario d (50 non-pleiotropic eQTL and 20 pleiotropic eQTL affecting 20 expressed genes each with no hidden factors), HEFT and LR had equivalent performance as expected (Table S4 and Figure S5-S6). For scenario c (no pleiotropy), HEFT-TS, PEER, and PANAMA had equivalent performance. We note LORS also had equivalent performance for this scenario when considering AUC in the 0-0.001 and 0-0.01 FPR range, where the slightly lower performance in the 0-0.05 FPR range occurs because the method pre-selects markers to include by linear regression, which caps the maximum number of true positives that can be identified (and hence the True Positive

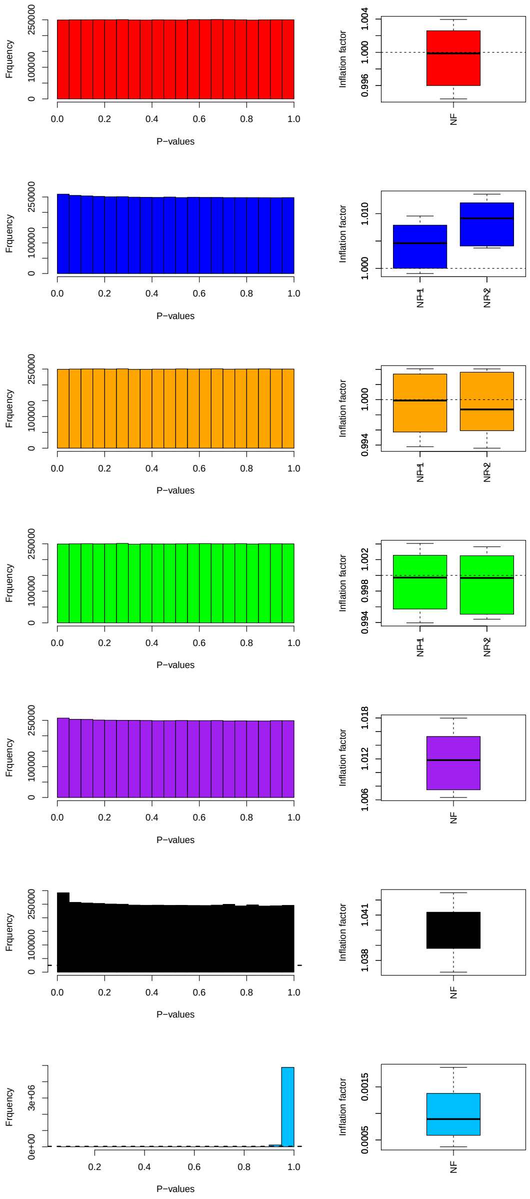


Figure S2: Histograms and boxplots showing the distributions of p-values for all SNP-gene tests of association for the scenario where there are no genotypic effects and no hidden factors (scenario a). The left column shows the histogram of the p-values for a specific simulation with factor number of 2 (when selection of factor number applies), and the right column shows the boxplots of the inflation factor for p-values for number of factors (NF) 1 and 2 for all ten simulations (again when factor number applies). From top to bottom are respectively LR (linear regression), HEFT, HEFT-TS, PEER, LMM, SVA and PANAMA. Note that LORS does not produce p-values and is therefore not included.

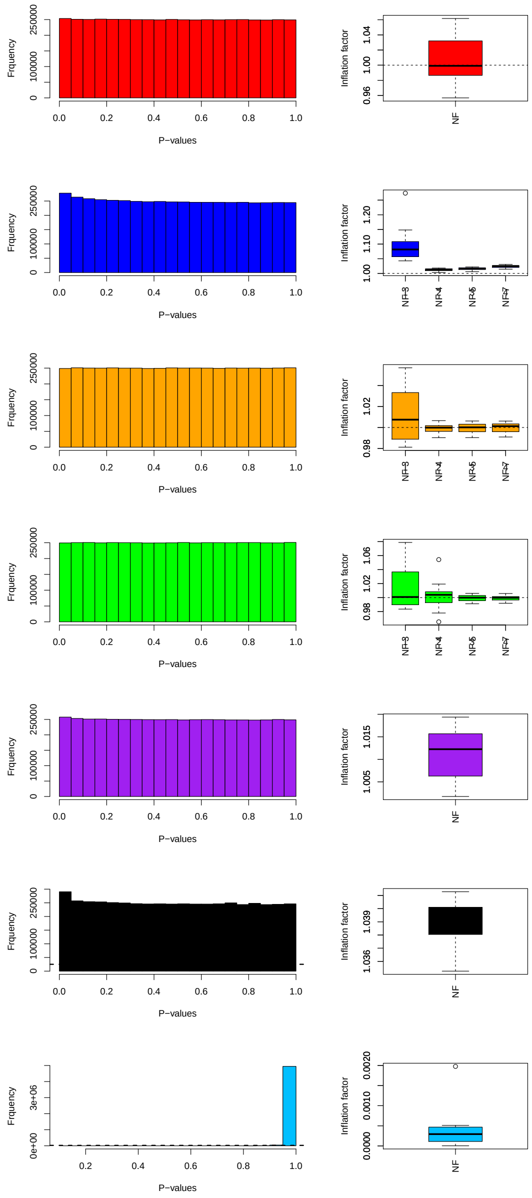


Figure S3: Histograms and boxplots showing the distributions of p-values for all SNP-gene tests of association for a scenario with no eQTL and hidden factors that are orthogonal to the SNPs (orthogonal scenario b). The left column shows the histogram of the p-values for a specific simulation with factor number of 7 (when selection of factor number applies), and the right column shows the boxplots of the inflation factor for p-values for number of factors (NF) 3, 4, 5, and 7 for all ten simulations. From top to bottom are respectively LR (linear regression), HEFT, HEFT-TS, PEER, LMM, SVA and PANAMA. Note that LORS does not produce p-values and is therefore not included.

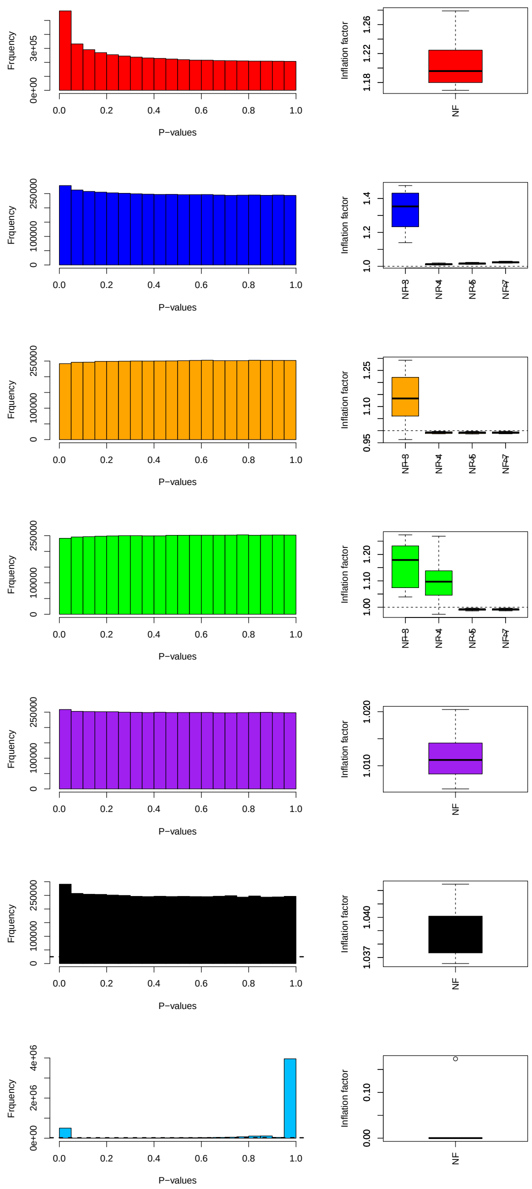


Figure S4: Histograms and boxplots showing the distributions of p-values for all SNP-gene tests of association for a scenario with no eQTL and hidden factors that are non-orthogonal to the SNPs (non-orthogonal scenario b). The left column shows the histogram of the p-values for a specific simulation with factor number of 7 (when selection of factor number applies), and the right column shows the boxplots of the inflation factor for p-values for number of factors (NF) 3, 4, 5, and 7 for all ten simulations. From top to bottom are respectively LR (linear regression), HEFT, HEFT-TS, PEER, LMM, SVA and PANAMA. Note that LORS does not produce p-values and is therefore not included.

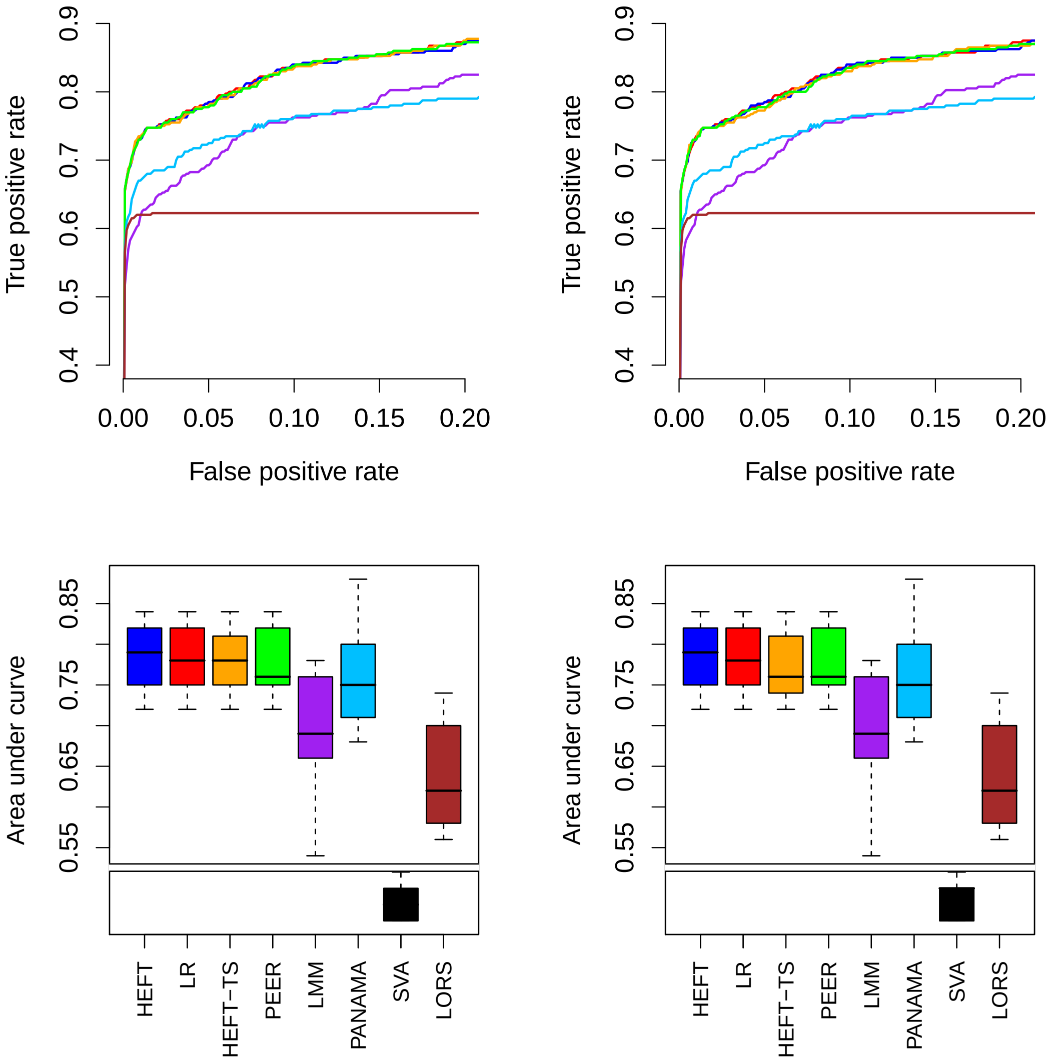


Figure S5: Average Receiver Operating Characteristic (ROC) curves (top) and boxplots of the area under the curve (AUC) for the ROC for a false positive rate in the range 0-0.05 (bottom) for simulated data in the case of non-pleiotropic eQTLs but no hidden factors (scenario c), where the left and right columns correspond to providing factor numbers of 1 and 2 (when factor number selection applies). The methods are color coded as: red=regression, blue=HEFT, orange=HEFT-TS, green=PEER, purple=LMM, skyblue=PANAMA, black=SVA, brown=LORS. Note the the leveling off of the ROC curve for LORS is a consequence of this method pre-selecting markers to include, which caps the maximum number of true positives that can be identified.

Table S3: The range of genomic inflation factor for genome-wide p-values returned by analysis methods when applied to data simulated under null scenarios a and b when being provided with different numbers of factors, where plots of a subset of these are presented in Figure S2-??. The values in the table show the average inflation factor for the analysis of each of the 10 simulated datasets for a given scenario with ± confidence intervals. For LR and LMM there is no factor number selection so the same numbers are repeated in each column. PANAMA is not shown as the output q values cannot be easily assessed with an inflation factor and LORS is not shown as it does not produce p-values

		a		b			
	#Factor	1	2	3	4	5	6
non-orthogonal	LR	1±0	1±0	1.21±0.04	1.21±0.04	1.21±0.04	1.21±0.04
	HEFT	1±0	1.01±0	1.33±0.12	1.01±0	1.02±0	1.02±0
	HEFT-TS	1±0	1±0	1.14±0.1	0.99±0	0.99±0	0.99±0
	PEER	1±0	1±0	1.16±0.08	1.1±0.08	0.99±0	0.99±0
	LMM	1.01±0	1.01±0	1.01±0	1.01±0	1.01±0	1.01±0
	SVA	1.04±0	1.04±0	1.04±0	1.04±0	1.04±0	1.04±0
orthogonal	LR	1±0	1±0	1±0.03	1±0.03	1±0.03	1±0.03
	HEFT	1±0	1.01±0	1.1±0.07	1.01±0	1.02±0	1.02±0
	HEFT-TS	1±0	1±0	1.01±0.02	1±0	1±0	1±0
	PEER	1±0	1±0	1.01±0.03	1±0.02	1±0	1±0
	LMM	1.01±0	1.01±0	1.01±0	1.01±0	1.01±0	1.01±0
	SVA	1.04±0	1.04±0	1.04±0	1.04±0	1.04±0	1.04±0

Rate) resulting in the leveling off of LORS performance in this higher FPR range. For scenario d (pleiotropy and no hidden factors), the two step methods (HEFT-TS, PEER) had equivalent performance when the assumed number of factors was <2 (where the true number was zero) but significantly worse performance than HEFT when the assumed number of factors was 2, indicating that fitting of additional hidden factors resulted in these two-step methods accounting for pleiotropic effects of eQTL, while the simultaneous fitting of eQTL and hidden factors in HEFT allows this method to be robust to the wrong number of factors. PANAMA and LORS also fit eQTL and hidden factor effects simultaneously but unlike scenario c (no pleiotropy and no hidden factors) when there are pleiotropic effects of eQTL in scenario d, these methods had significantly worse performance compared to HEFT. It therefore appears that PANAMA and LORS are fitting the pleiotropic effects of eQTL as hidden factors, where we discuss possible reasons for this behavior in the next section (S.3.1.2). We note that SVA and LMM also had significantly worse performance than HEFT, which we also suspect is due to over-fitting of these methods. For SVA this may be a consequence of the permutation approach for determining the number of factors while for LMM, it is possible that fitting a full rank random effect in the mixed model accounts for some of the eQTL effects and reduces the power of the method. We also discuss possible reasons in section S.3.1.2 below.

S.3.1.2 Performance for eQTL and hidden factors.

For the scenarios where there are both eQTL and hidden factors, performance depended heavily on the type of eQTL effects, specifically whether there was no pleiotropy (scenario e) or pleiotropy (scenario f). In scenario e where there were 50 eQTL each affecting an individual gene expression level (no pleiotropy), HEFT outperformed LR as expected (as measured by AUC) but had qualitatively equivalent performance compared to HEFT-TS, PEER, and PANAMA (Table S4), where these results were consistent regardless of whether the hidden factors were orthogonal or non-orthogonal to the eQTL effects and regardless of whether the correct (4) or incorrect number

Table S4: Comparisons of HEFT to each of the other methods when applying two-sided t tests to the area under the ROC curves (AUC), when considering False Positive Rates (FPR) in the range 0-0.05. The average total number of significant hits recovered is also provided, where each non-duplicated SNP-gene pairing found to have a significant association is considered a hit. The comparisons are performed when HEFT is provided two different factor numbers, where the correct number of factors for scenario c (non-pleiotropic eQTL and no hidden factors) and scenario d (both non-pleiotropic / pleiotropic eQTL and no hidden factors) is zero and the correct number of factors for scenario e (non-pleiotropic eQTL and no hidden factors) and scenario d (both non-pleiotropic / pleiotropic eQTL and hidden factors) is four. Note that for the non-pleiotropic scenarios there are 50 total eQTL associations and for non-pleiotropic / pleiotropic scenarios there are 450 total eQTL associations (50 non-pleiotropic eQTL and 20 pleiotropic eQTL each affecting 20 gene expression levels). Also note that the each AUC comparison and average number of hits is over 10 simulated data sets such that the same average number of hits can correspond to a different p-value depending on the standard error. Finally, note that SVA is not included as the performance was far worse than the other methods.

				HEFT	LR	HEFT-TS	PEER	LMM	PANAMA	LORS
c	Non-orth	df=1	p-value		0.765	0.848	0.843	0.00185	0.0576	1.02e-05
			#Hits	40.7	41	40.9	40.9	36	38.6	33.6
		df=2	p-value		0.765	1	0.843	0.00185	0.0576	1.02e-05
			#Hits	40.7	41	40.7	40.9	36	38.6	33.6
	Orth	df=1	p-value		0.907	0.812	0.732	0.0133	0.389	0.000236
			#Hits	39.2	39.1	39	38.9	34.6	38	31.9
		df=2	p-value		0.907	0.566	0.732	0.0133	0.389	0.000236
			#Hits	39.2	39.1	38.6	38.9	34.6	38	31.9
				HEFT	LR	HEFT-TS	PEER	LMM	PANAMA	LORS
d	Non-orth	df=1	p-value		1	0.159	0.116	3.77e-12	4.86e-07	6.58e-08
			#Hits	348	348	343	344	221	292	305
		df=2	p-value		0.864	0.00881	0.000509	6.2e-12	5.4e-07	6.69e-08
			#Hits	349	348	340	337	221	292	305
	Orth	df=1	p-value		0.831	0.257	0.275	5.98e-14	1.5e-09	8.68e-06
			#Hits	351	350	346	347	226	294	312
		df=2	p-value		0.722	0.0189	0.0193	5.51e-14	1.26e-09	7.22e-06
			#Hits	352	350	341	341	226	294	312
				HEFT	LR	HEFT-TS	PEER	LMM	PANAMA	LORS
e	Non-orth	df=4	p-value		0.00348	0.867	0.15	0.00925	0.0816	1.82e-05
			#Hits	40.4	34.9	40.7	38.6	36.6	37.8	31.4
		df=7	p-value		0.00265	0.927	0.93	0.00595	0.0629	1.3e-05
			#Hits	40.6	34.9	40.7	40.7	36.6	37.8	31.4
	Orth	df=4	p-value		0.0236	1	0.396	0.0363	0.179	0.000272
			#Hits	39.6	36.1	39.6	38.4	35.6	37.2	32
		df=7	p-value		0.0375	0.933	0.931	0.0531	0.25	0.000398
			#Hits	39.2	36.1	39.1	39.4	35.6	37.2	32
				HEFT	LR	HEFT-TS	PEER	LMM	PANAMA	LORS
f	Non-orth	df=4	p-value		7.9e-08	0.861	0.00333	4.74e-13	1.83e-10	2.41e-08
			#Hits	346	312	347	336	229	293	298
		df=7	p-value		8.26e-08	0.00293	0.000235	2.01e-13	1.8e-10	1.73e-08
			#Hits	347	312	332	332	229	293	298
	Orth	df=4	p-value		1.45e-05	0.934	0.00169	4.85e-11	1.84e-07	1.15e-05
			#Hits	349	325	349	337	235	293	300
		df=7	p-value		7.85e-06	0.000304	7.27e-05	1.23e-10	2.65e-07	1.37e-05
			#Hits	350	325	332	334	235	293	300

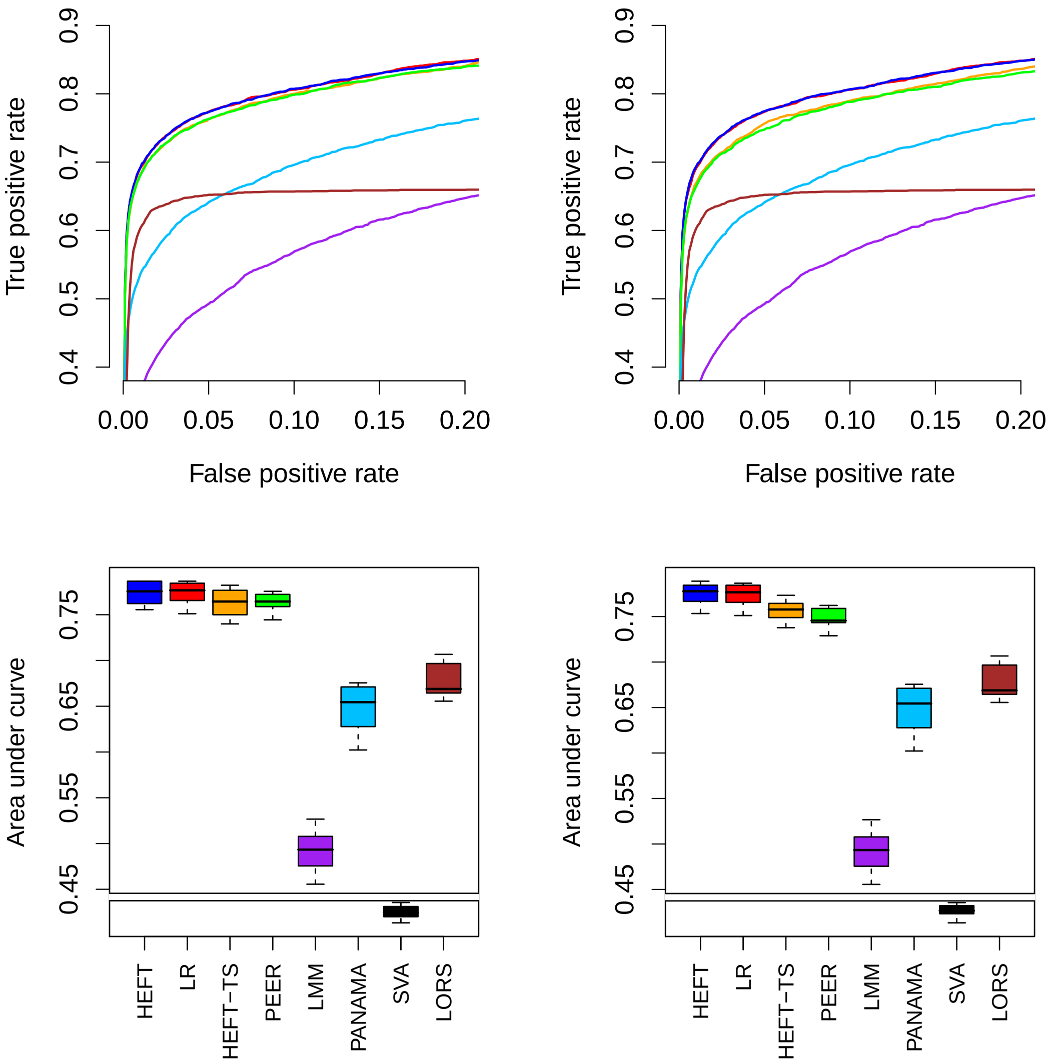


Figure S6: Average Receiver Operating Characteristic (ROC) curves (top) and boxplots of the area under the curve (AUC) for the ROC for a false positive rate in the range 0-0.05 (bottom) for simulated data in the case of pleiotropic eQTLs but no hidden factors (scenario d), where the left and right columns correspond to providing factor numbers of 1 and 2 (when factor number selection applies). The methods are color coded as: red=regression, blue=HEFT, orange=HEFT-TS, green=PEER, purple=LMM, skyblue=PANAMA, black=SVA, brown=LORS. Note the the leveling off of the ROC curve for LORS is a consequence of this method pre-selecting markers to include, which caps the maximum number of true positives that can be identified.

(3, 5, 7) of hidden factors was considered by HEFT (Figures S7-S8). We note that the hidden factor method LORS also had qualitatively equivalent performance to HEFT for scenario e when considering AUC in the 0-0.001 and 0-0.01 FPR range, where the leveling off of LORS performance occurs because this method pre-selects markers to include by linear regression, which caps the maximum number of true positives that can be identified (TPR) and therefore leading to the appearance of slightly lower performance when measured by AUC in the 0-0.05 FPR range. Interestingly, while HEFT and these other hidden factor methods had comparable performance for scenario e, the hidden factor methods SVA and LMM, had significantly worse performance (Table S4 and Figures S7-S8). In the case of SVA, we suspect this is due to fitting too many hidden factors as selected using their default permutation approach, although we note that it is not entirely clear why their factor selection heuristic should be over-fitting for this scenario. For LMM, the lower performance is almost certainly due to over-fitting as a consequence of using a full rank random effect matrix in the mixed model, since PANAMA, which allows for reduced rank random effects in the same framework, had equivalent performance compared to HEFT.

For scenario f, where there are a combination of 50 non-pleiotropic and 20 pleiotropic eQTL (where each of the pleiotropic eQTL affect 20 expressed genes) and either orthogonal or non-orthogonal hidden factors, HEFT had significantly better performance overall compared to LR and all hidden factor methods (Table S4 and Figures S9-S10). We note that when the two-step form HEFT-TS considers the correct number of factors (4), the performance of HEFT and HEFT-TS is equivalent as expected. However, when assuming too many factors (7), the simultaneous fitting of eQTL and hidden factors in HEFT allows this method to correctly reduce the variance attributable to these extra hidden factors, such that they do not fit the pleiotropic effects of the eQTLs, where HEFT-TS cannot make this correction and performs significantly worse. The simultaneous fitting in HEFT also explains why this method outperforms the two-step method PEER, where the difference in performance is more significant when HEFT and PEER assume too many hidden factors. In the cases of SVA and LMM, the better performance of HEFT is attributable to the same reasons for the observed better performance in scenario e where there is no pleiotropy, i.e. possible over-fitting of hidden factors using the permutation approach of SVA and over-fitting of hidden factor effects in LMM due to the full rank of the random effects in the mixed model.

Why HEFT has significantly better performance compared to PANAMA for scenario f is not as clear. We suspect this result is due to PANAMA fitting the effects of some of the pleiotropic eQTL as hidden factors but given that PANAMA simultaneous fits eQTL and hidden factors it is not as clear why this method cannot as accurately fit eQTL even when assuming too many hidden factors. One possibility is that PANAMA simultaneously fits a subset of eQTL at the same time and this could impact performance for identifying pleiotropic effects. Another possibility is that by integrating over both eQTL and hidden factor effects in the mixed model framework used by PANAMA, the resulting pooling of the hidden factor and error covariance (see section S.1.1 above) can lead to over-estimates of the variance component terms associated with hidden factors. However, disentangling the precise theoretical reasons for the better performance of HEFT compared to PANAMA for this scenario is a complex problem that is beyond the scope of this paper.

Similarly, it appears that LORS fits some of the pleiotropic eQTL effects in scenario f as hidden factors leading to significantly worse performance than HEFT regardless of the AUC considered (0-0.001, 0-0.1, 0-0.05) where we note that this is not a consequence of the marker pre-selection of LORS (the leveling off of the ROC curve). Again, it is not completely clear why this simultaneous method cannot correctly fit the pleiotropic effects of eQTL in this scenario. One possibility is that LORS places a penalty on the nuclear norm accounting for factor effects (the entire $\Lambda\mathbf{F}$ matrix in HEFT), as opposed to HEFT which places a penalty on the factor effects $\mathbf{F}$. LORS therefore penalizes all factors as a group (i.e. the model penalty does not have the flexibility to explicitly consider fitting of different numbers of factors), while HEFT separately penalizes the effects of each factor. This separate penalization in HEFT may induce a greater overall penalty and may prevent over-fitting, although determining whether this is the case and the theoretical basis if so, is a complex problem beyond the scope of the current work. Another possibility is that, LORS simultaneously fits a subset of eQTL at the same time making use of a lasso penalty or considers one SNP marker at a time, where for the latter approach, the method applies no penalty. In contrast HEFT always fits one marker at a time and always applies a ridge penalty to the possible genetic effects such that the genetic effects and the factors are regularized in the same way. We have observed that placing a ridge penalty on both the marker under consideration and the hidden factor effects is essential to achieving appropriate model fit in cases of pleiotropy for HEFT and this may therefore lead to the better performance of HEFT compared to LORS in scenario f. However, again, determining the precise theoretical reasons for this result is a complex problem beyond the scope of this current work.

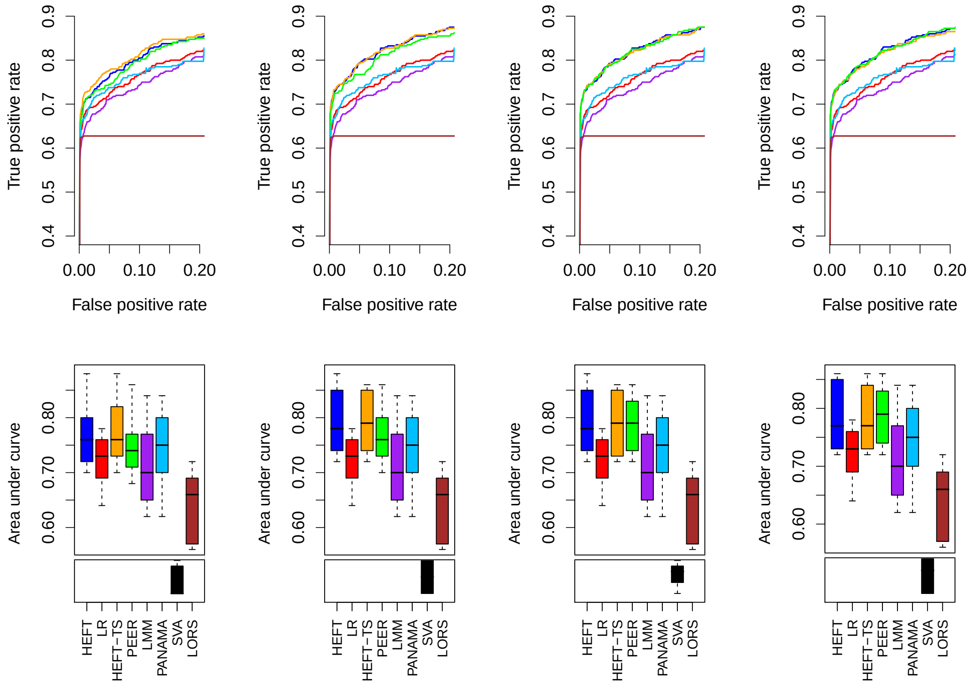


Figure S7: Average Receiver Operating Characteristic (ROC) curves (top) and boxplots of the area under the curve (AUC) for the ROC for a false positive rate in the range 0-0.05 (bottom) for simulated data in the case of non-pleiotropic eQTL effects and orthogonal hidden factors (scenario e), where from left to right correspond to provided factor numbers of 3, 4, 5, and 7 respectively (when factor number selection applies). The methods are color coded as: red=regression, blue=HEFT, orange=HEFT-TS, green=PEER, purple=LMM, skyblue=PANAMA, black=SVA, brown=LORS. Note the the leveling off of the ROC curve for LORS is a consequence of this method pre-selecting markers to include, which caps the maximum number of true positives that can be identified.

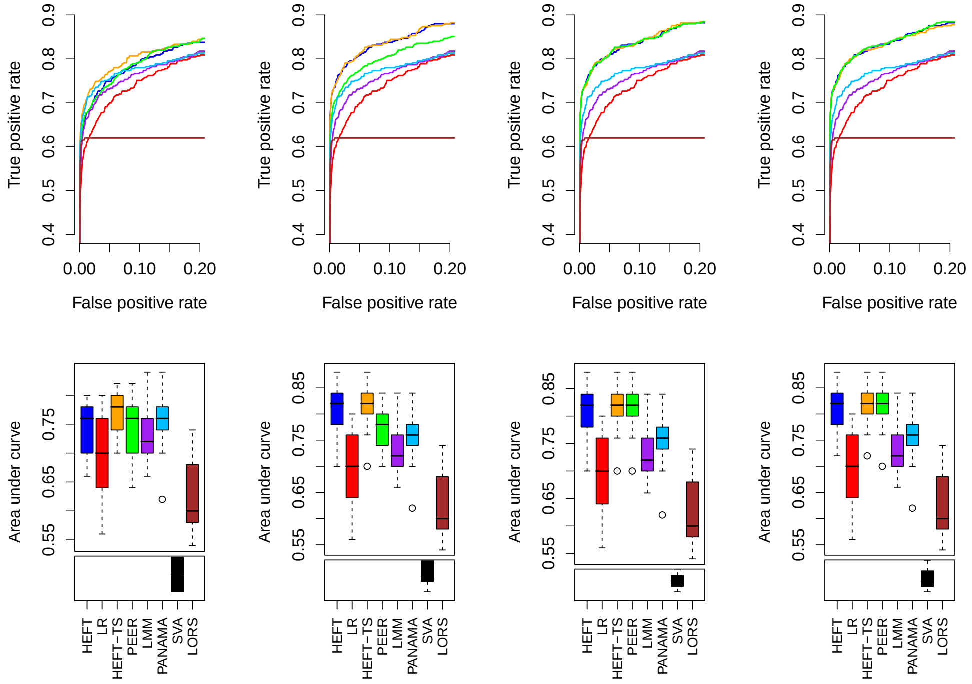


Figure S8: Average Receiver Operating Characteristic (ROC) curves (top) and boxplots of the area under the curve (AUC) for the ROC for a false positive rate in the range 0-0.05 (bottom) for simulated data in the case of non-pleiotropic eQTL effects and non-orthogonal hidden factors (scenario e), where from left to right correspond to provided factor numbers of 3, 4, 5, and 7 respectively (when factor number selection applies). The methods are color coded as: red=regression, blue=HEFT, orange=HEFT-TS, green=PEER, purple=LMM, skyblue=PANAMA, black=SVA, brown=LORS. Note the the leveling off of the ROC curve for LORS is a consequence of this method pre-selecting markers to include, which caps the maximum number of true positives that can be identified.

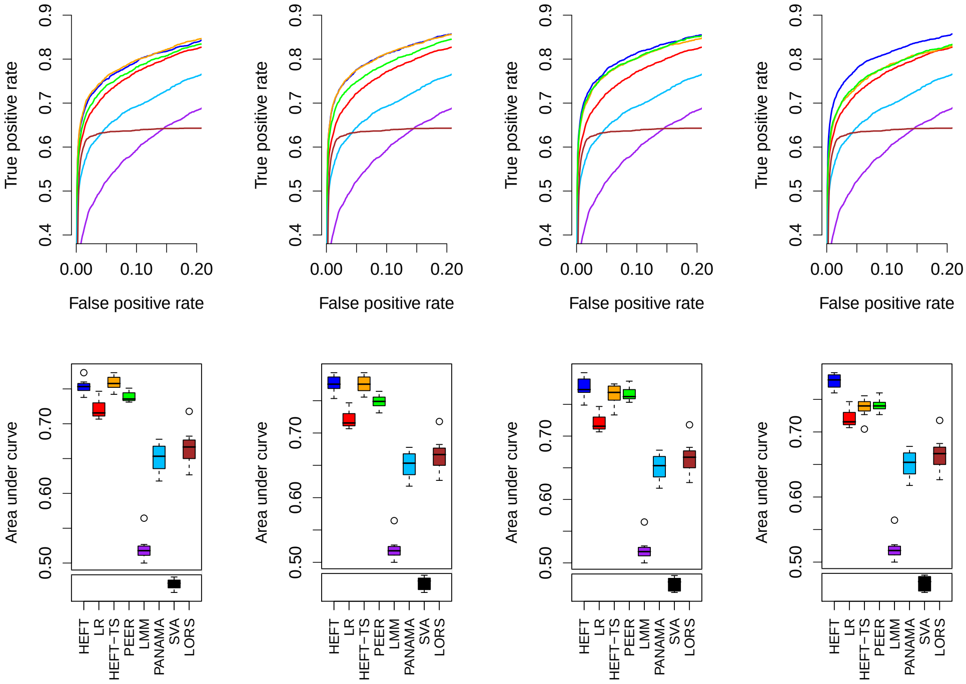


Figure S9: Average Receiver Operating Characteristic (ROC) curves (top) and boxplots of the area under the curve (AUC) for the ROC for a false positive rate in the range 0-0.05 (bottom) for simulated data in the case of pleiotropic eQTL effects and orthogonal hidden factors (scenario f), where from left to right correspond to provided factor numbers of 3, 4, 5, and 7 respectively (when factor number selection applies). The methods are color coded as: (red=regression, blue=HEFT, orange=HEFT-TS, green=PEER, purple=LMM, skyblue=PANAMA, black=SVA, brown=LORS. Note the the leveling off of the ROC curve for LORS is a consequence of this method pre-selecting markers to include, which caps the maximum number of true positives that can be identified.

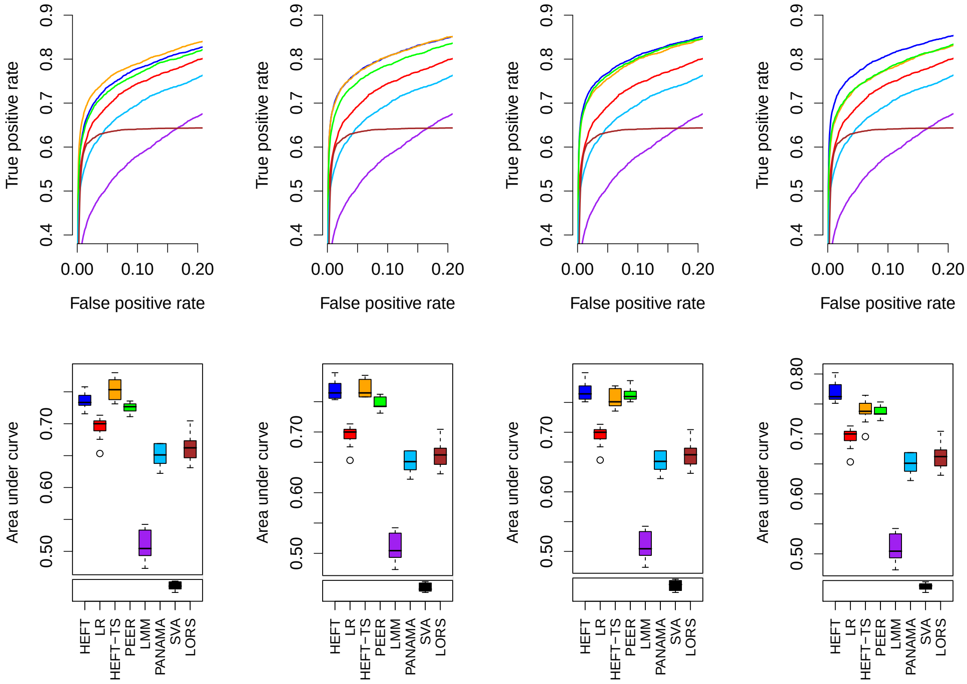


Figure S10: Average Receiver Operating Characteristic (ROC) curves (top) and boxplots of the area under the curve (AUC) for the ROC for a false positive rate in the range 0-0.05 (bottom) for simulated data in the case of pleiotropic eQTL effects and non-orthogonal hidden factors (scenario f), where from left to right correspond to provided factor numbers of 3, 4, 5, and 7 respectively (when factor number selection applies). The methods are color coded as: (red=regression, blue=HEFT, orange=HEFT-TS, green=PEER, purple=LMM, skyblue=PANAMA, black=SVA, brown=LORS. Note the the leveling off of the ROC curve for LORS is a consequence of this method pre-selecting markers to include, which caps the maximum number of true positives that can be identified.

S.3.1.3 Recovery of the smoking factor when treated as hidden

As an empirical assessment of the ability of HEFT to recover hidden factors, we used the factor learning component of HEFT to analyze the lung SAE gene expression data, where the known information about whether individuals were smokers or nonsmokers was treated as missing. Smoking has a well-characterized effect throughout the SAE transcriptome. For this analysis, HEFT was able to learn the effects of smoking status when this covariate was treated as a hidden factor, providing good separation of smokers and nonsmokers (Figure S11). From the analysis, the influence of smoking appears to be more complex than could be well modeled with a single fixed covariate, indicating that even in the unusual case where the critical factors are known and measured, it may be of value to learn hidden factors in an eQTL analysis.

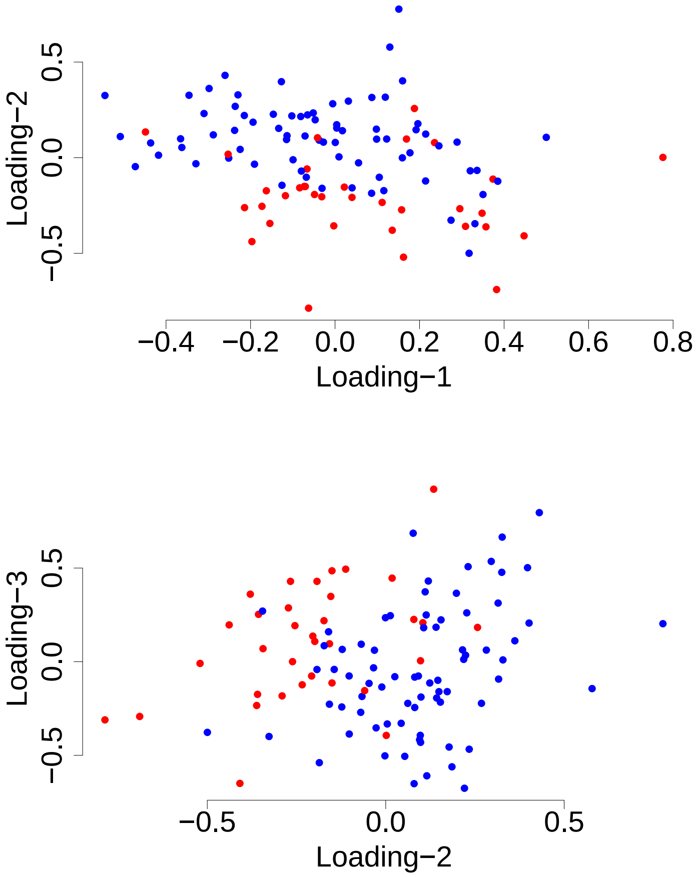


Figure S11: Plot showing the separation of smokers (blue) and nonsmokers (red) plotted on the hidden factors learned by HEFT, where loading 1 and 2 for the factor are plotted on the top and loading 2 and 3 are plotted at the bottom.

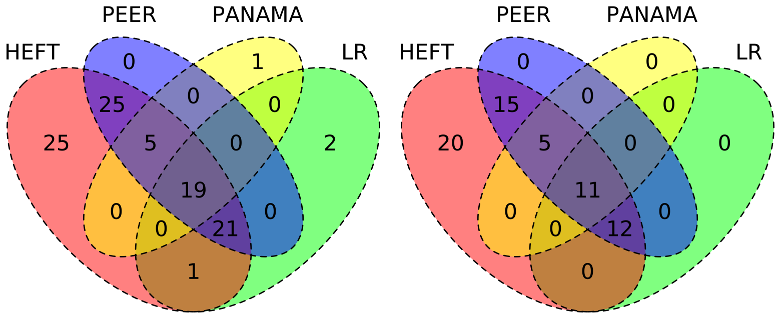


Figure S12: Venn diagram showing the total number of non-duplicate SNP-gene associations (left) and the subset of these that are *cis-* (right) identified by HEFT, PEER, PANAMA and LR when applying a Bonferroni correction for multiple tests.

Table S5: List of the 96 non-duplicated top eQTL associations identified by HEFT significant at a Bonferroni threshold, where only the top associated SNPs are listed. From left to right the columns represent respectively the ranking, ensemble code for the genes, the top SNP associated with this gene, and the p-values.

Ranking	Gene ID	Chr(SNP)	Position	SNP	P-Values
1	5906	1	112038561	rs1886498	1.46499485948337e-33
2	140686	20	43836276	rs2664529	2.27127502421512e-31
3	89778	18	59530818	rs4940595	7.75169142022207e-28
4	374491	13	24078151	rs943049	1.50554696527431e-25
5	80177	6	153052613	rs2250514	1.74101297874152e-25
6	6840	0	0	rs3858231	2.12530111251114e-25
7	63928	16	23697839	rs194788	8.46873720736218e-25
8	23421	1	63837368	rs855325	3.53979830579298e-24
9	114757	17	72058534	rs752049	7.33012802900279e-24
10	388335	17	10556076	rs397278	6.39394600924099e-23
11	1915	6	74287580	rs3822960	7.95367824562483e-22
12	5340	6	161064326	rs1321200	6.63854323347397e-21
13	116285	16	20587707	rs433598	2.02376243218272e-20
14	155368	7	72863628	rs4355658	2.35431495370645e-20
15	340542	23	101168380	rs2858353	2.61556705576623e-20
16	8000	8	143761003	rs2976396	5.75349616679876e-20
17	164781	2	228485182	rs3748863	6.47663988110386e-20
18	51144	11	43796511	rs10768983	9.78888533972432e-20
19	318	9	34271390	rs7045680	1.26766986219884e-19
20	5947	3	140736561	rs12485273	1.86700120545829e-19
21	26751	2	270819	rs7605824	4.43726175926121e-19
22	403314	1	181889002	rs6699011	1.26809312494076e-18
23	7976	8	28491587	rs11779401	1.77971301988934e-18
24	90637	7	1171226	rs2960840	2.07643459155093e-18

25	25961	10	74540513	rs2280369	5.3059275592548e-18
26	55278	6	107221054	rs1026619	1.17621599828576e-17
27	1965	14	66916971	rs8008724	1.33675165066165e-17
28	150142	21	42317483	rs220219	1.64955608207078e-17
29	286464	23	36068724	rs16987374	1.77494728717695e-17
30	22948	5	10318076	rs699113	1.86493789973556e-17
31	158158	9	84807488	rs1502682	3.86943395919504e-17
32	5268	18	59295033	rs3744941	5.34526015688907e-17
33	26503	6	74399417	rs9446964	8.09618716653111e-17
34	26090	20	25223843	rs2258719	1.45642163092577e-16
35	54879	1	112933916	rs6666579	1.66651287593059e-16
36	10230	17	38773860	rs4793229	1.72080169717462e-16
37	100506015	2	161907239	rs10197817	2.54805768511242e-16
38	100507580	22	24215353	rs6004673	4.35569747641881e-16
39	91612	14	64477410	rs2412065	4.44742133568217e-16
40	84930	10	27494117	rs7068375	7.17212767471488e-16
41	4649	15	69903832	rs2742323	1.59951937168628e-15
42	84545	10	102724768	rs4919510	2.19436097956699e-15
43	10781	19	9394805	rs6512121	2.6373840456275e-15
44	388407	17	56851431	rs2079795	2.91445531827876e-15
45	5889	17	54195586	rs8074016	3.72419479226112e-15
46	25854	4	187345974	rs4586997	3.77686506053501e-15
47	80150	11	61840106	rs1406384	7.16689214689067e-15
48	100131564	1	93561736	rs7555292	1.02301175982386e-14
49	55034	18	31980483	rs3737468	2.41283860647574e-14
50	121506	12	14961902	rs2193356	2.68918264415563e-14
51	26999	5	156687851	rs13155266	4.73999050614003e-14
52	100507540	9	74103814	rs7874628	4.77268615823305e-14
53	10558	9	93804054	rs7045602	5.39021021351054e-14
54	51531	9	99670770	rs7357707	5.90037022855741e-14
55	2882	1	52836600	rs835341	6.04084235032921e-14
56	54960	23	13929093	rs7055913	7.41803216920483e-14
57	27030	14	74615438	rs175490	9.02909469073108e-14
58	51703	10	114138679	rs12255316	1.16488754370987e-13
59	55020	22	45066872	rs6008552	1.42090376672065e-13
60	538	23	77217818	rs2643591	1.58701624661496e-13
61	146562	16	4734874	rs2075469	1.70605590922464e-13
62	6263	15	31881493	rs2115747	2.80777022150772e-13
63	100507316	8	144410365	rs7824894	2.90777614923962e-13
64	51816	22	16062294	rs1076106	3.04711413440298e-13
65	8624	21	39555501	rs2836965	3.54602705122609e-13
66	143241	10	82062290	rs10788562	4.6114142614977e-13
67	3631	2	98402962	rs17031139	5.2297407172801e-13
68	51016	14	23681818	rs3742500	5.62602288154044e-13
69	622	3	198745276	rs13077136	1.03436568736022e-12
70	201651	3	152973681	rs4679934	1.28939628281059e-12
71	133383	5	56214969	rs832584	1.92293301156847e-12
72	2965	11	18322286	rs4150622	2.14534949336851e-12
73	55253	7	66219599	rs17144722	2.52141749626425e-12

74	84221	21	46524284	rs2839195	2.64339920859971e-12
75	151525	2	159886945	rs174227	3.8214945233408e-12
76	643529	10	91582178	rs1125326	4.19750709824653e-12
77	64105	5	64951379	rs2161278	4.27951343576203e-12
78	654433	2	113700504	rs3748916	4.56219494155033e-12
79	100506707	2	113705738	rs2863243	6.6617384815837e-12
80	55728	4	39796815	rs17619330	6.95256644737194e-12
81	401491	9	2527815	rs588933	8.40770283486909e-12
82	80868	6	30026415	rs2508037	8.68017934939448e-12
83	7180	6	49811816	rs597544	1.10976804567039e-11
84	55125	18	12967206	rs8088313	1.21441064557634e-11
85	54847	3	114737833	rs4580515	1.24583836679942e-11
86	57545	4	15091907	rs6810461	1.45170694299707e-11
87	11102	3	58282684	rs6777105	1.45641838150676e-11
88	284323	19	45181310	rs8105066	1.52901564045426e-11
89	10783	9	126069557	rs12379417	1.59002247514929e-11
90	55256	2	3488694	rs9750132	1.62436743738252e-11
91	145957	15	73938020	rs2593280	1.89118641889377e-11
92	6006	1	25641524	rs10903129	2.2380711377044e-11
93	55733	1	208584705	rs6696657	2.34656975928403e-11
94	128344	1	111697010	rs1058530	2.58389549266189e-11
95	79772	5	93992505	rs10052066	2.95141473257443e-11
96	79618	8	28881393	rs4732896	3.24184877322828e-11

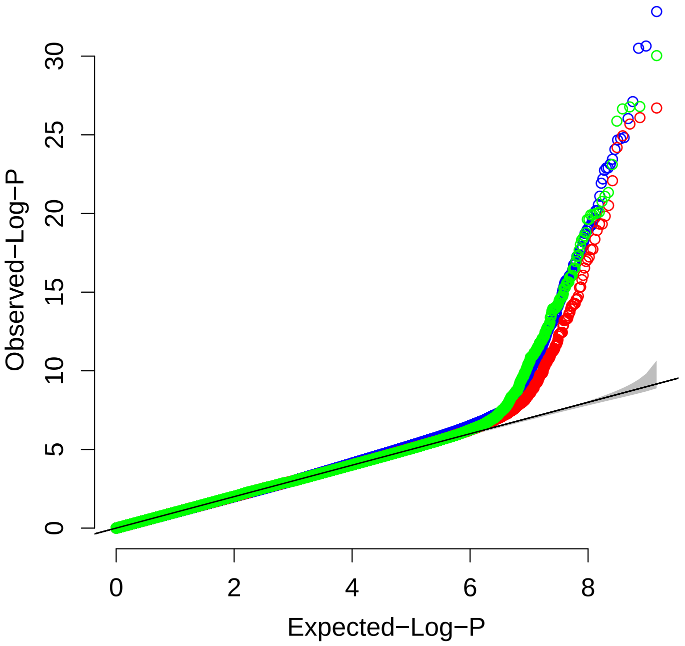


Figure S13: QQ plots showing the p-value distribution for all tests of association between the 191,959 SNPs and 7,575 genes (points have been thinned) expressed in human lung SAE for HEFT (blue), PEER (green), and LR (red). Note that PANAMA returns p-values automatically corrected for multiple tests by calculating q values, such that producing a QQ plot is not possible for this method. The grey bands in the QQ plots correspond to the 95% confidence interval (CI) of the order statistics, where we note that this is the CI for 1.45e9 p-values such that it is quite tight, and the slight early deviation of all methods beyond the estimated CI (including LR) is therefore likely an artifact and not a reflection of poor model fit.

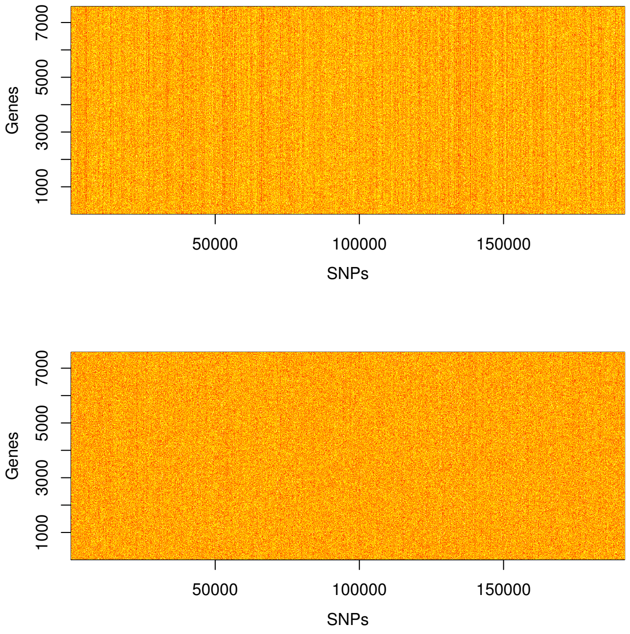


Figure S14: A genome-wide heat map representing p-values obtained from the analysis of associations of all 191,959 SNPs with all 7,575 genes expressed in human lung SAE using LR (top) and HEFT (bottom) where the p-values have been averaged for every 15 genes and 100 SNPs. Genes are arranged in rows and SNPs arranged in columns, where colors from yellow to red represent large to small (significant) p-values. Note that LR identified SNPs associated with almost all expressed genes, indicating unaccounted for hidden factors, where this trend is not observed with HEFT.

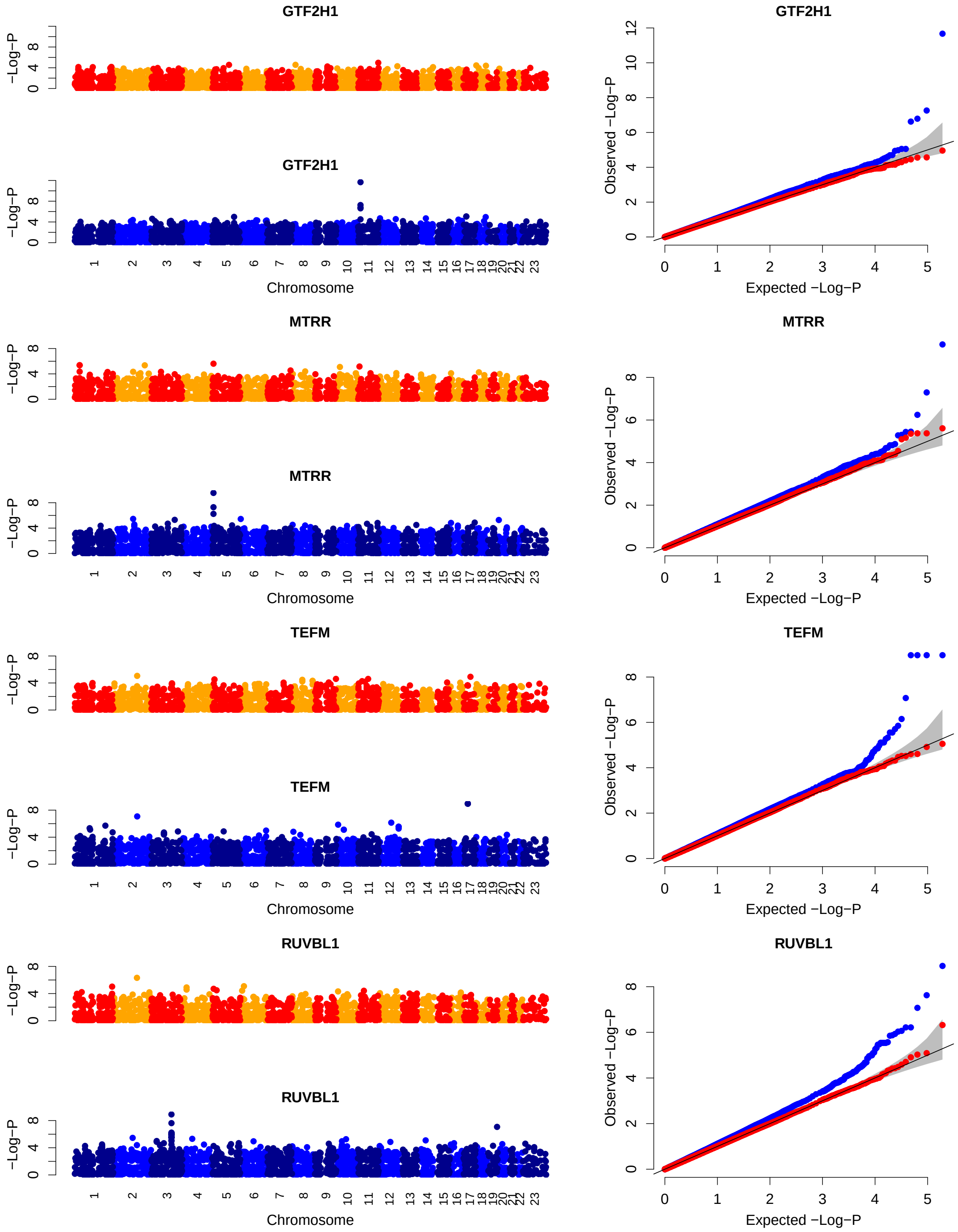


Figure S15: Manhattan plots (left column) and QQ plots (right column) for example genes where HEFT (blue plots) identified a significant *cis*-eQTL for a gene with a lung related phenotype or disease association that were not identified by a linear regression (red and orange/yellow plots), where the genes ordered from the top to bottom are: GTF2H1, MTRR, TEFM (C17orf42), and RUVBL1. The grey band corresponds to the 95% confidence interval.

References

Aulchenko, Y., Ripke, S., Isaacs, A., and van Duijn, C. (2007). Genabel: an r library for genome-wide association analysis. *Bioinformatics (Oxford, England)*, **23**(10), 1294–1296.

Bishop, C. (2007). *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer, 1st ed. 2006. corr. 2nd printing edition.

Chen, G., Marjoram, P., and Wall, J. (2009). Fast and flexible simulation of dna sequence data. *Genome research*, **19**(1), 136–142.

Dai, M., Wang, P., Boyd, A., Kostov, G., Athey, B., Jones, E., Bunney, W., Myers, R., Speed, T., Akil, H., Watson, S., and Meng, F. (2005). Evolving gene/transcript definitions significantly alter the interpretation of genechip data. *Nucleic Acids Research*, **33**(20), e175–e175.

Devlin, B., Bacanu, S., and Roeder, K. (2004). Genomic control to the extreme. *Nat Genet*, **36**(11).

Edgar, R., Domrachev, M., and Lash, A. (2002). Gene expression omnibus: Ncbi gene expression and hybridization array data repository. *Nucleic Acids Research*, **30**(1), 207–210.

Engelhardt, B. and Stephens, M. (2010). Analysis of population structure: A unifying framework and novel methods based on sparse factor analysis. *PLoS Genetics*, **6**(9), e1001117.

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

Harvey, Ben-Gary, Heguy, Adriana, Leopold, Philip, Carolan, Brendan, Ferris, Barbara, Crystal, and Ronald (2008). Modification of gene expression of the small airway epithelium in response to cigarette smoking. *Journal of Molecular Medicine*, **86**(7), 853–853.

Irizarry, R., Hobbs, B., Collin, F., Beazer-Barclay, Y., Antonellis, K., Scherf, U., and Speed, T. (2003a). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, **4**(2), 249–264.

Irizarry, R., Bolstad, B., Collin, F., Cope, L., Hobbs, B., and Speed, T. (2003b). Summaries of affymetrix genechip probe level data. *Nucleic acids research*, **31**(4), e15.

Kang, H. M., Zaitlen, N., Wade, C., Kirby, A., Heckerman, D., Daly, M., and Eskin, E. (2008). Efficient control of population structure in model organism association mapping. *Genetics*, **178**(3), 1709–1723.

Leek, J. and Storey, J. (2007). Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet*, **3**(9), e161–1735.

Listgarten, J., Kadie, C., Schadt, E., and Heckerman, D. (2010). Correction for hidden confounders in the genetic analysis of gene expression. *Proceedings of the National Academy of Sciences of the United States of America*, **107**(38), 16465–16470.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M., Bender, D., Maller, J., Sklar, P., de Bakker, P., Daly, M., and Sham, P. (2007). Plink: a tool set for whole-genome association and population-based linkage analyses. *American journal of human genetics*, **81**(3), 559–575.

Raman, T., O'Connor, T., Hackett, N., Wang, W., Harvey, B.-G., Attiyeh, M., Dang, D., Teater, M., and Crystal, R. (2009). Quality control in microarray assessment of gene expression in human airway epithelium. *BMC Genomics*, **10**, 493.

Rubin, D. and Thayer, D. (1982). Em algorithms for ml factor analysis. *Psychometrika*, **47**(1), 69–76.

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.

Voight, B., Adams, A., Frisse, L., Qian, Y., Hudson, R., and Di Rienzo, A. (2005). Interrogating multiple aspects of variation in a full resequencing data set to infer human population size changes. *PNAS*, **102**(51), 18508–18513.

Wigginton, J., Cutler, D., and Abecasis, G. (2005). A note on exact tests of hardy-weinberg equilibrium. *American journal of human genetics*, **76**(5), 887–893.

Yang, C., Wang, L., Zhang, S., and Zhao, H. (2013). Accounting for non-genetic factors by low-rank representation and sparse regression for eqtl mapping. *Bioinformatics*.