Bayesian Detection of Expression Quantitative Trait Loci Hot Spots

Leonardo Bottolo,*,† Enrico Petretto,*,† Stefan Blankenberg,‡ François Cambien,§ Stuart A. Cook,*, Laurence Tiret,§ and Sylvia Richardson†,††,1**

*MRC Clinical Sciences Centre, Imperial College, London W12 0NN United Kingdom, †Department of Epidemiology and Biostatistics, Imperial College, London W2 1PG, United Kingdom, ‡University Heart Center, D-20246 Hamburg, Germany, §INSERM UMRS 937, Pierre and Marie Curie University, 75013 Paris, France, **National Heart and Lung Institute, Imperial College, London W2 1PG, United Kingdom, and ††MRC–HPA Centre for Environment and Health, Imperial College, London-Harefield Hospital, Harefield, Middlesex UB9 6JH, United Kingdom

ABSTRACT High-throughput genomics allows genome-wide quantification of gene expression levels in tissues and cell types and, when combined with sequence variation data, permits the identification of genetic control points of expression (expression QTL or eQTL). Clusters of eQTL influenced by single genetic polymorphisms can inform on hotspots of regulation of pathways and networks, although very few hotspots have been robustly detected, replicated, or experimentally verified. Here we present a novel modeling strategy to estimate the propensity of a genetic marker to influence several expression traits at the same time, based on a hierarchical formulation of related regressions. We implement this hierarchical regression model in a Bayesian framework using a stochastic search algorithm, HESS, that efficiently probes sparse subsets of genetic markers in a high-dimensional data matrix to identify hotspots and to pinpoint the individual genetic effects (eQTL). Simulating complex regulatory scenarios, we demonstrate that our method outperforms current state-of-the-art approaches, in particular when the number of transcripts is large. We also illustrate the applicability of HESS to diverse real-case data sets, in mouse and human genetic settings, and show that it provides new insights into regulatory hotspots that were not detected by conventional methods. The results suggest that the combination of our modeling strategy and algorithmic implementation provides significant advantages for the identification of functional eQTL hotspots, revealing key regulators underlying pathways.

THE current focus of biological research has turned to high-throughput genomics, which encompasses large-scale data generation and a variety of integrated approaches that combine two or more -omics of data sets. An important example of integrative genomics analysis is the investigation of the genetic regulation of transcription, also called expression quantitative trait locus (eQTL) or "genetical genomics" studies (Cookson *et al.* 2009; Majewski and Pastinen 2011). A typical eQTL analysis follows a natural structure of parallel regressions between the large set of q responses (*i.e.*, expression phenotypes), and that of p explanatory variables (*i.e.*, genetic markers, often single nucleotide polymorphism, SNPs), where p is typically much larger than the number of observations n.

From a statistical point of view, the size and the complex multidimensional structure of eQTL data sets pose a significant challenge. Not only does one wish to detect the set of important genetic control points for each response (expression phenotype), including *cis*- and *trans*-acting control for the same transcript, but, ideally, one would wish to exploit the dependence between multiple expression phenotypes. This will facilitate the discovery of key regulatory markers, so-called hotspots (Breitling *et al.* 2008), *i.e.*, genetic loci or single polymorphisms that influence a large number of transcripts. Identification of hotspots can inform on network and pathways, which are likely to be controlled by major regulators or transcription factors (Yvert *et al.* 2003; Wu *et al.* 2008). Most importantly, there is mounting evidence that common diseases may be caused (or modulated) by changes

Copyright © 2011 by the Genetics Society of America
doi: 10.1534/genetics.111.131425
Manuscript received July 5, 2011; accepted for publication August 23, 2011
Supporting information is available online at http://www.genetics.org/cgi/content/full/genetics.111.131425/DC1.
[1]Corresponding author: Department of Epidemiology and Biostatistics, Imperial College, 1 Norfolk Place, London, W2 1PG, United Kingdom. E-mail: sylvia.richardson@imperial.ac.uk

at a few regulatory control points of the system (*i.e.*, hotspots), which can cause perturbations with large phenotypic effects (Chen *et al.* 2008; Schadt 2009).

In this article, we set out to perform hotspot and eQTL detection in an efficient manner, which exploits fully the multidimensional dependencies within the genome-wide gene expression and genetic data sets. We build upon our previous work (Bottolo and Richardson 2010), where we implemented Bayesian sparse multivariate regression for continuous response to search over the possible subsets of predictors in the large 2^p model space. For each expression phenotype, this corresponds to carrying out multipoint mapping within an inference framework, Bayesian variable selection (BVS), where model uncertainty is fully integrated. Here, we propose a novel structure for linking parallel multivariate regressions that borrows information in a hierarchical manner between the phenotypes to highlight the hotspots. To be precise, we propose a new multiplicative decomposition of the joint matrix of selection probabilities ω_{kj} that link marker j to phenotype k and demonstrate in a simulation study that this hierarchical structure and its Bayesian implementation (hierarchical evolutionary stochastic search or HESS algorithm) possess good characteristics in terms of sensitivity and specificity, outperforming current methods for hotspot and eQTL detection. Finally, we show the applicability of our method in two real-case eQTL studies, including animal models and human data. Our approach is broadly applicable and extendable to other high-dimensional genomic data sets and represents a first step toward a more reliable identification of functional eQTL hotspots and the underlying causal regulators.

Analysis models for eQTL data are linked to two strands of work: (i) methods for multiple mapping of QTL, where a single continuous response, referred to as a "trait," is linked to DNA variation at multiple genetic loci by using a multivariate regression approach, and (ii) models that exploit the pattern of dependence between the sets of responses associated with a predictor (*i.e.*, genetic marker). There is a vast literature on multi-mapping QTL (see the review by Yi and Shriner 2008); some of the models have been extended to the analysis of a small number of traits simultaneously (Banerjee *et al.* 2008; Xu *et al.* 2008). Several styles of approaches have been adopted ranging from adaptive shrinkage (Yi and Xu 2008; Sun *et al.* 2010) to variable selection within a composite model space framework that sets an upper bound on the number of effects (Yi *et al.* 2007). Most of the implemented algorithms sample the regression coefficients via Gibbs sampling. However, these have not been used with a substantial set of markers in the "large p small n" paradigm, but mostly in case of candidate genes or in small experimental cross-animal studies. To face the challenges typical of larger eQTL studies, we have chosen to build our multi-mapping model using a recently developed Bayesian sparse regression approach (Bottolo and Richardson 2010). In this approach, subset selection is implemented in an efficient way for vast (potentially multi-modes) model space by integrating out the regression coefficients and by using a purposely designed MCMC variable selection algorithm that enhances the model search with ideas and moves inspired by evolutionary Monte Carlo algorithms.

The first eQTL modeling approach that explicitly set out to borrow information from all the transcripts was proposed by Kendziorski *et al.* (2006). In the mixture over markers (MOM) method, each response (expression phenotype) $\mathbf{y}_k$, $1 \le k \le q$, is potentially linked to the marker j with probability p_j or not linked to any marker with probability p_0. All responses linked to the marker j are then assumed to follow a common distribution $f_j(\cdot)$, borrowing strength from each other, while those of nonmapping transcripts have distribution f_0. Inspired by models that have been successful for finding differential expression, the marginal distribution of the data for each response $\mathbf{y}_k$ is thus given by a mixture model: $p_0 f_0(\mathbf{y}_k) + \sum_{j=1}^{p} p_j f_j(\mathbf{y}_k)$. A basic assumption of the MOM model is that a response is associated with at most one predictor. For good identifiability of the mixture, MOM requires a sufficient number of transcripts to be associated with the markers. The authors use the EM algorithm to fit the mixture model and estimate the posterior probability of mapping nowhere or to any of the p locations. By combining information across the responses, MOM is more powerful and can achieve a better control of false discovery rates (FDR) by thresholding the posterior probabilities than pure univariate differential expression methods testing each transcript-marker pair. But as originally developed, it is not fully multivariate as it does not account for multiple effects of several markers on each expression trait.

To improve on identification of eQTL effects, Jia and Xu (2007) formulate a unifying $q \times p$ hierarchical model in which each transcript $\mathbf{y}_k$, $1 \le k \le q$, is potentially linked to the complete set of p markers $\mathbf{X}$ through a full linear model with regression coefficients, $\beta_k = (\beta_{k1},..., \beta_{kj},..., \beta_{kp})^{\mathrm{T}}$. Inspired by Bayesian shrinkage approaches already used in conventional QTL mapping, they propose using a mixture prior on each of the β_{kj}, also known as "spike and slab,"

$$\beta_{kj} \sim \left(1 - \gamma_{kj}\right) N(0, \delta) + \gamma_{kj} N\left(0, \sigma_j^2\right), \qquad (1)$$

with a fixed very small δ for the spike and an independent prior for the variance σ_j^2 of the slab in the jth marker. They then link the q responses through a hierarchical model of the Bernoulli indicators γ_{kj}, establishing what we refer to as a hierarchical regression set-up. They assume that $\gamma_{kj} \sim$ Bernoulli(ω_j), $1 \le k, \le q$, and give ω_j a Dirichlet(1, 1) prior. In this model, to improve detection of transcript-marker associations, strength is borrowed across all the transcripts via the common latent probability ω_j. Jia and Xu (2007) implement their hierarchical model in a fully Bayesian framework using an MCMC algorithm called BAYES, based solely on Gibbs sampling.

The high dimensionality of both gene expression and marker space has been alternatively addressed through the use of data reduction methods. In particular, Chun and Keleş

(2009) have proposed implementing sparse partial least-squares regression (M-SPLS eQTL) on preclustered group of transcripts. M-SPLS selects markers associated with each transcript cluster by evaluating the loadings on a set of latent components. As the dimension of each cluster is moderate, SPLS implements a multivariate formulation that takes into account the correlation between the transcripts in the same cluster. Sparsity of the latent direction vectors is achieved by imposing a combination of L_1 and L_2 penalties, similar to the elastic net. The tuning parameters K and η controlling the number of latent components and the convexity of the penalized likelihood are tuned together by cross-validation. The output of this method is the set of the regression coefficients of the markers belonging to the latent vectors that are significantly associated with a subset of transcripts, selected by bootstrap confidence interval.

Jia and Xu's linked regression set-up and fully Bayesian formulation is a natural starting point for eQTL detection, which shares common features with our approach. Here, we present a novel model structure and state of the art implementation based upon evolutionary Monte Carlo. We report the results of a simulation study comparing our HESS method to BAYES (Jia and Xu 2007), as well as to two alternative approaches: MOM (Kendziorski *et al.* 2006) and M-SPLS (Chun and Keleş 2009). Finally, we show the application of our method to two diverse genomic experiments in mouse and human genetic contexts.

Theory and Methods

Hierarchical related sparse regression

Let $\mathbf{Y} = \{\mathbf{y}_k^{\mathrm{T}}, 1 \le k \le q\}$ the $n \times q$ matrix of responses, with $\mathbf{y}_k = (y_{k1}, \ldots, y_{kn})^{\mathrm{T}}$ the sequence of n observations of the kth response, and let $\mathbf{X}$ be the $n \times p$ design matrix with ith row $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^{\mathrm{T}}$. We assume throughout that $\mathbf{x}_i$ is quantitative. It encompasses the case of continuous biomarkers, inbred genotypes {0, 1} for recombinant inbred (RI) strains and {0, 1, 2} genotype coding for F_2 animal crosses or human data. A linear model for the kth response can be described by the equation

$$\mathbf{y}_k = \alpha_k \mathbf{1}_n + \mathbf{X}\beta_k + \varepsilon_k,$$

where α is an unknown constant, $\mathbf{1}_n$ is a column vector of ones, $\boldsymbol{\beta}_k = (\beta_{k1}, \ldots, \beta_{kp})^{\mathrm{T}}$ is the $p \times 1$ vector of regression coefficients, and ε_k is the error term with $\varepsilon_k \sim N(\mathbf{0}, \sigma_k^2 \mathbf{I}_n)$, where $\mathbf{I}_n$ is the diagonal matrix of dimension n. BVS is performed by placing a constant prior density on α_k and a prior on $\boldsymbol{\beta}_k$, which depends on a latent binary vector $\gamma_k = (\gamma_{k1}, \ldots, \gamma_{kj}, \ldots, \gamma_{kp})^{\mathrm{T}}$, where $\gamma_{kj} = 1$ if $\beta_{kj} \neq 0$ and $\gamma_{kj} = 0$ if $\beta_{kj} = 0$, $j = 1, \ldots, p$. Conditionally on the latent binary vector, the linear model becomes

$$\mathbf{y}_k = \alpha_k \mathbf{1}_n + \mathbf{X}_{\gamma_k} \beta_{\gamma_k} + \varepsilon_k,$$

where $\boldsymbol{\beta}_{\gamma k}$ is the nonzero vector of coefficients extracted from $\boldsymbol{\beta}_k$, $\mathbf{X}_{\gamma k}$ is the design matrix of dimension $n \times p_{\gamma k}$, with columns corresponding to $\gamma_{kj} = 1$, and $p_{\gamma_k} \equiv \gamma_k^{\mathrm{T}} \mathbf{1}_p$ the number of selected covariates for the k response. For every regression k, we assume that, apart from the intercept α_k, $\mathbf{X}$ contains no variables that would be included in every possible model and that the columns of the design matrix have all been centered in 0.

Assuming independence of the q regression equations conditionally on the selected predictors modeled in the $q \times p$ latent binary matrix $\Gamma = \{\gamma_k^{\mathrm{T}}, 1 \le k \le q\}$, the likelihood becomes

$$\prod_{k=1}^{q} \phi_n\left(\mathbf{y}_k; \alpha_k \mathbf{1}_n + \mathbf{X}_{\gamma_k} \beta_{\gamma_k}, \sigma_k^2 \mathbf{I}_n\right), \tag{2}$$

where $\phi_n(\cdot)$ is the n -variate normal density function.

The description of the joint likelihood as the product of q regression equations is similar to the one proposed by Jia and Xu (2007). However, one important difference is the assignment in (2) of a regression specific error variance σ_k^2, allowing for transcript-related residual heterogeneity and making our formulation more flexible. A more general model, seemingly unrelated regressions (SUR) introduces additional dependence between the responses $\mathbf{Y}$ through the noise ε_k, modeling the correlation between the residuals of different responses (Banerjee *et al.* 2008). However, it becomes computational unfeasible when the size of q is large, which is typical in eQTL experiments.

Prior set-up

For a given k, we follow the same prior set-up for the regression coefficients and error variance as described in Bottolo and Richardson (2010). First, we treat the intercept α_k separately, assigning it a constant prior, $p(\alpha_k) \propto 1$. Second, conditionally on γ_k, we assign a g-prior structure on the regression coefficients and an inverse-gamma (Inv Ga) density to the residual variance

$$p\left(\beta_k | \gamma_k, \tau, \sigma_k^2\right) = N\left(\mathbf{0}, \sigma_k^2 \tau \left(\mathbf{X}_{\gamma_k}^{\mathrm{T}} \mathbf{X}_{\gamma_k}\right)^{-1}\right) \tag{3}$$

$$p\left(\sigma_k^2\right) = \text{Inv Ga}(a_\sigma, b_\sigma), \tag{4}$$

with a_σ, $b_\sigma > 0$, and $E(\sigma_k^2) = b_\sigma/(a_\sigma - 1)$. This conjugate prior set-up has many advantages. The most important is that, for a given k, the marginal likelihood $p(\mathbf{y}_k|\mathbf{X}, \gamma_k, \tau)$ can be written in a closed form that is particularly simple to compute once (3) and (4) are integrated out. Furthermore, it allows for more efficient MCMC exploration with correlated predictors than the nonconjugate case (*i.e.*, when the variance component σ_k^2 in (3) is different from the error variance) and it provides more accurate identification of the high-probability models among those visited during the MCMC (George and McCulloch 1997). Finally it leads to a simple and interpretable expression, $E(\beta_{kj}|\mathbf{Y}, \tau) = \tau/(1+\tau)\beta_{kj}^{\mathrm{OLS}}$ with $\beta_{kj}^{\mathrm{OLS}}$ the ordinary least-squares solution, of the level of shrinkage.

The hierarchical structure on the regression coefficients is completed by specifying a hyper-prior on the scaling coefficient τ, $p(\tau)$. We adopt the Zellner–Siow priors structure for the regression coefficients that can be thought as a scale mixture of g-priors and an inverse-gamma prior on τ, $p(\tau) = \text{Inv Ga}(1/2, n/2)$ with heavier tails than the normal distribution, $p(\beta_k|\gamma_k, \sigma_k^2) = \text{Cauchy}(\mathbf{0}, n\sigma_k^2(\mathbf{X}_{\gamma_k}^{\mathrm{T}}\mathbf{X}_{\gamma_k})^{-1})$. In general it has been observed (Bottolo and Richardson 2010) that data adaptivity of the degree of shrinkage conforms better to different variable selection scenarios than assuming standard fixed values (which can be easily implemented by using a point mass prior for τ). Since the level of shrinkage can influence the results of the variable selection procedure, in our model we force all the q regression equations to share the same common τ, linking the regression equations hierarchically through the variance of their nonzero coefficients.

The prior specification is concluded by assigning a Bernoulli prior on the latent binary value γ_{kj}, $p(\gamma_{kj}|\omega_{kj}) = \text{Bernoulli}(\omega_{kj})$. The prior chosen for ω_{kj} is of paramount importance in BVS since it controls the level of sparsity, *i.e.*, the association with a parsimonious set of important predictors. For a given response this task can be accomplished by specifying a common small-selection probability for all p predictors, $\omega_{kj} = \omega_k$ and giving $p(\omega_k) = \text{Beta}(a_k, b_k)$ (Bottolo and Richardson 2010). Inducing sparsity when all the responses are jointly considered is harder because further constraints are desirable. eQTL surveys (Cookson *et al.* 2009) suggest that only a fraction of expression traits are under genetic regulation and the number of their control points is usually small. This can be modeled by assigning a different probability for each marker $\omega_{kj} = \omega_j$ with an hyper-prior on ω_j. This solution, first proposed by Jia and Xu (2007) with the conjugate prior $p(\omega_j) = \text{Dirichlet}(d_{1j}, d_{2j})$, assumes that this selection probability is the same for all the responses. However, whatever the sensible choice of the hyperparameters d_{1j} and d_{2j}, $d_{1j}, d_{2j} = 1$ or $d_{1j}, d_{2j} = 0.5$, the posterior density greatly depends on the ratio between the number of transcripts associated to the marker j, q_j, and the total number the transcripts in the eQTL experiment, q, since $E(\omega_j|\boldsymbol{Y}) = (q_j + d_{1j})/(q + d_{1j} + d_{2j})$, where $q_j = \#\{j: \gamma_{kj} = 1\}$. In such formulation, the results are thus clearly influenced by the number of responses analyzed and sparsity of each kth regression cannot be controlled in the prior specification adopted for ω_{kj} of γ_{kj}.

In this article we propose a novel way of specifying the selection probability ω_{kj} to synthetize the benefits of both approaches, Bottolo and Richardson (2010) and Jia and Xu (2007). We propose decomposing this probability into its marginal effects

$$\omega_{jk} = \omega_k \times \rho_j \tag{5}$$

with ω_k and ρ_j the "row" and "column" effect, respectively, and $0 \le \omega_k \le 1$ and $\rho_j \ge 0$, but constrained so that $0 \le \omega_{jk} \le 1$. The idea behind this decomposition is to control the level of sparsity for each row k through a suitable choice of the hyperparameters a_k, b_k of $p(\omega_k) = \text{Beta}(a_k, b_k)$, while the parameter ρ_j captures the "relative propensity" of predictor j to influence several responses at the same time. Large values of ρ_j indicate that predictor j has a marked influence on ω_{jk} and thus likely to be a hotspot. The adopted multiplicative formulation has some similarity to the disease mapping paradigm where the relative risk level acts in a multiplicative fashion on an expected number of cases in a binomial or Poisson disease risk model. A gamma density on the jth latent hotspot effect, $p(\rho_j) = \text{Ga}(c, d)$, with $E(\rho_j) = c/d$, complete the hierarchical structure for the decomposition (5).

We conclude this section by describing the choice of the hyperparameters for ω_k and ρ_j. Since by construction $\omega_k \perp \rho_j$, $E(\omega_{jk}) = E(\omega_k)E(\rho_j)$. If we assume $c = d$, the hotspot propensity does not change the *a priori* row marginal expectation, $E(\omega_{jk}) = E(\omega_k)$. However, it inflates the *a priori* row marginal variance $\text{Var}(\omega_{jk}) > \text{Var}(\omega_k)$, with $\text{Var}(\omega_{jk}) = \text{Var}(\omega_k)(1 + d^{-1}) + d^{-1}E^2(\omega_k)$. For the specification of the hyperparameters a_k and b_k, we use the Beta-binomial approach illustrated in Kohn *et al.* (2001), after marginalizing over the column effect in (5). The two hyperparameters can be worked out once $E(p_{\gamma_k})$ and $\text{Var}(p_{\gamma_k})$, the expected number and the variance of the number of genetic control points for each response, are specified.

Posterior inference

After integrating out the intercepts, the regression coefficients and the error variances, the joint density can be factorized as

$$p(\mathbf{Y}, \mathbf{X}, \Gamma, \Omega, \tau) = p(\mathbf{Y}|\mathbf{X}, \Gamma, \tau)p(\Gamma|\Omega)p(\Omega)p(\tau), \tag{6}$$

where $p(\mathbf{Y}|\mathbf{X}, \Gamma, \tau) = \prod_{k=1}^{q} p(\mathbf{y}_k|\mathbf{X}, \gamma_k, \tau)$, $p(\Gamma|\Omega) = \prod_{k=1}^{q}\prod_{j=1}^{p} p(\gamma_{kj}|\omega_{kj})$, and $p(\Omega) = \prod_{k=1}^{q} p(\omega_k)\prod_{j=1}^{p} p(\rho_j)$. Posterior inference is carried out on the $q \times p$ latent binary matrix $\boldsymbol{\Gamma} = \{\gamma_{kj}, 1 \le k \le q, 1 \le j \le p\}$, on the $q \times p$ selection probability matrix $\boldsymbol{\Omega} = \{\omega_{kj}, 1 \le k \le q, 1 \le j \le p\}$, and on the scaling coefficient τ, if not fixed.

Sampling $\boldsymbol{\Gamma}$ is extremely challenging since complex dependence structures in the $\mathbf{X}$ create well-known problems of multimodality of the model space even for a single regression equation. Here the computational challenge is higher since we are aiming to explore a huge model space of dimension $(2^p)^q$. For this reason vanilla MCMC (MC^3, Gibbs sampler, simple dimension changing moves) cannot guarantee a reliable exploration of the model space in a limited number of iterations. In this article we use a sampling scheme introduced by Bottolo and Richardson (2010), evolutionary stochastic search (ESS) as a building block for our new algorithm HESS. For each response, HESS relies on running multiple chains with different "temperature" in parallel, which exchange information about the set of covariates that are selected in each chain. Since chains with higher temperatures flatten the posterior density, global moves (between chains) allow the algorithm to jump from one local

mode to another. Local moves (within chains) permit the fine exploration of alternative models, resulting in a combined algorithm that ensures that the chains mix efficiently and do not become trapped in local modes. Specific modifications of ESS were introduced to comply with the structure of ω_{kj}, which are sampled with the decomposition (5) rather than integrating them out as in ESS. This requires some modifications in the local moves (details in Supporting Information, File S1, Section S.2.2.)

For sampling the selection probability matrix $\mathbf{\Omega}$, we implemented a Metropolis-within-Gibbs algorithm for each element of the row effect $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_q)^{\mathrm{T}}$ and column effects $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_q)^{\mathrm{T}}$, rejecting proposed values outside the range [0, 1]. However, since the dimension of $\boldsymbol{\omega}$ and $\boldsymbol{\rho}$ is very large, tuning the proposal for each element of the two vectors is prohibitive. To make HESS fully automatic, we use the adaptive MCMC scheme proposed by Roberts and Rosenthal (2009), where the variance of the proposal density is tuned during the MCMC to reach a specified acceptance rate. To satisfy the asymptotic convergence of the adaptive MCMC scheme, mild conditions are imposed (details in File S1, Section S.2.3).

If not fixed, the scaling coefficient τ, which is common for all the q regression equations and all the L chains, is sampled using a Metropolis-within-Gibbs algorithm with random walk proposal and fixed proposal variance (details in File S1, Section S.2.4).

Finally, we describe a complete sweep of our algorithm. We assume that the design matrix is fully known. If missing values are present, these can be imputed in a preprocessing imputation step (for instance using the fill.geno function from the *qtl* R package for genetic crosses (Broman and Sen 2009) or *FastPhase* (Scheet and Stephens 2006) for human data). Without loss of generality, we assume that the responses and the design matrix have both been centered. The same notation is used when τ is fixed or given a prior distribution. For simplicity of notation we do not index variables by the chain index, but we emphasize that the description below applies to each chain:

- Given $\mathbf{\Omega}$ and τ we update γ_k, according to the ESS procedure, using global and local moves. During the burn-in, we sample the latent binary vector γ_k for each k to tune the regression specific temperature ladder (details in File S1, Section S.2.5). After the burn-in, at each sweep, we select at random without replacement a fraction ϕ of the regressions where to update γ_k.
- Given $\mathbf{\Gamma}$ and τ, we sample $\boldsymbol{\omega}$ and $\boldsymbol{\rho}$ with a random walk Metropolis with adaptive proposals.
- Given $\mathbf{\Gamma}$ and $\mathbf{\Omega}$, we sample τ with a random walk Metropolis with a fixed proposal. To balance the number of updates of the latent binary values γ_{kj} with those of the scaling coefficient, at each sweep, the number of times we sample τ is proportional to $q \times p \times L$.

The Matlab implementation of the HESS algorithm is available upon request from the authors.

Postprocessing analysis

In this section we present some of the postprocessing operations required to extract useful information from the rich output of our model. We stress that, while here for simplicity we are not using the output of the heated chains, following Gramacy *et al.* (2010), posterior inference could also be carried out using the information contained in all the chains.

The primary quantity of interest is the posterior propensity of each predictor to be a hotspot. In the spirit of cluster detection rules in disease mapping (Richardson *et al.* 2004), we use tail posterior probabilities of the propensities ρ_j, *i.e.*, declare the jth predictor to be a hotspot if

$$\Pr\left(\rho_j > 1 \middle| \boldsymbol{Y}\right) \geq t, \tag{7}$$

where t is a chosen threshold. We have found by empirical exploration and simulations that choosing a posterior threshold of $t = 0.8$ gives good performance across different scenarios with varying dimensions (data not shown).

The next quantity of interest is the posterior probability of inclusion for the pair (k, j). Following Petretto *et al.* (2010), the *marginal probability of inclusion* is

$$p\left(\gamma_{kj} = 1 \middle| \mathbf{y}_k\right) = C_k^{-1} \sum_{s=1}^{S} 1_{(\gamma_{kj}^{(s)}=1)}(\gamma_k) p(\gamma_k^{(s)} | \mathbf{y}_k), \tag{8}$$

where $\gamma_k^{(s)} = (\gamma_{k1}^{(s)}, \ldots, \gamma_{kj}^{(s)}, \ldots, \gamma_{kq}^{(s)})$ is the latent binary vector sampled at iteration s, $p(\gamma_k^{(s)}|\mathbf{y}_k)$is the model posterior probability obtained through inexpensive numerical integration in the full output (see File S1, Section S.3) and $C_k = \sum_{s=1}^{S} p(\gamma_k^{(s)}|\mathbf{y}_k)$ is the constant of normalization. The Bayes factor (BF) for the pair (k, j) is derived from (8) as the ratio between posterior odds and prior odds

$$\mathrm{BF}_{kj} = \frac{p(\gamma_{kj} = 1|\mathbf{y}_k)}{1 - p(\gamma_{kj} = 1|\mathbf{y}_k))} \Big/ \frac{E\left(p_{\gamma_k}\right)/p}{1 - \left(E\left(p_{\gamma_k}\right)/p\right)}, \tag{9}$$

where $E(p_{\gamma k})$ is the *a priori* expected number of genetic control points for the kth transcript.

Similarly to (8), if of interest, we can further evaluate the joint posterior probability of the set of predictors declared as hotspots as

$$p\left(\bigcap_{j=1}^{\mathcal{H}} (\gamma_{kj} = 1) \middle| \mathbf{y}_k\right) = C_k^{-1} \sum_{s=1}^{S} 1_{\left(\cap_{j=1}^{\mathcal{H}} (\gamma_{kj}=1)\right)}(\gamma_k) p(\gamma_k^{(s)} | \mathbf{y}_k) \tag{10}$$

with C^k as before and $\mathcal{H}$ the set of markers identified as hotspots.

Finally, the *best model visited* is defined as

$$\gamma_k^{\mathrm{B}} = \left\{ \gamma_k^{(s)} : \max_s p(\gamma_k^{(s)} | \mathbf{y}_k) \right\}. \tag{11}$$

Note that the configuration posterior probability $p(\boldsymbol{\Gamma}|\mathbf{Y})$ (see File S1, section S.3) can be used as an alternative weight in (8) and (10) or to derive the max *a posteriori (MAP) configuration visited*

$$\boldsymbol{\Gamma}^{\mathrm{B}} = \left\{\boldsymbol{\Gamma}^{(s)} : \max_{s} p(\boldsymbol{\Gamma}^{(s)}|\mathbf{Y})\right\}.$$

Results

Simulation studies

We carried out a simulation study to compare our algorithm with recently proposed multiple response models: MOM (Kendziorski *et al.* 2006), BAYES (Jia and Xu 2007), and M-SPLS (Chun and Keleş 2009).

To create more realistic examples, we decided not to simulate the **X** matrix, but to use real human-phased genotype data spanning 500 kb, region ENm014 (chromosome 7: 126,368,183–126,865,324 bp), from the Yoruba population used in the HapMap project (Altshuler *et al.* 2005) as the design matrix. After removing redundant variables, the set of SNPs is reduced to $P = 498$, with $n = 120$, giving a 120×498 **X** matrix. As noted by Chun and Keleş (2009), high correlations between markers might affect the performance of variable selection procedures that do not explicitly consider such a grouping structure. The benefit of using real human data are to test competing algorithms when the pattern of correlation, *i.e.*, linkage disequilibrium (LD), is complex and blocks of LD are not artificial, but they derive naturally from genetic forces, with a slow decay of the level of correlation between SNPs (see Figure S1).

In the simulated examples, we carefully select the SNPs that represent the hotspots (Figure S1): (i) all hotspots are inside blocks of correlated variables; (ii) the first four SNPs are weakly dependent ($r^2 < 0.1$); and (iii) the remaining two SNPs are correlated with each other ($r^2 = 0.46$) and also linked to one of the previous simulated hotspots ($r^2 = 0.52$ and $r^2 = 0.44$, respectively), creating a potential masking effect difficult to detect. Apart from the hotspots, no other SNPs are used to simulate transcript–SNP associations. We simulated four cases:

SIM1: In this example we simulated $q = 100$ transcripts from the selected six hotspots, with some transcripts predicted by multiple correlated markers (polygenic control): for instance transcripts 17–20 are regulated by three SNPs at the same time (see Figure S2). Altogether we simulated 94 transcript–SNP associations in 50 distinct transcripts. The effects were simulated from a normal density with smaller variance than in Jia and Xu (2007), $\beta_{kj} \sim N(0,0.3^2)$with $\boldsymbol{\varepsilon}_{\mathrm{k}} \sim N(\mathbf{0}, 0.1^2\mathbf{I}_n)$ to mimic the smaller signal-to-noise ratio expected in genetically heterogeneous human data.

SIM2: As in the previous example, we simulated 100 responses, but there are only three hotspots with the same simulated association as before, leading to 64 transcript–SNP associations in 30 distinct transcripts. Moreover we created potential false-positive associations by simulating transcripts 81–90 and 91–100 using a linear transformation of transcript 20 with a mild negative correlation (in the interval $[-0.5, -0.4]$) and of transcript 80 with a strong positive correlation (in the interval [0.8, 0.9]), respectively. Since we create false-positive associations, the scenario will inform on how different algorithms behave when correlations among some transcripts are not due to SNPs.

SIM3: This simulation set-up is identical to the first scenario for the first 100 responses, but we increase the number of simulated responses to $q = 1{,}000$, simulating the further 900 transcripts from the noise.

SIM4: This is the same as the second simulated data set for the first 100 responses, with additional 900 responses simulated from the noise, giving altogether $q = 1000$ responses.

We discuss here the hyperparameters set-up. Since *a priori*, in addition to a large effect of a SNP that is located close to the transcript (*cis*-eQTL), we expect only a few additional control points associated with the variation of gene expression (typically *trans*-eQTL); in HESS we set $E(p_{\gamma k}) = 2$ and $\mathrm{Var}(p_{\gamma \mathrm{k}}) = 2$, meaning the prior model size for each transcript response is likely to range from 0 to 6 (Petretto *et al.* 2010). Following Kohn *et al.* (2001), we fixed $a_\sigma = 10^{-10}$ and $b_\sigma = 10^{-3}$, giving rise to a noninformative prior on the error variance. We run the HESS algorithm for 6000 sweeps with 1000 as burn-in with three chains and $\varphi = 1/4$. Computational time is similar for the first two simulated examples, 6 hr, and 10 times greater for the last two simulated scenarios on a Intel Xeon CPU at 3.33 GHz with 24 Gb RAM.

We run BAYES for 15,000 sweeps with 5000 as burn-in, recording sampled values every 5 sweeps. The variance δ of the spike component 1 is set 10^{-4}, which is 100 times lower than the noise variance. Since the code available from the authors was written in SAS/IML, we recoded their Gibbs sampler in Matlab. We used the default parameters for MOM, while in M-SPLS the two tuning parameters are obtained through cross-validation selected in the interval $K = 1,\ldots, 10$ and $\eta = 0.01, \ldots, 0.99$. Each simulated example was replicated 25 times and we run the four algorithms on each replicate.

Power to detect hotspots: The identification of the hotspots is of primary interest for all the algorithms we are comparing. In HESS using the tail posterior probability $\Pr(\rho_j > 1|\mathbf{Y})$, we can rank the predictors according to their propensity to be a hotspot, while in BAYES the posterior mean of the common latent probability ω_j, $E(\omega_j|\mathbf{Y})$ is utilized to prioritize important markers. In MOM the strength for a predictor to be a hotspot is not directly available but, as suggested by the authors, given a marker, it can be obtained by taking a suitable quantile of the transcript–marker associations distribution across responses. We use their R function get.hotspots recording the average of the distribution for each predictor. M-SPLS, after cross-validation, provides a list of latent components that

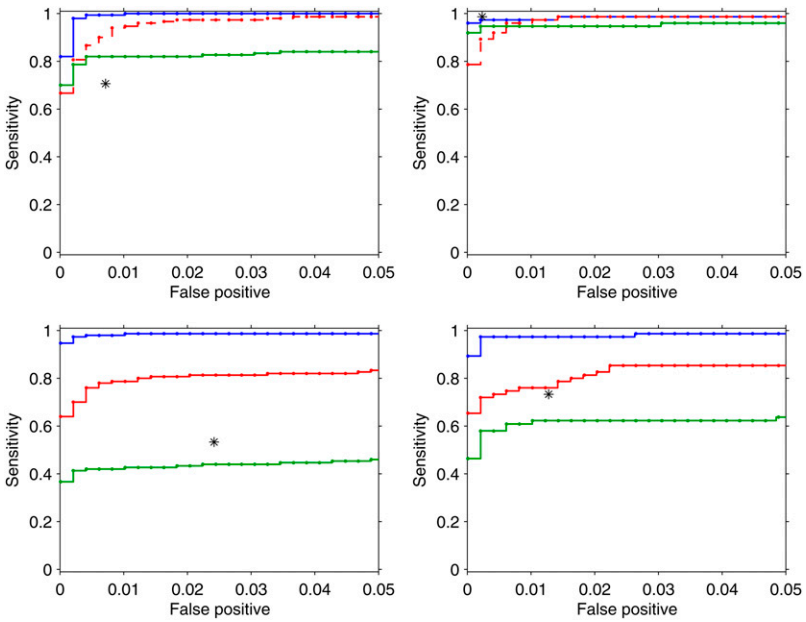

Figure 1 ROC curves for hotspots detection using HESS (blue line), MOM (red line), BAYES (green line), and M-SPLS (black star) in the four simulated scenarios (Figure S2). From top to bottom, left to right: SIM1, $q = 100$ and six hotspots; SIM2, $q = 100$ and three hotspots; SIM3, $q = 1000$ and six hotspots; SIM4, $q = 1000$ and three hotspots. For M-SPLS, type I error and power were calculated conditionally on the list of latent vector components. (Top) MOM is indicated by a red dashed line to highlight that it is not designed in the cases when the number of markers is larger than the number of traits.

predicts most of the variability of **Y**. While this group cannot be interpreted directly as the list of hotspots, we use it to test the existing overlap between the simulated hotspots and the latent components. Finally, differently from the analyses presented in Jia and Xu (2007) and Chun and Keleş (2009), in the HESS power calculation, we simply rank the evidence for being a hotspot provided by each algorithm across the 25 replicates. Therefore we are not using any method-specific procedure to call a hotspot, based for instance on FDR considerations, that could influence the comparison results.

Figure 1 shows the ROC curves for the four algorithms considered. HESS (blue lines) outperforms all the other methods with sizeable power on the simulated examples. It is not significantly affected by the dimension of the eQTL experiment (top, $q = 100$; bottom, $q = 1000$). This is somehow expected since the hotspot propensity does not depend directly on the number of transcripts analyzed (see File S1, section S.2.3). Spurious correlations among transcripts not due to SNPs (right) have a negligible effect on the HESS power, showing robust properties of our algorithm in detecting hotspots under different scenarios.

The other methods show good properties when $q = 100$, but their power degrades sensibly when $q = 1000$. This is expected for BAYES (green lines) since $E(\omega_j|\mathbf{Y})$ is affected by q, while the performance of MOM (red lines) is more stable. MOM (red dashed lines) shows good power in the simulated examples even when the number of markers is larger than the number of traits, a situation that MOM is not designed for (top). Finally M-SPLS has greater power than BAYES, but it is outperformed by both HESS and MOM in the more sparse scenarios (bottom). Looking more closely at the list of latent vectors identified by M-SPLS (data not shown), we noted that the simulated hotspots at SNP 362 and 466 that are linked to SNP 239 were rarely selected (false negative) in both SIM1 and SIM3. On the contrary, SNP 75 is very often included (false positive) in the list of latent vectors in all the scenarios. This might reflect the high correlation between SNP 362 and 466 with SNP 239, as well as the strong dependency between SNP 75 and 30 where we simulated a hotspot. This suggests that M-SPLS has limited efficiency in the presence of complex correlation patterns in the design matrix. In Figure S3, Figure S4, Figure S5, and Figure S6, interested readers can find the visual representation of the signals detected by each algorithm and averaged across the 25 replicates.

Power to detect transcript-marker associations: Figure 2 shows the ROC curves of the transcript–marker marginal associations detected by each method. Also in this case, to perform the power calculation, we are not using any method-specific way to declare a significant association, since we simply record the output from each algorithm and rank it across the 25 replicates. In particular we use the marginal probability of inclusion $p(\gamma_{kj} = 1|\mathbf{y}_k)$ (8) for HESS; the posterior frequency $\bar{\gamma}_{kj} = \Sigma_{s=1}^{S} \gamma_{kj}^{(s)}/S$ for BAYES, where $\gamma_{kj}^{(s)}$ is the value recorded at iteration s; the transcript–marker association provided the MOM object momObj; and finally the associations selected by bootstrap confidence interval at different type I error levels ($\alpha = 10^{-4}, 10^{-3}, 10^{-2}, 0.05$) for M-SPLS.

For transcript–marker association detection, we find that HESS has higher power than that of the other methods in all the simulated scenarios. As expected when more responses are included, the power decreases slightly (bottom), while spurious associations due to the correlation between transcripts do not seem to affect the ability of HESS to distinguish between true and false signals (right). MOM is quite stable across scenarios, but it reaches only half of the power of HESS. BAYES and M-SPLS have similar behavior and their performance degrades when $q = 1000$. BAYES, in particular, has very low power since the shrinkage to the null effect,

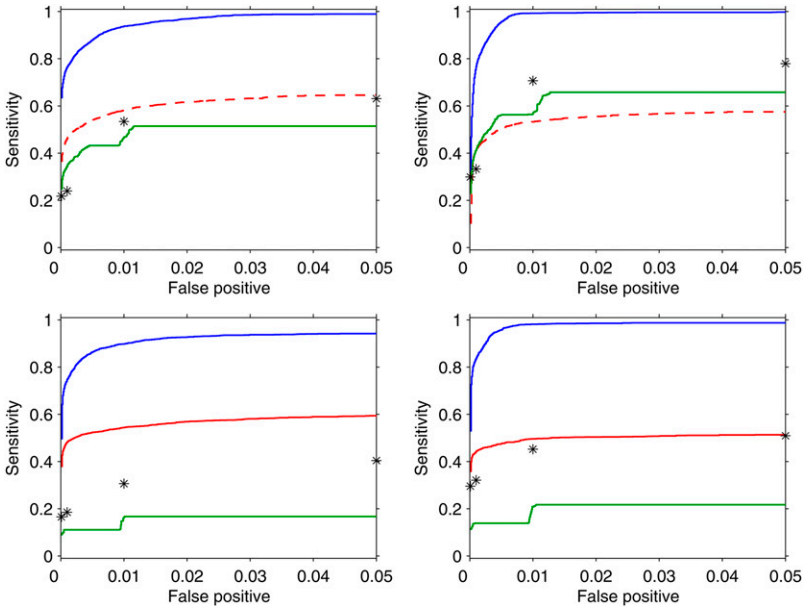

Figure 2 ROC curves for transcript–marker associations using HESS (blue line), MOM (red line), BAYES (green line), and M-SPLS (black star) in the four simulated scenarios (Figure S2). From top to bottom, left to right: SIM1, $q = 100$ and six hotspots; SIM2, $q = 100$ and three hotspots; SIM3, $q = 1000$ and six hotspots; SIM4, $q = 1000$ and three hotspots. For M-SPLS, power is calculated conditionally on the list of transcript–marker associations selected by bootstrap confidence interval at a fixed type I error ($\alpha = 10^{-4}, 10^{-3}, 10^{-2}, 0.05$). In the top, MOM is indicated by a red dashed line to highlight that it is not designed in the cases when the number of markers is larger than the number of traits.

caused by common latent probabilityω_j, is particularly strong in SIM3 and SIM4.

The different power of the methods considered can be better understood by looking at Figure S3, Figure S4, Figure S5, and Figure S6, where, for each simulated example, we averaged the evidence of transcript–marker association across replicates. HESS is able to identify the correct simulated pattern, with very few false positives. When false-positive associations are simulated, for instance, in SIM3 and SIM4, HESS assigns on average lower posterior probability of inclusion than for the true positive ones (Figure S4 and Figure S6, top left). While MOM is able to identify the simulated hotspots, it finds it difficult to separate the true transcript–marker associations from the spurious ones (Figure S3, Figure S4, Figure S5, and Figure S6, bottom left). The main limitation of M-SPLS is the correct identification of the latent vectors when highly correlated predictors are considered (Figure S3, Figure S4, Figure S5, and Figure S6, top right). Finally BAYES is able to identify the simulated pattern when $q = 100$ (Figure S3 and Figure S4, bottom right), but it seems to be too conservative when the number of responses is large, $q = 1000$ (Figure S5 and Figure S6, bottom right). The higher false-negative rate in BAYES may depend on the poor efficiency of the MCMC sampler (which is based exclusively on the Gibbs sampling that is not able to jump between distant competing models) and on the spike and slab prior that is not integrated out. The latter influences the sampling of γ_{kj} since the latent binary vector depends on the regression coefficients (see Figure S7 for an illustration).

Real case studies

Here we present two applications of HESS to: (i) mouse gene expression data published in Lan *et al.* (2006) that is commonly used as a benchmark data set for detection of eQTL (Chun and Keleş 2009) and eQTL hotspots (Kendziorski *et al.* 2006; Jia and Xu 2007) and (ii) human monocytes expression data set recently analyzed for disease susceptibility by Zeller *et al.* (2010).

Mouse data set: The mouse data set has been previously described in detail (Lan *et al.* 2006), and it consists of 45,265 probe sets the expression of which has been measured in the liver of 60 mice. Mice were collected from the F_2-*ob/ob* cross (B6 × BTBR) and genotype data were available for 145 microsatellite markers from 19 autosomal chromosomes. To make our analysis comparable with previously reported studies (Jia and Xu 2007; Chun and Keleş 2009), we focused on 1573 probe sets showing sizeable variation in gene expression in the mouse population (sample variance >0.12). Running HESS for 12,000 sweeps with 2000 as burn-in and the same choice of the hyperparameters described in the simulation studies, among the 145 markers 16 were identified with posterior tail probability >0.8, regulating a significant number of probe sets (Table S1). We report the genome location of the identified hotspots in Figure 3 and show transcript–marker associations in Figure 4. Since large hotspot propensity reveals that multiple traits are controlled by the same marker, we focused on biologically meaningful transcript–marker associations by using marginal probability of association >0.95 (corresponding to local FDR 5%, Ghosh *et al.* 2006). Six markers were found to control more than 5% of all analyzed probe sets as shown in Figure 3. While marker D15Mit63 was previously detected by BAYES and M-SPLS, three other major regulatory points were identified solely by our method: D13Mit91, D18Mit9, and D18Mit202, controlling 14.1, 10.6, and 9.7% of all analyzed probe sets, respectively (Table S2).

The regulatory hotspot at marker D13Mit91 is located within the *Kif13a* (kinesin family member 13A) gene, which is involved in intracellular protein transport and microtubule motor activity via direct interaction with the AP-1 adaptor complex (Nakagawa *et al.* 2000). This hotspot is

associated with 222 probe sets, representing 190 distinct well-annotated genes, that are enriched for specific gene ontology (GO) terms, including "protein localization" ($P = 4.2 \times 10^{-6}$), "protein transport" ($P = 5.7 \times 10^{-6}$), and "establishment of protein localization" ($P = 6.4 \times 10^{-6}$). Hence, given its molecular function *Kif13a* is likely to be a candidate master regulator of the genes implicated with protein transport, and whose expression is associated with marker D13Mit91.

The other two newly identified markers, D18Mit9 and D18Mit202, are located on mouse chromosome 18. D18Mit9 resides within a known QTL (*Hdlq30*) involved in HDL cholesterol levels (Korstanje *et al.* 2004) whereas D18Mit202 resides within a known diabetes susceptibility/resistance locus (Idd21, insulin-dependent diabetes susceptibility 21) (Hall *et al.* 2003).

Human data set: The human data set included 648 probe sets, representing 516 unique and well-annotated genes (Ensemble GRCh37), that were found to be coexpressed in monocytes, delineating a network driven by the *IRF7* transcription factor in 1490 individuals from the Gutenberg Heart Study (GHS) (for details on the network analysis, see Heinig *et al.* (2010)). This *IRF7*-driven inflammatory network (IDIN) was also reconstructed in a distinct population cohort: 758 subjects from the Cardiogenics Study showing significant overlap with the network in GHS. The "core" of the network consisted of a small gene set ($q = 17$), including *IRF7* and coregulated target genes, the expression of which was found to be *trans*-regulated by a locus on human chromosome 13q32 using MANOVA in Cardiogenics (Heinig *et al.* 2010). However, this *trans*-regulation was not found in the GHS study, using similar MANOVA analysis.

Here we take a new look and use HESS to analyze the larger IDIN with 648 probe sets in the GHS population ($n = 1490$ individuals) and the SNP set ($P = 209$) spanning 1 Mb on chromosome 13q32 (data available upon request from Stefan Blankenberg under the framework of a formalized collaboration via a Memorandum Transfer Agreement). While MOM and BAYES fail to detect any signal at this locus, using HESS we found two SNPs, rs9557207 and rs11616269, which are detected as hotspots for the IDIN expression with tail posterior probability 0.83 and 0.91, respectively (Figure 5). These SNPs are located 45.3 kb (rs9557207) and 25.1 kb (rs11616269) from SNP rs9585056, which previously showed significant *trans*-effect on the core gene set of the network in Cardiogenics ($P = 5.0 \times 10^{-3}$). This region was also associated with *EBI2* expression ($P = 2.2 \times 10^{-8}$), the candidate gene at this locus, and with type I diabetes (T1D) ($P = 7.0 \times 10^{-10}$) (Heinig *et al.* 2010). For the two identified hotspots, we looked in detail at each transcript–marker association and compute their BF as given in (9). We observe that 26 and 13 transcripts show clear evidence of associations (BF $>$ 10, Kass and Raftery 2007) in the two hotspots identified (Table S3) delineating the extent of regulatory effects. To further

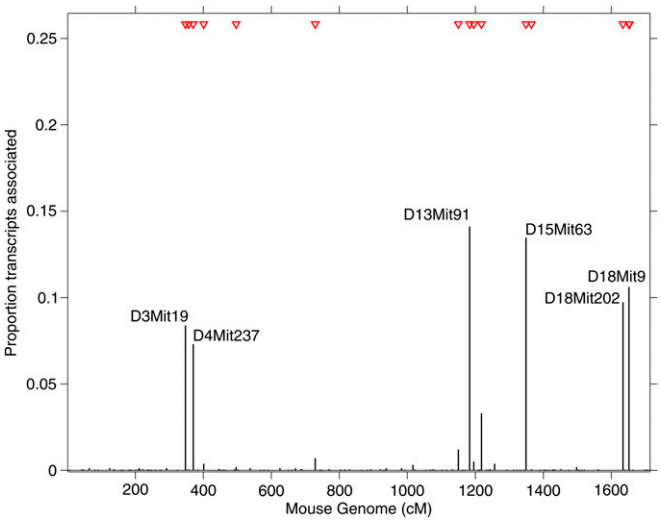

Figure 3 Proportion of transcripts associated with each marker in the mouse data example ($n = 60$, $P = 145$, and $q = 1573$). Transcript–marker association was declared at 5% local FDR with marginal probability of inclusion >0.95. The 16 red triangles indicate markers (two of them are overlapping and hence are not distinguishable) that have been identified as hotspots with tail posterior probability >0.8.

calibrate this evidence, we investigated BF for marker–transcript associations in a comparable simulated set-up, that of SIM3. Using the threshold BF $>$ 10 would lead to declaring <5% false positive marker–transcript associations in the identified hotspots (data not shown). Note that most of these transcripts (80%) are found only in the network inferred in GHS and not with the Cardiogenics network, suggesting a complex pattern of regulatory effects around locus rs9585056 which is highlighted in a specific manner in each population. These population-specific regulatory effects could reflect differences in monocytes selection protocols between GHS and Cardiogenics (see Heinig *et al.* (2010) for details). However, the identification of hotspots at the 13q32 locus by HESS in GHS represents a significant replication of the findings previously reported, which reflects the increased power of HESS over alternative methods.

Discussion

We have presented a new hierarchical model and algorithm, HESS, for regression analysis of a large number of responses and predictors and have applied this to hotspot discovery in eQTL experiments. Simulating a variety of complex scenarios, we have demonstrated that our approach outperforms currently used algorithms. In particular, HESS shows increased power to detect hotspots when a large number of transcripts are jointly analyzed. This is due to the propensity measure ρ_j that we use, which quantifies the latent hotspot effect independently of the response dimensionality. One improvement of HESS over vanilla MCMC-based algorithms is in the search procedure that efficiently probes alternative models and assesses their importance, thus providing a reliable model

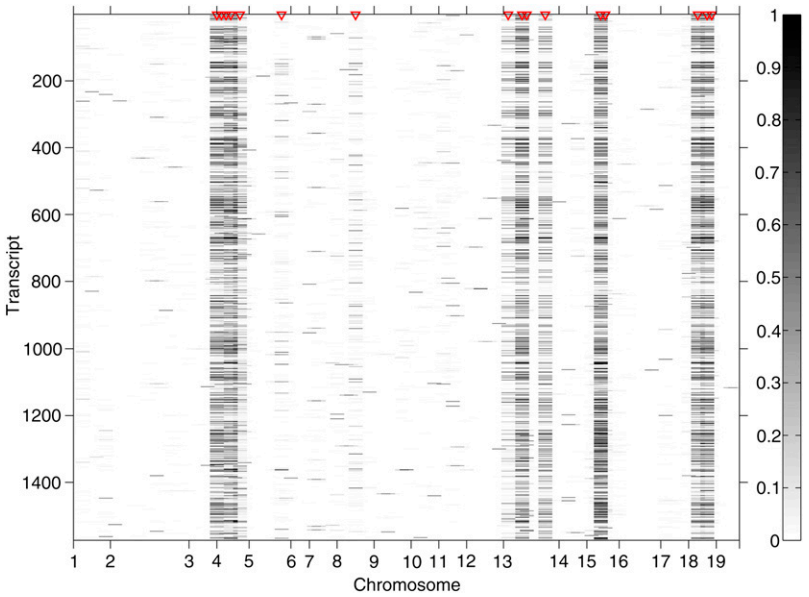

Figure 4 Heat map of the marginal probabilities of inclusion for each transcript–marker pair in the mouse data example ($n = 60$, $P = 145$, and $q = 1573$). The 16 red triangles indicate markers that have been identified as hotspots with tail posterior probability >0.8.

space exploration (Bottolo and Richardson 2010). We have also illustrated the potential of HESS to discover regulatory hotspots in two eQTL studies that encompass diverse genetic contexts (animal model and human data). In contrast to other methods, using HESS, we were able to replicate an established regulatory control of a large inflammatory network in humans (Heinig *et al.* 2010). Moreover, in the mouse data set, we identified a new candidate (*Kif13a*) for the regulation of a set of genes implicated in protein transport, which was not detected by other approaches.

Our model is embedded in the linear regression framework with additive effects. One distinct feature of our formulation is the multiplicative decomposition of the selection probabilities and its hierarchical set-up, which allows other structures and/or different types of prior information to be included. For example, specific weights π_{kj} (suitable normalized) could be added in (5), $\omega_{jk} = \omega_k \times \rho_j \times \pi_{kj}$, to provide additional prior information about the regulation of kth transcript. This may include *cis*-acting genetic control or auxiliary information on regulatory effects of the jth SNP (*i.e.*, evolutionary conservation, coding, noncoding, genomic location, etc.) (Lee *et al.* 2009). Likewise, additional structure on the responses (*e.g.*, KEGG pathways membership, predicted targets of transcription factors, protein complexes, etc.) could be included using k-indexed weights, ω_k, to favor detection of hotspots for similar responses.

Another possible extension of our method is the inclusion of interactions in the linear model and their efficient detection. Recent advances in this direction have employed either a stepwise search for interactions between preselected main effects (Wang *et al.* 2011) or partition models with which to discover modules or clusters of transcript–marker responses (Zhang *et al.* 2010). Such approaches could be embedded in our variable selection algorithm.

The current Matlab version of HESS represents a first step toward a more efficient implementation in high-level coding languages (currently undergoing), taking advantage of the existing C++ version of ESS algorithm (Bottolo *et al.* 2011). The approach that we propose here is ideally suited after prioritizing candidate genomic regions or gene networks, as shown in the discussed human case study. The flexibility to incorporate prior biological knowledge makes our method suitable for a wide range of analyses beyond eQTL hotspots detection, including genetic regulation of miRNA targets and metabolic and epigenetic phenotypes.

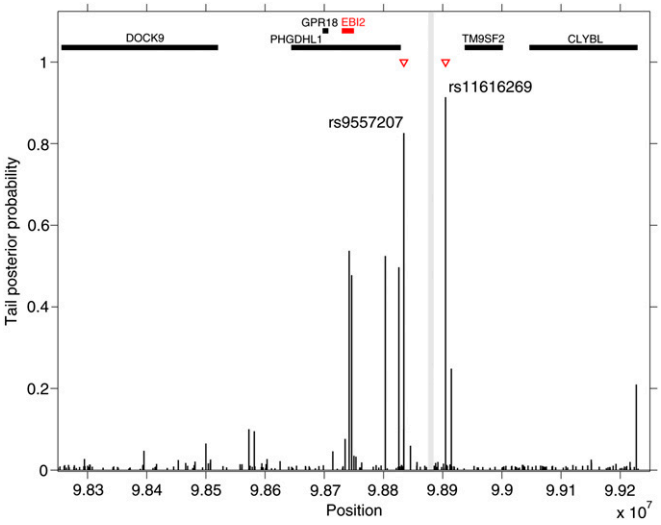

Figure 5 Tail posterior probability for each marker in the human data example (Gutenberg Heart Study, $n = 1490$, $P = 209$, and $q = 648$). Red triangles indicate markers that have been identified as hotspots with tail posterior probability >0.8. The vertical gray line highlights the physical position of annotated SNP rs9557217 and rs9585056 that were previously associated with IDIN network in the Cardiogenics Study cohort and *EBI2* expression (Heinig *et al.* 2010). Thick horizontal bars on the top of the figure display physical position of genes in the 1-Mb region obtained from Ensemble database.

Acknowledgments

We thank two anonymous referees and the associate editor for their comments that improved the manuscript. L.B. received funding from the Wellcome Trust Value-in-People award. S.R. gratefully acknowledges support from ESRC National Centre for Research Methods (BIAS II node, grant RES-576–25–0015). L.B. and S.R. acknowledge support from the Medical Research Council (grant G1002319). E.P. and S.A.C. acknowledge support from the Medical Research Council and the British Heart Foundation (grant FS/11/25/28740).

Literature Cited

Altshuler, D., L. D. Brooks, A. Chakravarti, F. S. Collins, M. D. Daly *et al.*, 2005 A haplotype map of the human genome. Nature 437: 1299–1320.

Banerjee, S., B. S. Yandell, and N. Yi, 2008 Bayesian quantitative trait loci mapping for multiple traits. Genetics 179: 2275–2289.

Bottolo, L., and S. Richardson, 2010 Evolutionary stochastic search for Bayesian model exploration. Bayesian Anal. 5: 583–618.

Bottolo, L., M. Chadeau-Hyam, D. I. Hastie, S. R. Langley, E. Petretto *et al.*, 2011 ESS++: a C++ objected-oriented algorithm for Bayesian stochastic search model exploration. Bioinformatics 27: 587–588.

Breitling, R., Y. Li, B. M. Tesson, J. Fu, T. Wiltshire *et al.*, 2008 Genetical genomics: spotlight on QTL hotspots. PLoS Genet. 4: e1000232.

Broman, K. W., and S. Sen, 2009 *A Guide to QTL Mapping with R/qtl*. Springer-Verlag, Berlin.

Chen, Y., J. Zhu, P. Y. Lum, X. Yang, S. Pinto *et al.*, 2008 Variations in DNA elucidate molecular networks that cause disease. Nature 452: 429–435.

Chun, H., and S. Keleş, 2009 Expression quantitative trait loci mapping with multivariate sparse partial least squares regression. Genetics 182: 79–90.

Cookson, W., L. Liang, G. Abecasis, M. Moffatt, and M. Lathrop, 2009 Mapping complex disease traits with global gene expression. Nat. Rev. Genet. 10: 184–194.

George, E. I., and R. E. McCulloch, 1997 Approaches for Bayesian variable selection. Statist. Sinica 7: 339–373.

Ghosh, D., W. Chen, and T. Raghunathan, 2006 The false discovery rate: a variable selection perspective. J. Statist. Plann. Inference 136: 2668–2684.

Gramacy, R. B., R. J. Samworth, and R. King, 2010 Importance tempering. Stat. Comput. 20: 1–7.

Hall, R. J., J. E. Hollis-Moffatt, M. E. Merriman, R. A. Green, D. Baker *et al.*, 2003 An autoimmune diabetes locus (Idd21) on mouse chromosome 18. Mamm. Genome 14: 335–339.

Heinig, M., E. Petretto, C. Wallace, L. Bottolo, M. Rotival *et al.*, 2010 A trans-acting locus regulates an anti-viral expression network and type 1 diabetes risk. Nature 467: 460–464.

Jia, Z., and S. Xu, 2007 Mapping quantitative trait loci for expression abundance. Genetics 176: 611–623.

Kass, R. E., and A. E. Raftery, 2007 Bayes factor. J. Am. Stat. Assoc. 90: 773–795.

Kendziorski, C. M., M. Chen, M. Yuan, H. Lan, and A. D. Attie, 2006 Statistical methods for expression quantitative trait loci (eQTL) mapping. Biometrics 62: 19–27.

Kohn, R., M. Smith, and D. Chan, 2001 Nonparametric regression using linear combinations of basis functions. Stat. Comput. 11: 313–322.

Korstanje, R., R. Li, T. Howard, P. Kelmenson, J. Marshall *et al.*, 2004 Influence of sex and diet on quantitative trait loci for hdl cholesterol levels in an SM/J by NZB/BlNJ intercross population. J. Lipid Res. 45: 881–888.

Lan, H., M. Chen, J. B. Flowers, B. S. Yandell, D. S. Stapleton *et al.*, 2006 Combined expression trait correlations and expression quantitative trait locus mapping. PLoS Genet. 2: e6.

Lee, S. I., A. M. Dudley, D. Drubin, P. Silver, N. J. Krogan *et al.*, 2009 Learning a prior on regulatory potential from eQTL data. PLoS Genet. 5: e1000358.

Majewski, J., and T. Pastinen, 2011 The study of eQTL variations by RNA-seq: from SNPs to phenotypes. Trends Genet. 27: 72–79.

Nakagawa, T., M. Setou, D. Seog, and N. Ogasawara Dohmae *et al.*, 2000 A novel motor, KIF13A, transports mannose-6-phosphate receptor to plasma membrane through direct interaction with AP-1 complex. Cell 103: 569–581.

Petretto, E., L. Bottolo, S. R. Langley, M. Heinig, M. C. McDermott-Roe *et al.*, 2010 New insights into the genetic control of gene expression using a Bayesian multi-tissue approach. PLOS Comput. Biol. 6: e1000737.

Richardson, S., A. Thomson, N. Best, and P. Elliott, 2004 Interpreting posterior relative risk estimates in disease mapping studies. Environ. Health Perspect. 112: 1016–1025.

Roberts, G. O., and J. S. Rosenthal, 2009 Examples of adaptive MCMC. J. Comput. Graph. Statist. 9: 349–367.

Schadt, E. E., 2009 A molecular networks as sensors and drivers of common human diseases. Nature 461: 218–223.

Scheet, P., and M. Stephens, 2006 A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. Am. J. Hum. Genet. 78: 629–644.

Sun, W., J. G. Ibrahim, and F. Zou, 2010 Genomewide multiple-loci mapping in experimental crosses by iterative adaptive penalized regression. Genetics 185: 349–359.

Wang, P., J. A. Dawson, M. P. Keller, B. S. Yandell, N. A. Thornberry *et al.*, 2011 A model selection approach for expression quantitative trait loci (eQTL) mapping. Genetics 187: 611–621.

Wu, C., D. L. Delano, N. Mitro, S. V. Su, J. Janes *et al.*, 2008 Gene set enrichment in eqtl data identifies novel annotations and pathway regulators. PLoS Genet. 5: e1000070.

Xu, C., X. Wang, Z. Li, and S. Xu, 2008 Mapping QTL for multiple traits using Bayesian statistics. Genet. Res. Camb. 91: 23–37.

Yi, N., and D. Shriner, 2008 Advances in Bayesian multiple QTL mapping in experimental designs. Heredity 100: 240–252.

Yi, N., and S. Xu, 2008 Bayesian lasso for quantitative trait loci mapping. Genetics 179: 1045–1055.

Yi, N., D. Shriner, S. Banerjee, T. Mehta, D. Pomp *et al.*, 2007 An efficient Bayesian model selection approach for interacting quantitative trait loci models with many effects. Genetics 176: 1865–1877.

Yvert, G., R. B. Brem, J. Whittle, J. M. Akey, E. Foss *et al.*, 2003 Trans-acting regulatory variation in *Saccharomyces cerevisiae* and the role of transcription factors. Nat. Genet. 35: 57–64.

Zeller, T., P. Wild, S. Szymczak, M. Rotival, A. Schillert *et al.*, 2010 Genetics and beyond: the transcriptome of human monocytes and disease susceptibility. PLoS ONE 5: e10693.

Zhang, W., J. Zhu, E. E. Schadt, and J. S. Liu, 2010 A Bayesian partition method for detecting pleiotropic and epistatic eQTL modules. PLOS Comput. Biol. 6: e1000642.

Communicating editor: G. A. Churchill

GENETICS

Supporting Information
http://www.genetics.org/cgi/content/full/genetics.111.131425/DC1

Bayesian Detection of Expression Quantitative Trait Loci Hot Spots

Leonardo Bottolo, Enrico Petretto, Stefan Blankenberg, François Cambien, Stuart A. Cook, Laurence Tiret, and Sylvia Richardson,

Copyright © 2011 by the Genetics Society of America
DOI: 10.1534/genetics.111.131425

File S1

Supporting Information

S.1 Notation

We briefly recall here the notation that was used along the paper. Moreover we introduce some new notation to easy the the illustration of the MCMC scheme.

Let $\boldsymbol{Y}$ and $\boldsymbol{X}$ the $n \times q$ and $n \times p$ matrix of the responses and predictors, respectively. Let $\boldsymbol{\Gamma} = \{\gamma_{lkj}, 1 \leq l \leq L, 1 \leq k \leq q, 1 \leq j \leq p\}$ the matrix of latent binary values, where L is the number of simulated chains, q is the number of responses and p is the number of predictors and let $\boldsymbol{\Gamma}_k = (\boldsymbol{\gamma}_{1k}, \ldots, \boldsymbol{\gamma}_{lk}, \ldots, \boldsymbol{\gamma}_{Lk})^T$ the $L \times p$ latent binary matrix for the kth response in expanded state-space, where $\boldsymbol{\gamma}_{lk} = (\gamma_{lk1}, \ldots, \gamma_{lkj}, \ldots, \gamma_{lkp})^T$. Similarly let $\boldsymbol{\Omega} = \{\omega_{lkj}, 1 \leq l \leq L, 1 \leq k \leq q, 1 \leq j \leq p\}$ the matrix of selection probability with $\omega_{lkj} = \omega_{lk} \times \rho_{lj}$ and let $\boldsymbol{\Omega}_k = (\boldsymbol{\omega}_{1k}, \ldots, \boldsymbol{\omega}_{lk}, \ldots, \boldsymbol{\omega}_{Lk})^T$ the $L \times p$ selection matrix for the kth response in expanded state-space, where $\boldsymbol{\omega}_{lk} = (\omega_{lk1}, \ldots, \omega_{lkj}, \ldots, \omega_{lkp})^T$. For a given chain l, let $\boldsymbol{\omega}_l = (\omega_{l1}, \ldots, \omega_{lk}, \ldots, \omega_{lq})^T$ and $\boldsymbol{\rho}_l = (\rho_{l1}, \ldots, \rho_{lj}, \ldots, \rho_{lp})^T$ the 'row' and the 'column' effect, respectively. Finally the temperature ladder for each regression equation k is denoted by $\boldsymbol{t}_k = (t_{1k}, \ldots, t_{lk}, \ldots, t_{Lk})^T$ with $1 = t_{1k} < t_{2k} < \ldots < t_{Lk}$.

S.2 Technical details of MCMC implementation

S.2.1 Full conditionals

Given (6), to sample the binary latent value γ_{lkj}, the selection probability $\omega_{lkj} = \omega_{lk} \times \rho_{lj}$ and the scaling coefficient τ, the tempered full conditionals in the expanded state-space are:

- $p(\boldsymbol{\gamma}_{lk}|\cdots) \propto p(\boldsymbol{y}_k|\boldsymbol{X}, \boldsymbol{\gamma}_{lk}, \tau)^{1/t_{lk}} \prod_{j=1}^{p} p(\gamma_{lkj}|\omega_{lkj})^{1/t_{lk}}$
- $p(\omega_{lk}|\cdots) \propto p(\omega_{lk})^{1/t_{lk}} \prod_{j=1}^{p} p(\gamma_{lkj}|\omega_{lkj})^{1/t_{lk}}$
- $p(\rho_{lj}|\cdots) \propto p(\rho_{lj}) \prod_{k=1}^{q} p(\gamma_{lkj}|\omega_{lkj})^{1/t_{lk}}$
- $p(\tau|\cdots) \propto p(\tau) \prod_{l=1}^{L} \prod_{k=1}^{q} p(\boldsymbol{y}_k|\boldsymbol{X}, \boldsymbol{\gamma}_{lk}, \tau)^{1/t_{lk}}$

Note that in the full conditional $p(\rho_{lj}|\cdots)$ the prior density $p(\rho_{lj})$ is not tempered and the reason will be explained in Supporting Information S.2.3.

S.2.2 $\boldsymbol{\Gamma}$ update

The update of the elements of the $q \times p$ latent binary matrix $\boldsymbol{\Gamma}$ is of paramount importance and efficient algorithms are required in order to visit the very large model space $(2^p)^q$ and to escape from local modes. In the following we provide some technical details omitted from the main text of the local and global moves that we found useful to implement. At each sweep of the algorithm each/both of moves can be applied to *all* the q regression equations or to a random without replacement subgroup of them (see Richardson et al. (2011) for alternative subgroup selection with adaptive probability).

Local move

We first introduce the single chain sampling scheme and then we extend the results for multiple chains. There are many ways to update locally $\boldsymbol{\gamma}_k$, but we found useful to apply an extension of Bottolo and Richardson (2010) proposal, where traditional samplers used in Bayesian variable selection (*i.e.* MC^3, Gibbs sampler and Reversible Jump) are replaced by a Metropolis-within-Gibbs sampler known as Fast Scan Metropolis-Hastings (FSMH). Let $L_{k(j=1)} = p(\boldsymbol{y}_k|\boldsymbol{X}, \boldsymbol{\gamma}_{k(j=1)}, \tau)$ and $L_{k(j=0)} = p(\boldsymbol{y}_k|\boldsymbol{X}, \boldsymbol{\gamma}_{k(j=0)}, \tau)$ with $\boldsymbol{\gamma}_{k(j=1)} = (\gamma_{k1}, \ldots, \gamma_{kj} = 1, \ldots, \gamma_{kp})^T$ and $\boldsymbol{\gamma}_{k(j=0)} = (\gamma_{k1}, \ldots, \gamma_{kj} = 0, \ldots, \gamma_{kp})^T$ the marginal likelihood once the regression coefficients $\boldsymbol{\beta}_k$ and the residual error variance σ_k^2 are integrated out. Moreover let $p(\gamma_{kj} = 1|\omega_{kj}) = \omega_{kj}$ and $p(\gamma_{kj} = 0|\omega_{kj}) = 1 - \omega_{kj}$. If a Gibbs sampler update is performed, a new value of γ_{kj} is drawn from a Bernoulli distribution with probability

$$\theta_{kj} = \frac{\omega_{kj} L_{k(j=1)}}{(1-\omega_{kj}) L_{k(j=0)} + \omega_{kj} L_{k(j=1)}} \tag{S.1}$$

if, in the previous iteration, $\gamma_{kj} = 0$ since by independence $p(\gamma_{kj} = 1|\boldsymbol{\gamma}_{k\setminus j}, \boldsymbol{\omega}_k) = p(\gamma_{kj} = 1|\omega_{kj})$ (with an obvious modification if $\gamma_{kj} = 1$ in the previous iteration). However in a sparse framework, where $p_{\boldsymbol{\gamma}_k} \ll p$, this probability is dominated by ω_{kj} and if ω_{kj} is small (because for instance ω_k or ρ_j or both are small) also θ_{kj} will be small. For instance, it easy to show that when $p_{\boldsymbol{\gamma}_k} \ll p$ and therefore by Kohn et al. (2001) $a_k \ll b_k$, the sampled value of ω_k is, on average, very small

$$E(\omega_k|\boldsymbol{y}_k) = \frac{p_{\boldsymbol{\gamma}_k} + a_k}{p + a_k + b_k}.$$

It turns out that, if $\gamma_{kj} = 0$, it is likely that also the new sampled value will be zero. Kohn et al. (2001) propose to split the acceptance probability of the Metropolised version of (S.1) (to add a

new covariate in the regression)

$$1 \wedge \frac{\omega_{kj} L_{k(j=1)}}{(1-\omega_{kj}) L_{k(j=0)}} \frac{Q_{kj}(1 \to 0)}{Q_{kj}(0 \to 1)},$$

where $Q_{kj}(\cdot \to \cdot)$ is the proposal density, into two parts: firstly, sampling a proposed value of γ_{kj}, γ^*_{kj}, from a Bernoulli distribution with probability ω_{kj} and then, if $\gamma^*_{kj} \neq \gamma_{kj}$, accept the new value with probability

$$1 \wedge \frac{L_{k(j=1)}}{L_{k(j=0)}}$$

since $Q_{kj}(0 \to 1) = \omega_{kj}$ and $Q_{kj}(1 \to 0) = 1 - \omega_{kj}$, with an obvious modification if a deletion is proposed. The advantage of this scheme is that the time consuming evaluation of the marginal likelihood L_{kj} is limited to the set of variables where $\gamma^*_{kj} \neq \gamma_{kj}$.

The same sampling scheme can be extended to a parallel tempering set-up as illustrated in Bottolo and Richardson (2010). In this case the Metropolis-within-Gibbs acceptance probability of the jth predictor in the kth regression and the lth chain is

$$1 \wedge \frac{L^{1/t_{lk}}_{lk(j=1)}}{L^{1/t_{lk}}_{lk(j=0)}},$$

where $L^{1/t_{lk}}_{lk(j=1)} = [p(\boldsymbol{y}_k|\boldsymbol{X}, \boldsymbol{\gamma}_{lk(j=1)}, \tau)]^{1/t_{lk}}$ and similarly for $L^{1/t_{lk}}_{lk(j=0)}$, since adding (deleting) a covariate in the regression equation is proposed with probability $Q_{lkj}(0 \to 1|t_{lk}) = \tilde{\omega}_{lkj}(t_{lk})$ ($Q_{lkj}(1 \to 0|t_{lk}) = 1 - \tilde{\omega}_{lkj}(t_{lk})$), with

$$\tilde{\omega}_{lkj}(t_{lk}) = \frac{\omega^{1/t_{lk}}_{lkj}}{\omega^{1/t_{lk}}_{lkj} + (1-\omega_{lkj})^{1/t_{lk}}}$$

the renormalised probability $[p(\gamma_{lkj} = 1|\omega_{lkj})]^{1/t_{lk}} = \omega^{1/t_{lk}}_{lkj}$ and t_{lk} the temperature attached to the kth regression in the lth chain. Further discussion and advantages of this sampling scheme over MC^3, Reversible Jump and Gibbs sampler in a multiple chain set-up when the number of

predictors is very large with respect to the number of truly associated variables are presented in Bottolo and Richardson (2010).

Global moves

We recall that global moves are bold moves that try to swap part or the whole state of two randomly selected chains among the population of chains (Liang and Wong, 2000). In the following we present the accepted probability of crossover operator (partial swap), exchange operator and all-exchange operator (full swap).

Suppose that in the kth regression two new latent binary vectors $\boldsymbol{\gamma}_{lk}^*$ and $\boldsymbol{\gamma}_{rk}^*$ are generated from two preselected chains, l and r, according to some crossover operator (Liang and Wong, 2000; Bottolo and Richardson, 2010). The proposed population of chains in the kth regression $\boldsymbol{\Gamma}_k^* = (\boldsymbol{\gamma}_{1k}, \ldots, \boldsymbol{\gamma}_{lk}^*, \ldots, \boldsymbol{\gamma}_{rk}^* \ldots, \boldsymbol{\gamma}_{Lk})^T$ is accepted with probability

$$1 \wedge \frac{\exp\left\{f(\boldsymbol{\gamma}_{lk}^*|\boldsymbol{\omega}_{lk},\tau)/t_{lk} + f(\boldsymbol{\gamma}_{rk}^*|\boldsymbol{\omega}_{rk},\tau)/t_{rk}\right\} Q_k(\boldsymbol{\Gamma}_k^*, \boldsymbol{\Gamma}_k|\boldsymbol{\Omega}_k,\tau,\boldsymbol{t}_k)}{\exp\left\{f(\boldsymbol{\gamma}_{lk}|\boldsymbol{\omega}_{lk},\tau)/t_{lk} + f(\boldsymbol{\gamma}_{rk}|\boldsymbol{\omega}_{rk},\tau)/t_{rk}\right\} Q_k(\boldsymbol{\Gamma}_k, \boldsymbol{\Gamma}_k^*|\boldsymbol{\Omega}_k,\tau,\boldsymbol{t}_k)},$$

where $f(\boldsymbol{\gamma}_{lk}|\boldsymbol{\omega}_{lk},\tau) = \log(p(\boldsymbol{y}_k|\boldsymbol{X},\boldsymbol{\gamma}_{lk},\tau)) + \sum_j \log(p(\gamma_{lkj}|\omega_{lkj}))$ and $Q_k(\boldsymbol{\Gamma}_k,\cdot|\boldsymbol{\Omega}_k,\tau,\boldsymbol{t}_k)$ is the proposal density which is defined as the product of the selection probability and the crossover operator probability (Liang and Wong, 2000). The transition density depends on the selection probabilities $\boldsymbol{\Omega}_k$ in the kth regression, the scaling coefficient τ and the kth regression temperature ladder $\boldsymbol{t}_k$.

The exchange operator can be seen as special case of the crossover operator where the whole information contained in the two preselected chains with uniform probability l and r are tentatively swapped with probability

$$1 \wedge \frac{\exp\left\{f(\boldsymbol{\gamma}_{rk}|\boldsymbol{\omega}_{lk},\tau)/t_{lk} + f(\boldsymbol{\gamma}_{lk}|\boldsymbol{\omega}_{rk},\tau)/t_{rk}\right\}}{\exp\left\{f(\boldsymbol{\gamma}_{lk}|\boldsymbol{\omega}_{lk},\tau)/t_{lk} + f(\boldsymbol{\gamma}_{rk}|\boldsymbol{\omega}_{rk},\tau)/t_{rk}\right\}}$$

since $Q_k(\boldsymbol{\Gamma}_k, \boldsymbol{\Gamma}_k^*|\boldsymbol{\Omega}_k, \tau, \boldsymbol{t}_k) = Q_k(\boldsymbol{\Gamma}_k^*, \boldsymbol{\Gamma}_k|\boldsymbol{\Omega}_k, \tau, \boldsymbol{t}_k)$ because the selection probability is uniform over the L chains (random selection without replacement).

Finally, in the all-exchange operator the chains whose states are swopped are selected at random with probability equal to

$$p_{hk} = \frac{\tilde{p}_{hk}}{\sum_{h=1}^{1+L(L-1)/2} \tilde{p}_{hk}}, \tag{S.2}$$

where in (S.2) each pair $(l, r < l)$ is denoted by a single number h, $\tilde{p}_{hk} = \tilde{p}_{(l,r)k}$, including the rejection move, $h = 1$ with $\tilde{p}_{(l,r)k} = \exp\{(f(\boldsymbol{\gamma}_{rk}|\boldsymbol{\omega}_{rk}, \tau) - f(\boldsymbol{\gamma}_{lk}|\boldsymbol{\omega}_{lk}, \tau))(1/t_{lk} - 1/t_{rk})\}$.

S.2.3 $\boldsymbol{\Omega}$ update

For each chain l, $l = 1, \ldots, L$, we update the elements of the $q \times p$ selection probability matrix $\boldsymbol{\Omega}$ by using a Metropolis-within-Gibbs sampler with adaptive proposals. Let ω_{lk}^* and ρ_{lj}^* the proposed new values of the kth row effect and jth column effect in the lth chain respectively. The acceptance probability of the two parameters is

$$1 \wedge \left[\frac{(\omega_{lk}^*)^{p_{\boldsymbol{\gamma}_{lk}}}(1-\omega_{lk}^*)^{p-p_{\boldsymbol{\gamma}_{lk}}}}{\omega_{lkj}^{p_{\boldsymbol{\gamma}_{lk}}}(1-\omega_{lkj})^{p-p_{\boldsymbol{\gamma}_{lk}}}} \frac{Beta(\omega_{lk}^*; a_{\omega_k}, b_{\omega_k})|J(\lambda^{-1}(\omega_{lk}^*))|}{Beta(\omega_{lk}; a_{\omega_k}, b_{\omega_k})|J(\lambda^{-1}(\omega_{lk}))|} \right]^{1/t_{lk}} \frac{Q_{lk}(\lambda_{lk}^*, \lambda_{lk})}{Q_{lk}(\lambda_{lk}, \lambda_{lk}^*)} \tag{S.3}$$

and

$$1 \wedge \frac{Ga(\rho_{lj}^*; c_{\rho_j}, d_{\rho_j})|J(\varphi^{-1}(\rho_{lj}^*))|}{Ga(\rho_{lj}; c_{\rho_j}, d_{\rho_j})|J(\varphi^{-1}(\rho_{lj}))|} \prod_{k=1}^{q} \left[\frac{(\omega_{lkj}^*)^{\gamma_{lkj}}(1-\omega_{lkj}^*)^{1-\gamma_{lkj}}}{\omega_{lkj}^{\gamma_{lkj}}(1-\omega_{lkj})^{1-\gamma_{lkj}}} \right]^{1/t_{lk}} \frac{Q_{lj}(\varphi_{lj}^*, \varphi_{lj})}{Q_{lj}(\varphi_{lj}, \varphi_{lj}^*)}, \tag{S.4}$$

where in (S.3) $p_{\boldsymbol{\gamma}_{lk}} = \boldsymbol{\gamma}_{lk}^T \mathbf{1}_p$, $\lambda_{lk} = \text{logit}(\omega_{lk})$, $J(\lambda^{-1}(\omega_{lk}))$ is the Jacobian of the inverse transformation evaluated in ω_{lk} and $Beta(\cdot)$ is the beta density function, while in (S.4) $J(\varphi^{-1}(\rho_{lj}))$ is the Jacobian of the inverse transformation evaluated in ρ_{lj}, $\omega_{lkj}^* = \omega_{lk} \times \rho_{lj}^*$, and $Ga(\cdot)$ is the gamma density function. As a technical point, since the prior density $p(\rho_{lj})$ cannot be indexed

by k, in order to write the acceptance probability (S.4), in our model the prior for ρ_{lj} is not tempered.

We sample the proposed new values ω_{lk}^* and ρ_{lj}^* after suitable transformation from $Q_{lk}(\lambda_{lk}, \cdot) = \phi(\lambda_{lk}, s_{lk}^2(b))$ and $Q_{lj}(\varphi_{lj}, \cdot) = \phi(\varphi_{lj}, s_{lj}^2(b))$, respectively, where $s_{lk}(b)$ and $s_{lk}(b)$ are the adaptive proposals' standard deviations at batch b and $\phi(\cdot)$ is the normal density function. Following Roberts and Rosenthal (2009), *asymptotic convergence* is obtained enforcing the *diminishing adaptation condition* and imposing the *bounded convergence condition*. For the former condition, after the batch bth of 50 sweeps, say, the proposals' standard deviation are updated as follow: $s_{lk}(b+1) = s_{lk}(b) \pm \delta_s(b)$ and $s_{lj}(b+1) = s_{lj}(b) \pm \delta_s(b)$, where we add (subtract) to the current values $s_{lk}(b)$ and $s_{lj}(b)$ the quantity $\delta_s(b) = \min\{0.01, b^{-1/2}\}$ if the acceptance frequency of (S.3) and (S.4) are higher (lower) than the optimal acceptance rate (0.44), respectively. The latter condition is fulfilled assuming that $L_\lambda < s_{lk} < U_\lambda$ and $L_\varphi < s_{lj} < U_\varphi$ for some large positive (negative) values of U_λ and U_φ (L_λ and L_φ).

S.2.4 τ updates

The variable scaling coefficient is common to *all* the q regression equations and to *all* L chains. A new value τ^* is obtained using a Metropolis-with-Gibbs with acceptance probability

$$1 \wedge \frac{Ga(\tau^*; 1/2, n/2)|J(\psi^{-1}(\tau^*))| \prod_{l=1}^{L} \prod_{k=1}^{q} p(\boldsymbol{y}_k|\boldsymbol{X}, \boldsymbol{\gamma}_{lk}, \tau^*)^{1/t_{lk}}}{Ga(\tau; 1/2, n/2)|J(\psi^{-1}(\tau))| \prod_{l=1}^{L} \prod_{k=1}^{q} p(\boldsymbol{y}_k|\boldsymbol{X}, \boldsymbol{\gamma}_{lk}, \tau)^{1/t_{lk}}} \frac{Q(\psi^*, \psi)}{Q(\psi, \psi^*)},$$

where $\psi = \log(\tau)$, $J(\psi^{-1}(\tau))$ is the Jacobian of the inverse transformation evaluated in τ, $Ga(\cdot)$ is the gamma density function and $Q(\psi, \cdot) = \phi(\psi, 1)$. As in (S.4), the prior density is not tempered since we are sampling a common value across the q regressions and the L chains. The

rational of this choice, for a given k, is illustrated in detail in Bottolo and Richardson (2010).

S.2.5 Temperature placement

During the burn-in, for each regression equation k, we automatically tune the temperature ladder in order to reach a specified acceptance rate of the exchange operator. In particular we chose as temperature ladder the geometric scale, such that the ratio of two consecutive temperatures is constant, $t_{(l+1)k}/t_{lk} = r_k$. Then after batch bth, say 100 sweeps, we update r_k as follows: $r_k(b+1) = r_k(b) \pm \delta_r$, where we add (subtract) to the current values $r_k(b)$ the quantity δ_r if the acceptance frequency of the exchange operator are higher (lower) than the optimal acceptance rate (0.50). For details on how to fix the value of δ_r interested reader can refer to Bottolo and Richardson (2010). For a discussion of different temperature scales, see Atchadé et al. (2010).

S.3 Post-processing

For a fixed k,

$$p(\boldsymbol{\gamma}_k^{(t)}|\boldsymbol{y}_k) = \frac{1}{S}\sum_{s=1}^{S} p(\boldsymbol{y}_k|\boldsymbol{X}, \boldsymbol{\gamma}_k^{(t)}, \tau^{(s)}) p(\tau^{(s)}) \prod_{j=1}^{p} p(\gamma_{kj}^{(t)}|\omega_{kj}^{(s)}) p(\rho_j^{(s)})$$

is the model posterior probability for the kth regression, where $\boldsymbol{\gamma}_k^{(t)} = (\gamma_{k1}^{(t)}, \ldots, \gamma_{kq}^{(t)})^T$ is latent binary vector recorded at the tth sweep of the algorithm, $p(\boldsymbol{y}_k|\boldsymbol{X}, \boldsymbol{\gamma}_k^{(t)}, \tau^{(s)})$ is the marginal likelihood and $\tau^{(s)}$, $\omega_{kj}^{(s)} = \omega_k^{(s)} \times \rho_j^{(s)}$ and $\rho_j^{(s)}$ are the values of the parameters recorded at the sth sweep.

When the q regressions are jointly considered, the configuration posterior probability is de-

fined as

$$p(\mathbf{\Gamma}^{(t)}|\boldsymbol{Y}) = \frac{1}{S}\sum_{s=1}^{S} p(\tau^{(s)}) \prod_{k=1}^{q} p(\boldsymbol{y}_k|\boldsymbol{X}, \boldsymbol{\gamma}_k^{(t)}, \tau^{(s)}) p(\omega_k^{(s)}) \prod_{j=1}^{p} p(\gamma_{kj}^{(t)}|\omega_{kj}^{(s)}) p(\rho_j^{(s)})$$

with $\mathbf{\Gamma}^{(t)}$ the configuration of the latent binary matrix at sweep tth.

Further literature cited

Atchadé, Y. F., G. O. Roberts and J. S. Rosenthal (2010). Towards optimal scaling of metropolis-coupled markov chain monte carlo. *Stat. Comput.*. To appear.

Bottolo, L. and S. Richardson (2010). Evolutionary Stochastic Search for Bayesian model exploration. *Bayesian Analysis 5*, 583–618.

Kohn, R., M. Smith and D. Chan (2001). Nonparametric regression using linear combinations of basis functions. *Stat. Comput. 11*, 313–322.

Liang, F. and W. H. Wong (2000). Evolutionary Monte Carlo: application to C_p model sampling and change point problem. *Stat. Sinica 10*, 317–342.

Richardson, S., L. Bottolo and J. S. Rosenthal (2011). Bayesian models for sparse regression analysis of high dimensional data (with discussion). In J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. Dawid, D. Heckerman, A. F. M. Smith, and M. West (Eds.), *Bayesian Statistics: Proc. 9th Int. Meeting*. Oxford University Press.

Roberts, G. O. and J. S. Rosenthal (2009). Examples of adaptive MCMC. *J. Comput. Graph. Stat. 9*, 349–367.

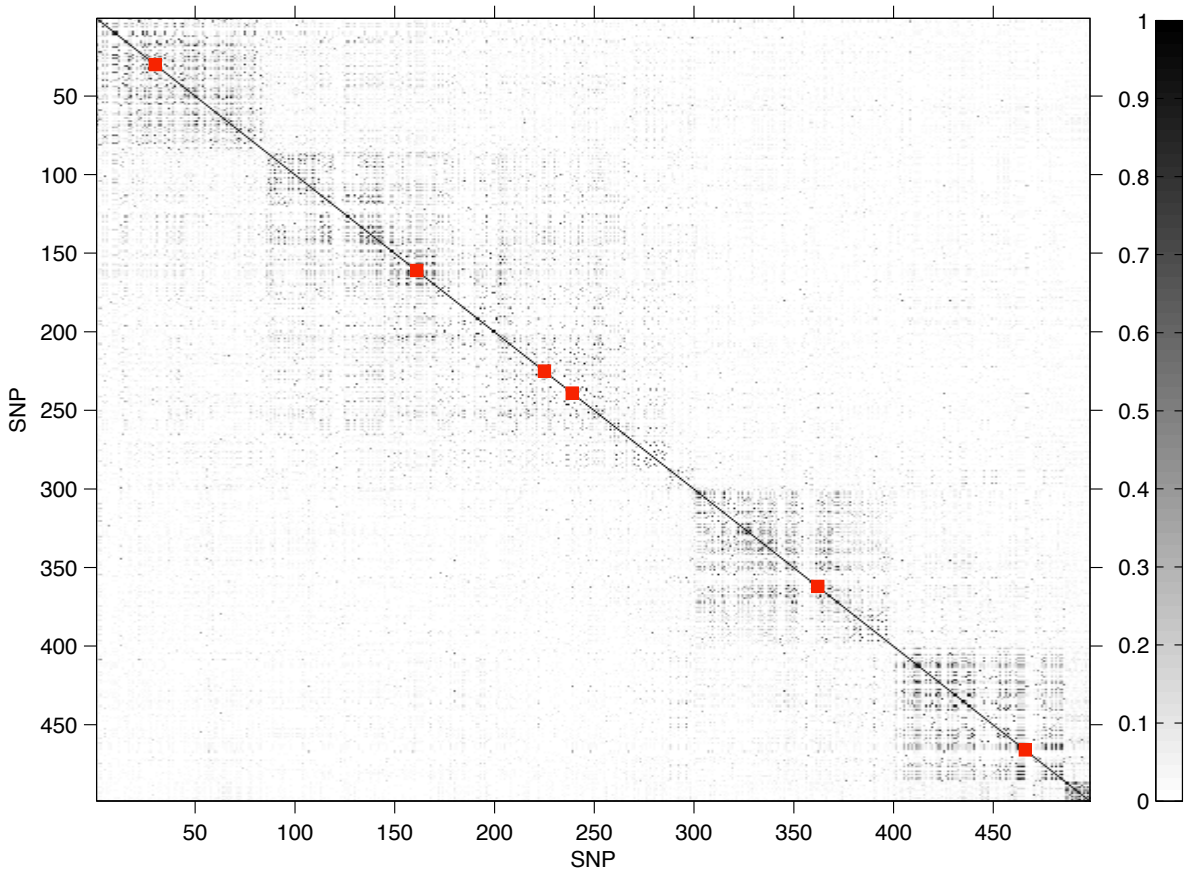

Figure S.1: Heat-map of the pattern of correlation, linkage disequilibrium (LD) for Yoruba population, HapMap project, in the region ENm014 spanning 500-Kb (chrom 7: 126,368,183-126,865,324 bp). Red squares indicate the marker where the hot-spots have been simulated (SNP 30, 161, 225, 239, 362 and 466).

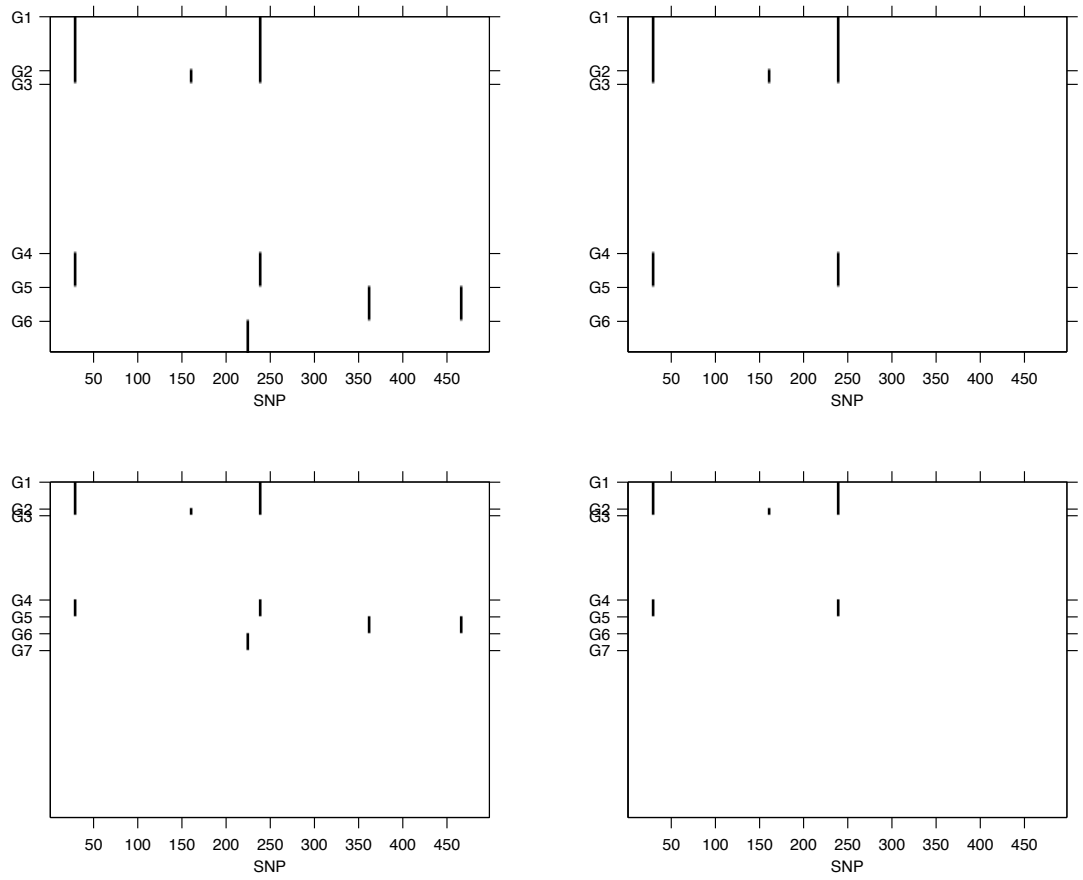

Figure S.2: Map configuration in the four simulated scenarios. From top to bottom, left to right: SIM1, SIM2, SIM3 and SIM4. SIM1, $q = 100$ transcripts simulated with SNP 30 and 239 influencing transcripts 1-20 and 71-80, SNP 161 influencing transcripts 17-20, SNP 225 influencing transcripts 91-100, and finally eQTLs 362 and 466 influencing transcripts 81-90. Altogether 94 transcript-SNP associations are simulated in 50 distinct transcripts; SIM2, 100 responses simulated with only three hot spots (30, 161, 239) and the same simulated pattern of association as in the first scenario leading to 64 transcript-SNP associations in 30 distinct transcripts; SIM3, the simulation set-up is identical to the first scenario for the first 100 responses, but the number of simulated responses is increased to $q = 1,000$, simulating further 900 transcripts from the noise; SIM4, as in the second simulated data set for the first 100 responses, with additional 900 responses simulated from the noise, and altogether $q = 1,000$. The symbol 'G' in the y-axis identifies groups of transcripts that are influenced by the same pattern of markers. SIM1 and SIM2, G_1: transcripts 1-16; G_2: transcripts 17-20; G_3: 21-70; G_4: transcripts 71-80; G_5: transcripts 81-90; G_6: transcripts 91-100. SIM3 and SIM4 as before with G_7: transcripts 101-1000.

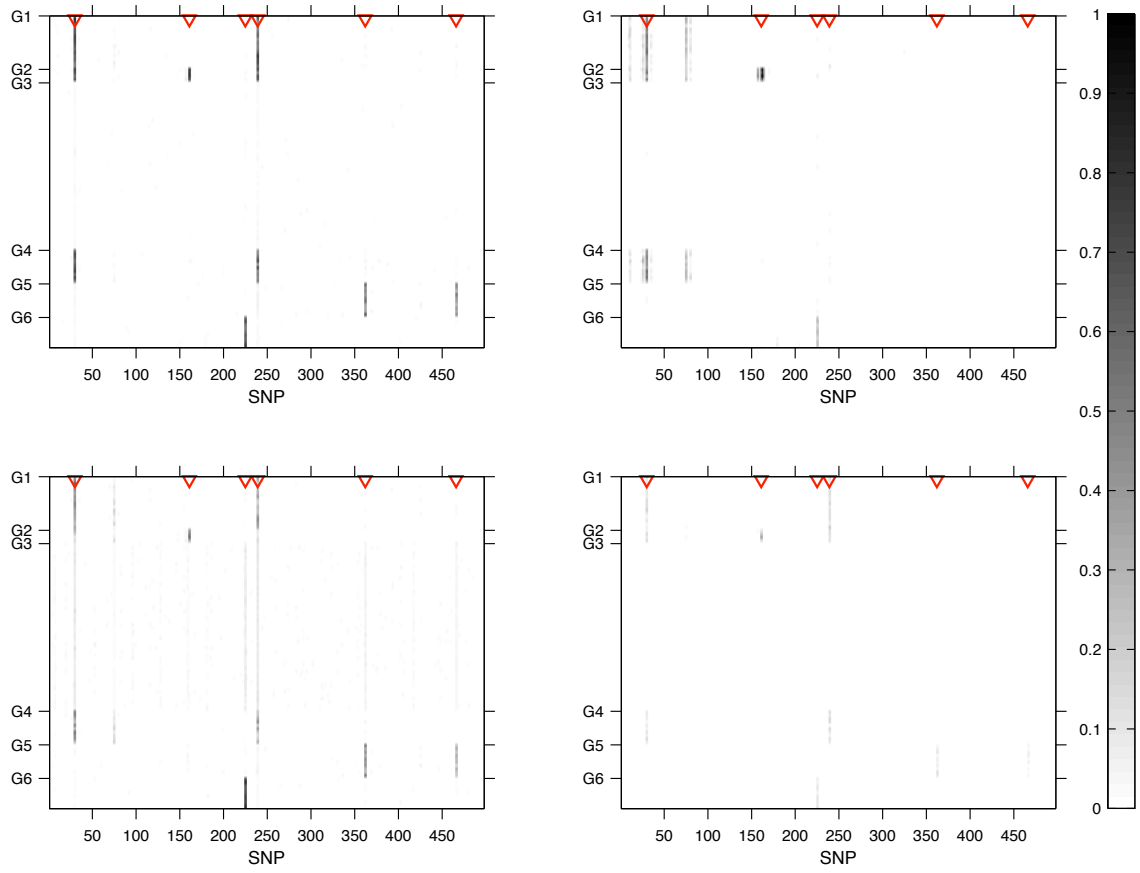

Figure S.3: Heat-map of the signals detected by each method in the first simulated example, SIM1, and averaged across the 25 replicates. In M-SPLS the significant (non-significant) transcript-marker association is recoded as 1 (0). From top to bottom, left to right: HESS, M-SPLS, MOM and BAYES. The symbol 'G' in the y-axis identifies groups of transcripts that are influenced by the same pattern of markers. Red triangles indicate where the hot-spots have been simulated.

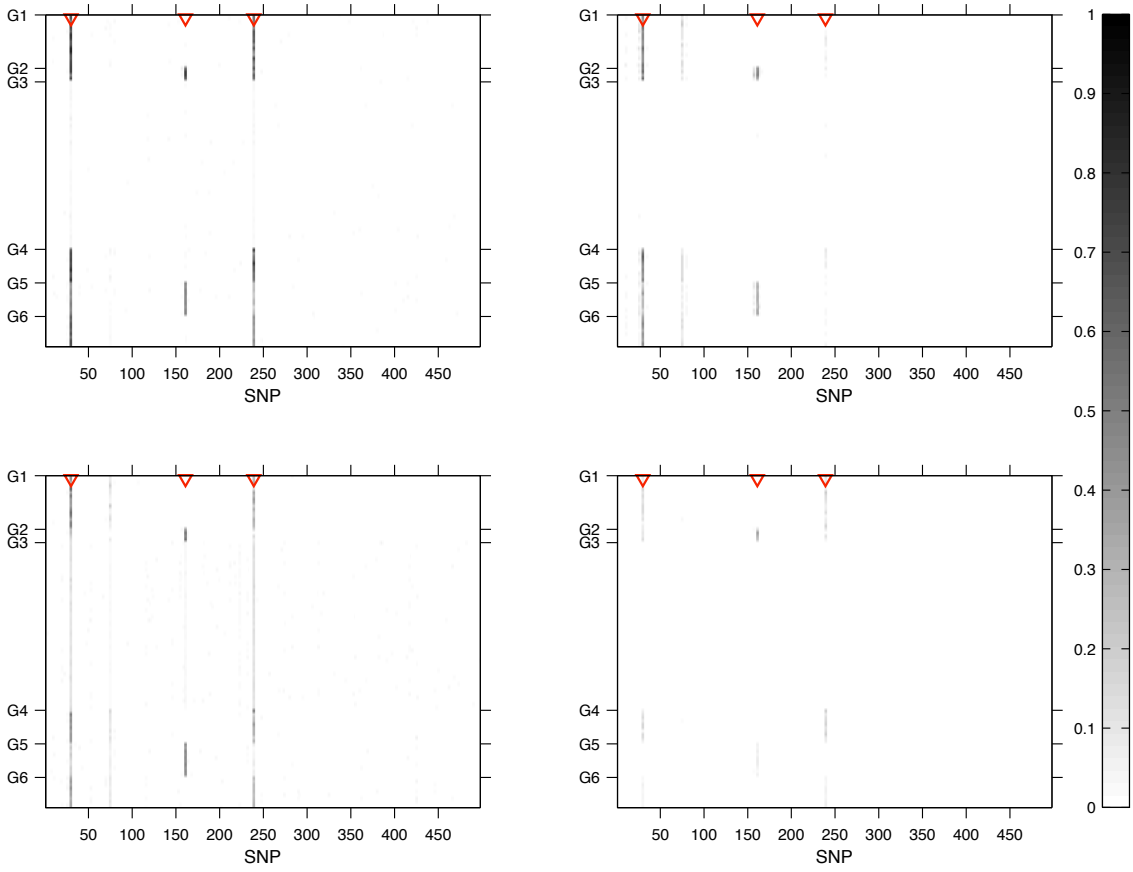

Figure S.4: Heat-map of the signals detected by each method in the second simulated example, SIM2, and averaged across the 25 replicates. In M-SPLS the significant (non-significant) transcript-marker association is recoded as 1 (0). From top to bottom, left to right: HESS, M-SPLS, MOM and BAYES. The symbol 'G' in the y-axis identifies groups of transcripts that are influenced by the same pattern of markers. Red triangles indicate where the hot-spots have been simulated.

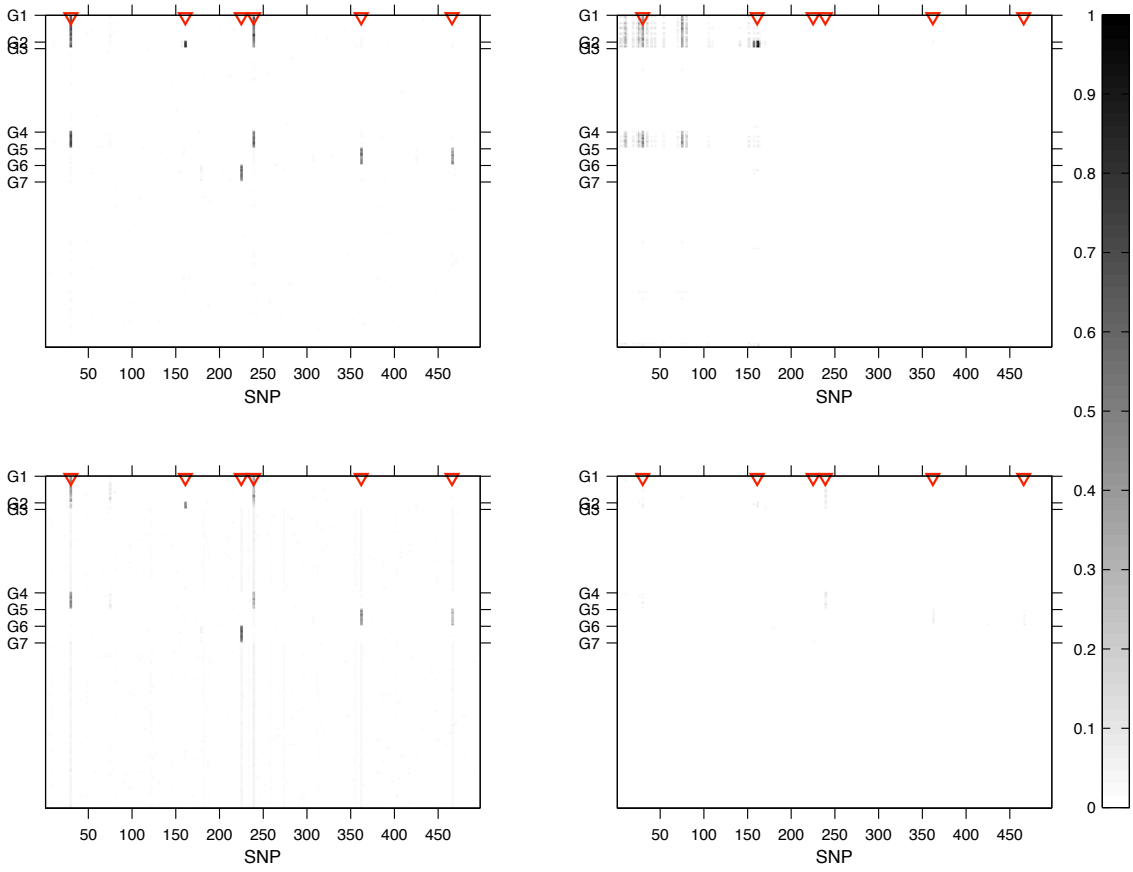

Figure S.5: Heat-map of the signals detected by each method in the third simulated example, SIM3, and averaged across the 25 replicates. In M-SPLS the significant (non-significant) transcript-marker association is recoded as 1 (0). From top to bottom, left to right: HESS, M-SPLS, MOM and BAYES. The symbol 'G' in the y-axis identifies groups of transcripts that are influenced by the same pattern of markers. Red triangles indicate where the hot-spots have been simulated.

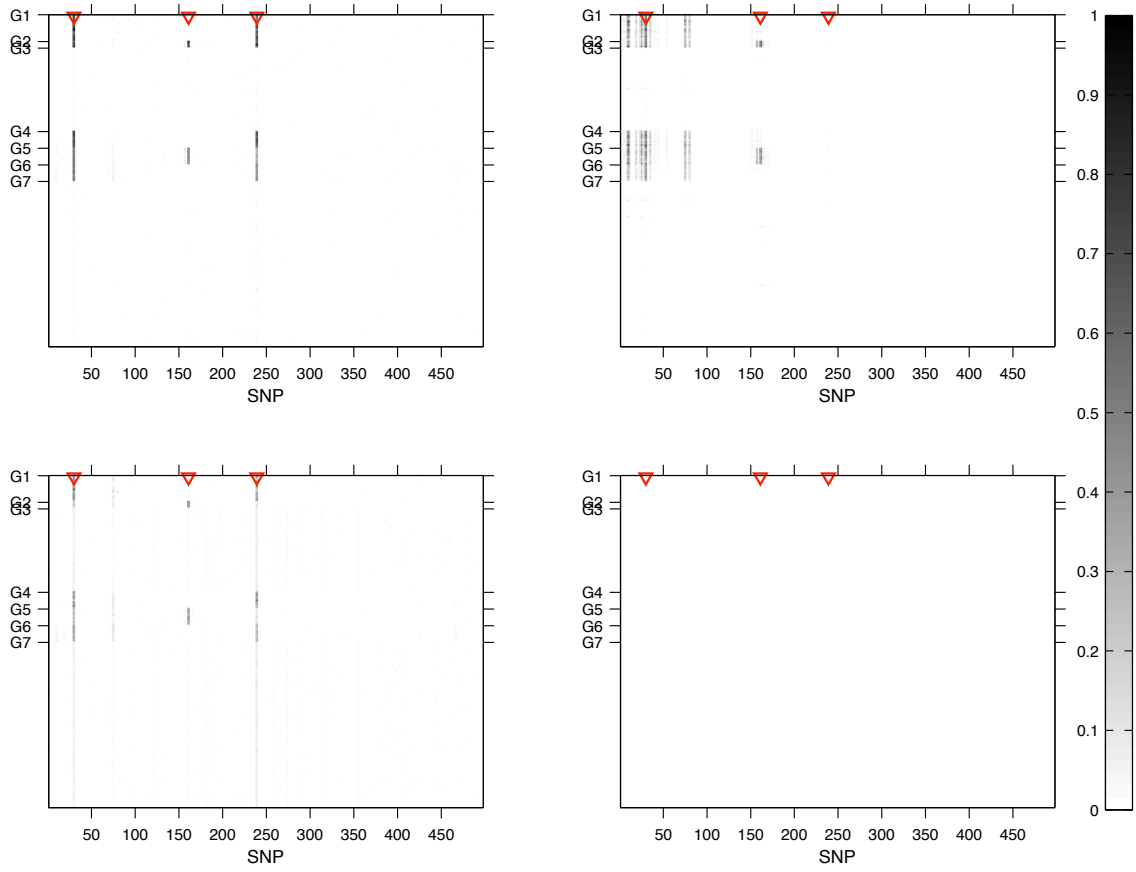

Figure S.6: Heat-map of the signals detected by each method in the fourth simulated example, SIM4, and averaged across the 25 replicates. In M-SPLS the significant (non-significant) transcript-marker association is recoded as 1 (0). From top to bottom, left to right: HESS, M-SPLS, MOM and BAYES. The symbol 'G' in the y-axis identifies groups of transcripts that are influenced by the same pattern of markers. Red triangles indicate where the hot-spots have been simulated.

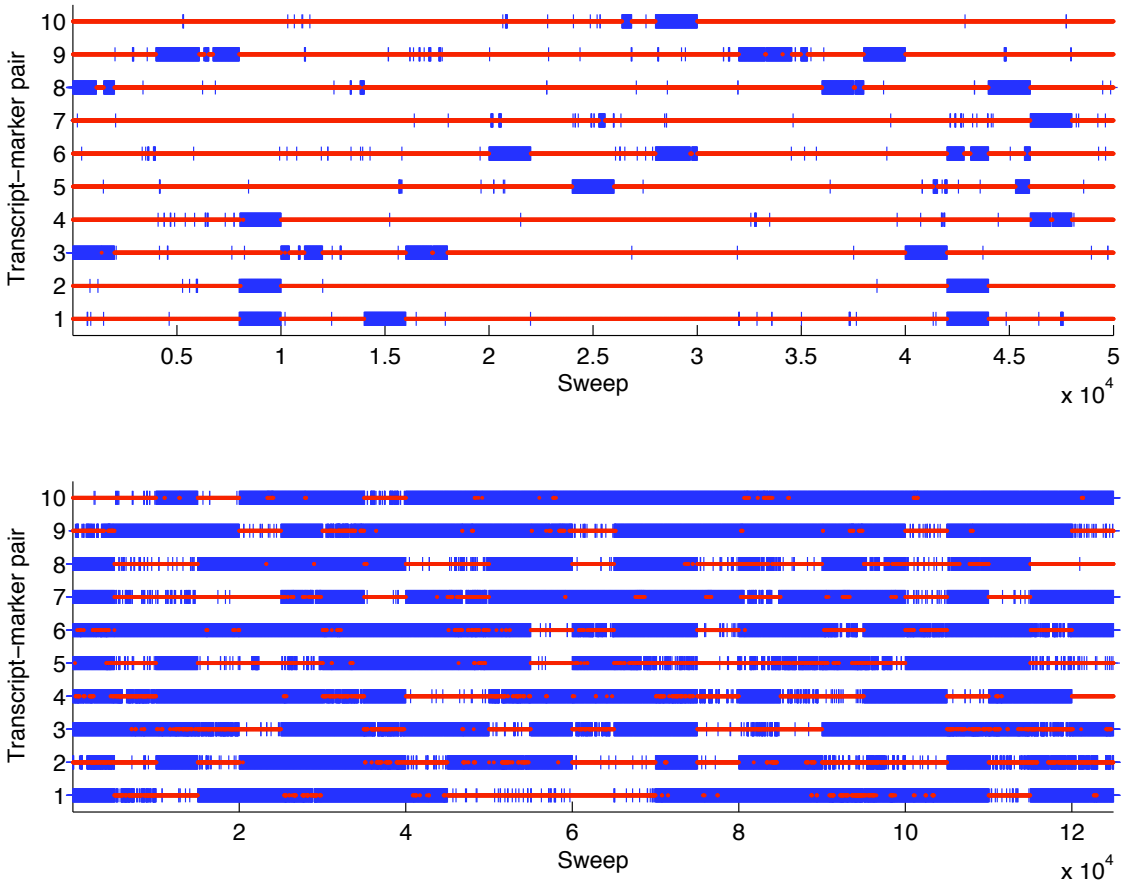

Figure S.7: Trace plot of the latent binary values obtained from BAYES (top) and HESS (bottom) in SIM1 for the 10 true positive associations simulated in the third hot-spot ($j = 225, k = 91, \ldots, 100$). For the 25 replicates, the output (γ_{kj}) of each algorithm was piled up giving rise to a vector of 50,000 (2,000 $\times$ 25) and 125,000 (5,000 $\times$ 25) sweeps, respectively. Red dot and blue cross indicate $\gamma_{kj} = 0$ and $\gamma_{kj} = 1$, respectively. HESS correctly identifies the 10 transcript-marker associations as indicated by a large majority of blue crosses. Good MCMC mixing is clear from the sequence of blue crosses interrupted by red dots and *vice versa*. On the contrary, BAYES misses the simulated associations (false negative) and gets stuck in $\gamma_{kj} = 0$ producing long stripes of consecutive red dots. Overall, the different efficiency in the MCMC mixing between BAYES and HESS is apparent from the diverse coloured stripe patterns.

Tables S1-S3

Tables S1-S3 are available for download at
http://www.genetics.org/cgi/content/full/genetics.111.131425/DC1 as Excel files.