The Annals of Applied Statistics
2020, Vol. 14, No. 2, 905–928
https://doi.org/10.1214/20-AOAS1332

A GLOBAL-LOCAL APPROACH FOR DETECTING HOTSPOTS IN MULTIPLE-RESPONSE REGRESSION

BY HÉLÈNE RUFFIEUX[1], ANTHONY C. DAVISON[2], JÖRG HAGER[3], JAMIE INSHAW[4], BENJAMIN P. FAIRFAX[5], SYLVIA RICHARDSON[1,6] AND LEONARDO BOTTOLO[7,6,1]

[1]*MRC Biostatistics Unit, University of Cambridge, helene.ruffieux@mrc-bsu.cam.ac.uk*
[2]*Ecole Polytechnique Fédérale de Lausanne (EPFL), anthony.davison@epfl.ch*
[3]*Nestlé Research, EPFL Innovation Park, jorg.hager@rd.nestle.com*
[4]*Wellcome Centre for Human Genetics, Oxford, University of Oxford, jinshaw@well.ox.ac.uk*
[5]*Department of Oncology, MRC Weatherall Institute for Molecular Medicine, University of Oxford, benjamin.fairfax@oncology.ox.ac.uk*
[6]*Alan Turing Institute, sylvia.richardson@mrc-bsu.cam.ac.uk*
[7]*Department of Medical Genetics, University of Cambridge, lb664@cam.ac.uk*

We tackle modelling and inference for variable selection in regression problems with many predictors and many responses. We focus on detecting *hotspots*, that is, predictors associated with several responses. Such a task is critical in statistical genetics, as hotspot genetic variants shape the architecture of the genome by controlling the expression of many genes and may initiate decisive functional mechanisms underlying disease endpoints. Existing hierarchical regression approaches designed to model hotspots suffer from two limitations: their discrimination of hotspots is sensitive to the choice of top-level scale parameters for the propensity of predictors to be hotspots, and they do not scale to large predictor and response vectors, for example, of dimensions 10^3–10^5 in genetic applications. We address these shortcomings by introducing a flexible hierarchical regression framework that is tailored to the detection of hotspots and scalable to the above dimensions. Our proposal implements a fully Bayesian model for hotspots based on the horseshoe shrinkage prior. Its global-local formulation shrinks noise globally and, hence, accommodates the highly sparse nature of genetic analyses while being robust to individual signals, thus leaving the effects of hotspots unshrunk. Inference is carried out using a fast variational algorithm coupled with a novel simulated annealing procedure that allows efficient exploration of multimodal distributions.

1. Introduction. Understanding the genetic architecture of complex human traits is crucial for predicting health risks and developing effective therapies. Over the past two decades thousands of genome-wide association studies have assessed the effects of millions of genetic variants on disease susceptibility. Among other important findings, these studies have revealed that most of the genetic variants involved in associations lie in noncoding regions of the genome (Ward and Kellis (2012), Tak and Farnham (2015)) which renders their functional interpretation difficult and suggests studying how they may affect clinical traits through changes in gene regulation. This observation stimulated much of the current focus in statistical genetics on expression quantitative trait locus (eQTL) analyses which assess how genetic variants control intermediate gene expression phenotypes. Genetic variants can act locally, affecting the expression of a nearby gene (*cis*-eQTL), or they can alter expression of remote transcripts (*trans*-eQTL). Understanding by which mechanisms *trans* regulation can

Received October 2018; revised February 2020.
Key words and phrases. Annealed variational inference, hierarchical model, horseshoe prior, molecular quantitative trait locus analyses, multiplicity control, normal scale mixture, regulation hotspot, shrinkage, statistical genetics, variable selection.

take place, via a local *cis* gene that acts on a whole network or via other means, is a subject of active debate (Westra et al. (2013), Solovieff et al. (2013), Brynedal et al. (2017), Yao et al. (2017)). In particular, the detection of *pleiotropic* variants, regulating the expression of tens or possibly hundreds of transcripts, is of great interest: such "*trans*-hotspot" genetic variants may provide insight into the regulatory landscape of the transcriptome and, hence, into the mechanisms shaping the evolution of the human genome. They may also shed light on important functional processes underlying clinical traits and diseases.

Despite these promises, the locations and abundance of master regulatory sites on the genome remain largely unknown. Indeed, most eQTL studies rely on conventional univariate screening, such as provided by MatrixEQTL (Shabalin (2012)), and have focused on detecting proximal *cis* associations, either to limit the multiple testing burden or because the distal *trans* associations uncovered would fail to replicate. Existing joint modelling approaches that directly model the response covariance (e.g., Yin and Li (2011), Bhadra and Mallick (2013)) only provide partial solutions to modelling hotspots. For computational reasons they are typically limited to the analysis of a few clinical phenotypes or require drastic preliminary dimension reduction that often dilutes or even discards weak but relevant signals.

The present paper aims to provide an effective statistical tool to bridge this gap: it describes a joint modelling framework that is tailored to the detection of *trans*-regulatory hotspots while scaling to tens of thousands of molecular expression levels. The model consists of a series of sparse regressions linked in a hierarchical manner, which allows the borrowing of strength across all responses (molecular expression levels) and candidate predictors (genetic variants), a key benefit of the Bayesian hierarchical framework adopted. It provides information beyond pairwise associations of predictors and responses and yields interpretable posterior measures of the propensity of predictors to be hotspots. These modelling features were introduced and discussed in Richardson, Bottolo and Rosenthal (2011), Bottolo et al. (2011) and Ruffieux et al. (2017), wherein the gain in statistical power over certain existing approaches was demonstrated.

This work focuses on realistic molecular quantitative trait locus settings, where a very large number of responses is analysed. It characterizes a parameter sensitivity issue, which was not highlighted in previous work and can be especially damaging for large response dimensions, and develops a robust solution based on a second-stage continuous shrinkage model that allows automatic discrimination of hotspots. The sensitivity concerns the specification of hyperparameters for a top-level variance parameter controlling hotspot propensity. Specifying variance components in hierarchical models is often difficult. Gelman (2006) discusses the relevance of several noninformative and weakly informative priors on random effect variances. In *large n* settings the Bernstein–von Mises theorem suggests that the choice of prior may be unimportant in practice, but in high-dimensional settings priors may have a strong impact on inferences. When the variance is close to zero, which is the case in sparse scenarios such as molecular QTL studies, Gelman cautions that badly chosen priors may severely distort posterior inferences. This observation is at the heart of work on scale-mixture priors such as the Strawderman–Berger prior (Strawderman (1971), Berger (1980)), the Student-t prior (Gelman et al. (2008)) or the horseshoe prior (Carvalho, Polson and Scott (2010)). These shrinkage priors differ in the modelling of the scale parameter, and all have substantial mass near zero in order to achieve good recovery of the overall sparsity pattern while being sufficiently heavy-tailed to capture strong signals. Fully noninformative priors (e.g., whereby the scale parameter would be assigned a Jeffreys prior) are ruled out, as they would fail to regularize. We lean on this body of work and overcome the sensitivity issue by introducing a fully Bayesian framework for hotspot detection based on the horseshoe prior. Because it entails both global and local scale parameters, our proposal flexibly adapts to the pleiotropic level and the number of responses associated with each genetic variant and robustly identifies large individual hotspot effects, whatever the overall sparsity level.

The detection of hotspots in molecular QTL studies would not be feasible without fast inference procedures, yet scalability should not be at the expense of accurate posterior exploration. This is particularly important in very high-dimensional settings, where posteriors are difficult to explore because they are highly multimodal. Building on previous work, we propose a computationally advantageous variational inference scheme for our global-local framework; the accuracy of such a scheme was validated and benchmarked against MCMC inference in Ruffieux et al. (2017). Here, we pay particular attention to problems with strongly-correlated predictors which further exacerbate multimodality. Such settings are typically encountered in genetics, as genetic variants exhibit local correlation structures along the genome. We augment the state space of our algorithm with a simulated annealing procedure which allows it to escape more easily from local modes and, thus, increases the chances of converging to the global mode (Rose, Gurewitz and Fox (1990), Ueda and Nakano (1998)).

The paper is organized as follows. Section 2 presents the dataset used throughout the paper and provides a data-driven motivation for our work. Section 3 states the problem in light of Richardson, Bottolo and Rosenthal (2011), Bottolo et al. (2011) and Ruffieux et al. (2017) and formalizes its consequences for sensitivity and multiplicity control. Section 4 presents our modelling framework and discusses its properties. Section 5 describes our annealed variational inference procedure. Section 6 assesses the performance of our approach in simulations, and Section 7 applies it to real eQTL data. Section 8 summarizes the results and gives some general discussion. Our approach is implemented in the publicly available R package `atlasqtl`.

2. Data and motivating example. We introduce an eQTL study which serves both to demonstrate the need for tailored modelling of hotspots and to illustrate the merits of our proposal throughout the paper. This study differs from most molecular QTL analyses, as it involves expression from $CD14^+$ monocytes before and after immune stimulation, performed by exposing the monocytes to the inflammation proxies interferon-γ (IFN-γ) or differing durations of lipopolysaccharide (LPS 2h or LPS 24h). The genetic variants are single nucleotide polymorphisms (SNPs) determined using Illumina arrays, and the samples were obtained from 432 healthy European individuals.

Related work (Fairfax et al. (2014), Kim et al. (2014), Lee et al. (2014)) has suggested that gene stimulation may trigger substantial *trans*-regulatory activity, creating favourable conditions for the manifestation of hotspot genetic variants. Indeed, while hotspots often exhibit associations with genes in their vicinity, they are evidenced by their capacity to influence (*trans*-act on) many remote genes. In addition to monocyte expression, we consider B-cell expression data for the same samples to contrast the hotspot activity for the two cell types.

To recall the known drawbacks of the basic univariate screening approach when used for detecting of hotspots, we regressed each unstimulated monocyte level on each genetic variant from chromosome one. This led to the following observations (Table 1 and Supplemental Material S.1, Ruffieux et al. (2020)): first, as expected, the estimated effect sizes of *trans* associations uncovered at Benjamini–Hochberg false discovery rate of 20% were substantially smaller than those of the *cis* effects. Second, although this screening uncovered about 2.5 times more *cis* associations than *trans* associations, about one-third of the former were essentially redundant: because of the local correlation structure on the genome (*linkage disequilibrium*), a single transcript was often assessed as under control by several genetic variants at the same locus, yet these genetic variants are likely to be proxies for a single causal variant. Such scenarios were much less represented among the uncovered *trans* associations, as they concerned only about 2% of them. Hence, the large number of false positive *cis* associations reported by the marginal screening is likely to have hampered the detection of, weaker, *trans* effects.

TABLE 1

Detection of cis *and* trans *associations by univariate screening using a Benjamini–Hochberg false discovery rate threshold of* 0.2. *Effects between a transcript and a SNP located less than* two *megabases* (*Mb*) *to it were defined as in* cis *effects*; *the remaining effects were defined as* trans *effects*. *Left*: *Number of detected pairwise associations*. *Middle*: *Number of detected pairwise associations after grouping those between a given transcript and several SNPs in linkage disequilibrium* (*LD*) *using* $r^2 > 0.5$ *and window size* 2 *Mb*. *Right*: *Average magnitude of regression estimates*, *and standard deviation in parentheses*

	Number	Number after LD pruning	Magnitude of estimated effects
Cis effects	1611	1049	0.11 (0.10)
Trans effects	655	641	0.04 (0.03)

There is a broad consensus about the generality of the above remarks when using marginal approaches (Gilad, Rifkin and Pritchard (2008), Mackay, Stone and Ayroles (2009), Nica and Dermitzakis (2013)). It may be tempting to view them as consequences of the multiplicity burden entailed by molecular QTL problems; false discovery rate techniques with different corrections for *cis* and *trans* effects have indeed been proposed (Peterson et al. (2016)) and may alleviate the issue. Rather than pursue this approach, we anticipate and tackle the question upfront, at the modelling stage, by building a model for hotspots that can directly borrow information across genes. Hierarchical regression models along this line exist, but none of them allow a fully Bayesian treatment of the hotspot propensities that is computationally feasible at the scale required by current eQTL studies. We now show that adequate calibration of hotspot sizes is difficult and uncertain if not properly learnt from the data.

3. Problem statement. We consider a series of hierarchically related regressions, with q centered responses, $y = (y_1, \ldots, y_q)$ and p centered candidate predictors, $X = (X_1, \ldots, X_p)$, for n samples ($n \ll p$),

$$
\begin{aligned}
&y_t \mid \beta_t, \tau_t \sim \mathcal{N}_n\big(X\beta_t, \tau_t^{-1} I_n\big), \quad t = 1, \ldots, q,\\
&\beta_{st} \mid \gamma_{st}, \tau_t, \sigma^2 \sim \gamma_{st}\mathcal{N}\big(0, \sigma^2\tau_t^{-1}\big) + (1 - \gamma_{st})\delta_0, \quad s = 1, \ldots, p,\\
&\gamma_{st} \mid \omega_{st} \sim \text{Bernoulli}(\omega_{st}),
\end{aligned} \tag{1}
$$

where δ_0 is the Dirac distribution and where τ_t and σ^{-2} are assigned Gamma priors. In the molecular QTL setting on which we focus, the predictors represent p genetic variants, typically SNPs, and the responses are q molecular expression levels for n individuals. The regression parameters, β_{st}, are specific to each pair of predictor X_s and response y_t and have spike-and-slab priors to induce sparsity (Mitchell and Beauchamp (1988), George (2000)). Hence, the binary latent variables γ_{st} take value unity in case of association and are zero otherwise. The global variance of effects, σ^2, allows information sharing across responses associated with overlapping sets of predictors. This specification will be complemented with a second-level model on the probabilities of association ω_{st} in Section 4.

Model formulation (1) and variants thereof have been employed by authors such as Jia and Xu (2007), Richardson, Bottolo and Rosenthal (2011), Bottolo et al. (2011) and Ruffieux et al. (2017). Their proposals differ primarily in the prior specification for the probability of association parameter, ω_{st}. Richardson, Bottolo and Rosenthal (2011) and Bottolo et al. (2011) decouple the predictor and response effects by setting $\omega_{st} = \omega_s \times \omega_t$ and place prior distributions on each of ω_s and ω_t, whereas Jia and Xu (2007) and Ruffieux et al. (2017) use the simpler formulation $\omega_{st} \equiv \omega_s$. A suitable specification of the predictor-specific parameter ω_s is crucial, as ω_s controls the propensity of each predictor X_s to be a hotspot, that is, to be simultaneously associated with several responses. As we now explain, the discrimination of

TABLE 2
Ratios (4) for a grid of variances σ_ω^2 and numbers of associated responses q_s. The total number of responses is $q = 20{,}000$, and the base rate is $\mu_\omega = 0.1$. The penalty varies greatly depending on the chosen value for σ_ω^2 and increases roughly linearly with q_s

σ_ω^2: q_s	5	10	50	100
10^{-4}	1.0	1.1	1.5	2.1
10^{-3}	1.4	2.0	6.5	12.2
10^{-2}	6.0	12.3	62.4	125.4

hotspots can be very sensitive to the choice of prior distribution for ω_s, and this sensitivity becomes particularly severe in very large response settings, where the detection of hotspots is a key task.

For the sake of discussion, we illustrate our point with the formulation of Ruffieux et al. (2017), whereby

$$\omega_{st} \equiv \omega_s \overset{\text{i.i.d.}}{\sim} \text{Beta}(a, b), \quad a, b > 0, s = 1, \ldots, p; \tag{2}$$

similar considerations apply to the models of Jia and Xu (2007), Bottolo et al. (2011) and Richardson, Bottolo and Rosenthal (2011). We discuss the choice of the hyperparameters a and b through the prior expectation and variance for ω_s. The expectation corresponds to the prior base rate of associated pairs, $\mu_\omega = \text{E}(\omega_s) = \text{pr}(\gamma_{st} = 1)$. Its value should be small to induce sparsity, typically $\mu_\omega \ll 1$ for $p, q \gg n$, and may be fixed using an estimate of the overall signal sparsity. In contrast, there is no prior state of knowledge about $\sigma_\omega^2 = \text{Var}(\omega_s)$, and its choice turns out to impact the prior size of hotspots when q is large. To formalize this, it is helpful to study prior odds ratios, as Scott and Berger (2010) did when discussing expected model sizes in single-response sparse regression. For a given predictor X_s, write $\gamma_s(q_s)$ the q-variate indicator vector whose first $0 < q_s \leq q$ entries are unity and the following $q - q_s$ are zero. The prior odds ratio,

$$\text{POR}(q_s - 1 : q_s) = \frac{\text{pr}\{\gamma_s(q_s - 1)\}}{\text{pr}\{\gamma_s(q_s)\}} = \frac{b + q - q_s}{a + q_s - 1}, \tag{3}$$

quantifies the penalty induced by the prior when moving from $q_s - 1$ to q_s responses associated with X_s. The penalty increases with the total number of responses in the model (for fixed a, b and q_s), but it also decreases monotonically as q_s increases, so that it is a priori easier to add a response when X_s is already associated with many responses. More insight into this phenomenon can be obtained by looking at the quantity

$$\frac{\text{POR}(0 : 1)}{\text{POR}(q_s - 1 : q_s)}, \tag{4}$$

which compares the cost of adding a further response association with X_s when moving from the null model or from a model with $q_s - 1$ associations already.

In molecular QTL problems, q_s is typically much smaller than q, as each SNP is believed to control just a few molecular entities. For $q_s \ll q$, (4) behaves roughly linearly in q_s with slope $\approx a^{-1} = \sigma_\omega^2\{\mu_\omega^2(1 - \mu_\omega) - \mu_\omega\sigma_\omega^2\}^{-1}$. Hence, large σ_ω^2 favours large hotspots while small σ_ω^2 tends to give an association pattern that is more scattered across predictors. In the latter case, strong shrinkage towards $\mu_\omega \ll 1$ may be induced and the resulting hotspot sizes may be underestimated, whereas, in the former case artifactual hotspots may appear when

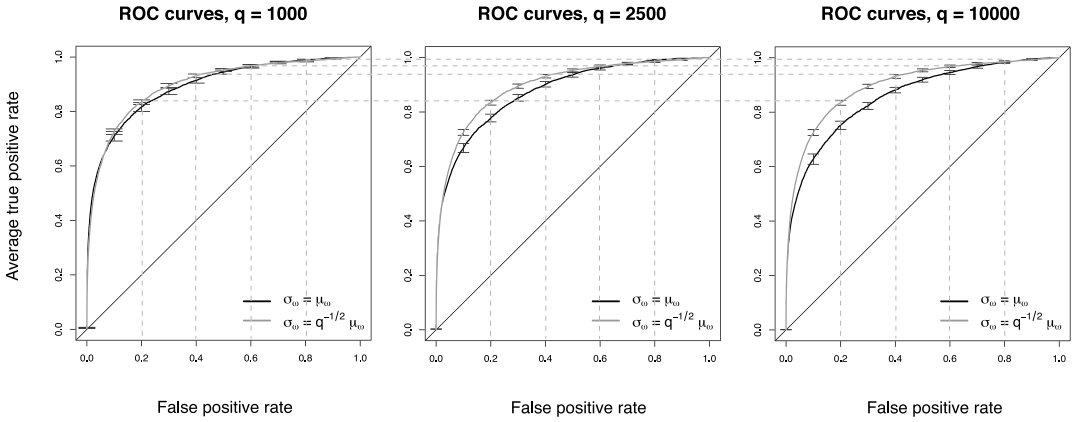

FIG. 1. *Variable selection performance with and without multiplicity adjustment, measured by average receiver operating characteristic (ROC) curves with* 95% *confidence intervals obtained from* 100 *replicates. Three problems are simulated, with an increasing number of response variables,* $q = 1000$ *(left),* $q = 2500$ *(middle),* $q = 10{,}000$ *(right) and* $p = 100$ *candidate predictors for* $n = 100$ *samples. The pattern of associations is the same for all three scenarios*: 50 *responses are chosen randomly among the first* 1000 *responses to be associated with at least one of* 10 *predictors*; *the rest of the responses are drawn from Gaussian noise. For a given response, the proportion of its variance explained by the predictors does not exceed* 15%. *Two implementations of model* (1)–(2) *are compared*: *one uses a fixed choice of variance* $\sigma_\omega^2 = \mu_\omega^2$ *(black curves); its performance deteriorates as* q *increases, from left to right. The other uses the proposed adjustment for the total number of responses* q, *that is,* $\sigma_\omega^2 = q^{-1}\mu_\omega^2$ *(grey curves); its performance remains unchanged as* q *increases (see grid). The base rate is fixed to the simulated proportion of associated predictors, that is,* $\mu_\omega = 0.1$.

data are insufficiently informative to dominate the prior specification. Table 2 shows that the penalties (4) can differ drastically for different choices of σ_ω^2.

To evaluate the extent to which this could impact inference in flat likelihood scenarios, it is helpful to also study the case where q_s is of order q, even though this is unlikely to be encountered in our applications. When $q_s \sim q$ (i.e., when q_s/q tends to a strictly positive constant as $q \to \infty$), (4) is of order $O(q)$, so that, in weakly informative data settings, the sensitivity may lead to the manifestation of massive spurious hotspots associated with nearly all responses. Such undesired "pile-up" effects highlight the need to adjust for the dimensionality of the response.

The sensitivity of inferences to the hotspot propensity variance relates to the well-known issue of specifying prior distributions for variance components, as ω_s can be viewed as a random effect. While this sensitivity and its related response multiplicity burden are important problems that affect any hierarchically related regression model such as (1), they have been neither formalized nor investigated in the literature. In fact, the number of responses presented in numerical experiments is usually rather small (10–1000), mainly limited by the heavy computational load of MCMC sampling, so that this sensitivity issue typically goes unnoticed. Another aspect is that "pile-up" effects can be avoided by choosing a small hotspot propensity variance at the risk of giving up substantial hotspot selection performance. The very sparse nature of molecular QTL analyses also rules out the use of simple empirical Bayes estimates which typically collapse to the degenerate case $\hat{\sigma}_\omega^2 = 0$; see, for example, Scott and Berger (2010), van de Wiel, Te Beest and Münch (2019). Thus, a tailored solution is needed.

Our proposal resolves the above issues, based on two considerations. First, we argue that "pile-up" effects can be prevented by suitably linking the hotspot propensity variance to the number of responses, in effect performing multiplicity adjustment. Indeed, choosing $\sigma_\omega^2 = O(q^{-1})$, ratio (4) is $O(1)$ when $q_s \sim q$. For small values of μ_ω, typically chosen in

sparse association problems, this adjustment amounts to enforcing small and similar numbers of response associations for all predictors, with the degree of shrinkage depending on the number of responses q (in the limiting case $q \to \infty$, we obtain $\omega_s \equiv \mu_\omega$). Figure 1 illustrates the degradation of the variable selection performance in moderately informative problems with increasing q and shows how the proposed penalty addresses the issue.

Second, we embed and *relax* this multiplicity adjustment in a fully Bayesian framework involving a second-stage model on the probability of association, ω_{st}, and, hence, infer the hotspot propensity variances from the data in a fully automatic way with no ad hoc choice or compromise that would bias the hotspot sizes.

4. Global-local modelling framework.

4.1. *Second-stage probit model on the probability of association.* As a first step in detailing our proposal, we complement (1) with a hierarchical probit model on the probability of association, that is,

$$\omega_{st} = \Phi(\theta_s + \zeta_t), \quad \zeta_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(n_0, t_0^2), \quad s = 1, \dots, p, t = 1, \dots q, \tag{5}$$

where $\Phi(\cdot)$ is the standard normal cumulative distribution function and where we assume, for now, that $\theta_s \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, s_0^2)$. This second-stage model offers a flexible and interpretable representation of the association probability in multiresponse settings: it involves a response-specific parameter, ζ_t, which adapts to the sparsity pattern corresponding to each response, and a propensity parameter, θ_s, which encodes predictor-specific modulations of the probability of association, as in Richardson, Bottolo and Rosenthal (2011) and Bottolo et al. (2011). The hyperparameters n_0 and t_0^2 are set to match a selected expectation and variance for the prior number of associated predictors per response (see Supplementary Material S.2, Ruffieux et al. (2020)). The variance parameter s_0^2 essentially plays the role of σ_ω^2, presented in Section 3, in influencing the prior odds ratios; in particular, an application of the delta method shows that if $s_0^2 \sim O(q^{-1})$ as $q \to \infty$, then $\text{Var}\{\Phi(\theta_s)\} \sim O(q^{-1})$. While no closed form can be obtained for prior odds ratios (3) based on model (5), numerical experiments suggest that (4) indeed behaves independently of q when $s_0^2 = q^{-1}$, for $q_s \approx q$ large. Formulation (5) sets the stage for introducing our new multiplicity-adjusted hotspot model which combines the benefits of both global and local control and adaptation.

4.2. *Horseshoe prior on hotspot propensities.* Our proposed specification for the hotspot propensity adds flexibility in modelling the scale of θ_s in (5) by letting

$$\theta_s \mid \lambda_s, \sigma_0 \sim \mathcal{N}(0, \sigma_0^2 \lambda_s^2), \quad \lambda_s \overset{\text{i.i.d.}}{\sim} \text{C}^+(0, 1), \quad s = 1, \dots, p, \tag{6}$$

where $\text{C}^+(\cdot, \cdot)$ is a half-Cauchy distribution. This corresponds to placing a horseshoe prior (Carvalho, Polson and Scott (2010)) on the hotspot propensities, $\theta_s \mid \sigma_0 \overset{\text{i.i.d.}}{\sim} \text{HS}(0, \sigma_0)$. The global scale σ_0 adapts to the overall sparsity pattern, while the Cauchy tails of the predictor-specific scale parameters λ_s flexibly capture the hotspot effects.

The horseshoe prior is a popular example of absolutely continuous shrinkage priors, with newly established theoretical guarantees, such as near-minimaxity in estimation (van der Pas, Szabó and van der Vaart (2017)). It also belongs to the class of global-local shrinkage priors that have an infinite spike at the origin and regularly-varying tails (Polson and Scott (2011), Bhadra et al. (2016)).

4.3. *Multiplicity-adjusted shrinkage profile.* While the local scale parameters λ_s are essential to suitably detect the few large signals, the choice of the global scale σ_0 is no less important, as σ_0 controls the ability of the model to discriminate signal from noise. Piironen and Vehtari (2017) propose to choose σ_0 based on specific sparsity assumptions; we extend their considerations to our multiresponse setting and further highlight how the dimension of the response needs to be accounted for in order to recover the beneficial shrinkage properties conferred by the horseshoe prior when used in the classical normal means model. For a given predictor X_s, we reparametrize the probit link formulation,

$$\gamma_{st} \mid \theta_s, \zeta_t \sim \text{Bernoulli}\big\{\Phi(\theta_s + \zeta_t)\big\}, \quad t = 1, \ldots, q,$$

by introducing a q-variate auxiliary variable $z_s = (z_{s1}, \ldots, z_{sq})$, as

$$\gamma_{st} = \mathbb{1}\{z_{st} > 0\}, \quad z_{st} \mid \theta_s, \zeta_t \sim \mathcal{N}(\theta_s + \zeta_t, 1), \quad t = 1, \ldots, q. \tag{7}$$

In this second-stage probit model, z_{st} can be understood as data and θ as a sparse parameter. Given the hyperparameters n_0 and t_0^2 for ζ_t, we have

$$z_{st} \mid \theta_s \sim \mathcal{N}\big(n_0 + \theta_s, 1 + t_0^2\big),$$

so that

$$\mathrm{E}(\theta_s \mid z_s, \sigma_0, \lambda_s) = (1 - \kappa_s)\frac{1}{q}\sum_{t=1}^{q}(z_{st} - n_0) + \kappa_s \times 0 = (1 - \kappa_s)\bar{z}_s',$$

where $\bar{z}_s' = \bar{z}_s - n_0$ and

$$\kappa_s = \frac{1}{1 + \alpha(\sigma_0)\lambda_s^2}$$

is the *shrinkage factor* for hotspot propensities, with $\alpha(\sigma_0) = q(1 + t_0^2)^{-1}\sigma_0^2$ (Lemma S.3.1 of Supplemental Material S.3, Ruffieux et al. (2020)).

In the horseshoe prior literature with half-Cauchy priors on the local scales as well as unit global scale and error variance, this factor has a Beta$(1/2, 1/2)$ prior whose shape resembles a horseshoe, hence, the name. As this prior density is unbounded at 0 and 1, one expects a priori, either large effects, with κ_s close to zero, or no effects, with κ_s close to one. In our case it can be shown that

$$p(\kappa_s \mid \sigma_0) = \pi^{-1}\alpha(\sigma_0)^{1/2}\kappa_s^{-1/2}(1 - \kappa_s)^{-1/2}\big[1 + \kappa_s\{\alpha(\sigma_0) - 1\}\big]^{-1}, \quad 0 < \kappa_s < 1,$$

using $\lambda_s \overset{\text{i.i.d.}}{\sim} \mathrm{C}^+(0, 1)$; this prior density reduces to Beta$(1/2, 1/2)$ when $\alpha(\sigma_0) = 1$, that is, when $\sigma_0^2 \approx q^{-1}$, as $t_0^2 \ll 1$ under sparse assumptions (Lemma S.3.2 and Figure S.4 of Supplemental Material S.3, Ruffieux et al. (2020)). This formulation therefore enjoys the shrinkage properties of the horseshoe prior. Critically, using a default choice of $\sigma_0^2 = O(1)$ as $q \to \infty$ would yield $\mathrm{E}(\kappa_s \mid \sigma_0) \approx 0$ for q large, so that, on average, θ_s would be unregularized given z_s. These two choices can be read in light of the discussion in Section 3: the latter mirrors the absence of any correction for the dimensionality of the response, possibly creating spurious "pile-up" effects, whereas the former satisfies the multiplicity adjustment condition with the proposed scaling factor q^{-1} for σ_0^2.

Fixing $\sigma_0^2 = q^{-1}$ would stop the global scale from adapting to the degree of signal sparsity. We instead place a hyperprior on σ_0 which embeds the penalty. Following Carvalho, Polson and Scott (2010), we choose a half-Cauchy prior,

$$\sigma_0 \sim \mathrm{C}^+\big(0, q^{-1/2}\big). \tag{8}$$

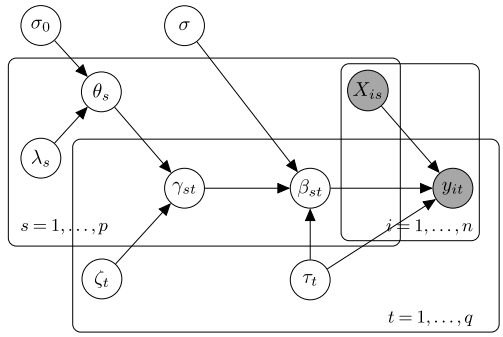

FIG. 2. *Graphical representation of model* (10). *The shaded nodes are observed, the others are inferred*; β_{st} *is the regression coefficient for association between predictor* X_s (*SNP*) *and response* y_t (*expression level*), *and* γ_{st} *is the latent binary indicator for the presence or absence of this effect. The probability of association is decoupled into response-specific,* ζ_t, *and predictor-specific,* θ_s, *contributions. The latter entails the global-local second-stage model for hotspots.*

An equivalent parametrization of (6) and (8) is

$$\theta_s \mid \lambda_s, \sigma_0 \sim \mathcal{N}\big(0, q^{-1}\lambda_s^2\sigma_0^2\big), \quad \lambda_s \overset{\text{i.i.d.}}{\sim} \mathrm{C}^+(0,1), \quad \sigma_0 \sim \mathrm{C}^+(0,1), \tag{9}$$

from which one clearly sees how the multiplicity factor rescales the hotspot propensity variance. For clarity, we gather the complete specification of our global-local hierarchical model; it combines (1) and the decomposition of the probability parameter (5) with (6) and (8),

$$\begin{aligned}
&y_t \mid \beta_t, \tau_t \sim \mathcal{N}_n\big(X\beta_t, \tau_t^{-1} I_n\big), \quad t = 1, \ldots, q,\\
&\beta_{st} \mid \gamma_{st}, \tau_t, \sigma \sim \gamma_{st}\mathcal{N}\big(0, \sigma^2\tau_t^{-1}\big) + (1-\gamma_{st})\delta_0, \quad s = 1, \ldots, p,\\
&\gamma_{st} \mid \theta_s, \zeta_t \sim \text{Bernoulli}\big\{\Phi(\theta_s + \zeta_t)\big\}, \quad \zeta_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}\big(n_0, t_0^2\big),\\
&\theta_s \mid \lambda_s, \sigma_0 \sim \mathcal{N}\big(0, \lambda_s^2\sigma_0^2\big), \quad \lambda_s \overset{\text{i.i.d.}}{\sim} \mathrm{C}^+(0,1), \quad \sigma_0 \sim \mathrm{C}^+\big(0, q^{-1/2}\big),
\end{aligned} \tag{10}$$

with Gamma prior distributions for τ_t and σ^{-2}; a graphical representation is provided in Figure 2.

5. Annealed variational inference. Joint inference on molecular QTL models is particularly difficult, a serious complication being the high dimensionality of the predictor and the response spaces. In our proposal, as well as in those of Jia and Xu (2007), Richardson, Bottolo and Rosenthal (2011), Bottolo et al. (2011) and Ruffieux et al. (2017), the binary latent matrix $\Gamma = \{\gamma_{st}\}$ creates a discrete search space of dimension $2^{p\times q}$, $p, q \gg n$, and the quality of inferences hinges on successful exploration of this space. Several sampling schemes have been proposed for spike-and-slab models. Most of them involve drawing each latent component from its marginal posterior distribution and, therefore, require costly evaluations of marginal likelihoods at each iteration. Mixing problems also arise, mainly caused by the difficulty that the sampler has in jumping between the states defined by the spike and the slab components. The resulting sample autocorrelations are high, so many iterations are usually needed to collect enough independent samples.

There are two paths towards scaling up Bayesian inference. The first is to design more efficient Markov chain Monte Carlo algorithms. While there is a growing literature on the *large n* case with proposals involving partition-based parallelization (e.g., Wang and Dunson (2013)) or data subsampling (e.g., Bardenet, Doucet and Holmes (2014)), research is limited for the high-dimensional case, apart from work on approximating transition kernels (O'Brien and

Dunson (2004), Bhattacharya and Dunson (2010), Guhaniyogi, Qamar and Dunson (2018)), so effectively scaling MCMC methods for dimensions such as those involved in molecular QTL analyses is still largely out of reach. Therefore, the second path investigates deterministic alternatives to sampling-based approaches. These include expectation-maximization algorithms, expectation-propagation inference and variational inference. Effectively implemented, these approaches require only reasonable computing resources. A legitimate concern, however, is whether fast deterministic inference can be sufficiently accurate for variable selection in genome-wide association studies. In the case of variational inference, Carbonetto and Stephens (2012) and Ruffieux et al. (2017) provide positive evidence for accurate posterior exploration, notably through extensive comparisons with MCMC inference. Here, we build on this previous work and develop an efficient variational inference scheme for our global-local modelling framework. We further improve posterior exploration by coupling our algorithm with a simulated annealing procedure (Rose, Gurewitz and Fox (1990), Ueda and Nakano (1998)).

Let v be the parameter vector of interest. Variational posterior approximations are obtained by considering a tractable analytical approximation, $q(v)$, to the true posterior distribution, $p(v \mid y)$. The *mean-field* approximation (Opper and Saad (2001)) assumes that $q(v)$ factorizes over some partition of v, $\{v_j\}_{j=1,\dots,J}$, that is,

$$q(v) = \prod_{j=1}^{J} q(v_j), \tag{11}$$

with no assumption on the functional forms of the $q(v_j)$. One then performs inference by maximizing the following lower bound on the marginal log-likelihood,

$$\mathcal{L}(q) = \int q(v) \log\left\{\frac{p(y, v)}{q(v)}\right\} \mathrm{d}v, \tag{12}$$

which is a tractable alternative to minimizing the Kullback–Leibler divergence (see, e.g., Blei, Kucukelbir and McAuliffe (2017)),

$$\mathrm{KL}(q \parallel p) = -\int q(v) \log\left\{\frac{p(v \mid y)}{q(v)}\right\} \mathrm{d}v. \tag{13}$$

Quantity (12) is often called ELBO, for *evidence lower bound*, in the machine learning literature.

With each v_j modelled as independent a posteriori of the other parameters given the observations and the hyperparameters, mean-field variational inferences (11) trade off posterior dependence assumptions and computational complexity. For our model, independence assumptions between β_{st} and γ_{st} would be particularly problematic: they would make $q(\beta_t)$ a unimodal representation of the marginal distribution $p(\beta_t \mid y)$ and, thus, a poor proxy for the highly multimodal posterior distribution implied by the spike-and-slab prior on β_t. Considering model reparametrization (7), we instead employ a structured factorization, whereby we model β_{st}, γ_{st} and z_{st} jointly, that is, for each fixed $t \in \{1, \dots, q\}$, we seek a variational distribution of the form

$$\prod_{s=1}^{p} q(\beta_{st}, \gamma_{st}, z_{st}). \tag{14}$$

This structured factorization induces point mass mixture factors and, hence, retains the multimodal behaviour of the spike-and-slab distribution. It is also a faithful representation of the true posterior distribution when predictors are only weakly dependent, since the latter factorizes as (14) when using an orthogonal design matrix, as pointed out by Carbonetto and Stephens (2012).

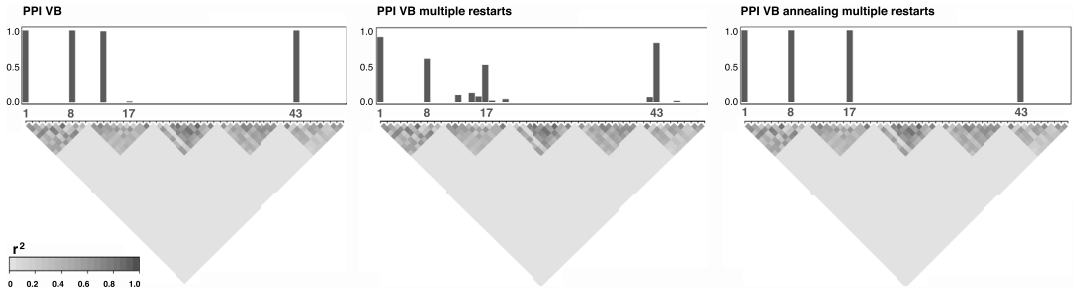

FIG. 3. *Variable selection under high multicollinearity. Problem with a single response and* 1000 *SNPs autocorrelated by blocks as candidate predictors* (*first* 50 *shown*). *The SNPs simulated as associated with the response explain* 30% *of its variance*; *their positions are marked by the numeric labels. The bars show the posterior probabilities produced by the variational algorithm*, *one run* (*left*) *and average of* 1000 *runs with different starting values* (*middle*), *and by the annealed variational algorithm with initial temperature* $T = 5$ *and grid of* 100 *temperatures*, *average of* 1000 *runs with different starting values* (*right*).

Other fast deterministic inference procedures based on expectation-maximization algorithms have been proposed for spike-and-slab models (Ročková and George (2014)). While these are limited to producing point estimates, variational procedures infer full approximating distributions and, thus, also estimate posterior parameter uncertainty. Variational inference is frequently decried as underestimating posterior variances; however, as a result of both the mean-field independence assumptions (Wainwright and Jordan (2008)) and the optimization of a *reverse* Kullback–Leibler divergence, $p(\cdot)$ and $q(\cdot)$ are swapped in (13) compared to the standard Kullback–Leibler divergence. As the variational objective function (12) is not concave, underestimated variances may also affect the ability of the algorithm to retain relevant variables. Indeed, in highly multimodal settings, such as those induced by strongly correlated predictors, the approximation tends to concentrate mass on a local mode corresponding to a single configuration of variables in groups of correlated predictors. Figure 3 considers a problem with highly correlated predictors, where the vanilla variational algorithm completely misses one active SNP and instead picks one of its correlated neighbours with high confidence. Figure 3 also suggests that averaging posterior probabilities across multiple runs with different starting values can produce "diluted" posterior summaries which better reflect the uncertainty of SNP selection in regions of high linkage disequilibrium (i.e., with strong local correlation). But although averaging mitigates the problem, it results in increased computational costs, becoming quite substantial for typical molecular QTL problems.

Simulated annealing directly targets the improved exploration of multimodal posteriors. It introduces a so-called *temperature* parameter which indexes a series of *heated* distributions and controls the degree of separation of their modes. The procedure starts with large temperatures that flatten the distribution of interest, thereby sweeping most of its local modes away and facilitating the search for the global optimum. Temperatures are then progressively decreased until the *cold* distribution, corresponding to the original multimodal distribution, is reached.

Optimization via simulated annealing was first described in Metropolis et al. (1953) and Kirkpatrick, Gelatt and Vecchi (1983) for Metropolis algorithms and was then adapted for expectation-maximization by Ueda and Nakano (1998) and for variational inference by Katahira, Watanabe and Okada (2008). Variational inference lends itself to simulated annealing principles. Indeed, the objective function (12) can be rewritten as the sum of two terms: the expected value of the log joint distribution, which encourages the approximation to put mass on configurations of the variables that best explain the data, and an entropy term, which prefers the approximation to be more dispersed. Annealing inflates the entropy by

multiplying it by the temperature parameter,

$$\mathcal{L}_T(q) = \int q(v) \log p(y, v)\, \mathrm{d}v - T \int q(v) \log q(v)\, \mathrm{d}v, \quad T \geq 1;$$

it penalizes the first term (when $T > 1$) and gradually relaxes this penalty until the original variational algorithm is obtained (when $T = 1$).

There is no consensus on the type of temperature schedule to use. We follow Kirkpatrick, Gelatt and Vecchi (1983) in their choice of geometric schedule and use the specific implementation of Gramacy, Samworth and King (2010),

$$T_j = (1 + \Delta)^{j-1}, \quad \Delta = T_J^{1/(J-1)} - 1, \quad j = J, \ldots, 1,$$

where T_J is the hottest temperature. Our experiments suggest that initial temperatures between 2 and 5 and grids of 10 to 100 temperatures, depending on the computational resources at hand, are sufficient for good exploration. A final purpose of Figure 3 is to illustrate the benefits of annealed variational inference over classical variational inference: while selection based on the former may suffer from poorly chosen starting values, selection based on the latter consistently identifies the relevant SNPs across all 1000 restarts.

In *large n* regimes the scalability of variational algorithms can often greatly benefit from data subsampling, which may be implemented generically in stochastic gradient ascent schemes; this is not the case in high dimensions. In this latter regime we believe that tailored, model-specific derivations aiming for closed-form updates are important. Taking advantage of the conditional conjugacy properties of our model and resorting to suitable reparametrizations, we obtained all the variational updates analytically, albeit using special functions, such as the incomplete gamma and exponential integral functions. In particular, obtaining closed-form updates for the horseshoe's half-Cauchy scale parameters hinged on introducing auxiliary variables (see, e.g., Neville, Ormerod and Wand (2014)) to arrive at variational distributions in the Gamma family or involving cheap-to-compute special functions, and this was somewhat complicated by the annealing. The full derivation of the annealed variational updates is in the Supplementary Material S.4 (Ruffieux et al. (2020)).

We then implemented a block coordinate ascent optimization procedure, where the variational parameters are updated in turn and by blocks for all the responses, exploiting the concavity of $\mathcal{L}(q)$ in each of these blocks. This scheme combined with the rapidly computable updates produces a highly effective algorithm. The algorithm returns the variational parameters after convergence, some of which can be directly employed to perform variable selection, for example, the variational posterior probability of association for each pair, $\mathrm{E}_q(\gamma_{st})$, the variational posterior mean of the corresponding regression coefficient, $\mathrm{E}_q(\beta_{st})$ and the variational posterior means of the hotspot propensities, $\mathrm{E}_q(\theta_s)$, where $\mathrm{E}_q(\cdot)$ is the expectation with respect to the variational posterior distribution $q(\cdot)$ (see Supplementary Material S.4, Ruffieux et al. (2020)).

6. Simulations.

6.1. *Data generation for pleiotropic QTL problems.* The numerical experiments presented below are meant to closely reproduce real genetic data scenarios demonstrating pleiotropy, that is, the control of several outcomes by a single SNP, for which our method is primarily designed. They also broadly illustrate the characteristic features of the method when applied to association studies with a large number of correlated responses. Simulated data mimic molecular QTL data based on general principles of statistical genetics. We either extract SNPs from real datasets (for Sections 6.4 and 6.5) or simulate them as in Hardy–Weinberg equilibrium and autocorrelated by blocks. In the latter case, we form the blocks

using realisations from multivariate Gaussian latent variables of dimension 50 and with autocorrelation coefficient drawn uniformly at random in a preselected interval; for simulations of Sections 6.2 and 6.3, we use $(0.75, 0.95)$. We then use a quantile thresholding rule to code the number of minor alleles as 0, 1 or 2 according to a SNP-specific minor allele frequency drawn from a uniform distribution, $\text{Unif}(0.05, 0.5)$. We also generate block-dependent responses using multivariate normal variables; the blocks consist of 10 equicorrelated responses, with residual correlation drawn from the interval $(0, 0.25)$ for simulations of Sections 6.2 and 6.3, and from the interval $(0, 0.5)$ for simulations of Sections 6.4 and 6.5.

The pleiotropic association pattern is constructed as follows. To model large and functionally inert genomic regions, we partition the SNPs into N chunks of size 200 and leave $\lfloor N/2 \rfloor$ chunks with no associations. From the remaining chunks we randomly select labels for the "active" predictors, namely, SNPs associated with at least one response. Similarly, we select labels for the "active" responses, namely, responses associated with at least one SNP. We then randomly associate each active predictor with one active response and each active response with one active predictor. For each active SNP s, we draw a "propensity" parameter ω_s from a $\text{Beta}(1, 5)$ distribution, and further associate the SNP with other active responses whose labels are sampled with probability ω_s; these SNP-specific propensities $\{\omega_s\}$ therefore create hotspots of different "sizes". We effectively generate the associations under an additive dose-effect scheme, whereby each copy of the minor allele results in a uniform and linear increase in risk, and we draw the proportion of a response's variance explained by individual SNPs from a left-skewed Beta distribution to favour the generation of smaller effects. We then rescale these proportions so that the proportion of response variance attributable to genetic variants does not exceed a certain value; the magnitude of SNP effects derives from this value, and the sign of the effects is altered with probability 0.5. These choices imply an inverse relationship between minor allele frequencies and effect sizes, as expected under natural selection (selection against SNPs with large penetrance is stronger; see, e.g., Park et al. (2011)). For a given experiment, we keep the same association pattern across all replicates, but we regenerate the SNPs (if not real), effect sizes and responses for each replicate. The remaining settings (e.g., numbers of variables and of samples, proportion of response variance explained by the SNPs) vary, so will be detailed in the text corresponding to each experiment. Data-generation functions are implemented in the publicly available R package `echoseq`.

6.2. *Variable selection performance with global-local modelling.* In this section we evaluate the performance of our proposal for discriminating hotspots and selecting pairs of associated predictor and response variables. We simulated a "reference" data scenario with hotspots associated with approximately 35 responses on average and whose cumulated effect sizes are responsible for at most 25% of the variability of each response. We also generated four variants of this scenario: with smaller or larger hotspots (average sizes ≈ 17 and 85, respectively), and with weaker or stronger effects (response variance explained by hotspots below 20% and 30%, respectively). Each problem involves $p = 1000$ SNPs and $q = 20{,}000$ responses (which corresponds the estimated number of protein-coding genes in humans), for $n = 300$ samples. We simulated 20 hotspots, and, depending on the hotspot size scenario, 100, 200 or 500 responses had at least one association.

We benchmarked our global-local model (10) against four alternatives. The first three are based on the proposal (1)–(2) of Ruffieux et al. (2017) with three choices of hotspot propensity variance, σ^2_ω. These choices were made without assuming any prior state of knowledge, as would be faced in real data situations: we set the base rate of associated pairs to $\mu_\omega = 0.002$, so that two predictors are a priori associated with each response, on average. Then, for each model we set the hotspot propensity scale to a different fraction of this base rate. The fourth

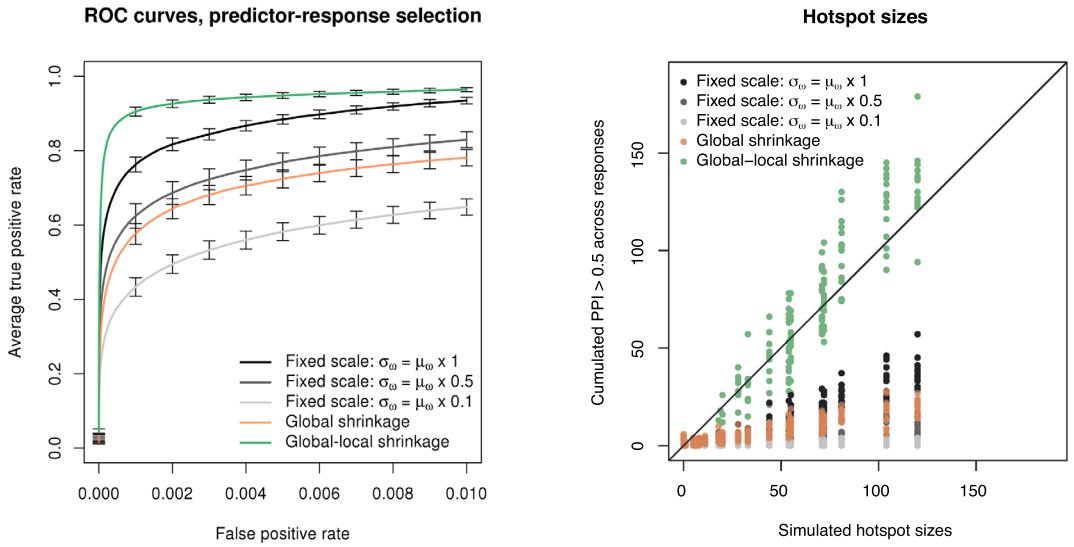

FIG. 4. *Performance of five hotspot modelling approaches, for the "reference" data generation case. Left: truncated average ROC curves for predictor-response selection with* 95% *confidence intervals obtained from* 64 *replicates. Right: sizes of recovered hotspots based on the median probability model rule (Barbieri and Berger (2004)) applied to the variational posterior probabilities of association,* $\mathrm{E}_q(\gamma_{st})$; 16 *replicates are superimposed. The data comprise* $p = 1000$ *simulated SNPs with* 20 *hotspots,* $q = 20{,}000$ *responses, of which* 200 *are associated with at least one hotspot, leaving the rest of the responses unassociated. The block-autocorrelation coefficients for SNPs were drawn from the interval* $(0.75, 0.95)$*, and the residual block-equicorrelation coefficients for responses were drawn from the interval* $(0, 0.25)$*. At most* 25% *of each response variance is explained by the hotspots. For the fixed-variance models, we used a base rate* $\mu_\omega = 0.002$*, and scales* $\sigma_\omega = \mu_\omega \times \{1, 0.5, 0.1\}$*, as explained in the text.*

model places a *global* Gamma prior on the hotspot propensity precision and embeds the multiplicity penalty used in our proposal, that is,

$$\theta_s \mid \sigma_0 \sim \mathcal{N}(0, q^{-1}\sigma_0^2), \quad \sigma_0^{-2} \sim \text{Gamma}\left(\frac{1}{2}, \frac{1}{2}\right),$$

which can be reparametrized as

$$\theta_s \mid \sigma_0 \sim \mathcal{N}(0, \sigma_0^2), \quad \sigma_0^{-2} \sim \text{Gamma}\left(\frac{1}{2}, \frac{1}{2q}\right).$$

With this choice the propensity parameter has a Cauchy marginal prior distribution, $\theta_s \sim \mathrm{C}(0, q^{-1/2})$. Both the Cauchy and the horseshoe models rely on the base rate level used for the three fixed-variance models to define the prior expected number of predictors associated with each response as $\mathrm{E}_p = \mu_\omega \times p = 2$; the prior variance for this number is set to $\mathrm{V}_p = 100$ which is large enough to cover a wide range of configurations. We use annealed variational inference on all five models; the geometric schedule consists of a grid of 100 temperatures with initial temperature $T = 5$.

Figure 4 and Table 3 compare the five models in terms of selection of associated pairs of predictors and responses, selection of predictors (in our case, hotspots) and hotspot size estimation. They suggest several comments.

First, they illustrate our motivating statement: selection is sensitive to the choice of hotspot propensity variance; the pairwise selection performance of the three models with fixed variances varies greatly. The model with small variance strongly shrinks the hotspot sizes which prevents the detection of many associations. The model with large variance identifies more pairs but fails to uncover the smallest hotspots; their estimated signals are overwhelmed by

TABLE 3

Average standardized partial areas under the curve $\times 100$ *with false positive threshold* 0.01 *for predictor-response selection performance and predictor* (*hotspot*) *selection performance. Different hotspot size and effect size scenarios are reported, each based on* 64 *replicates*; *the "reference" case is displayed in Figure* 4. *Standard errors are in parentheses and, for each scenario, the best two performances are in bold*

	$\sigma_\omega = \mu_\omega \times 0.1$	$\sigma_\omega = \mu_\omega \times 0.5$	$\sigma_\omega = \mu_\omega \times 1$	Global	Global-local
Pairwise selection					
Reference	55.5 (1.2)	74.1 (1.3)	**85.9 (0.7)**	69.5 (1.2)	**93.6 (0.4)**
Smaller hotspots	56.7 (1.0)	67.7 (1.3)	**82.6 (0.8)**	74.4 (0.9)	**90.4 (0.6)**
Larger hotspots	57.7 (1.0)	84.4 (0.5)	**89.6 (0.3)**	65.3 (0.8)	**96.7 (0.1)**
Weaker hotspots	44.7 (1.0)	57.5 (1.4)	**77.3 (0.8)**	53.0 (1.2)	**81.5 (1.0)**
Stronger hotspots	64.0 (1.1)	82.6 (0.7)	**90.2 (0.4)**	78.6 (0.7)	**96.2 (0.1)**
Predictor selection					
Reference	65.8 (2.7)	68.2 (2.6)	68.2 (2.2)	**71.7 (2.1)**	**74.6 (1.6)**
Smaller hotspots	53.1 (3.4)	54.7 (3.3)	54.9 (3.2)	**61.0 (3.1)**	**64.0 (3.0)**
Larger hotspots	80.2 (2.9)	**84.3 (2.6)**	84.0 (2.7)	83.7 (2.4)	**87.1 (2.1)**
Weaker hotspots	52.9 (3.1)	54.6 (3.2)	56.0 (2.8)	**59.2 (2.9)**	**63.2 (2.5)**
Stronger hotspots	75.6 (3.1)	78.3 (2.8)	77.4 (2.7)	**81.3 (2.4)**	**84.7 (2.1)**

noise as a result of insufficient sparsity being induced (also see Supplementary Material S.5.1, Ruffieux et al. (2020)). Moreover, arbitrarily fixing hotspot propensity variances to large values may trigger artifactual "pile-up" effects when the data are less informative, as discussed in Section 3.

Second, the Cauchy model (global shrinkage only) is often able to discriminate the small hotspot signals from the noise, thanks to its global scale inferred from the data, but is not as good for pairwise selection and estimation of hotspot sizes; because it is mostly informed by SNPs with no simulated associations, the global scale concentrates towards zero which over-penalizes large hotspots, hampering the detection of pairwise associations with these hotspots. This phenomenon is of particular concern when signals are extremely sparse, as is thought to be the case in molecular QTL problems. Degeneracy issues can also arise in empirical Bayes settings; see, for example, Scott and Berger (2010), van de Wiel, Te Beest and Münch (2019) for a discussion and van der Pas, Szabó and van der Vaart (2016), van der Pas, Salomond and Schmidt-Hieber (2016), van der Pas, Szabó and van der Vaart (2017) for solutions based on prior distributions with support truncated away from zero. One may also attempt to improve the Cauchy specification by acknowledging the presence of genomic regions with diverse degrees of functional plausibility and introducing region-specific variance parameters to adapt to these degrees. Although inference may then be marginally impacted by the overall signal sparsity, such a formulation raises questions on the sensitivity to the chosen genome partition.

Our proposal performs well for selection of both response-predictor pairs and hotspots. Unlike the fixed-scale models, it can clearly separate the small hotspots from the noise. Moreover, the hotspot sizes are well inferred overall: there is some variability depending on the simulated effects (redrawn for each replicate), with the very small hotspots often underestimated, but the estimated sizes are much closer to the truth than those of the other models, which all strongly overshrink. We obtained the hotspot sizes by thresholding the variational posterior probabilities of association at 0.5, a threshold which corresponds to the *median probability model* rule described by Barbieri and Berger (2004) as having optimal prediction performance. Hence, the flexibility offered by the horseshoe's heavy-tailed local scale parameters improves on global scale parameter formulations, whether the parameter values are fixed or inferred.

6.3. *Null model scenario*. We examine the behaviour of our approach on data with neither hotspots nor individual associations. We took the data simulated for the first replicate of the "reference" scenario discussed in Section 6.2, but randomly shuffled the response sample labels, thus leaving the response correlation structure untouched. We ran the method on eight such permuted datasets and observed no hotspot using the 0.5-thresholding rule on the variational posterior probabilities of association: there were at most four associated responses per predictor. The average proportion of false positive pairwise associations was 2×10^{-5}.

6.4. *Annealed variational inference in presence of strong multicollinearity.* The present numerical experiment focuses on data exhibiting strong predictor and response multicollinearity. To best reproduce conditions encountered in molecular QTL studies, we used real SNP data from the eQTL study described in Section 2. We considered a 1.7 megabase (Mb) region located ≈ 1 Mb upstream of the MHC region and comprising 200 variants for which $n = 413$ observations were available. We distributed five active SNPs across the blocks and simulated 500 "active" responses. Effects were small, with each response having at most 10% of its variability explained by genetic variation. We added another 19, 500 inactive responses, drawn from Gaussian noise. The residual correlation of the responses spanned larger values than in Section 6.2, with block-correlation coefficients $\rho \in (0, 0.5)$.

Figure 5 indicates that the annealed variational algorithm clearly discriminates hotspots. Moreover, when declaring associations using a threshold of 0.5 on the marginal posterior probabilities, the hotspot sizes were well estimated, except for SNP id 105. In contrast, the nonannealed version of the algorithm struggled to single out the relevant SNPs from their correlated neighbours, especially around SNP id 110. Hence, the behaviour observed in the small experiment of Section 5 also arises in multiple-response settings. We also applied the algorithm with and without annealing on the data from the first replicate, performing 500 runs each using different starting values. We found that the optimal value reached by the objective function (12) was consistently higher and less variable in the annealed case (Figure 5). This was expected, as (12) is a lower bound on the marginal log-likelihood, but further suggests that this bound may indeed represent a good proxy for the marginal log-likelihood.

6.5. *Comparison with other approaches*. We conclude this series of simulation experiments by comparing the method with existing approaches. We choose two competing methods, MatrixEQTL (Shabalin (2012)) and HESS (Richardson, Bottolo and Rosenthal (2011), Bottolo et al. (2011)) as representative of two types of approaches: a univariate screening algorithm that tests the SNP-response pairs one by one, and joint hierarchical modelling coupled with parallel chain MCMC inference. We restrict the number of simulated responses to 10,000, in order to ensure a reasonable convergence time for the HESS MCMC run, and involve 15 SNPs in associations.

We rely on the default settings proposed in the MatrixEQTL and HESS implementations: these correspond, for the former, to using an additive linear model for the genotype effects and t-statistics for significance tests and, for the latter, to running three parallel chains for 22,000 iterations, discarding the first 2000 as burn-in. Our annealed variational inference procedure was about 30 times faster than the MCMC inference implemented in HESS, with an average runtime for one replicate of four hours and 17 minutes for the former and five days and 10 hours for the latter on an Intel Xeon CPU, 2.60 GHz.

As expected, the ROC curves of Figure 5 indicate that MatrixEQTL performs worse than the two joint approaches. It correctly identifies the strong associations but also declares many spurious associations involving SNPs in high linkage disequilibrium. This agrees with the motivating example in Section 2; marginal screening often provides satisfactory answers

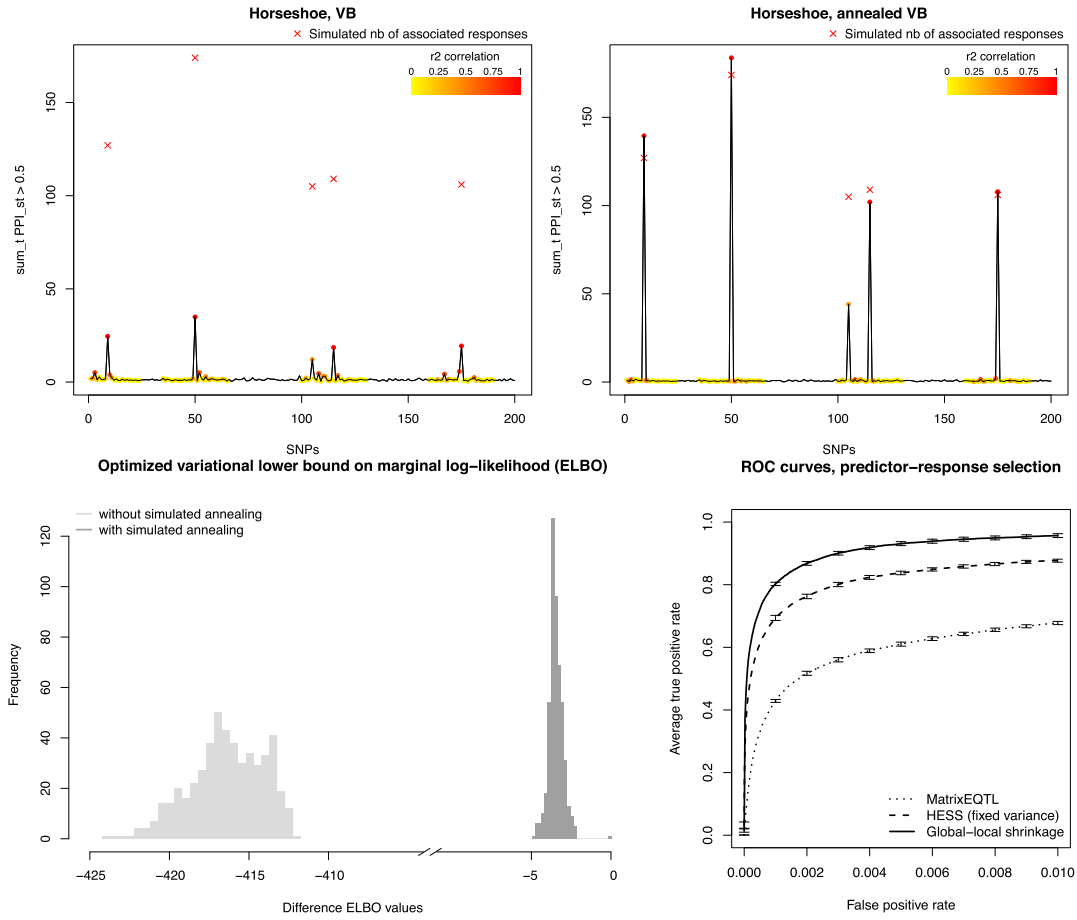

FIG. 5. *Performance comparisons between classical and annealed variational inferences, and with competing methods. Problem with responses equicorrelated by blocks with residual correlation* $\rho \in (0, 0.5)$; 500 *of them are under genetic control. Candidate predictors are* $p = 200$ *SNPs from a cohort of European ancestry, for* $n = 413$ *individuals. Top: Hotspot discrimination achieved by classical (left) and annealed (right) variational inferences, for a problem with* $q = 20{,}000$ *responses. The plots show the cumulated number of responses associated per SNP, after thresholding the variational posterior probabilities of association at* 0.5, *and averaging over* 16 *replicates. The crosses show the simulated sizes of the five hotspots (whose cumulated effects account for at most* 10% *of the variability of a response). The highlighted regions quantify the linkage disequilibrium structure in* r^2 *computed with respect to hotspots* 9, 50, 115 *and* 175, *respectively; hotspot* 105 *is correlated with hotspot* 115. *Bottom, left: Histograms of optimized lower bound on the marginal log-likelihood (ELBO) with classical and annealed variational inferences,* 500 *replicates; the* x*-axis shows the maximum ELBO value subtracted from all other values. Bottom, right: Truncated average ROC curves with* 95% *confidence intervals for the MatrixEQTL and HESS methods, and our proposal. The same settings as above are used, except for the number of responses, limited to* $q = 10{,}000$, *and the number of hotspots,* 15, *whose cumulated effects account for at most* 20% *of the variability of a response. Both HESS and our proposal have prior expectation* $\mathrm{E}_p = 1$ *and variance* $\mathrm{V}_p = 10$ *for the number of SNPs associated with a response; the value of* E_p *is smaller than in Sections* 6.2 *and* 6.3 *because there are fewer candidate predictors, and so is the value of* V_p, *to limit the computational costs of HESS.*

when the aim is to highlight *cis* associations at the level of loci, but, because of the multiplicity burden, it often fails to declare weaker effects such as those involved in *trans* associations. By borrowing information across all SNPs and responses, HESS achieves much better association recovery. The HESS run is based on a specific choice of hotspot propensity variance which is hard-coded and not accessible to the user; we expect the performance to vary with other choices of variances, similarly to what was shown in Figure 4 for the approach of Ruffieux et al. (2017). With its global and local variances inferred from the data, our proposal

performs best. Confronting this performance with MCMC inference further suggests that the independence assumptions underlying the variational mean-field formulation do not degrade the quality of variable selection, as shown by Ruffieux et al. (2017). The coupling with simulated annealing results in an excellent selection in our experiments and in a fraction of the time required by MCMC techniques; this is particularly remarkable in highly multimodal settings.

7. A targeted study of hotspot activity with stimulated monocyte expression. We return to the eQTL data presented in Section 2. As discussed there, stimulation of monocytes may boost *trans*-regulatory activity, so the analysis of stimulated eQTL data should benefit from a method tailored to the detection of hotspots. In this section we analyse three genomic regions comprising genes thought to play a central role in the pathogenesis of immune disorders (Fairfax et al. (2012), Fairfax et al. (2014)): *NFE2L3* on chromosome 7, *IFNB1* on chromosome 9 and *LYZ* on chromosome 12. Each region involves 1500 SNPs and spans from 7.5 to 12 Mb.

The following quality control steps were performed prior to the analyses. For the genotyping we applied standard filters that exclude SNPs with call rate $< 95\%$, violate the Hardy–Weinberg equilibrium assumption (at nominal p-value level 10^{-4}) or have minor allele frequency < 0.05. For the transcripts we considered the top 30% quantile of the interquartile range distributions in each (un)stimulated condition. In order to work with a common set of transcripts across conditions, we then retained the intersection of the transcripts selected in each condition and checked that no highly varying transcript was dropped in this process. Finally, we discarded samples with unusual transcript values, separately for each condition. The numbers of individuals thus retained were 413 for unstimulated monocytes, 366 for IFN-γ, 260 for LPS 2 h, 321 for LPS 24 h and 275 for B-cells; the number of transcripts was 24,461.

We ran our method on each of the three regions and for all four monocyte conditions, as well as for the B cells, resulting in 15 separate analyses. We employed the same prior base rate of associated pairs as in the simulation of Sections 6.2 and 6.3, giving a prior average of $\mathrm{E}_p = 0.002 \times p = 3$ SNPs associated with each transcript, and used a variance of $\mathrm{V}_p = 25$. Figure 6 compares the evidence for hotspots produced by our proposal and plain univariate screening. It shows the nominal $-\log_{10}$ p-values of a univariate screening against the raw posterior probabilities, both summed across responses, and suggests that the two approaches agree on the small or moderate evidence but also that our proposal appears to boost and better distinguish hotspot effects.

In order to derive empirical false discovery rates, we ran a permutation analysis with 30 replicates for each region and condition by shuffling the sample labels of the expression matrix; this was computationally feasible thanks to the efficiency of our variational procedure. We then obtained Bayesian false discovery rates for a fine grid of thresholds on the variational posterior probabilities of association and fitted a spline in order to derive thresholds corresponding to a false discovery rate of 20%.

Figure 6 indicates increased *trans*-regulatory activity under stimulation with IFN-γ and LPS 24 h. This activity was endorsed by the absence of hotspots in the B-cell analysis; indeed, previous studies comparing B cells and monocytes on the three regions suggested that QTL activity was specific to the latter (Fairfax et al. (2012)), so the former may be used as negative controls in our analyses. The degree of activity also varies greatly across the three regions: the *NFE2L3* region is essentially inactive; its largest hotspot is of size 8 and appears under IFN-γ stimulation, in line with previous observations (Fairfax et al. (2014)). The *IFNB1* region shows more activity under LPS 24 h stimulation; this confirms existing work (Fairfax et al. (2014)) but also reveals more associations with transcripts. The top LPS 24 h hotspot in the *IFNB1* region, rs3898946, is an eQTL reported in the GTEx database, for genes *FOCAD*

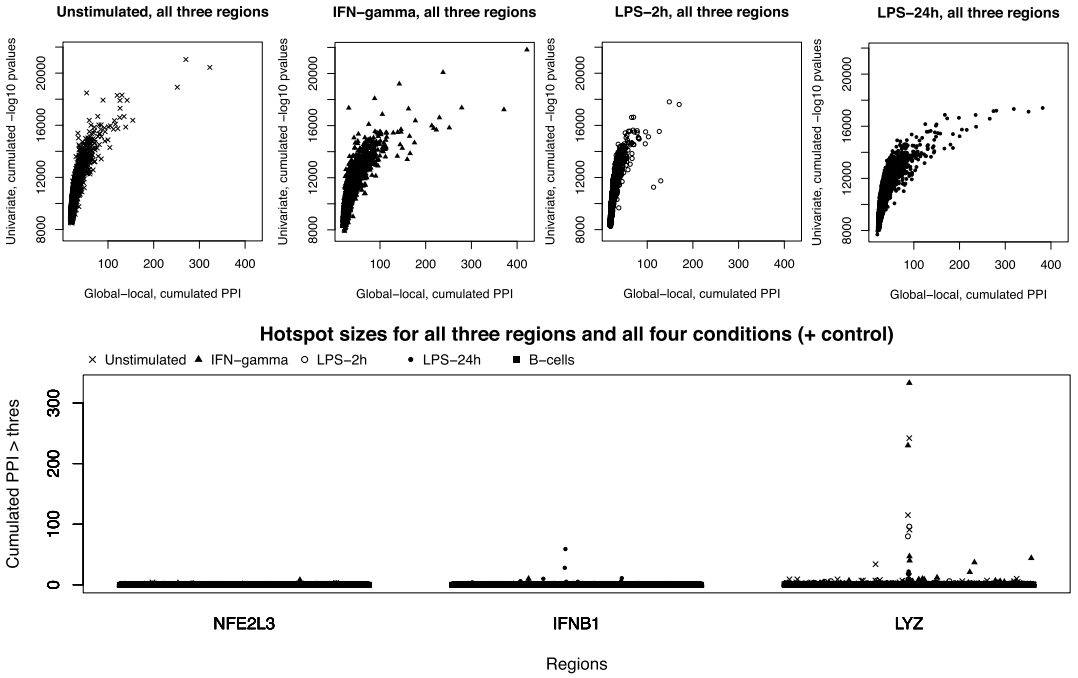

FIG. 6. *Hotspots from stimulated eQTL analyses, for the* NFEL2L3, IFNB1 *and* LYZ *genomic regions with the four monocyte conditions and the B-cell negative controls. Top: for each condition, raw hotspot evidence for all three regions comprising* NFEL2L3, IFNB1 *and* LYZ. *Scatterplots with* $-\log_{10}$ *p-values of univariate screening, summed across responses, versus variational posterior probabilities of association obtained by our proposal, summed across responses. Bottom: Hotspot sizes, as declared using a permutation-based FDR of* 20%.

and *MLLT3* in the tibial artery and for gene *PTPLAD2* in skin tissues; this provides further support for a mechanistic role of this hotspot (to be confirmed in further work). The *LYZ* region is known for its high degree of pleiotropy (Rotival et al. (2011)) and is indeed very active in our analyses.

Although Fairfax et al. (2014) mostly report stimuli-specific *trans*-regulatory activities, our top hotspot hit, rs6581889, located only *nine* Kb downstream of the *LYZ* gene, is persistent across all four conditions: it is the largest hotspot in the unstimulated condition with size 242, in the IFN-γ condition with size 333 and in the LPS 2 h condition with size 96; it is the second largest hotspot in the LPS 24 h condition with size 18. A Venn diagram showing the transcript overlap across conditions is given in the Supplementary Material S.6 (Ruffieux et al. (2020)). Hence, the SNP activity was triggered by the IFN-γ stimulation but was also substantial after *two* hours and 24 hours of LPS stimulation. The B-cell data provide a good negative control, as they show no activity in the *LYZ* region; the largest number of responses associated with a given SNP is three, and the signal does not colocalize with any hotspot uncovered in monocytes. Finally, rs6581889 is a known *cis* eQTL for *LYZ* and *YEATS4* in multiple tissues, two associations which our analyses confirmed.

8. Conclusion. We have introduced a new approach for the efficient detection hotspots in regression problems with tens of thousands of response variables. Our proposal makes novel contributions to both modelling and inference: it introduces a flexible fully Bayesian model for hotspots and implements an efficient variational inference procedure coupled with simulated annealing. It accommodates three essential characteristics of molecular QTL: extreme sparseness of association patterns, strong multimodality induced by locally correlated genetic variants and very high dimensions of both the predictor and the response vectors.

Our simulations indicated that severe sparsity renders ineffective models based only on a global variance for the hotspot propensity. Our global-local model provides sufficient refinement to properly identify the locations and sizes of individual hotspots; it is free of ad hoc variance choices and automatically adapts to different signal sparsity degrees.

Collinearity exacerbates posterior multimodality and often causes unstable estimates when obtained by joint inference. As accurate inference is critical to the effective use of the hotspot model in high dimensions, we developed a simulated annealing scheme to improve the exploration of multimodal posterior spaces. In our numerical experiments, the resulting inferences were robust to different algorithm initializations, even on data with marked correlation structures. It yielded satisfactory estimates of hotspot sizes in situations where classical variational inference would strongly overshrink.

Our formulation of the first-level model involves two-group mixture priors of the spike-and-slab form for the regression coefficients β_{st} rather than one-group continuous priors, such as the Laplace or the horseshoe prior, reserving the latter for the second-level modelling of the probability parameters ω_{st}. The relative merits of one-group and two-group shrinkage priors for testing purposes have been a subject of considerable discussion over the past few years (see, e.g., Li and Pati (2017), Piironen and Vehtari (2017)). Testing associations for each predictor and response pair can be effectively achieved in a variety of fashions (while permitting some borrowing of information across the responses), including with one-group priors on β_{st}. Indeed, good theoretical guarantees for testing and uncertainty quantification are now available for the one-group prior framework. In particular, van der Pas, Kleijn and van der Vaart (2014), van der Pas, Szabó and van der Vaart (2016), van der Pas, Szabó and van der Vaart (2017) studied posterior concentration under the horseshoe prior. Datta and Ghosh (2013) established optimal asymptotic error rates for a multiple testing decision rule based on the horseshoe shrinkage factor, and Ghosh et al. (2016) extended their results to a rich class of one-group priors. Interestingly, Datta and Ghosh pointed out that two-group priors are conceptually very natural for testing tasks, thanks to their noise-signal components, and that the horseshoe decision rule is built on analogies with the two-group model. At the same time, the computational advantages of one-group continuous priors are also often contrasted to the burden caused by the large discrete search space induced by two-group mixture priors.

In our model we used both a two-group prior and a one-group prior, at different levels. We bypassed the computational burden of the two-group formulation by developing a variational approach that permits fast inference, even under the spike-and-slab prior of the first level. Crucially, the spike-and-slab formulation directly serves the primary aim of our method: the detection of hotspot effects via a dedicated parameter, θ_s, that allows further borrowing of information across the responses. We saw how this leads to a natural representation of the hotspot propensities as predictor-specific modulations of the probability of association. We then made use of the adaptive properties of the global-local horseshoe formulation to embed a penalty that adjusts for the response dimension and prevents the manifestation of artifactual hotspots when the likelihood is relatively flat; we provided two complementary justifications for its choice.

Several extensions may be considered. First, the illustration on stimulated-monocyte eQTL data suggests extending the model to jointly account for the multiple stimulated states. Other types of conditions may benefit from such joint modelling: for instance, molecular QTL data are nowadays often collected for multiple tissues or time points. See Petretto et al. (2010) and Lewin et al. (2015) for examples based on the model of Richardson, Bottolo and Rosenthal (2011). Second, on the algorithmic side a natural enhancement would be to embed the annealing temperature as an auxiliary parameter to be inferred. This would permit adaptive and dynamic control of the temperature schedule and may help to balance the number of temperatures used and, hence, the use of resources, with the level of entropy needed for

good exploration. Mandt et al. (2016) have a procedure based on this idea, but their proposal requires precomputing an approximation of the joint distribution normalizing constant. Sensible cheap estimates may be envisioned for *large n* cases but are unrealistic for high-dimensional regression. Obtaining theoretical guarantees for our algorithm would also be beneficial; several recent results on tempered variational approximation for simpler models suggest that desirable convergence properties may be provable for our annealed variational updates (Alquier, Ridgway and Chopin (2016), Alquier and Ridgway (2017), Yang, Pati and Bhattacharya (2017)).

We do not claim that our approach can provide direct conclusive evidence on the functional consequences of the identified hotspots, as this always requires follow-up studies at the level of individual loci. We do argue, however, that it is well suited to highlight promising candidate variants for functional studies which may save substantial investment in prospective research. Our method applies to any type of molecular QTL problem. In particular, it may be used with proteomic and lipidomic expression data which are gaining in popularity because they may be more closely linked with clinical phenotypes.

Software. The software `atlasqtl` is written in R with C++ subroutines.

Acknowledgements. We are grateful to the editor and the two anonymous referees for their valuable comments that improved the presentation of the paper. We thank Armand Valsesia for his helpful comments, and also thank Colin Star, Bruce O'Neel and Jaroslaw Szymczak for giving us access to computing resources.

This research was funded by Nestlé Research (H.R., J.H.), the Alan Turing Institute under the Engineering and Physical Sciences Research Council grant EP/N510129/1 (L.B.), the MRC grant MR/M013138/1 "Methods and tools for structural models integrating multiple high-throughput omics data sets in genetic epidemiology" (L.B.), the UK Medical Research Council programme MRC MC UU 00002/10 (S.R.) and the Alan Turing Institute Fellowship number TU/B/000092 (S.R.).

SUPPLEMENTARY MATERIAL

Supplement A (DOI: 10.1214/20-AOAS1332SUPPA; .pdf). Supplementary Material contains technical appendices and complements to the motivating real data example and numerical experiments.

Supplement B (DOI: 10.1214/20-AOAS1332SUPPB; .zip). R package `atlasqtl` v.0.1.3 implementing the method.

REFERENCES

ALQUIER, P. and RIDGWAY, J. (2017). Concentration of tempered posteriors and of their variational approximations. Preprint. Available at arXiv:1706.09293.

ALQUIER, P., RIDGWAY, J. and CHOPIN, N. (2016). On the properties of variational approximations of Gibbs posteriors. *J. Mach. Learn. Res.* **17** Art. ID 239. MR3595173

BARBIERI, M. M. and BERGER, J. O. (2004). Optimal predictive model selection. *Ann. Statist.* **32** 870–897. MR2065192 https://doi.org/10.1214/009053604000000238

BARDENET, R., DOUCET, A. and HOLMES, C. (2014). Towards scaling up Markov chain Monte Carlo: An adaptive subsampling approach. In *International Conference on Machine Learning* (*ICML*) 405–413.

BERGER, J. (1980). A robust generalized Bayes estimator and confidence region for a multivariate normal mean. *Ann. Statist.* **8** 716–761. MR0572619

BHADRA, A. and MALLICK, B. K. (2013). Joint high-dimensional Bayesian variable and covariance selection with an application to eQTL analysis. *Biometrics* **69** 447–457. MR3071063 https://doi.org/10.1111/biom.12021

BHADRA, A., DATTA, J., POLSON, N. G. and WILLARD, B. (2016). Default Bayesian analysis with global-local shrinkage priors. *Biometrika* **103** 955–969. MR3620450 https://doi.org/10.1093/biomet/asw041

BHATTACHARYA, A. and DUNSON, D. B. (2010). Nonparametric Bayesian density estimation on manifolds with applications to planar shapes. *Biometrika* **97** 851–865. MR2746156 https://doi.org/10.1093/biomet/asq044

BLEI, D. M., KUCUKELBIR, A. and MCAULIFFE, J. D. (2017). Variational inference: A review for statisticians. *J. Amer. Statist. Assoc.* **112** 859–877. MR3671776 https://doi.org/10.1080/01621459.2017.1285773

BOTTOLO, L., PETRETTO, E., BLANKENBERG, S., CAMBIEN, F., COOK, S. A., TIRET, L. and RICHARDSON, S. (2011). Bayesian detection of expression quantitative trait loci hot spots. *Genetics* **189** 1449–1459.

BRYNEDAL, B., CHOI, J., RAJ, T., BJORNSON, R., STRANGER, B. E., NEALE, B. M., VOIGHT, B. F. and COTSAPAS, C. (2017). Large-scale trans-eQTLs affect hundreds of transcripts and mediate patterns of transcriptional co-regulation. *Am. J. Hum. Genet.* **100** 581–591. https://doi.org/10.1016/j.ajhg.2017.02.004

CARBONETTO, P. and STEPHENS, M. (2012). Scalable variational inference for Bayesian variable selection in regression, and its accuracy in genetic association studies. *Bayesian Anal.* **7** 73–107. MR2896713 https://doi.org/10.1214/12-BA703

CARVALHO, C. M., POLSON, N. G. and SCOTT, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika* **97** 465–480. MR2650751 https://doi.org/10.1093/biomet/asq017

DATTA, J. and GHOSH, J. K. (2013). Asymptotic properties of Bayes risk for the horseshoe prior. *Bayesian Anal.* **8** 111–131. MR3036256 https://doi.org/10.1214/13-BA805

FAIRFAX, B. P., MAKINO, S., RADHAKRISHNAN, J., PLANT, K., LESLIE, S., DILTHEY, A., ELLIS, P., LANGFORD, C., VANNBERG, F. O. et al. (2012). Genetics of gene expression in primary immune cells identifies cell type-specific master regulators and roles of HLA alleles. *Nat. Genet.* **44** 502–510.

FAIRFAX, B. P., HUMBURG, P., MAKINO, S., NARANBHAI, V., WONG, D., LAU, E., JOSTINS, L., PLANT, K., ANDREWS, R. et al. (2014). Innate immune activity conditions the effect of regulatory variants upon monocyte gene expression. *Science* **343** Art. ID 1246949.

GELMAN, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Anal.* **1** 515–533. MR2221284 https://doi.org/10.1214/06-BA117A

GELMAN, A., JAKULIN, A., PITTAU, M. G. and SU, Y.-S. (2008). A weakly informative default prior distribution for logistic and other regression models. *Ann. Appl. Stat.* **2** 1360–1383. MR2655663 https://doi.org/10.1214/08-AOAS191

GEORGE, E. I. (2000). The variable selection problem. *J. Amer. Statist. Assoc.* **95** 1304–1308. MR1825282 https://doi.org/10.2307/2669776

GHOSH, P., TANG, X., GHOSH, M. and CHAKRABARTI, A. (2016). Asymptotic properties of Bayes risk of a general class of shrinkage priors in multiple hypothesis testing under sparsity. *Bayesian Anal.* **11** 753–796. MR3498045 https://doi.org/10.1214/15-BA973

GILAD, Y., RIFKIN, S. A. and PRITCHARD, J. K. (2008). Revealing the architecture of gene regulation: The promise of eQTL studies. *Trends Genet.* **24** 408–415.

GRAMACY, R., SAMWORTH, R. and KING, R. (2010). Importance tempering. *Stat. Comput.* **20** 1–7. MR2578072 https://doi.org/10.1007/s11222-008-9108-5

GUHANIYOGI, R., QAMAR, S. and DUNSON, D. B. (2018). Bayesian conditional density filtering. *J. Comput. Graph. Statist.* **27** 657–672. MR3863766 https://doi.org/10.1080/10618600.2017.1422431

JIA, Z. and XU, S. (2007). Mapping quantitative trait loci for expression abundance. *Genetics* **176** 611–623.

KATAHIRA, K., WATANABE, K. and OKADA, M. (2008). Deterministic annealing variant of variational Bayes method. *J. Phys., Conf. Ser.* **95** Art. ID 012015.

KIM, S., BECKER, J., BECHHEIM, M., KAISER, V., NOURSADEGHI, M., FRICKER, N., BEIER, E., KLASCHIK, S., BOOR, P. et al. (2014). Characterizing the genetic basis of innate immune response in TLR4-activated human monocytes. *Nat. Commun.* **5** Art. ID 5236.

KIRKPATRICK, S., GELATT, C. D. JR. and VECCHI, M. P. (1983). Optimization by simulated annealing. *Science* **220** 671–680. MR0702485 https://doi.org/10.1126/science.220.4598.671

LEE, M. N., YE, C., VILLANI, A.-C., RAJ, T., LI, W., EISENHAURE, T. M., IMBOYWA, S. H., CHIPENDO, P. I., RAN, F. A. et al. (2014). Common genetic variants modulate pathogen-sensing responses in human dendritic cells. *Science* **343** Art. ID 1246980.

LEWIN, A., SAADI, H., PETERS, J. E., MORENO-MORAL, A., LEE, J. C., SMITH, K. G. C., PETRETTO, E., BOTTOLO, L. and RICHARDSON, S. (2015). MT-HESS: An efficient Bayesian approach for simultaneous association detection in OMICS datasets, with application to eQTL mapping in multiple tissues. *Bioinformatics* **32** 523–532.

LI, H. and PATI, D. (2017). Variable selection using shrinkage priors. *Comput. Statist. Data Anal.* **107** 107–119. MR3575062 https://doi.org/10.1016/j.csda.2016.10.008

MACKAY, T. F. C., STONE, E. A. and AYROLES, J. F. (2009). The genetics of quantitative traits: Challenges and prospects. *Nat. Rev. Genet.* **10** 565–577.

MANDT, S., MCINERNEY, J., ABROL, F., RANGANATH, R. and BLEI, D. (2016). Variational tempering. In *Proceedings of the* 19*th International Conference on Artificial Intelligence and Statistics*. *Proceedings of Machine Learning Research* (*PMLR*) **51** 704–712.

METROPOLIS, N., ROSENBLUTH, A. W., ROSENBLUTH, M. N., TELLER, A. H. and TELLER, E. (1953). Equation of state calculations by fast computing machines. *J. Chem. Phys.* **21** 1087–1092.

MITCHELL, T. J. and BEAUCHAMP, J. J. (1988). Bayesian variable selection in linear regression. *J. Amer. Statist. Assoc.* **83** 1023–1036. MR0997578

NEVILLE, S. E., ORMEROD, J. T. and WAND, M. P. (2014). Mean field variational Bayes for continuous sparse signal shrinkage: Pitfalls and remedies. *Electron. J. Stat.* **8** 1113–1151. MR3263115 https://doi.org/10.1214/14-EJS910

NICA, A. C. and DERMITZAKIS, E. T. (2013). Expression quantitative trait loci: Present and future. *Philos. Trans. R. Soc. B* **368** Art. ID 20120362.

O'BRIEN, S. M. and DUNSON, D. B. (2004). Bayesian multivariate logistic regression. *Biometrics* **60** 739–746. MR2089450 https://doi.org/10.1111/j.0006-341X.2004.00224.x

OPPER, M. and SAAD, D., eds. (2001). *Advanced Mean Field Methods*: *Theory and Practice*. *Neural Information Processing Series*. MIT Press, Cambridge, MA. MR1863214

PARK, J.-H., GAIL, M. H., WEINBERG, C. R., CARROLL, R. J., CHUNG, C. C., WANG, Z., CHANOCK, S. J., FRAUMENI, J. F. and CHATTERJEE, N. (2011). Distribution of allele frequencies and effect sizes and their interrelationships for common genetic susceptibility variants. *Proc. Natl. Acad. Sci. USA* **108** 18026–18031.

PETERSON, C. B., BOGOMOLOV, M., BENJAMINI, Y. and SABATTI, C. (2016). TreeQTL: Hierarchical error control for eQTL findings. *Bioinformatics* **32** 2556–2558.

PETRETTO, E., BOTTOLO, L., LANGLEY, S. R., HEINIG, M., MCDERMOTT-ROE, C., SARWAR, R., PRAVENEC, M., HÜBNER, N., AITMAN, T. J. et al. (2010). New insights into the genetic control of gene expression using a Bayesian multi-tissue approach. *PLoS Comput. Biol.* **6** Art. ID e1000737.

PIIRONEN, J. and VEHTARI, A. (2017). Sparsity information and regularization in the horseshoe and other shrinkage priors. *Electron. J. Stat.* **11** 5018–5051. MR3738204 https://doi.org/10.1214/17-EJS1337SI

POLSON, N. G. and SCOTT, J. G. (2011). Shrink globally, act locally: Sparse Bayesian regularization and prediction. In *Bayesian Statistics* 9 (J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith and M. West, eds.) 501–538. Oxford Univ. Press, Oxford. MR3204017 https://doi.org/10.1093/acprof:oso/9780199694587.003.0017

RICHARDSON, S., BOTTOLO, L. and ROSENTHAL, J. S. (2011). Bayesian models for sparse regression analysis of high dimensional data. In *Bayesian Statistics* 9 (J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith and M. West, eds.) 539–568. Oxford Univ. Press, Oxford. MR3204018 https://doi.org/10.1093/acprof:oso/9780199694587.003.0018

ROČKOVÁ, V. and GEORGE, E. I. (2014). EMVS: The EM approach to Bayesian variable selection. *J. Amer. Statist. Assoc.* **109** 828–846. MR3223753 https://doi.org/10.1080/01621459.2013.869223

ROSE, K., GUREWITZ, E. and FOX, G. (1990). A deterministic annealing approach to clustering. *Pattern Recogn. Lett.* **11** 589–594.

ROTIVAL, M., ZELLER, T., WILD, P. S., MAOUCHE, S., SZYMCZAK, S., SCHILLERT, A., CASTAGNÉ, R., DEISEROTH, A., PROUST, C. et al. (2011). Integrating genome-wide genetic variations and monocyte expression data reveals trans-regulated gene modules in humans. *PLoS Genet.* **7** Art. ID e1002367.

RUFFIEUX, H., DAVISON, A. C., HAGER, J. and IRINCHEEVA, I. (2017). Efficient inference for genetic association studies with multiple outcomes. *Biostatistics* **18** 618–636. MR3984082 https://doi.org/10.1093/biostatistics/kxx007

RUFFIEUX, H., DAVISON, A. C., HAGER, J., INSHAW, J., FAIRFAX, B., RICHARDSON, S. and BOTTOLO, L. (2020). Supplement to "A global-local approach for detecting hotspots in multiple-response regression". https://doi.org/10.1214/20-AOAS1332SUPPA, https://doi.org/10.1214/20-AOAS1332SUPPB

SCOTT, J. G. and BERGER, J. O. (2010). Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. *Ann. Statist.* **38** 2587–2619. MR2722450 https://doi.org/10.1214/10-AOS792

SHABALIN, A. A. (2012). Matrix eQTL: Ultra fast eQTL analysis via large matrix operations. *Bioinformatics* **28** 1353–1358.

SOLOVIEFF, N., COTSAPAS, C., LEE, P. H., PURCELL, S. M. and SMOLLER, J. W. (2013). Pleiotropy in complex traits: Challenges and strategies. *Nat. Rev. Genet.* **14** 483–495.

STRAWDERMAN, W. E. (1971). Proper Bayes minimax estimators of the multivariate normal mean. *Ann. Math. Stat.* **42** 385–388. MR0397939 https://doi.org/10.1214/aoms/1177693528

TAK, Y. G. and FARNHAM, P. J. (2015). Making sense of GWAS: Using epigenomics and genome engineering to understand the functional relevance of SNPs in non-coding regions of the human genome. *Epigenet. Chromatin* **8** Art. ID 57. https://doi.org/10.1186/s13072-015-0050-4

UEDA, N. and NAKANO, R. (1998). Deterministic annealing EM algorithm. *Neural Netw.* **11** 271–282.

VAN DE WIEL, M. A., TE BEEST, D. E. and MÜNCH, M. M. (2019). Learning from a lot: Empirical Bayes for high-dimensional model-based prediction. *Scand. J. Stat.* **46** 2–25. MR3915265 https://doi.org/10.1111/sjos.12335

VAN DER PAS, S. L., KLEIJN, B. J. K. and VAN DER VAART, A. W. (2014). The horseshoe estimator: Posterior concentration around nearly black vectors. *Electron. J. Stat.* **8** 2585–2618. MR3285877 https://doi.org/10.1214/14-EJS962

VAN DER PAS, S. L., SALOMOND, J.-B. and SCHMIDT-HIEBER, J. (2016). Conditions for posterior contraction in the sparse normal means problem. *Electron. J. Stat.* **10** 976–1000. MR3486423 https://doi.org/10.1214/16-EJS1130

VAN DER PAS, S., SZABÓ, B. and VAN DER VAART, A. (2017). Adaptive posterior contraction rates for the horseshoe. *Electron. J. Stat.* **11** 3196–3225. MR3705450 https://doi.org/10.1214/17-EJS1316

VAN DER PAS, S., SZABÓ, B. and VAN DER VAART, A. (2016). How many needles in the haystack? Adaptive inference and uncertainty quantification for the horseshoe. Preprint. Available at arXiv:1607.01892.

WAINWRIGHT, M. J. and JORDAN, M. I. (2008). Graphical models, exponential families, and variational inference. *Found. Trends Mach. Learn.* **1** 1–305.

WANG, X. and DUNSON, D. B. (2013). Parallelizing MCMC via Weierstrass sampler. Preprint. Available at arXiv:1312.4605.

WARD, L. D. and KELLIS, M. (2012). Interpreting noncoding genetic variation in complex traits and human disease. *Nat. Biotechnol.* **30** 1095–1106. https://doi.org/10.1038/nbt.2422

WESTRA, H.-J., PETERS, M. J., ESKO, T., YAGHOOTKAR, H., SCHURMANN, C., KETTUNEN, J., CHRISTIANSEN, M. W., FAIRFAX, B. P., SCHRAMM, K. et al. (2013). Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nat. Genet.* **45** 1238–1243.

YANG, Y., PATI, D. and BHATTACHARYA, A. (2017). Alpha-variational inference with statistical guarantees. Preprint. Available at arXiv:1710.03266.

YAO, C., JOEHANES, R., JOHNSON, A. D., HUAN, T., LIU, C., FREEDMAN, J. E., MUNSON, P. J., HILL, D. E., VIDAL, M. et al. (2017). Dynamic role of trans regulation of gene expression in relation to complex traits. *Am. J. Hum. Genet.* **100** 571–580.

YIN, J. and LI, H. (2011). A sparse conditional Gaussian graphical model for analysis of genetical genomics data. *Ann. Appl. Stat.* **5** 2630–2650. MR2907129 https://doi.org/10.1214/11-AOAS494