Please cite this article in press as: Ruffieux et al., EPISPOT: An epigenome-driven approach for detecting and interpreting hotspots in molecular QTL studies, The American Journal of Human Genetics (2021), https://doi.org/10.1016/j.ajhg.2021.04.010

ARTICLE

EPISPOT: An epigenome-driven approach for detecting and interpreting hotspots in molecular QTL studies

Hélène Ruffieux,[1,*] Benjamin P. Fairfax,[2] Isar Nassiri,[2] Elena Vigorito,[1] Chris Wallace,[1,3] Sylvia Richardson,[1,4] and Leonardo Bottolo[1,4,5]

Summary

We present EPISPOT, a fully joint framework which exploits large panels of epigenetic annotations as variant-level information to enhance molecular quantitative trait locus (QTL) mapping. Thanks to a purpose-built Bayesian inferential algorithm, EPISPOT accommodates functional information for both *cis* and *trans* actions, including QTL *hotspot* effects. It effectively couples simultaneous QTL analysis of thousands of genetic variants and molecular traits with hypothesis-free selection of biologically interpretable annotations which directly contribute to the QTL effects. This unified, epigenome-aided learning boosts statistical power and sheds light on the regulatory basis of the uncovered hits; EPISPOT therefore marks an essential step toward improving the challenging detection and functional interpretation of *trans*-acting genetic variants and hotspots. We illustrate the advantages of EPISPOT in simulations emulating real-data conditions and in a monocyte expression QTL study, which confirms known hotspots and finds other signals, as well as plausible mechanisms of action. In particular, by highlighting the role of monocyte DNase-I sensitivity sites from >150 epigenetic annotations, we clarify the mediation effects and cell-type specificity of major hotspots close to the lysozyme gene. Our approach forgoes the daunting and underpowered task of one-annotation-at-a-time enrichment analyses for prioritizing *cis* and *trans* QTL hits and is tailored to any transcriptomic, proteomic, or metabolomic QTL problem. By enabling principled epigenome-driven QTL mapping transcriptome-wide, EPISPOT helps progress toward a better functional understanding of genetic regulation.

Introduction

Molecular datasets and annotation databases are growing in size and in diversity. In particular, genetic data are now routinely collected along with gene, protein, or metabolite level measurements and analyzed in molecular quantitative trait locus (QTL) studies, with the aim of unravelling the regulatory mechanisms underlying common diseases. However, these studies present additional complexities compared to classical genome-wide association studies (GWASs). First, they entail a very different statistical paradigm: while GWASs consider a single or a few related clinical traits, molecular QTL studies typically involve hundreds or thousands of molecular traits, regressed on hundreds of thousands of genetic variants. Second, they need to accommodate two types of genetic control: a variant may affect molecular products of genes in its vicinity (*cis* action) or products of remote genes (*trans* action), where the latter mode of control is typically much weaker and, hence, harder to uncover than the former. In particular, pleiotropic or hotspot genetic variants may exert weak *trans* effects on many molecular traits.

The current mapping practice only partially embraces the features of QTL studies. Indeed, widely used marginal screening approaches[1,2] suffer from a large multiplicity burden and tend to lack statistical power as they do not exploit the regulation patterns shared by the molecular entities, whereas joint modeling approaches[3,4] are often limited by the computational burden implied by the exploration of high-dimensional spaces of candidate variants and traits. To manage this tension between scalable inference and comprehensive joint modeling, we recently proposed a variational inference approach, called ATLASQTL,[5] which explicitly borrows information across thousands of molecular traits controlled by shared pathways and offers a robust fully Bayesian parametrization of hotspots; its increased sensitivity and that of earlier related models have been demonstrated in different molecular QTL studies.[4–7]

In complement to the actual mapping task, biologists increasingly try to capitalize on the wealth of available epigenetic annotation sources to infer the functional potential of genetic variants. The standard strategy uses epigenetic marks mostly for prioritization of hits derived from marginal screening: it consists in looping through all the loci with statistically significant associations and, for each locus, inspecting marks to decide on "a most promising" functional candidate genetic variant among all those in linkage disequilibrium (LD). This approach has the following disadvantages. First, publicly available databases nowadays contain several hundred epigenetic annotations. Preselecting just a few may involve omitting others that are relevant, which may bias the conclusions. Second, even if a comprehensive inspection were feasible,

[1]MRC Biostatistics Unit, University of Cambridge, Cambridge CB2 0SR, UK; [2]Department of Oncology, MRC Weatherall Institute for Molecular Medicine, University of Oxford, John Radcliffe Hospital, Oxford OX3 9DS, UK; [3]Cambridge Institute of Therapeutic Immunology and Infectious Disease, Jeffrey Cheah Biomedical Centre, Cambridge Biomedical Campus, University of Cambridge, Cambridge CB2 0AW, UK; [4]The Alan Turing Institute, London NW1 2DB, UK; [5]Department of Medical Genetics, University of Cambridge, Cambridge CB2 0QQ, UK
*Correspondence: helene.ruffieux@mrc-bsu.cam.ac.uk
https://doi.org/10.1016/j.ajhg.2021.04.010.

the degrees of relevance of the annotations may be very uneven and may depend on the conditions, cell types, tissues, and even genomic regions considered, so it is unclear how to weight each contribution. In response to this, a number of model-based approaches leveraging epigenetic annotations have been proposed over the past decade, whether for genome-wide association studies (e.g., iBMU,[8] bfGWAS,[9] FINDOR[10]) or fine mapping (e.g., PAINTOR,[11] RiVIERA[12]).

Despite this extensive development, no existing method provides a solution to our problem, namely, modeling the functional enrichment of *trans*-QTLs and hotspots, a task which is substantially more complex and elusive than for the functional enrichment of *cis*-QTLs or GWA signals for a series of related phenotypes. All available modeling tools are designed for genetic mapping with one[8–11] or a few[12] traits at a time, while *trans*-QTL and hotspot mapping requires considering thousands of traits simultaneously.

It is also worth noting that many approaches accommodate only small numbers of candidate annotations by computational or statistical stability constraints,[8,9] or take as input GWA summary statistics rather than individual-level data, thereby not benefiting from the added statistical power obtained from jointly modeling the latter, along with the functional information.[10–12]

Our work enables large-scale inference for *cis*- and *trans*-QTL regulation using whole panels of external epigenetic annotations and argues that the epigenome can serve both to increase statistical power for QTL mapping and to shed light on the biology underlying the uncovered genetic map in a systematic manner. Specifically, it couples a fully Bayesian QTL mapping strategy, in which all loci and molecular traits are analyzed jointly, with a principled leveraging of epigenetic information by treating this information as complementary predictor-level data that may inform the probability of genetic variants to be involved in QTL associations. As successfully demonstrated in the context of genetic mapping with clinical traits, suitable use of epigenetic information can boost the detection of weak associations and help in discriminating genuine signals from spurious ones caused by LD or other confounding factors.[13,14]

Our modeling framework, called EPISPOT, directly infers the role of sparse sets of annotations—from hundreds of candidate functional annotations—in the activation of both *cis* and *trans* mechanisms affecting hundreds to thousands of molecular traits. Importantly, it combines this epigenome-driven feature with a flexible hotspot modeling feature inspired from our previous work,[5] thereby offering a unified toolkit to refine the detection of hotspots, aided by the epigenetic information at hand. The base version of EPISPOT assesses the action of the annotations uniformly for the full set of analyzed transcripts. However, for cases where a sensible partition into subsets of co-expressed molecular traits (modules[15]) is available, we also develop a module version of EPISPOT, which accommodates module-specific epigenetic action by estimating the contribution of the epigenetic marks to the QTL associations in each module.

Our take is that fully joint modeling is paramount to borrow information across loci, epigenetic marks, and molecular traits with complex dependences, but this requires careful algorithmic considerations to ensure scalable inference while retaining accuracy. EPISPOT implements an adaptive and parallel variational expectation-maximization (VBEM) algorithm, augmented with a simulated annealing scheme which effectively explores the multimodal parameter spaces induced by highly structured data. This optimization routine is purposely tailored to the analysis of genetic data with strong LD blocks, for which the inclusion of the epigenetic data has the greatest impact.

Our framework also constitutes an effective tool for interpreting (1) the detected *trans*-acting and hotspot variants based on their overlap with the selected epigenetic marks and (2) the molecular traits under genetic control in light of these marks. This additional purpose of EPISPOT is key given that elucidating the mechanisms of action of hotspots is often as challenging as mapping them in the first place. Indeed, there is accumulating evidence that most genetic variants acting in *trans* lie in intergenic regions,[16–18] where functional roles are difficult to decipher. Moreover, the massive *trans*-gene networks under genetic control are thought to be subject to subtle interplays, and researchers are often left with a variety of possible strategies to try to understand the interacting pathways between the genotype and underlying disease endpoints.[19] These strategies range from hypothesis-driven bottom-up approaches that start from isolated mechanisms and try to generalize them (e.g., based on *cis*-mediation hypotheses) to agnostic top-down approaches that directly model the whole system in view of teasing apart its fundamental components (e.g., based on graphical modeling approaches).[20] Our approach provides an alternative anchor toward decoding the complex networks controlled by hotspots, namely via the epigenetic marks found to be informative for the genetic mapping.

EPISPOT is not targeted at genome-wide discovery but at effecting refined QTL mapping and hotspot prioritization, based on genomic regions—hereafter called candidate loci—harboring SNPs thought to be involved in QTL regulation. A crucial distinction with the existing enrichment approaches is that the candidate loci do not correspond to a previously determined list of QTL hits but are whole genomic regions, which can involve hundreds of genetic variants (most of them with no QTL activity). EPISPOT exploits shared epigenetic signals across these regions to then select QTL hits with an increased statistical power.

Importantly, fruitful applications of EPISPOT, which can successfully decipher part of the molecular regulation machinery, require problems where the signal-to-noise and density of epigenetic/QTL signals are sufficient. In this work, we will describe extensive simulation experiments to highlight the benefits of using epigenetic information when available for a panel of regulation scenarios, and we will question the conditions under which inference is adequately powered to leverage this information. We will therefore formulate guidelines for practical use and

Please cite this article in press as: Ruffieux et al., EPISPOT: An epigenome-driven approach for detecting and interpreting hotspots in molecular QTL studies, The American Journal of Human Genetics (2021), https://doi.org/10.1016/j.ajhg.2021.04.010

provide a software implementation of EPISPOT along with documented code for the data-generation procedure used in the simulation experiments.

Another key component of the present paper concerns illustrating and exploiting the advantages of EPISPOT in real molecular QTL conditions. We will conduct and discuss the findings of a thorough monocyte expression QTL (eQTL) study leveraging a panel of annotations, including DNase-I sensitivity sites identified in different tissues and cell types, Ensembl gene annotations, and chromatin state data from ENCODE. In particular, by pinpointing context-relevant marks in a hypothesis-free manner, EPISPOT will allow us to disentangle key mechanisms pertaining to the lysozyme pleiotropic activity of chromosome 12—an activity which, although reported in several studies, is so far left unexplained in terms of its functional and mediation processes. Obtaining such evidence without EPISPOT would involve the daunting task of evaluating the enrichment of candidate eQTL hits in each individual epigenetic mark; this would also have no guarantee of success since one-at-a-time inspection strategies are deprived of the enhanced statistical power obtained with a unified joint epigenome/QTL mapping strategy.

Material and methods

Two-level hierarchical regression model

We consider a Bayesian model linking three data sources (Figure 1A) with two levels of hierarchy. The bottom level parametrizes the QTL effects and the top level parametrizes the epigenetic modulations of the primary QTL effects.

Specifically, the bottom level hierarchy uses a series of conditionally independent spike-and-slab regressions to model the regulation of q molecular traits by p candidate genetic variants or single-nucleotide polymorphisms (SNPs) for n samples:

$$\begin{aligned}&\boldsymbol{y}_t|\boldsymbol{\beta}_t,\tau_t \sim \mathrm{N}_n(\mathbf{X}\boldsymbol{\beta}_t,\tau_t^{-1}\mathbf{I}_n), \quad t=1,\ldots,q,\\ &\beta_{st}|\gamma_{st},\sigma^2,\tau_t \sim \gamma_{st}\,\mathrm{N}(0,\sigma^2\,\tau_t^{-1})+(1-\gamma_{st})\delta_0, \quad s=1,\ldots,p,\end{aligned} \qquad \text{(Equation 1)}$$

where $\boldsymbol{y}=(\boldsymbol{y}_1,\ldots,\boldsymbol{y}_q)$ is an $n\times q$ matrix of centered responses (molecular traits) and $\boldsymbol{X}=(\boldsymbol{X}_1,\ldots,\boldsymbol{X}_p)$ is an $n\times p$ matrix of centered candidate predictors for them (SNPs). Here, δ_0 is the Dirac distribution and to each regression parameter β_{st} corresponds a binary latent parameter γ_{st} taking value 1 if and only if SNP s is associated with trait t. Taking the posterior means of the latent parameters γ_{st} then yields marginal posterior probabilities of inclusion (qtl-PPIs, Figure 1C), $\mathrm{pr}(\gamma_{st}=1|\boldsymbol{y})$, from which Bayesian false discovery rate (FDR) estimates can be obtained. Moreover, the precision parameters τ_t and σ^{-2} are assigned diffused Gamma priors.

The top-level hierarchy parametrizes the effects of the epigenetic marks on the QTL probability of association via a second-stage probit regression on the probability of effects:

$$\begin{aligned}&\gamma_{st}|\theta_s,\zeta_t,\boldsymbol{\xi} \sim \mathrm{Bernoulli}\{\Phi(\zeta_t+\theta_s+\boldsymbol{V}_s^T\boldsymbol{\xi})\},\\ &\theta_s \sim \mathrm{N}(0,s_{0s}^2), \quad \zeta_t \sim \mathrm{N}(n_0,t_0^2),\\ &\xi_l|\rho_l \sim \rho_l\,\mathrm{N}(0,s^2)+(1-\rho_l)\delta_0,\\ &\rho_l \sim \mathrm{Bernoulli}(\omega_l), \quad l=1,\ldots,r,\end{aligned} \qquad \text{(Equation 2)}$$

where $\Phi(\cdot)$ is the standard normal cumulative distribution function and $\boldsymbol{V}=(\boldsymbol{V}_1,\ldots,\boldsymbol{V}_r)$ is a $p\times r$ matrix of (centered) predictor-level covariates (epigenetic marks). The epigenetic marks therefore represent external annotations that directly annotate the SNPs, rather than sample-specific annotations.

Although prior information on the relevance of the marks for the QTL control can be accommodated if desirable, this is not required, as the use of a sparse prior on the mark effects $\boldsymbol{\xi}$ allows incorporating a large number of marks even though only a fraction may be responsible for genetic activity. In particular, if none of the marks are relevant, the QTL mapping will not suffer any bias from modeling the candidate marks (see simulation studies hereafter). Moreover, similarly as for the QTL effects, mark selection is easily achieved using posterior probabilities of inclusion, $\mathrm{pr}(\rho_l=1|\boldsymbol{y})$, corresponding to the posterior means of the binary latent inclusion indicators ρ_l (epi-PPIs, Figure 1C). This typically yields a sparse subset of marks, whose biological interpretation may help in understanding the mechanisms of action of the SNPs involved in the QTL associations.

A parametrization tailored to the detection of hotspots

In addition to embedding the predictor-level regression for the epigenetic effects, the top-level probit model in Equation 2 also decouples the contributions of the predictors (SNPs) and the responses (molecular traits), namely, by involving a response-specific parameter, ζ_t, which adapts to the sparsity level linked with each response $\boldsymbol{y}_t$ and a predictor-specific parameter, θ_s, which encodes modulations of the probability of association according to the overall effect of each predictor $\boldsymbol{X}_s$. Parameter θ_s has a central role in pleiotropic molecular QTL settings as it represents the propensity of each predictor to be associated with multiple responses, i.e., its propensity to be a hotspot. Its Gaussian prior specification ensures closed-form updates, which is critical to the efficiency of the algorithm on large datasets. It also conveniently permits using a local-scale representation (via s_{0s}) to prevent overshrinkage of large hotspot signals; see our previous work on the hierarchical modeling of hotspots, from which this formulation is borrowed.[5]

Here, the value of s_{0s} is set by empirical Bayes, and so are the epigenetic effect hyperparameters ω_l and s. The values of the hyperparameters n_0 and t_0 are chosen to induce sparsity, by specifying a prior expectation and a prior variance for the number of predictors associated with each response (supplemental material and methods).

Hence, the EPISPOT model (Equations 1 and 2) borrows information across the three types of entities (epigenetic marks, SNPs, and molecular traits) in a unified manner, while providing interpretable posterior quantities, in particular qtl-PPIs and epi-PPIs, for the selection of each type of variable. It leverages the epigenome for two complementary purposes: (1) to enhance statistical power for QTL and hotspot mapping and (2) to shed light on the biology underlying the genetic control, via the inspection of the selected marks.

A modification for module-specific epigenetic contributions

The machinery of genetic control is complex and it is unlikely that the action of the epigenome on QTL regulation will uniformly affect the transcriptome. In particular, different groups of molecular traits may be governed by different functional mechanisms, involving different sets of epigenetic marks, to different degrees. When a partition into modules of genes (proteins or metabolites for pQTL or

Please cite this article in press as: Ruffieux et al., EPISPOT: An epigenome-driven approach for detecting and interpreting hotspots in molecular QTL studies, The American Journal of Human Genetics (2021), https://doi.org/10.1016/j.ajhg.2021.04.010

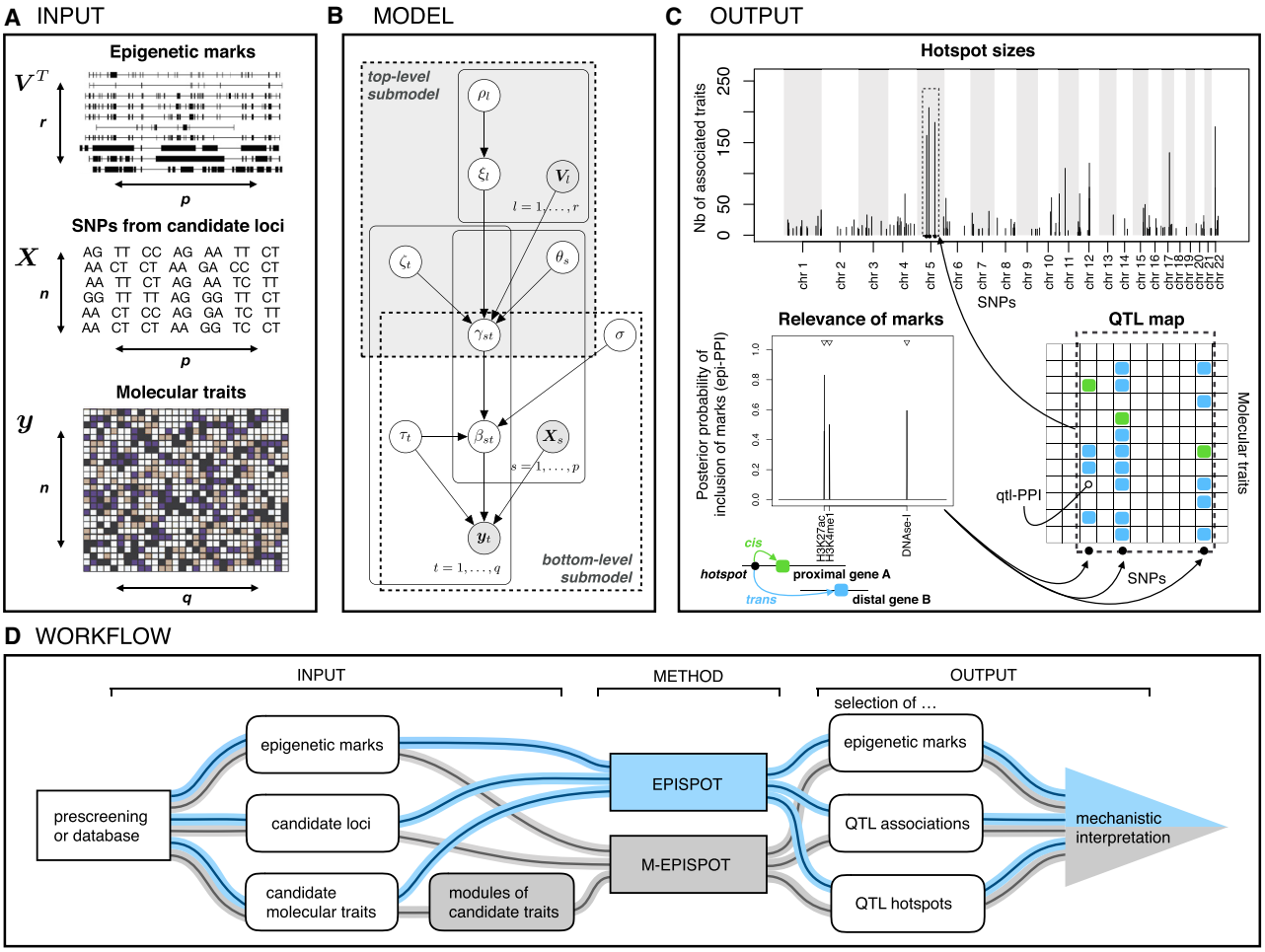

Figure 1. Overview of EPISPOT

(A) Data input. Epigenetic annotations (predictor-level information) $\boldsymbol{V}$, genetic variants from candidate loci (candidate predictors) $\boldsymbol{X}$, molecular traits (responses) $\boldsymbol{y}$.

(B) Graphical representation for the two-level hierarchical model. The shaded nodes are observed, and the others are inferred. The top-level regression corresponds to the top plate; the probability of association is decoupled into a trait-specific contribution, ζ_t, a SNP-specific contribution with a "hotspot propensity parameter" θ_s and an epigenome-specific contribution, ξ_l, where $\boldsymbol{V}_l$ is the vector gathering the observations of predictor-level epigenetic covariate l for all candidate SNP predictors $\boldsymbol{X}_s$, $s = 1,\ldots,p$. Parameter β_{st} models the effect between SNP $\boldsymbol{X}_s$ and trait $\boldsymbol{y}_t$, and γ_{st} and ρ_l are binary latent indicators for the QTL associations and epigenetic mark involvement, respectively. Parameter σ models the typical size of QTL effects and τ_t^{-1} models the residual variability of trait $\boldsymbol{y}_t$.

(C) Posterior output. Selection of epigenetic marks with a role in QTL regulation is carried out using the posterior probabilities of inclusion (epi-PPIs), $\mathrm{pr}(\rho_l = 1|\boldsymbol{y})$, $l = 1,\ldots,r$ (bottom left) and selection of associated SNP-trait pairs (aided by the marks) is carried out using the posterior probabilities of inclusion (qtl-PPIs), $\mathrm{pr}(\gamma_{st} = 1|\boldsymbol{y})$, $s = 1,\ldots,p; t = 1,\ldots,q$ (bottom right). The hotspot Manhattan plot (top) reports the number of traits associated with each SNP ("hotspot size"), after using a selection threshold on the qtl-PPIs (e.g., FDR-based).

(D) EPISPOT workflow. Candidate loci and molecular traits are obtained from a preliminary screening or from existing databases and supplied as input to the method along with epigenetic marks at the variants harbored by the loci. The algorithm is used with or without the module option depending on whether the traits are gathered into modules or not (M-EPISPOT in gray, resp. EPISPOT in blue). The output consists of sets of associated variants and traits, QTL hotspots, and epigenetic marks relevant to the primary QTL associations for given significance thresholds. It is then interpreted to generate mechanistic hypotheses about the functional processes underpinning the QTL associations.

mQTL analyses, respectively) likely to be co-regulated is available to the analyst, it can be provided as input to the method which will then infer the annotation effects in a module-specific fashion, based on the following modification of top-level Equation 2:

$$\begin{aligned}
&\gamma_{st}|\theta_{m,s},\zeta_t,\boldsymbol{\xi}_m \sim \text{Bernoulli}\{\Phi(\zeta_t + \theta_{m,s} + \boldsymbol{V}_s^T\boldsymbol{\xi}_m)\},\\
&\theta_{m,s} \sim \mathrm{N}\left(0, s^2_{0m,s}\right), \quad \zeta_t \sim \mathrm{N}(n_0, t_0^2),\\
&\xi_{m,l}|\rho_{m,l} \sim \rho_{m,l}\,\mathrm{N}(0, s_m^2) + (1-\rho_{m,l})\delta_0,\\
&\rho_{m,l} \sim \text{Bernoulli}(\omega_{m,l}), \quad l = 1,\ldots,r,
\end{aligned} \qquad \text{(Equation 3)}$$

where $m \in \mathcal{M}$ is a module of traits, with $\mathcal{M}$ a partition of $\{1,\ldots,q\}$ and $m \ni t$. Parameter ξ_m then represents the epigenetic contribution of the r marks for the QTL associations involving the traits from module m. The hotspot parameter $\theta_{m,s}$ also accounts for the module structure: it represents the propensity of SNP s to be associated with few or many traits from module m. This encodes module-specific pleiotropic levels and also reflects the fact that a SNP controlling a given trait in a module is more likely to be also associated with related traits from the same module compared to traits outside the module.

The corresponding version of the algorithm—implementing Equations 1 and 3—is hereafter called M-EPISPOT when an explicit distinction with the base, module-free version—implementing Equations 1 and 2—is needed.

Different approaches, based on some prior state of knowledge, on specific optimization methods, or both, will typically yield complementary definitions of modules. In some instances, there will be obvious biological reasons backing up the obtained grouping; in others, no clear partitioning will emerge, in which case the analyst may choose to use the module-free version of the model. As there is no generic strategy for forming modules, it is important to understand the impact of such choices on inference. In particular, from a modeling point of view, a given module should ideally comprise co-regulated molecular traits, i.e., traits with shared genetic control, triggered by common epigenetic mechanisms. The top-level regression (Equation 3) will then represent the possible epigenetic effects underlying the functional mechanisms in the module, and module-specific epi-PPIs will be useful to select the marks involved in the regulation of each module. In particular, shared signals will be best leveraged when the molecular traits controlled by a given SNP belong to a same module. The simulation studies and the eQTL analysis will provide practical recommendations as well as analyses of sensitivity to module misspecification.

A scalable purpose-built algorithm

The hierarchical model described above couples two levels of spike-and-slab regression, which accommodate three large spaces of SNPs, molecular traits and epigenetic marks, with possibly thousands of variables each. Careful algorithmic strategies are therefore critical to ensure that inference is accurate and scalable. To meet both requirements, we implement an adaptive variational expectation-maximization (VBEM) algorithm and augment it with a simulated annealing procedure that efficiently explores the highly multimodal variable spaces formed by data with strong dependence structures.

VBEM algorithms were introduced by Blei et al.[21] in the context of Dirichlet allocation modeling. In short, they iterate between optimizing empirical Bayes estimates (in our case for the hotspot propensity and epigenetic effect hyperparameters) and running a variational algorithm for the remaining parameters, given the updated empirical Bayes estimates.

We present hereafter the algorithm in its general module-based form (M-EPISPOT); omitting the index m and taking $M = 1$ gives the base version with no module partitioning (EPISPOT).

Let $\boldsymbol{\nu} = (\boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\gamma}, \sigma^2, \boldsymbol{\theta}, \boldsymbol{\zeta}, \boldsymbol{\xi}, \boldsymbol{\rho})$ denote the parameters for Equations 1 and 3, and let $\boldsymbol{\eta} = (\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_M)$ denote the second-stage model hyperparameters, with $\boldsymbol{\eta}_m = (s^2_{0m}, s^2_m, \boldsymbol{\omega}_m)$ for module $m = 1, \ldots, M$. We propose estimating $\boldsymbol{\eta}$ via an empirical Bayes procedure, by finding

$$\widehat{\boldsymbol{\eta}} = \arg\max_{\boldsymbol{\eta}} \ell(\boldsymbol{\eta}; \boldsymbol{\gamma}), \qquad \text{(Equation 4)}$$

where $\ell(\boldsymbol{\eta}; \boldsymbol{\gamma}) = \log p(\boldsymbol{\gamma}|\boldsymbol{\eta})$ is the marginal log-likelihood. Computing Equation 4 analytically for our model would require high-dimensional integration and thus is infeasible. Our VBEM algorithm circumvents this by coupling the empirical Bayes estimation of the hyperparameter $\boldsymbol{\eta}$ with a variational inference scheme that simultaneously infers the model parameter vector $\boldsymbol{\nu}$. The procedure implements alternating optimizations of the variational lower bound

$$\mathcal{L}(q; \boldsymbol{\eta}) = \mathrm{E}_q \log p(\boldsymbol{\gamma}, \boldsymbol{\nu}|\boldsymbol{\eta}) - \mathrm{E}_q \log q(\boldsymbol{\nu}), \qquad \text{(Equation 5)}$$

where $q(\boldsymbol{\nu})$ is the variational density for $p(\boldsymbol{\nu}|\boldsymbol{\gamma}, \widehat{\boldsymbol{\eta}})$ for a current estimate $\widehat{\boldsymbol{\eta}}$ and $\mathrm{E}_q(\cdot)$ is the expectation with respect to $q(\boldsymbol{\nu})$. More precisely, it initializes the parameter and hyperparameter vectors $\boldsymbol{\nu}^{(0)}$ and $\boldsymbol{\eta}^{(0)}$, and alternates between the E-step,

$$q^{(t)} = \arg\max_q \mathcal{L}(q; \boldsymbol{\eta}^{(t-1)}),$$

using the variational algorithm for obtaining $q^{(t)}$ at iteration t, and the M-step,

$$\boldsymbol{\eta}^{(t)} = \arg\max_{\boldsymbol{\eta}} \mathcal{L}(q^{(t)}; \boldsymbol{\eta}),$$

until convergence of $\boldsymbol{\eta}^{(t)}$. In our case, the updates for the M-step are obtained analytically by setting to zero the first derivative of $\mathcal{L}(q^{(t)}; \boldsymbol{\eta})$ with respect to each component of $\boldsymbol{\eta}$. This only requires computing and differentiating the joint likelihood term $\mathrm{E}_q \log p(\boldsymbol{\gamma}, \boldsymbol{\nu}|\boldsymbol{\eta})$ in Equation 5, as the entropy term $-\mathrm{E}_q \log q(\boldsymbol{\nu})$ is a function of $\boldsymbol{\eta}^{(t-1)}$ and is constant with respect to $\boldsymbol{\eta}$.

Variational inference is typically orders of magnitude faster than classical Markov chain Monte Carlo inference[5,6,22] for comparisons on GWA and molecular QTL models. Some computational cost is added for VBEM algorithms as each E-step requires running the variational algorithm until convergence. Moreover, the two regression levels of our Equations 1 and 2 or Equations 1 and 3 necessitate the exploration of a very large parameter space, which is complex and time consuming for any type of inference.

We consider two strategies to overcome this burden. First, we substantially reduce the runtime of the within-EM variational runs by using an adaptive stopping criterion, namely, starting with a large tolerance and dynamically decreasing it according to the convergence state of the overall EM algorithm. The second strategy applies to the module version of our algorithm: the specification in Equation 3 suggests that its hyperparameters may be estimated reasonably well by restricting the VBEM scheme to subproblems corresponding to each module, i.e., applying Equations 1 and 2 to the subsets of responses $\boldsymbol{\gamma}_m$ separately for obtaining the corresponding empirical Bayes estimates $\boldsymbol{\eta}_m$, $m = 1, \ldots, M$. In addition to accelerating hyperparameter estimation for each module (as the model is much smaller), this has the advantage of allowing parallelization across modules. Once all module hyperparameters are estimated, they are inserted into Equations 1and 3 and variational inference is run on the entire dataset.

Strong posterior multimodality can be induced by dense genotyping panels with marked LD structures, whereby the inclusion of epigenetic information is particularly beneficial to disentangle the genetic contributions. To robustly infer signals from problems with strong data dependence structures, we augment all variational schemes with a simulated annealing routine.[23,24] Annealing introduces a so-called temperature parameter to index the variational distributions and control the level of separation between their modes, thereby easing the progression to the global optimum. In practice, we start with a temperature T_0 to flatten the posterior distribution and sweep most local modes away, and we then lower it at each iteration, until the original multimodal distribution, called the cold distribution, is reached. Finally, to ensure stable inference, our routine excludes redundant SNPs and marks (i.e., displaying perfect collinearity with other SNPs/marks) prior to the run. Moreover, constant marks or marks that concern less than a given proportion of SNPs (default 5%) are also discarded before the analysis as insufficiently informative.

A sketch of the algorithm and the full derivation of the annealed VBEM updates are in the supplemental material and methods. The

algorithm is implemented as a publicly available R package with C++ subroutines (see Web Resources). Both the EPISPOT and M-EPISPOT versions run within seconds to few hours depending on the numbers of loci, molecular traits, and epigenetic marks (see the runtime profiling in the supplemental material and methods). We also provide simulation studies that demonstrate the robustness of EPISPOT to different degrees of LD and the benefit of coupling VBEM inference with simulated annealing in case of strong LD (supplemental material and methods).

Recommended use

EPISPOT is a refining tool for the detection and interpretation of QTL and hotspot effects. It is meant to be used for joint analysis of preselected genomic regions (candidate loci) and transcripts believed to be under genetic control (Figure 1D). Different approaches can be considered to obtain loci of interest. Public databases can be employed to form loci of given size around previously identified hits, provided this information is available for the condition, tissue, or cell type at hand. An alternative approach is based on a preliminary application of ATLASQTL[5] or another screening method, ideally on an independent dataset. If no independent dataset is available to the analyst, useful research hypotheses may still be obtained by running the prescreening step on the same dataset, prior to running EPISPOT. However, results should then be considered as exploratory, since this procedure interrogates the same data twice, which is subject to overfitting.

The effectiveness of EPISPOT for detecting and exploiting the relevant epigenetic marks for QTL mapping depends on multiple conditions that have a coordinated effect on statistical power. The number of loci analyzed must be reasonably large to hope for the marks to be sufficiently represented at causal loci. These loci must also be densely genotyped or imputed to ensure that the causal SNPs, and the epigenetic marks they may fall into, are included in the analysis. The frequency of each relevant mark among causal SNPs, as well as the strength of its contribution to initiating the QTL effects and the quality of the mark annotation also play a role, as do the degree of co-regulation of traits by the same SNPs, the sample size of the analyzed dataset, the individual effect sizes of QTL associations, and the correlation structures among marks, traits, and SNPs (LD). We examine the impact of these different parameters in a series of simulation studies described in the results and in the supplemental material and methods.

Were these conditions not sufficiently met for EPISPOT to borrow information across the loci and learn the mark contributions, the QTL mapping would not benefit from further level of information provided by the marks (no mark selected) but it would nonetheless benefit from the joint analysis of SNPs and traits. Notably, the sparse modeling of the marks implies that the inclusion of marks, were these insufficiently informative, has no risk of deteriorating the QTL mapping (see the "Null scenario" section in the supplemental material and methods); this is a major advantage of our method.

Results

Data generation and simulation set-up

The series of simulation studies presented in the next sections have the dual purpose of (1) illustrating the effectiveness of EPISPOT in learning from the epigenome when the epigenetic annotations at hand are sufficiently informative (first simulation study), and (2) evaluating the method in weakly informative scenarios (second simulation study) or scenarios where the module partition supplied to M-EPISPOT is misspecified (third simulation study).

We simulate data so as to best emulate molecular QTL regulation and the role of the epigenome in triggering this regulation; the general data-generation procedure is detailed in the supplemental material and methods and we further tailor it to each simulation experiment in their dedicated sections. Here, for simplicity, we represent the presence or absence of a mark at each SNP using a binary variable. In real case scenarios, all types of continuous annotations can be considered without modification since they are encoded as predictors in the second-level regression framework employed by EPISPOT, hence with no distributional assumption.

We use the following terminology when referring to the simulated association patterns:

- an "active SNP" has at least one association with a molecular trait
- an "active locus" involves at least one active SNP
- an "active trait" has at least one association with a SNP
- an "active module" contains at least one trait involved in QTL associations
- an "active mark" triggers at least one SNP-trait QTL association
- the "hotspot size" is the number of traits associated with a given hotspot SNP.

We benchmark our approach against two representative state-of-the-art methods for QTL mapping, namely, the fully joint Bayesian QTL method ATLASQTL,[5] which is also tailored to the modeling of hotspots but does not accommodate the epigenetic marks, and the widely used marginal screening approach MATRIXEQTL,[2] which tests each SNP-trait pair one-by-one and does not involve any epigenetic information.

A first illustration

We first describe the type of posterior output produced by EPISPOT and its performance in a simple problem where no modules are involved, i.e., the active epigenetic marks exert their influence on all associated SNP-trait pairs.

We simulate 32 datasets with an average of 600 molecular traits, $r = 500$ candidate epigenetic marks and 60 candidate loci, each comprising an average of 20 real SNPs for 413 subjects. These are initial choices are meant to reflect plausible scenarios encountered in real applications, after preselecting candidate loci and candidate traits likely to be controlled by these loci. A subset of 100 SNPs are active (between 0 and 3 per locus; see Table 1) and their QTL effects are triggered by $r_0 = 3$ active marks. This is a strong assumption, which permits a direct illustration of our algorithm in a simple setting but, since it may be unrealistic, we will only use it as a starting point for the more complex

Table 1. Average number of simulated loci stratified by the number of active SNPs in the first simulation study

Total number of loci	60
Inactive loci	9.1 (2.7)
Loci with 1 active SNP	17.6 (3.9)
Loci with 2 active SNPs	17.6 (2.6)
Loci with 3 active SNPs	15.8 (2.2)

Standard deviations are in parentheses (32 simulated datasets).

numerical experiments that follow. To help interpretability in the context of the simulations, we also generate marks with positive effects only, i.e., inducing QTL activity and not repressing it (supplemental material and methods). Moreover, the large number of candidate marks and the low number of active marks are used to illustrate the ability of EPISPOT to discriminate sparse subsets of relevant marks from whole panels of marks (most of which with no contribution to the QTL effects). The QTL signals are relatively weak: for any given trait, the cumulated QTL effects are responsible for at most 25% of its total variance. Many active SNPs are hotspots; across all 32 replicates, the active SNPs are associated with a number of traits ranging from 1 (isolated QTL association) to 96 (large hotspot), with an average of 27 active traits per active SNP.

All these choices will be varied in the subsequent simulation experiments; for an extensive comparison over a grid of scenarios, see the supplemental material and methods.

Figure 2 shows that EPISPOT could clearly discriminate the three active marks contributing to the QTL associations from the remaining $r - r_0 = 497$ inactive marks. The partial receiver operating characteristic (ROC) curves also show that it outperforms ATLASQTL in terms of selecting associated SNP-trait pairs and hotspots. It is unsurprising given that ATLASQTL does not use any predictor-level information, yet it nevertheless confirms that EPISPOT can effectively exploit the marks to enhance the estimation of the primary QTL associations. MATRIXEQTL performs poorly compared to the two joint approaches EPISPOT and ATLASQTL, which is expected since, by design, it does not exploit the shared association signals across traits.

We checked that EPISPOT and ATLASQTL display similar performance under simulation scenarios with no active mark: their 95% confidence intervals for the standardized partial area under the curve (pAUC) overlap, i.e., (0.74, 0.78) and (0.76, 0.79) for ATLASQTL, resp. EPISPOT (supplemental material and methods). This further supports the observation that the improvement of EPISPOT seen in Figure 2 is attributable to an effective use of the three informative marks and not to other intrinsic differences between the two models; more evidence on this is provided in the next simulation experiment.

Performance under varying degrees of epigenome involvement

Effectiveness in QTL mapping is subject to a number of interdependent factors pertaining to (1) the sparsity of the studied QTL network and magnitude of the QTL effects, (2) the amount of information contained in the data at hand, and (3) the ability of the statistical approach to interrogate the data, i.e., by both leveraging and being robust to the dependence structures within and across genetic variants and molecular traits. When it comes to exploiting the epigenome to enhance statistical power, an additional level of complexity is introduced for determining the impact of the above factors on the analysis, and new questions arise as to whether the signal present in the data is sufficient to inform inference on the location of the relevant epigenetic marks and of the QTL associations potentially triggered by these marks.

In the previous simulation experiment, we generated data under the simplifying assumption that all QTL associations were induced by the epigenome, and to a degree to which the relevant marks would be detectable, as evidenced by the high epi-PPIs for the active marks and the power gained from leveraging this signal (Figure 2). Here, we focus on evaluating how the level of involvement of the epigenome in QTL activity impacts the detection of QTL effects and of the marks responsible for these effects.

We consider a series of QTL problems, each generated by replicates of 32, for a grid of response numbers and degrees of involvement of the epigenome in activating QTL control. More precisely, we simulate data with a number of traits sampled from a Poisson distribution with mean $\lambda = 200, 400, 600, 800, 1000$, or 1,600 and 60 loci with 20 SNPs each and involving 100 active SNPs in total. We vary the proportion of active SNPs whose activity is triggered by epigenetic marks from $p_{\text{epi}} = 0$ (all QTL associations simulated independently of the action of the epigenome) to $p_{\text{epi}} = 1$ (all QTL associations simulated as the result of the action of the epigenome); see the supplemental material and methods for the data-generation details. The typical pleiotropic pattern simulated is displayed in Figure 3 for the different choices of p_{epi} and problems with an average of $\lambda = 600$ traits.

Figure 3 also shows the performance for the selection of QTL effects in terms of standardized pAUC. It provides two separate layers of information: first, it illustrates again how EPISPOT is able to leverage the epigenetic marks to improve QTL mapping, and more so when the number of active SNPs triggered by these marks increases (top to bottom rows) since EPISPOT is then able to effectively borrow information across the mark-activated SNPs. This underlines the need for the relevant epigenetic marks to be sufficiently

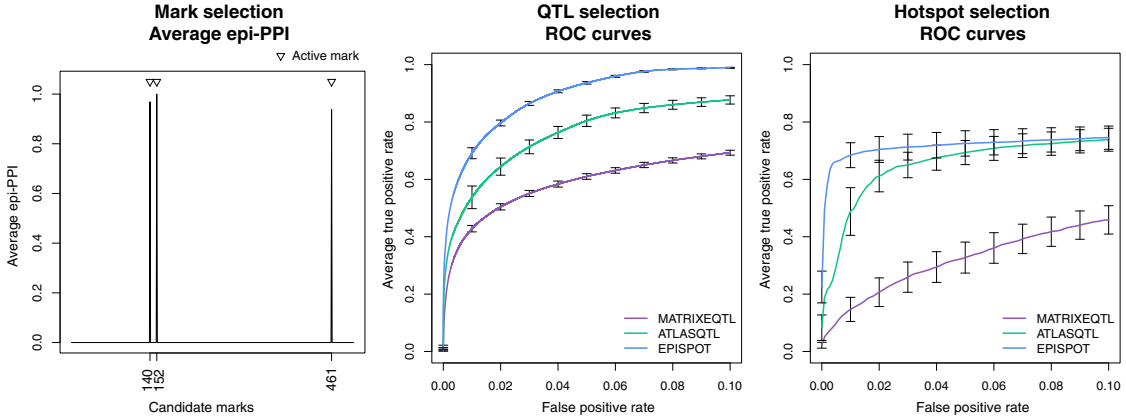

Figure 2. Performance for selection of epigenetic marks, pairs of associated SNPs and traits, and hotspots
Left: Epi-PPIs for the marks averaged over 32 replicates. The three marks simulated as active are indicated by the triangles. Middle: Average partial ROC curves for SNP-trait selection with 95% confidence intervals obtained from 32 replicates. EPISPOT is compared to the joint hotspot-QTL mapping method, ATLASQTL,[5] and the univariate screening method, MATRIXEQTL,[2] none of which makes use of the epigenetic marks. Right: idem for the selection of active SNPs (here, mainly hotspots).

represented at causal variants so that the analyzed data are informative about their involvement. It is therefore advised to use a reasonably large number of loci thought to be active and dense SNP panels (e.g., imputed SNPs, see the eQTL case study section), so the active SNPs are more likely to be included. Second, it shows that the joint modeling of all traits permits exploiting shared signals across these traits, thereby also improving statistical power, as reflected by the increased pAUCs for problems with larger numbers of traits in Figure 3. This is particularly true in the presence of co-regulated molecular traits, a special case of which is the regulation of these traits by a single hotspot.

Figure 3 also indicates that, when the epigenetic signal is moderate to large ($p_{\text{epi}} = 0.4$, 0.6, 0.8, or 1), EPISPOT is able to pick the active epigenetic marks from a large number of candidate marks, while setting the epi-PPIs of the inactive marks to zero. However, when the signal is weak ($p_{\text{epi}} = 0.2$), the active marks are barely detected, as expected. Importantly, though, in the null scenario where the epigenome plays no role ($p_{\text{epi}} = 0$), modeling the $r = 500$ inactive marks does not deteriorate the performance (supplemental material and methods).

These experiments also suggest that annotations which are more likely to trigger QTL associations at numerous causal SNPs, such as cell-type-specific enhancers, could have increased opportunities to be picked up and leveraged. This may imply that the QTL mapping would benefit more from the use of general annotations than from that of more specific types of marks, such as ChIP-seq binding sites of transcription factors, which may display a lower degree of sharing between hotspots. Further investigations on real datasets would need to confirm this. However, as there is no intrinsic limitation on the number of candidate annotations supplied to EPISPOT, nothing prevents the analyst from using both general and more specific annotations, and letting the model select the annotations which are sufficiently informative.

Finally, the quality of the mark annotation will have a similar impact on performance. We show in complementary simulations (supplemental material and methods) that EPISPOT will not take full advantage of the epigenome if the supplied annotations are of poor quality: the QTL mapping performance declines with the level of noise in the annotations, but EPISPOT remains superior to alternative approaches for which no annotation information is supplied.

We also tested the impact of other data scenarios on the ability of EPISPOT to detect and utilize the marks for improving QTL mapping. More precisely, we ran simulations for a grid of configurations, varying: the number of active SNPs, the average QTL effect sizes, the degree of co-regulation of the traits and the hotspot sizes; see section "QTL mapping performance for a grid of simulated data scenarios" of the supplemental material and methods. These experiments show that (1) these parameters have a coordinated effect on statistical power, and (2) thanks to its flexible hierarchical representation, EPISPOT is very effective at taking advantage of shared functional patterns, yielding a substantial mapping performance gain.

Inferring module-specific epigenetic action

The simulation experiments presented next focus on evaluating M-EPISPOT, i.e., the module version of the algorithm which models module-specific epigenetic effects. They illustrate how statistical power and interpretability are enhanced when the structure underlying epigenome-driven QTL associations is exploited. They also evaluate the robustness of inference when misspecified module partitions are supplied to M-EPISPOT. This is particularly important given the uncertainty that often surrounds the definition of modules, as reflected by fact that different co-expression inferential tools often produce different module specifications.

We start with a simple example involving 60 concatenated loci of average size 40 SNPs and two modules of 50

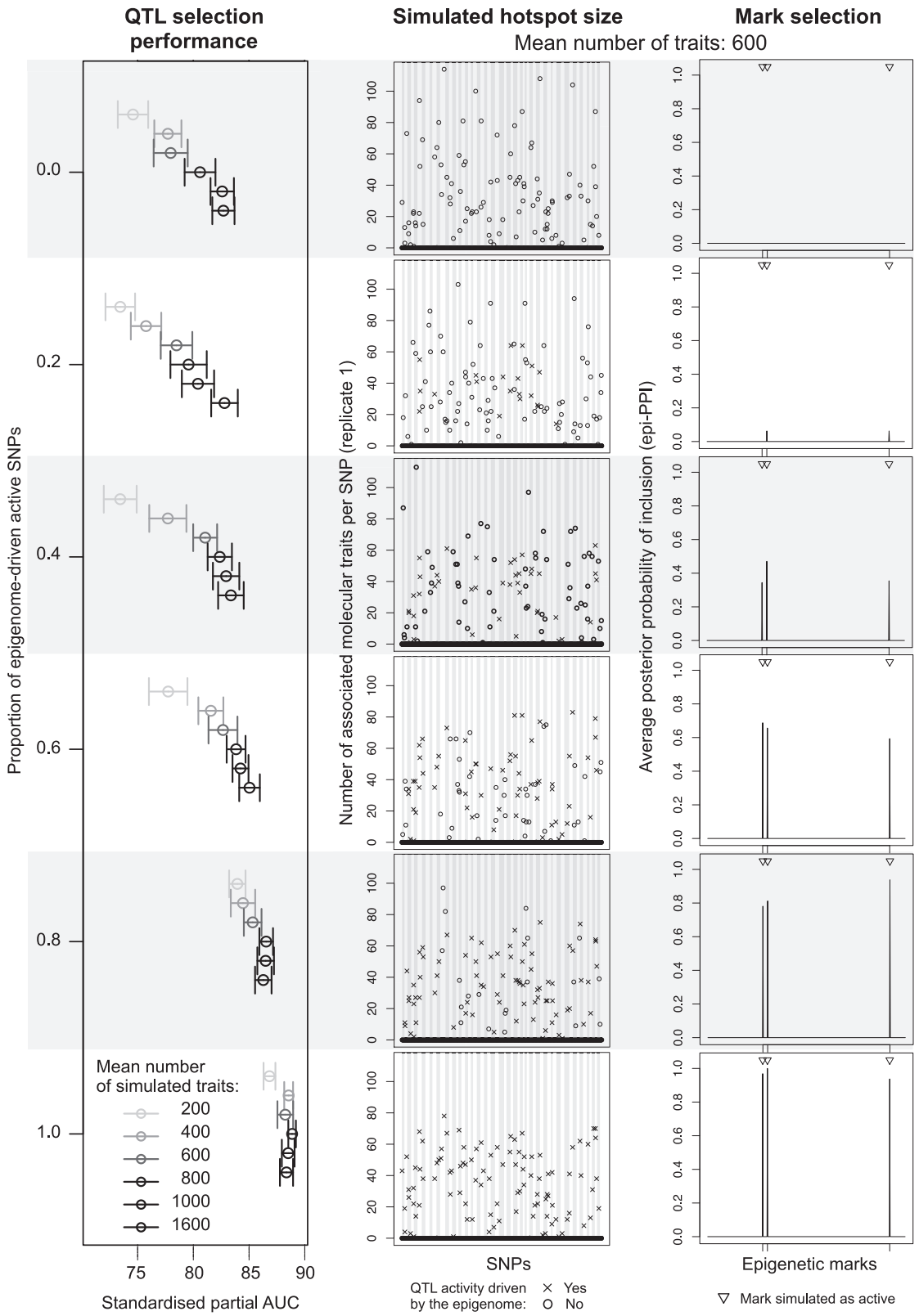

Figure 3. Performance of EPISPOT for a grid of numbers of traits and proportions p_{epi} of epigenome-driven active SNPs
Left: Standardized pAUCs for the QTL selection performance with 95% confidence intervals. Middle: Simulated hotspot QTL pattern for problems with an average of 600 traits (first replicate for each value of p_{epi}). The crosses indicate hotspots whose activity is triggered by the epigenome and the circles indicate hotspots whose activity is independent of the epigenome. Right: Average epi-PPIs, as inferred by EPISPOT for the simulated scenarios with an average of 600 traits.

Please cite this article in press as: Ruffieux et al., EPISPOT: An epigenome-driven approach for detecting and interpreting hotspots in molecular QTL studies, The American Journal of Human Genetics (2021), https://doi.org/10.1016/j.ajhg.2021.04.010

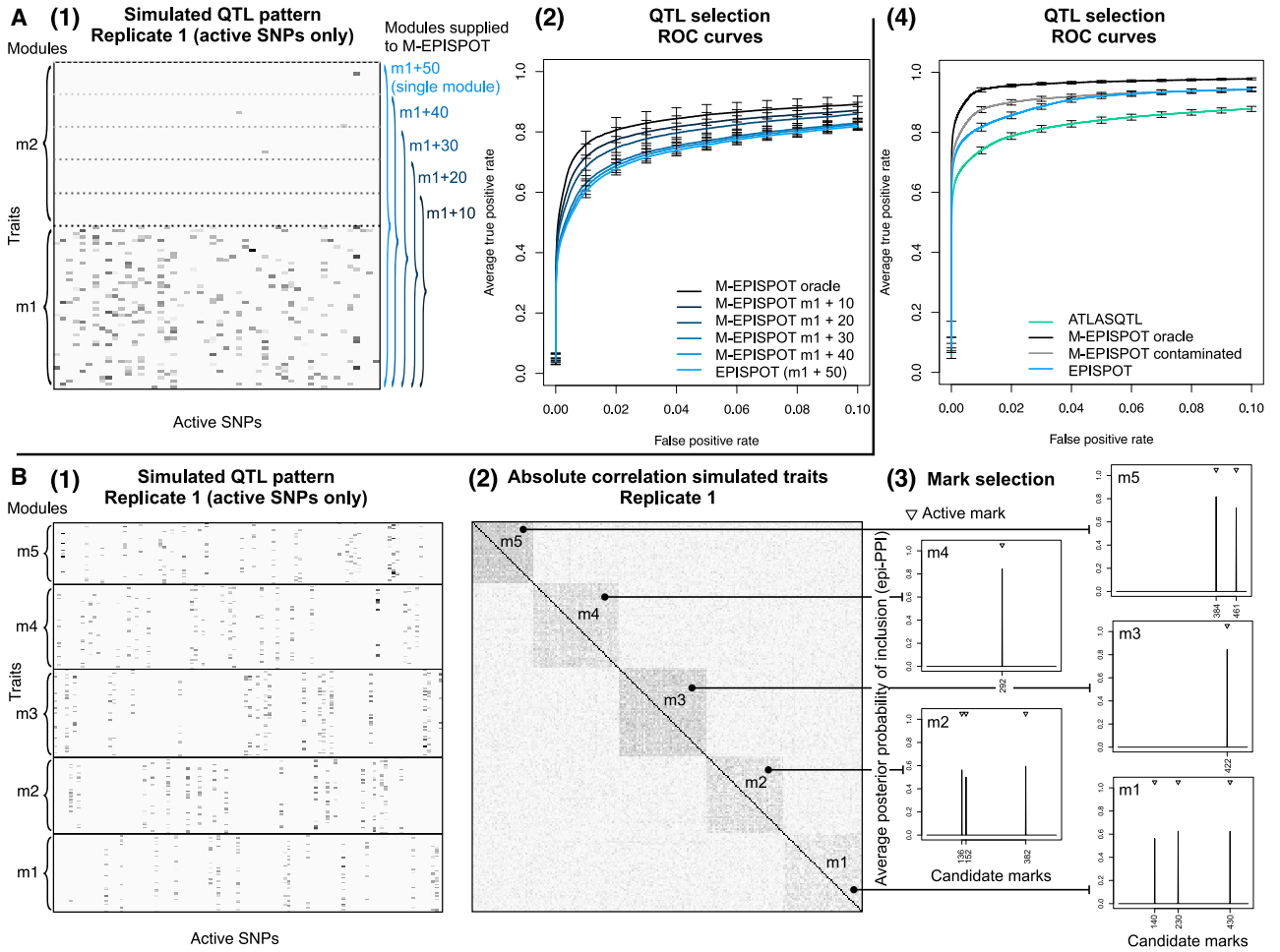

Figure 4. Performance of M-EPISPOT

(A) Simulated scenario with two modules, whereby the first module m_1 is contaminated by an increasing number of traits from the second module m_2. Panel A(1) shows the simulated pleiotropic pattern for one replicate. The gray levels suggest the different QTL effect strengths of each active SNP (x axis) with the traits (y axis) from modules m_1 and m_2. The horizontal dotted lines mark the boundary between m_1 and m_2 for the misspecified module partitions supplied to M-EPISPOT. Panel A(2) shows the partial ROC curves (with 95% confidence intervals based on 32 replicates) for the QTL mapping performance obtained when supplying the different misspecified partitions shown in A(1) to M-EPISPOT.

(B) Simulation with five pleiotropic modules. Panel B(1) shows the simulated pattern for the active SNPs of one replicate. Panel B(2) panel shows the dependence structure of the simulated traits for one replicate. Panel B(3) shows the module-specific average epi-PPIs for the contribution of the epigenetic marks to the QTL effects. Panel B(4) shows the partial ROC curves for the QTL mapping, with 95% confidence intervals based on 32 replicates.

simulated traits each. In the first module (m_1), the traits are largely co-regulated by hotspots whose activity is imputable to the epigenome. In the second module (m_2), only few traits are involved in isolated QTL associations, with no implication of the epigenome. Figure 4A illustrates the corresponding simulated QTL pattern restricted to the active SNPs, for the first data replicate. We evaluate the performance of M-EPISPOT with the following settings:

1. The oracle case, where we assume the simulated module partition $\mathcal{M} = \{m_1, m_2\}$ to be known and provided it as input to M-EPISPOT;
2. the module-free case, where we perform inference with the base model EPISPOT which does not exploit the module partition;
3. a series of intermediate cases, where the module partition supplied to M-EPISPOT is misspecified, i.e., module m_1 is contaminated with $10, 20, 30$ or 40 traits from module m_2 (Figure 4A). This mimics a real data scenario whereby the assignment of some traits to modules is difficult.

The ROC curves of Figure 4A show that leveraging information about the underlying module partition can improve significantly the detection of QTL effects. They also confirm the intuition that the impact of misspecified partitions on performance is a function of the degree of misspecification: for a given specificity, the power decreases smoothly with the number of inactive traits from module m_2 contaminating module m_1. From a modeling point of view, leaving all traits controlled by a same

hotspot in a single module permits maximizing the opportunities to learn the epigenetic contribution to the QTL activity by borrowing strength across co-regulated traits. It is advised to make use of prior information on pleiotropy when available in order to avoid splitting hotspot-controlled networks of traits into distinct modules.

The second simulation experiment considers a more general setting with 5 modules of average size 50. It compares ATLASQTL, EPISPOT, and M-EPISPOT with the oracle module partition supplied and M-EPISPOT with a contaminated module partition supplied, i.e., where a fifth of the traits in the simulated modules are randomly re-assigned to the other modules.

Figure 4B leads to a conclusion similar to that of the previous example: the idealized scenario of the oracle module partition provided to M-EPISPOT yields the best performance, followed, in order, by the more realistic case of the contaminated partition, the EPISPOT run (with no module information) and finally, the ATLASQTL run which does not make use of any epigenetic information. Importantly, the fact that the module-free version EPISPOT outperforms ATLASQTL indicates that even when the module structure is not employed, the method is still able to leverage the epigenome in order to improve the QTL mapping.

Figure 4B also shows how the marks responsible for the activation of the different modules are correctly recovered by M-EPISPOT. An inspection of these separate sets of marks provides a refined level of interpretability for a module-specific understanding of the genetic control. We will see in the eQTL analysis presented next how this can be particularly helpful to shed light on the mechanistic action of *trans* hotspots, when such hotspots are thought to control gene modules in a context-specific way.

An epigenome-driven monocyte eQTL case study

In this section, we take advantage of EPISPOT in a targeted eQTL study to refine the detection and characterization of genetic regulation in monocytes. Specifically, we analyze two independent datasets with transcript levels measured in $CD14^+$ monocytes. Our study workflow is described in Figure 5A: we discover active loci in a prescreening step using the joint hotspot QTL mapping approach ATLASQTL[5] in the first dataset (n = 413 samples[25]), and we then leverage the epigenome using EPISPOT in the second dataset (CEDAR cohort, n = 286 samples[26]) for an in-depth analysis of the genetic activity in the preselected loci.

The epigenetic information consists of a panel of 168 annotation variables, compiling DNase-I sensitivity sites from different tissues and cell types, Ensembl gene annotations, and chromatin state data from ENCODE. These variables display strong correlation structures within annotation types, as well as within tissues and cell types at a finer granularity level (Figure 5C). Details about the prescreening step, as well as the epigenetic, genetic, and expression datasets are given in the supplemental material and methods, and the eQTL associations for the prescreening and subsequent analyses are listed in Tables S1, S2, S3, and S4.

In this case study, we concentrate our attention on the following key finding revealed by the prescreening step: chromosome 12 is highly pleiotropic, notably around the gene *LYZ* (MIM: 153450). This gene encodes lysozyme, a highly conserved enzyme with peptidoglycan-lytic activity that is robustly expressed in monocytes. The *LYZ* locus has already been reported as pleiotropic using several monocyte datasets,[27–29] but its functional role remains unclear. We will therefore exploit the epigenetic annotations within EPISPOT to shed light on the mechanisms of action of this locus as well as of other surrounding *cis*- and *trans*-acting loci.

Importantly, while our discussion will mainly concentrate on a few pleiotropic loci of interest, EPISPOT will be applied on a whole collection of loci from chromosome 12, which display QTL signal according to the ATLASQTL prescreening at 5% FDR. By borrowing strength across all the loci (learning from hotspot signals, as well as isolated *cis* and *trans* signals), EPISPOT will infer the epigenetic contributions to the QTL activity of the different regions.

The LYZ-region pleiotropy defines two modules of transcripts

A total of 977 eQTL associations, involving 350 unique SNPs on chromosome 12 and 430 unique transcripts genome-wide, were identified at FDR 5% from the ATLASQTL prescreening analysis of the first dataset. When mapped to the CEDAR dataset, the ATLASQTL eQTLs corresponded to 195 independent loci, expected to involve distinct eQTL signals and comprising a total of $p = 1{,}540$ SNPs (see Figure 5A and data-preparation details in the supplemental material and methods). As highlighted in the second simulation study (section "performance under varying degrees of epigenome involvement"), supplying a dense panel of SNPs (here imputed SNPs) to EPISPOT is important to ensure a sufficient representation of the relevant epigenetic marks among the analyzed SNPs.

We also mapped the prescreened transcripts to the CEDAR dataset. The *LYZ*-region pleiotropy defines two natural modules of transcripts, based on whether they are associated with SNPs in the vicinity of *LYZ* (<1 Mb from it) or not, and further augmenting these modules with highly correlated transcripts (supplemental material and methods). This module partition is driven by the following biological consideration: the peculiar pleiotropic QTL activity arising from the *LYZ* region may be triggered by specific epigenetic influences, which may differ from those triggering isolated (scattered) *cis* or *trans* effects outside the *LYZ* region; to reflect this, the modules are hereafter referred to as the pleiotropic module and the scattered module, respectively (Figure 5A).

The correlation structure within and across the two modules supports this partitioning (Figure 5B). Namely, it indicates a strong co-expression of transcripts within the pleiotropic module, suggesting a dense network of genes whose connections may be attributed in large part to the shared QTL control exerted by the *LYZ* hotspots. Conversely, the transcripts in the scattered module display

Please cite this article in press as: Ruffieux et al., EPISPOT: An epigenome-driven approach for detecting and interpreting hotspots in molecular QTL studies, The American Journal of Human Genetics (2021), https://doi.org/10.1016/j.ajhg.2021.04.010

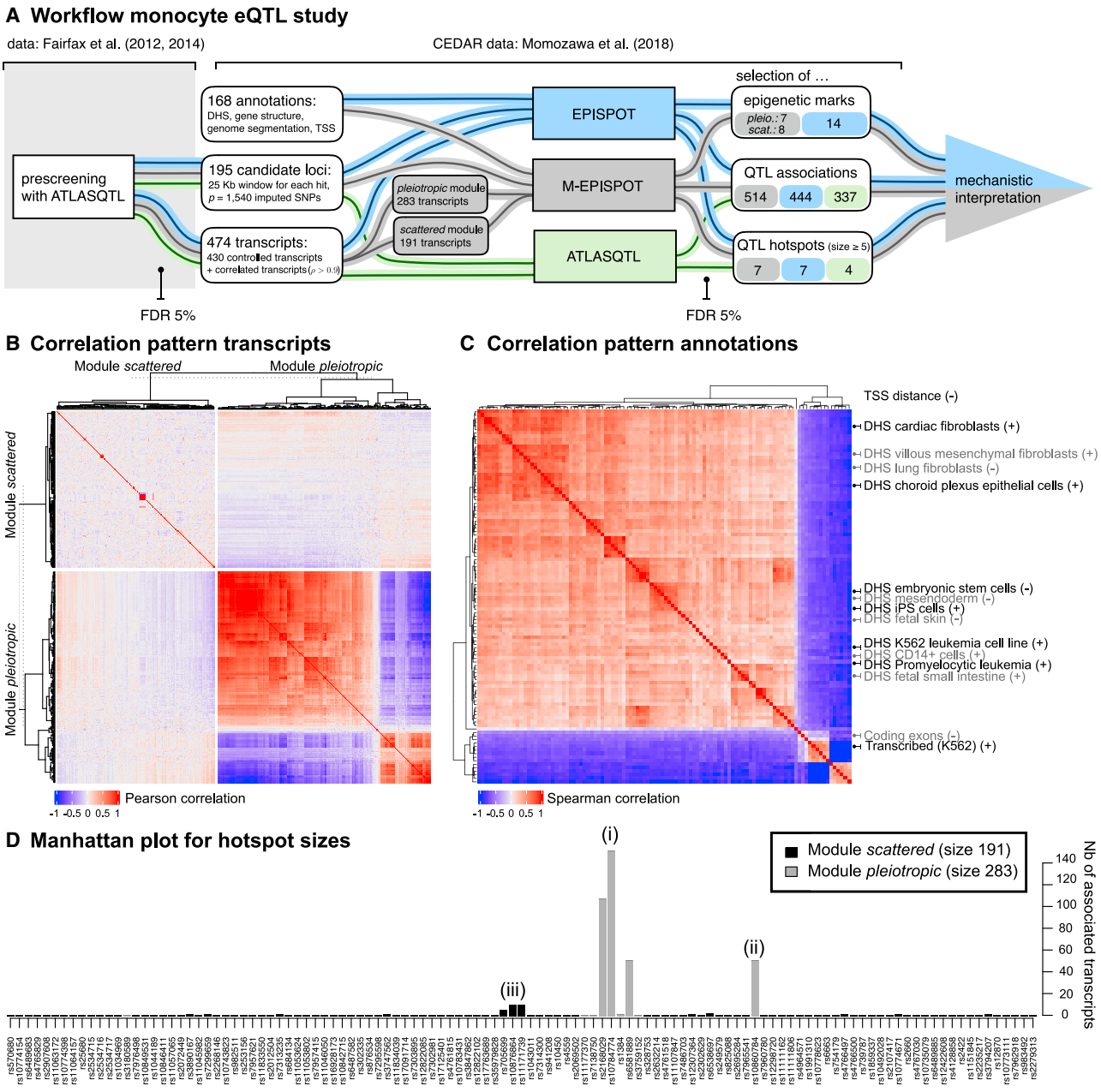

Figure 5. Overview of the monocyte eQTL case study

(A) Workflow for the monocyte eQTL case study. Candidate loci from chromosome 12 and transcripts are obtained from a preliminary prescreening in the first dataset[25] using the joint eQTL mapping approach ATLASQTL[5] with a permutation-based Bayesian false discovery rate (FDR) of 5% for selecting pairs of associated SNP-transcript. The analysis is then performed in the second dataset (CEDAR).[26] EPISPOT and M-EPISPOT select associated SNP-transcript pairs, QTL hotspots, and epigenetic marks relevant to the primary QTL associations. This output is then interpreted as a whole to generate hypotheses about the mechanisms of action underlying these associations.

(B) Correlation of the analyzed transcripts according to their module membership. The "pleiotropic module" displays a strong dependence pattern, reflecting dense connections in the network controlled by the hotspots; the traits in the "scattered module" are mostly uncorrelated, which is unsurprising given that they are mainly controlled via isolated *cis* mechanisms.

(C) Correlation of the epigenetic annotations supplied to the method. All variables are binary, except the distance to the closest transcription start site (TSS) which is not included in the heatmap. Only the labels of the marks retained by M-EPISPOT are displayed; a heatmap with the full labels is provided in the supplemental material and methods. The majority of the marks are DNase-I hypersensitivity sites (DHSs) in different tissues and cell types. They tend to cluster together on the top left 4/5 of the heatmap, and DHSs in similar tissues and cell types also form subgroups. The remaining marks relate to gene structures and genome segmentation annotations. The labels indicated on the right are in gray and black depending on whether they were selected by M-EPISPOT as relevant for the pleiotropic, resp. scattered module. The + and − indicate positive, resp. negative effects of the marks, i.e., their triggering or repressive action on the primary QTL effects. Their relevance is discussed in the main text and in the supplemental material and methods.

(D) Hotspot sizes (i.e., number of associated transcripts per SNP) as inferred by M-EPISPOT. Only the active SNPs (i.e., associated with ≥ 1 transcripts) are displayed. The gray and black colors indicate the module membership of the controlled transcripts. The numbers in parentheses refer to the discussion of the main text.

Table 2. Number of hits and replication rates

	PRESCREENING	CEDAR		
	($n = 413$, $p = 28{,}100$, $q = 22{,}827$)	($n = 286$, $p = 1{,}540$, $q = 474$)		
	ATLASQTL	M-EPISPOT	EPISPOT	ATLASQTL
Nb eQTL associations	977	514	444	337
cis replication (%)		78.2	77.9	77.9
trans replication (%)		55.8	54.9	54.9

Number of eQTL associations discovered by the ATLASQTL prescreening (chromosome 12) and by each of the three (M-EPISPOT, EPISPOT, and ATLASQTL) analyses of the CEDAR data, along with the replication rates for the associations discovered at the prescreening stage. All analyses use an FDR threshold of 5%. The numbers of samples *n*, SNPs *p*, and transcripts *q* are indicated for each dataset. Full lists of eQTL associations for the different methods are provided in Tables S1, S2, S3, and S4.

little co-expression, which is unsurprising given that they tend to be involved in isolated QTL effects (most transcripts are controlled by distinct genetic variants).

Overall comparison of methods and replication rates

We next refined our understanding of the eQTL structure in this region using the CEDAR dataset. To assess the sensitivity of inference to this module partition, we compared the results of the module-based algorithm, M-EPISPOT, with those of the base algorithm, EPISPOT, i.e., with no module provided as input. Moreover, to highlight the benefits of using epigenetic information, we also confronted these two runs with an ATLASQTL analysis of the same data. We employed the same settings for all three runs to set common grounds for comparison. In particular, we used a same permutation-based Bayesian FDR threshold of 5% for declaring QTL associations (Figure 5A and supplemental material and methods). Importantly, the simulated annealing scheme implemented as part of the EPISPOT algorithm is specifically designed to handle the strong LD structures present in the dense SNP panel data and the block correlation structures among transcript levels (Figure 5B) and epigenetic marks (Figure 5C); an illustration for different degrees of LD is given in the supplemental material and methods.

In the CEDAR dataset, the M-EPISPOT analysis of the two modules ($q = 283+191$ transcripts) and the 195 candidate loci ($p = 1{,}540$ SNPs) identified 514 eQTL associations, involving a total of 267 unique transcripts and 82 unique loci (Table S2). In terms of independent replication of the prescreening hits, this corresponds to rates of 78.2% and 55.8% for the *cis* and *trans* QTL associations, respectively. Using ATLASQTL instead of M-EPISPOT on the CEDAR data yielded 262 unique active transcripts and 80 unique active loci, with slightly lower *cis* and *trans* replication rates, namely 77.9% and 54.9%, respectively (Table 2, Table S4). Similar observations were obtained for EPISPOT (Table 2, Table S3). Given the well-known difficulty to validate *trans* effects and the relatively small sample size of the CEDAR dataset (n = 289), these appreciable independent replication rates may result from the efficient joint modeling of all transcripts and SNPs achieved by M-EPISPOT, EPISPOT, and ATLASQTL.

A focus on two susceptibility loci

We next discuss two examples of pleiotropic loci. First, not only does M-EPISPOT confirm the *LYZ* pleiotropic activity (Figure 5D-i), but it also uncovers associations of this locus with four additional genes compared to the ATLASQTL run, namely, *COPZ1* (MIM: 615472), *DPY30* (MIM: 612032), *KLHL28*, and *OSTC* (MIM: 619023). The EPISPOT run (with no module partitioning) reports the exact same list as ATLASQTL, also missing the above four genes.

The second example is a pleiotropic locus uncovered by M-EPISPOT and for which only isolated effects were detected at the prescreening stage (Figure 5D-ii). This locus is located 32 Mb downstream to the *LYZ* locus and entails a hotspot of size 52 in the gene body of *GNPTAB* (MIM: 607840), namely, rs10860784 ($r^2 = 0.001$ with the lead hotspot rs10784774 of the *LYZ* locus). The *trans* network formed by the controlled transcripts has not been previously described and neither has any *trans*-acting effect involving rs10860784 (up to proxies using $r^2 > 0.8$). However, rs10860784 is known to be *cis*-acting on *DRAM1* (MIM: 610776) (located 98 Kb downstream) in multiple tissues,[30] an association which M-EPISPOT also confirms using a looser FDR of 15%. Moreover, the UK Biobank PheGWAS also reported[31] a strong association between this SNP and height (MIM: 606255) ($p = 1.47 \times 10^{-14}$).

The module-free version EPISPOT run also finds a *trans* network for the exact same SNP, yet slightly smaller, as it involves 31 transcripts at FDR 5%; ATLASQTL finds no signal. This example suggests that the added value of epigenome-driven inference is particularly striking for the detection of weak *trans* signals. Indeed, a comparison of the estimated QTL effects attributable to rs10860784 with those attributable to *LYZ* pleiotropic locus (Figure 5D-i) shows that the former are significantly smaller in magnitude compared to the latter (t test $p < 2 \times 10^{-16}$).

The selected epigenetic annotations reveal possible genetic mechanisms of action

The above figures suggest that the M-EPISPOT and EPISPOT runs allow for more powerful QTL mapping compared to ATLASQTL. This probably results from their ability to leverage the epigenetic marks, as we next discuss.

For each module, M-EPISPOT identifies a subset of epigenetic annotations with a potential to induce or inhibit the

QTL associations (depending on the sign of the posterior mean of each annotation effect); these annotations are highlighted in Figure 5C. For instance, DNase-I hypersensitivity sites (DHS) in fibroblasts and epithelial cells of different tissues tend to promote the QTL effects. Interestingly, DHS in $CD14^+$ monocytes are found to be enhancers of eQTL effects in both the M-EPISPOT and EPISPOT runs, with epi-PPI > 0.99. The two runs also estimate a negative effect of the distance to transcription start sites (TSSs, epi-PPI > 0.99), in line with the frequently reported decay in abundance of eQTL signals with the distance to TSS.[32] These last two observations are helpful to interpret the uncovered QTL signals, as we next discuss.

$CD14^+$ cell DHS: Hints to a monocyte-specific pleiotropic activity in LYZ

We first focus on the *LYZ* pleiotropic region. Previous studies have highlighted distinct lead hotspots around *LYZ*,[33] yet none provided a functional characterization that would allow a clear prioritization of one variant over another. The lead hotspots revealed by the M-EPISPOT and EPISPOT runs are intergenic variants, rs10784774 (size 154) and rs2168029 (size 109, $r^2 = 0.89$ with rs10784774; see Figure 5D-i). They differ from the lead hotspot flagged by the ATLASQTL run, namely, rs1384 (size 149, $r^2 = 0.99$ with rs10784774). We next examine the possible biology behind these candidates, starting with the ATLASQTL top hotspot.

The fact that rs1384 is located within the 3′ UTR of *LYZ* may suggest a *trans* action mediated by *LYZ*. This hypothesis is plausible given that the locus associates with *LYZ* in all M-EPISPOT, EPISPOT, and ATLASQTL runs and that GTEx also reported this *cis* association in whole blood and different tissues. Conversely, regressing out the effect of *LYZ* on the expression matrix does not explain away the hotspot effects (the size of the top hotspot in the *LYZ* locus is only marginally reduced: 134 versus 154 in the original M-EPISPOT analysis, supplemental material and methods). This does not rule out *LYZ* expression initiating the formation of the hotspot, but the downstream consequential changes in expression are too complex to simply regress out in a linear manner, and so only reduced mediation is observed.

The monocyte-specific DHS annotation selected by M-EPISPOT for the pleiotropic module suggests a complementary scenario. Namely, the pleiotropic activity of the locus may be triggered by cell-type-specific enhancers in open chromatin regions, which are known to be key players in activating the transcription in *trans*.[34] This hypothesis of monocyte-specific pleiotropy would also explain why no hotspot was reported so far in cell types and tissues other than monocytes.[25,35] To investigate this further, we performed an additional enrichment analysis using the multiple tissue- and cell-type histone modification marks of the ENCODE catalog: we found that the two sets of genes associated with the M-EPISPOT's lead hotspots rs10784774 and rs2168029, respectively, are enriched in H3K27ac enhancers, again in $CD14^+$ monocytes only, which further supports cell-type-specific activation. One notable gene in this group is the transcription factor *CREB1* (MIM: 123810), which has previously been suggested as a putative mediator of the *LYZ* pleiotropic network.[25] Notably, regressing out the effect of *CREB1* on the expression matrix substantially reduces the pleiotropy of the locus (the size of the top hotspot in the *LYZ* locus is 36, versus 154 in the original M-EPISPOT analysis). Moreover, the connectivity of the transcript conditional independence network is also markedly lower (supplemental material and methods).

It seems most plausible, however, that the *trans*-mediation effect by *CREB1* may be preceded by a *cis* effect on *LYZ* or an isoform-specific effect. This possibility is supported by a strong divergent allele-specific correlation between *LYZ* and *CREB1*, which we observed when conditioning on the genotype of the lead hotspot rs10784774 (supplemental material and methods). This indicates an indirect *cis-trans-cis* mediation of the *trans* network by *LYZ*-mediated *CREB1* expression differentially feeding back onto *LYZ*, an observation replicated in both datasets analyzed. Notably, scanning SNP effects on transcription factor binding motifs identifies putative divergence in *CREB1* binding dependent upon allelic carriage at rs10784774, in keeping with the allele-specific correlation observation (supplemental material and methods). While our analyses of residual values cannot completely resolve this, such a feedback circuit might explain why the effect of regressing for *CREB1* is greater than the effect of regressing for *LYZ*. Finally, it has previously been noted that EP300 (MIM: 602700), a binding partner of CREB1, shows allelic effect on *LYZ* expression,[25] although this in an opposing manner to that observed for CREB1 alone, and importantly, the effect size of the EP300 association is markedly less than that for CREB1. In total, these observations lend further weight to allele-specific regulation via rs10784774, although, given that CREB1 and EP300 may form components of multi-protein complexes, the fine mechanistic details of this regulation fall outside the scope of this publication.

We further explored whether the two sets of genes associated with either rs10784774 and rs2168029 were enriched in transcription factor binding sites (TFBS) using the ENCODE data in K562 cells. We found a profound enrichment of a number of TBFS, including ATF3, CREB1, and c-Myc (Table S5). The networks of transcription factors for rs10784774 and rs2168029 are similar, indicating conserved regulatory networks, although unlike with rs10784774, rs2168029 does not overlap a *CREB1* binding site and therefore would not be proposed to feedback here.

Interestingly, ATF transcription factors are CREB-binding proteins, in line with the *CREB1*-mediation hypothesis, but the strong enrichment for many other transcription factors suggests that the same loci can be targeted by different processes and the co-occupancy of these loci in primary monocytes may resolve this further, although is

important to note that, unlike the very significant association between *LYZ* and *CREB1*, there is no association between *LYZ* and *ATF3* expression, so we can discount this gene playing a role in this genomic circuit. The c-Myc transcription factor is involved in cell division and has broad transcriptional consequences,[36] which is sensible given the large pleiotropy observed at the *LYZ* locus, for rs10784774 and rs2168029. Consistent with this, the UK Biobank data further reveal strong associations of these two SNPs with monocyte counts and other myeloid cell counts.[31]

Although by no means conclusive, these observations corroborate the context specificity of the *trans* effects controlled by the *LYZ* locus, and indeed may be more representative of other unresolved *trans* loci across the genome that, while of potential high biological importance, lack the pleiotropic effect of the *LYZ* locus. They also suggest that the epigenome-driven EPISPOT runs found promising candidate hotspots, whose presumed mechanisms of action on the massive *LYZ* gene network would merit experimental follow up.

Distance to TSSs: Examples of cis *and hotspot signals shared across cell types*

Another interesting result concerns the negative effect of the annotation coding the distance to TSSs, this time for transcripts belonging to the scattered module. As active transcripts in this module are mostly involved in *cis* associations, the module specificity of this annotation aligns with the previous observation that the distance to TSSs associates with an enrichment of *cis* eQTLs.[32,37] Moreover, an empirical assessment of this enrichment in our dataset shows that the SNPs selected with M-EPISPOT are on average significantly closer to TSSs compared to SNP subsets of the same size randomly drawn within the analyzed loci ($p = 0.017$). Such an enrichment is unsurprising and actually also present in the EPISPOT and ATLASQTL results, but the importance of the distance to TSS is nevertheless made explicit by the selection of the TSS variable by both EPISPOT and M-EPISPOT.

For instance, three candidate hotspots, rs10876864, rs11171739 ($r^2 = 0.94$ with rs10876864), and rs705699 ($r^2 = 0.86$ with rs10876864), located 13 Mb upstream of the *LYZ* locus, are representative of this enrichment as they are within a TFBS, a 5′ UTR and an exon, respectively (Figure 5D-iii). Our ATLASQTL prescreening and EPISPOT analyses find that they control a small network of size 11 involving transcripts mapping to the *cis* gene *RPS26* (MIM: 603701) and other distal genes, including *IP6K2* (MIM: 606992) on chromosome 3.

This locus has been linked with several autoimmune diseases[38–41] including type 1 diabetes (MIM: 222100), where evidence exists that *RPS26* transcription does not mediate the disease association.[42] Interestingly, previous studies have reported the *RPS26 cis* effect as an isolated association in monocytes. The *trans* activity, in particular on *IP6K2*, was unknown in monocytes, but is known in B and T cells.[25,43] This suggests that it has so far gone unnoticed in monocytes using standard univariate mapping approaches, but our fully joint, annotation-driven method has enabled its detection. Moreover, unlike the monocyte-specific *LYZ* pleiotropic locus discussed above, this locus is an example of *trans*-hotspot eQTL present in several cell types. The genomic location also aligns with the observation that eQTLs common to multiple cell types or tissues tend to be closer to TSSs compared to eQTLs only detectable in a single cell type or tissue.[44]

Discussion

Large panels of epigenetic marks are nowadays collected along with genetic data and employed as part of different modeling approaches, whether for single-trait association studies or fine mapping.[8–12] However, their use to enhance molecular QTL mapping remains mostly heuristic. Thanks to its hypothesis-free mark selection routine which is fully integrated within a joint QTL mapping framework, EPISPOT can identify the relevant epigenetic marks from thousands of candidates, while also directly refining estimation in large molecular QTL studies.

Specifically, EPISPOT brings important modeling and algorithmic contributions. First, it implements a flexible hierarchical model which enables parametrizing both *cis* and *trans* actions on thousands of molecular traits, whereas existing epigenome-based approaches are limited to GWAS or *cis* QTL mapping for one or a handful of traits.[8–12] Second, it is both fully joint and scalable, accounting for all epigenetic marks, genetic variants and molecular levels, and their shared signals, in a single modeling framework. Third, it combines this information to perform an automated selection of the epigenetic marks relevant to the QTL effects of the problem at hand, thereby providing direct insight into the functional basis of the signals. Fourth, its crafted annealed variational algorithm ensures a robust exploration of complex parameters spaces, such as induced by candidate SNPs in high LD, corresponding to scenarios for which the use of epigenetic information is particularly beneficial. Finally, EPISPOT allows for module-specific learning of the epigenetic action.

We showed in a series of simulation experiments emulating epigenome-driven QTL problems that EPISPOT effectively scales to large datasets, while retaining the accuracy necessary for a powerful QTL mapping. We demonstrated that our method was not only able to pinpoint the correct marks with high posterior probability, but that it could also leverage these marks to improve the detection of weak QTL signals. In particular, we saw that the spike-and-slab representation of the epigenome contribution ensures that the irrelevant epigenetic marks are effectively discarded as "noise," so panels with hundreds of candidate marks can be considered without the risk of worsening inferences. This allows skipping the delicate process of pre-filtering marks, whose practical grounds are often blurry and disconnected from the QTL dataset under consideration. Moreover, although in a strict sense

epigenetic marks represent a subset of functional annotations, it is possible to interpret this terminology more loosely and supply other types of annotations or scores that may carry information about the involvement of SNPs in QTL regulation.

Our work attaches special importance to acknowledging the complexity of the learning task (selection of hotspots, pairwise QTL associations between variants and molecular traits, selection of epigenetic marks relevant to these QTL associations) and possible biological scenarios (pattern of regulation, importance of the epigenome in this regulation, dependence structures among variants, marks and molecular traits, and between them). Our simulations examined under what conditions inference is well powered to leverage the epigenetic information and evaluated the sensitivity to different input choices, in particular when gene modules are provided. Importantly, our method is not meant to be used as a black box to fish genetic variants involved in *trans* regulation and their epigenetic roots, but rather is predicated on a careful analysis design that takes into account the dataset, the biological question of interest, and the expected statistical power. Further assessments for specific problem settings (sparsity levels, association patterns, and epigenetic control) can be made using the code provided online (see EPISPOT and ECHOSEQ in web resources).

Finally, we showed how our simulation studies prefigured the efficiency of EPISPOT in a large monocyte eQTL study (high replication in an independent sample, previously unreported pleiotropic loci, refined list of candidate lead hotspots). We further illustrated how the EPISPOT posterior output can be used to both select interpretable annotations underlying the QTL activity and reduce the range of hypotheses about the functional mechanisms involved, particularly for hotspots. We also showed how the localized nature of QTL activity could be accounted for when inferring annotations in a module-specific fashion using M-EPISPOT (the monocyte-specific enhancer activity affecting the pleiotropic module, the enrichment of QTL hits closer to TSSs affecting the scattered module). Altogether, this thorough case study demonstrates that QTL analyses may largely benefit from the use of rich complementary data sources annotating the primary genotyping data, provided principled joint approaches are used to capture shared association patterns.

EPISPOT offers perspectives for robust and interpretable molecular QTL mapping, toward a better understanding of the functional basis of genetic regulation. Thanks to its efficient annealed VBEM algorithm with adaptive and parallel schemes, it enables information sharing across epigenetic marks, genetic variants, and molecular traits governed by complex regulatory mechanisms, at scale. In particular, its use of selection indicators in a spike-and-slab framework allows for a systematic identification of sparse sets of epigenetic annotations which are directly relevant for the QTL regulation of the problem at hand.

We envision holistic approaches such as EPISPOT to be increasingly adopted in an age where large molecular datasets and annotation information become widely available. EPISPOT is applicable to any type of molecular QTL problem, involving genomic, proteomic, lipidomic, or metabolomic levels, but also to genome-wide association with several clinical endpoints. In particular, exploiting the epigenome to build finer maps of hotspots across the genome holds great promises, as these master regulators are likely to be triggered by tissue- and cell-type-specific epigenetic functions.

Data and code availability

Fairfax et al.[25,28] provide gene expression in $CD14^+$ monocytes and genotyping data from individuals with European ancestry. The raw expression data were generated with Illumina HumanHT-12 v4 arrays and downloaded from ArrayExpress[45] (accession E-MTAB-2232), while the raw genotyping data were generated by Illumina HumanOmniExpress-12 arrays and have been deposited at the European Genome-Phenome Archive (accessions: EGAD00010000144 and EGAD00010000520). The expression data are freely available, but the genotyping data require a data access agreement, as detailed in Fairfax et al.[25,28] and https://www.well.ox.ac.uk/research/research-groups/julian-knight-group/research-projects/data-access.

The CEDAR dataset[26] consists of gene expression data from $CD14^+$ monocytes and genotyping data from individuals with European ancestry. The raw expression data were generated with Illumina HumanHT-12 v4 arrays and downloaded from ArrayExpress[45] (accession: E-MTAB-6667), while the raw genotyping data were generated by Illumina HumanOmniExpress-12 v1_A arrays and downloaded from ArrayExpress (accession: E-MTAB-6666). Both the expression and genotyping data are freely available.

Both studies were approved by the local human research ethic committees, namely, the Oxfordshire Research Ethics Committee (COREC reference 06/Q1605/55)[28] and the University of Liège Academic Hospital Ethics Committee.[26] Participants provided informed written consent, and all procedures were conducted in accordance with the Declaration of Helsinki.

All statistical analyses were performed using the R environment (v.3.6.1)[46] and the synthetic datasets were generated using the freely available R package ECHOSEQ (v.0.3.0). The R package EPISPOT implements the method.

Supplemental information

Supplemental information can be found online at https://doi.org/10.1016/j.ajhg.2021.04.010.

Acknowledgments

We are grateful to the editor and the two anonymous referees for their helpful comments. We thank Verena Zuber for her help in setting up the epigenetic annotation panel and Colin Starr for managing computational resources.

This research was funded by the UK Medical Research Council program MRC MC UU 00002/10 (H.R., S.R.), MC UU 00002/4 (E.V., C.W.), and MR M0 13138/1, MR S0 2638X/1 (L.B.); the Engineering and Physical Sciences Research Council EP/R018561/1 (S.R.); the BHF-Turing Cardiovascular Data Science Awards 2017

& the Alan Turing Institute under the Engineering and Physical Sciences Research Council grant EP/N510129/1 (L.B.); the Alan Turing Institute Fellowship number TU/B/000092 (S.R.), and the Wellcome Trust WT107881 (E.V., C.W.). This work was also supported by the NIHR Cambridge BRC. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR, or the Department of Health and Social Care. B.P.F. and I.N. are funded by a Wellcome Intermediate Clinical Fellowship to B.P.F. (no. 201488/Z/16/Z).

Declaration of interests

The authors declare no competing interests.

Received: November 13, 2020
Accepted: April 8, 2021
Published: April 27, 2021

Web Resources

ATLASQTL (v.0.1.4), https://github.com/hruffieux/atlasqtl
ECHOSEQ (v.0.3.0), https://github.com/hruffieux/echoseq
EnrichR (v.2.1), https://amp.pharm.mssm.edu/Enrichr
Ensembl, http://grch37.ensembl.org/index.html
EPISPOT, implemented as an R package with C++ subroutines and publicly available under the GNU General Public License version 3 (GPL3), https://github.com/hruffieux/epispot
GTEx, https://gtexportal.org/home
GWAS Catalog, https://www.ebi.ac.uk/gwas
MATRIXEQTL (v.2.3), http://www.bios.unc.edu/research/genomic_software/Matrix_eQTL
OMIM, https://www.omim.org
PhenoScanner, http://www.phenoscanner.medschl.cam.ac.uk
PLINK (v.v1.90b5.3), http://zzz.bwh.harvard.edu/plink
R (v.3.6.1), https://www.r-project.org

References

1. Stegle, O., Parts, L., Durbin, R., and Winn, J. (2010). A Bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies. PLoS Comput. Biol. *6*, e1000770.
2. Shabalin, A.A. (2012). Matrix eQTL: ultra fast eQTL analysis via large matrix operations. Bioinformatics *28*, 1353–1358.
3. Jia, Z., and Xu, S. (2007). Mapping quantitative trait loci for expression abundance. Genetics *176*, 611–623.
4. 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.
5. Ruffieux, H., Davison, A.C., Hager, J., Inshaw, J., Fairfax, B., Richardson, S., and Bottolo, L. (2020a). A global-local approach for detecting hotspots in multiple response regression. Ann. Appl. Stat. *14*, 905–928.
6. Ruffieux, H., Davison, A.C., Hager, J., and Irincheeva, I. (2017). Efficient inference for genetic association studies with multiple outcomes. Biostatistics *18*, 618–636.
7. Ruffieux, H., Carayol, J., Popescu, R., Harper, M.E., Dent, R., Saris, W.H.M., Astrup, A., Hager, J., Davison, A.C., and Valsesia, A. (2020b). A fully joint Bayesian quantitative trait locus mapping of human protein abundance in plasma. PLoS Comput. Biol. *16*, e1007882.
8. Quintana, M.A., and Conti, D.V. (2013). Integrative variable selection via Bayesian model uncertainty. Stat. Med. *32*, 4938–4953.
9. Yang, J., Fritsche, L.G., Zhou, X., Abecasis, G.; and International Age-Related Macular Degeneration Genomics Consortium (2017). A scalable Bayesian method for integrating functional information in genome-wide association studies. Am. J. Hum. Genet. *101*, 404–416.
10. Kichaev, G., Bhatia, G., Loh, P.-R., Gazal, S., Burch, K., Freund, M.K., Schoech, A., Pasaniuc, B., and Price, A.L. (2019). Leveraging polygenic functional enrichment to improve GWAS power. Am. J. Hum. Genet. *104*, 65–75.
11. Kichaev, G., Yang, W.-Y., Lindstrom, S., Hormozdiari, F., Eskin, E., Price, A.L., Kraft, P., and Pasaniuc, B. (2014). Integrating functional data to prioritize causal variants in statistical fine-mapping studies. PLoS Genet. *10*, e1004722.
12. Li, Y., and Kellis, M. (2016). Joint Bayesian inference of risk variants and tissue-specific epigenomic enrichments across multiple complex human diseases. Nucleic Acids Res. *44*, e144–e144.
13. Leek, J.T., and Storey, J.D. (2007). Capturing heterogeneity in gene expression studies by surrogate variable analysis. PLoS Genet. *3*, 1724–1735.
14. Stegle, O., Parts, L., Piipari, M., Winn, J., and Durbin, R. (2012). Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. Nat. Protoc. *7*, 500–507.
15. Langfelder, P., and Horvath, S. (2007). Eigengene networks for studying the relationships between co-expression modules. BMC Syst. Biol. *1*, 54.
16. Borel, C., Deutsch, S., Letourneau, A., Migliavacca, E., Montgomery, S.B., Dimas, A.S., Vejnar, C.E., Attar, H., Gagnebin, M., Gehrig, C., et al. (2011). Identification of cis- and trans-regulatory variation modulating microRNA expression levels in human fibroblasts. Genome Res. *21*, 68–73.
17. Fagny, M., Paulson, J.N., Kuijjer, M.L., Sonawane, A.R., Chen, C.-Y., Lopes-Ramos, C.M., Glass, K., Quackenbush, J., and Platig, J. (2017). Exploring regulation in tissues with eQTL networks. Proc. Natl. Acad. Sci. USA *114*, E7841–E7850.
18. Zhu, L., Tripathi, J., Rocamora, F.M., Miotto, O., van der Pluijm, R., Voss, T.S., Mok, S., Kwiatkowski, D.P., Nosten, F., Day, N.P.J., et al.; Tracking Resistance to Artemisinin Collaboration I (2018). The origins of malaria artemisinin resistance defined by a genetic and transcriptomic background. Nat. Commun. *9*, 5158.
19. Wainberg, M., Sinnott-Armstrong, N., Mancuso, N., Barbeira, A.N., Knowles, D.A., Golan, D., Ermel, R., Ruusalepp, A., Quertermous, T., Hao, K., et al. (2019). Opportunities and challenges for transcriptome-wide association studies. Nat. Genet. *51*, 592–599.
20. Sieberts, S.K., and Schadt, E.E. (2019). Inferring causal associations between genes and disease via the mapping of expression quantitative trait loci. In Handbook of Statistical Genomics, D.J. Balding, I. Moltke, and J. Marioni, eds. (John Wiley & Sons), pp. 697–733.
21. Blei, D.M., Ng, A.Y., and Jordan, M.I. (2003). Latent Dirichlet allocation. J. Mach. Learn. Res. *3*, 993–1022.
22. 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*, 73–108.

23. Rose, K., Gurewitz, E., and Fox, G. (1990). A deterministic annealing approach to clustering. Pattern Recognit. Lett. *11*, 589–594.
24. Ueda, N., and Nakano, R. (1998). Deterministic annealing EM algorithm. Neural Netw. *11*, 271–282.
25. Fairfax, B.P., Makino, S., Radhakrishnan, J., Plant, K., Leslie, S., Dilthey, A., Ellis, P., Langford, C., Vannberg, F.O., and Knight, J.C. (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.
26. Momozawa, Y., Dmitrieva, J., Théâtre, E., Deffontaine, V., Rahmouni, S., Charloteaux, B., Crins, F., Docampo, E., Elansary, M., Gori, A.-S., et al.; International IBD Genetics Consortium (2018). IBD risk loci are enriched in multigenic regulatory modules encompassing putative causative genes. Nat. Commun. *9*, 2427.
27. Rotival, M., Zeller, T., Wild, P.S., Maouche, S., Szymczak, S., Schillert, A., Castagné, R., Deiseroth, A., Proust, C., Brocheton, J., et al.; Cardiogenics Consortium (2011). Integrating genome-wide genetic variations and monocyte expression data reveals trans-regulated gene modules in humans. PLoS Genet. *7*, e1002367.
28. Fairfax, B.P., Humburg, P., Makino, S., Naranbhai, V., Wong, D., Lau, E., Jostins, L., Plant, K., Andrews, R., McGee, C., and Knight, J.C. (2014). Innate immune activity conditions the effect of regulatory variants upon monocyte gene expression. Science *343*, 1246949.
29. Rakitsch, B., and Stegle, O. (2016). Modelling local gene networks increases power to detect trans-acting genetic effects on gene expression. Genome Biol. *17*, 33.
30. GTEx Consortium (2015). Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science *348*, 648–660.
31. Canela-Xandri, O., Rawlik, K., and Tenesa, A. (2018). An atlas of genetic associations in UK Biobank. Nat. Genet. *50*, 1593–1599.
32. Wen, X., Luca, F., and Pique-Regi, R. (2015). Cross-population joint analysis of eQTLs: fine mapping and functional annotation. PLoS Genet. *11*, e1005176.
33. Kolberg, L., Kerimov, N., Peterson, H., and Alasoo, K. (2020). Co-expression analysis reveals interpretable gene modules controlled by *trans*-acting genetic variants. eLife *9*, e58705.
34. Chen, L., Ge, B., Casale, F.P., Vasquez, L., Kwan, T., Garrido-Martín, D., Watt, S., Yan, Y., Kundu, K., Ecker, S., et al. (2016). Genetic drivers of epigenetic and transcriptional variation in human immune cells. Cell *167*, 1398–1414.e24.
35. Kerimov, N., Hayhurst, J.D., Manning, J.R., Walter, P., Kolberg, L., Peikova, K., Samovica, M., Burdett, T., Jupp, S., Parkinson, H., et al. (2020). eQTL Catalogue: a compendium of uniformly processed human gene expression and splicing QTLs. bioRxiv. https://doi.org/10.1101/2020.01.29.924266.
36. Miller, D.M., Thomas, S.D., Islam, A., Muench, D., and Sedoris, K. (2012). c-Myc and cancer metabolism. Clin. Cancer Res. *18*, 5546–5553.
37. Gaffney, D.J., Veyrieras, J.-B., Degner, J.F., Pique-Regi, R., Pai, A.A., Crawford, G.E., Stephens, M., Gilad, Y., and Pritchard, J.K. (2012). Dissecting the regulatory architecture of gene expression QTLs. Genome Biol. *13*, R7.
38. Todd, J.A., Walker, N.M., Cooper, J.D., Smyth, D.J., Downes, K., Plagnol, V., Bailey, R., Nejentsev, S., Field, S.F., Payne, F., et al.; Genetics of Type 1 Diabetes in Finland; and Wellcome Trust Case Control Consortium (2007). Robust associations of four new chromosome regions from genome-wide analyses of type 1 diabetes. Nat. Genet. *39*, 857–864.
39. Craddock, N.J., Jones, I.R.; and Wellcome Trust Case Control Consortium (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature *447*, 661–678.
40. Hakonarson, H., Qu, H.-Q., Bradfield, J.P., Marchand, L., Kim, C.E., Glessner, J.T., Grabs, R., Casalunovo, T., Taback, S.P., Frackelton, E.C., et al. (2008). A novel susceptibility locus for type 1 diabetes on Chr12q13 identified by a genome-wide association study. Diabetes *57*, 1143–1146.
41. Nair, R.P., Duffin, K.C., Helms, C., Ding, J., Stuart, P.E., Goldgar, D., Gudjonsson, J.E., Li, Y., Tejasvi, T., Feng, B.-J., et al.; Collaborative Association Study of Psoriasis (2009). Genome-wide scan reveals association of psoriasis with IL-23 and NF-kappaB pathways. Nat. Genet. *41*, 199–204.
42. Plagnol, V., Smyth, D.J., Todd, J.A., and Clayton, D.G. (2009). Statistical independence of the colocalized association signals for type 1 diabetes and RPS26 gene expression on chromosome 12q13. Biostatistics *10*, 327–334.
43. Kasela, S., Kisand, K., Tserel, L., Kaleviste, E., Remm, A., Fischer, K., Esko, T., Westra, H.-J., Fairfax, B.P., Makino, S., et al. (2017). Pathogenic implications for autoimmune mechanisms derived by comparative eQTL analysis of CD4+ versus CD8+ T cells. PLoS Genet. *13*, e1006643.
44. Ongen, H., Andersen, C.L., Bramsen, J.B., Oster, B., Rasmussen, M.H., Ferreira, P.G., Sandoval, J., Vidal, E., Whiffin, N., Planchon, A., et al. (2014). Putative cis-regulatory drivers in colorectal cancer. Nature *512*, 87–90.
45. Athar, A., Füllgrabe, A., George, N., Iqbal, H., Huerta, L., Ali, A., Snow, C., Fonseca, N.A., Petryszak, R., Papatheodorou, I., et al. (2019). ArrayExpress update - from bulk to single-cell expression data. Nucleic Acids Res. *47* (D1), D711–D715.
46. R Core Team (2020). R: A Language and Environment for Statistical Computing (R Foundation for Statistical Computing).