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

Hélène Ruffieux[a,*], Benjamin P. Fairfax[b], Isar Nassiri[b], Elena Vigorito[a], Chris Wallace[a,c], Sylvia Richardson[a,d], Leonardo Bottolo[e,d,a]

[a]MRC Biostatistics Unit, University of Cambridge, Cambridge, United Kingdom

[b]Department of Oncology, MRC Weatherall Institute for Molecular Medicine, University of Oxford, United Kingdom

[c]Cambridge Institute of Therapeutic Immunology and Infectious Disease, Jeffrey Cheah Biomedical Centre, Cambridge Biomedical Campus, University of Cambridge, Cambridge, United Kingdom

[d]Alan Turing Institute, London, United Kingdom

[e]Department of Medical Genetics, University of Cambridge, Cambridge, United Kingdom

September 21, 2020

Abstract

We present EPISPOT, a fully joint framework which exploits large panels of epigenetic marks as variant-level information to enhance molecular quantitative trait locus (QTL) mapping. Thanks to a purpose-built Bayesian inferential algorithm, our approach effectively couples simultaneous QTL analysis of thousands of genetic variants and molecular traits genome-wide, and hypothesis-free selection of biologically interpretable marks which directly contribute to the QTL effects. This unified learning approach boosts statistical power and sheds light on the regulatory basis of the uncovered associations. EPISPOT is also tailored to the modelling of *trans*-acting genetic variants, including QTL *hotspots*, whose detection and functional interpretation are challenging with standard approaches. We illustrate the advantages of EPISPOT in simulations emulating real-data conditions and in an epigenome-driven monocyte expression QTL study which confirms known hotspots and reveals new ones, as well as plausible mechanisms of action. In particular, based on monocyte DNase-I sensitivity site annotations selected by the method from > 150 epigenetic annotations, we clarify the mediation effects and cell-type specificity of well-known master hotspots in the vicinity of the lyzosyme gene. EPISPOT is radically new in that it makes it possible to forgo the daunting and underpowered task of one-mark-at-a-time enrichment analyses for the prioritisation of QTL hits. Our method can be used to enhance the discovery and functional understanding of signals in QTL problems with all types of outcomes, be they transcriptomic, proteomic, lipidomic, metabolic or clinical.

Keywords: Epigenetic annotations; Hierarchical modelling; Large-scale multivariate mapping; Molecular QTL studies; Scalable variational EM algorithm; *Trans* hotspots.

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 analysed in molecular quantitative trait locus (QTL) studies, with the aim of unravelling the

*Corresponding author: helene.ruffieux@mrc-bsu.cam.ac.uk

regulatory mechanisms underlying common diseases. However, these studies present additional complexities compared to classical genome-wide association studies (GWAS). First, they entail a very different statistical paradigm: while GWAS 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 of statistical power as they do not exploit the regulation patterns shared by the molecular entities, whereas joint modelling 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 modelling, 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 parametrisation 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 capitalise on the wealth of available *epigenetic annotation sources* to infer the functional potential of genetic variants. The standard strategy uses epigenetic marks mostly for prioritisation of hits derived from marginal screening: it consists in looping through all the loci with statistically significant associations and, for each locus, inspecting a few marks to decide on "a most promising" functional candidate genetic variant among all those in linkage disequilibrium. This approach presents the following disadvantages: first, publicly available databases nowadays contain several hundreds of epigenetic marks for each variant. Preselecting just a few may involve omitting others that are relevant, which may bias the conclusions. Second, even if a comprehensive inspection were feasible, the degrees of relevance of the marks 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.

Here, we argue 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. We propose to couple a fully Bayesian QTL mapping strategy, in which all loci and molecular traits are analysed 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 GWAS [8, 9], suitable use of epigenetic information can boost the detection of weak associations and help in discriminating genuine signals from spurious ones caused by linkage disequilibrium or other confounding factors [10, 11].

We introduce a novel approach, called EPISPOT, which infers the role of sparse sets of marks — from *hundreds of candidate epigenetic marks* — in the activation of both *cis* and *trans* mechanisms affecting *a whole network of molecular traits*. Importantly, our proposal combines this epigenome-driven feature with the flexible hotspot modelling feature from our previous work [5], thereby

offering a unified toolkit to refine the detection of master *trans*-hotspots from panels of candidate loci, aided by the epigenetic information at hand. The base version of EPISPOT assesses the action of the marks uniformly for the full set of analysed transcripts. However, for cases where a sensible partition into subsets of co-expressed molecular traits (*modules* [12]) is available, we also develop a *module* version of EPISPOT, which accommodates module-specific epigenetic action by estimating the contribution of the marks to the QTL associations in each module.

Our take is that fully joint modelling 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-maximisation (VBEM) algorithm, augmented with a simulated annealing scheme which effectively explores the multimodal parameter spaces induced by highly-structured data. This optimisation routine is purposely tailored to the analysis of genetic data with strong linkage disequilibrium blocks, for which the inclusion of the epigenetic data has the greatest impact.

EPISPOT is not targeted at genome-wide discovery but at effecting refined QTL mapping and hotspot prioritisation, based on genomic regions — or *candidate loci* — believed to be involved in QTL regulation. The candidate loci can be obtained from an application of ATLASQTL or from another preliminary screening, ideally on an independent dataset (see Figure 1D for a typical analysis workflow and Methods section for details). 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 harbour hundreds of genetic variants (most of them with no QTL activity). EPISPOT exploits shared epigenetic signals across these regions to then select QTL hits from them with an increased statistical power.

Our framework also constitutes a novel tool for *interpreting* (i) the detected *trans*-acting and *hotspot* variants based on their overlap with the selected epigenetic marks, and (ii) 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 [e.g., 13–15], where functional roles are difficult to decipher. Moreover, the massive *trans*-gene networks under genetic control are thought to be subject to subtle interplays [16], 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. These strategies range from hypothesis-driven bottom-up approaches that start from isolated mechanisms and try to generalise 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 modelling approaches); see [17] for a review. Our approach provides an alternative anchor towards decoding the complex networks controlled by hotspots, namely via the epigenetic marks found to be informative for the genetic mapping.

Importantly, fruitful applications of EPISPOT, that can successfully decipher part of the molecular regulation machinery, require problems where the signal-to-noise and density of epigenetic/QTL signals are sufficient. We will describe extensive simulation experiments to highlight the benefits of using epigenetic information when available for a panel of regulatory program 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 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 lyzosyme pleiotropic activity of chromosome 12 — an activity which, although reported in several studies, was 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.

Results

Two-level hierarchical regression model. We consider a Bayesian model linking three data sources (Figure 1A) with two levels of hierarchy, each involving a spike-and-slab regression formulation (Figure 1B and Methods section). It is meant to be used as a refining analysis tool, on candidate loci and molecular traits for which some degree of involvement in genetic regulation has been evidenced (Figure 1D and Methods section).

The bottom level of the model uses a series of conditionally independent regressions to parametrise the QTL effects, i.e., the regulation of q molecular traits — gathered in an $n \times q$ matrix $\boldsymbol{y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_q)$ — by p candidate genetic variants or single nucleotide polymorphisms (SNPs) — gathered in an $n \times p$ matrix $\boldsymbol{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_p)$ — for n samples. It models the QTL associations using a binary latent parameter γ_{st} taking value 1 if and only if SNP s is associated with trait t. The posterior means of the γ_{st} then correspond to marginal posterior probabilities of inclusion of the SNP-trait pairs, $\mathrm{pr}(\gamma_{st} = 1 \mid \boldsymbol{y})$ (qtl-PPIs, Figure 1C), from which Bayesian false discovery rate (FDR) estimates can be obtained.

The top-level hierarchy then lets the primary QTL associations be informed by r candidate epigenetic marks for the p SNPs — gathered in a $p \times r$ matrix $\boldsymbol{V} = (\boldsymbol{V}_1, \ldots \boldsymbol{V}_r)$ — via a spike-and-slab probit regression on the probability of effects. 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 modelling the candidate marks (see simulations studies hereafter). Moreover, similarly as for the QTL effects, mark selection is easily achieved using posterior probabilities of inclusion, $\mathrm{pr}(\rho_l = 1 \mid \boldsymbol{y})$, 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.

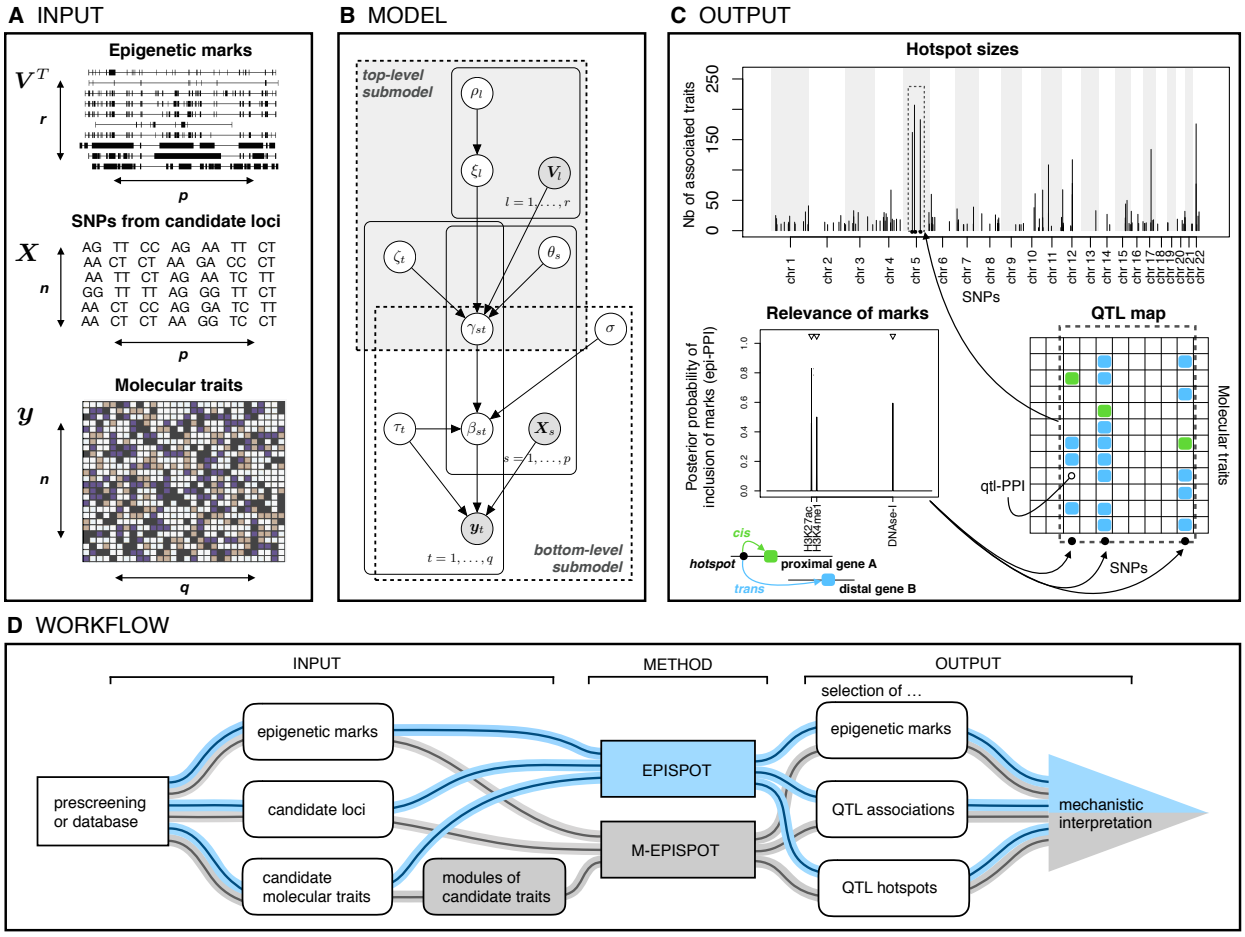

Figure 1: **Overview of EPISPOT.** A: Data input. Epigenetic annotations (predictor-level information) $\boldsymbol{V}$, genetic variants from candidate loci (candidate predictors) $\boldsymbol{X}$, molecular traits (responses) $\boldsymbol{y}$. B: Graphical representation for the two-level hierarchical 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 models the residual variability of trait $\boldsymbol{y}_t$. See also Methods section. C: Posterior output. Selection of epigenetic marks with a role in QTL regulation is carried out using the posterior probabilities of inclusion (epi-PPIs), $\text{pr}(\rho_l = 1 \mid \boldsymbol{y})$, $l = 1, \ldots, r$ (bottom left) and selection of associated SNP-trait pairs (aided by the marks) is carried out using the the posterior probabilities of inclusion (qtl-PPIs), $\text{pr}(\gamma_{st} = 1 \mid \boldsymbol{y})$, $s = 1, \ldots, p$; $t = 1, \ldots, q$ (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 present in 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 grey, 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.

Finally, the propensity of each SNP to be associated with many traits, i.e., to be a hotspot, is explicitly modelled using a SNP-specific parameter, θ_s, similarly as in our earlier work [5].

In summary, the EPISPOT model 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 variables. It lever-

ages 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. The model is described in full in the Methods section and Additional file 1.

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 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. The modification of the top-level model hierarchy is detailed in Methods section and the corresponding version of the algorithm is hereafter called M-EPISPOT when an explicit distinction with the base, module-free version is needed.

Different approaches, based on some prior state of knowledge, on specific optimisation 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 modelling point of view, a given module should ideally comprise co-regulated molecular traits, i.e., traits under genetic control and correlated to common epigenetic marks. The top-level regression 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 guidelines 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-maximisation (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. This annealing feature is designed to robustly infer signals from genotyping data with marked linkage disequilibrium structures, whereby the inclusion of epigenetic information is particularly beneficial to disentangle the genetic contributions. Moreover, for M-EPISPOT, inference can be performed in two stages and in parallel across the modules, saving substantial computational time. Both the EPISPOT and M-EPISPOT versions run within minutes to few hours depending on the numbers of loci, molecular traits and epigenetic marks.

The details and a sketch of the algorithm are provided in the Methods section. The full derivation of the annealed VBEM updates is in the Additional file 2. The algorithm is implemented as a publicly available R package with C++ subroutines [18].

Data generation and simulation set up. The series of simulation studies presented in the next sections have the dual purpose of (i) illustrating the effectiveness of EPISPOT in learning from the epigenome when the epigenetic annotations at hand are sufficiently informative (first simulation study), and (ii) 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 Methods section and we further tailor it to each simulation experiment in their dedicated sections. We use the following terminology when referring to the simulated association patterns:

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

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. A subset of 100 SNPs are active (between 0 and 3 per locus) 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 numerical experiments that follow. To help interpretability in the context of the simulations, we also generated marks with positive effects only, i.e., *inducing* QTL activity and not repressing it (Methods section). 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.

We benchmark our approach against our earlier joint model, ATLASQTL [5], also tailored to the modelling of hotspots but which does not accommodate the epigenetic marks, as well as with the purely marginal screening approach, MATRIXEQTL [2], which tests each SNP-trait pair one-by-one and does not involve any epigenetic information.

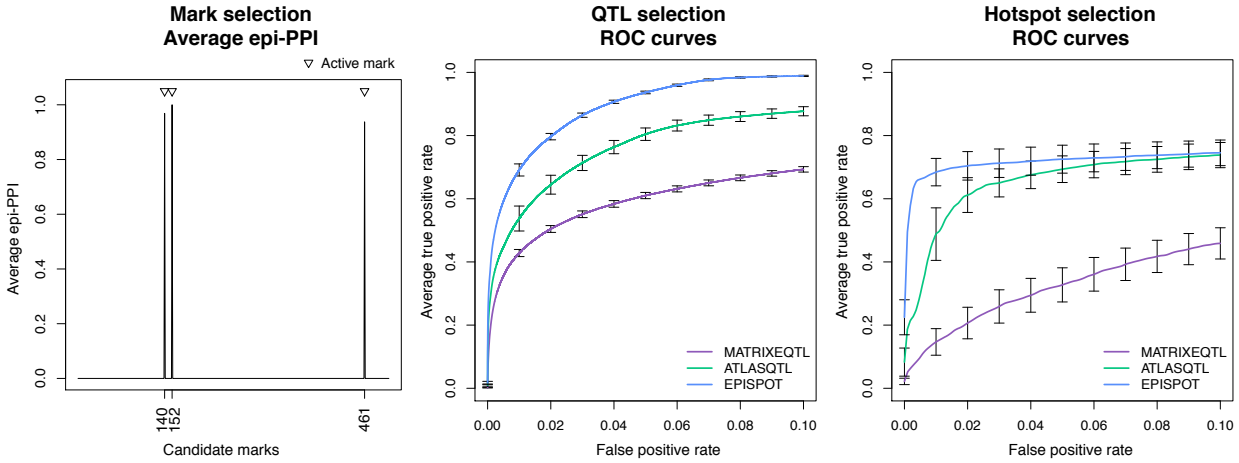

Figure 2: **Performance for selection of epigenetic marks, pairs of associated SNPs and traits, and hotspots.** Left: epi-PPIs for the marks averaged over 32 replicates. The three marks simulated as active are indicated by the triangles. Middle: average partial ROC curves for SNP-trait selection with 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).

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 standardised partial area under the curve (pAUC) overlap, i.e., $(0.74, 0.78)$ and $(0.76, 0.79)$ for ATLASQTL, resp. EPISPOT (Additional file 3). 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 (i) the sparsity of the studied QTL network and magnitude of the QTL effects (ii) the amount of information contained in the data at hand (iii) 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.

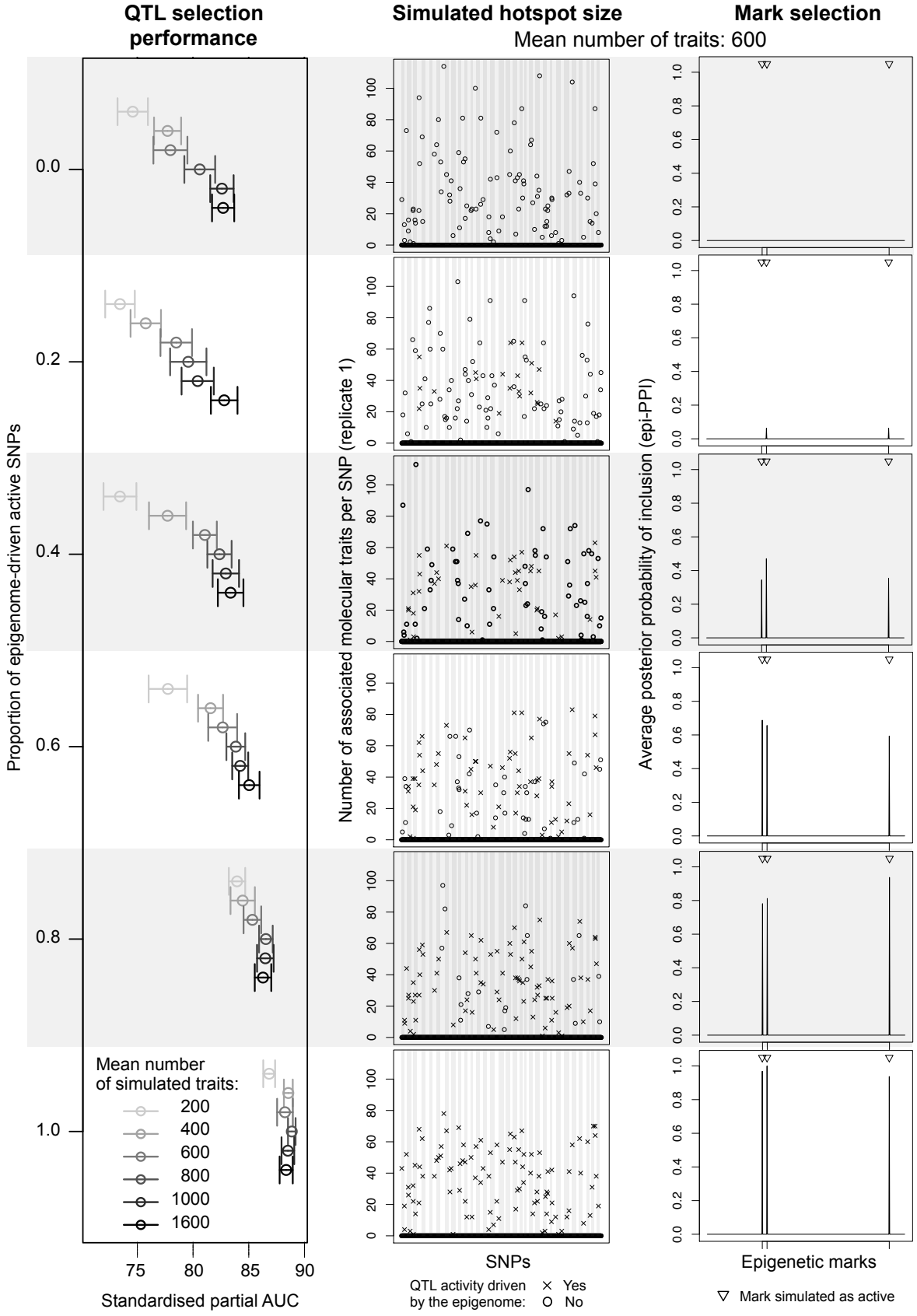

Figure 3: **Performance of EPISPOT for a grid of numbers of traits and proportions $p_{\texttt{epi}}$ of epigenome-driven active SNPs.** Left: standardised pAUCs for the QTL selection performance. Middle: simulated hotspot QTL pattern for problems with an average of 600 traits (first replicate for each value of $p_{\texttt{epi}}$). 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.

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 1600, respectively, 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 Methods section 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 standardised 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 represented at causal variants so that the analysed 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 modelling 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 presence of co-regulated molecular traits, a special case of which is the regulation of these traits by a single hotspot.

Finally, Figure 3 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 ($p_{\text{epi}} = 0$), modelling the $r = 500$ inactive marks does not deteriorate the performance (Additional file 3).

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.

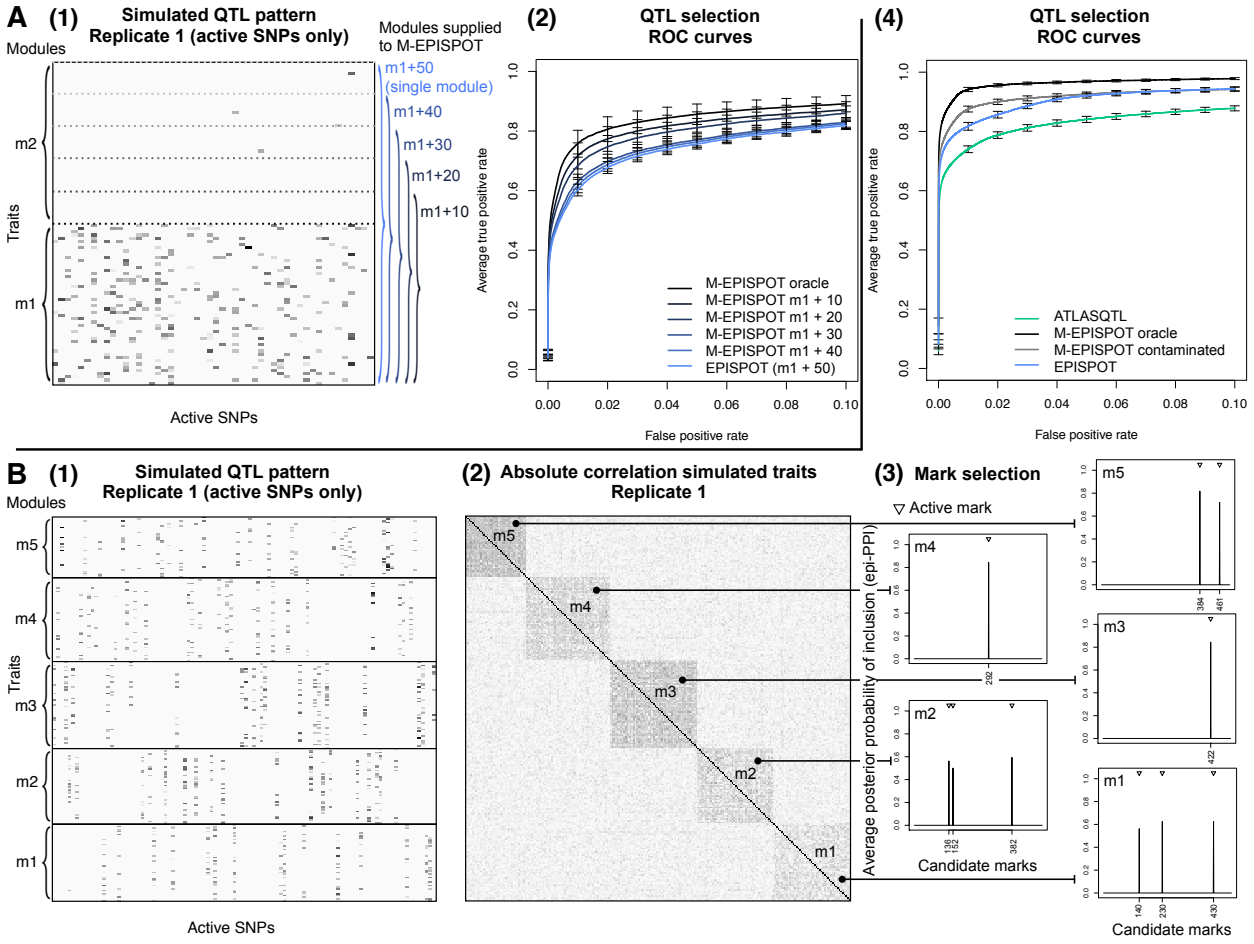

Figure 4: **Performance of M-EPISPOT.** A: Simulated scenario with two modules, whereby the first module m_1 is contaminated by an increasing number of traits from the second module m_2. Panel A(1) shows the simulated pleiotropic pattern for one replicate. The grey levels suggest the different QTL effect 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.

We start with a simple example involving 60 concatenated loci of average size 40 SNPs and two modules of 50 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 modelling point of view, leaving all traits controlled by a same hotspot in a single module permits maximising 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 five modules of average size 50. It compares ATLASQTL, EPISPOT, 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 idealised 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 the that 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 interpretation of genetic regulation in monocytes. Specifically, we analyse 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 in the first dataset ($n = 413$ samples, see [19]), and we then leverage the epigenome using EPISPOT in the second dataset (CEDAR cohort, $n = 286$ samples, see [20]) 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

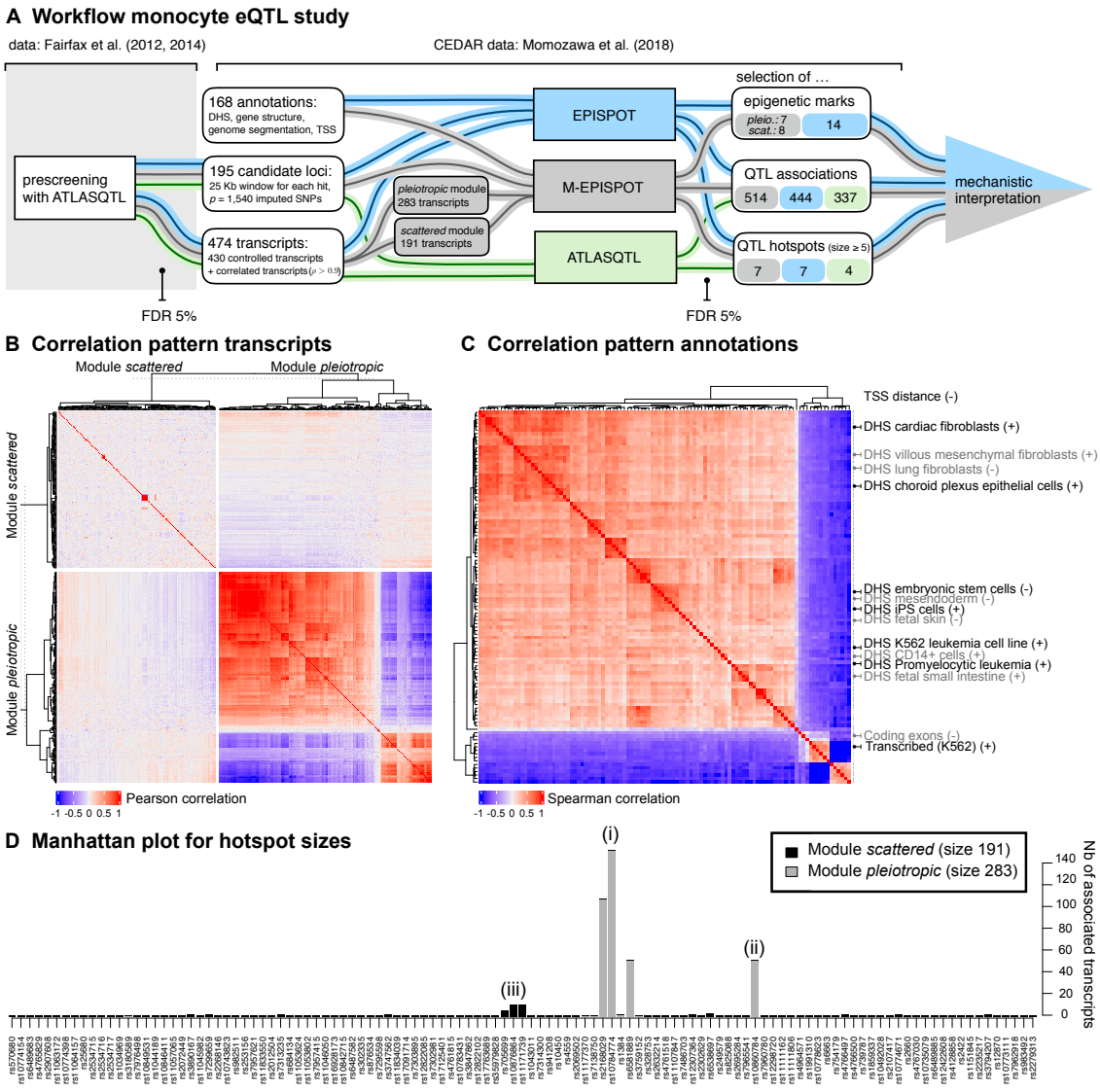

Figure 5: **Overview of the monocyte eQTL case study.** A: Workflow for the monocyte eQTL case study. Candidate loci from chromosome 12 and transcripts are obtained from a preliminary prescreening in the first dataset [19] using the joint eQTL mapping approach ATLASQTL [5] with a permutation-based Bayesian false discovery rate (FDR) of 5% for selecting pairs of associated SNP-transcript. The analysis is then performed in the second dataset [CEDAR, see 20]. 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 analysed 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 Additional file 4. The majority of the marks are DNase-I hypersensitivity sites (DHS) in different tissues and cell types. They tend to cluster together on the top-left 4/5 of the heatmap, and DHS in similar tissues/cell types also form subgroups. The remaining marks relate to gene structures and genome segmentation annotations. The labels indicated on the right are in grey 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 Additional file 4. 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 grey and black colours indicate the module membership of the controlled transcripts. The numbers in parentheses refer to the discussion of the main text.

in the Methods section and Additional file 4, and the eQTL associations for the prescreening and subsequent analyses are listed in Additional files 5–8 (Tables S1–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*. This gene encodes lysozyme, which is associated with the monocyte-macrophage system and enhances the activity of immunoagents. The *LYZ* locus has already been reported as pleiotropic in several monocyte datasets [21–23], 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.

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, these eQTLs corresponded to 195 independent loci, comprising a total of $p = 1,540$ imputed SNPs (Figure 5A and Methods section). As highlighted in the second simulation study, 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 analysed 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 (Methods section). 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 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. 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 Methods section).

Table 1: **Number of hits and replication rates.** 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 Additional files 5–8 (Tables S1–S4).

	PRESCREENING ($n = 413, p = 28100, q = 22827$)	**CEDAR** ($n = 286, p = 1540, q = 474$)		
	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

Importantly, the simulated annealing scheme implemented as part of the EPISPOT algorithm is specifically designed to handle effectively the strong linkage disequilibrium structures present in the dense SNP panel data and the block correlation structures among transcript levels (Figure 5B) and epigenetic marks (Figure 5C).

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 (Additional file 6: 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 1, Additional file 8:Table S4). Similar observations were obtained for EPISPOT (Table 1, Additional file 7: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 modelling of all transcripts and SNPs achieved by M-EPISPOT, EPISPOT and ATLASQTL.

A focus on two susceptibility loci

We next discuss two examples resulting from a closer inspection of the analysed 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*, *DPY30*, *KLHL28* and *OSTC*. 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 novel 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 3′ UTR of the gene *GNPTAB*, 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 nor has any *trans*-acting association involving rs10860784 (up to proxies using $r^2 > 0.8$). However rs10860784 is known to be *cis*-acting on gene *DRAM1* (located 98 Kb downstream) in multiple tissues [24], an association which M-EPISPOT also confirms using a looser FDR of 15%. Moreover,

the UK Biobank PheGWAS also reported a strong association between this SNP and height ($p = 1.47 \times 10^{-14}$, see [25]).

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-value $< 2 \times 10^{-16}$).

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 associations 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 [26]. 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* (see [27] for a list), yet none provided a functional characterisation that would allow a clear prioritisation 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. However regressing out the effect of *LYZ* on the expression matrix does not explain away the hotspot effects, which argues somewhat against a mediation by *LYZ* (the size of the top hotspot in the *LYZ* locus is only marginally reduced: 134 versus 154 in the original M-EPISPOT analysis, see Additional file 4).

The monocyte-specific DHS annotation selected by M-EPISPOT for the *pleiotropic module* suggests another 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* [28]. This hypothesis of monocyte-specific pleiotropy would also explain

why no hotspot was reported so far in cell types and tissues other than monocytes [19, 29]. To investigate this further, we performed a complementary enrichment analysis using the multiple tissue/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. In addition, the distal gene *CREB1*, located on chromosome 2, overlaps this histone mark. This gene has previously been suggested as a putative mediator of the *LYZ* pleiotropic network [19]. Unlike for *LYZ*, 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 substantially lower (Additional file 4).

We further explored whether the two sets of genes associated with rs10784774, respectively 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 transcription factors, including ATF3, CREB1, c-Myc (Additional file 9: Table S5). The networks of transcription factors for rs10784774 and rs2168029 are similar, indicating that the two SNPs may be proxies for a same causal variant or, at least, that they share a same biological function. 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. The c-Myc transcription factor is involved in cell division and has broad transcriptional consequences [30], 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 [25].

Although by no means conclusive, these observations corroborate the context specificity of the *trans* effects controlled by the *LYZ* locus, possibly with *CREB1* acting as a monocyte-specific mediator of the hotspot network. They also suggest that our epigenome-driven EPISPOT runs found promising candidate hotspots, whose presumed mechanisms of action on the massive *LYZ* gene network would merit experimental follow up.

Finally, our TFBS analysis also indicated a somewhat weaker enrichment for the set of genes controlled by the smaller hotspot rs10860784 (Figure 5D-ii) located downstream to the above *LYZ* locus. Although its network of TFBS partly overlaps with those of the lead hotspots rs10784774 and rs2168029 from the *LYZ* locus, there are notable differences. In particular, no strong enrichment is seen for ATF transcription factors, which may hint towards distinct pathways for the two loci (Figure 5D-i & ii). rs10860784 may control a more generic gene network, as opposed to the monocyte-specific network of the *LYZ* locus.

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 [26, 31]. 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 analysed 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.

As it is located in a 3′ UTR, the rs10860784 hotspot discussed above is a good representative of this enrichment. In addition, three candidate hotspots, namely, 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 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* and other distal genes, including *IP6K2* on chromosome 3.

This locus has been linked with several autoimmune diseases [32–35] including type 1 diabetes, where evidence exists that *RPS26* transcription does not mediate the disease association [36]. 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 [19, 37]. 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 [38].

Discussion

Large panels of epigenetic marks are nowadays collected along with genetic data and expression levels, however their use to enhance the detection of QTL effects remains mostly heuristic. Thanks to its hypothesis-free mark selection routine which is fully integrated within a joint QTL mapping framework, EPISPOT can tell apart the relevant epigenetic marks from thousands of candidates while also directly refining the estimation and interpretation of associations in large molecular QTL studies.

Epigenetic marks have been used in different modelling approaches, such as for single-trait association studies (e.g., FINDOR [39]) or fine mapping (e.g., FINEMAP [40]). EPISPOT, however, is unique in several fundamental respects. First, to the best of our knowledge, it is the first approach that exploits the epigenome to enhance molecular QTL mapping, with a special emphasis on detection of hotspots. The fully Bayesian model implemented in EPISPOT parametrises proximal *cis and* distal *trans* action 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, 9, 41]. Second, it is both fully joint and scalable, accounting simultaneously for all epigenetic marks, genetic variants and molecular levels and their shared signals. Third, it combines this information to perform an *automated selection of the epigenetic marks relevant to the primary QTL associations*, 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 linkage disequilibrium, 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.

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 which 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 at GitHub [18, 42].

Finally, we showed how our simulation studies prefigured the efficiency of EPISPOT in a large monocyte eQTL study (high replication in an independent sample, novel 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 localised 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.

Conclusion

EPISPOT opens new perspectives for robust and interpretable molecular QTL mapping. 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. Importantly, with EPISPOT it

is finally possible to forgo the daunting and underpowered task of one-mark-at-a-time enrichment analyses for the prioritisation of QTL signals.

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 applies to any type of molecular QTL problem, involving genomic, proteomic, lipidomic or metabolic 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.

Methods

Model. The Bayesian hierarchical model implemented in EPISPOT comprises two levels of hierarchy (see graphical representation in Figure 1B). The bottom level parametrises the QTL effects and the top level parametrises the epigenetic modulations of the primary QTL effects.

Specifically, the bottom-level hierarchy consists of the following multiple-response spike-and-slab regression model:

$$\begin{aligned} \boldsymbol{y}_t \mid \boldsymbol{\beta}_t, \tau_t &\sim \mathcal{N}_n\left(\boldsymbol{X}\boldsymbol{\beta}_t, \tau_t^{-1}\boldsymbol{I}_n\right), && t = 1, \ldots, q, \\ \beta_{st} \mid \gamma_{st}, \sigma^2, \tau_t &\sim \gamma_{st}\,\mathcal{N}\left(0, \sigma^2\,\tau_t^{-1}\right) + (1-\gamma_{st})\,\delta_0\,, && s = 1, \ldots, p, \end{aligned} \tag{1}$$

where $\boldsymbol{y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_q)$ is an $n \times q$ matrix of centred responses (molecular traits), $\boldsymbol{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_p)$ is an $n \times p$ matrix of centred candidate predictors for them (SNPs), for n samples. 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 γ_{st} then yields marginal posterior probabilities of inclusion (qtl-PPIs), $\mathrm{pr}(\gamma_{st} = 1 \mid \boldsymbol{y})$. Moreover, the precision parameters τ_t and σ^{-2} are assigned diffused Gamma priors.

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

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

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 (centred) predictor-level covariates (epigenetic marks). This probit representation also allows decoupling the contributions of the predictors and the responses: it involves 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 influences 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 involves a local-scale feature (via s_{0s}) which effectively prevents overshrinkage; see our previous work on the modelling of hotspots from which this formulation is borrowed [5]. The value of s_{0s} is set by empirical Bayes, and so are the epigenetic effect hyperparameters ω_l and s (see hereafter). 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 (Additional file 1).

When a partition of the responses into modules is available, it can be supplied to the algorithm (then called M-EPISPOT) which will infer module-specific epigenetic effects based on the following modification of the top-level model (2):

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

where $m \in \mathcal{M}$ is a module of traits, with $\mathcal{M}$ a partition of $\{1, \ldots, q\}$ and $m \ni t$. Parameter $\boldsymbol{\xi}_m$ 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, which 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.

Annealed variational expectation-maximisation inference. The algorithm implemented in EPISPOT is designed to meet accuracy and scalability requirements in problems with thousands of molecular traits, SNPs and epigenetic marks. It consists of an annealed variational expectation-maximisation (VBEM or variational-EM) approach, with adaptive and parallel schemes.

VBEM algorithms were introduced by Blei et al. (2003) [43] in the context of Dirichlet allocation modelling. In short, they iterate between optimising 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 $\boldsymbol{v} = \left(\boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\gamma}, \sigma^2, \boldsymbol{\theta}, \boldsymbol{\zeta}, \boldsymbol{\xi}, \boldsymbol{\rho}\right)$ denote the parameters for model (1)–(3), and let $\boldsymbol{\eta} = (\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_M)$ denote the second-stage model hyperparameters, with $\boldsymbol{\eta}_m = (\boldsymbol{s}_{0m}^2, s_m^2, \boldsymbol{\omega}_m)$ for module $m = 1, \ldots, M$. We propose estimating $\boldsymbol{\eta}$ via an empirical Bayes procedure, by finding

$$\hat{\boldsymbol{\eta}} = \arg\max_{\boldsymbol{\eta}} \ell\left(\boldsymbol{\eta}; \boldsymbol{y}\right), \tag{4}$$

where $\ell\left(\boldsymbol{\eta}; \boldsymbol{y}\right) = \log p\left(\boldsymbol{y} \mid \boldsymbol{\eta}\right)$ is the marginal log-likelihood. Computing (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 $\boldsymbol{v}$. The procedure implements alternating optimisations of the variational lower bound

$$\mathcal{L}(q; \boldsymbol{\eta}) = \mathrm{E}_q \log p(\boldsymbol{y}, \boldsymbol{v} \mid \boldsymbol{\eta}) - \mathrm{E}_q \log q(\boldsymbol{v}), \tag{5}$$

where $q(\boldsymbol{v})$ is the variational density for $p\left(\boldsymbol{v} \mid \boldsymbol{y}, \hat{\boldsymbol{\eta}}\right)$ for a current estimate $\hat{\boldsymbol{\eta}}$ and $\mathrm{E}_q\left(\cdot\right)$ is the expectation with respect to $q(\boldsymbol{v})$. More precisely, it initialises the parameter and hyperparameter

Algorithm 1: VBEM algorithm

Define: Parameters $\boldsymbol{v} = (v_1, \ldots, v_J)$, hyperparameters $\boldsymbol{\eta} = (\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_M)$, $\boldsymbol{\eta}_m = (\boldsymbol{s}^2_{0m}, s^2_m, \boldsymbol{\omega}_m)$.

1. VBEM runs for hyperparameter estimation

for *m = 1, . . . , M (parallel loop)* **do**

Input: Responses for module m, predictors and predictor-level covariates: $\boldsymbol{y}_m$, $\boldsymbol{X}$, $\boldsymbol{V}$
Output: Empirical Bayes hyperparameter estimate: $\hat{\boldsymbol{\eta}}_m$
initialise: $\boldsymbol{\eta}_m^{(0)}$, $t \leftarrow 0$
repeat
$t \leftarrow t + 1$

E-step (details in Additional file 2):

Input: Current hyperparameter value: $\boldsymbol{\eta}_m^{(t-1)}$
Output: Top-level model variational parameters: $\boldsymbol{\mu}_\theta$, $\boldsymbol{\sigma}^2_\theta$, $\boldsymbol{\mu}_\xi$, $\boldsymbol{\sigma}^2_\xi$, $\boldsymbol{\rho}^{(1)}$ (dropping label m)
repeat
for $j = \text{shuffle}(1, \ldots, J_m)$ **do**
$q_m\left(v_j; \boldsymbol{\eta}_m^{(t-1)}\right) \propto \exp\left\{\mathrm{E}_{-j} \log p\left(\boldsymbol{v}, \boldsymbol{y}_m \mid \boldsymbol{\eta}_m^{(t-1)}\right)\right\}$,
end
until convergence of all variational parameters (with adaptive tolerance);

M-step:

Input: Current variational parameter values: $\boldsymbol{\mu}_\theta$, $\boldsymbol{\sigma}^2_\theta$, $\boldsymbol{\mu}_\xi$, $\boldsymbol{\sigma}^2_\xi$, $\boldsymbol{\rho}^{(1)}$
Output: Updated hyperparameter value: $\boldsymbol{\eta}_m^{(t)}$
$s^2_{0m,s} \leftarrow \mu^2_{\theta,s} + \sigma^2_{\theta,s}, \quad s = 1, \ldots, p,$
$s^2_m \leftarrow \dfrac{\sum_{l=1}^r \rho_l^{(1)} \left(\mu^2_{\xi,l} + \sigma^2_{\xi,l}\right)}{\sum_{l=1}^r \rho_l^{(1)}}$
$\omega_{m,l} \leftarrow \rho_l^{(1)}, \quad l = 1, \ldots, r,$
$\boldsymbol{\eta}_m^{(t)} \leftarrow (\boldsymbol{s}^2_{0m}, s^2_m, \boldsymbol{\omega}_m)$

until convergence of $\boldsymbol{\eta}_m^{(t)}$;
$\hat{\boldsymbol{\eta}}_m \leftarrow \boldsymbol{\eta}_m^{(t)}$
end

2. Final variational run

Input: All responses, predictors and predictor-level covariates: $\boldsymbol{y}$, $\boldsymbol{X}$, $\boldsymbol{V}$, empirical Bayes hyperparameter: $\hat{\boldsymbol{\eta}}$
Output: Variational parameters
repeat
for $j = \text{shuffle}(1, \ldots, J)$ **do**
$q\left(v_j; \hat{\boldsymbol{\eta}}\right) \propto \exp\left\{\mathrm{E}_{-j} \log p\left(\boldsymbol{v}, \boldsymbol{y} \mid \hat{\boldsymbol{\eta}}\right)\right\}$
end
until convergence of all variational parameters;

vectors $\boldsymbol{v}^{(0)}$ and $\boldsymbol{\eta}^{(0)}$, and alternates between the E-step,

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

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

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

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 $\mathcal{L}\left(q^{(t)}; \boldsymbol{\eta}\right)$ with respect to each component of $\boldsymbol{\eta}$. This only

requires computing and differentiating the joint likelihood term $\mathrm{E}_q \log p(\boldsymbol{y}, \boldsymbol{v} \mid \boldsymbol{\eta})$ in (5), as the entropy term $-\mathrm{E}_q \log q(\boldsymbol{v})$ 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, see, e.g., [5, 6, 44] 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 model (1)–(2) or (1)–(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 (3) suggests that its hyperparameters may be estimated reasonably well by restricting the VBEM scheme to subproblems corresponding to each module, i.e., applying model (1)–(2) to the subsets of responses $\boldsymbol{y}_m$ 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 parallelisation across modules. Once all module hyperparameters are estimated, they are inserted into model (1)–(3) and variational inference is run on the entire dataset. A sketch of the procedure is given in Algorithm 1, and the full derivation is in Additional file 2.

To efficiently deal with the posterior multimodality induced by strong data dependence structures, we augment all variational schemes (within the E-step and the final run) with a simulated annealing routine, although for brevity this is not described in Algorithm 1; see Additional file 2 for details. Annealing introduces a so-called *temperature* parameter to index the variational distributions and control the level of separations 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 which concern less than a given proportion of SNPs (default 5%) are also discarded as insufficiently informative before the analysis.

Recommended use. EPISPOT is a refining tool for the detection and interpretation of hotspots. It is meant to be used for joint analysis of preselected genomic regions (*candidate loci*) and transcripts believed to be under genetic control. 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 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, as this procedure interrogates the same data twice, which may be subject to overfitting.

Data-generation design for the simulation studies. Given the remarkably complex and multifaceted biochemical processes involved in genetic regulation, multiple interrelated steps are required to generate realistic datasets with epigenome-induced QTL associations. For each of these steps, we take special care to accommodate a wide range of parameter settings in order to cover a variety of plausible regulatory programs. We will also focus on producing scenarios demonstrating pleiotropy, with hotspots of diverse "sizes" (numbers of associated traits).

The code for generating simulated traits, SNPs (real or simulated) and epigenetic marks (real or simulated), as well as association patterns across these three data types, can be found under the form of documented functions in the R package `echoseq` freely available online [42]; it can be employed to generate alternative association maps to those presented hereafter, under a panel of assumptions from which the user can choose. This resource can also be used as an independent tool to generate epigenome-driven synthetic QTL data which mimic real data conditions.

We base all our numerical experiments on real genetic data which we supplement with data simulated according to generally-accepted principles of population genetics. To fix ideas, we present the steps for the general scenario with M modules; the canonical scenario with no module is readily obtained by choosing $M = 1$.

- **Simulation step 1 — independent loci from real genotyping data**. We start by building the $n \times p$ SNP matrix $\boldsymbol{X}$ by concatenating loci from quality-checked genotyping data for $n = 413$ healthy European individuals with minor allele frequency > 0.05 [19, 22]. Namely, we draw the locus sizes from a Poisson distribution with a prespecified mean, and sample the loci from chromosome one, making sure that they are sufficiently far apart (i.e., separated by at least 150 SNPs which corresponds to a median size of 1 Mb) so as to be reasonably assumed "independent loci".

- **Simulation step 2 — epigenetic control map**. We then form the "control map" between the epigenetic marks and the SNPs. We randomly select up to three active SNPs from each locus, stopping when a prespecified total number of active SNPs has been reached; hence some loci may contain no active SNP. We similarly select r_0 active marks among r binary epigenetic marks to be simulated (see below); the remaining $r - r_0$ marks will have no role in the generation of the QTL associations. Not all QTL associations are expected to result from epigenetic modifications. To accommodate this, we specify a proportion of active SNPs whose QTL associations will be triggered by active marks; the remaining active SNPs will have QTL associations simulated independently of the action of the marks. The epigenome control map design is slightly more involved when $M > 1$ modules are simulated, since it must reflect the fact that distinct modules can be governed by distinct epigenetic processes. In other words, the set of active marks and their action on SNP activity are simulated as module-specific, i.e., the active traits within a given module are associated with SNPs whose activity is triggered by specific marks, and these active marks may differ from those triggering QTL associations with traits from another module. The number, r_{0m}, of active marks controlling each module corresponds to the minimum between r_0 and a draw from a zero-truncated Poisson distribution with parameter 1. All module-specific active marks are then placed randomly among the r_0 active marks, ensuring that each of the r_0 marks are assigned to at least one module.

- **Simulation step 3 — QTL control map**. We next generate the pleiotropic QTL association pattern. For each active SNP s, we choose a subset of active modules (in the non-module

case $M = 1$, this step is skipped). Then, for each of these active modules, we draw the proportion of its traits controlled by the SNP from a uniform distribution (first and third simulation study) or from a right-skewed Beta distribution, favouring large hotspots (second simulation study). We then randomly select the traits associated with SNP s within the module according to this proportion. These module-specific *hotspot propensities* therefore produce hotspots of different sizes within and across the active modules.

- **Simulation step 4 — epigenetic marks**. We then effectively generate a $p \times r$ binary matrix of marks $\boldsymbol{V}$ as follows. For each mark, we draw a proportion of SNPs falling in the mark from a left-skewed Beta distribution (so the mark concerns relatively few SNPs), and we randomly select the SNPs concerned by the mark. We code this mapping in $\boldsymbol{V}$ by assigning the value unity if and only if the SNP (row of $\boldsymbol{V}$) falls within the mark (column of $\boldsymbol{V}$) and we make sure that each mark concerns at least two SNPs. We then enforce that each entry of $\boldsymbol{V}$ corresponding to a pair of active mark and SNP whose activity is triggered by the mark is set to unity.

- **Simulation step 5 — molecular traits**. Given this matrix $\boldsymbol{V}$ and the above epigenome- and QTL-association maps, we generate a series of auxiliary variables in view of simulating the traits as the sum of a genetic component and an independent noise component. For each module m, we first obtain an $r \times 1$ regression vector $\boldsymbol{\xi}_m$ for the epigenetic marks, such that its nonzero entries correspond to active marks (i.e., inducing QTL associations with traits from module m) and have a log-normal distribution. Hence the marks have non-negative effects, thereby only increasing the potential of SNPs to be involved in QTL activity, but not decreasing it ("enrichment effect"). We then draw $z_{st} \sim \mathcal{N}(\zeta, 1)$, for all $s = 1, \ldots, p$, $t = 1, \ldots, q$, with a large negative mean $\zeta = -2.5$ to induce overall sparsity. For each active SNP s and module m controlled by it, we next proceed as follows: if the QTL activity is triggered by the epigenome, we set $z_{st} \leftarrow z_{st} + \boldsymbol{V}_s^T \boldsymbol{\xi}_m$ for all traits $t \in m$ associated with SNP s; if the activity is not triggered by the epigenome, we set $z_{st} \leftarrow z_{st} + \theta_{m,s}$ for all traits $t \in m$ associated with SNP s, where $\theta_{m,s}$ is a SNP-specific effect drawn from a log-normal distribution. We then obtain binary variables specifying the QTL association pattern, $\gamma_{st} = \mathbb{1}(z_{st} > q_{1-\alpha})$, for all $s = 1, \ldots, p$, $t = 1, \ldots, q$, where $q_{1-\alpha}$ is the $1 - \alpha$ empirical quantile of z_{st} ($s = 1, \ldots, p$, $t = 1, \ldots, q$), with α, a chosen proportion of pairwise associations. We next use these variables to generate the regression coefficients β_{st}. If $\gamma_{st} = 0$, we set $\beta_{st} = 0$. To set the β_{st} for which $\gamma_{st} = 1$, we first draw the proportion of a trait's variance explained by individual SNPs from a left-skewed Beta distribution to favour the generation of smaller effects and we then rescale these proportions so that the proportion of genetic variance of each trait does not exceed a prescribed value. The magnitude of the β_{st} derives from this value, and its sign is altered with probability 0.5. This implies an inverse relationship between minor allele frequencies and effect sizes, as expected under natural selection [45]. Finally, we build the $n \times q$ matrix of traits, $\boldsymbol{y} \leftarrow \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where the noise $\boldsymbol{\varepsilon}$ is a centred multivariate normal variable with covariance such that the traits are equicorrelated with coefficient drawn from the interval $(0, 0.25)$; for module-based scenarios, the noise component is modelled using a multivariate normal variable with block equicorrelation.

Here the number of traits, q, is either prespecified or drawn from a Poisson distribution with a given mean; in the third simulation study, we simulate five modules of traits, and the number of traits in each module corresponds to a Poisson draw with mean 50.

We evaluate statistical performance based on 32 data replicates for each scenario, and use the annealing-augmented version of our algorithm, with a geometric schedule for a grid of 100 inverse temperatures and with initial temperature $T = 5$.

Monocyte eQTL datasets, epigenetic annotations and selection of candidate susceptibility loci. The first dataset consists of genotyping and microarray expression measurements for $n = 413$ healthy European individuals [19, 22]. The genotyping data are preprocessed using standard quality control filters (SNPs with call rate $< 95\%$, violating the Hardy–Weinberg equilibrium assumption at nominal p-value level 10^{-4} and with minor allele frequency $< 5\%$ are discarded). The transcripts are also quality checked and only the levels with mean expression and IQR above the first quartile of their corresponding empirical distribution are kept. This results in retaining $p \approx 550$K SNPs and $q = 22,827$ transcripts for analysis. Known confounders, namely, age, gender and batch, are regressed out from the expression matrix. To account for hidden confounders, we first derive principal components for the transcripts screened as free of any genetic control (i.e., using a preliminary univariate screening with MATRIXEQTL [2] and retaining the transcripts whose association p-values with SNPs are all $> 10^{-6}$) and we regress out the first $k = 10$ components from the expression matrix, where $k = 10$ is obtained by maximising the number of *cis* and *trans* associations. Here and throughout the analysis, an effect between a SNP and a transcript is called *cis* eQTL if the SNP and the transcript are on the same chromosome and no more than 1 Mb apart, and it is called *trans* eQTL otherwise.

The application of ATLASQTL at a genome-wide level reveals substantial pleiotropy around the gene *LYZ*, on chromosome 12. To shed light on the genetic mechanisms underlying this pleiotropy, we focus our analysis on the genetic variants located on chromosome 12. We next describe the data preparation steps for the EPISPOT analysis.

- **Analysis step 1 — prescreening with ATLASQTL**. We first consider the ATLASQTL run on the first dataset and take note of all QTL associations involving SNPs from chromosome 12 based on a permutation-based Bayesian FDR threshold of 5% (using spline-based interpolation) on the posterior probabilities of inclusion. We obtain 382 *cis*, resp. 595 *trans* associations, involving 350 unique SNPs on chromosome 12 and 430 unique transcripts overall. We also define the *LYZ region* as encompassing all SNPs located < 1 Mb upstream or downstream to *LYZ*. SNPs in this region are responsible for 515 *trans* associations and only 22 *cis* associations, reflecting the high level of pleiotropy; in comparison SNPs outside the *LYZ* region control 360 *cis* and 80 *trans* associations.
- **Analysis step 2 — candidate loci**. We then prepare the second, independent monocyte eQTL dataset involving $n = 286$ healthy European individuals from the CEDAR study [20]. These data consist of imputed SNPs (Sanger Imputation Services with the UK10K+ 1000 Genomes Phase 3 Haplotype panels, see details in [20]) and $q = 12,771$ quality-checked microarray transcript levels, which are preprocessed using the same filtering- and confounding-adjustment procedure as for the first dataset. We look up all 350 active SNPs in this dataset and we form loci by gathering all imputed SNPs in a 25 Kb neighbourhood of each active SNP and then merging overlapping regions. This results in a total of $p = 1,543$ SNPs distributed into 195 loci on chromosome 12; we concatenate all the loci to form an $n \times p$ matrix $\boldsymbol{X}$ of candidate SNP predictors.
- **Analysis step 3 — epigenetic annotations**. We then retrieve epigenetic information for SNPs from the 1000 genome project, using a curated database [46]. This database gathers different genomic annotations, namely, DNase-I hypersensitivity sites (DHS) for a range of tissues

and cell lines, annotations on gene structures ($3'$ and $5'$ UTRs, protein-coding exons), as well as genome segmentation annotations reporting whether nearby histone modifications are in line with transcription start sites (TSSs), CTCF binding sites, enhancer activity, promoter-flanking regions or repressed chromatin. Moreover, the distance of genetic variants to their nearest TSS in the Ensembl gene database is also provided. With the exception of the distance to TSSs, each entry of a given mark is coded unity if the corresponding variant falls in the mark and zero otherwise. We further merge experimental replicates by taking the union of the annotations derived from the same tissue or cell type. We hence obtain a total of $r = 168$ candidate epigenetic annotations for all our candidate SNPs but three, which are excluded from the $\boldsymbol{X}$ matrix and from all downstream analyses. We last drop all binary marks which concern $< 5\%$ of the SNPs in $\boldsymbol{X}$ and gather the remaining marks in a $p \times r$ matrix $\boldsymbol{V}$, where now $p = 1,540$ and $r = 107$; $\boldsymbol{V}$ has no missing entry. Figure 5C shows the correlation structure among the candidate annotations in $\boldsymbol{V}$; a more detailed heatmap is provided in Additional file 4.

- **Analysis step 4 — modules of transcripts**. We then look up all 430 active transcripts in the second dataset; 50 transcripts are missing in this dataset and we assign the remaining 380 transcripts to two "modules" based on whether they were controlled by SNPs from the *LYZ* region in the first dataset (*pleiotropic module*, for "pleiotropic" QTL control) or not (*scattered module*, for "scattered" QTL control). We then augment each module by adding all transcripts highly correlated with any transcript in the module, starting with the *pleiotropic module* (Pearson correlation $\rho > 0.9$). This results in a partition with $q = 283 + 191$ transcripts in the *pleiotropic* and *scattered modules*, respectively. We gather all transcripts to form an $n \times q$ matrix $\boldsymbol{y}$ of traits.
- **Analysis step 5 — settings for the QTL methods**. Finally, to ensure common comparative grounds, we use the same settings for all the methods (M-EPISPOT, EPISPOT and ATLASQTL), i.e., annealing schemes with same schedule, as well as a prior average number of SNPs associated with each trait of 2 and a corresponding prior variance of 4 (see Additional file 1).

We base the replication of the ATLASQTL prescreening results on the transcript levels available in CEDAR and we map the hits up to proxy SNPs in a 1 Mb window.

Software

EPISPOT is implemented as an R package with C++ subroutines and is publicly available under the GNU General Public License version 3 (GPL3) [18].

Ethics approval and consent to participate

The studies were approved by the local human research ethic committees, namely, the Oxfordshire Research Ethics Committee (COREC reference 06/Q1605/55) [22] and the University of Liège Academic Hospital Ethics Committee [20]. Participants provided informed written consent, and all procedures were conducted in accordance with the Declaration of Helsinki.

Availability of data and materials

Fairfax et al. (2012, 2014) [19, 22] provide gene expression in $CD14^+$ monocytes and genotyping data from individuals with European ancestry. The raw expression data were generated with Illu-

mina HumanHT-12 v4 arrays and downloaded from ArrayExpress [47] (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 [19, 22] and https://www.well.ox.ac.uk/research/research-groups/julian-knight-group/research-projects/data-access.

The CEDAR dataset [20] 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 [47] (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. Preprocessed data are available at http://cedar-web.giga.ulg.ac.be/.

All statistical analyses were performed using the R environment (version 3.6.1) [48] and the synthetic datasets were generated using the freely available R package ECHOSEQ (version 0.3.0) [42]. The following resources were also employed for the data processing, method comparison and eQTL analysis, in addition to the EPISPOT software [18].

- ATLASQTL (version 0.1.4) [49];
- EnrichR (version 2.1) [50];
- Ensembl [51];
- Gene ATLAS (UK Biobank) [25];
- GTEx [52];
- GWAS Catalog [53];
- MATRIXEQTL (version 2.3) [54];
- PhenoScanner [55];
- PLINK (version v1.90b5.3) [56].

Funding

This research was funded by the UK Medical Research Council programme MRC MC UU 00002/10 (HR, SR), MC UU 00002/4 (EV, CW) and MR M0 13138/1 (LB), the Alan Turing Institute Fellowship number TU/B/000092 (HR, SR) and EP/N510129/1 (LB), and the Wellcome Trust WT107881 (EV, CW). BPF and IN are funded by a Wellcome Intermediate Clinical Fellowship to BPF (no. 201488/Z/16/Z).

Author's contributions

HR, SR and LB designed and developed the EPISPOT method. IN and HR contributed to the monocyte eQTL data preprocessing, with input from BPF. HR ran the simulation experiments, and HR and BPF performed the eQTL data analysis. BPF, HR, EV, CW interpreted the results

of the eQTL study. HR, LB and SR wrote the manuscript with input from all authors, and have primary responsibility for final content. All authors read and approved the final manuscript.

Acknowledgements

We thank Colin Starr for managing computational resources.

References

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

[2] A. A. Shabalin. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics*, 28:1353–1358, 2012.

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

[4] L. Bottolo, E. Petretto, S. Blankenberg, F. Cambien, S. A. Cook, L. Tiret, and S. Richardson. Bayesian detection of expression quantitative trait loci hot spots. *Genetics*, 189:1449–1459, 2011.

[5] H. Ruffieux, A. C. Davison, J. Hager, J. Inshaw, B. Fairfax, S. Richardson, and L. Bottolo. A global-local approach for detecting hotspots in multiple response regression. *The Annals of Applied Statistics*, 14:905–928, 2020.

[6] H. Ruffieux, A. C. Davison, J. Hager, and I. Irincheeva. Efficient inference for genetic association studies with multiple outcomes. *Biostatistics*, 18:618–636, 2017.

[7] H. Ruffieux, J. Carayol, R. Popescu, M. E. Harper, R. Dent, W. H. M. Saris, A. Astrup, A. C. Davison, J. Hager, and A. Valsesia. A fully joint Bayesian quantitative trait locus mapping of human protein abundance in plasma. *PLOS Computational Biology*, 16:e1007882, 2020.

[8] M. A. Quintana and D. V. Conti. Integrative variable selection via Bayesian model uncertainty. *Statistics in Medicine*, 32:4938–4953, 2013.

[9] J. Yang, L. G. Fritsche, X. Zhou, G. Abecasis, and International Age-Related Macular Degeneration Genomics Consortium. A scalable Bayesian method for integrating functional information in genome-wide association studies. *The American Journal of Human Genetics*, 101: 404–416, 2017.

[10] J. T. Leek and J. D. Storey. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genetics*, 3:e161, 2007.

[11] O. Stegle, L. Parts, M. Piipari, J. Winn, and R. Durbin. Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nature Protocols*, 7:500, 2012.

[12] P. Langfelder and S. Horvath. Eigengene networks for studying the relationships between co-expression modules. *BMC Systems Biology*, 1:54, 2007.

[13] C. Borel, S. Deutsch, A. Letourneau, E. Migliavacca, S. B. Montgomery, A. S. Dimas, C. E. Vejnar, Attar H., M. Gagnebin, and C. Gehrig. Identification of cis-and trans-regulatory variation modulating microRNA expression levels in human fibroblasts. *Genome research*, 21: 68–73, 2011.

[14] M. Fagny, J. N. Paulson, M. L. Kuijjer, A. R. Sonawane, C.-Y. Chen, C. M. Lopes-Ramos, K. Glass, J. Quackenbush, and J. Platig. Exploring regulation in tissues with eQTL networks. *Proceedings of the National Academy of Sciences*, 114:E7841–E7850, 2017.

[15] L. Zhu, J. Tripathi, F. M. Rocamora, O. Miotto, R. van der Pluijm, T. S. Voss, S. Mok, D. P. Kwiatkowski, F. Nosten, and N. P. J. Day. The origins of malaria artemisinin resistance defined by a genetic and transcriptomic background. *Nature communications*, 9:1–13, 2018.

[16] M. Wainberg, N. Sinnott-Armstrong, N. Mancuso, A. N. Barbeira, D. A. Knowles, D. Golan, R. Ermel, A. Ruusalepp, T. Quertermous, and K. Hao. Opportunities and challenges for transcriptome-wide association studies. *Nature Genetics*, 51:592–599, 2019.

[17] S. K. Sieberts and E. E. Schadt. Inferring causal associations between genes and disease via the mapping of expression quantitative trait loci. In D. J. Balding, I. Moltke, and J. Marioni, editors, *Handbook of Statistical Genomics*, volume 2, pages 697–733. John Wiley & Sons, Oxford, United Kingdom, 2019.

[18] Epispot r package. URL `https://github.com/hruffieux/epispot`. Accessed 17 July 2020.

[19] B. P. Fairfax, S. Makino, J. Radhakrishnan, K. Plant, S. Leslie, A. Dilthey, P. Ellis, C. Langford, F. O. Vannberg, and J. C. Knight. Genetics of gene expression in primary immune cells identifies cell type–specific master regulators and roles of HLA alleles. *Nature Genetics*, 44: 502, 2012.

[20] Y. Momozawa, J. Dmitrieva, E. Théâtre, V. Deffontaine, S. Rahmouni, B. Charloteaux, F. Crins, E. Docampo, M. Elansary, and A.-S. Gori. IBD risk loci are enriched in multigenic regulatory modules encompassing putative causative genes. *Nature communications*, 9: 2427, 2018.

[21] M. Rotival, T. Zeller, P. S. Wild, S. Maouche, S. Szymczak, A. Schillert, R. Castagné, A. Deiseroth, C. Proust, and J. Brocheton. Integrating genome-wide genetic variations and monocyte expression data reveals trans-regulated gene modules in humans. *PLoS Genetics*, 7:e1002367, 2011.

[22] B. P. Fairfax, P. Humburg, S. Makino, V. Naranbhai, D. Wong, E. Lau, L. Jostins, K. Plant, R. Andrews, and C. McGee. Innate immune activity conditions the effect of regulatory variants upon monocyte gene expression. *Science*, 343:1246949, 2014.

[23] B. Rakitsch and O. Stegle. Modelling local gene networks increases power to detect trans-acting genetic effects on gene expression. *Genome Biology*, 17:33, 2016.

[24] GTEx Consortium. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science*, 348:648–660, 2015.

[25] O. Canela-Xandri, K. Rawlik, and A. Tenesa. An atlas of genetic associations in UK Biobank. *Nature Genetics*, 50:1593–1599, 2018.

[26] X. Wen, F. . . Luca, and R. Pique-Regi. Cross-population joint analysis of eQTLs: fine mapping and functional annotation. *PLOS Genetics*, 11, 2015.

[27] L. Kolberg, N. Kerimov, H. Peterson, and K. Alasoo. Co-expression analysis reveals interpretable gene modules controlled by trans-acting genetic variants. *BioRxiv*, 2020.

[28] L. Chen, B. Ge, F. P. Casale, L. Vasquez, T. Kwan, D. Garrido-Martín, S. Watt, Y. Yan, K. Kundu, and S. Ecker. Genetic drivers of epigenetic and transcriptional variation in human immune cells. *Cell*, 167:1398–1414, 2016.

[29] N. Kerimov, J. D. Hayhurst, J. R. Manning, P. Walter, L. Kolberg, K. Peikova, M. Samovič a, T. Burdett, S. Jupp, and H. Parkinson. eQTL Catalogue: a compendium of uniformly processed human gene expression and splicing QTLs. *BioRxiv*, 2020.

[30] Miller, D. M. and Thomas, S. D. and Islam, A. and Muench, D. and Sedoris, K. c-Myc and cancer metabolism. *Clinical Cancer Research*, 18:5546–53, 2012.

[31] D. J. Gaffney, J.-B. Veyrieras, J. F. Degner, R. Pique-Regi, A. A. Pai, G. E. Crawford, M. Stephens, Y. Gilad, and J. K. Pritchard. Dissecting the regulatory architecture of gene expression QTLs. *Genome Biology*, 13:R7, 2012.

[32] J. A. Todd, N. M. Walker, J. D. Cooper, D. J. Smyth, K. Downes, V. Plagnol, R. Bailey, S. Nejentsev, S. F. Field, and F. Payne. Robust associations of four new chromosome regions from genome-wide analyses of type 1 diabetes. *Nature Genetics*, 39:857–864, 2007.

[33] N. J. Craddock and I. R. Jones. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447:661–678, 2007.

[34] H. Hakonarson, H.-Q. Qu, J. P. Bradfield, L. Marchand, C. E. Kim, J. T. Glessner, R. Grabs, T. Casalunovo, S. P. Taback, and E. C. Frackelton. A novel susceptibility locus for type 1 diabetes on Chr12q13 identified by a genome-wide association study. *Diabetes*, 57:1143–1146, 2008.

[35] R. P. Nair, K. C. Duffin, C. Helms, J. Ding, P. E. Stuart, D. Goldgar, J. E. Gudjonsson, Y. Li, T. Tejasvi, and B.-J. Feng. Genome-wide scan reveals association of psoriasis with IL-23 and NF-κB pathways. *Nature Genetics*, 41:199, 2009.

[36] V. Plagnol, D. J. Smyth, J. A. Todd, and D. G. Clayton. Statistical independence of the colocalized association signals for type 1 diabetes and RPS26 gene expression on chromosome 12q13. *Biostatistics*, 10:327–334, 2009.

[37] S. Kasela, K. Kisand, L. Tserel, E. Kaleviste, A. Remm, K. Fischer, T. Esko, H.-J. Westra, B. P. Fairfax, and S. Makino. Pathogenic implications for autoimmune mechanisms derived

by comparative eQTL analysis of CD4+ versus CD8+ T cells. *PLOS genetics*, 13:e1006643, 2017.

[38] H. Ongen, C. L. Andersen, J. B. Bramsen, B. Oster, M. H. Rasmussen, P. G. Ferreira, J. Sandoval, E. Vidal, N. Whiffin, and A. Planchon. Putative cis-regulatory drivers in colorectal cancer. *Nature*, 512:87–90, 2014.

[39] G. Kichaev, G. Bhatia, P.-R. Loh, S. Gazal, K. Burch, M. K. Freund, A. Schoech, B. Pasaniuc, and A. L. Price. Leveraging polygenic functional enrichment to improve GWAS power. *The American Journal of Human Genetics*, 104:65–75, 2019.

[40] C. Benner, C. C. A. Spencer, A. S. Havulinna, V. Salomaa, S. Ripatti, and M. Pirinen. FINEMAP: efficient variable selection using summary data from genome-wide association studies. *Bioinformatics*, 32(10):1493–1501, 2016.

[41] W. Chen, S. K. McDonnell, S. N. Thibodeau, L. S. Tillmans, and D. J. Schaid. Incorporating functional annotations for fine-mapping causal variants in a Bayesian framework using summary statistics. *Genetics*, 204:933–958, 2016.

[42] Echoseq r package. URL `https://github.com/hruffieux/echoseq`. Accessed 17 July 2020.

[43] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[44] P. Carbonetto and M. Stephens. Scalable variational inference for Bayesian variable selection in regression, and its accuracy in genetic association studies. *Bayesian analysis*, 7:73–108, 2012.

[45] J.-H. Park, M. H. Gail, C. R. Weinberg, R. J. Carroll, C. C. Chung, Z. Wang, S. J. Chanock, J. F. Fraumeni, and N. Chatterjee. Distribution of allele frequencies and effect sizes and their interrelationships for common genetic susceptibility variants. *Proceedings of the National Academy of Sciences*, 108:18026–18031, 2011.

[46] J. K. Pickrell. Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. *The American Journal of Human Genetics*, 94:559–573, 2014.

[47] A. Athar, A. Füllgrabe, N. George, H. Iqbal, L. Huerta, A. Ali, C. Snow, N. A. Fonseca, R. Petryszak, and I. Papatheodorou. Arrayexpress update–from bulk to single-cell expression data. *Nucleic acids research*, 47:D711–D715, 2019.

[48] The r project for statistical computing. URL `https://www.r-project.org/`. Accessed 17 July 2020.

[49] Atlasqtl r package. URL `https://github.com/hruffieux/atlasqtl`. Accessed 17 July 2020.

[50] enrichr. URL `https://amp.pharm.mssm.edu/Enrichr/`. Accessed 17 July 2020.

[51] Ensembl database. URL `http://grch37.ensembl.org/index.html`. Accessed 17 July 2020.

[52] The genotype-tissue expression (gtex) database. URL `https://gtexportal.org/home`. Accessed 17 July 2020.

[53] The gwas catalog database. URL `https://www.ebi.ac.uk/gwas/`. Accessed 17 July 2020.

[54] Matrix eqtl. URL `http://www.bios.unc.edu/research/genomic_software/Matrix_eQTL/`. Accessed 17 July 2020.

[55] Phenoscanner v2: A database of human genotype-phenotype associations. URL `http://www.phenoscanner.medschl.cam.ac.uk/`. Accessed 17 July 2020.

[56] Plink: Whole genome association analysis toolset. URL `http://zzz.bwh.harvard.edu/plink/`. Accessed 10 October 2018.

Additional Files

Additional file 1 — Hyperparameter specification for top-level priors. Format: PDF.

Additional file 2 — Derivation of the variational-EM algorithm. Details of the variational distributions and updates, variational evidence lower bound and EM hyperparameter updates. Format: PDF.

Additional file 3 — Addendum to simulation studies: null scenario. Format: PDF.

Additional file 4 — Addendum to monocyte eQTL case study.. Epigenetic annotations for the monocyte eQTL study and network analysis for the *LYZ* hotspot mediation effects Format: PDF.

Additional file 5 — Table S1: Annotated 5% FDR eQTL associations using the ATLASQTL prescreening. Format: XLSX.

Additional file 6 — Table S2: Annotated 5% FDR eQTL associations using M-EPISPOT (CEDAR) and posterior summary for epigenetic marks. Format: XLSX.

Additional file 7 — Table S3: Annotated 5% FDR eQTL associations using EPISPOT (CEDAR) and posterior summary for epigenetic marks. Format: XLSX.

Additional file 8 — Table S4: Annotated 5% FDR eQTL associations using ATLASQTL (CEDAR). Format: XLSX.

Additional file 9 — Table S5: Transcription factor enrichment analysis results for the sets of genes controlled by the top hotspots. Format: XLSX.