Genetic Variant Selection: Learning Across Traits and Sites

Laurel Stell* and Chiara Sabatti§

* Department of Health Research and Policy, Stanford University, Stanford, CA 94305
§ Departments of Health Research and Policy and Statistics, Stanford University, Stanford, CA 94305

ABSTRACT We consider resequencing studies of associated loci and the problem of prioritizing sequence variants for functional follow-up. Working within the multivariate linear regression framework helps us to account for the joint effects of multiple genes; and adopting a Bayesian approach leads to posterior probabilities that coherently incorporate all information about the variants’ function. We describe two novel prior distributions that facilitate learning the role of each variable site by borrowing evidence across phenotypes and across mutations in the same gene. We illustrate their potential advantages with simulations and re-analyzing a dataset of sequencing variants.

1. Introduction
Genome-Wide Association Studies (GWAS) have allowed human geneticists to compile a rather long list of loci where DNA variation appears to be reproducibly associated to phenotypic variability (NHGRI 2015). While these might represent only a subset of the portion of the genome that is important for the traits under study (Manolio et al. 2009), there is little doubt that understanding the characteristics and mechanisms of functional variants at these loci is a necessary next step. As resequencing becomes ever more affordable, follow-up investigations of GWAS loci often start with a comprehensive catalogue of their genetic variants in a sample of thousands of individuals, raising the question of how to sort through these results.

Among the many challenges, let us discuss two. First, common variants are often correlated
and it is difficult to distinguish their roles without accounting for the broader genetic background of the individuals who carry them. Second, rare variants are present in a small enough portion of the sample that statistical statements become impossible. With this in mind, it has been noted that (a) it is important to account for correlation between variants to obtain useful ranking; (b) we should increasingly be able to take advantage of the information gathered through other studies; and (c) Bayesian models provide a principled approach to guide variant prioritization. To adequately select among variants in the same locus (defined as a genomic region that might encompass multiple genes but that corresponds to the same association signal in a GWAS study), researchers have resorted to model selection approaches (Valdar et al. 2012) or approximations of the joint distribution of univariate test statistics (Faye et al. 2013; Hormozdiari et al. 2014). Prior information on variant annotation has been incorporated in models for eQTL (Veyrieras et al. 2008) and more recently for general traits (Pickrell 2014; Kichaev et al. 2014; Chung et al. 2014), and annotation programs increasingly attempt to include information on identified genetic loci (Wang et al. 2010). Prioritization often relies on Bayes’s theorem, and Bayesian methods have received renewed attention in the context of GWAS data analysis (Guan and Stephens 2011; Peltola et al. 2012a,b), genomic prediction (Gianola 2013), and the evaluation of heritability (Zhou et al. 2013).

In this context, we explore the advantages of a careful specification of the prior distributions on variants, by allowing sharing of information across multiple phenotypes and across neighboring rare variants. We are motivated by the analysis of an exome resequencing study (Service et al. 2014) in which individual level data is available for exomic variants at multiple genomic loci that have demonstrated evidence in GWAS of association to lipid traits. By design, the vast majority of measured variants is coding or in UTR, that is, in portions of the genome with high prior probability of harboring functional mutations. Annotation can help distinguish the role of synonymous variants, and conservation scores can be used to predict the effect of nonsynonymous ones; but annotation cannot be used to discount the importance of a large number of non-coding variants that one can expect to occur in a whole genome sequencing dataset. Measures on levels of High Density Lipoprotein (HDL), Low Density Lipoprotein (LDL), and triglycerides (TG) are available for the study subjects, and we are interested in capitalizing on the multidimensional nature of the phenotype. Prior
analyses of this dataset (Service et al. 2014; Bogdan et al. in press) have illustrated the importance and the challenges of multivariate linear models, and we explore here the advantages offered by carefully selecting priors for Bayesian models. Abstracting from the specifics of this dataset, we show how hierarchical prior distributions can be adapted to learn about the functionality of a variant by (i) looking across multiple phenotypes and (ii) aggregating the effects of multiple rare variants in the same gene. Since the power of Bayesian methods in borrowing information is well known, it is not surprising that others have explored their application in this context. For example, Yi et al. (2011) illustrate the use of priors to model a group effect for multiple rare variants, while Stephens (2013) describes models for the analysis of multiple traits. Our approach, however, is distinct from others in that it strives to achieve all of the following: (1) constructing a multivariate linear model that simultaneously accounts for the contributions of multiple genes and genomic loci; (2) providing inference on variant-specific effects—while linking information across traits and genomic sites; and (3) accounting for the large number of variants tested, effectively enacting a form of multiple comparison adjustment.

This paper is organized as follows. We devote Section 2 to introducing the novel priors in the context of the genetic model, using an approximation of the posterior distribution to illustrate their inferential implications. Section 3 describes the MCMC scheme used to sample the posterior, the setting used for simulations, and the criteria for comparison of methods. Section 4 presents the results of simulation studies highlighting the potential of our proposal, as well as the description of the analysis of the motivating dataset.

2. Prior distributions on genetic variants

One characteristic of a genetic study based on resequencing, as contrasted to genotyping, is that researchers aim to collect a comprehensive catalogue of all genetic variants. This has implications for the statistical models used to analyze the data and the prior assumptions. Let \( n \) be the number of subjects in the study and \( p \) the number of polymorphic sites assayed. We will use \( y_i \) to indicate the phenotypic value of subject \( i \) and \( x_{iv} \) the genotype of this subject at variant \( v \) (typically coded as
minor allele count). The simplest genetic model for a heritable phenotype is of the form

$$y_i = \sum_{k \in T} G_{ik} + \zeta_i,$$

where \(\zeta_i\) incapsulate all nongenetic effects and \(G_{ik}\) for \(k \in T\) represent the contributions of a set \(T\) of genes that act additively and independently. Without loss of generality and following a standard practice in GWAS, we will assume that the effects of nongenetic determinants of the phenotypes have been regressed out from \(y_i\) so that \(\zeta_i\) can be considered independent ‘error’ terms. Let us assume that the genetic effects are a linear function of minor allele counts so that

$$y_i = \sum_{v \in V} \beta_v X_{iv} + \epsilon_i,$$

for a set \(V\) of causal variants with \(\epsilon_i \text{iid } \mathcal{N}(0, 1/\rho)\). Although this assumption is substantial, it only has the role of simplifying notation. While (1) represents the true genetic architecture of the trait, the membership of \(V\) is unknown in a typical association study, so the relation between the phenotype and genetic variants is expressed as

$$y_i = \sum_{v=1}^{p} \beta_v X_{iv} + \epsilon_i, \quad \epsilon_i \text{iid } \mathcal{N}
\left(0, \frac{1}{\rho}\right),$$

summing over all variable sites and with the understanding that only an (unknown) subset of \(\beta = (\beta_1, \ldots, \beta_p)\) is different from 0. Below we will use the compact matrix notation \(y = X\beta + \epsilon\). Using (2) to describe the relation between traits and genotypes depends heavily on the assumption that a resequencing study assays all variants. In GWAS, on the other hand, causal variants might be untyped, which means their contributions are partially captured by correlated variants and partially included in the error term. It would still be meaningful in that context to use a linear model to link phenotype and genotypes. However, in GWAS, the errors cannot be assumed independent, and the interpretation of the coefficients of \(X\)—as well as their prior distribution—is substantially more complicated. We note that mixed effects models can be used to address the first concern (Kang et al. 2010).
The parameters in model (2) are $\beta$ and $\rho$; we now focus on their prior distribution. Following standard practice, we take $\rho \sim \text{Gamma}(\alpha_{\rho}, \lambda_{\rho})$. (See Guan and Stephens (2011) for another approach that specifically targets GWAS and relies on heritability information.) On the vector $\beta$, we want a prior that reflects our model selection goals and our understanding of the genetic architecture. There are several aspects to consider: (a) given the exhaustive nature of the genotyping process, we believe that most of the variants available do not directly influence the trait; (b) it seems reasonable that a variant that influences one trait (so that its effect size is definitely not zero) might also influence other traits; and finally (c) it appears likely that if a rare variant influences the outcome, other nearby rare variants might also have an effect. Our main goal is to describe prior distributions on $\beta$ that incorporate these beliefs. We start by recalling one class of priors that reflect (a) and then move on to generalizations that account for the sharing of information implied by (b) and (c). In what follows, we assume that the allele counts in the column of $X$ have been standardized to have mean zero and variance one.

### A. Priors incorporating sparsity

The prior belief that only a fraction of the typed variants has an effect on the phenotype is but one instance of what is a common assumption in high-dimensional statistics, i.e. that the parameter $\beta$ of interest is sparse. To specify a prior on $\beta$ that gives positive probability to vectors with a number of coordinates equal to zero, we rely on a construction by George and McCulloch (1993) and introduce a vector of indicator variables $Z$ such that $Z_v = 0$ implies $\beta_v = 0$. The $Z_v$ are iid Bernoulli with parameter $\omega$, which governs the sparsity of the model and has a Beta($A_{\omega}, B_{\omega}$) prior. Let $\beta_Z$ indicate the collection of elements of $\beta$ corresponding to nonzero elements of $Z$, and let $X_Z$ be the corresponding columns of $X$. It has been found useful to assume $(\beta_Z | Z, \rho, \tau) \sim \mathcal{N} \left(0, \frac{\tau^2}{\rho} \Sigma_Z \right)$, where $\Sigma_Z$ is a known matrix and $\tau \sim \text{Unif}(\tau_1, \tau_2)$ links the error variance to the size of the $\beta$ coefficients. In the literature, $\Sigma_Z$ mainly has one of two forms: $I_{|Z|}$ (the identity matrix of size $|Z|$, where $|Z|$ indicates the number of nonzero components of the vector $Z$) or $n(X_Z^T X_Z)^{-1}$, which is referred to as the g-prior (Zellner 1986) (and is a viable choice only when $|Z| < n$). Various views on the choice of $\Sigma_Z$ have been put forth (Chipman et al. 2001; Heaton and Scott 2010; Guan and Stephens 2011), but the strongest argument for the g-prior is that it provides computational benefits (see below). For
either choice of $\Sigma_Z$, all of its diagonal entries are equal, resulting in an equal prior variance for each of the $\beta_v$. Given the standardization of the columns of $X$, this implies that the original effect sizes are expected to be larger for rare variants than for common variants, which is reasonable.

![Sparse $\beta$ prior](image)

**Figure 1** Schematic representation of the *Sparse* prior distribution on $\beta$. Hyperparameters are indicated in blue. The red portion describes all random objects and their dependency structure; unless explicitly indicated, random variables are independent. Boxes identify variables that share one of the distributions depicted in black.

One of the advantages of the prior summarized in Figure (1) is that the derived posterior distribution can be analytically integrated with respect to $\omega$, $\beta$, and $\rho$. While a MCMC is still needed to fully explore the posterior and carry out inference, we can rely on a collapsed Gibbs sampler that focuses only on $\tau$ and the indicator variables $Z$. This reduces the computation at each iteration and improves its convergence rate (Liu 1994). The prior densities for $\tau$ and $Z$ will be denoted, respectively, as $f_\tau(\tau)$ and $f_Z(Z)$—the latter being easily obtained from the beta-binomial distribution assumed for $|Z|$. As shown in the appendix, integrating $\beta$ and $\rho$ out gives the marginal posterior density

$$f_{Z,\tau}(Z, \tau|y) \propto f_\tau(\tau)f_Z(Z) \left( \lambda_\rho + \frac{S_Z^2}{2} \right)^{-\left(\frac{n}{2} + \lambda_\rho\right)} \frac{\det(\Omega_Z)^{1/2}}{\tau^{|Z|} \det(\Sigma_Z)^{1/2}},$$

where $\Omega_Z^{-1} = X^T_Z X_Z + \tau^{-2} \Sigma_Z^{-1}$ and $S_Z^2 = y^T y - y^T X_Z \Omega_Z X^T_Z y$. Choosing $\Sigma_Z$ as in the g-prior leads to a simplification of the ratio in (3), thereby avoiding the evaluation of one determinant at each iteration.

Despite the need to evaluate numerically interesting summaries of the posterior of $Z$, we obtained an approximation (whose derivation and applicability is described in the appendix) to gain a general understanding of how hyperparameters and data contribute to the final inferential results. Specifically, we focus on the posterior expected value of $Z_v$, indicator of variant $v$, conditional on the indicators of
all other variants $Z_{[-v]}$. In the case of orthogonal regressors, this expectation can be approximated as

$$
E[Z_v | Z_{[-v]}, \tau, y]^{-1} \approx 1 + \tau \sqrt{n} \frac{B_\omega + p - |Z_{[-v]}| - 1}{A_\omega + |Z_{[-v]}|} (1 - \eta_v^2)^{n/2},
$$

(4)

where $\eta_v = x_v^T y / \sqrt{ny^T y}$ is approximately the correlation between variant $v$ and the trait. From (4), one gathers that increasing $|Z_{[-v]}|$, which is the number of variants already used to explain the trait, increases the chance of an additional variant $v$ to be considered relevant. This is a consequence of the fact that the parameter $\omega$, which describes the sparsity of $\beta$ and hence the degree of poligenicity of the trait, is learned from the data (rather than set at a predetermined value). When a large number of variants have been found relevant, the trait is estimated to be highly poligenic and hence it is judged more likely that an additional variant might contribute to its variability. On the other hand, augmenting the total number of genotyped sites $p$ will make it harder for any specific variant $v$ to be judged important; this is adjusting for the look-everywhere effect, an important step in gene mapping studies.

Now that we have introduced this basic framework, we can consider modifications that facilitate learning about the role of a variant across multiple traits and in the context of neighboring sites. We start with the first problem.

**B. Learning across traits**

One of the characteristics of current genetic datasets is the increased availability of multidimensional phenotypes. This is due partly to the automation with which many traits are measured and partly to the increased awareness that precise phenotypic measurements are needed to make progress in our understanding of the underlying biological pathways. Having records of multiple traits in the same dataset allows for cross-pollination of genetic information. On the one hand, if a genetic variant is functional, it can be expected to impact more than one phenotype. On the other hand, even if noise in one phenotype makes it hard to distinguish the predictive power of a causal variant from that of a non-causal neighboring variant, it is much less likely that multiple traits would have noise correlated in such a way that causal and non-causal variants are indistinguishable for all of them. With this in mind, let us generalize the variant selection problem described in the previous section to handle
multiple traits.

Extending the notation, let $y_t$ be the standardized values for trait $t$, $\beta_t$ the coefficients of $X$ in the mean $E[y_t]$, and $Z_t$ the corresponding indicator vector. We organize these by column in a $n \times q$ matrix $Y$, a $p \times q$ matrix $\beta$, and a $p \times q$ matrix $Z$. Also let $s_v = \sum_t Z_{vt}$ denote the number of traits associated with variant $v$, let $\beta_{Z_t}$ be the entries of $\beta_t$ corresponding to entries equal to one in $Z_t$, and let $X_{Z_t}$ be the corresponding columns of $X$. The data-generating model is $y_t = X_{Z_t} \beta_{Z_t} + \epsilon_t$ with $\epsilon_t \sim \mathcal{N}(0, \frac{1}{\rho_t}I_n)$, and the priors on $\rho_t$ are simple extensions of the one used previously: $\rho_t \ iid \ Gamma(\alpha_\rho, \lambda_\rho)$. Note that this model assumes that, conditionally on the genetic variants that influence them, the traits are independent; specifically, there are no shared environmental effects. This assumption might or might not be appropriate depending on context, but the prior distribution on $\beta$ that we are about to describe can be used also for models that do not rely on this assumption.

We want a prior for $\beta$ that continues to enforce sparsity but that allows learning about the role of a variant across traits. One possibility, first proposed by Jia and Xu (2007), is to introduce a variant-specific probability of functional effect $\nu_v$, constant across traits and a priori independent with $\nu_v \sim Beta(A_v, B_v)$, where $A_v$ and $B_v$ can capture annotation information. Following the setup of the previous section, we then take $(Z_{vt} | \nu_v)$ independent Bernoulli($\nu_v$), set $\beta_{vt} = 0$ whenever $Z_{vt} = 0$, and let $(\beta_{Z_t} | Z_t, \rho_t, \tau)$ be independent across $t$ with distribution $\mathcal{N}(0, \frac{\tau^2}{\rho_t^2} \Sigma_{Z_t})$. As before, $\tau \sim \text{Unif}(\tau_1, \tau_2)$.

As detailed in the appendix, we can derive an approximation analogous to (4):

$$E[Z_{vt} | Z_{[-(vt)}], \tau, Y]^{-1} \approx 1 + \tau \sqrt{\frac{B_v + q - s_{v,[-t]} - 1}{A_v + s_{v,[-t]}}} (1 - \eta_{vt}^2)^{n/2}, \quad (5)$$

where $\eta_{vt} = \frac{x^T y_t}{\sqrt{y_t^T y_t}}$ and $s_{v,[-t]} = \sum_{\ell \neq t} Z_{vt}$ tallies the number of phenotypes for which the variant $v$ has been judged relevant. This highlights a consequence of the selected prior distribution: as the total number of phenotypes $q$ here has taken the role of $p$ in (4), the role of each variant is judged not in reference to all the other variants but only in comparison to the effect of the same variant across traits. In other words, while there is learning across phenotypes, there is no adjustment for the multiplicity of queried variants. Bottolo et al. (2011) previously observed that sparsity of $Z_t$ could not be controlled by specification of the priors in this approach, and proposed letting $Z_{vt}$ have Bernoulli parameter $\nu_v \omega_t$ with independent priors on each factor.
We propose a different remedy by introducing another layer in the hierarchical priors. Let $W$ be a vector of indicator variables of length $p$: if $W_v = 0$, then $v_0 = 0$; if $W_v = 1$, $v_0 \sim \text{Beta}(A_v, B_v)$. We take $W_v$ iid Bernoulli($\omega_W$) with $\omega_W \sim \text{Beta}(A_W, B_W)$; the $(Z_{vt}|v_0)$ are independent Bernoulli($v_0$), as before. The schematic in Figure 2 summarizes this prior proposal. The existence of a specific

Across Traits $\beta$ prior

\[
\begin{align*}
W_v \sim \text{Bern}(\omega_W) & \quad Z_{vt} \sim \text{Bern}(v_0) \\
W_1 & \quad Z_{11} \quad \cdots \quad Z_{1q} \\
\vdots & \quad \vdots \quad \ddots \quad \vdots \\
W_p & \quad Z_{p1} \quad \cdots \quad Z_{pq} \\
v_0 & \begin{cases} 
0 & W_v = 0 \\
\text{Beta}(A_v, B_v) & W_v = 1 
\end{cases} \\
\beta_i & \begin{cases} 
\beta_1 = 0 & Z_{vi} = 0 \\
\beta_i \sim \mathcal{N}(0, \frac{\Sigma Z_i}{p_i}) & Z_{vi} = 1 
\end{cases}
\end{align*}
\]

Figure 2 Schematic representation of the Across Traits prior distribution on $\beta$. Hyperparameters are indicated in blue. The red portion describes all random objects and their dependency structure; unless explicitly indicated, random variables are independent. Boxes identify variables that share one of the distributions depicted in black.

parameter $v_0$ for each site $v$ allows variation in the average number of impacted traits per variant; some variants can be highly pleiotropic, while others are relevant for one trait only. The sparsity parameter $\omega_W$ is once again estimated from the data, allowing for multiplicity adjustment. The introduction of $W$ effectively specifies a hierarchical prior on $v_1, \ldots, v_p$; among the many possible ways to accomplish this, the one we adopt emphasizes the role of the sparsity parameter $\omega_W$ and is easily interpretable. The appendix presents an indicative approximation of the posterior conditional expected values of $W_v$ similar to (4) and (5). It depends on all phenotypes (enabling learning across traits), but the total number of variants $p$ has again become the leading factor for effective multiplicity correction. We will refer to this prior as learning Across Traits. We include the first proposal in some comparison studies, indicating it as the Unadjusted approach to emphasize the fact that it does not include an adjustment for multiplicity.

This may be an appropriate point at which to clarify the relation between the prior we are proposing

Genetic variant selection 9
and the traditional investigation of pleiotropy versus coincident linkage. The latter terminology
derives from linkage studies, where the nature of the signal is such that the localization of variants
with possible pleiotropic effects is possible only up to a certain genomic interval. This interval might
contain one variant affecting multiple traits or contain different variants, each affecting a subgroup
of the traits. First, it is worth noting that in this paper we are working in the context of association
studies, which allow for a much finer resolution than linkage studies. The occurrence of multiple
variants affecting multiple traits within the same linkage disequilibrium block is less likely, given that
LD blocks are shorter than linkage regions. Secondly, ours is a fixed effects model using sequence
data. We are aiming to estimate the specific effect of each variant rather than simply identifying a
locus with a random effects model. Our framework, then, automatically considers two options: one
variant affecting multiple traits or multiple variants affecting separate traits. The choice between
these two alternatives is made on the basis of the posterior probability of the two models. This being
said, it is important to recall that if two neighboring variants in LD affect two separate traits, the
posterior probabilities of the two alternative models might be similar. The prior we introduce favors
pleiotropy in the sense that it recognizes as likely that some variants affect multiple genes, but it does
not exclude the alternative explanation, allowing the data to tilt the posterior in either direction. We
have investigated this with simulations in the supplementary material in File S1.

C. Learning across sites

We now consider another form of ‘learning from the experience of others’ to improve our ability
to identify functional variants. We focus on rare variants, which are observed in just a handful of
subjects and for which it might be impossible to estimate individual effects. It is reasonable to assume
that if one rare variant in a gene has an impact on a trait, other rare variants in the same gene also
might be functional; with an appropriate hierarchical prior we might increase our ability to learn
the effects of these variants. Of course a similar assumption might also be reasonable for common
variants, but given that we observe these in a sufficiently large sample, we aim to estimate their
individual effect without convolving their signal with that of others.

The data-generating model is again (2). We define $r$ groups of variants, and we use $\gamma(v)$ to
indicate the group to which variant $v$ belongs. Let $G = (G_1, \ldots, G_r)$ be a vector of indicator variables
associated to the groups; we use these to link information from different variants. Specifically, if \( G_g = 0 \), then the proportion \( v_g \) of causal variants in group \( g \) is equal to zero; otherwise, \( v_g \sim \text{Beta}(A_g, B_g) \) (setting \( v_g = 1 \) for groups comprised of only one variant). The variant-specific indicators \( Z_v \) are iid Bernoulli with parameter \( v_{\gamma(v)} \). Similarly to prior specifications, \( (G_g|\omega_G) \) are iid Bernoulli(\( \omega_G \)) with \( \omega_G \sim \text{Beta}(A_G, B_G) \). This results in the partially exchangeable prior on \( \beta \) represented in Figure 3; the parameter \( v_g \) allows sharing information on functionality across all variants in the same group.

**Across Sites \( \beta \) prior**

\[
\begin{aligned}
G_g &\sim \text{Bern}(\omega_G) \\
Z_v &\sim \text{Bern}(v_{\gamma(v)}) \\
\rho &\sim \text{Gamma}(\alpha_p, \lambda_p) \\
\omega_G &\sim \text{Beta}(A_G, B_G) \\

v_g &\sim \begin{cases} 0 & G_g = 0 \\ \text{Beta}(A_g, B_g) & G_g = 1 \& p_g > 1 \\ 1 & G_g = 1 \& p_g = 1 \end{cases} \\
\tau &\sim \text{Unif}[\tau_1, \tau_2] \\
\beta &\sim \begin{cases} \beta_{1-Z} = 0 \\ \beta_Z \sim \mathcal{N}(0, \frac{\tau^2}{\tau} \Sigma_Z) \end{cases}
\end{aligned}
\]

*Figure 3* Schematic representation of the *Across Sites* prior distribution on \( \beta \). Hyperparameters are indicated in blue. The red portion describes all random objects and their dependency structure; unless explicitly indicated, random variables are independent. Boxes identify variables that share one of the distributions depicted in black.

As described in the appendix, the posterior conditional probability that a variant \( v \) belongs to the model depends on the overall number of groups, the number of groups considered relevant, and the number \( s_g = \sum_{\gamma(v) = g} Z_v \) of variants in the same group that are deemed functional. The prior distribution in Figure 3, which we refer to as learning *Across Sites*, allows one to achieve an effect similar to that of burden tests, while still providing some variant-specific information (which is in contrast, for example, to the proposal in Yi *et al.* (2011)).

### 3. Methods

#### A. MCMC sampling

While we have resorted to some analytical approximation for expository convenience, we explore the posterior distribution with MCMC. As previously mentioned, we can focus on sampling \( \tau \) and all indicator variables. We use a Metropolis-within-Gibbs scheme, with the proposal distributions
described below. For \( \tau \), the common practice of using a truncated Gaussian works well. The discrete indicator variables pose a greater challenge, even though having integrated out \( \beta \) allows us to work with sample space of fixed dimension, eliminating the need for a reversible jump MCMC. When there is only one layer of indicator variables \( Z \), the proposal consists of first choosing with equal probability whether to add or remove a variant and then choosing uniformly among the candidate variants the one for which to propose a change of status. If the prior distribution is described using higher level indicators as well, then proposed changes to both levels must be consistent. If an entry of \( W \) is changed from one to zero, the associated entries of \( Z \) also have to be zeroed; when proposing to change an entry of \( W \) from zero to one, the associated entries of \( Z \) are selected from the prior marginal. Additionally, there are proposal moves that leave \( W \) unchanged but then randomly select one of its nonzero entries and draw a proposal for the associated entries of \( Z \) in a fashion analogous to that described previously. Details of the algorithm are in the supplementary material.

These simple proposal distributions will have trouble in two situations. The most common is when two or more variants are strongly associated with a phenotype but are also strongly correlated with each other due to LD. Any specific Markov chain will tend to include one of the variants in the model, leaving out the rest. Another problematic situation is when the effects of two variants on a phenotype depend upon each other, so neither variant is likely to enter the model by itself, even if their joint inclusion would be favored by the posterior distribution. Others (Guan and Stephens 2011; Peltola et al. 2012a,b) have described proposal distributions that overcome these difficulties and that can be reasonably applied to our setting—even though we do not investigate this in detail, focusing on the description of novel priors.

The average \( \bar{Z} \) of realized values of \( Z \) can be used to summarize the evidence in favor of each variant. Given its practical importance, the basic convergence checks incorporated in our package are based on \( \bar{Z} \). By default, the R code distributed in the package ptycho starts four chains from different points, runs each chain for a specified number of MCMC iterations, computes the averages for each chain separately, and then checks the range \( \Delta \bar{Z} \) of these averages. Details on the MCMC can be found in the supplementary material. The algorithm is implemented in the R package ptycho (Stell 2015).
B. Evaluation of variant selection performance

To investigate the performance of the proposed priors, we apply them to simulated and real data. The posterior distribution can be summarized in multiple ways. One can look for the indicator configuration that receives the highest posterior, for example, or make marginal inference on each variant. Both computational and robustness considerations make it practical to rely on posterior averages $Z_{vt}$ for comparisons. In the Bayesian models, then, we consider selecting variant $v$ for trait $t$ if the posterior average $Z_{vt}$ is larger than a certain threshold $\xi \in (0, 1)$: $S_t \equiv \{v : Z_{vt} > \xi\}$.

For benchmarking purposes, we will also analyze the datasets with some non-Bayesian approaches. Specifically we will consider (a) the Lasso (Tibshirani 1996); (b) a set of univariate linear regressions (one for each trait and variant), leading to $t$-statistics used to test the hypotheses of no association $H_{vt} : \beta_{vt} = 0$ with multiplicity adjustment for the $pq$ hypotheses via the Benjamini-Hochberg (BH) procedure at level $\alpha$ (Benjamini and Hochberg 1995); and (c) multivariate regression including all possible variants, with subsequent tests on the $pq$ null hypotheses for each coefficient incorporating adjustment via the BH procedure at level $\alpha$. The set of selected variants is equivalent in (a) to the set of estimated nonzero coefficients and in (b) and (c) to the set of variants for which the $H_{vt} : \beta_{vt} = 0$ are rejected. We will refer to these approaches as (a) Lasso, (b) BH marginal, and (c) BH full.

The threshold $\xi$ for Bayesian selection, the penalty of the Lasso, and the level $\alpha$ of BH can all be considered tuning parameters. We will compare the results of different procedures as these are varied (see details in the supplementary material). We base our comparison on an empirical evaluation of power and FDR associated with the different methods. Specifically, for each simulation and each method of analysis, we calculate the proportion of causal variants that are identified and the proportion of selected variants that are in fact false discoveries. The average of these values across multiple simulations is what we refer to as power and FDR in the results. The Bayesian methods also provide an estimate of FDR: if $Z_{vt}$ is approximately the probability that variant $v$ is causal for trait $t$, then the mean of $(1 - Z_{vt})$ over the selected variants is the Bayesian False Discovery Rate. We let $\hat{BFDR}$ denote this mean and explore how well it approximates (or not) the realized FDP, evaluated across all traits and variants.
C. Genotype and phenotype data

Our work has been partially motivated by a resequencing study: Service et al. (2014) analyzed targeted exome resequencing data for 17 loci in subjects of Finnish descent (from the 1966 Northern Finland Birth Cohort (NFBC) and the Finland-United States Investigation of NIDDM Genetics study (FUSION)). While the original study considered six quantitative metabolic traits, we focus here on the fasting levels of High Density Lipoprotein (HDL), Low Density Lipoprotein (LDL), and triglycerides (TG), transformed and adjusted for confounders as in the initial analyses (see supplementary material). The genotype data was obtained by sequencing the coding regions of 78 genes from 17 loci that had been found by previous GWAS meta-analyses to have a significant association to one of the six traits. In addition, we had access to the first five principal components of genome-wide genotypes. The goal in Service et al. (2014) is to identify which variants in these loci are most likely to directly influence the observed variability in the three distinct lipid traits.

Data cleansing and filtering are described in detail in the supplement; here we limit ourselves to note that for the purpose of the simulation study, the collection of variants was pruned to eliminate 550 variants observed only once and to obtain a set of variants with maximal pairwise correlation equal to 0.3 by removing another 558 variants. We excluded singletons from consideration since it would not be possible to make inference on their effect without strong assumptions. Multiple considerations motivated our choice of selecting a subset with only modest correlations: (a) correlated variants make the convergence of MCMC problematic, which might impair our ability to understand the inference derived from the posterior distribution; more importantly, (b) it is very difficult to evaluate and compare the performance of model selection methods in the presence of a high correlation between variants; and finally, (c) statistical methods cannot really choose between highly correlated variants and the selection among these needs to rely on experimental studies. Let us expand on these last two points. Procedures that build a multivariate linear model, such as the Lasso, would select one out of multiple highly correlated variants that have some explanatory power for the response; approaches such as BH marginal would instead tend to select them all; and Bayesian posterior probabilities for each of the variants would reflect the fact that substitutes are available: there will be multiple variants with elevated (if not very high in absolute terms) posterior probability.
It becomes difficult to meaningfully compare FDR and power across these methods, substantially reflecting the fact that the problem is somewhat ill-posed: if multiple highly correlated variants are available, any of them can stand for the others, and it is arbitrary to decide on purely statistical grounds that one belongs to the model while the others do not. Since our goal here is to understand the operating characteristics of the procedures, we found it useful to analyze them in a context where the target is well-identified and the results easily comparable.

After the described pruning, the genetic data used in the simulations contains 5335 subjects and 768 variants. Genotypes were coded with minor allele counts, and missing values (0.04% of genotypes) were imputed using variant average counts for consistency with previous analysis. Observed minor allele frequencies range from $2 \times 10^{-4}$ to 0.5, with a median of 0.0009 and a mean of 0.02. There are 628 variants with MAF<0.01. Annotation information was obtained as in Service et al. (2014), resulting in 61% coding, 34% UTR, and the remainder intragenic. Prior to applying the selection methods, the five genetic principal components along with the intercept were regressed out of both $X$ and $Y$, and the columns of both were then standardized.

When studying a real dataset, however, investigators might not be comfortable with such a stringent level of pruning; one might be concerned that variants with important effect are eliminated and that one is essentially reducing the information content of the sample. Indeed, when analyzing real data, we used a much more comprehensive approach, as described in the case study section.

D. Simulation scenarios

We constructed two simulation scenarios: one to simply illustrate the advantages of the proposed priors and the other to investigate their potential in a set-up that models a real genetic investigation.

**Illustrative example: orthogonal $X$.** We set $n = 5000$, $p = 50$, $q = 5$, and $X = \sqrt{\frac{n-1}{n/p}} (I_p I_p \cdots I_p)^T$ so that $X^T X = (n - 1)I_p$. In generating $\beta$ and the responses, we want to cover a range of different signal-to-noise ratios. To achieve this, we sample values of the parameters using the distributional assumptions that we described in the specification of the priors. To explore the performance of the Across Traits and Across Sites models—both when they provide an accurate description of reality as well as when they do not—we use three rules to generate the probability with which each variant is
associated to each trait. (a) We sample one sparsity parameter $\omega$ for each trait and keep it constant across variants. (b) We sample a probability $\nu_v$ for each variant and keep it constant across traits. Finally, (c) we define groups of five variants and sample one probability $\nu_g$ of causality for each group of variants and each trait. Rules (a)–(c) are most closely reflected in the prior structure of the basic, Across Traits and Across Sites models, respectively; and we indicate them as exchangeable variants, pleiotropy, and gene effect. We generate 100 datasets per rule, each with $q$ responses, and analyze them with the described set of approaches. When using Bayesian methods, we rely on non-informative priors (see supplement for details).

**Actual genotypes $X$.** To explore the potential power and FDR in the analysis of the dataset with three lipid traits, we generate artificial phenotypes starting from the available pruned genotypes. We consider a mixture of possible genetic architectures. In the construction of each dataset, (a) one gene is selected uniformly at random for each phenotype and 3–4 of its rare variants are causal (gene effect); (b) 40 distinct common variants are selected uniformly at random and each has probability equal to 0.1 to be causal for each of the phenotypes (thereby substantially representing trait-specific variants); and, finally, (c) 10 additional common variants are selected uniformly at random and each has a probability 0.9 to be causal for each phenotype (pleiotropic effects). This results in traits that are on average determined by 3–4 rare variants in one gene, 4 common variants with effects on one trait only, and 9 common variants with effects across multiple traits. We generated a total of 100 such datasets, as detailed in the supplementary material.

**E. Data Availability**

The sequencing and phenotype data are available on dbGaP. The Northern Finland Birth Cohort 1966 (NFBC1966) study accession number is phs000276.v2.p1. The Finland-United States Investigation of NIDDM Genetics (FUSION) study accession number is phs000867.v1.p1, with the sequencing data in substudy phs000702.v1.p1. In both cases, the sequencing data used in this paper has molecular data type equal to ‘Targeted Genome’ rather than ‘Whole Exome’.
4. Results

A. Simulations

![Diagram showing plots comparing empirical FDR and power](image.png)

**Figure 4** Power (top) and BFDR (bottom) as a function of empirical FDR in the illustrative example. Each color indicates a different variant selection approach (see legend at the top). Displays in different columns are for different data-generating mechanisms.

**Illustrative example.** Figure 4 showcases the possible advantages of the priors we have described. The plots on the top row compare the empirical FDR and power of the different variant selection methods on the datasets with orthogonal X. Points along the curves are obtained by varying tuning parameters and averaging the resulting FDP and power across 100 simulated datasets. Our setting is such that BH full, BH marginal, Lasso and the basic Bayes model have very similar behaviors: the Across Traits and Unadjusted models achieve the highest power per FDR in the presence of pleiotropy and the worst power per FDR in the presence of gene effects; in contrast, the Across Sites model has maximal power in the presence of gene effects and worse power in the presence of pleiotropy. While it is not surprising that the most effective prior is the one that matches more closely the structural
characteristics of the data, it is of note that the loss of power deriving from an incorrect choice of the *Across Traits* or the *Across Sites* model is minimal for FDR values lower than 0.2, which are arguably the range scientists might consider acceptable (see File S1, Figure C for a detail of these values). On the bottom row of Figure 4, we compare the estimated BFDR with actual FDR for the Bayesian models; here the most serious mistake is in underestimating FDR, which would lead to an anti-conservative model selection. Once again it can be seen that the best performance is obtained with the prior that matches the data-generating process. Besides this, it is useful to analyze the behavior of the *Unadjusted* approach: its power increase per FDR in the presence of pleiotropy is less pronounced than that of the *Across Traits* model, substantially because the *Unadjusted* approach is too liberal, with a BFDR which is significantly underestimated. This is in agreement with the lack of adjustment for multiplicity indicated by (5). Results for alternate hyperparameters are in File S1, Figure D and Figure E.

**Generating phenotypes from actual genotype data.** Figure 5 shows the performance of the variant selection methods in the analysis of traits generated from actual genotype data, further emphasizing the potential gains associated with the proposed strategies. For given FDR, both the *Across Traits* and *Across Sites* priors lead to an increase in power over the other methods. This is due to the fact that phenotypes are generated assuming both pleiotropy and contributions from multiple rare variants in the same gene (gene effects). In the lower portion of Figure 5, we separate the power to recover rare variants with gene effects from that for trait-specific common variants ($\omega = 0.1$) and from that for common variants with pleiotropic effects ($\omega = 0.9$). As expected, the gains of *Across Traits* and *Across Sites* are for the portion of genetic architecture that is accurately reflected in these priors. The estimates BFDR are accurate, indicating that all three Bayesian priors correctly learned $\tau$ and the probabilities of function.

Finally, while we have relied on ROC-like curves to compare different approaches as the value of their tuning parameters vary, it is useful to focus on the operating characteristics of the standard ways of selecting the tuning parameters. By convention, the target FDR for BH is usually 0.05. For Lasso selection, the function *cv.glmnet* provides two choices for $\lambda$: minimizing the cross-validation error and using the one-standard error rule. In Bayesian approaches, one can select variants so that
Figure 5 In the top portion, power and BFDR as a function of empirical FDR in the simulation from actual genotype data. In the lower panels, power is calculated separately for the different types of variants: rare, common with trait-specific effects, and common with pleiotropic effects.
Table 1 FDR and power for specific choices of selection parameters applied to simulated traits with actual genotype data.

<table>
<thead>
<tr>
<th>Variant selection criteria</th>
<th>FDR</th>
<th>power</th>
</tr>
</thead>
<tbody>
<tr>
<td>BH full with $\alpha = 0.05$</td>
<td>0.037</td>
<td>0.32</td>
</tr>
<tr>
<td>BH marginal with $\alpha = 0.05$</td>
<td>0.079</td>
<td>0.37</td>
</tr>
<tr>
<td>Lasso min error $\lambda$</td>
<td>0.810</td>
<td>0.63</td>
</tr>
<tr>
<td>Lasso 1-se $\lambda$</td>
<td>0.046</td>
<td>0.26</td>
</tr>
<tr>
<td>basic $\hat{BFDR} \leq 0.05$</td>
<td>0.039</td>
<td>0.37</td>
</tr>
<tr>
<td>Across Traits $\hat{BFDR} \leq 0.05$</td>
<td>0.054</td>
<td>0.45</td>
</tr>
<tr>
<td>Across Sites $\hat{BFDR} \leq 0.05$</td>
<td>0.051</td>
<td>0.40</td>
</tr>
</tbody>
</table>

global $\hat{BFDR} \leq 0.05$. Table 1 compares FDR and power for these selection parameters; the Bayesian methods appear to control the target FDR and arguably result in better power. Analogous summaries for other decision rules are in File S1, Table B; here we simply remark that including variants such that $\hat{BFDR} \leq 0.05$ in this dataset was practically equivalent to selecting variants with posterior probability larger than 0.7. We will capitalize on this observation for the real data analysis.

File S1, Figure F shows the results of another set of simulations along the lines of a traditional investigation of pleiotropy versus coincident linkage; we give a very brief summary here. In the case of separate causal variants, the Across Traits prior may have a slight loss of power but is still much better than BH with p-values from the full model. In the case of pleiotropy, however, the Across Traits prior clearly has greater power per FDR.

B. Case Study: the influence of 17 genomic loci on lipid traits

We now turn to the analysis of the three lipid traits in the Finnish dataset. While resequencing data comes from 17 loci identified via GWAS, prior evidence of association is available only between some of these loci and some traits. In particular, four loci have no documented association with any of the three lipid traits we study; we include variants from these loci in the analysis as negative controls. (This is different from the work in Service et al. (2014), which examines only variants in loci specifically associated with each trait.)
Table 2 Summary of selections for BH with $\alpha = 0.05$, Lasso with $\lambda$ chosen by cv.glmnet, and Bayesian approaches with $\xi = 0.7$. The columns labeled $R$ and $V^*$ give, respectively, the number of variants selected across the entire study and in the four loci with no prior evidence of association to any of the lipid traits analyzed (CRY2, G6PC2, MTNR1B, and PANK1). BFDR reports, for Bayesian methods, the Bayesian FDR computed separately for each trait.

<table>
<thead>
<tr>
<th>Variant Selection</th>
<th>HDL</th>
<th>LDL</th>
<th>TG</th>
</tr>
</thead>
<tbody>
<tr>
<td>BH full p-values</td>
<td>13</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>BH marginal p-values</td>
<td>22</td>
<td>6</td>
<td>10</td>
</tr>
<tr>
<td>Lasso min error $\lambda$</td>
<td>134</td>
<td>40</td>
<td>80</td>
</tr>
<tr>
<td>Lasso 1-se $\lambda$</td>
<td>16</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>basic</td>
<td>21</td>
<td>0.046</td>
<td>8</td>
</tr>
<tr>
<td>Across Traits</td>
<td>19</td>
<td>0.084</td>
<td>8</td>
</tr>
<tr>
<td>Across Sites</td>
<td>25</td>
<td>0.064</td>
<td>5</td>
</tr>
</tbody>
</table>

Service et al. (2014) relied on univariate regression to test the association between each trait and each variant with MAF>0.01 and on burden tests to evaluate the role of nonsynonymous rare variants. Bogdan et al. (in press) re-analyzed the data relative to HDL with a set of model selection approaches, including the novel methodology SLOPE; to facilitate comparison with their results, we add SLOPE to the analysis methods considered so far. Groups for the Across Sites model were defined so as to mimic the burden tests in Service et al. (2014), which means a group with more than one variant contains all nonsynonymous variants with MAF<0.01 in the same gene.

We start by analyzing the pruned subset of variants used in the simulation studies and postpone a more exhaustive search, noting again that this allows for a more straightforward comparison of the variants selected by different methodologies. Table 2 compares the number of variants selected by various methods with specified tuning parameters. The column labeled $V^*$ shows the number of selected variants that are in a locus lacking any prior evidence of association to lipid traits. The Lasso with $\lambda$ chosen to minimize cross-validated prediction error clearly results in far too many selections, so we discard this approach for the remaining results. For Bayesian approaches, the threshold $\xi = 0.7$ results in average BFDR approximately controlled at the 0.05 level.
**Figure 6** Estimated variant effects on HDL. Each panel corresponds to a locus, the $x$-axis indicates the variant’s genomic position and the $y$-axis its regression coefficient (with the exception of BH marginal, only nonzero coefficients are represented). The color code of the panel titles indicates the presence/absence of prior evidence of association between the locus and HDL (turquoise/orange, respectively). Model selection methods are distinguished using plotting symbols, as indicated in the legend at the top.
Figure 6 illustrates the model selection results for HDL. (Analogous displays for the other two phenotypes are in File S1, Figure G and Figure H. Also, File S1, Table C, Table D, and Table E detail differences in selections between approaches.) Each display corresponds to a locus, with turquoise shading (rather than orange) used to indicate prior evidence of association to HDL. Variants are arranged according to their genomic positions in the loci, and the values of their estimated coefficients are plotted on the $y$-axis; with the exception of marginal BH, we display only nonzero coefficients. When available, a vertical black line indicates the position of the SNPs originally used to select the locus (‘Array SNP’).

There is substantial overlap among the results of various methods. Model selection approaches seem to generally agree with the findings in Service et al. (2014) (with Lasso 1-se the most discrepant, missing a number of the associations identified in Service et al. (2014); see File S1, Table C). Still, we can point to some significant differences. With the Across Traits approach we select two variants in two loci where no other method identifies any signal: in CELSR2 and FADS1. These two loci have prior evidence of association to LDL and to all three lipid traits, respectively, and the Across Traits approach identifies pleiotropic effects. In contrast, the Across Traits approach does not select four very rare (MAF<0.001) variants considered relevant by more than one alternative method. While we do not know where the truth lies at this point, it is very hard to evaluate the effect of a rare variant on purely statistical grounds, and the outcome of the Across Traits model might well be the more reliable.

The Across Sites approach identifies four variants that other approaches overlook. Three are rare variants in ABCA1: two missense rare (0.01>MAF>0.001) and one nonsense very rare (MAF 0.00016); their discovery is facilitated by the fact that they are included in a group with multiple other significant variants. The fourth is a common variant in the MVK locus, for which there is prior evidence of association to HDL. Other approaches do not recover this simply because the signal in the locus is, in the dataset analyzed here, barely below the detection threshold; Across Sites has a slight advantage over the other Bayesian methods because grouping reduces the number of comparisons to account for. We note that ABCA1 is a gene in which rare variants were found to have a role by the burden test analysis in Service et al. (2014).

We have relied on the pruned dataset since low correlation across variants greatly facilitates the
comparison of the selection results by different methods. However, mindful of the concerns of scientists unwilling to pre-screen the genotypic information, we have also carried out a more comprehensive analysis of this dataset, showcasing that this is indeed an option available to researchers. Details for these analyses are in the supplement, but we summarize them here. First, we have compared the results of four different levels of pruning (correlation less than 0.3, 0.5, 0.7 and 0.9). We have found that very few values of \( Z_{vt} \) change by more than 0.05 when different levels of pruning are used and that less stringent levels of pruning do not lead to substantially more findings—unlike when applying BH to marginal p-values. In fact, there is a greater tendency for variants to drop out of the selection set as correlated variants are added to \( X \). Second, to completely eliminate pruning, we analyzed all variants in one locus along the lines of Servin and Stephens (2007); Hormozdiari et al. (2014); Kichaev et al. (2014); Chen et al. (2015), using the basic prior and assuming that the number of significant regressors \( |Z| \) is no greater than five or six (depending on the total number of possible variants in the locus). We have restricted our attention to two loci only, those that showed stronger evidence of influencing HDL via multiple variants. File S1, Table H offers a precise comparison of results, but it suffices here to note that the set \( Z \) of variants with the largest posterior density is equal to the variants selected among the original pruned set (correlation less than 0.3) by the basic prior for one locus and, for the other locus, the two sets substantially overlap.

5. Discussion

As the genetics community devotes increasing effort to follow up GWAS hits with resequencing, a number of suggestions on how to prioritize variants have emerged. In truth, while dealing with the same broad scientific goals, many of these contributions address different aspects of the problem and therefore should be seen as complementary rather than alternatives; taken together they provide scientists with useful suggestions. Annotation information has been shown to be quite useful when the set of variants under consideration is sufficiently diverse. It is important to account for the correlation across variants to avoid paying attention to SNPs that are only ‘guilty by association.’ Bayes’s theorem and Bayesian statistics are a natural way of dealing with the decision of which variants to pursue. In this context, others have studied (a) how to choose priors that incorporate...
annotation information, tuning their parameters with available datasets; (b) how to approximate posterior distributions of variant effects; (c) how to sample from the posterior distribution using efficient MCMC or variational schemes; and (d) how to efficiently evaluate posterior probabilities for a set of variants. Here we focus on another aspect of prior selection: describing how partial exchangeability assumptions can be used to borrow information across traits and neighboring sites, while maintaining an effective control for multiplicity across variants and fitting multivariate regression models that estimate the specific contribution of each associated site, while accounting for others. We briefly refer to some of the most direct antecedents of our proposed priors to underscore relevant differences.

Yi et al. (2011) proposed the use of hierarchical priors to capture effects of rare variants through groups, similar to the Across Sites model. However, their proposal does not incorporate sparsity considerations, resulting in the estimate of a nonzero effect for each variant and each group and therefore not engaging in model selection. Quintana et al. (2011, 2012) took an additional step towards the Across Sites model by incorporating sparsity via the indicator variable $Z$. They considered only rare variants, used the same effect size for all rare variants in a genomic region, used the MLE for the effect sizes rather than integrating them out, and, most importantly, controlled sparsity by using $A_g = 1$ and $B_g = p_g$ in the prior for $\nu_g$ rather than introducing another layer of indicator variables in the hierarchical prior—all of which means their approach has less flexibility and less learning.

The Across Sites prior also echoes the proposal of Zhou et al. (2010) who suggested the use of group penalization in Lasso to estimate multivariate sparse models while encouraging coordinated selection of rare variants in the same gene. This computationally appealing approach has not become as popular in genetics as in many other fields, possibly because of the difficulties connected with the selection of its tuning parameters when model selection is the goal. Cross validation is often used to determine the appropriate level of penalization; while this works well for prediction purposes, its performance is less than satisfactory in identifying variants with truly nonzero coefficients (as illustrated by our case study). Alexander and Lange (2011), Valdar et al. (2012), and, most recently, Sabourin et al. (2015) explore coupling resampling techniques with Lasso penalization to improve model selection. This not only increases computational costs but also greatly reduces the initial
simplicity of the model. As documented in Bogdan et al. (in press), identifying a single \( \lambda \) value that performs FDR control is challenging; Yi et al. (2015) investigates this task in the context of GWAS and provides guidelines. The final model selection of these machine learning approaches uses complex rules; in contrast, the Bayesian models we described are based on easy to interpret parameters.

The use of hierarchical Bayesian methods has ample precedents in eQTL studies, where they have been used to correct for multiplicity (Kendziorski et al. 2006) and to increase power of detecting variants affecting multiple traits. In our presentation of the Unadjusted and Across Traits approaches, we referred to methods proposed by Jia and Xu (2007) and Bottolo et al. (2011). More recent work (Flutre et al. 2013; Li et al. 2013) has focused on the identification of local (cis) effects across tissue, and considered models with only one functional variant. The recent contribution by Chung et al. (2014)—which appeared while this work was in preparation—underscores as we do the importance of learning both across sites and across traits to prioritize variants. These authors, however, work with p-values from GWAS studies, rather than actual resequencing data.

Having clarified the scope of our contribution, we want to briefly mention how it could be extended and combined with suggestions by others. First, let us point out that while in the simulations and in the analytical approximations we assumed \( n > p \), this restriction is by no means necessary to the Bayesian model we describe. On the contrary, the priors we propose—by learning sparsity and giving positive probabilities to configurations with some \( \beta_v = 0 \)—are well-suited to the case \( n < p \). The real challenge in dealing with GWAS-type data would be from a computational standpoint: increased mixing for MCMC as described in Xu et al. (2014) and Guan and Stephens (2011) or other algorithmic improvements (Carbonetto and Stephens 2012; Hormozdiari et al. 2014) would make our approach more widely applicable.

Another extension that is easily achieved is the combination of the Across Traits and Across Sites priors. Most immediately, the group indicators \( G \) in Figure 3 can be made trait-specific and linked across phenotypes with the same approach used to link the \( Z_t \) in Figure 2.

It is certainly possible to combine the partial exchangeability aspects of our models with a prior that incorporates annotation information. Refer, for example, to the Across Traits prior in Figure 2. Currently, the distribution on \( W_1, \ldots, W_p \), indicators of functionality of the variants, is a beta-binomial.
However, it is trivial to change it to a mixture of independent logits, with the linear model component including an intercept effect—which would capture the overall sparsity—and a linear combination of annotation indicators (Veyrieras et al. 2008; Kichaev et al. 2014; Pickrell 2014).

Since our focus has been on specification of the prior, we have not paid much attention to the data-generating model, which could certainly be improved. Specifically, we want to underscore the fact that using a mixed-model approach might be advisable to account for population structure (Kang et al. 2010) and when analyzing many phenotypes whose quantitative value might be influenced by confounders (Zhou and Stephens 2014) or simply by genetic variants not included in the model.

In conclusion, we want to emphasize the increasing importance in human genetics of models that account for pleiotropy. ‘Big data’ in genetics has often been equated with the abundance of sequences, and these certainly pose a number of management and interpretation challenges. Our increased acquisition capacity will also result, however, in the collection of a large number of phenotypes; gene expression, MRI scans, and mass spectrometry are just some examples of the large-scale phenotyping efforts underway. Now that DNA variation has been extensively described, annotating this appears as a fundamental challenge; the rich phenotypic collections increasingly available have a major role to play. After all, what better way of establishing if a variant has some functional impact than looking for its association with any trait available? Bayesian models that allow one to estimate the probability with which a variant has functional effects across phenotypes are likely to be very useful. In this paper, we have described a first step in this direction.

**Acknowledgements**

L. Stell and C. Sabatti were partially supported by NIH grants HG006695, HL113315, MH105578, and MH095454. We thank the authors of Service et al. (2014) for letting us use their data during the completion of dbGaP release. Finally our gratitude to H. Tombropoulos for unwavering editorial assistance.

**Literature Cited**


Heaton, M. J. and J. G. Scott, 2010 Bayesian computation and the linear model. In Frontiers of Statistical


Peltola, T., P. Marttinen, A. Jula, V. Salomaa, M. Perola, and A. Vehtari, 2012a Bayesian variable selection

Genetic variant selection
selection in searching for additive and dominant effects in genome-wide data. PLOS ONE 7: e29115.


Valdar, W., J. Sabourin, A. Nobel, and C. C. Holmes, 2012 Reprioritizing genetic associations in hit


Yi, N., N. Liu, D. Zhi, and J. Li, 2011 Hierarchical generalized linear models for multiple groups of rare and common variants: jointly estimating group and individual-variant effects. PLOS Genet. 7: e1002382.


A. Mathematical details for the basic prior

First we integrate $\beta$ out of the posterior distribution as in (Chen et al. 2015). Since $y = XZ\beta + \epsilon$ with $(\beta|Z, \rho, \tau) \sim \mathcal{N} \left( 0, \frac{\tau^2}{\rho} \Sigma_Z \right)$ and $\epsilon \sim \mathcal{N} (0, \frac{1}{\rho}I_n)$,

$$(y|Z, \rho, \tau) \sim \mathcal{N} \left( 0, \frac{1}{\rho}I_n + \frac{\tau^2}{\rho}X_Z\Sigma_Z X_Z^T \right).$$

(6)

While the likelihood can be written directly from this, a few manipulations give it in a more convenient form. Sylvester’s determinant theorem (Harville 2008, p. 420) implies

$$\det(I_n + \tau^2X_Z\Sigma_Z X_Z^T) = \det(I|Z| + \tau^2\Sigma_Z X_Z^T X_Z) = \tau^2|Z| \frac{\det(\Sigma_Z)}{\det(\Omega_Z)},$$

and a generalization of the Woodbury matrix identity (Harville 2008, p. 428) implies

$$(I_n + \tau^2X_Z\Sigma_Z X_Z^T)^{-1} = I_n - X_Z(\tau^{-2}\Sigma_Z^{-1} + X_Z^T X_Z)^{-1}X_Z^T = I_n - X_Z\Omega_Z X_Z^T.$$ 

Consequently, $\Pr(y|Z, \rho, \tau) = \left(\frac{\rho}{2\pi}\right)^{n/2} \frac{\det(\Omega_Z)^{1/2}}{|Z|^{1/2} \det(\Sigma_Z)^{1/2}} e^{-\rho S_Z^2/2}$. For the null model $Z = 0$, (6) shows that the covariance matrix is $\rho^{-1}I_n$, so in this case $S_Z^2 = y^T y$ and the ratio of determinants is set equal to one.

Next multiply $\Pr(y|Z, \rho, \tau)$ by the prior density function for $\rho$ and integrate to obtain

$$\frac{\det(\Omega_Z)^{1/2}}{\tau|Z| \det(\Sigma_Z)^{1/2}} \int \rho^{n/2} e^{-\rho S_Z^2/2} \rho^{a_\rho} e^{-\lambda_\rho \rho} d\rho \propto \frac{\det(\Omega_Z)^{1/2}}{\tau|Z| \det(\Sigma_Z)^{1/2}} \left( \lambda_\rho + \frac{S_Z^2}{2} \right)^{-\left(\frac{n}{2} + a_\rho\right)},$$

since the integrand is the density function of Gamma $\left( a_\rho + \frac{n}{2}, \lambda_\rho + \frac{1}{2}S_Z^2 \right)$, up to a normalizing factor. Hence, the marginal posterior for $Z$ and $\tau$ is given by (3).

Along the lines of Malsiner-Walli and Wagner (2011), we present an approximation of the posterior expected value $\mathbb{E}[Z_v|Z_{[-v]}, \tau, y]$. If $Z$ and $\tilde{Z}$ are equal except that $Z_v = 0$ and $\tilde{Z}_v = 1$ for one $v$, then

$$f_{Z, \tau}(\tilde{Z}, \tau|y) = \frac{\Pr(Z_v = 1|Z_{[-v]}, \tau, y)}{\Pr(Z_v = 0|Z_{[-v]}, \tau, y)} = \frac{\mathbb{E}[Z_v|Z_{[-v]}, \tau, y]}{1 - \mathbb{E}[Z_v|Z_{[-v]}, \tau, y]}.$$  

(7)
We will use several assumptions in order to simplify the expression on the left and then solve for $E[Z_0|Z_{-0}, \tau, y]$. Consider the case when the columns of $X$ are orthogonal, which implies $X^TX \approx nI_p$ and the two choices of $\Sigma_Z$ are essentially the same. Furthermore, $\langle x_v, y \rangle = \langle x_v, X\beta + \epsilon \rangle \approx n\beta_v + \langle x_v, \epsilon \rangle$, which is distributed as $N(0, \frac{n}{p}(Z_vn\tau^2 + 1))$; in this context, distinguishing signal from noise requires $n\tau^2 \gg 1$, so we assume this to be the case. Consequently, $\Omega^{-1}_Z \approx (n + \tau^{-2})I_p \approx nI_p$, which in turn implies $S^2_Z$ is approximately equal to the residual sum of squares (RSS) for the model indicated by $Z$. Finally, reflecting the results of current GWAS, we assume that the portion of variance explained (PVE) by the loci in consideration is rather small, so RSS is not much less than $y^Ty \approx n$ for any model. If one further chooses $\alpha_\rho = \lambda_\rho \ll \frac{n}{2}$, then

$$
\frac{f_{Z,\tau}(Z, \tau|y)}{f_{Z,\tau}(Z, \tau|y)} \approx \frac{1}{\tau} \frac{f_{Z}(Z)}{f_{Z}(Z)} \left( \frac{\det(\Omega_Z) \det(\Sigma_Z)}{\det(\Sigma_Z) \det(\Omega_Z)} \right)^{1/2} \left( \frac{S^2_Z}{S^2_Z} \right)^{-n/2}.
$$

Properties of the beta and gamma functions give that the ratio of $f_Z$ values is $(A_\omega + |Z|)/(B_\omega + p - |Z| - 1)$. Furthermore, $\sqrt{\frac{\det(\Omega_Z) \det(\Sigma_Z)}{\det(\Sigma_Z) \det(\Omega_Z)}} \approx 1/\sqrt{n}$. Finally, $S^2_Z \approx y^Ty - \frac{1}{\bar{n}} \sum_{u:Z_u=1}(x^T_u y)^2$, so

$$
\frac{S^2_Z}{S^2_Z} \approx \frac{S^2_Z - \frac{1}{\bar{n}}(x^T \bar{y})^2}{S^2_Z} \approx 1 - \frac{(x^T \bar{y})^2}{\bar{y}^T \bar{y}} \equiv 1 - \eta^2_v.
$$

Substituting these results into (7) and (8) gives (4).

**B. Mathematical details for learning across traits**

While the *Unadjusted* prior is not useful, we include its marginal posterior density here for completeness. Its derivation is very similar to that of the basic model, so we focus on the differences. A priori the rows of $Z$ are independent and each has a beta-binomial distribution, so $f_Z(Z) = \prod_{v=1}^p B(A_v + s_v, B_v + q - s_v)/B(A_v, B_v)$. Furthermore, the columns of $Y$ are independent given $Z$, $\beta$, and $\rho$, and similarly for the columns of $\beta$; so

$$
f_{Z,\tau}(Z, \tau|Y) \propto f_\tau(\tau)f_Z(Z) \prod_{t=1}^q \left( \lambda_\rho + \frac{S^2_{Z_t}}{2} \right)^{-\frac{(\frac{q}{2} + \alpha_\rho)}{2}} \frac{\det(\Omega_{Z_t})^{1/2}}{\tau|Z_t| \det(\Sigma_{Z_t})^{1/2}}.
$$
If $Z$ and $\tilde{Z}$ are equal except that $Z_{vt} = 0$ and $\tilde{Z}_{vt} = 1$ for one $v$ and one $t$, then the same calculations as for the basic model give that (8) simplifies as

$$f_{Z,\tau}(\tilde{Z}, \tau | y) \approx \frac{1}{\tau \sqrt{n}} \frac{A_v + s_v}{B_v + q - s_v - 1} (1 - \eta_{vt}^2)^{-n/2}.$$ 

This leads to the approximation (5).

Next we consider the Across Traits prior. The posterior density is the same as in (9) except that $f_Z$ is replaced by

$$f_{W,Z}(W, Z) = \int \Pr(Z | W, v) \Pr(W | \omega_W) \Pr(\nu) \Pr(\omega_W) dv d\omega_W$$

$$= \frac{B(A_W + |W|, B_W + p - |W|)}{B(A_W, B_W)} \prod_{v: W_v = 1} \frac{B(A_v + s_v, B_v + q - s_v)}{B(A_v, B_v)},$$

provided that $Z_{vt} = 0$ for all $v$ such that $W_v = 0$—otherwise, $f_{W,Z}(W, Z) = 0$.

To derive the approximation for the posterior expected values, consider $W$ and $\tilde{W}$ that are equal except that $W_v = 0$ and $\tilde{W}_v = 1$ for one $v$. Choose $Z$ consistent with $W$, which means $f_{W,Z}(W, Z) \neq 0$. Since we are using a subscript to denote a column of a matrix, we use a superscript as in $Z^v$ to denote row $v$ of $Z$. Furthermore, $Z^{[-v]}$ denotes the sub-matrix of $Z$ obtained by deleting row $v$.

Straightforward modification of (7) gives

$$\sum_\tilde{Z} f_{W,Z,\tau}(\tilde{W}, \tilde{Z}, \tau | Y) f_{W,Z,\tau}(W, Z, \tau | Y) = \frac{\mathbb{E}[W_v | W_{[-v]}, Z^{[-v]}, \tau, Y]}{1 - \mathbb{E}[W_v | W_{[-v]}, Z^{[-v]}, \tau, Y]}',$$

where the summation is over all $\tilde{Z}$ such that $\tilde{Z}^{[-v]} = Z^{[-v]}$. Furthermore, for any such $\tilde{Z}$,

$$f_{W,Z,\tau}(\tilde{W}, \tilde{Z}, \tau | Y) \approx \frac{A_W + |W_{[-v]}|}{B_W + p - |W_{[-v]}| - 1} \frac{B(A_v + s_v, B_v + q - s_v)}{B(A_v, B_v)} \prod_{t: Z_{vt} = 1} \frac{1}{\tau \sqrt{n}} (1 - \eta_{vt}^2)^{-n/2}.$$ 

Hence, $\mathbb{E}[W_v | W_{[-v]}, Z^{[-v]}, \tau, Y]^{-1} \approx 1 + 1/h_v(W_{[-v]}, \tau, Y)$ with

$$h_v(W_{[-v]}, \tau, Y) = \frac{A_W + |W_{[-v]}|}{B_W + p - |W_{[-v]}| - 1} \sum_{Z^v} \left[ \frac{B(A_v + |Z^v|, B_v + q - |Z^v|)}{B(A_v, B_v)(\tau \sqrt{n})Z^v} \prod_{t: Z_{vt} = 1} (1 - \eta_{vt}^2)^{-n/2} \right].$$
where the summation is over all $2^q$ possible values of $Z^v$.

### C. Mathematical details for learning across sites

The derivation of the marginal posterior for learning Across Sites consists of straightforward modifications of previous calculations. The marginal posterior distribution is as in (3) except that the prior $f(Z)$ is replaced by the joint prior for $Z$ and $G$, which is

$$f(G, Z) = \frac{B(A_G + |G|, B_G + r - |G|)}{B(A_G, B_G)} \prod_{g: G^g = 1, p_g > 1} \frac{B(A_g + s_g, B_g + p_g - s_g)}{B(A_g, B_g)}$$

provided that $Z^v = 0$ if $G_{\gamma(v)} = 0$, and $Z^v = 1$ if $G_{\gamma(v)} = 1$ and $p_{\gamma(v)} = 1$—otherwise, $f(G, Z) = 0$. Similar to the preceding approximations, $E[G_g | G_{[-g]^1}, Z_{[-g]^1}, \tau, y]^{-1} \approx 1 + 1/h(G_{[\neg g]}, \tau, y)$, where

$$h(G_{[\neg g]}, \tau, y) = \frac{A_G + |G_{[\neg g]}|}{B_G + r - |G_{[\neg g]}| - 1} \sum_{Z_g} \left[ \frac{B(A_g + |Z_g|, B_g + p_g - |Z_g|)}{B(A_g, B_g)(\tau \sqrt{n})^{Z_g}} \prod_{v: \gamma(v) = g, Z^v = 1} (1 - \eta_v^2)^{-n/2} \right]$$

with the summation being over all $2^{p_g}$ possible values of $Z_g$. When $p_g = 1$, however, the summation is replaced by $(1 - \eta_v^2)^{-n/2}/(\tau \sqrt{n})$, where $v$ is the lone variant in group $g$. Hence, the posterior conditional probability of $G_g$ depends upon the overall number of groups and the number of groups considered relevant, while the posterior conditional probability of $Z^v$ given that $G_g = 1$ depends on the number $s_g$ of variants in the same group that are deemed functional.