

Imputation-based analysis of association studies: candidate regions and quantitative traits

Bertrand Servin^{1,2} and Matthew Stephens^{1,3}

¹Department of Statistics, University of Washington

Box 354322, Seattle, WA 98195, USA.

May 8, 2007

² Present affiliation: INRA Laboratoire de Génétique Cellulaire. 31386 Castanet-Tolosan, France.

³ Present affiliation: Departments of Statistics and Human Genetics, University of Chicago. Chicago, IL 60637, USA.

Corresponding Author:

Bertrand Servin

INRA – Laboratoire de Génétique Cellulaire.

Chemin de Borde-Rouge - Auzeville

BP 52627

31326 Castanet-Tolosan, France.

e-mail: bservin@toulouse.inra.fr

phone: +33 5-6128-5431

fax: +33 5-6128-5308

Abstract

We introduce a new framework for the analysis of association studies, designed to allow untyped variants to be more effectively and directly tested for association with a phenotype. The idea is to combine knowledge on patterns of correlation among SNPs (e.g. from the International HapMap project, or resequencing data in a candidate region of interest), with genotype data at tag SNPs collected on a phenotyped study sample, to estimate (“impute”) unmeasured genotypes, and then assess association between the phenotype and these estimated genotypes. Compared with standard single-SNP tests, this approach results in increased power to detect association, even in cases where the causal variant is typed, with the greatest gain occurring when multiple causal variants are present. It also provides more interpretable explanations for observed associations, including assessing, for each SNP, the strength of the evidence that it (rather than another correlated SNP) is causal. Although we focus on association studies with quantitative phenotype and a relatively restricted region (e.g. a candidate gene), the framework is applicable, and computationally practical, for whole genome association studies. Methods described here are implemented in a software package, **Bim-Bam**, available from the Stephens Lab website <http://stephenslab.uchicago.edu/software.html>.

Author Summary

Ongoing association studies are evaluating the influence of genetic variation on phenotypes of interest (hereditary traits and susceptibility to disease) in large patient samples. However, although genotyping is relatively cheap, most association studies genotype only a small proportion of SNPs in the region of study, with many SNPs remaining untyped. Here we present methods for assessing whether these untyped SNPs are associated with the phenotype of interest. The methods exploit information on patterns of multi-marker correlation (“linkage disequilibrium”) from publically-available databases, such as the International HapMap project or the SeattleSNPs resequencing studies, to estimate (“impute”) patient genotypes at untyped SNPs, and assesses the estimated genotypes for association with phenotype. We show that, particularly for common causal variants, these methods are highly effective: compared with standard methods, they provide both greater power to detect associations between genetic variation and phenotypes, and also better explanations of detected associations, in many cases closely approximating results that would have been obtained by genotyping all SNPs.

Introduction

Although the development of cheap high-throughput genotyping assays have made large-scale association studies a reality, most ongoing association studies genotype only a small proportion of SNPs in the region of study (be that the whole genome, or a set of candidate regions). Because of correlation (“Linkage Disequilibrium”; LD) among nearby markers, many untyped SNPs in a region will be highly correlated with one or more nearby typed SNPs. Thus, intuitively, testing typed SNPs for association with a phenotype will also have some power to pick up associations between the phenotype and untyped SNPs. In practice, typical analyses involve testing each typed SNP individually, and in some cases combinations of typed SNPs jointly (e.g. haplotypes), for association with phenotype, and hoping that these tests will indirectly pick up associations due to untyped SNPs. Here we present a framework for more directly and effectively interrogating untyped variation.

In outline, our approach improves on standard analyses by exploiting available information on LD among untyped and typed SNPs. Partial information on this is generally available from the International HapMap project [?]; in some cases more detailed information (e.g. resequencing data) may also be available, either through public databases (e.g. SeattleSNPs [?]), or through data collected as a part of the association study design (e.g. [?]). Our approach combines this background knowledge of LD with genotypes collected at typed SNPs in the association study, to explicitly predict (“impute”) genotypes in the study sample at untyped SNPs, and then tests for association between imputed genotypes and phenotype. We use statistical models for multi-marker LD to perform the genotype imputation, with uncertainty, and a Bayesian regression approach to perform the test for association, allowing for potential errors in the imputed genotypes. Although we focus specifically on methods for analyzing quantitative phenotypes in candidate gene studies, the same general framework can also be applied to discrete traits, and/or genome-wide scans.

These imputation-based methods can be viewed as a natural *analysis* complement to the “tag SNP” *design* strategy for association studies, which attempts to choose SNPs that are highly correlated with, and hence good predictors of, untyped SNPs. We are simply directly exploiting this property, together with recently-developed statistical models for multi-locus LD ([? ?]) to infer

the untyped SNP genotypes. Our approach is also somewhat analogous to multipoint approaches to linkage mapping (e.g. [?]), where observed genotypes at multiple markers predict patterns of identity by descent (IBD) at nearby positions without markers, and test for correlation between these patterns of IBD and observed phenotypes. In the association context, we are predicting identity by state (IBS) rather than IBD, and the methods of predicting IBS vs IBD differ greatly, but the approaches share the idea of using multipoint information to predict single-point information, and, at least in their simplest form, subsequently assessing correlation with phenotype at the single-point level. This strategy provides a clean and rigorous way to avoid the “curse of dimensionality” that can plague haplotype-based analyses, without making *ad hoc* decisions such as pooling rare haplotypes into a single class.

Although our methods are developed in a Bayesian framework, they can also be used to compute p values assessing significance of observed genotype-phenotype associations. Our approach should therefore be of interest to practitioners whether or not they favor Bayesian procedures in general. It has two main advantages over more standard approaches. First, it provides greater power to *detect* associations. Part of this increased power comes from incorporating extra information (knowledge on patterns of LD among typed and untyped SNPs), but, unexpectedly, we also found an increased power of our Bayesian approach even when all SNPs were actually typed. Second, and perhaps more importantly, it provides *more interpretable explanations* for potential associations. Specifically, for each SNP (typed and untyped), it provides a probability that it is causal. This contrasts with standard single-SNP tests, which provide a p -value for each SNP, but no clear way to decide which SNPs with small p values might be causal.

Models and Methods

We focus on an association study design in which genotype data are available for a dense set of SNPs on a “panel” of individuals, and genotypes are available for a subset of these SNPs (which for convenience we refer to as “tag SNPs”) on a “cohort” of individuals who have been phenotyped for a univariate quantitative trait. We assume the cohort to be a random sample from the population,

and consider application to other designs in the discussion.

Our strategy is to use patterns of LD in the panel, together with the tag SNP genotypes in the cohort, to explicitly predict the genotypes at all markers for members of the cohort, and then analyse the data *as if the cohort had been genotyped at all markers*. There are thus two components to our approach: i) predicting (“imputing”) cohort genotypes, and ii) analysing association between cohort genotypes and phenotypes. For i) we use existing models for population genetic variation across multiple markers [? ?], which perform well at estimating missing genotypes, and provide a quantitative assessment of the uncertainty in these estimates [?]. For ii) we introduce a new approach based on Bayesian regression, and describe how this approach can yield not only standard Bayesian inference, but also p -values for testing the null hypothesis of no genotype-phenotype association. We chose to take a Bayesian approach partly because it provides a natural way to consider uncertainty in estimated genotypes. However, the Bayesian approach has other advantages, particularly it provides a measure of the strength of the evidence for an association (the Bayes Factor) which is, in some respects, superior to conventional p values. Furthermore, in our simulations, p values from our Bayesian approach provide more powerful tests than standard tests, even if the cohort is actually genotyped at *all* markers (including all causal variants).

Bayesian Regression Approach

We now provide further details of our Bayesian regression approach. The literature on Bayesian regression methods is too large to review here, but papers particularly relevant to our work include [?], [?] and [?].

For simplicity we focus on the situation where cohort genotypes are known at all SNPs (tag and non-tag). Extension to the situation where the cohort is genotyped only at tag SNPs, and other genotypes are imputed using sampling-based algorithms such as PHASE [? ?], or fastPHASE [?], is relatively straightforward (see below).

Let G denote the cohort genotypes for all n individuals in the cohort, and $\mathbf{y} = (y_1, \dots, y_n)$ denote the corresponding (univariate, quantitative) phenotypes. We model the phenotypes by a

standard linear regression:

$$y_i = \mu + \sum_j \mathbf{x}_{ij} \boldsymbol{\beta}_j + \epsilon_i, \quad (1)$$

where y_i is the phenotype measurement for individual i , μ is the phenotype mean of individuals carrying the “reference” genotype, the \mathbf{x}_{ij} s are the elements of a design matrix \mathbf{X} (which depends on the genotype data; see below), the $\boldsymbol{\beta}_j$ s are the corresponding regression coefficients, and ϵ_i is a residual. We assume ϵ_i ’s are i.i.d. $\sim \mathcal{N}(0, 1/\tau)$, where τ denotes the inverse of the variance, usually referred to as the *precision* (we choose this parameterisation to simplify notation in later derivations). Thus $y_i | \mu, \mathbf{x}_i, \boldsymbol{\beta}, \tau \sim \mathcal{N}(\mu + \sum_j \mathbf{x}_{ij} \boldsymbol{\beta}_j, 1/\tau)$ and:

$$P(y_i | \mathbf{x}_i, \mu, \boldsymbol{\beta}, \tau) \propto \sqrt{\tau} \exp \left[-0.5\tau (y_i - (\mu + \sum_j \mathbf{x}_{ij} \boldsymbol{\beta}_j))^2 \right]. \quad (2)$$

We assume a genetic model where the genetic effect is additive across SNPs (i.e. no interactions) and where the three possible genotypes at each SNP (major allele homozygote, heterozygote, and minor allele homozygote) have effects 0, $a + ak$ and $2a$ respectively [?]. We achieve this by including two columns in the design matrix for each SNP, one column being the genotypes (coded as 0, 1 or 2 copies of the minor allele), and the other being indicators (0 or 1) for whether the genotype is heterozygous. The effect of SNP j is then determined by a pair of regression coefficients (β_{j1}, β_{j2}) , which are, respectively, the SNP additive effect a_j and dominance effect $d_j = a_j k_j$. While there are other ways to code the correspondance between genotypes and the design matrix, we chose this coding to aid specifying sensible priors (see below).

Priors for $(\boldsymbol{\beta}, \mu, \tau)$

Prior specification is intrinsically subjective, and specifying priors that satisfy everyone is probably a hopeless goal. Our aim is to specify “useful” priors, which avoid some potential pitfalls (discussed below), facilitate computation, and have some appealing properties, while leaving some room for context-specific subjective input. In particular we describe two priors below, which we refer to as prior D_1 and D_2 , that were developed based on the following considerations: i) inference should not depend on the units in which the phenotype is measured; ii) even if the phenotype is

affected by SNPs in this region, the majority of SNPs will likely not be causal; iii) for each causal variant there should be some allowance for deviations from additive effects (i.e. dominant/recessive effects), without entirely discarding additivity as a helpful parsimonious assumption; iv) computations should be sufficiently rapid to make application to genome-wide studies practical (this last consideration refers to prior D_2).

Priors on the phenotype mean and variance The parameters μ and τ relate to the mean and variance of the phenotype, which depend on units of measurement. It seems desirable that estimates (and, more generally, posterior distributions) of these parameters scale appropriately with the units of measurement, so, for example, multiplying all phenotypes by 1,000 should also multiply estimates of μ by 1,000. Motivated by this, for prior D_1 we used Jeffreys’ prior for these parameters:

$$P(\mu, \tau) \propto 1/\tau. \tag{3}$$

This prior is well known to have the desired scaling properties in the simpler context where observed data are assumed to be $N(\mu, 1/\tau)$ [?], and we conjecture that our prior D_1 also possesses these desired scaling properties in the more complex context considered here, although we have not proven this.

For prior D_2 we used a similar, but slightly different prior, based on assuming a prior for (μ, τ) of the form

$$\begin{aligned} \tau &\sim \Gamma(\kappa/2, \lambda/2) \\ \mu|\tau &\sim \mathcal{N}(0, \sigma_\mu^2/\tau). \end{aligned} \tag{4}$$

Specifically, our prior D_2 assumes the limiting form of this prior as $\kappa, \lambda \rightarrow 0$ and $\sigma_\mu^2 \rightarrow \infty$. In Appendix B we show that the posterior distributions obtained using this limiting prior scale appropriately.

Both prior distributions above are “improper” (meaning that the densities do not integrate to a finite value). Great care is necessary before using improper priors, particularly where one intends to compute Bayes Factors (BFs) to compare models, as we do here. However, we believe results obtained using these priors are sensible. For prior D_2 , as we show in the Appendix, the posteriors

are proper, and the BF tends to a sensible limit. For prior D_1 we believe this to be true, although we have not proven it.

Prior on SNP effects For brevity, we refer to SNPs that affect phenotype as QTNs, for Quantitative Trait Nucleotides. Our prior on the SNP effects has two components: a prior on which SNPs are QTNs, and a prior on the QTN effect sizes.

Prior on which SNPs are QTNs We assume that with some probability, p_0 , none of the SNPs is a QTN: that is, the “null model” of no genotype-phenotype associations holds. Otherwise, with probability $(1 - p_0)$, we assume there are l QTNs, where l has some distribution $p(l)$ on $\{1, 2, \dots, n_S\}$ where n_S denotes the number of SNPs in the region. Given l , we assume all subsets of l SNPs are equally likely. Both p_0 and $p(l)$ can be context-dependent, and choice of appropriate values is discussed below.

Prior on effect sizes If SNP j is a QTN then its effect is modelled by two parameters, a_j and $d_j = a_j k_j$. The parameter a_j measures a deviation from the mean μ and will depend on the unit of measurement of the phenotype. To reflect this, we scale the prior on a_j by the phenotypic standard deviation within each genotype class, $\sqrt{1/\tau}$. Specifically, our prior on a_j is $\mathcal{N}(0, \sigma_a^2/\tau)$, where σ_a reflects the typical size of a QTN effect compared with the phenotype standard deviation within each genotypic class. Choice of σ_a may be context-dependant, and is discussed below.

The parameter $d_j = a_j k_j$ measures the dominance effect of a QTN. If $k_j = 0$ then the QTN is additive: the heterozygote mean is exactly between the means of the two homozygotes. If $k_j = 1$ (respectively -1), allele 1 (respectively 0) is dominant. The case $|k_j| > 1$ corresponds to overdominance of allele 1 or allele 0. We investigate two different priors for the dominance effect:

1. **Prior D_1 :** We assume that k_j is a priori independent of a_j , with $k_j \sim \mathcal{N}(0, \sigma_k^2)$. We chose $\sigma_k = 0.5$, which gives $P(|k_j| > 1) \approx 0.05$, reflecting a belief that over-dominance is relatively rare.
2. **Prior D_2 :** We assume that d_j is a priori independent of a_j , with $d_j \sim \mathcal{N}(0, \sigma_d^2/\tau)$, where we

took $\sigma_d = 0.5\sigma_a$. This prior on d_j induces a prior on k_j in which k_j is not independent of a_j .

Prior D_1 has the attractive property that the prior probability of overdominance is independent of the QTN additive effect a_j . However the posterior distributions of a_j and k_j must be estimated via a computationally intensive MCMC scheme (see Appendix A). (An alternative, which we have not yet pursued, would be to approximate BFs under prior D_1 by numerical methods, such as Laplace Approximation; e.g. [?]). Prior D_2 is more convenient, as, when combined with the priors on μ and τ in equation (4), posterior probabilities of interest can be computed analytically (Appendix B).

For both priors D_1 and D_2 we assume effect parameters for different SNPs are, a priori, independent (given the other parameters).

Choice of $p_0, p(l)$ and σ_a The above priors include “hyperparameters”, p_0 and σ_a , and a distribution $p(l)$ that must be specified. The hyperparameter p_0 gives the prior probability that the region contains no QTNs. While choice of appropriate value is both subjective, and context-specific, for candidate regions we suggest p_0 will typically fall in the range 10^{-2} to 0.5. If data on multiple regions are available then it might be possible to estimate p_0 from the data, although we do not pursue this here. Instead, we mostly sidestep the issue of specifying p_0 by focussing on the Bayes Factor (described below), which allows readers of an analysis to use their own value for p_0 when interpreting results.

In specifying the prior, $p(l)$, for the number of QTNs, we suggest concentrating most of the mass on models with relatively few QTNs. Indeed, here we focus mainly on the extreme case where $p(l)$ is entirely concentrated on $l = 1$: that is, the “alternative” model is that the region contains a single QTN. Although rather restrictive, this seems a good starting point in practice, particularly since our results show that it can perform well even if multiple QTNs are present. Nonetheless there are advantages in considering models with multiple QTNs, and so we also consider a prior where $p(l)$ puts equal mass on $l = 1, 2, 3$ or 4. This prior suffices to illustrate the potential of our approach, although in practice it would probably be preferable to place decreasing probabilities on larger numbers of QTNs (e.g. $p(l = 2) < p(l = 1)$). An alternative would be to sidestep

specifying $p(l)$ by computing BFs comparing, say, 4-QTN, 3-QTN, 2-QTN, and 1-QTN models vs the “null” model. However, interpreting and acting on these BFs will inevitably correspond to implicit assumptions about the relative prior plausibility of these multi-QTN models.

Finally, specification of the standard deviation of the effect size, σ_a , involves subtle issues. Although it may seem tempting to use “large” σ_a , to reflect relative “ignorance” about effect sizes [?], we believe this is inadvisable. Although large σ_a yields a flat prior on effect sizes, this prior is far from “uninformative”, in that it places almost all its mass on large effect sizes. The result would essentially allow only zero effects (i.e. the “null” model), or large effects (the “alternative” model). If in truth the causal SNPs have relatively small effect, which is probably generally realistic, then (for realistic sample sizes) the null model would be strongly favoured over the alternative, because the data would be more consistent with zero effects than with large effects. Choice of σ_a can thus strongly effect inference, particularly the Bayes Factor, which we use to summarise evidence for the region containing any QTNs. Partly because of this, in practice we suggest averaging results over several values for σ_a (equivalent to placing a prior on σ_a). It may also be helpful to examine sensitivity of results to σ_a . For example, if the BF is small for all values of σ_a then there is no evidence for any QTN in the region; if the BF is large for some values and small for others then the evidence depends on the extent to which you believe in large vs small effects. However, for simplicity all results in this paper were obtained using a fixed value of $\sigma_a = 0.5$.

Inference

We focus on two key inferential problems: (i) *detecting* association between genotypes and phenotype; and (ii) *explaining* observed associations. In model (1), these translate to answering (i) are any β_j 's non-zero? and (ii) which β_j 's are non-zero and how big are they? We view the ability to address both questions within a single framework to be an advantage of our approach.

Detecting association To measure the evidence for *any* association between genotypes and phenotypes, we use the Bayes Factor, BF, [?] given by

$$\text{BF} = P(\mathbf{y}|\mathbf{G}, H_1)/P(\mathbf{y}|\mathbf{G}, H_0), \quad (5)$$

where H_0 denotes the null hypothesis that none of the SNPs is a QTN ($a_j = d_j = 0$ for all j), and H_1 denotes the complementary event (i.e. at least one SNP is a QTN). Computing the BF involves integrating out unknown parameters, as described in the Appendices. In interpreting a BF, it is helpful to bear in mind the formula “posterior odds = prior odds \times BF”, so, for example, if the prior odds are 1:1 (i.e. $p_0 = 0.5$, so association with genetic variation in the region is considered equally plausible, *a priori*, as no association) then a BF of 10 gives posterior odds of 10:1, or $\sim 91\%$ probability of an association.

In the special case where we allow at most one QTN, (5) reduces to

$$\text{BF} = (1/n_S) \sum_{j=1}^{n_S} P(\mathbf{y}|\mathbf{G}, H_j)/P(\mathbf{y}|\mathbf{G}, H_0) \quad (6)$$

where H_j denotes the event that SNP j is the QTN. The j th term in this sum corresponds to the BF for H_j vs the null model, and involves the genotype data at SNP j only. We refer to these terms as the “single-SNP” BFs, so in this special case the overall BF is the mean of the single-SNP BFs. This natural way for combining information across (potentially correlated) SNPs is an attractive property of BFs compared with single-SNP p values. Furthermore, in terms of detecting a genotype-phenotype association it can work well even if multiple QTNs are present (see results).

The Bayes/non-Bayes compromise From a Bayesian viewpoint, the BF provides *the* measure of the strength of evidence for genotype-phenotype association. That is, if one accepts our prior distributions and modeling assumptions, then the BF is all that is necessary to decide whether a genuine association is present. However, given the potential for debate over prior distributions, and for deviations from modelling assumptions, it is helpful to note that a p value for testing H_0 can be obtained from a BF through permutation. Specifically, one can compute the BF for the observed data, and for artificial data sets created by permuting observed phenotypes among cohort individuals, and obtain a p value as the proportion of permuted data sets for which the BF exceeds the BF for the observed data. Being based on permutation, the resulting p value is valid *irrespective of whether the model or priors are appropriate*. This p value also provides a helpful way to compare our approach with standard tests of association, and, as we show below, tests based

on BF appear to perform well in a wide variety of situations. Using BFs as test statistics to obtain p values is referred to as the “Bayes/non-Bayes compromise” by Good [?].

Explaining and Interpreting Associations To “explain” observed associations we compute posterior distributions for SNP effects ($a_j + d_j$ and $2a_j$ for the heterozygote and minor-allele homozygote respectively), with particular focus on the posterior probability that each SNP is a QTN ($P(a_j \neq 0)$). Here, our Bayesian regression approach has an important qualitative advantage over standard multiple regression. Specifically, if a genetic region contains multiple highly-correlated SNPs, each highly correlated with the phenotype, then the correct conclusion would be any of these SNPs could be causal, without identifying which one. This will be reflected in the posterior distribution of the effects: the overall probability that at least one SNP is a QTN will be high, but (at least in the simplest case where we assume at most one QTN) this probability will be spread out over the multiple correlated SNPs. In contrast, if multiple highly correlated SNPs are included in a standard multiple regression it is possible that no one of them will produce a significant p value.

We also argue that the imputation-based approach brings us closer to being able to interpret estimated effects for each SNP as actual *causal effects*, rather than simply associations. Indeed, the key to making the leap from association to causality is controlling for all potential confounding factors, and by imputing genotypes at nearby SNPs the imputation-based approach controls for one important set of confounding factors (the nearby SNPs) which would otherwise be ignored. Thus, while functional studies provide the ultimate route to convincingly demonstrating causal effects, our approach may help target such studies on the most plausible candidate SNPs.

Imputing genotypes

In the tagSNP design, observed genotypes G_{obs} consist of panel genotypes at all SNPs and cohort genotypes at tagSNPs only. To apply our methods in this situation, we use sampling-based algorithms (PHASE [? ?], or fastPHASE [?]) to generate multiple imputations for the complete genotype data (all individuals at all SNPs) by sampling from $P(G|G_{obs})$. We then incorporate these imputations into our inference: for prior D_1 , this involves adding a step in the MCMC

scheme to sample the imputed genotypes from their posterior distribution given all the data; for prior D_2 it involves simply averaging relevant calculations over imputations. Details are given in the Appendices.

Results

“Power” and comparisons with other approaches

We compared the power of our approach to other common approaches via simulation. We simulated genotype and phenotype data (with $\mu = 0$ and $\tau = 1$) for genetic regions of length 20kb containing a single QTN, and genetic regions of length 80kb containing four QTNs, as follows:

- Using a coalescent-based simulation program, *msHOT* [?], simulate 600 haplotypes from a constant-sized random mating population, under an “infinite sites” mutation model, with (population-scaled) mutation rate $\theta = 0.4/\text{kb}$ and “background” recombination rate $\rho = 0.8/\text{kb}$, and a recombination hotspot (width 1kb; recombination rate 50ρ per kb) in the center of the region.
- Form genotypes for a “panel” of 100 individuals by randomly pairing 200 haplotypes, and a “cohort” of 200 individuals by randomly pairing the other 400 haplotypes.
- Select tag SNPs from the panel data using the approach of Carlson et al. [?] with an r^2 cutoff of 0.8. As in Carlson et al. [?], SNPs with panel minor allele frequency (MAF) < 0.1 were not tagged.
- Select which SNPs are QTNs, and their effect sizes, and simulate phenotype data for each cohort individual according to (1). We considered four scenarios: A) a “common” (MAF > 0.1) QTN, with a range of effect sizes $a = 0.2, 0.3, 0.4, 0.5$ and “mild” dominance for the minor allele ($d = 0.4a$); B) a common QTN, with $a = 0.3$ and “strong” dominance for the major allele ($d = -a$); C) a “rare” (MAF $0.01 - 0.05$) QTN, with $a = 1$ and no dominance ($d = 0$); D) four common, relatively uncorrelated, QTNs, each with $a = 0.3$ and $d = 0.4a$.

In each situation we randomly chose a QTN satisfying the relevant MAF requirements (in the 600 sampled haplotypes), except under scenario D we first chose four tag SNP “bins” at random and then randomly chose a QTN satisfying the MAF requirement in each bin, thereby ensuring the four QTNs were relatively uncorrelated. (While real data may contain multiple highly-correlated QTNs, we did not explicitly consider this case, since their effect would be similar to a single QTN.)

We compared power of tests based on the BF (under prior D_2 , allowing at most one QTN, using expression (6)) with four other significance tests:

- Two tests based on p_{\min} , the minimum p value obtained from testing each SNP individually (via standard ANOVA-based methods) for association with the phenotype. These two tests differed in whether the single SNP p -values were obtained using the 1df “allelic” test, which assumes an additive model where the mean phenotype of heterozygotes lies midway between the two homozygotes (equivalent to linear regression of phenotype on genotype), or the 2df “genotype” test which treats the mean of the heterozygotes as a free parameter.
- A test based on p_{reg} , the global p -value obtained from linear regression of phenotype on all SNP genotypes (using the standard F statistic, coding the genotypes as 0,1 and 2 at each SNP, and assuming additivity across SNPs). See Chapman et al. [?] for example.
- A test based on BF_{\max} , the *maximum* single-SNP BF. We included this test for comparison with the *mean* single-SNP BF (6), to examine whether averaging information across SNPs in (6) improved power.

For each test, we analysed each data set in two ways: as if data had been collected using (i) a “resequencing design” (i.e. all individuals were completely resequenced, so genotype data are available at all SNPs in all individuals); and (ii) a “tag SNP design” (i.e. in panel individuals genotype data are available at all SNPs, but in cohort individuals genotype data are available at tag SNPs only). For the tag SNP design, we assumed haplotypic phase is known in the panel (as it is, mostly, for the HapMap data for example), but not in the cohort; however our approach can also

deal with unknown phase in the panel. For p_{reg} and p_{min} , tests were performed on all SNPs for the resequencing design, and on tag SNPs only for the tag SNP design. For BF and BF_{max} single-SNP BFs were computed for all SNPs in both designs (averaging over imputed genotypes for non-tag SNPs in the tag SNP design). For p_{reg} we computed a p value assessing significance using the standard asymptotic distribution for the F statistic; for the other tests we found p values by permutation, using 200-500 random permutations of phenotypes assigned to cohort individuals. (The relatively small number of permutations limits the size of the smallest possible p value, causing discontinuities near the origin in Figure 1).

[Figure 1 about here.]

Figure 1 shows power of each test vs type I error under both resequencing and tag SNP designs. For scenario A (a single common QTN) the relative performances of methods were similar for all four effect sizes examined (data not shown), and so we pooled these results in the figure.

Comparing p_{min} and p_{reg} , the single-SNP tests (p_{min}) were more powerful when all variants (including the causal variant) were typed, or when the QTN was a common SNP and therefore “tagged” by a tag SNP, while the regression-based approach (p_{reg}) was more powerful when the QTN was a rare SNP not “tagged” by any tag SNP. Among the two single-SNP tests, the 1df allelic test performed as well as, or better than, the 2df genotypic test, except in scenario B where the major allele exhibits strong dominance. In particular, for scenario A, where the causal variant exhibits dominance, the allelic test (which assumes no dominance), performed better than the genotypic test. This is presumably because, with the effect and sample sizes considered, the extra parameter estimated in the genotypic test does not sufficiently improve model fit. Although relative performance of p_{min} and p_{reg} in the tag SNP design could depend on tag SNP selection scheme (and the one we used, based on pairwise LD, would seem to favor p_{min}), it seems reasonable to expect single-SNP tests to be effective at detecting “direct” associations between the phenotype and a causal variant, or “near-direct” association between a SNP that tags a causal variant, and the regression-based approach to be better at detecting indirect associations between a phenotype and a variant not “tagged” by a single SNP (the intuition, from Chapman et al. [?], is that

such variants can be highly correlated with linear combinations of tag SNPs, and thus be detected by linear regression). In principle, p_{reg} could also effectively capture “direct” associations, but our empirical results suggest that it is less effective at this than the single SNP tests. (However, poor performance of p_{reg} under the resequencing design may be due in part to inadequacy of the asymptotic theory when large numbers of correlated covariates are used. This might be alleviated by assessing significance of p_{reg} by permutation.)

Turning now to our approach, except for Scenario B in the tag SNP design, the test based on the BF is as powerful or more powerful than the other tests. Thus, unlike p_{reg} and p_{min} , the BF performs well in detecting both “direct” and “indirect” associations: if the QTN is typed, the BF detects it using observed genotype data at that SNP; otherwise it detects it using the imputed genotype data at the QTN. In Scenario B, where the major allele exhibits strong dominance, our approach suffered slightly in power compared with the genotypic test, presumably because our prior places relatively low weight on strong dominance. However, the power loss was small compared with that of the allelic test. Thus our prior “allows” for dominance without suffering the full penalty incurred by the extra parameter in the genotypic test when dominance is less strong (Scenario A).

In Scenario D, which involved multiple QTNs, tests based on the BF clearly outperformed other tests considered, even though the BF was computed allowing at most one QTN. Our explanation is that the BF, being the *average* of single-SNP BFs, has greater opportunity to capture the presence of multiple QTNs than does the minimum p value. This explanation is supported by the fact that the maximum BF, BF_{max} , performs less well than BF. To examine whether power might be further increased by explicitly allowing for multiple QTNs we compared power for BFs computed using 1-QTN and 2-QTN models (in the 2-QTN model $p(l = 1) = p(l = 2) = 0.5$). We found little difference in power, although BFs for the 2-QTN model tended to be larger than BFs for the 1-QTN model, so allowing for multiple QTNs may help if the BF itself, rather than a p -value based on the BF, is used to measure the strength of evidence for association. In addition, considering multiple-QTN models should have advantages when attempting to *explain* an association (see below).

A second, and perhaps more surprising, situation where the BF outperforms other methods is when all SNPs are typed and tested (i.e. Scenario A, resequencing design). Here, in contrast to

Scenario D, BF_{max} performs similarly to the standard BF, suggesting that the power gain is due not to averaging, but to an intrinsic property of single-SNP BFs that makes them better measures of evidence than single-SNP p values. Our explanation is that the BF tends to be less influenced by less informative SNPs (e.g. those with very small MAF, of which there are many in the resequencing design), whereas p values tend to give equal weight to all SNPs, regardless of information content. Specifically, BFs for relatively uninformative SNPs will always lie close to 1, and should not greatly influence either the maximum or the average of the single-SNP BFs (or, more precisely, will not greatly influence differences in these test statistics among permutations of phenotypes). In contrast, p values for each SNP are forced, by definition, to have a uniform distribution under H_0 , and so p values from a large number of uninformative SNPs unassociated with the phenotype could swamp any signal generated by a single informative SNP associated with the phenotype. Although the resequencing design is currently uncommon, this observation suggests that it may generally be preferable to rank SNPs according to their BFs, rather than by p values (e.g. in genome scans). It also highlights a general (rarely considered, and perhaps underappreciated) drawback of p values as a measure of evidence: the strength of evidence of a given p value depends on the informativeness of the test being performed, or more specifically on the distribution on the p values under the alternative hypothesis, which is generally not known. Thus, for example, a p -value of 10^{-5} in a study involving few individuals may be less impressive than the same p value in a larger study. In contrast, the interpretation of a BF does not depend on study size or similar factors.

Resequencing vs tag SNP designs

[Figure 2 about here.]

An important feature of Figure 1 is that, for scenarios A), B) and D) where the causal SNPs are common, power is similar for the resequencing and tag SNP designs. Indeed, in these cases most other aspects of inference are also similar. For example, Figure 2 shows that, under scenarios A) and B), estimated effect sizes, BFs, and posterior probability that the actual causal variant is a QTN, are typically similar for both designs. Thus under these scenarios, our imputation-based approach *effectively recreates results that would have been obtained by resequencing all individuals*.

In contrast, when the causal variant is rare, there is a noticeable drop in power for the tag SNP design vs the resequencing design, and the BFs, posterior probabilities, and effect size estimates under the two designs often differ substantially (data not shown). This may seem slightly disappointing: one might have hoped that, even with tag SNPs chosen to capture common variants, they might also capture some rare variants. Indeed, this can happen: in some simulated data sets the rare causal variant was clearly identified by our approach, presumably because it was highly correlated with a particular haplotype background, and could thus be accurately predicted by tag SNPs. However, this occurred relatively rarely (just a few simulations out of 100).

We wondered whether a different tagging strategy, aimed at capturing rare variants, might improve performance when the causal variant is rare. The development of such strategies lies outside the scope of this paper, but, to assess potential gains that *might* be achieved, we analyzed rare-variant simulations assuming that all SNPs *except the causal variant* were typed in the cohort. Power from this approach (Figure 3) gives a conservative upper bound on what could be achieved using a more effective tagging design, without actually typing the causal variant. Although power was higher than with the r^2 -based tag SNP selection, it remained substantially lower than in the resequencing design, where the causal variant is typed.

[Figure 3 about here.]

We also wondered whether a different approach to impute missing genotypes (in the cohort at non-tag SNPs) might improve performance. For results above we used the software fastPHASE [?] to impute the genotypes, so we re-ran the analysis using a different imputation algorithm [PHASE: ? ?]. Results for these two approaches (Figure 4) show little difference in terms of power, consistent with previous results [?] suggesting the two approaches have similar accuracy in imputing missing genotypes.

[Figure 4 about here.]

In summary, imputation-based methods appear to increase power of the tag SNP design to detect rare variants, but nevertheless remain notably less powerful than BFs based on the complete resequencing data.

Comparison of prior D_1 and D_2

Priors D_1 and D_2 differ in their assumed correlation between the dominance effect ($d = ak$) and main effect a : in D_1 the prior probability of over-dominance is independent of a , whereas under D_2 over-dominance is more likely for small a than for large a (Figure 5). In this respect, D_1 is perhaps more sensible than D_2 ; however, D_2 is computationally much simpler. To examine the effects of these priors on inference, we compared i) the BF; and ii) the posterior probability assigned to the actual causal variant; under each prior for the datasets from scenarios A and B. Results agreed quite closely (Figure 6), suggesting prior D_2 provides a reasonable approximation to prior D_1 in the scenarios considered. This is important since prior D_2 is computationally practical for computing BFs for very large datasets (e.g. genome-wide association studies with hundreds of thousands of SNPs), for which sampling posterior distributions of parameters using an MCMC scheme would be computationally daunting.

[Figure 5 about here.]

[Figure 6 about here.]

Allowing for multiple causal variants

When analysing a candidate region one would ideally like not only to detect any association, but also to identify the causal variants (QTNs). Since a candidate region could contain multiple QTNs, we implemented an MCMC scheme (using prior D_1) to fit multi-QTN models where the number of QTNs is estimated from the data; here we consider a multi-QTN model with equal prior probabilities on 1, 2, 3 or 4 QTNs. (A similar MCMC scheme could also be implemented for prior D_2 , and could exploit the analytical advantages of this prior to reduce computation. Indeed, for regions containing a modest number of SNPs it would be possible to examine all subsets of SNPs, and entirely avoid MCMC.)

We compare this multi-QTN model with a 1-QTN model on a dataset simulated with four QTNs (scenario D). The estimated BF for a 1-QTN model was ~ 6000 , while for the multi-QTN model it was $> 10^5$ (we did not perform sufficient iterations to estimate how much bigger than

10^5). Thus, if a region contains multiple causal variants then allowing for this possibility may provide substantially higher BFs. Figure 7 shows the *marginal* posterior probabilities for each SNP being a QTN, under the 1-QTN and multi-QTN models, conditional on at least one SNP in the region being a QTN. (Summarising the more complex information on posterior probabilities for *combinations* of SNPs is an important future challenge.) Under the 1-QTN model only one of the four causal SNPs has a large marginal posterior probability, whereas under the multi-QTN model all four are moderately large. Of course, other SNPs correlated with the four QTNs were also associated with the phenotype, and so have elevated posterior probabilities. This example illustrates the potential for the multi-QTN model to provide fuller explanations for associations.

[Figure 7 about here.]

SCN1A polymorphism and maximum dose of carbamazepine

We applied our method to data from association studies involving the SCN1A gene and the maximum dose of carbamazepine in epileptic patients [? ?]. For this analysis, the “panel” consisted of parents from 32 trios of European descent from the CEPH Utah collection [from ?] and the “cohort” consisted of 425 patients of European descent for whom the maximum dose of carbamazepine had been determined [from ?]. Genetic data on the trios were available for 15 polymorphisms, 14 SNPs and 1 indel, corresponding to snp1 to snp15 and indel12 in Table 2 of Weale et al. [?]. For cohort individuals genotype data are available at four tag SNPs (snp1 (rs590478), snp5 (rs8191987), snp7 (rs3812718) and snp9 (rs2126152)) chosen to summarize haplotype diversity at the 15 panel polymorphisms (for details, see Tate et al. [?]).

We first estimated haplotypes in 64 parents using the trio option in PHASE [?]. Since trio information allows haplotypes to be accurately determined [?] we assumed these estimated panel haplotypes were correct in subsequent analyses. We then applied our method to compute a BF for overall association between genetic data and the phenotype, and to compute, for each SNP, the posterior probability that it was a QTN. In applying our method we used PHASE to impute the genotypes in the cohort at non tag SNPs, and performed analyses under both priors D_1 and D_2 .

BFs for priors D_1 and D_2 were respectively 3.15 and 2.33, and the corresponding p values (estimated using 1000 permutations) were 0.006 and 0.019. We also computed p values using single SNP tests at tag SNPs and obtained 0.007 for the allelic test, and 0.019 for the genotype test. (These are essentially the two tests performed by Tate et al. [?], who reported the smallest p -values uncorrected for multiple comparisons). These BFs represent only modest evidence for an association. If one were initially even somewhat sceptical about SCN1A as a candidate for influencing this phenotype, one might remain somewhat sceptical after analysing these data. For example, with a 20% prior probability on variation in SCN1A influencing phenotype, the posterior probability of association under either prior is $< 50\%$. (Prior probability of 0.2 gives prior odds of 0.2:(1-0.2), or 1:4; a BF of 3 then gives posterior odds of 3:4, which translates to a posterior probability of 3/7.) On the other hand, SCN1A might be considered a relatively good candidate for influencing response to carbamazepine, since it is the drug's direct target. And, depending on follow-up costs and potential benefits of finding a functional variant, posterior probabilities of very much $< 50\%$ might be deemed worth following-up.

[Figure 8 about here.]

Among the 15 SNPs analyzed, snp7 was assigned the highest posterior probability of being a QTN (Figure 8). This SNP, which is a tag SNP, was also implicated by the analysis in Tate et al. [?]. However, the posterior probability of this SNP represents only 34 % of the posterior mass. Six additional SNPs are needed to encompass 90% of the posterior mass: snp6 (rs3812719), snp8 (rs490317), snp9 (rs2126152), snp10 (rs7601520), snp11 (rs2298771) and snp13 (rs7571204). The posterior distributions of the main effect, a , for each of these seven SNPs, conditional on it being a QTN, are very similar (Figure 8).

In summary, these data provide modest evidence of association between SCN1A and maximum dose of carbamazepine, and, among the SNPs analyzed, snp7 (rs3812718) appears the best candidate for being causal. A recent follow-up study appears to confirm this variant as being functionally important [?].

(see Appendix C). A second distinctive contribution is that we compare our Bayesian approach directly with standard p -value based approaches, providing both qualitative insight and quantitative support for several advantages of single-SNP BFs over single-SNP p values. These advantages include: i) the BF allows for both additive and dominant effects without the additional degree of freedom incurred by the general 2df hypothesis test; ii) the BF better reflects the informativeness of each SNP, in particular that SNPs with small MAF are typically less informative than SNPs with larger MAF (this advantage presumably being greatest for SNP panels containing many SNPs with small MAF); iii) it provides a principled way to take into account prior information on each SNP, e.g. whether it lies in or near a gene whose function is believed likely to influence the trait; and iv) averaging single-SNP BFs provides a convenient, and in some ways effective, approach to combining information across multiple SNPs in a region.

Perhaps the most important *disadvantage* of BFs compared with p values, is that a BF is strictly “valid” only under the assumption that both the prior and the model are “correct”. Since this is never the case in practice, BFs are never strictly valid, Our hope is to make the prior and model sufficiently accurate that resulting BFs are “useful”. (Note that p values may be valid but useless: e.g. p values simulated from a uniform distribution independent of phenotype and genotype data are valid, in that they are uniformly distributed under the null hypothesis, but useless.) Here it is helpful to distinguish two different uses of BFs: as test statistics to compute permutation-based p values, as in the power comparisons in this paper; and as direct measures of evidence (e.g. in “posterior odds = BF \times prior odds”). Our limited experience is that p values obtained from BFs are relatively robust to prior and modeling assumptions, but that the absolute values of BFs are substantially more sensitive. In particular, BFs tend to be sensitive to both i) choice of σ_a, σ_d ; and ii) the normality assumption in the phenotype model. We now discuss each of these issues in turn.

Choice of σ_a, σ_d corresponds to quantifying prior beliefs about likely additive and dominance effect sizes. In this paper we used (in prior D_2) $\sigma_a = 0.5$ and $\sigma_d = \sigma_a/2$. We now believe these values are likely larger than appropriate for most studies of complex phenotypes, placing too little weight on small, but realistic, effect sizes. Our current suggested “default” procedure is to average BFs computed with $\sigma_a = 0.05, 0.1, 0.2$ and 0.4 , and $\sigma_d = \sigma_a/4$, which places more weight on

smaller effect sizes, and less weight on overdominance. We would expect to modify these values in the light of further information about typical effect sizes for particular traits. It could also be argued that, in addition to allowing a continuum of deviations from the additive model it may make sense to specify prior probabilities for “pure” recessive or dominant models (i.e. $d = -a, a$). BFs under these models can be computed easily by simply replacing all heterozygous genotypes with homozygous genotypes for the major or minor allele.

Regarding the normality assumption, following a suggestion by Mathew Barber (personal communication), in practical applications we are currently applying a normal quantile transform to phenotypes (replacing the r th biggest of n observations with the $r/(n + 1)$ th quantile of the standard normal distribution) before applying our methods and computing BFs. Imposing normality on our phenotype in this way is different from the normality assumption in our phenotype model, which states that the *residuals* are normally distributed. However, in this context, where effect sizes are expected to be generally rather small, normality of phenotype and normality of residuals are somewhat similar assumptions, suggesting that this transform may be effective.

Throughout this paper we have assumed a “population” sampling design where phenotype and genotype data are available on a random sample from a population, and perform analyses conditional on the observed genotype data. An alternative common design involves collecting genotypes only on individuals whose phenotypes lie in the tails of the distribution [?]. To apply our methods to such designs we suggest conditioning on *unordered* observed phenotypes, denoted $\{\mathbf{y}\}$, in addition to conditioning on the genotypes \mathbf{G} , and to perform inference based on the observed correspondence between phenotypes and genotypes, $P(\mathbf{y}|\{\mathbf{y}\}, \mathbf{G})$. The BF under this conditioning, denoted $\widetilde{\text{BF}}$, is

$$\begin{aligned}\widetilde{\text{BF}} &= P(\mathbf{y}|\{\mathbf{y}\}, \mathbf{G}, H_1)/P(\mathbf{y}|\{\mathbf{y}\}, \mathbf{G}, H_0) \\ &= [P(\mathbf{y}|\mathbf{G}, H_1)/P(\{\mathbf{y}\}|\mathbf{G}, H_1)]/[P(\mathbf{y}|\mathbf{G}, H_0)/P(\{\mathbf{y}\}|\mathbf{G}, H_0)] \\ &= \text{BF}/(1/n!) \sum_{\nu} P(\nu(\mathbf{y})|\mathbf{G}, H_1)/P(\mathbf{y}|\mathbf{G}, H_0)\end{aligned}\tag{7}$$

where the sum is over all permutations ν of the observed phenotypes, and BF is the standard BF used throughout this paper (Equation (5)). A naive Monte-Carlo estimate of $\widetilde{\text{BF}}$ can be obtained as the BF computed using observed phenotype data, divided by the average BF computed

using multiple random permutations of phenotypes. However, more sophisticated methods, such as importance sampling, may be required to obtain accurate estimates. (Use of $\widetilde{\text{BF}}$ may also provide an alternative way to improve robustness to the normality assumption, since conditioning on observed phenotype values should reduce the influence of assumptions about their distribution.) Adapting our approach to standard case-control designs will require development of appropriate (and computationally-tractable) priors and represents an important area for future work.

Accession Numbers

The Entrez Gene ID of the SCN1A gene is 6323.

Availability of Software

Methods described here are implemented in a software package, **Bim-Bam** (Bayesian IMputation-Based Association Mapping), available from the Stephens Lab website <http://stephenslab.uchicago.edu/software.html>.

Acknowledgements

We thank D. Goldstein for access to the SCN1A data, and M. Weale and S. Tate for providing the data in a convenient electronic form. We thank N. Patterson for pointing us to the I. Good reference, and J. Marchini and P. Donnelly for helpful conversations. Computing support was provided by the University of Washington Center for Study of Demography and Ecology, High Performance Computing Cluster Cooperative. This work was supported by NIH grant RO1 HG02585-01 to MS.

A MCMC sampling for prior D_1

When using prior D_1 , we estimate BF's and posterior distributions via MCMC. Specifically, given observed phenotypes \mathbf{y} and observed genotype data \mathbf{G}_{obs} (which for the tag SNP design will consist of genotypes of all SNPs in the panel, and genotypes of tag SNPs in the cohort), we sample from the joint distribution of the model parameters, $(\mu, \tau, \beta = (\mathbf{a}, \mathbf{k}))$, and of the “complete” genotypes \mathbf{G} (which will consist of genotypes at all SNPs in all individuals, including particularly the genotypes at the non tag SNPs in the cohort).

In outline the approach is:

1. Update \mathbf{G} given the genotype information available \mathbf{G}_{obs} .
2. Update the genetic effects parameters $\mathbf{a}, \mathbf{k}, \mu$
3. Update τ from $\tau|\mathbf{a}, \mathbf{k}, \mu, \mathbf{G}, \mathbf{y}$

These steps are iterated many times to obtain samples from a Markov-Chain whose stationary distribution is the joint posterior distribution of all model parameters.

Updating the genotypes To update \mathbf{G} we first propose a new value \mathbf{G}' from $P(\mathbf{G}|\mathbf{G}_{obs})$, and use a Metropolis-Hastings step to accept or reject it. The new configuration is accepted with probability:

$$a = \min\left(1, \frac{P(\mathbf{y}|\mu, \mathbf{a}, \mathbf{k}, \tau, \mathbf{G}')}{P(\mathbf{y}|\mu, \mathbf{a}, \mathbf{k}, \tau, \mathbf{G})}\right). \quad (8)$$

(Here the proposal probability has cancelled with the prior distribution to yield this acceptance probability.)

In practice, we actually generated a large number of samples from $P(\mathbf{G}|\mathbf{G}_{obs})$, using PHASE [? ?] or fastPHASE [?] and propose new configurations by choosing uniformly at random from this sample.

Update of the genetic effects To describe this update we introduce additional notation. Let γ denote the set of SNPs which are QTNs, L denote the maximum number of QTNs allowed under

our prior, and n_S denote the total number of SNPs in the region. To update the genetic effect parameters we first propose a new value γ^* for γ , as follows. With probability 0.2 we set $\gamma^* = \gamma$. Otherwise we propose a new value γ^* by adding and/or removing a SNP from γ :

1. If γ contains no SNPs, we add a new SNP at random.
2. If γ includes L SNPs: if $L = n_S$ then remove a SNP at random; otherwise with probability 0.5 remove a SNP at random, and with probability 0.5 remove a SNP at random from γ and add a randomly-chosen SNP currently not in γ .
3. In all other configurations for γ , we either change the status (*i.e.* from included to not included or from not included to included) of a SNP at random (with probability 0.5) or switch a SNP included with a non-included SNP (probability 0.5).

Then, given the proposed new set of QTNs, γ^* , we jointly propose new values for their respective regression coefficients and the reference mean μ , by sampling from the proposal distribution

$$q_{ak}(\mu^*, \mathbf{a}^*, \mathbf{k}^* | \gamma^*) \sim \mathcal{N}(\mathbf{B}, \mathbf{V}), \quad (9)$$

where $\mathbf{B} = \mathbf{V} \mathbf{X}^t \mathbf{y}$, $\mathbf{V} = (\tau \mathbf{X}^t \mathbf{X} + \mathbf{v}^{-1})^{-1}$ and $\mathbf{v} = \text{diag}(\sigma_\mu^2, \sigma_a^2/\tau, \sigma_k^2 \sigma_a^2/\tau, \dots, \sigma_a^2, \sigma_k^2 \sigma_a^2/\tau)$. Here, unlike in the main paper, we assume the design matrix, \mathbf{X} , has the first column a vector of 1s, to incorporate the intercept term. The dimensions of \mathbf{X} and \mathbf{v} are function of the number of QTNs. We took σ_μ^2 to be very large.

The idea here is that q_{ak} is an approximation to the conditional distribution of $\mu, \mathbf{a}, \mathbf{k}$ given all the other parameters. Specifically, it would be the posterior distribution of the regression coefficients if priors on the additive effect and dominance effect were joint normal, with prior distribution $\hat{p}(\mu, \mathbf{a}, \mathbf{k} | \gamma, \tau) \sim \mathcal{N}(\mathbf{0}, \mathbf{v})$. As a result,

$$q_{ak}(\mu, \mathbf{a}, \mathbf{k} | \gamma) = \frac{P(\mathbf{y} | \mu, \mathbf{a}, \mathbf{k}, \tau, \gamma, \mathbf{G}) \hat{p}(\mu, \mathbf{a}, \mathbf{k} | \gamma, \tau)}{\hat{p}(\mathbf{y} | \gamma, \tau, \mathbf{G})} \quad (10)$$

where the denominator $\hat{p}(\mathbf{y} | \gamma, \tau, \mathbf{G})$ is the integral of the numerator over $\mu, \mathbf{a}, \mathbf{k}$, which can be computed analytically as in prior D_2 below, leading to:

$$\hat{p}(\mathbf{y}|\gamma, \tau, \mathbf{G}) = (2\pi)^{-n/2} \tau^{n/2} \frac{|\mathbf{V}|^{1/2}}{|\mathbf{v}|^{1/2}} \exp\left[-0.5(\mathbf{y}^t \mathbf{y} - \mathbf{B}^t \mathbf{V}^{-1} \mathbf{B})\right]. \quad (11)$$

The new proposed values are then accepted with probability:

$$\begin{aligned} a &= \min\left(1, \frac{P(\mathbf{y}|\mu^*, \mathbf{a}^*, \mathbf{k}^*, \gamma^*, \tau, \mathbf{G})P(\mu^*, \mathbf{a}^*, \mathbf{k}^*|\gamma^*, \tau)P(\gamma^*)}{P(\mathbf{y}|\mu, \mathbf{a}, \mathbf{k}, \gamma, \tau, \mathbf{G})P(\mu, \mathbf{a}, \mathbf{k}|\gamma, \tau)P(\gamma)} \frac{q_{ak}(\mu, \mathbf{a}, \mathbf{k}|\gamma)}{q_{ak}(\mu^*, \mathbf{a}^*, \mathbf{k}^*|\gamma^*)} \frac{q(\gamma|\gamma^*)}{q(\gamma^*|\gamma)}\right) \\ &= \min\left(1, \frac{\hat{p}(\mathbf{y}|\gamma^*, \tau, \mathbf{G})P(\mu^*, \mathbf{a}^*, \mathbf{k}^*|\gamma^*, \tau)}{\hat{p}(\mathbf{y}|\gamma, \tau, \mathbf{G})\hat{p}(\mu^*, \mathbf{a}^*, \mathbf{k}^*|\gamma^*, \tau)} \frac{\hat{p}(\mu, \mathbf{a}, \mathbf{k}|\gamma, \tau)}{P(\mu, \mathbf{a}, \mathbf{k}|\gamma, \tau)} \frac{P(\gamma^*)}{P(\gamma)} \frac{q(\gamma|\gamma^*)}{q(\gamma^*|\gamma)}\right). \end{aligned} \quad (12)$$

As the effect of a QTN typically depends substantially on which other SNPs are QTNs, this joint update of all the QTNs effects at once is essential [to achieve a good mixing of the chain].

Update of τ We update τ by sampling from its full conditional distribution:

$$\tau|\mu, \boldsymbol{\beta}, \mathbf{y}, \mathbf{G} \sim \Gamma\left(n/2, \left(\sum_i (y_i - (\mu + \mathbf{x}_i(\mathbf{G})\boldsymbol{\beta}))^2\right)/2\right) \quad (13)$$

where $\mathbf{x}_i(\mathbf{G})$ is the i th row of the design matrix formed from genotypes \mathbf{G} .

Approximation of BFs from MCMC output To approximate the BF we applied the MCMC scheme with a prior odds of 1 (i.e. probability of 0.5 on each of the null and alternative models), and then estimate the BF for the alternative vs the null model using the estimated posterior odds, being the ratio of the number of iterations in which γ contains at least one SNP to the number of iterations in which γ contains no SNPs (adding one to both the numerator and denominator to deal with potential 0 counts).

B Analytical computations for prior D_2

With prior D_2 , we can evaluate some posterior quantities of interest by exact computations (*i.e.* without the need for a MCMC approach). For simplicity below we focus on the model where there is a single causative SNP (SNP s say) in the model; however the case where more SNPs are

included is easily handled, by extending the design matrix \mathbf{X} and adding further (σ_a^2, σ_d^2) pairs to the diagonal matrix $\boldsymbol{\nu}$ below.

Under prior D_2 the prior distribution for τ, β is of the form

$$\tau \sim \Gamma(\kappa/2, \lambda/2) \quad (14)$$

with density

$$p(\tau; \kappa, \lambda) = (\lambda/2)^{\kappa/2} \frac{\tau^{\kappa/2-1} \exp[-(\lambda/2)\tau]}{\Gamma(\kappa/2)}, \quad (15)$$

and

$$\beta|\tau \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\nu}/\tau) \quad (16)$$

where $\boldsymbol{\nu} = \text{diag}(\sigma_\mu^2, \sigma_a^2, \sigma_d^2)$.

The joint posterior distribution for τ, β is available analytically as

$$\tau|\mathbf{y}, \mathbf{G} \sim \Gamma((n + \kappa)/2, 0.5(\mathbf{y}^t \mathbf{y} - \mathbf{B}^t \boldsymbol{\Omega}^{-1} \mathbf{B} + \lambda)) \quad (17)$$

$$\beta|\tau, \mathbf{y}, \mathbf{G} \sim \mathcal{N}(\mathbf{B}, (1/\tau)\boldsymbol{\Omega}) \quad (18)$$

where

$$\mathbf{B} = \boldsymbol{\Omega} \mathbf{X}^t \mathbf{y} \quad (19)$$

$$\boldsymbol{\Omega} = (\boldsymbol{\nu}^{-1} + \mathbf{X}^t \mathbf{X})^{-1}. \quad (20)$$

Unlike in the main paper, we assume the first column of the design matrix \mathbf{X} is a vector of 1s, to incorporate the intercept term. Thus \mathbf{X} is a matrix with n rows (one for each individual), the first column containing all 1s (corresponding to an intercept term in the regression), and then two columns for each SNP included in the model, the first column being the SNP genotype (corresponding to the additive effect) and the second being a 0/1 indicator for whether the SNP genotype is a heterozygote (corresponding to the dominance effect).

We note that, in the limit $\sigma_\mu \rightarrow \infty$ and $\lambda \rightarrow 0$ the posteriors for τ, β changes appropriately with shifts and scaling operations on \mathbf{y} . In particular, in this limit:

1. The posterior mean for β , changes appropriately with shifts in \mathbf{y} . That is, adding c to each element of \mathbf{y} will add c to the first element of \mathbf{B} (the “mean” parameter, μ), leaving the other elements (the “effect” parameters, a and d) unchanged. This follows from the fact that $\mathbf{\Omega}^{-1}\mathbf{B} = \mathbf{X}^T\mathbf{y}$ implies $\mathbf{\Omega}^{-1}(\mathbf{B} + (c, 0, 0)^T) = \mathbf{X}^T(\mathbf{y} + c\mathbf{1})$, where $\mathbf{1}$ is the vector of length n whose elements are all 1s, as can be verified by straightforward algebra.
2. The posterior for τ is invariant to shifts in \mathbf{y} (ie adding some constant c to each element of \mathbf{y} does not change the posterior for τ), because the term $\mathbf{y}^t\mathbf{y} - \mathbf{B}^t\mathbf{\Omega}^{-1}\mathbf{B}$ does not change with shifts in \mathbf{y} . This follows from the fact that this term is equal to $(\mathbf{y} - \mathbf{X}\mathbf{B})^T\mathbf{y}$; that $(\mathbf{y} - \mathbf{X}\mathbf{B})^T$ does not change with shifts in \mathbf{y} (easily shown using 1 above); and $(\mathbf{y} - \mathbf{X}\mathbf{B})$ is orthogonal to $\mathbf{1}$, which can be checked by using the definition of \mathbf{B} to show that $(\mathbf{y} - \mathbf{X}\mathbf{B})^T\mathbf{X} = \mathbf{B}^T\boldsymbol{\nu}$, and then noting that the first column of \mathbf{X} is $\mathbf{1}$, and the first element of $\mathbf{B}^T\boldsymbol{\nu}$ is 0.
3. The posterior for τ scales appropriately with \mathbf{y} (that is, multiplying all elements of \mathbf{y} by some constant c essentially divides τ by c^2). This follows from elementary properties of the Gamma distribution, and because multiplying \mathbf{y} by c has the effect of multiplying the term $\mathbf{y}^t\mathbf{y} - \mathbf{B}^t\mathbf{\Omega}^{-1}\mathbf{B}$ by c^2 .
4. The posterior for β changes appropriately with shifts and scaling of \mathbf{y} . This follows from 1-3 above.

These invariance properties motivated us to use these limits ($\sigma_\mu \rightarrow \infty$ and $\lambda \rightarrow 0$) for the results in this paper. We also took $\kappa \rightarrow 0$. (In practice the results in this paper were obtained using a “large” value of σ_μ^2 although it is possible to derive the limiting results explicitly.)

As we now show, the Bayes Factor, $BF(s)$, that SNP s is a QTN vs. the “null” that no SNP is a QTN is also available analytically, and behaves sensibly in the limit $\sigma_\mu \rightarrow \infty$ and $\lambda, \kappa \rightarrow 0$.

First, integrating out β :

$$\begin{aligned}
P_s(\mathbf{y}|\tau, \mathbf{G}) &= \frac{P_s(\mathbf{y}|\boldsymbol{\beta}, \tau, \mathbf{G})P_s(\boldsymbol{\beta}|\tau)}{P_s(\boldsymbol{\beta}|\mathbf{y}, \tau, \mathbf{G})} \\
&= (2\pi)^{-n/2}\tau^{n/2}\frac{|\boldsymbol{\Omega}|^{1/2}}{|\boldsymbol{\nu}|^{1/2}}\exp\left[-0.5(\mathbf{y}^t\mathbf{y} - \mathbf{B}^t\mathbf{\Omega}^{-1}\mathbf{B})\tau\right].
\end{aligned} \tag{21}$$

Now integrating out τ :

$$\begin{aligned} P_s(\mathbf{y}|\mathbf{G}) &= \int_0^\infty P_s(\mathbf{y}|\tau, \mathbf{G})P(\tau) d\tau \\ &= (2\pi)^{-n/2} \frac{|\boldsymbol{\Omega}|^{1/2}}{|\boldsymbol{\nu}|^{1/2}} \int_0^\infty \tau^{(n+\kappa)/2-1} \exp\left[-0.5(\mathbf{y}^t\mathbf{y} - \mathbf{B}^t\boldsymbol{\Omega}^{-1}\mathbf{B} + \lambda)\tau\right] d\tau. \end{aligned} \quad (22)$$

Recognizing the above integral as the normalising constant of a $\Gamma((n + \kappa)/2, 0.5(\mathbf{y}^t\mathbf{y} - \mathbf{B}^t\boldsymbol{\Omega}^{-1}\mathbf{B} + \lambda))$ distribution, we obtain:

$$P_s(\mathbf{y}|\mathbf{G}) = (2\pi)^{-n/2} \frac{|\boldsymbol{\Omega}|^{1/2}}{|\boldsymbol{\nu}|^{1/2}} (\lambda/2)^{\kappa/2} \frac{\Gamma((n + \kappa)/2)}{\Gamma(\kappa/2)} \left(\frac{\mathbf{y}^t\mathbf{y} - \mathbf{B}^t\boldsymbol{\Omega}^{-1}\mathbf{B} + \lambda}{2} \right)^{-(n+\kappa)/2}. \quad (23)$$

The above expression gives the probability of the observed phenotype data under the hypothesis that SNP s is a QTN. It can also be used to obtain the probability of the phenotype data under the “null” of no effect, by setting $\nu = \sigma_\mu^2$, and X to be the vector of all 1s, and substituting these into (19) and (20) to compute \mathbf{B} and $\boldsymbol{\Omega}$. This gives

$$P_0(\mathbf{y}) = (2\pi)^{-n/2} \frac{\Omega_0^{1/2}}{\sigma_\mu} (\lambda/2)^{\kappa/2} \frac{\Gamma((n + \kappa)/2)}{\Gamma(\kappa/2)} \left(\frac{\mathbf{y}^t\mathbf{y} - \Omega_0 n^2 \bar{y}^2 + \lambda}{2} \right)^{-(n+\kappa)/2} \quad (24)$$

where $\Omega_0 = ((\sigma_\mu^2)^{-1} + n)^{-1}$.

The Bayes Factor $BF(s)$ is then the ratio of (23) to (24):

$$BF(s) = \frac{P_s(\mathbf{y}|\mathbf{G})}{P_0(\mathbf{y})} = \frac{|\boldsymbol{\Omega}|^{1/2}}{\Omega_0^{1/2}} \cdot \frac{1}{\sigma_a \sigma_d} \cdot \left[\frac{\mathbf{y}^t\mathbf{y} - \mathbf{B}^t\boldsymbol{\Omega}^{-1}\mathbf{B} + \lambda}{\mathbf{y}^t\mathbf{y} - \Omega_0 n^2 \bar{y}^2 + \lambda} \right]^{-(n+\kappa)/2}. \quad (25)$$

Limiting value of the BF Our priors for β and τ are obtained by taking the limit $+\infty$ for σ_μ and 0 for both λ and κ . The limit of $BF(s)$ with respect to λ and κ is:

$$\lim_{\substack{\lambda \rightarrow 0 \\ \kappa \rightarrow 0}} BF(s) = \frac{|\boldsymbol{\Omega}|^{1/2}}{\Omega_0^{1/2}} \cdot \frac{1}{\sigma_a \sigma_d} \cdot \left[\frac{\mathbf{y}^t\mathbf{y} - \mathbf{B}^t\boldsymbol{\Omega}^{-1}\mathbf{B}}{\mathbf{y}^t\mathbf{y} - \Omega_0 n^2 \bar{y}^2} \right]^{-n/2} \quad (26)$$

Finally, taking the limit when $\sigma_\mu \rightarrow \infty$ is straightforward because $\boldsymbol{\Omega}$ has a finite limit as $\sigma_\mu \rightarrow \infty$.

Bayes Factor for multiple SNPs If we assume the effects of multiple SNPs are additive, the computation of the Bayes Factor, $BF(s_1, s_2, \dots, s_p)$ that p SNPs (s_1, \dots, s_p) are QTNs vs. no SNP is a QTN can be done following the same approach. $BF(s_1, s_2, \dots, s_p)$ is computed with equation

(25), after modification of the design matrix \mathbf{X} and the prior matrix $\boldsymbol{\nu}$. \mathbf{X} is then a $(n \times (2p + 1))$ matrix: the first column of \mathbf{X} is filled with 1's and each pair of column $\{(2i, 2i + 1); i \in [1, p]\}$ relates the individuals with their genotypes. $\boldsymbol{\nu}$ is a $((2p + 1) \times (2p + 1))$ diagonal matrix with $\nu_{1,1} = \sigma_\mu^2$, $\nu_{2i,2i} = \sigma_a^2$ and $\nu_{2i+1,2i+1} = \sigma_d^2$ for $i \in [1, p]$.

Bayes Factor for a region Here we show how to compute the Bayes Factor, BF, for association ($H1$) vs no association ($H0$). Given a prior distribution on the number of QTNs $p(l)$ on $[1..L]$, we have:

$$\text{BF} = \sum_{l=1}^L p(l) \frac{1}{\binom{n_s}{l}} \sum_{(s_1, \dots, s_l) \in c(l, n_s)} \text{BF}(s_1, \dots, s_l) \quad (27)$$

where $c(l, n_s)$ denotes the ensemble of all possible combinations of l SNPs taken from all n_s SNPs.

In the particular case where $L = 1$, this reduces to:

$$\text{BF} = (1/n_s) \sum_{s=1}^{n_s} \text{BF}(s)$$

which is the mean of the single SNP Bayes Factors over the region.

C R code to compute BF for prior D_2

The following R code computes the $\log_{10}(\text{BF})$ for a single-SNP, whose genotypes (coded as 0, 1 or 2 copies of the minor allele) are contained in the vector \mathbf{g} , given a corresponding vector of phenotypes \mathbf{y} , and user-supplied values for σ_a and σ_d . (Individuals with missing data in either \mathbf{g} or \mathbf{y} are ignored in this calculation.)

```
logBF = function(g, y, sigmaa, sigmad) {
  subset = complete.cases(y) & complete.cases(g)
  y=y[subset]
  g=g[subset]
  n=length(g)
```

```

X = cbind(rep(1,n),g,g==1)
invnu = diag(c(0,1/sigmaa^2,1/sigmad^2))
invOmega = invnu + t(X) %*% X
B = solve(invOmega, t(X) %*% cbind(y))
invOmega0 = n
return(-0.5*log10(det(invOmega)) + 0.5*log10(invOmega0) - log10(sigmaa)
- log10(sigmad) - (n/2) * (log10( t(y- X %*% B) %*% y)
- log(t(y) %*% y - n*mean(y)^2) ))
}

```

List of Figures

1	Power comparisons. Each colored line shows power of test varying with significance threshold (type I error). Black : Bayes Factor from our method (prior D_2); Green : p_{\min} (allelic test) ; Red : p_{\min} (genotype test) ; Blue : p_{reg} , multiple regression; Grey : BF_{max} . Each column of figures shows results for data analysed under the “resequencing design” (left) and the “tag SNP design” (right); Each row shows results for the four different simulation scenarios. Top row: A) single common variant, modest dominance; Bottom row: B) single common variant, strong dominance for minor allele	37
1	Continued. Top row: C) single rare variant, no dominance; Bottom row: D) multiple common variants.	38
2	Comparison of results for resequencing design (x-axis) and tag SNP design (y-axis). Panels show a) errors in the estimates (posterior means) of the heterozygote effect ($a + d$); b) errors in the estimates (posterior means) of the main effect (a); c) posterior probability of being a QTN ($P((a, d) \neq (0, 0))$) assigned to the causal variant	39
3	Examination of potential effect of different tag SNP strategies on power, when the causal variant is rare ($0.01 < \text{MAF} < 0.05$). Solid line: Resequencing design; dashed line: tag SNP design, with tags selected using LDselect [?]; dotted line: tag SNP design, with all SNPs except the causal SNP as tags.	40
4	Power of the multipoint approach in the rare variant scenario for two different imputation algorithms.	41
5	Scatter plot of samples from prior distribution of a (x-axis) and $a + d$ (y-axis), for priors D_1 (black) and D_2 (blue). The solid yellow line corresponds to $d = 0$ (additivity). The dashed red lines are the limits above and below which a SNP exhibits over-dominance.	42
6	Comparison of inferences using prior D_1 and D_2 for the Bayes Factor (left) and the posterior probability assigned to the causal locus being a QTN (right). Results shown are for all datasets for the common variant scenario A and B and for both the resequencing design and the tag SNP design. The discrepancy between the larger estimated BFs is caused by the fact we used insufficient MCMC iterations to accurately estimate very large BFs ($> 10^6$) under prior D_1	43
7	Illustration of how a multi-QTN model can provide fuller explanations than a 1-QTN model for observed associations. The figure shows, for each SNP in a dataset simulated under scenario D, the estimated posterior probability that it is a QTN, conditional on an association being observed. Left: Results from 1-QTN model. Right: Results from multi-QTN model allowing up to four QTNs. The four actual QTNs are indicated with a star. Colors of the vertical lines indicate tag SNP “bins” (i.e. groups of SNPs tagged by the same variant).	44

- 8 Results for the SCN1A dataset. Left panel shows the posterior probability assigned to each SNP being a QTN, with filled circles denoting tag SNPs, and open circles denoting non-tag SNPs. The right panel shows (in gray) estimated posterior densities of the additive effect for each of the 7 SNPs assigned the highest posterior probabilities of non-zero effect (representing 90% of the posterior mass). The average of these curves is shown in black. 45

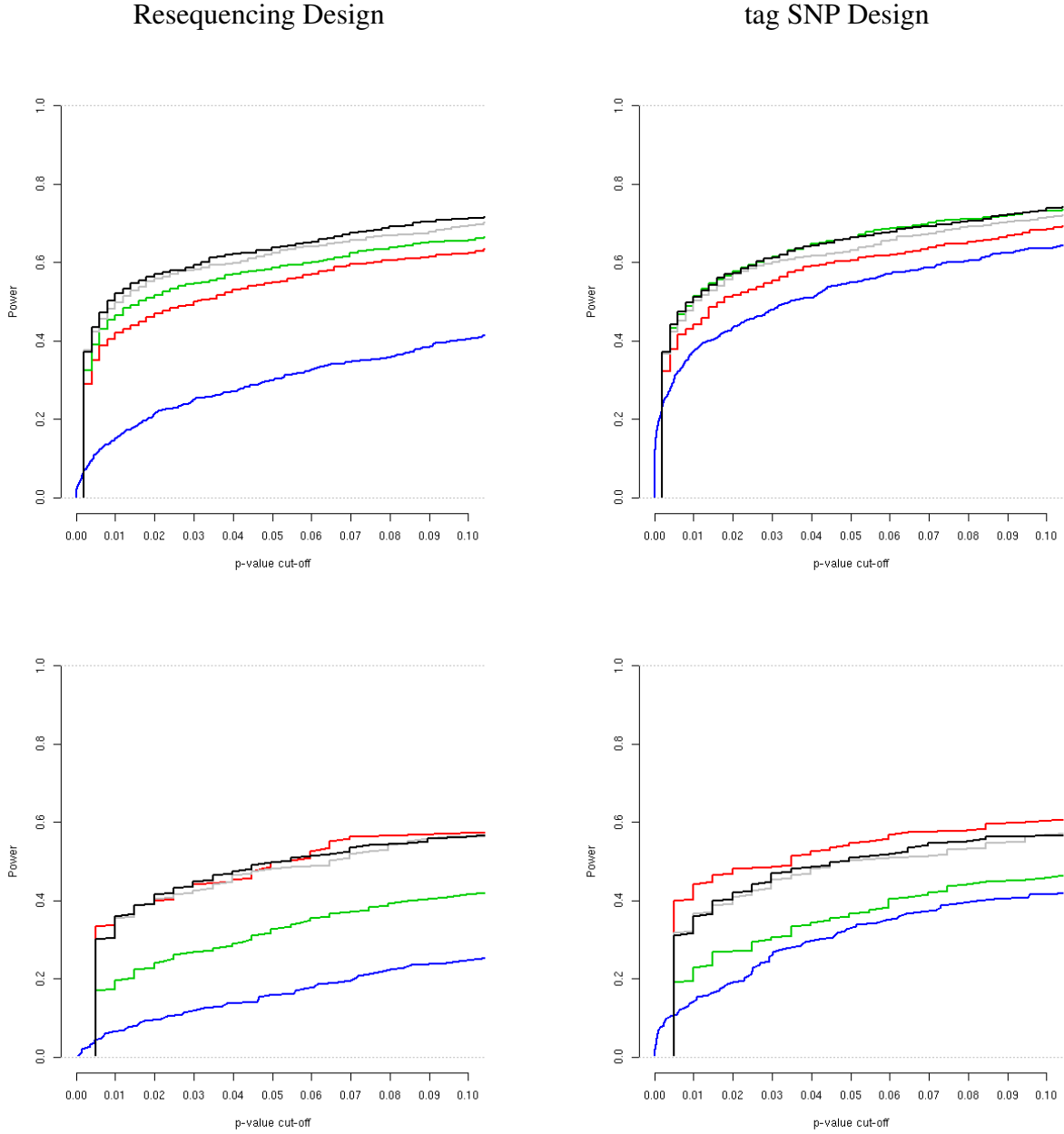


Figure 1: Power comparisons. Each colored line shows power of test varying with significance threshold (type I error). Black : Bayes Factor from our method (prior D_2); Green : p_{\min} (allelic test) ; Red : p_{\min} (genotype test) ; Blue : p_{reg} , multiple regression; Grey : BF_{max} . Each column of figures shows results for data analysed under the “resequencing design” (left) and the “tag SNP design” (right); Each row shows results for the four different simulation scenarios. Top row: A) single common variant, modest dominance; Bottom row: B) single common variant, strong dominance for minor allele

Resequencing Design

tag SNP Design

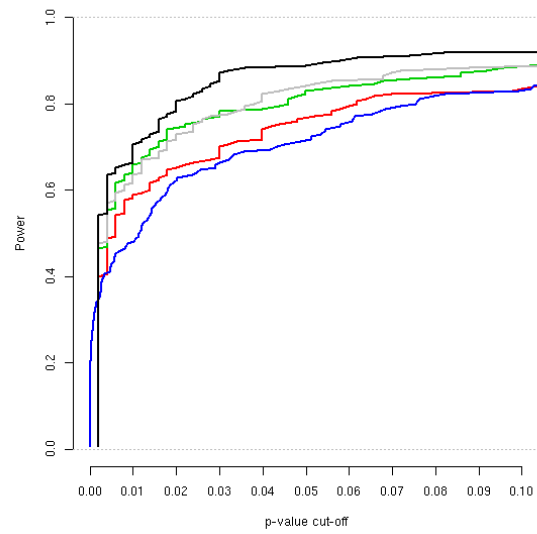
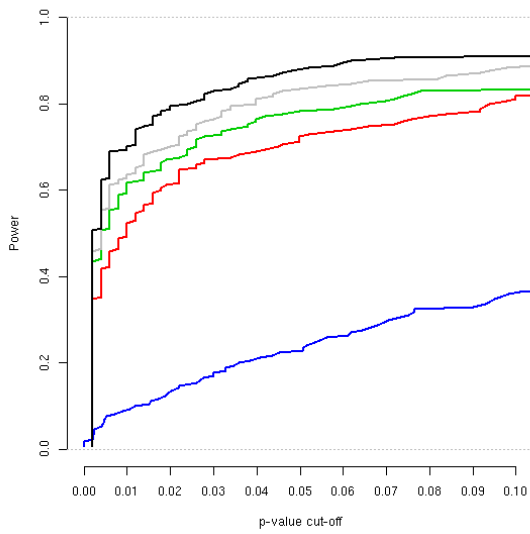
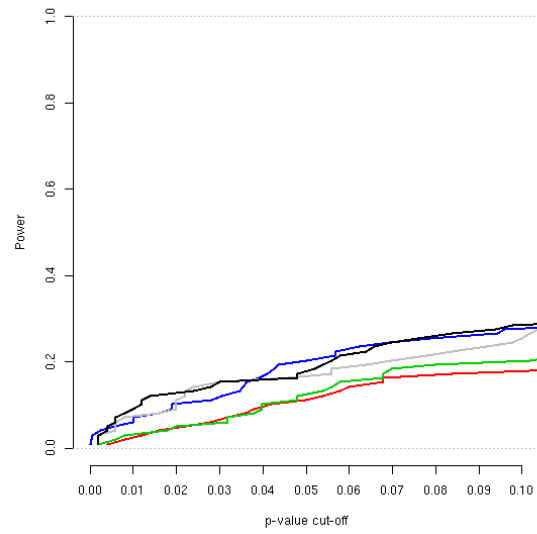
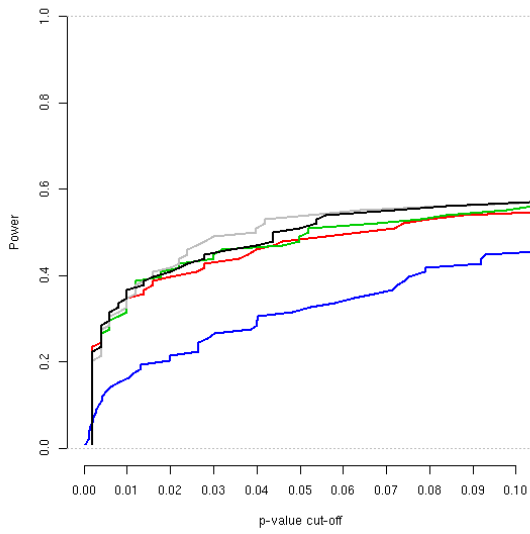


Figure 1: Continued. Top row: C) single rare variant, no dominance; Bottom row: D) multiple common variants.

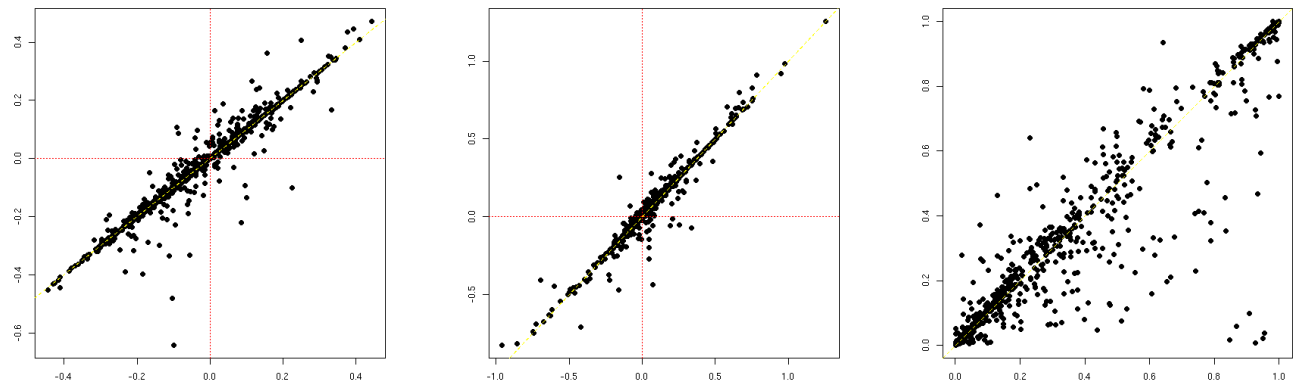


Figure 2: Comparison of results for resequencing design (x-axis) and tag SNP design (y-axis). Panels show a) errors in the estimates (posterior means) of the heterozygote effect ($a + d$); b) errors in the estimates (posterior means) of the main effect (a); c) posterior probability of being a QTN ($P((a, d) \neq (0, 0))$) assigned to the causal variant .

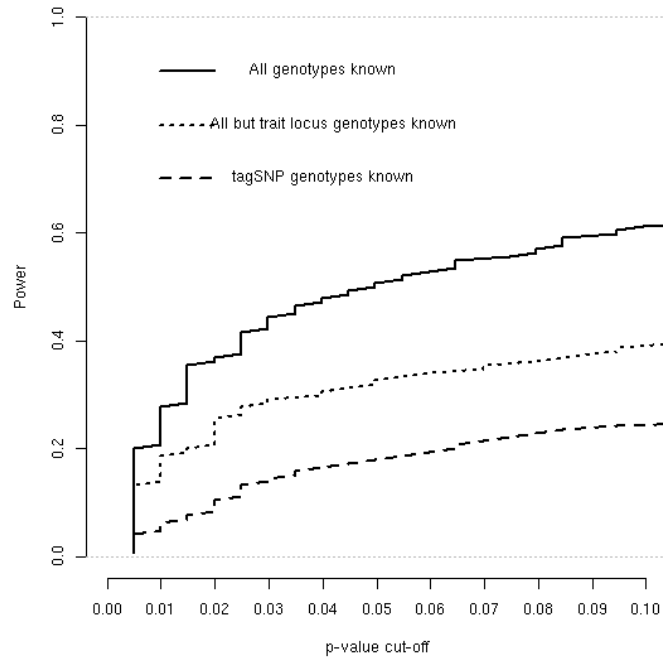


Figure 3: Examination of potential effect of different tag SNP strategies on power, when the causal variant is rare ($0.01 < \text{MAF} < 0.05$). Solid line: Resequencing design; dashed line: tag SNP design, with tags selected using method from [?]; dotted line: tag SNP design, with all SNPs except the causal SNP as tags.

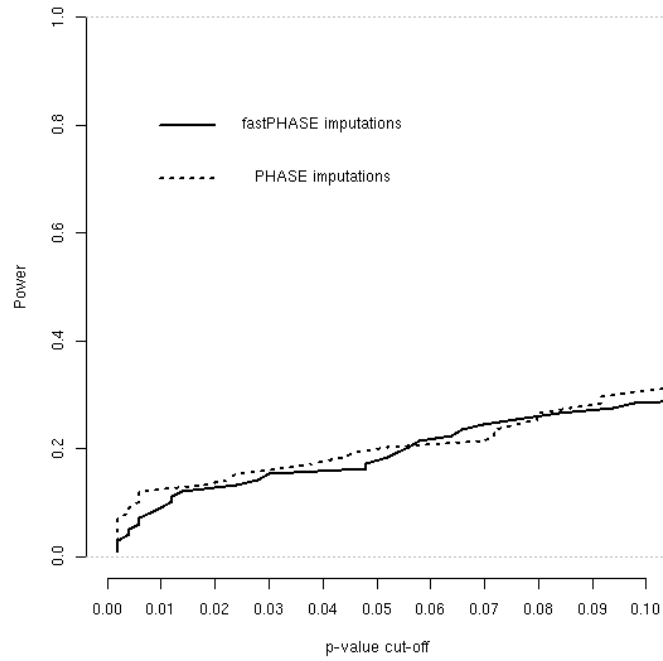


Figure 4: Power of the multipoint approach in the rare variant scenario for two different imputation algorithms.

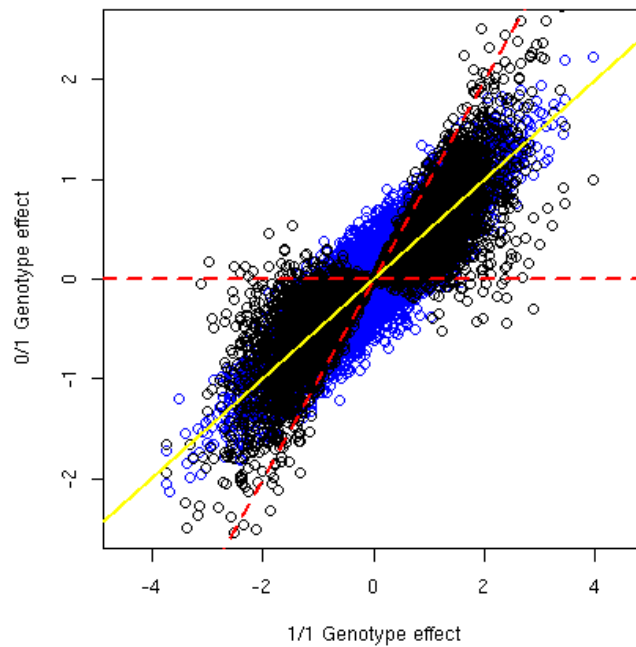


Figure 5: Scatter plot of samples from prior distribution of a (x-axis) and $a + d$ (y-axis), for priors D_1 (black) and D_2 (blue). The solid yellow line corresponds to $d = 0$ (additivity). The dashed red lines are the limits above and below which a SNP exhibits over-dominance.

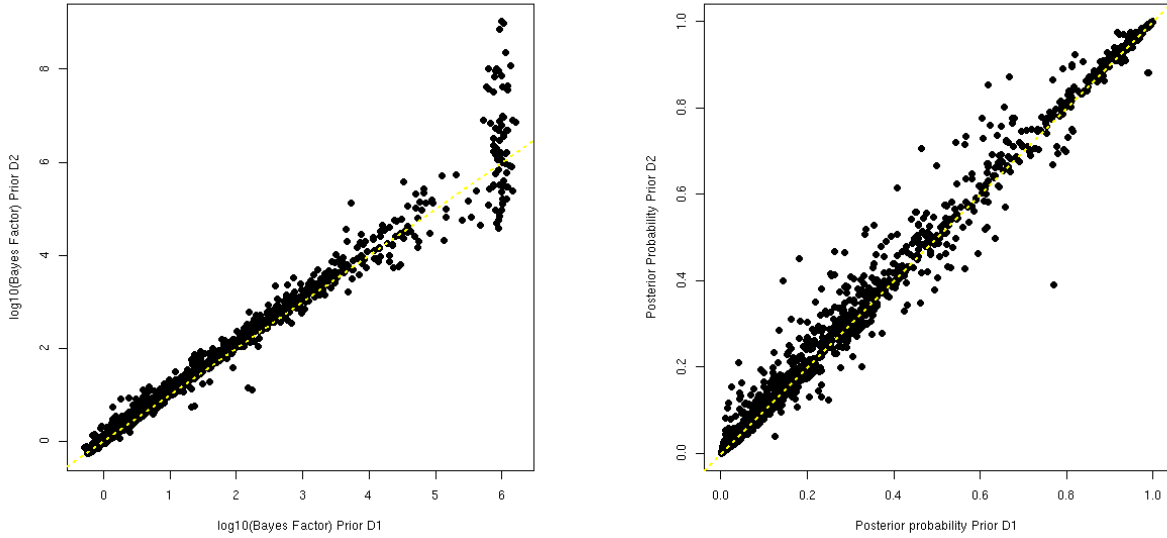


Figure 6: Comparison of inferences using prior D_1 and D_2 for the Bayes Factor (left) and the posterior probability assigned to the causal locus being a QTN (right). Results shown are for all datasets for the common variant scenario A and B and for both the resequencing design and the tag SNP design. The discrepancy between the larger estimated BFs is caused by the fact we used insufficient MCMC iterations to accurately estimate very large BFs ($> 10^6$) under prior D_1 .

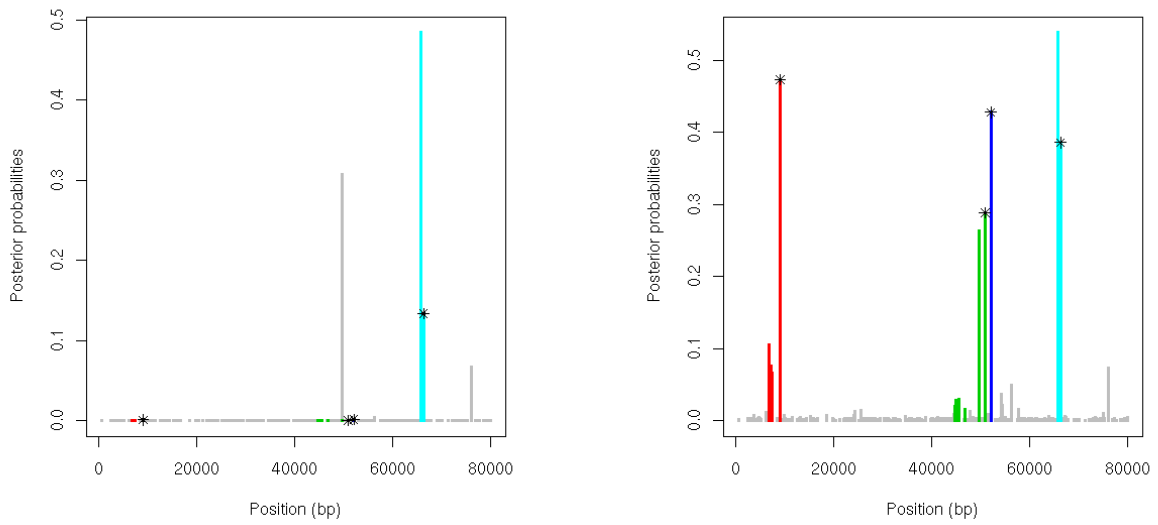


Figure 7: Illustration of how a multi-QTN model can provide fuller explanations than a 1-QTN model for observed associations. The figure shows, for each SNP in a dataset simulated under scenario D, the estimated posterior probability that it is a QTN, conditional on an association being observed. Left: Results from 1-QTN model. Right: Results from multi-QTN model allowing up to four QTNs. The four actual QTNs are indicated with a star. Colors of the vertical lines indicate tag SNP “bins” (i.e. groups of SNPs tagged by the same variant).

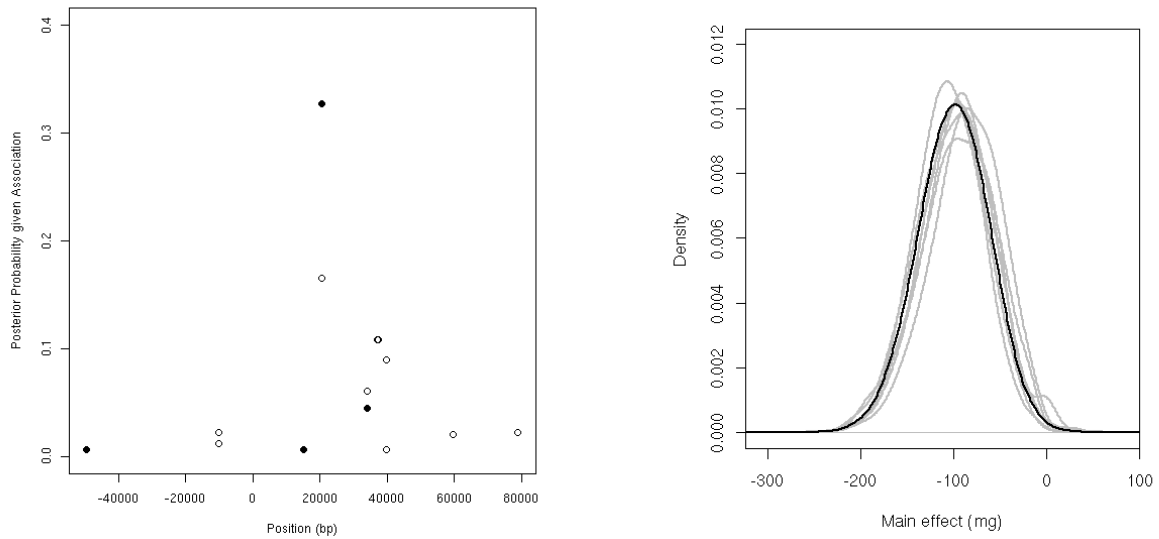


Figure 8: Results for the SCN1A dataset. Left panel shows the posterior probability assigned to each SNP being a QTN, with filled circles denoting tag SNPs, and open circles denoting non-tag SNPs. The right panel shows (in gray) estimated posterior densities of the additive effect for each of the 7 SNPs assigned the highest posterior probabilities of non-zero effect (representing 90% of the posterior mass). The average of these curves is shown in black.