The Annals of Applied Statistics
2011, Vol. 5, No. 2A, 873–893
DOI: 10.1214/10-AOAS411

SPARSE PARTITIONING: NONLINEAR REGRESSION WITH BINARY OR TERTIARY PREDICTORS, WITH APPLICATION TO ASSOCIATION STUDIES

BY DOUG SPEED[1] AND SIMON TAVARÉ[2]

University of Cambridge

This paper presents *Sparse Partitioning*, a Bayesian method for identifying predictors that either individually or in combination with others affect a response variable. The method is designed for regression problems involving binary or tertiary predictors and allows the number of predictors to exceed the size of the sample, two properties which make it well suited for association studies.

Sparse Partitioning differs from other regression methods by placing no restrictions on how the predictors may influence the response. To compensate for this generality, *Sparse Partitioning* implements a novel way of exploring the model space. It searches for high posterior probability partitions of the predictor set, where each partition defines groups of predictors that jointly influence the response.

The result is a robust method that requires no prior knowledge of the true predictor–response relationship. Testing on simulated data suggests *Sparse Partitioning* will typically match the performance of an existing method on a data set which obeys the existing method's model assumptions. When these assumptions are violated, *Sparse Partitioning* will generally offer superior performance.

Introduction. In recent years association studies have surged in popularity, driven by the ability to interrogate the genome in ever-increasing detail [McCarthy et al. (2008)]. The common aim of these studies is to detect genomic variants that are linked with a particular phenotype. It is hoped that detecting such variants will bring us closer to understanding the biological processes at work.

So far these studies have had mixed results. While variants with strong effects are picked up fairly readily [e.g., The Wellcome Trust Case Control Consortium (2007)], there is speculation that more subtle associations are being missed [Cordell (2009)]. This suggests the need to develop more sophisticated tools that are able to explore beyond the obvious [Stephens and Balding (2009)].

Formally, an association study can be viewed as a regression problem consisting of n data points (the samples) and N predictors (the variants). In this paper

Received December 2009; revised October 2010.

[1] Supported by an EPSRC doctoral fellowship.

[2] Supported in part by NIH ENDGAME Grant U01 HL084706 and NIH Grant P50 HG02790.

Key words and phrases. Sparse Bayesian modeling, nonlinear regression, association studies, large p small n problems.

we consider the case when each predictor takes either two or three unique values. This is common in association studies. For example, a predictor might record presence or absence of a mutation, or whether a variant is in a neutral, amplified or deleted state. We also allow for "large p, small n problems" in which the number of predictors exceeds the sample size. Again, this is often the case with association studies, owing to the abundance of genetic variants available to examine.

Currently available regression tools can be characterized by how they permit predictors to influence the response. For example, many fit an additive model, which overlooks the possibility that interactions between predictors might affect the response. The methods which permit interactions will generally specify the type of interactions they allow. A key factor affecting performance is whether the data set being examined conforms to the restrictions the method imposes. *Sparse Partitioning* tries to avoid placing restrictions on the underlying model relationship. This should enable it to maintain power in scenarios where other methods might fail.

Section 1 describes some of the existing methods suitable for processing high-dimensional data. Sections 2 and 3 briefly outline the *Sparse Partitioning* methodology. Sections 4 and 5 test the performance of *Sparse Partitioning* compared to existing methods, while Section 6 concludes the paper. Additional details are provided in the Appendix and supplementary material.

1. Existing methods.

The task of a regression method is to infer how the predictors influence the response. Let the vector $\mathbf{Y}$ (size $n \times 1$) contain the response values and the matrix $\mathbf{X}$ (size $n \times N$) contain the predictors. For the ith data point, Y_i denotes its response, while $X_{i,1}, \dots, X_{i,g}, \dots, X_{i,N}$ denote its predictor values. If we write the regression model as $l(\mathbb{E}(\mathbf{Y})) = f(\mathbf{X})$, where l is a specified link function, the aim is to deduce properties of $f(\mathbf{X})$, the "underlying relationship." In particular, we wish to identify the subset of predictors that contribute toward $f(\mathbf{X})$.

Consider writing the underlying relationship as

$$f(\mathbf{X}) = f_1(X_{G_{1,1}}, \dots, X_{G_{1,s_1}}) + \cdots + f_K(X_{G_{K,1}}, \dots, X_{G_{K,s_K}}).$$

Under this representation, $f(\mathbf{X})$ is influenced by additive contributions from groups of interacting predictors. f_k describes the contribution of predictors $G_{k,1}, \dots, G_{k,j}, \dots, G_{k,s_k}$ to $f(\mathbf{X})$. In this paper, additivity is not considered an interaction. Therefore, the predictors in each group are said to interact with each other, but not to interact with a predictor in a different group. For the most general relationship, all predictors feature in one group. In practice, however, we suspect $f(\mathbf{X})$ is far simpler.

We have distinguished existing methods based on two features of their underlying relationships: whether they permit more than one group of predictors to contribute to $f(\mathbf{X})$ and whether they permit interactions between contributing predictors. Figure 1 demonstrates the four possibilities, using the case when the predictors are binary and the response is continuous.

	ONE GROUP OF PREDICTORS	**MULTIPLE GROUPS OF PREDICTORS**
NO INTERACTIONS	$\mathbf{Y} = \alpha + \beta X_g$ e.g., *Single*	$\mathbf{Y} = \alpha + \sum_1^N \beta_g X_g$ e.g., *SSS*
INTERACTIONS	$\mathbf{Y} = f(X_{G_1}, \dots, X_{G_s})$ e.g., *Pairs*, *CART*, *RF*	$\mathbf{Y} = f_1(X_{G_{1,1}}, \dots, X_{G_{1,s_1}}) + \cdots +$ $f_K(X_{G_{K,1}}, \dots, X_{G_{K,s_K}})$ e.g., *Logic*, *MARS*, *Sparse Partitioning*

FIG. 1. *Regression methods can be categorized according to two features of their underlying relationship. This table shows the four possibilities, for the case of binary predictors and a continuous response. Explanations of the existing methods, Single, Pairs, CART, RF, SSS, Logic and MARS, are provided in the main text.*

1.1. *One group, maximum group size one.*

$$f(\mathbf{X}) = f_1(X_{G_{1,1}}).$$

The simplest assumption supposes the response is influenced by only one predictor. Most methods in this category are equivalent to performing a maximum likelihood test comparing a null hypothesis, $f(\mathbf{X}) =$ constant, with an alternative, $f(\mathbf{X}) = f_1(X_g)$. *Single* is our implementation of such a method. Considering that these methods can only detect an associated predictor by its marginal effect, they are surprisingly successful. They are also extremely fast to run and therefore very popular [e.g., Stranger et al. (2007)].

Bayesian alternatives are possible [e.g., Balding (2006)] and useful if certain predictors are thought *a priori* more likely to be associated. Otherwise they will generally produce the same results as classical methods.

1.2. *One group, maximum group size greater than one.*

$$f(\mathbf{X}) = f_1(X_{G_{1,1}}, \dots, X_{G_{1,s_1}}).$$

Even for very high-dimensional problems (>500,000 predictors) it is possible to test exhaustively all pairwise models [cf. Marchini, Donnelly and Cardon (2005)]. The method *Pairs* is our extension of *Single*, performing a maximum likelihood test for each pair of predictors. While the method could be extended further to consider three or four way interactions, this is often infeasible due to computation time.

A second method in this category is *CART* [Classification and Regression Trees; Breiman et al. (1984)]. *CART* differs from *Pairs* in not insisting on the full interaction model for associated predictors. For example, a *CART* model containing two associated predictors might have only 3 degrees of freedom, even though there are 4 unique vector values present. Random Forest [Breiman (2004)] offers a stochastic interpretation of this method, constructing a large number of trees in a quasi-random fashion and summarizing their properties.

1.3. *More than one group, maximum group size one.*

$$f(\mathbf{X}) = f_1(X_{G_{1,1}}) + \cdots + f_K(X_{G_{K,1}}).$$

This underlying relationship allows more than one predictor to be causal, but restricts the causal predictors to contributing additively. When there are more predictors than samples, the standard multiple regression model will become oversaturated and fail.

The classical solution, adopted by Variable Subset Selection, Lasso and Ridge Regression [described in Hastie, Tibshirani and Friedman (2001)], is to introduce a penalty term that limits the number of contributing predictors. However, this penalty term can appear quite arbitrary. An alternative is offered by Bayesian methods [Wang et al. (2005); Zhang et al. (2005); Hoggart et al. (2008)]. These methods allow our preference for sparse models to be reflected in the prior distribution. We picked Shotgun Stochastic Search [Hans, Dobra and West (2007)] to represent this category of methods in the simulation studies.

1.4. *More than one group, maximum group size greater than one.*

$$f(\mathbf{X}) = f_1(X_{G_{1,1}}, \dots, X_{G_{1,s_1}}) + \cdots + f_K(X_{G_{K,1}}, \dots, X_{G_{K,s_K}}).$$

Allowing both interactions and multiple groups of predictors to contribute to the underlying relationship has the potential of most accurately describing the true model. However, both decisions increase the size of the model space and so the difficulty of identifying the true model.

Logic Regression [Ruczinski, Kooperberg and LeBlanc (2003)] and Multivariate Adaptive Regression Splines are two of the few methods in this class. Both methods place restrictions on the permitted functions which reduce the size of the model space; *Logic* insists on Boolean operators, while *MARS* uses products of hinge functions. *Sparse Partitioning* falls into this category, but places no restriction on the types of functions allowed.

2. Formulating the regression as a partitioning.

In order to describe *Sparse Partitioning*'s methodology, it is convenient to formulate the regression problem as a search for high scoring partitions. Consider how the underlying relationship groups predictors:

$$\begin{aligned} f(\mathbf{X}) &= f_1(X_{G_{1,1}}, \dots, X_{G_{1,s_1}}) + \cdots + f_K(X_{G_{K,1}}, \dots, X_{G_{K,s_K}}) \\ &= f_1(X_{\mathbf{G}_1}) + \cdots + f_K(X_{\mathbf{G}_K}). \end{aligned}$$

The disjoint sets $\mathbf{G}_1, \mathbf{G}_2, \dots, \mathbf{G}_K$ index groups of associated predictors. If we let $\mathbf{G}_0$ index the "null group"—the group of predictors in no way associated with the response—then $\mathbb{G} = \{\mathbf{G}_0, \mathbf{G}_1, \mathbf{G}_2, \dots, \mathbf{G}_K\}$ defines a partitioning of $\{1, 2, \dots, N\}$.

In the representation above, predictors are not allowed to feature in more than one nonnull group. To avoid this restriction, while at the same time maintaining

$$\begin{aligned} \mathbf{I} &= (\overbrace{1\ 1}^{\mathbf{G}_1}\ \overbrace{0\ 0}^{\mathbf{G}_0}\ \overbrace{2}^{\mathbf{G}_2}) \\ f(\mathbf{X}) &= f_1(X_1, X_2) + f_2(X_5) \end{aligned} \quad \Rightarrow \quad \begin{aligned} &\text{For binary predictors and a continuous response:} \\ &\mathbf{Y} = \alpha_0 + \alpha_{1,1}X_1(1 - X_2) + \alpha_{1,2}(1 - X_1)X_2 \\ &\qquad + \alpha_{1,3}X_1X_2 + \alpha_{2,1}X_5 \end{aligned}$$

FIG. 2. *An example of a partitioning for a problem containing five binary predictors* (*each valued* 0 *or* 1) *and a continuous response.*

a partitioning, the predictor set is expanded to contain C copies of each predictor and N is increased accordingly. For the remainder of this paper, we describe the method supposing $C = 1$, then explain the changes required when this is not the case.

A partition can also be described by the vector $\mathbf{I} = (I_1, I_2, \dots, I_N)$, where I_g indicates to which group predictor g belongs. This notation will be useful later on and also reminds us that the ordering within groups is not important. Figure 2 gives an example of a simple partitioning and the underlying relationship to which it refers.

The focus of *Sparse Partitioning* is to determine properties of the partitioning defined by the underlying relationship. Our main desire is to identify which predictors are not in the null group. However, it is also useful to know whether predictors feature in the same nonnull group, indicating interactions. The advantage of formulating the problem in terms of partitions is that the model space is likely too vast to hope to detect accurately the explicit underlying relationship (i.e., determine $\mathbf{f} = \{f_1, f_2, \dots, f_K\}$ as well as $\mathbb{G}$). Fortunately, we are usually more interested in detecting which predictors are involved, rather than exactly how they contribute (the latter can be saved for follow-up analysis).

3. Sparse Partitioning methodology. *Sparse Partitioning* is a Bayesian methodology, so it follows the usual steps of deciding a prior, calculating the likelihood, then computing the posterior distribution of models through use of Bayes formula:

$$\mathbb{P}(\text{Model}|\text{Data}) \propto \mathbb{P}(\text{Model}) \times \mathbb{P}(\text{Data}|\text{Model}).$$

Each model is defined by $\{\mathbb{G}, \mathbf{f}\}$, a partition and a corresponding set of functions. However, $\mathbf{f}$ is considered a nuisance parameter, so we are interested in determining the marginal posterior $\mathbb{P}(\mathbb{G}|\text{Data})$.

3.1. *Prior.*

$$\mathbb{P}(\text{Model}) = \mathbb{P}(\mathbb{G}, \mathbf{f}) = \mathbb{P}(\mathbb{G}) \times \mathbb{P}(\mathbf{f}|\mathbb{G}).$$

The prior for the partition is based on our belief in p_g, the probability that predictor g is associated with the response. Therefore, $\mathbb{P}(\mathbb{G}) = \mathbb{P}(\mathbf{I})$ is constructed such that $\sum_{\mathbf{I}:I_g \neq 0} \mathbb{P}(\mathbf{I}) = p_g$. Two partitions containing the same associations are given equal weight. When multiple copies of each predictor are permitted, p_g equals the probability that at least one copy of predictor g is associated. *Sparse Partitioning* keeps fixed the values of p_g, however, it is straightforward to allow them to vary if more detailed prior information is available.

Given $\mathbf{G}_k$, the relevant information of function f_k can be described by the values it assigns each node (unique vector value) of $X_{\mathbf{G}_k}$. For the example in Figure 2, $\mathbf{f}$ can be described by $\boldsymbol{\alpha} = (\alpha_0, \alpha_{1,1}, \alpha_{1,2}, \alpha_{1,3}, \alpha_{2,1})$. Therefore, the prior on $\mathbf{f}$ is equivalent to a prior on $\boldsymbol{\alpha}$, for which *Sparse Partitioning* uses a multivariate normal distribution.

3.2. *Likelihood.* The likelihood is determined by the regression model. When the response is continuous (e.g., a quantitative trait), *Sparse Partitioning* supposes the residuals are normally distributed. When the response is binary (e.g., a case-control experiment with affected and unaffected patients), *Sparse Partitioning* uses a logit link function. The marginal likelihood is obtained by integrating across the function parameters:

$$\mathbb{P}(\text{Data}|\mathbb{G}) = \int_{\mathbf{f}} \mathbb{P}(\text{Data}|\mathbf{f}, \mathbb{G})\mathbb{P}(\mathbf{f}|\mathbb{G})\, d\mathbf{f} = \int_{\boldsymbol{\alpha}} \mathbb{P}(\text{Data}|\boldsymbol{\alpha}, \mathbb{G})\mathbb{P}(\boldsymbol{\alpha}|\mathbb{G})\, d\boldsymbol{\alpha}.$$

3.3. *Posterior.* Explicit calculation of $\mathbb{P}(\mathbb{G}|\text{Data})$ would require an exhaustive search of the space of partitions, which is infeasible even for reasonably sized problems. Therefore, *Sparse Partitioning* uses Markov Chain Monte Carlo (MCMC) techniques to estimate statistics of this posterior distribution. Within each MCMC iteration, two sampling stages are used: the first proposes, in turn, a change to each component of $\mathbf{I}$; the second proposes a change to one element of $\mathbb{G}$. The bulk of *Sparse Partitioning*'s processing time is spent sampling from the posterior distribution. Therefore, it is convenient that the two stages can be parallelized with an almost linear speed-up.

Full details of the methodology are provided in Appendix and Sections 1, 2 and 3 of the Supplementary Material [Speed and Tavaré (2010)].

4. Simulation studies. In total, ten simulation studies were carried out, designed to test various aspects of *Sparse Partitioning*'s performance and make comparisons with existing methods. This section presents results from the first study. Further details of the methods used to simulate data and the results from the remaining nine studies are provided in Section 4 of the Supplementary Material [Speed and Tavaré (2010)].

Typically, each simulated data set consisted of 100 samples, each of 1000 predictors, three of which were causal for the response. Each regression method was asked to identify its top three associations and was then scored by how many causal predictors it correctly identified. Empirical estimates were obtained by averaging over 100 data sets.

TABLE 1
The first simulation study considered three different underlying relationships

Model	Underlying relationship
I	$Y = X_1 + 1.5X_2 - 2X_3$
II	$Y = 1.5X_1 \times X_2 + X_3$
III	$Y = f(X_1, X_2) + X_3$; $f(0,0) = 0$, $f(1,0) = 1$, $f(0,1) = 2$, $f(1,1) = -1$

The first study examined the case of (uncorrelated) binary predictors and a continuous response. Each scenario concentrated on a particular underlying relationship (Models I, II or III) for a particular frequency of the causal predictors (0.05, 0.1, 0.2, 0.4 or random). The first model was designed so that each causal predictor contributed additively, the second featured a multiplicative interaction, while the third involved a "weird" interaction (see Table 1)

Figure 3 presents results from the first study. Each plot relates to a different underlying relationship. Within each plot, the lines display the average number of causal predictors correctly identified by each method for different frequencies of the causal predictors. Figure 4 provides an alternative interpretation of these results, reporting how often each method successfully detected 0, 1, 2 or 3 causal predictors for each scenario.

Under Model I, *SSS*, *Logic*, *MARS* and *Sparse Partitioning* were the four best performing methods across different frequencies, pulling clear of *Single*, *Pairs* and *RF* as the causal predictor frequency increased. Under Model II, this order was essentially maintained, with the exception of *SSS*, which dropped into the second tier of performers. However, under Model III, *Sparse Partitioning* has emerged on top, comprehensively beating six of its rivals, with only *Pairs* coming close.

Sparse Partitioning has performed best in this simulation study, proving itself most robust across the different models. It has triumphed under Model III, when

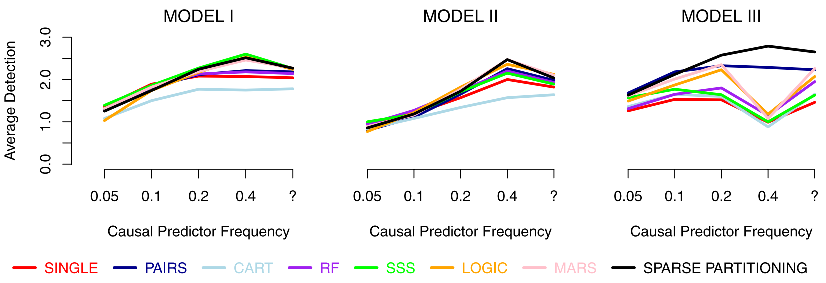

FIG. 3. *Each plot considers a different underlying relationship (described in the main text). Within each plot, the lines report the average number of causal predictors correctly detected by each method for different causal predictor frequencies ("?" denotes random).*

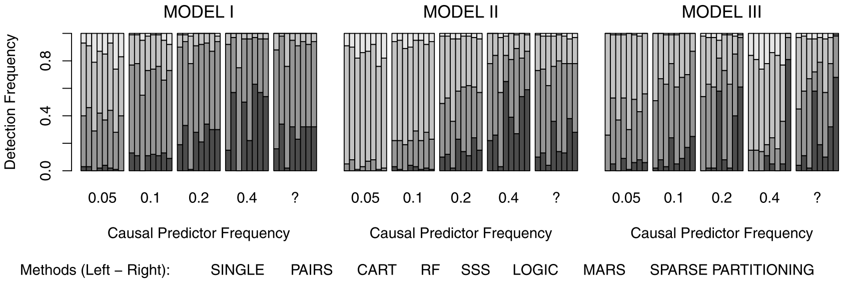

FIG. 4. *Each group of eight vertical bars compares the methods for a particular underlying relationship and causal predictor frequency* (*"?" denotes random*). *Within each bar, the lengths of the shaded sections* (*from top to bottom*) *indicate the proportion of time the method correctly detected* 0, 1, 2 *and* 3 *causal predictors. For example, the lengths of the darkest gray bars show how often each method successfully identified all three causal predictors. For all scenarios, the ordering of methods is the same* (*from left to right*): Single, Pairs, CART, RF, SSS, Logic, MARS *and* Sparse Partitioning.

the underlying relationship assumptions of all other methods have been violated. Note that its generality has not prevented it from at least matching the performance of the existing methods under Models I and II.

5. Application to real data sets. We applied *Sparse Partitioning* to four previously analyzed association studies: the first looked at expression data for 109 individuals in the HapMap project (http://hapmap.ncbi.nlm.nih.gov); the second and third examined data sets from the "2010 Project" (http://walnut.usc.edu/2010), a large-scale study of the plant *Arabidopsis thaliana*; the fourth used data provided by the Flint laboratory at the University of Oxford (http://www.well.ox.ac.uk/flint-2). Extended versions of all results are provided in Figures 12–16 of the Supplementary Material [Speed and Tavaré (2010)].

5.1. *HapMap data.* Dr. Antigone Dimas kindly provided us with a sample of 109 individuals, each typed for 1,186,075 Single Nucleotide Polymorphisms (SNPs) and measured for expression levels of 2682 genes [Dimas (2009)]. We applied *Sparse Partitioning* to the four genes for which Dr. Dimas found strongest evidence for an interaction, copying her decision to consider only SNPs within one million base pairs (1 Mbp) of each gene. Figure 5 presents the results for MTHFR, the third of these genes, located approximately 11.8 Mbp along Chromosome 1. For each SNP in the 2 Mbp region, the top plot displays the p-value obtained by *Single*, while the bottom plot reports the posterior probability of association from *Sparse Partitioning* (circles correspond to run one result, triangles to run two). The solid vertical line marks the location of the gene, while the two dashed vertical lines mark the locations of the SNPs declared interacting by Dr. Dimas. The dashed horizontal lines provide estimates of the 5, 25 and 50% significance thresholds for the top association of each method, calculated using permutation tests.

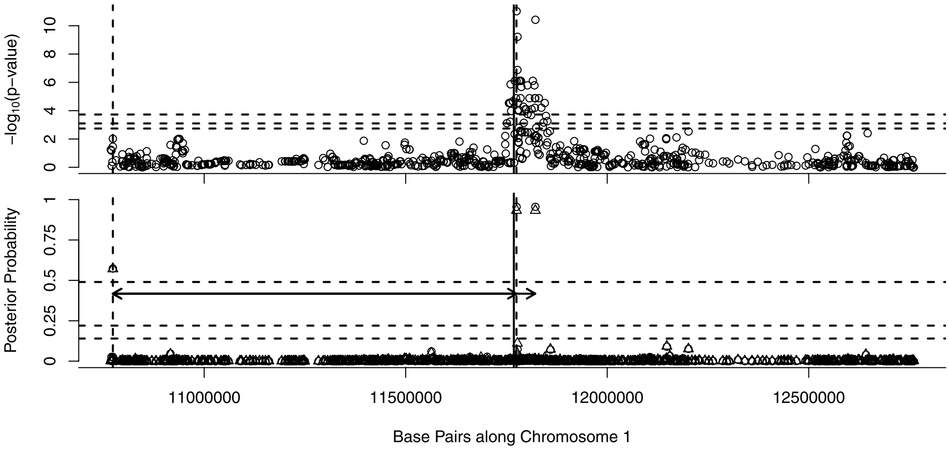

FIG. 5. *Analysis of expression levels of* MTHFR *using HapMap data. The top plot shows results from* Single, *the bottom plot shows results from two runs of* Sparse Partitioning. *Full details are provided in the main text.*

Sparse Partitioning found three promising SNPs, rs2286139, rs2643888 and rs2279703, with posterior probabilities of association 0.57, 0.96 and 0.96, respectively. It is no coincidence that the second and third hits have matching probabilities. Before analysis, *Sparse Partitioning* searches for highly correlated predictors, as is often the case with fine-scale genetic data. SNP rs2643888 was found to be highly correlated with SNP rs2279703, with matching values for 106 of the 109 individuals. Therefore, the former SNP was removed from analysis, and subsequently given the same posterior estimates as the latter. *Sparse Partitioning* returned a posterior probability of interaction of 0.42 between SNPs rs2286139 and rs2643888/rs2279703 (indicated by the horizontal arrows), offering some support for Dr. Dimas' findings of an interaction.

5.2. 2010 *project*: *Pilot data*. The project's pilot data set looked at 95 accessions, genotyped for 5419 SNPs and measured for ten phenotypic traits. We focused on the tenth phenotype, expression levels of the FRIGIDA gene. We decided to remove eight accessions whose genotypes were either almost identical to remaining accessions or were flagged as suspicious by principal component analysis. Using methods similar to the original analysis [Zhao et al. (2007)], we first adjusted the phenotype to correct for confounding due to population structure and relatedness of accessions. By contrast, we chose not to impute missing values, meaning approximately 10% of the genotypes were supplied to *Sparse Partitioning* as unobserved.

Figure 6 compares the p-values obtained from *Single* to the posterior probabilities of association of *Sparse Partitioning*. Our method identified just one strong association, coinciding with the third strongest hit of *Single* and suggesting that,

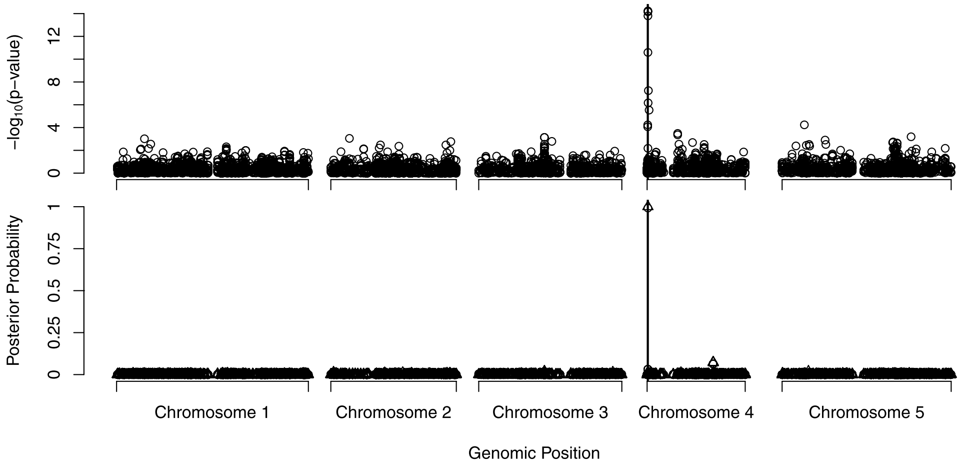

FIG. 6. *Analysis of expression levels of* FRIGIDA *for Arabidopsis thaliana. The top plot shows results from* Single, *the bottom plot shows results from two runs of* Sparse Partitioning. *Full details are provided in the main text.*

in this case, the simple underlying relationship of *Single* might be appropriate. For both methods the strong associations lay very close to the FRIGIDA region, marked by a solid vertical line, suggesting the results are accurate.

A possible concern is that *Sparse Partitioning*'s generality might lead to overfitting on occasions when simple models are more appropriate. Here that does not appear to be the case, with *Sparse Partitioning* declaring only one strong association. We repeated the analysis using imputed data, which allowed us to compare the prediction accuracy of each method via leave-one-out cross-validation. The linear model containing only the top hit from *Single* explained 44% of the variance, agreeing closely with *Sparse Partitioning*'s estimate of 42% variance explained.

5.3. 2010 *project*: *Release* 3.04. We examined how *Sparse Partitioning* would deal with a problem encountered in the 2010 project's most recent paper [Atwell et al. (2010)]. The expression level of the FLC gene is known to be affected by polymorphisms in the FRIGIDA region [Johanson et al. (2000); Shindo et al. (2005)]. Atwell et al. performed a one-SNP-at-a-time association study using FLC expression as the response. Its analysis produced results similar to our analysis by *Single*, shown in the top plot of Figure 7. While some SNPs within the FRIGIDA region (which is marked by a vertical line) achieved genome-wide significance, two stronger groups of associations were detected approximately 200 kbp and 1 Mbp to the right. Prior knowledge would suggest these downstream associations are spurious. When Atwell et al. repeated the analysis, but this time including in the regression model two alleles of the FRIGIDA gene known to affect FLC, the downstream associations vanished, increasing suspicion that they were false positives. For the rest of this section, we assume this to be the case.

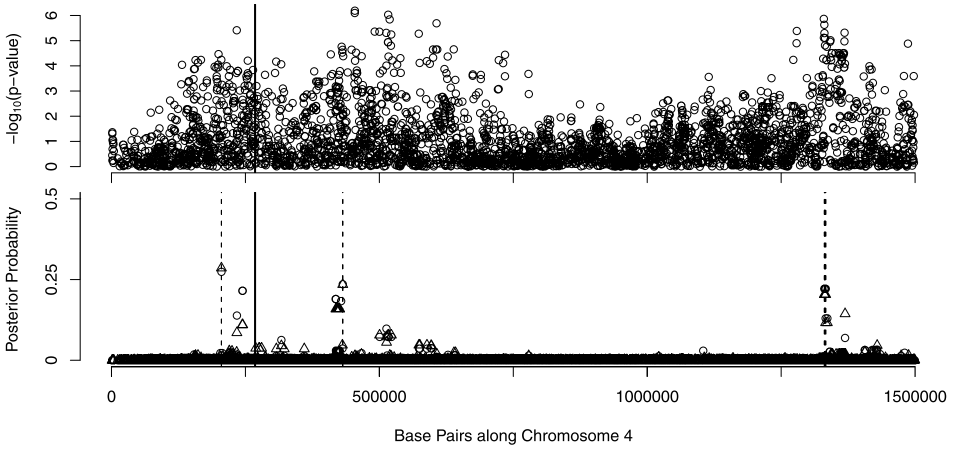

FIG. 7. *Analysis of expression levels of* FLC *for Arabidopsis thaliana. The top plot shows results from* Single, *the bottom plot, which has a truncated y-axis, shows results from two runs of* Sparse Partitioning. *Full details are provided in the main text.*

The project's latest data release provides typing for 214,553 SNPs across five chromosomes. We picked the 3509 SNPs located within the first 1.5 Mbp of Chromosome 4. As this subset was a biased selection (e.g., it contained over two-thirds of those SNPs with marginal p-values less than 10^{-4}), it was necessary to reflect this when choosing the prior probability of association for *Sparse Partitioning*. In the event, we settled upon a prior probability of 1 in 3500.

We used imputed data for this analysis, as the increased SNP density allowed missing values to be inferred more reliably. Similar to the analysis of Atwell et al., we decided to correct only for relatedness, as discussions with members of the Nordborg group convinced us that adjusting for population structure risked removing too much true signal. The bottom plot of Figure 7 shows the results of *Sparse Partitioning*. The dashed vertical lines indicate the three regions where our method found most evidence of association. While two false positives remained, *Sparse Partitioning* gave greatest recognition to the FRIGIDA region, identifying a possible association approximately 60 kbp upstream of the gene.

In this example, we had knowledge of the true causal region, allowing us to identify the false associations. The concern is that this example is one of many, and that most times we will not know the correct answer. In these cases, the best we can hope is that a method acknowledges the true and false positives, but recognizes the uncertainty. This is what *Sparse Partitioning* has done here. Furthermore, our method more precisely identified peaks than *Single* which should speed up the verification process.

5.4. *Mouse data.* Jon Krohn, from Professor Jonathan Flint's group at the University of Oxford, kindly provided us with CD4 counts for 1274 "heterogeneous stock" mice [Solberg et al. (2006)], along with genotypic values for

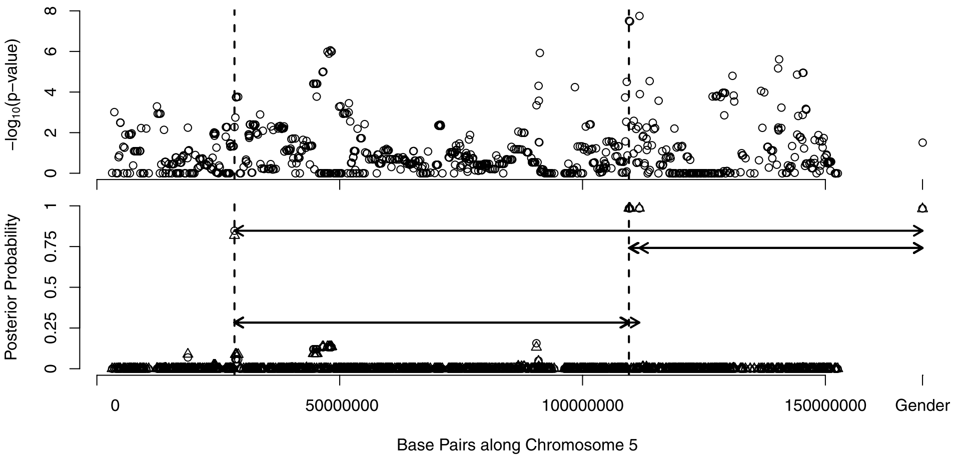

FIG. 8. *Analysis of CD4 count in mice. The top plot shows results from* Single, *the bottom plot shows results from two runs of* Sparse Partitioning. *Full details are provided in the main text.*

770 SNPs covering the length of Chromosome 5. Krohn had previously analyzed this data set using *Bagphenotype*, software designed by Dr. William Valdar (http://www.unc.edu/~wvaldar/bagphenotype.html). The response values were continuous, while the predictors were tertiary. Only a tiny proportion of genotypes (0.1%) were missing, so we saw no need to impute values and instead left them as unobserved. In addition, we were provided with the gender of each mouse, which we coded as a binary variable and included in the set of predictors.

As the chromosomal region was a subsection of a genome-wide study, we decided a prior probability of association of 1 in 10,000 was appropriate for each SNP. There is overwhelming prior knowledge that CD4 counts are linked to gender [e.g., Maini et al. (1996)], so we decided upon a prior probability of 0.5. We run *Sparse Partitioning* allowing three copies of each predictor ($C = 3$). As Figure 8 demonstrates, the top hits from *Single*, SNPs CEL-5_106584673 and rs13478460, which due to linkage disequilibrium are almost identical, persisted in *Sparse Partitioning*. In addition, our method declares associated SNP rs13478156. As indicated by the horizontal arrows, *Sparse Partitioning* found evidence of interactions between gender and the top SNPs. To test the effect of our prior choices, we repeated the analysis with prior probabilities $\{10^{-4}, 0.1\}$, $\{10^{-3}, 0.5\}$ and $\{10^{-3}, 0.1\}$, and obtained very similar results on each occasion (results not shown).

Maximum likelihood tests offer justification for why rs13478156 was found by *Sparse Partitioning*, but largely overlooked by *Single*. The supplementary material provides plots for two sets of tests. The first set compares, for each SNP, the pairwise interaction with gender against a null model of no association. We find that the top hits of *Single* remain the most significant hits here. However, these tests provide only limited information about the strength of the interaction terms. Therefore, the second set compares the pairwise interaction model against the additive model for that SNP and gender. We see that the interaction between rs13478156 and gender is highly significant. This supports *Sparse Partitioning*'s claim that this SNP acts in a gender specific way, which also agrees with findings from Krohn's analysis.

This data set demonstrated the advantage of allowing multiple copies of each predictor. When *Sparse Partitioning* is run without this option (results for $C = 1$ are shown in the supplementary material), the best fitting partition features three associated predictors in a single nonnull group. The posterior estimates of pairwise interactions cannot be trusted because the method is unable to distinguish between, say, a single three-way interaction and a pair of two-way interactions. Allowing multiple copies of predictors requires only a small increase in computation time, so we recommend this option is used.

6. Discussion.

It is fairly easy to design a regression method that is finely tuned for a specific underlying relationship and then demonstrate its superior power on data sets which obey this model. If one were presented only with the results of the Model III simulations, it would be easy to think that *Sparse Partitioning* is such a method. We have tried to show this is not case. We believe that *Sparse Partitioning* offers a robust alternative to existing methods. It fares equally well under simple models, but comes into its own as the model becomes more complex.

6.1. *Prospects for nonlinear regression.*

Nonlinear regression methods are competing over a fairly small share of the market, bounded on the one side by the performance of methods such as *Single* and on the other by the strength of signal present in the data. Despite these limitations, there remains a demand for such methods. There are many examples where standard linear methods fail to explain a satisfactory percent of the variation, so it is quite possible that nonadditive systems are at work. *Sparse Partitioning* should not be viewed as a search for interactions, but rather as a regression method which bears interactions in mind. Even for situations in which it cannot pinpoint an interaction with certainty, its detection power should benefit for having considered their existence.

6.2. *Generality.*

Sparse Partitioning's strength derives from the generality of its underlying relationship. Therefore, it is perhaps a surprise that the method does not appear to suffer in situations where this relationship is overly complicated. The results in Section 4 suggest there is no inherent disadvantage to using such a general underlying relationship. While *Sparse Partitioning* will almost certainly overfit the true model at some points in the MCMC sampling, its posterior estimates are based on model averages, rather than the single highest scoring model visited. For this reason, it should not matter if a nonassociated predictor is occasionally declared associated, as these errors are likely to be spread thinly across the noncausal predictors. Additionally, as *Sparse Partitioning* seeks only to estimate marginal posterior probabilities, using an underlying relationship too general should not upset the Bayesian mechanics. The prior probability that a predictor is associated remains constant (equal to p_g) regardless of the size of the model space. Even if excessive generality does affect some aspects of the posterior distribution, the marginal posterior probabilities should remain correct.

6.3. *Diagnosis*. The only way to calculate the posterior distribution exactly is through an exhaustive search of the space of partitions. Unfortunately, this is feasible for only the smallest data sets, so instead *Sparse Partitioning* is forced to explore the model space in a stepwise fashion. In this respect, *Sparse Partitioning*'s search holds an advantage over deterministic algorithms. When deciding which model in the neighborhood to visit next, *Sparse Partitioning* is not forced to move to the highest scoring model. Instead it is able to try a lower scoring move, in the hope that this is a gateway to a higher scoring region.

The drawback of this stochasticity is the variability it introduces into *Sparse Partitioning*'s results. The analysis in Section 5 provides some tips for gauging *Sparse Partitioning*'s performance. It is sensible to compare the results with those of *Single*, as we would expect very strong associations to be found by both methods. Repeating the analysis with a new random seed will highlight obvious lack of convergence, as should examination of trace plots. Additionally, if time permits, repeating the analysis with the response values permuted will provide significance thresholds under a model of no true associations.

6.4. *Limitations*. The processing time required for each iteration scales linearly with N. We speculate that the number of iterations required for convergence scales approximately with the 1.5th power of N (based on the stepwise nature of *Sparse Partitioning*'s search) and exponentially with the number of true associations (based on the growth of the model space).

As a rule of thumb, we consider *Sparse Partitioning* suitable for problems with no more than 20,000 predictors, or cases where $N/n < 100$. This is not to say that *Sparse Partitioning* cannot be applied on, say, a genome-wide scale, but it may be necessary to filter the predictors first. We suggest picking, for example, the highest 10% of hits from *Single*. Of course, this is not ideal. It is certainly possible that true associations are concealed within the remaining 90% of predictors. But considering standard practice involves picking, perhaps, the top 100 hits of *Single* for further analysis, the ability to consider instead the top few thousand hits should offer a significant advantage. As we experienced in Section 5, it is important to realize when we have selected a biased subset of predictors and reflect this in the prior probability of association. The easiest solution is to pick the priors as if analyzing the complete set of predictors.

We have identified situations in which *Sparse Partitioning* will struggle. Examples were found in Simulation Studies Five and Ten (see Supplementary Material). The latter was an almost unavoidable situation because the true relationship heavily contradicted our prior beliefs. The former demonstrated the drawback of treating tertiary predictors as categorical variables, when in fact their values have a natural ordering. We suspect that this problem can be overcome by application of a Bayesian version of Projection Pursuit [described in Hastie, Tibshirani and Friedman (2001)] that we are now developing.

Additionally, consider the case in which the response is influenced by an interaction of two predictors, but the inclusion of neither predictor on its own significantly improves the model fit. For one of these predictors to have a realistic chance of being included in a nonnull group, the improvement in fit must offset the penalty of inclusion implied by $\mathbb{P}(\mathbb{G})$. Because of the single-step nature of *Sparse Partitioning*'s search, it is unlikely that either predictor will appear in the current model, which is required for the method to consider their interaction. For our method to be successful in this case, it would have to permit two-step moves or resort to an exhaustive search.

Sparse Partitioning can be used when the predictors are continuous, provided a suitable transformation exists. For example, we have applied our method to copy number values, by first reducing each continuous measurement to one of three classes (neutral, increased or decreased). In the same way, we hope our method can be applied to a whole range of problems.

APPENDIX: DETAILS OF BAYESIAN FRAMEWORK

The regression model is written as $l(\mathbb{E}(\mathbf{Y})) = f(\mathbf{X})$, where l is a specified link function and $f(\mathbf{X})$ is the "underlying relationship." Without loss of generality, the underlying relationship can be expressed as the sum of functions of groups of associated predictors:

$$f(\mathbf{X}) = f_1(X_{\mathbf{G}_1}) + f_2(X_{\mathbf{G}_2}) + \cdots + f_K(X_{\mathbf{G}_K}).$$

The disjoint sets $\mathbf{G}_1, \mathbf{G}_2, \dots, \mathbf{G}_K$ index groups of associated predictors. Let $\mathbf{G}_0$, the "null group," index the predictors not associated. Therefore, $\mathbb{G} = \{\mathbf{G}_0, \mathbf{G}_1, \mathbf{G}_2, \dots, \mathbf{G}_K\}$ partitions $\{1, 2, \dots, N\}$. Equivalently, the partition can be described by the vector $\mathbf{I} = (I_1, I_2, \dots, I_N)$, where I_g indicates to which group the gth predictor belongs. Only unique partitions are considered, so the ordering of elements within groups is irrelevant, as is the ordering of nonnull groups.

A single model will be $\{\mathbb{G}, \mathbf{f}\}$, a partition and a corresponding set of functions $\{f_1, f_2, \dots, f_K\}$. The model space will be all such permissible pairs. If we wish to allow predictors to feature in more than one group of associations, the predictor set is expanded to contain C copies of each predictor. An alternative approach is to keep one copy of each predictor, but relax the condition on disjoint groups. However, we felt this approach created a greater amount of duplication within the space of underlying relationships, making it more challenging to define a prior. The description of the method supposes $C = 1$, with the alternative case discussed when necessary.

We are interested in the posterior distribution of $\mathbb{G}$ and $\mathbf{f}$, given the observed values for $\mathbf{X}$ and $\mathbf{Y}$. To be fully Bayesian, we must also consider the distribution of the predictors, which can be written as $\mathbb{P}(\mathbf{X}|\varepsilon)$, for some parameter vector ε:

$$\mathbb{P}(\mathbb{G}, \mathbf{f}, \varepsilon|\mathbf{X}, \mathbf{Y}) \propto \mathbb{P}(\mathbb{G}, \mathbf{f}, |\mathbf{X}, \mathbf{Y}) \times \mathbb{P}(\varepsilon|\mathbb{G}, \mathbf{f}, \mathbf{X}, \mathbf{Y}).$$

If we assume ε is unaffected by $\mathbb{G}$ and $\mathbf{f}$ [Gelman et al. (2004)], its posterior can be ignored in the calculation of $\mathbb{P}(\mathbb{G}, \mathbf{f}|\mathbf{X}, \mathbf{Y})$. Similarly, as we only wish to estimate properties of the posterior distribution of partitions, we treat the functions as nuisance parameters:

$$\begin{aligned}\mathbb{P}(\mathbb{G}|\mathbf{X}, \mathbf{Y}) &\propto \mathbb{P}(\mathbb{G}|\mathbf{X}) \times \mathbb{P}(\mathbf{Y}|\mathbf{X}, \mathbb{G})\\ &= \mathbb{P}(\mathbb{G}|\mathbf{X}) \times \int_{\mathbf{f}} \mathbb{P}(\mathbf{Y}|\mathbf{f}, \mathbf{X}, \mathbb{G})\mathbb{P}(\mathbf{f}|\mathbf{X}, \mathbb{G})\, d\mathbf{f},\end{aligned}$$

with $\mathbb{P}(\mathbb{G}|\mathbf{X})$ reducing to $\mathbb{P}(\mathbb{G})$, as we suppose the prior distribution of $\mathbb{G}$ does not depend on the observed values of the predictors.

A.1. Partition prior, $\mathbb{P}(\mathbb{G})$. The prior for the partition is constructed so the probability that predictor g is associated equals p_g. For partition $\mathbf{I}$, we can define the equivalence class $[\mathbf{I}]$ containing all partitions that declare the same predictors associated. To ensure the marginal probability that predictor g is associated equals p_g, we desire

$$\mathbb{P}([\mathbf{I}]) = \sum_{\mathbf{I}' \in [\mathbf{I}]} \mathbb{P}(\mathbf{I}') = \prod_{j:I_j=0} (1 - p_j) \prod_{j:I_j \neq 0} p_j = \prod_{j \in \mathbf{G}_0} (1 - p_j) \prod_{j \notin \mathbf{G}_0} p_j,$$

because then

$$\begin{aligned}\mathbb{P}(I_g \neq 0) &= \sum_{\mathbf{I}:I_g \neq 0} \mathbb{P}(\mathbf{I}) = \sum_{[\mathbf{I}]:I_g \neq 0} \left(\prod_{j:I_j=0} (1 - p_j) \prod_{j:I_j \neq 0} p_j \right)\\ &= p_g \prod_{j \neq g} [(1 - p_j) + p_j],\end{aligned}$$

equaling p_g, as required.

Assigning equal weighting to members of $[\mathbf{I}]$, we can calculate $\mathbb{P}(\mathbf{I})$ explicitly by counting the size of each equivalence class. If $\mathbf{I}$ declares $s = N - |\mathbf{G}_0|$ predictors associated, then the size of $[\mathbf{I}]$ will be the number of ways s elements can be partitioned. Unrestricted, this would equal the sth Bell number. Instead, *Sparse Partitioning* limits each partition to no more than K nonnull groups, each containing at most S elements. These "truncated" Bell numbers, $B(s, K, S)$, can be calculated in a recursive fashion. Let a_j denote the number of groups of size j for $j = 1, 2, \dots, S$. Then

$$\begin{aligned}&B(s, K, S|a_1, a_2, \dots, a_{S-1}, a_S)\\ &\qquad = B(s, K, S|a_1 - 1, a_2, \dots, a_{S-1}, a_S)\end{aligned}$$

$$+ B(s, K, S|a_1 + 1, a_2 - 1, \dots, a_{S-1}, a_S)(a_1 + 1)$$
$$+ \cdots + B(s, K, S|a_1, a_2, \dots, a_{S-1} + 1, a_S - 1)(a_{S-1} + 1),$$

with boundary condition

$$B(0, K, S|a_1, a_2, \dots, a_{S-1}, a_S) = \begin{cases} 1, & \text{if } a_j = 0\ \forall j, \\ 0, & \text{otherwise.} \end{cases}$$

Equally weighting each member of $[\mathbf{I}]$ places a high probability $(1 - |[\mathbf{I}]|^{-1})$ on the existence of interactions, even though few interactions have so far been found and verified. It would be straightforward to alter the partition weightings. For example, we could choose to favor partitions containing fewer interactions. However, the lack of known interactions must largely be due to how hard they are to identify, coupled with how rarely they are searched for. Therefore, we are satisfied that a uniform weighting is a reasonable choice.

Sparse Partitioning requires that K and S are set in advance, to allow sufficient memory to be allocated and pre-calculation of $B(s, K, S)$. Theoretically, K and S should be no smaller than N, to ensure the two most extreme underlying relationships are possible (either N groups of size one or one group of size N). In practice, these values would require vast amounts of unnecessary computation. Therefore, we suggest K and S are set to the smallest values possible, without impacting the direction of the MCMC chain.

The calculation of $\mathbb{P}(I_g \neq 0)$ assumes $K \times S \geq N$, as the last summation supposes all 2^N equivalence classes are achievable. When this condition does not hold, the error involved can be calculated for the case that all prior probabilities are equal:

$$\mathbb{P}(I_g \neq 0) = p_g \sum_{s=0}^{KS-1} \binom{N-1}{s} p_g^s (1 - p_g)^{N-1-s} \Big/ \sum_{s=0}^{KS} \binom{N}{s} p_g^s (1 - p_g)^{N-s}$$
$$= p_g \mathbb{P}(s \leq KS - 1 \mid s \sim \mathbb{B}(p_g, N-1)) / \mathbb{P}(s \leq KS \mid s \sim \mathbb{B}(p_g, N)).$$

Using a normal approximation for each binomially distributed variable, we obtain

$$\mathbb{P}(I_g \neq 0) = p_g \Phi\biggl(\frac{KS - 1/2 - (N-1)p_g}{\sqrt{p_g(1 - p_g)(N-1)}}\biggr) \Big/ \Phi\biggl(\frac{KS + 1/2 - Np_g}{\sqrt{p_g(1 - p_g)N}}\biggr),$$

where Φ is the cumulative probability function for a standard normal. For small p_g, the value of $\mathbb{P}(I_g \neq 0)$ is affected most by the prior mean, Np_g. We suggest setting $K = 4$ and $S = 4$. Entering these values into the equation above, we find that the actual prior probability of association used by *Sparse Partitioning* lies within 1% of the desired value, p_g, even when the prior mean is as high as 9.

When multiple copies of predictors are allowed ($C > 1$), the prior probability of association for each copy of predictor g is set to $1 - \sqrt[C]{(1 - p_g)}$. This ensures the probability that one or more copies of predictor g are associated remains equal to p_g. Allowing multiple copies of predictors creates an element of duplication within the space of partitions. For example, a partition in which two copies of a predictor feature in the same nonnull group effects the same underlying relationship as the partition obtained when one of these copies is removed. As a result, the prior weighting for this underlying relationship is increased. However, for small values of p_g this effect will be negligible. As with K and S, it is necessary to specify C in advance. Its value has minimal effect on computation time, so we recommend a conservative setting, such as $C = 3$.

A.2. Function prior, $\mathbb{P}(\mathbf{f}|\mathbb{G})$. To ensure identifiability of the functions, one value of $X_{\mathbf{G}_k}$ is considered the base value and its mapping is absorbed into the overall intercept (denoted by α_0). Therefore, f_k has degree of freedom one less than d_k, the number of unique values (nodes) of $X_{\mathbf{G}_k}$. Let $V_{k,1}, V_{k,2}, \dots, V_{k,d_k-1}$ be dummy binary variables that distinguish the remaining $d_k - 1$ nodes; these map to $\alpha_{k,1}, \alpha_{k,2}, \dots, \alpha_{k,d_k-1}$, respectively. The underlying relationship can be written in standard regression form:

$$\begin{aligned} f(\mathbf{X}) = \alpha_0 &+ (\alpha_{1,1} V_{1,1} + \cdots + \alpha_{1,d_1-1} V_{1,d_1-1}) \\ &+ \cdots + (\alpha_{K,1} V_{K,1} + \cdots + \alpha_{K,d_K-1} V_{K,d_K-1}). \end{aligned}$$

All the relevant information of the functions is contained in the vector $\boldsymbol{\alpha} = \{\alpha_0, \alpha_{1,1}, \dots, \alpha_{1,d_1-1}, \dots, \alpha_{K,1}, \dots, \alpha_{K,d_K-1}\}$, of size $D = 1 + \sum(d_k - 1)$. *Sparse Partitioning* assigns independent normal priors with mean 0 to each element of $\boldsymbol{\alpha}$. These can be viewed as a penalty on smoothness, but one which accepts that with categorical predictors there is no ordering to the nodes. This agrees with a belief in parsimony, which prefers simple functions to complicated ones.

In the continuous response case the variance of these normal priors is σ^2/r; in the binary response case the variance is $1/r$. In both cases, the choice of r controls the extent by which smoothness is applied. Typically we set r to 1.

A.3. Likelihood, $\mathbb{P}(\mathbf{Y}|\mathbf{f}, \mathbf{X}, \mathbb{G})$. When the response is continuous, the link function is the identity and the residuals are assumed to be independent draws from a normal distribution with mean zero and variance σ^2:

$$\begin{aligned} \mathbb{P}(\mathbf{Y}|\mathbf{f}, \mathbf{X}, \mathbb{G}) &= \int_{\sigma^2} \mathbb{P}(\mathbf{Y}|\sigma^2, \mathbf{f}, \mathbf{X}, \mathbb{G}) \mathbb{P}(\sigma^2)\, d\sigma^2 \\ &= \int_{\sigma^2} (2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{Y} - f(\mathbf{X}))^T (\mathbf{Y} - f(\mathbf{X}))\right\} \sigma^{-2}\, d\sigma^2. \end{aligned}$$

This integral incorporates a prior for σ^2 of the form σ^{-2}, which reflects a preference for smaller variances. It does not matter that this prior is improper as it is common to all models.

When the response is binary, a logit link function is used, $l(a) = \log(\frac{a}{1-a})$:

$$\mathbb{P}(\mathbf{Y}|\mathbf{f}, \mathbf{X}, \mathbb{G}) = \prod_i [l^{-1} f(X_{i.})]^{Y_i} [1 - l^{-1} f(X_{i.})]^{(1-Y_i)}.$$

A.4. Marginal likelihood, $\mathbb{P}(\mathbf{Y}|\mathbf{X}, \mathbb{G})$.

$$\begin{aligned}\mathbb{P}(\mathbf{Y}|\mathbf{X}, \mathbb{G}) &= \int_{\mathbf{f}} \mathbb{P}(\mathbf{Y}|\mathbf{f}, \mathbf{X}, \mathbb{G}) \mathbb{P}(\mathbf{f}|\mathbf{X}, \mathbb{G})\, d\mathbf{f} \\ &= \int_{\boldsymbol{\alpha}} \mathbb{P}(\mathbf{Y}|\boldsymbol{\alpha}, \mathbf{X}, \mathbb{G}) \mathbb{P}(\boldsymbol{\alpha}|\mathbf{X}, \mathbb{G})\, d\boldsymbol{\alpha}.\end{aligned}$$

With a continuous response, $\mathbb{P}(\mathbf{Y}|\mathbf{X}, \mathbb{G})$ can be calculated explicitly. When the response is binary, *Sparse Partitioning* uses a Laplace approximation. Let $W(\boldsymbol{\alpha}) = \mathbb{P}(\mathbf{Y}|\boldsymbol{\alpha}, \mathbf{X}, \mathbb{G}) \mathbb{P}(\boldsymbol{\alpha}|\mathbf{X}, \mathbb{G})$ and $w(\boldsymbol{\alpha}) = \log(W(\boldsymbol{\alpha}))$:

$$w(\boldsymbol{\alpha}) \approx w(\boldsymbol{\alpha}') + (\boldsymbol{\alpha} - \boldsymbol{\alpha}')^T \frac{dw(\boldsymbol{\alpha}')}{d\boldsymbol{\alpha}} + \frac{1}{2} (\boldsymbol{\alpha} - \boldsymbol{\alpha}')^T \frac{d^2 w(\boldsymbol{\alpha}')}{d\boldsymbol{\alpha}^2} (\boldsymbol{\alpha} - \boldsymbol{\alpha}').$$

If $\hat{\boldsymbol{\alpha}}$ is the maximum likelihood estimate of $w(\boldsymbol{\alpha})$, then

$$W(\boldsymbol{\alpha}) \approx W(\hat{\boldsymbol{\alpha}}) \exp\left\{-\frac{1}{2} (\boldsymbol{\alpha} - \hat{\boldsymbol{\alpha}})^T \left(-\frac{d^2 w(\hat{\boldsymbol{\alpha}})}{d\boldsymbol{\alpha}^2}\right) (\boldsymbol{\alpha} - \hat{\boldsymbol{\alpha}})\right\}.$$

Therefore,

$$\mathbb{P}(\mathbf{Y}|\mathbf{X}, \mathbb{G}) \approx \mathbb{P}(\mathbf{Y}|\hat{\boldsymbol{\alpha}}, \mathbf{X}, \mathbb{G}) \mathbb{P}(\hat{\boldsymbol{\alpha}}|\mathbf{X}, \mathbb{G}) (2\pi)^{D/2} \left| -\frac{d^2 w(\hat{\boldsymbol{\alpha}})}{d\boldsymbol{\alpha}^2} \right|^{-1/2}.$$

Alternatively, *Sparse Partitioning* allows the user to select a probit link function, in which case a latent variable representation of the likelihood can be used [Albert and Chib (1993)]. Essentially, each binary response is replaced by a continuous "pseudo response." The regression model is then treated as if it were linear, except the new response values are resampled once per iteration.

When there are two or more functions present, the marginal likelihood will be affected (very slightly) by which node is considered the base value for each function. For consistency, the node removed is chosen according to a defined rule (and is the zero vector of $X_{\mathbf{G}_k}$ if available). In addition, before analysis begins, continuous response values are transformed to have mean 0 and variance 1 to reduce variability caused by the choice of base value.

Software. *Sparse Partitioning* has been implemented and is available at http://www.compbio.group.cam.ac.uk/software.html.

Acknowledgments. The authors thank Antigone Dimas, Keyan Zhao, Bjarni Vilhjálmsson and Jon Krohn for providing data from their respective laboratories. We thank Alexandra Jauhiainen, Terry Speed, Michael Stein and an anonymous reviewer for helpful suggestions on earlier versions of this article, and we acknowledge the support of The University of Cambridge, Cancer Research UK and Hutchison Whampoa Limited.

SUPPLEMENTARY MATERIAL

Supplement: Extra material (DOI: 10.1214/10-AOAS411SUPP; .pdf). Provides additional details of *Sparse Partitioning*'s methodology, full explanation of the simulation studies and extended results from applying the method to real data sets.

REFERENCES

ALBERT, J. and CHIB, S. (1993). Bayesian analysis of binary and polychotomous response data. *J. Amer. Statist. Assoc.* **88** 669–679. MR1224394

ATWELL, S., HUANG, Y., VILHJÁLMSSON, B., WILLEMS, G., HORTON, M. and LI, Y. (2010). Genome-wide association study of 107 phenotypes in *Arabidopsis thaliana* inbred lines. *Nature* **465** 627–631.

BALDING, D. (2006). A tutorial on statistical methods for population association studies. *Nat. Rev. Genet.* **7** 781–791.

BREIMAN, L. (2004). Random Forests. *Machine Learning* **45** 5–32.

BREIMAN, L., FRIEDMAN, J., OLSHEN, R. and STONE, C. (1984). *Classification and Regression Trees*. Wadsworth, Belmont, CA. MR0726392

CORDELL, H. (2009). Detecting gene-gene interactions that underlie human diseases. *Nat. Rev. Genet.* **10** 392–404.

DIMAS, A. (2009). The role of regulatory variation in sculpting gene expression across human populations and cell types. Ph.D. thesis, Darwin College, Univ. Cambridge.

GELMAN, A., CARLIN, J., STERN, H. and RUBIN, D. (2004). *Bayesian Data Analysis*. Chapman and Hall/CRC, Boca Raton, FL. MR2027492

HANS, C., DOBRA, A. and WEST, M. (2007). Shotgun stochastic search for "large *p*" regression. *J. Amer. Statist. Assoc.* **102** 507–516. MR2370849

HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. (2001). *The Elements of Statistical Learning*. Springer, New York. MR1851606

HOGGART, C., WHITTAKER, J., DE IORIO, M. and BALDING, D. (2008). Simultaneous analysis of all SNPs in genome-wide and re-sequencing association studies. *PLoS Genet.* **4** e10000130.

JOHANSON, U., WEST, J., LISTER, C., MICHAELS, S., AMASINO, R. and DEAN, C. (2000). Molecular analysis of FRIGIDA, a major determinant of natural variation in *Arabidopsis* flowering time. *Science* **290** 344–347.

MAINI, M., GILSON, R., CHAVDA, N., GILL, S., FAKOYA, A., ROSS, E., PHILLIPS, A. and WELLER, I. (1996). Reference ranges and sources of variability of CD4 counts in HIV-seronegative women and men. *Genitourin. Med.* **72** 27–31.

MARCHINI, J., DONNELLY, P. and CARDON, L. (2005). Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nat. Genet.* **37** 413–417.

MCCARTHY, M., ABECASIS, G., CARDON, L., GOLDSTEIN, D., LITTLE, J., IOANNIDIS, J. and HIRSCHHORN, J. (2008). Genome-wide association studies for complex traits: Consensus, uncertainty and challenges. *Nat. Rev. Genet.* **10** 356–369.

RUCZINSKI, I., KOOPERBERG, C. and LEBLANC, M. (2003). Logic regression. *J. Comput. Graph. Stat.* **12** 475–511. MR2002632

SHINDO, C., ARANZANA, M., LISTER, C., BAXTER, C., NICHOLLS, C., NORDBORG, M. and DEAN, C. (2005). Role of FRIGIDA and FLOWERING LOCUS C in determining variation in flowering time of *Arabidopsis thaliana. Plant Physiol.* **138** 1163–1173.

SOLBERG, L., VALDAR, W., GAUGUIER, D., NUNEZ, G., TAYLOR, A., BURNETT, S., ARBOLEDAS-HITA, C., HERNANDEZ-PLIEGO, P., DAVIDSON, S., BURNS, P., BHATTACHARYA, S., HOUGH, T., HIGGS, D., KLENERMAN, P., COOKSON, W., ZHANG, Y., DEACON, R., RAWLINS, J., MOTT, R. and FLINT, J. (2006). A protocol for high-throughput phenotyping, suitable for quantitative trait analysis in mice. *Mamm. Genome* **17** 129–146.

SPEED, D. and TAVARÉ, S. (2010). Supplement to “Sparse Partitioning: Nonlinear regression with binary or tertiary predictors with application to association studies.” DOI: 10.1214/10-AOAS411SUPP.

STEPHENS, M. and BALDING, D. (2009). Bayesian statistical methods for genetic association studies. *Nat. Rev. Genet.* **10** 681–690.

STRANGER, B., FORREST, M., DUNNING, M., INGLE, C., BEAZLEY, C. and THORNE, N. (2007). Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science* **315** 848–853.

THE WELLCOME TRUST CASE CONTROL CONSORTIUM (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447** 661–678.

WANG, H., ZHANG, Y., LI, X., MASINDE, G., MOHAN, S., BAYLINK, D. and XU, S. (2005). Bayesian shrinkage estimation of quantitative trait loci parameters. *Genetics* **170** 465–480.

ZHANG, M., MONTOOTH, K., WELLS, M., CLARK, A. and ZHANG, D. (2005). Mapping multiple quantitative trait loci by Bayesian classification. *Genetics* **169** 2305–2318.

ZHAO, K., ARANZANA, M., KIM, S., LISTER, C., SHINDO, C., TANG, C., TOOMAJIAN, C., ZHENG, H., DEAN, C., MARJORAM, P. and NORDBORG, M. (2007). An *Arabidopsis* example of association mapping in structured samples. *PLoS Genet.* **3** e4.

DEPARTMENT OF APPLIED MATHS
AND THEORETICAL PHYSICS
CENTRE FOR MATHEMATICAL SCIENCES
UNIVERSITY OF CAMBRIDGE
WILBERFORCE ROAD
CAMBRIDGE CB3 OWA
UK
E-MAIL: doug.speed@ucl.ac.uk

DEPARTMENT OF ONCOLOGY
LI KA SHING CENTRE
UNIVERSITY OF CAMBRIDGE
ROBINSON WAY
CAMBRIDGE CB2 ORE
UK
E-MAIL: st321@cam.ac.uk

SUPPLEMENTARY MATERIAL

SPARSE PARTITIONING: NONLINEAR REGRESSION WITH BINARY OR TERTIARY PREDICTORS, WITH APPLICATION TO ASSOCIATION STUDIES

By Doug Speed and Simon Tavaré

Sections 1, 2 and 3 provide further details of *Sparse Partitioning*'s MCMC sampling, while Section 4 explains fully the simulation studies. Additionally, Figures 12, 13, 14, 15 and 16, located at the end, contain plots relating to the real data set examples.

1. Details of MCMC Sampling.

The aim of Markov Chain Monte Carlo sampling is to create a Markov chain whose stationary distribution matches the posterior distribution. To effect this, it is necessary to control the move probabilities. Metropolis-Hastings' theory [Hastings (1970)] allows us to propose a new model however we please, then provides us with a probability with which to accept this proposal. If $\mathbb{Q}(\mathbb{G} \to \mathbb{G}')$ is the probability of proposing a move from model $\mathbb{G}$ to $\mathbb{G}'$, then it should be accepted with probability

$$\min\left(1, \frac{\mathbb{P}(\mathbb{G}'|\boldsymbol{X},\boldsymbol{Y})}{\mathbb{P}(\mathbb{G}|\boldsymbol{X},\boldsymbol{Y})}\ \frac{\mathbb{Q}(\mathbb{G}' \to \mathbb{G})}{\mathbb{Q}(\mathbb{G} \to \mathbb{G}')}\right).$$

Sparse Partitioning uses two sampling stages. First, in a random order, it samples new values for each I_g. Second it resamples a component of $\mathbb{G}$. In both cases, because the proposal distributions match the conditional posterior distributions, the acceptance probabilities will always be one (Gibbs' Sampling).

1.1. *Sampling New Values of I_g.*

A new value for I_g is sampled from its conditional posterior distribution $\mathbb{P}(I_g|I_{-g},\boldsymbol{X},\boldsymbol{Y})$. This distribution is calculated explicitly by finding the posterior scores for all partitions that differ from the current partition only in the value of I_g. The order that these partitions are searched mimics the way the truncated Bell numbers are calculated. First predictor g is removed from the current partition. Then it is added, in turn, to the null group, each non-empty non-null group (if space) and then as a singleton non-null group (if space).

When $C > 1$, the number of samplings required increases by a factor of C. However, when no copies of a predictor are associated, the conditional posterior distribution will be the same for each copy of that predictor. As a result, for sparse problems increasing C will have minimal effect on computation time.

1.2. *Sampling a New Component of $\mathbb{G}$.*

One of the elements of a non-null group, $G_{k,j}$, is picked uniformly at random and its value resampled from $\mathbb{P}(G_{k,j}|\boldsymbol{G}_1,\ldots,\boldsymbol{G}_{k,\text{-}j},\ldots,\boldsymbol{G}_K,\boldsymbol{X},\boldsymbol{Y})$. This distribution is calculated by first removing predictor $G_{k,j}$ from the current model, then scoring the neighborhood of all partitions that differ only by their value of $G_{k,j}$.

The conditional posterior probability for $G_{k,j}$ is the same for each copy of a predictor. Therefore, as with the first sampling, increasing C has minimal effect on computation time.

It is possible to explore the entire space of partitions by repeatedly changing single elements of $\boldsymbol{I}$, so, at first glance the second sampling method is not required. However, suppose we wish to replace an associated predictor with one that is not associated. To achieve this requires changes to two values of $\boldsymbol{I}$ and so would take at least two steps if only using the first sampling method. This

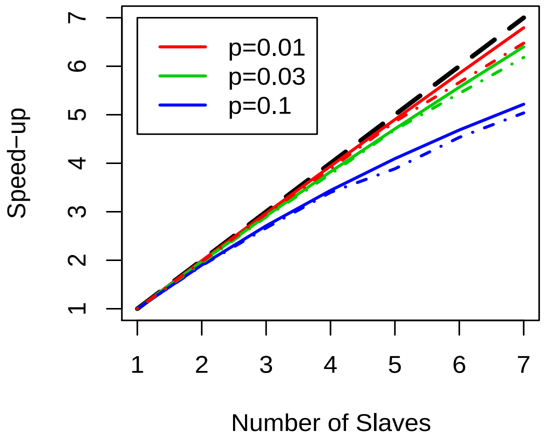

FIG 1. *During the resampling of $\boldsymbol{I}$, the master processor sends each servant a predictor, g, and instructs it to resample I_g. The expected number of valid samplings the master will receive after looping once through the servants depends on p, the likelihood that a sampling will report a changed value of I_g. The solid colored lines represent the theoretical speed-ups; the dashed colored lines some speed-ups observed in practice. For comparison, the black dashed line represents a perfect linear speed-up.*

will severely reduce the chance of such a move happening, especially if the move must pass through a low scoring model. The second sampling method corrects this, as the move can be achieved by changing just one element of $\mathbb{G}$.

1.3. *Obtaining Posterior Estimates.* After a predetermined burn-in period, *Sparse Partitioning* keeps count of how often each predictor features in a non-null group. This frequency provides the posterior estimate that the predictor is associated. *Sparse Partitioning* also keeps track of pairs of predictors that appear in the same non-null group. This information forms the posterior estimates that pairs of predictors interact. For diagnostic use, the posterior score of each partition in the Markov chain is recorded, as is the number of associations at each iteration.

In the case that two predictors are identical, one is removed before analysis and N adjusted accordingly. At the end of the analysis this predictor is given the same posterior score as its duplicate. Therefore, strictly speaking, this posterior score should be interpreted as the posterior probability that one of these duplicate predictors is associated. *Sparse Partitioning* can be instructed to apply a similar filtering to almost identical predictors. In the case of highly correlated predictors, this option provides a trade-off between performance and speed. The greater the filtering, the faster the method will converge, but the higher the chance that the true signal is overlooked.

1.4. *Prior Probability of Association, p_g.* In *Sparse Partitioning* the value of p_g is fixed throughout. An alternative approach is to treat p_g as a variable and resample its value at each iteration. This would provide a sampling from $\mathbb{P}(p_g|\boldsymbol{X},\boldsymbol{Y})$, which could instead be used as the posterior estimate that predictor g is associated. However, we see no advantage to this approach. The conditional posterior distribution of p_g is proportional to $\mathbb{P}(I_g|p_g) \times \mathbb{P}(p_g)$, yet is is hard to conceive a situation in which we have detailed prior information about p_g, above a belief in its mean. Regardless of this, the conjugate prior $\beta(1,b)$ is often used out of convenience [Zhang *et al.* (2005); Carvalho *et al.* (2008)]. Doing so produces results more difficult to interpret because the posterior mean will necessarily lie between $1/(2+b)$ and $2/(2+b)$.

2. Parallelization of MCMC Sampling. The bulk of *Sparse Partitioning*'s processing time is spent sampling from the posterior distribution. Therefore it is convenient that the two sampling methods can be parallelized. Consider a set-up with one master node and H servants. It is straightforward to parallelize the resampling of a component of $\mathbb{G}$. When scoring the neighborhood of partitions that differ in their value of $G_{k,j}$, simply assign each servant a portion.

Parallelization of the first sampling method utilizes the sparse nature of the problem. *Sparse Partitioning* instructs the $(h+1)$th servant processor to sample I_{g+h+1} using the current values of $I_{g+1}, I_{g+2}, \ldots, I_{g+h}$. The sampling of I_{g+h+1} is valid if the first h servants make no change to the current model. Only in the rare case that one of the preceding servants has altered the current model, must the system backtrack to the last correct sampling. During each iteration, most predictors remain unassociated, so this parallelization is very efficient. This is demonstrated in Figure 1.

3. Confounding and Missing Data. *Sparse Partitioning* is designed to cope with imperfect data, as is often the case with association studies. We consider two cases: when we wish to include additional confounding covariates and when some data are missing.

3.1. *Confounding.* When we wish to include additional factors in the regression model, the underlying relationship is transformed to $f(\boldsymbol{X}, \boldsymbol{\Theta}) + \boldsymbol{\Psi}\omega$, where the columns of $\boldsymbol{\Theta}$ and $\boldsymbol{\Psi}$ contain the extra covariates. Those contained in $\boldsymbol{\Theta}$ must be coded as binary or tertiary variables and are treated the same as standard predictors. The method will then consider interactions with these covariates, which is useful if, say, we wished to explore predictor-environment interactions. The user will probably wish to amend the prior probabilities of associations for these variates, as there will generally be a significantly stronger belief in their inclusions.

The covariates contained in $\boldsymbol{\Psi}$ are assumed not to interact with $\boldsymbol{X}$ or $\boldsymbol{\Theta}$. *Sparse Partitioning* includes them in the underlying relationship and allows for their effect each time it calculates a partition score. Normally it will make negligible difference to the results if the effect of these covariates is adjusted for pre-analysis (this will be the case either if these covariates are orthogonal to the main predictors or if the prior probabilities of association for the main predictors are much less than one). These covariates are not restricted to being tertiary, so in association studies might, for example, represent population structure or relatedness of samples.

3.2. *Missing Predictors.* The predictor set can be augmented as $\boldsymbol{X} = \{\boldsymbol{O}, \boldsymbol{U}\}$ where $\boldsymbol{O}$ and $\boldsymbol{U}$ represent the observed and unobserved predictors, respectively. $\boldsymbol{U}$ is treated as a variable during analysis, so the revised posterior distribution becomes

$$\mathbb{P}(\mathbb{G}, \boldsymbol{U}|\boldsymbol{Y}, \boldsymbol{O}) \propto \mathbb{P}(\boldsymbol{Y}|\mathbb{G}, \boldsymbol{U}, \boldsymbol{O}) \times \mathbb{P}(\boldsymbol{U}|\boldsymbol{O}) \times \mathbb{P}(\mathbb{G}).$$

Let $U_{g,j}$ be the jth missing value of the gth predictor. *Sparse Partitioning* uses a prior distribution of the form $\mathbb{P}(U_{g,j}|O_g)$, equal to the frequencies that 0, 1 and 2 appear in the observed values for the gth predictor. At each iteration, *Sparse Partitioning* resamples each missing value $U_{g,j}$ from its conditional posterior distribution $\mathbb{P}(U_{g,j}|U_{-g,j}, \mathbb{G}, \boldsymbol{Y}, \boldsymbol{O})$ using a Gibbs' Sampler. This sampling is very fast when predictor g is not associated, as $U_{g,j}$ will be sampled directly from its prior. Otherwise, it is necessary to calculate $\mathbb{P}(U_{g,j}{=}u|U_{-g,j}, \mathbb{G}, Y, \boldsymbol{O}) \propto \mathbb{P}(\boldsymbol{Y}|U_{g,j}{=}u, U_{-g,j}, \mathbb{G}, \boldsymbol{O}) \times \mathbb{P}(U_{g,j}{=}u|O_g)$ for $u = 0, 1, 2$.

Sparse Partitioning's method of sampling missing predictors assumes the predictors are uncorrelated. This is not normally the case for association study data, which typically display strong patterns of linkage disequilibrium (LD). For these situations, software packages exist for estimating missing genotypes based on the observed patterns of LD [Clark (1990); Stephens, Smith and Donnelly (2001)], which can be imputed in advance of analysis. When the variants are densely mapped, we recommend imputation, as it should make fuller use of the information available, as well as speed up convergence.

3.3. *Missing Responses.* When some response values are missing, the simplest solution is to omit all data for the affected samples. However, they can be included, if the user prefers. Each unobserved Y_i is sampled once per iteration using Metropolis-Hastings' theory. When the response is continuous, a new value y' is proposed from a standard normal distribution (with density function ϕ) and replaces the current value y with probability $\min(1, \rho)$, where

$$\rho = \frac{\mathbb{P}(\mathbb{G}, Y_i = y'|Y_{-i}, \boldsymbol{X})}{\mathbb{P}(\mathbb{G}, Y_i = y|Y_{-i}, \boldsymbol{X})} \; \frac{\phi(y)}{\phi(y')} = \frac{\mathbb{P}(Y_i = y', Y_{-i}|\mathbb{G}, \boldsymbol{X})}{\mathbb{P}(Y_i = y, Y_{-i}|\mathbb{G}, \boldsymbol{X})} \; \frac{\phi(y)}{\phi(y')}.$$

With a binary response, it is possible to obtain the conditional posterior explicitly by calculating $\mathbb{P}(Y_i{=}y|\mathbb{G}, Y_{-i}, \boldsymbol{X}) \propto \mathbb{P}(Y_i{=}y, Y_{-i}|\mathbb{G}, \boldsymbol{X})$ for $y = 0, 1$, from which a new value can be sampled directly (Gibbs' Sampling).

3.4. *Cross Validation.* Typically, cross-validation is performed post-analysis by first using a training data set to select a best model, then assessing this model's suitability on a test data set. Even ignoring time considerations, this is not possible with *Sparse Partitioning* because the method neither reports a single best partition, nor provides estimates of $\mathbb{P}(\boldsymbol{f}|\mathbb{G}, \boldsymbol{X}, \boldsymbol{Y})$. However, this problem can be overcome by treating response values as missing. Suppose $Y_{\boldsymbol{T}}$ corresponds to the response values in a test data set; the user should first set these to missing. At each iteration, *Sparse Partitioning* resamples these "missing responses" from their conditional posterior distributions. Once collected, these samples provides an estimate of $\mathbb{P}(Y_{\boldsymbol{T}}|\boldsymbol{X}, Y_{-\boldsymbol{T}})$.

Alternatively, *Sparse Partitioning* provides an estimate of leave-one-out cross-validation. At each iteration, the method calculates the expected value of each response, given the current partition. When the response is continuous, we obtain

$$\mathbb{P}(\boldsymbol{Y}|\sigma^2, \boldsymbol{X}, \mathbb{G}) \propto \int_{\boldsymbol{\alpha}} \exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{Y} - V\boldsymbol{\alpha})^T(\boldsymbol{Y} - V\boldsymbol{\alpha}) - \frac{r}{2\sigma^2}\boldsymbol{\alpha}^T\boldsymbol{\alpha}\right\},$$

where V is a design matrix, such that $f(\boldsymbol{X}) = V\boldsymbol{\alpha}$. Therefore

$$\begin{aligned}\mathbb{P}(\boldsymbol{Y}|\sigma^2, \boldsymbol{X}, \mathbb{G}) &\propto \exp\left\{-\frac{1}{2\sigma^2}\boldsymbol{Y}^T[I - V\,(V^TV + rI)^{-1}V^T]\,\boldsymbol{Y}\right\} \\ &= \exp\left\{-\frac{1}{2\sigma^2}\boldsymbol{Y}^T E\,\boldsymbol{Y}\right\},\end{aligned}$$

which has the form of a multivariate normal distribution. Although its variance depends on σ^2, its expectation does not. The same is true for the conditional posteriors $\mathbb{P}(Y_i|Y_{-i}, \sigma^2, \boldsymbol{X}, \mathbb{G})$, so we obtain

$$\mathbb{E}(Y_i|Y_{-i}, \boldsymbol{X}, \mathbb{G}) = -E_{i,i}^{-1}\sum_{j\neq i} E_{i,j}Y_j.$$

For a binary response, the expectation of Y_i is the probability it equals 1:

$$\mathbb{E}(Y_i|Y_{-i}, \boldsymbol{X}, \mathbb{G}) = \mathbb{P}(Y_i = 1|Y_{-i}, \boldsymbol{X}, \mathbb{G}) \propto \mathbb{P}(Y_i = 1, Y_{-i}|\boldsymbol{X}, \mathbb{G}).$$

Therefore

$$\mathbb{E}(Y_i|Y_{-i}, \boldsymbol{X}, \mathbb{G}) = \frac{\mathbb{P}(Y_i = 1, Y_{-i}|\boldsymbol{X}, \mathbb{G})}{\mathbb{P}(Y_i = 0, Y_{-i}|\boldsymbol{X}, \mathbb{G}) + \mathbb{P}(Y_i = 1, Y_{-i}|\boldsymbol{X}, \mathbb{G})}$$

Using these expectations, *Sparse Partitioning* is able to estimate $\mathbb{E}(Y_i|Y_{-i}, \boldsymbol{X})$ by Monte Carlo integration.

4. Further Details of the Simulation Studies. The main paper presents results from the first simulation study. This study examined the performance of *Sparse Partitioning* for the case of 100 individuals each typed for 1000 uncorrelated, binary predictors. A continuous response was generated from three predictors, according to a particular underlying relationship, with the introduction of normally distributed noise. In total, nine further studies were performed, using the first study as a template. The table below highlights how each subsequent study altered one aspect of the set-up:

Study Two	Causal predictors unobserved.
Study Three	10% of predictor values missing.
Study Four	Non-normal noise.
Study Five	Tertiary predictors, 2 causal loci.
Study Six	Binary response, 3 or 4 causal loci.
Study Seven	Correlated predictors, 4 causal loci.
Study Eight	Examine effect of prior choice.
Study Nine	Examine effect of number of iterations.
Study Ten	Non-disjoint underlying relationship, 3 or 4 causal loci.

Within each study, we compared methods under a range of different scenarios. Each scenario focused on a particular underlying relationship (chosen from Models I to XV) and a particular frequency for the causal predictors (0.05, 0.1, 0.2, 0.4 or random). In general, each study tested an underlying relationship which was additive, one containing a "simple interaction" and one containing a "full interaction."

In most cases, we assessed a method's performance by asking it to report the three predictors for which it found most evidence of association, then counting how many causal predictors this list contained. For each scenario, we obtained empirical estimates of detection accuracy by repeating this procedure for 100 data sets.

4.1. *Existing Methods and Settings.*

In order to offer a fair assessment of *Sparse Partitioning*'s performance, we tried to identify existing methods that could be considered rivals. Generally, a regression method dealing with a continuous response can be adapted to handle a binary value by adding in a suitable link function. However, the converse is not true. Therefore some methods of this type were omitted from comparison [Hahn, Ritchie and Moore (2002); Verzilli, Stallard and Whittaker (2006); Mailund, Besenbacher and Schierup (2006); Zhang and Liu (2007); Mukherjee *et al.* (2009)].

Single performs a maximum likelihood test for each predictor, comparing the null model, $f(X_g) = \alpha$, with the alternative, $f(X_g) = \beta_{X_g}$. In the case of a continuous response, the residual sums of squares, TSS_0 and TSS_1, are calculated for each model. The test statistic $n \log(TSS_0/TSS_1)$ is compared to a χ^2 distribution with degrees of freedom either 1 (binary predictors) or 2 (tertiary predictors). The method returns a p-value for each predictor, which represents the probability of observing under the null model a test statistic at least as extreme as that seen.

Pairs is an extension of *Single*, using the alternative model $f(X_g, X_{g'}) = \beta_{X_g, X_{g'}}$ with degrees of freedom equal to the number of unique values of $(X_g, X_{g'})$. The final score for each predictor corresponds to the lowest p-value obtained across the N tests involving that predictor.

CART is available in R [R Development Core Team (2008)] via the package `tree`. The function *tree* returns a model with an unrestricted number of associations. We reduced this to the required size by *prune*.

RF is implemented in the R package `randomForest`. The function *randomForest* returns importance weightings for each predictor, from which the top associations can be selected.

SSS is provided at the authors' website. For the studies, the prior belief in the number of associations present, *priormeanp*, was set to 5, while the number of iterations, *iters*, was set to 100. *nbest*=1000 determined that the posterior estimates were based on the 1000 top scoring models.

Logic was run using the R package `LogicReg`. The function *logreg*, with parameters *select*=2 and *nleaves*=s, returns a logic tree of size s. The package offers the ability to run using a MCMC based method [Kooperberg and Ruczinski (2005)] by setting *select*=7. However, the size of the simulated data sets proved too large for R to handle, so this method could not be used.

MARS was run using the R package `mda`. The function *mars*, with parameters *nk*=2*s*+1 and *degree*=3, returns the best model with at most s predictors and allowing for at most three-way interactions.

Sparse Partitioning was run for 200 iterations, which typically takes less than one minute. K, S, C and p_g were set to 4, 4, 3 and $5/N$, respectively.

4.1.1. *Generating Data sets.* The following text explains the construction of each data set in Study One, using as an example the underlying relationship of Model I. For Models II and III, $\boldsymbol{f}$ was changed accordingly.

Let $\boldsymbol{Z} = (Z_1, Z_2, Z_3)$ be the vector of the three causal predictors for a particular individual. Therefore $f(\boldsymbol{X}) = f(\boldsymbol{Z}) = Z_1 + 1.5Z_2 - 2Z_3$. Let $\boldsymbol{F}$ be the set of unique possible values of $f(\boldsymbol{Z})$ and F a variable drawn at random from this set. Each response value was generated as $F + \mathbb{N}(0, \sigma^2)$. This technique ensured they were sampled evenly across their full range of values, rather than according to their prevalence. From these response values, $\boldsymbol{Z}$ was sampled from its posterior distribution:

$$\mathbb{P}(\boldsymbol{Z}|\boldsymbol{Y}) \propto \mathbb{P}(\boldsymbol{Z}) \times \mathbb{P}(\boldsymbol{Y}|\boldsymbol{Z}) \propto m_1^{Z_1} m_2^{Z_2} m_3^{Z_3} \times \mathbb{N}(\boldsymbol{Y}|f(\boldsymbol{Z}), \sigma^2),$$

where m_1, m_2 and m_3 were the causal allele frequencies under consideration. σ^2 was picked to give realistic levels of heritability (ranging between about 1/4 and 1/2). For Models I, II and III, it was set to 1.7, 0.85 and 1.1, respectively.

In the final data set, $X_{200} = Z_1$, $X_{500} = Z_2$ and $X_{800} = Z_3$. The remaining 997 predictors were simulated from Bernoulli $(100, \rho)$ with $\rho \sim \mathrm{U}(0.05, 0.95)$.

Code for generating the data sets for each study is provided at the method's homepage, located at `http://www.compbio.group.cam.ac.uk/software.html`.

4.2. *Study One: Additional Results.* The top two plots of Figure 2 are enlarged version of those in the main paper. The bottom plot shows the effect of changing some of *Sparse Partitioning*'s input parameters. When $S = 1$ (*SP Additive*), the maximum group size is limited to one, so the method considers only additive models. When $K = 1$ (*SP Interaction*), only one tree is permitted, forcing the full interaction model to be fitted at each step.

As expected, the performance of *SP Additive* is nearly identical to that of *SSS*, and their lines almost exactly coincide for Model III. Also to be expected, the performance of *Sparse Partitioning* is damaged when K is set to one, as then the method will necessarily overfit the true model. However, it is worth noting that *SP Interaction* is the second best performing method for Model III. This supports our belief that using an underlying relationship too general is less of a penalty than using one too restrictive.

Additionally, the dashed lines in the bottom plot mark the average of the highest posterior probability of interaction for the standard version of *Sparse Partitioning.* This provides an insight into *Sparse Partitioning*'s mechanics. For Model I, this line is very flat and close to zero, as desired when the true underlying relationship contains no interactions. For Model II, the line mirrors the detection accuracy; the point at which *Sparse Partitioning* begins to detect the interaction is the point that it begins to successfully detect all three predictors. The same effect is seen for Model III, except now the signal from the interaction is stronger, so it is detected sooner.

4.3. *Study Two: Causal Predictors Unobserved.* Due to high correlations between genetic variants, it is often possible in an association study to infer the location of causal predictors even if they have not been typed directly. For this reason it is permissible, and usually more efficient, to type just a subset of variants. This study considered the case in which the causal predictors are not observed, but are highly correlated with observed predictors.

For these scenarios, Z_1, Z_2 and Z_3 were considered unobserved predictors and replaced in the final data set by $\boldsymbol{Z}' = (Z_1', Z_2', Z_3')$. For each causal predictor, Z_j' was initialized to Z_j and then its genotypic values randomly toggled until the linkage disequilibrium (LD) between Z_j and Z_j' (measured in terms of correlation squared) dropped below the desired level (either $r^2 = 0.9$ or 0.8).

The results for each level of correlation are shown in Figure 3. The shapes of the plots closely match those of Study One, albeit, as expected, with lower average detection accuracy. The noticeable exception is method *Pairs*, whose performance has caught up with that of *Sparse Partitioning*. When there are just two causal predictors, *Pairs* approaches the gold standard method, as it performs an exhaustive search of all two predictor models. Reducing the correlation between the causal and observed predictors has the effect of increasing the noise in the model. This might explain why the gap between *Pairs* and *Sparse Partitioning* has decreased. Once the noise increases to the extent that one causal predictor becomes "unfindable," we would expect *Pairs* to perform best. Bearing this in mind, it is reassuring that *Sparse Partitioning* is at no point overtaken.

4.4. *Study Three: 10% of Predictor Values Missing.* As documented in Section 3, *Sparse Partitioning* is designed to cope with missing predictor or response values. In this study, after simulating each data set, 10% of the predictor values were selected at random and recorded as missing. The detection accuracy was compared with *Single*, *Pairs* and *SSS*, the three existing methods able to accept missing values. The results, shown in Figure 4, closely mirror the corresponding plots for Study One.

4.5. *Study Four: Non-normal Noise.* For a continuous response, *Sparse Partitioning* calculates a likelihood under the assumption of normal residuals. Therefore we tested the impact when this assumption is violated. In this study we simulated data sets using first exponential, then uniform, noise. In both cases the distribution parameters were set to produce heritabilities similar to those in Study One.

Figure 5 displays the results. The introduction of exponential noise, shown in the top plot, does not have a marked effect on the results; the plots still appear to closely resemble those of Study One. This is not the case with uniform noise, where each model, and in fact each method, has responded differently to its introduction. Nonetheless, with the exception of the low frequency end of Model II data sets, *Sparse Partitioning* has maintained its lead, and has actually dramatically improved under Model III.

4.6. *Study Five: Tertiary Predictors.* Two causal tertiary predictors were created by summing two pairs of causal binary predictors, generated using the methodology of Study One. Three models were used, both additive in the two causal predictors.

Model	Underlying Relationship
IV	$Y = I_{X_1>0} + I_{X_2>1}$
V	$Y = f_1(X_1) + f_2(X_2)$ with each $f_k(0)$, $f_k(1)$ and $f_k(2)$ chosen at random
VI	$Y = f_1(X_1) + f_2(X_2)$ with f_k additive $(f_k(0) + f_k(2) = 2f_k(1))$

The results for each model are shown in Figure 6. For Models IV and V, *Sparse Partitioning* has performed best. *CART* has performed very poorly for Model IV. Its underlying relationship is unable to consider additive contributions, which forces the method to pick between the two causal predictors. Although *RF* encounters the same problem when fitting individual trees, its stochastic nature allows both predictors to feature highly, provided both are chosen sufficiently often.

It is typical for association studies to prefer models which are linear in the number of alleles present. This is a fair assumption if the effects of chromosomes within homologous pairs are independent, but less so if these pairs act in partnership. Model VI considers two linear models. It is no surprise that *SSS* performs best, as its underlying relationship matches the model. *Sparse Partitioning* has performed less well. The contrast between its performance for Models V and VI would suggest this is a consequence of an underlying relationship too general. It is confusing why this effect is apparent for Model VI, when it has not been a problem elsewhere. Nonetheless, we are considering introducing an option which allows the user to "encourage" additivity. Alternatively, we are finalizing a related method, a Bayesian version of Projection Pursuit, which views predictors as quantitative, so should be able to overcome this problem.

4.7. *Study Six: Binary Response.* A binary response generally contains less information than its continuous counterpart. Therefore, to maintain reasonable power for non-trivial models, we reduced the number of predictors and increased the number of samples. We extended the study of [Mukherjee *et al.* (2009)], which generated data sets of 100 predictors and 200 individuals for each of three different models. For each underlying relationships, a Boolean function was used to determine $\mathbb{P}(Y_i = 1)$; if the function evaluated true, $\mathbb{P}(Y_i = 1)$ was set to 0.9, if false, $\mathbb{P}(Y_i = 1)$ was set to 0.1. This corresponds to setting $f(\boldsymbol{X})$ to 2.2 or -2.2 when using a logit link function.

Model	Underlying Relationship
VII	$\mathbb{E}(Y) = 0.1 + 0.8 \times X_1 \wedge (X_2 \Leftrightarrow X_3)$
VIII	$\mathbb{E}(Y) = 0.1 + 0.8 \times (X_1 \wedge X_2^C) \oplus X_3$
IX	$\mathbb{E}(Y) = 0.1 + 0.8 \times (X_1 \wedge X_2) \oplus (X_3 \wedge X_4^C)$

The results for each model are shown in Figure 7. *Sparse Partitioning* and *MARS* are the best performing methods in this study, sharing the top two places across the three models. At the moment *Sparse Partitioning* can not be applied to a response with more than two categories, so this might be a useful extension to develop.

4.8. *Study Seven: Correlated Predictors.* To generate data sets displaying realistic patterns of linkage disequilibrium (LD), we used the program `ms` [Hudson (2002)], provided on its author's website. The command `ms 1000 1 -s 20 -r 10 20 -F 100` simulates 1000 individuals typed wildtype or mutant for twenty Single Nucleotide Polymorphisms (SNPs). We concatenated the results of 100 runs, with every second SNP removed, to obtain a data set of 1000 individuals typed for 1000 SNPs. To give an indication of the levels of LD this generated, if we filtered this data set so that no pair of predictors remains with a squared correlation greater than 0.8, this would remove approximately half the predictors. Again, three different underlying relationships were used, similar in nature to those in Study One, except this time there were four causal predictors.

Model	Underlying Relationship
X	$Y = aX_1 + bX_2 + cX_3 + dX_4$
XI	$Y = f_1(X_1, X_2) + f_2(X_3, X_4)$ where f_1 maps to $\{0, a, a, b\}$ and f_2 to $\{0, 0, 0, c\}$
XII	$Y = f_1(X_1, X_2) + f_2(X_3, X_4)$ where f_1 maps to $\{0, a, b, c\}$ and f_2 to $\{0, d, e, f\}$

The coefficients in each model were generated randomly for each data set. Unlike the other studies, this one used a prospective method of sampling. We created the predictors first, then used four of these to generate the response values. From the 1000 response values, we picked 100 representing a sufficiently broad spectrum. If this was not possible, we picked four new causal predictors and tried again. By generating the data sets in this fashion, it was not possible to fix the causal predictor frequency and likely that the method produced a bias toward frequencies closer to 0.5.

We applied two scoring systems. The first, identical to that used in the other studies, made no allowance for correlations. This could be considered overly harsh. Suppose a method identifies as associated a predictor near to, and so in high correlation with, a causal predictor. Strictly speaking, this would be a false positive and score zero, even though its detection would still be a helpful indicator of the region likely to contain the true association. Therefore the second system scored each block of ten predictors. For the method *Single*, which considers one predictor at a time, so makes no allowance for LD, blocks were scored according to their best scoring predictor. For *Pairs*, *SSS*, *RF* and *Sparse Partitioning*, blocks were scored by summing over their ten predictors. *CART*, *Logic* and *MARS* only return precise models, rather than weightings for each predictor, so could not be scored under this system.

Figure 8 provides the results for this study. As is necessarily the case, the detection accuracy has improved using the second scoring system, and overall *Single* and *RF* benefited most from this change. In most cases, however, the ranking of the five methods scored for both systems has been preserved. *Sparse Partitioning* has performed admirably, coming top in five cases, beaten only by *Pairs* in the sixth.

4.9. *Study Eight: Examine Effect of Prior Choice.* The most important input setting for *Sparse Partitioning* is p_g, the prior probability of association for each predictor. The other variable parameters, such as maximum number and size of groups, or variances of prior distributions, can almost always be left at their defaults. In this study, we investigated the effect of different values for p_g, as opposed to keeping it at 0.005, the *status quo* for other studies.

Figure 9 presents the results for four choices, $p_g = 0.0005$, 0.001, 0.002 and 0.005. For the first two models, the difference is slight, but as expected the latter two choices, which are closest to the true case, perform best. The difference is more noticeable for higher causal predictor frequencies under Model III. The results suggest it is advisable to verge on the cautious side when setting p_g, which agrees with the general message that less restrictive is better.

4.10. *Study Nine: Examine Effect of Number of Iterations.* Naturally, the more iterations that can be afforded, the better. In general we picked 200 iterations, as this allowed data sets to be analyzed by *Sparse Partitioning* in under a minute. To compare this number to other methods, which often sample for upward of 100,000 iterations, it is worth remembering that *Sparse Partitioning* performs approximately $2N$ samplings per iteration.

Figure 10 shows the results of varying the number of iterations from 100 to 800. As with many of the studies, the differences show up more as the models become more complicated. We can see that greater performance could have been obtained through use of more iterations and possibly this would have overcome some of the limitations of *Sparse Partitioning* found in other studies.

4.11. *Study Ten: Non-Disjoint Underlying Relationship.* It is conceivable that a predictor features more than once in the underlying relationship. This might, for example, correspond to a genetic variant involved in two or more pathways. This study considered three models where this is the case.

Model	Underlying Relationship
XIII	$Y = X_1 \times X_2 + X_2 \times X_3$
XIV	$Y = f_1(X_1, X_2) + X_2 \wedge X_3 + 2X_4$ where f_1 maps to $\{0, 1, 2, -1\}$
XV	$Y = f_1(X_1, X_2) + 2X_2 \oplus X_3$ where f_1 maps to $\{0, 1, 2, -1\}$

Although *Sparse Partitioning* has generally performed best, this study has uncovered scenarios where *Sparse Partitioning* is significantly beaten. In Model XIII, the ability to recover the true underlying relationship is highly dependent on how often the causal predictors' values match. The individuals for which either one or two of the causal predictors equal 1 will provide most information. When the causal predictor frequencies are rare, the correlation between their values increases. When this happens, each causal predictor will well predict the underlying relationship. Therefore the method *Single*, which considers predictors independently, will have a large advantage, as all three predictors will score highly. Conversely, *Sparse Partitioning* will perform badly, as the improvement in fit of deducing the true relationship is not sufficient to offset its prior belief that fewer predictors are associated. Hopefully, therefore, our prior belief in the rarity of such a relationship is correct.

References.

Carvalho, C., Chang, J., Lucas, J., Nevins, J., Wang, Q. and West, M. (2008). High-dimensional sparse factor modeling: applications in gene expression genomics. *J. Amer. Statist. Assoc.* **103** 1438-1456.

Clark, A. (1990). Inference of haplotypes from PCR-amplified samples of diploid populations. *Mol. Biol. Evol.* **7** 111-122.

Hahn, L., Ritchie, M. and Moore, J. (2002). Multifactor dimensionality reduction software for detecting gene-gene and gene-environment interactions. *Bioinformatics* **19** 376-382.

Hastings, W. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57** 97-109.

Hudson, R. (2002). Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* **18** 337-338.

Kooperberg, C. and Ruczinski, I. (2005). Identifying interacting SNPs using Monte Carlo logic regression. *Genet. Epidemiol.* **28** 157-170.

Mailund, T., Besenbacher, S. and Schierup, M. (2006). Whole genome association mapping by incompatibilities and local perfect phylogenies. *Bioinfomatics* **7** 454.

Mukherjee, S., Pelech, S., Neve, R., Kuo, W., Ziyad, S., Spellman, P., Gray, J. and Speed, T. (2009). Sparse combinatorial inference with an application in cancer biology. *Bioinformatics* **25** 265-271.

R Development Core Team, (2008). R: A language and environment for statistical computing ISBN 3-900051-07-0.

Stephens, M., Smith, N. and Donnelly, P. (2001). A new statistical method for haplotype reconstruction from population data. *Am. J. Hum. Genet.* **68** 978-989.

Verzilli, J., Stallard, N. and Whittaker, J. (2006). Bayesian graphical models for genomewide association studies. *Am. J. Hum. Genet.* **79** 100-112.

Zhang, Y. and Liu, J. (2007). Bayesian inference of epistatic interactions in case-control studies. *Nat. Genet.* **39** 1167-1173.

Zhang, M., Montooth, K., Wells, M., Clark, A. and Zhang, D. (2005). Mapping multiple quantitative trait loci by Bayesian classification. *Genetics* **169** 2305-2318.

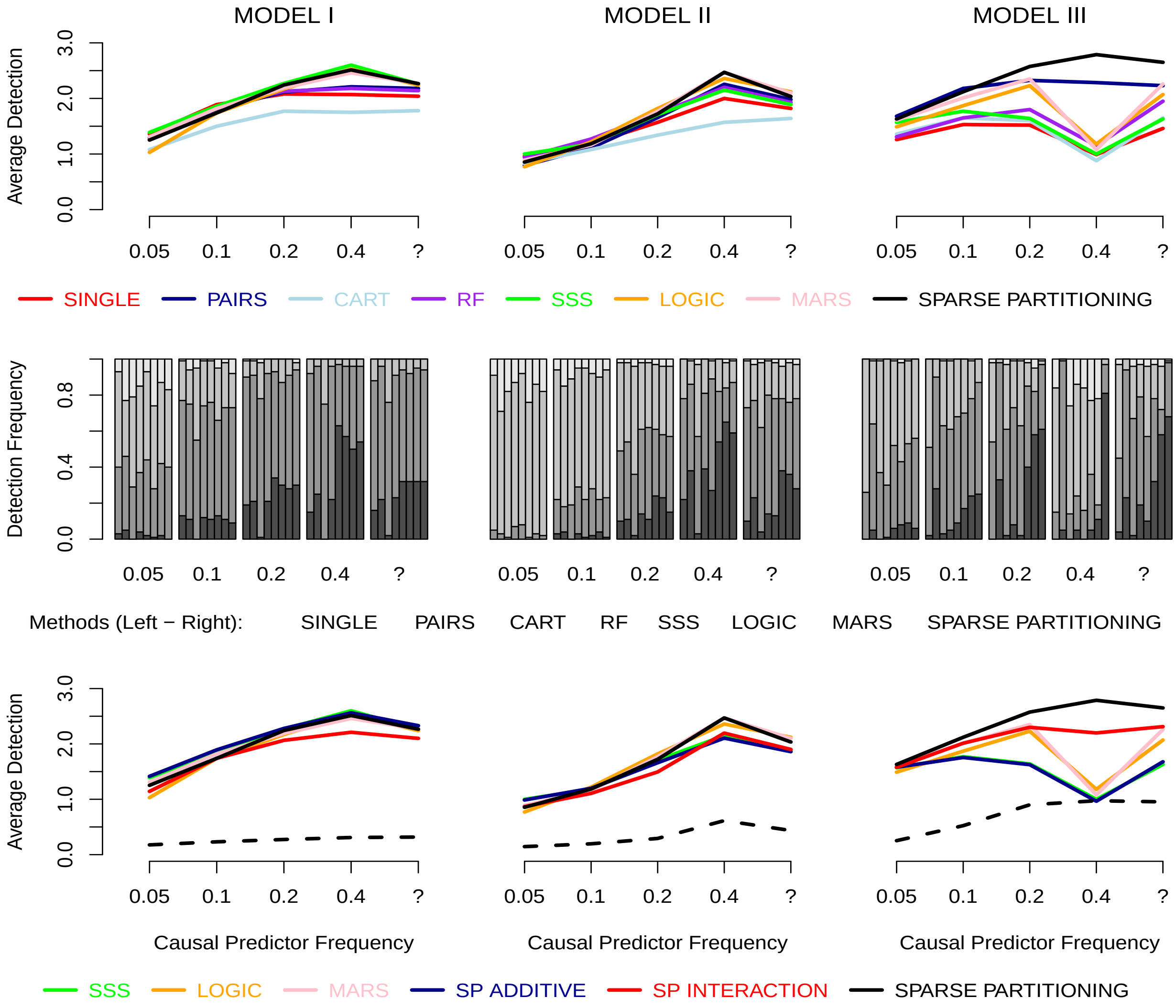

Fig 2. *Results of Study One. The top two plots provide enlarged versions of those in the main paper. The bottom plot shows the effect of restricting* Sparse Partitioning*'s underlying relationship by varying S and K.* SP Additive *($S = 1$) is only allowed to consider additive models, while* SP Interaction *($K = 1$) insists on the full interaction model at each step. Additionally, the dashed lines mark the average posterior probability of the top pairwise interaction for the standard version of* Sparse Partitioning.

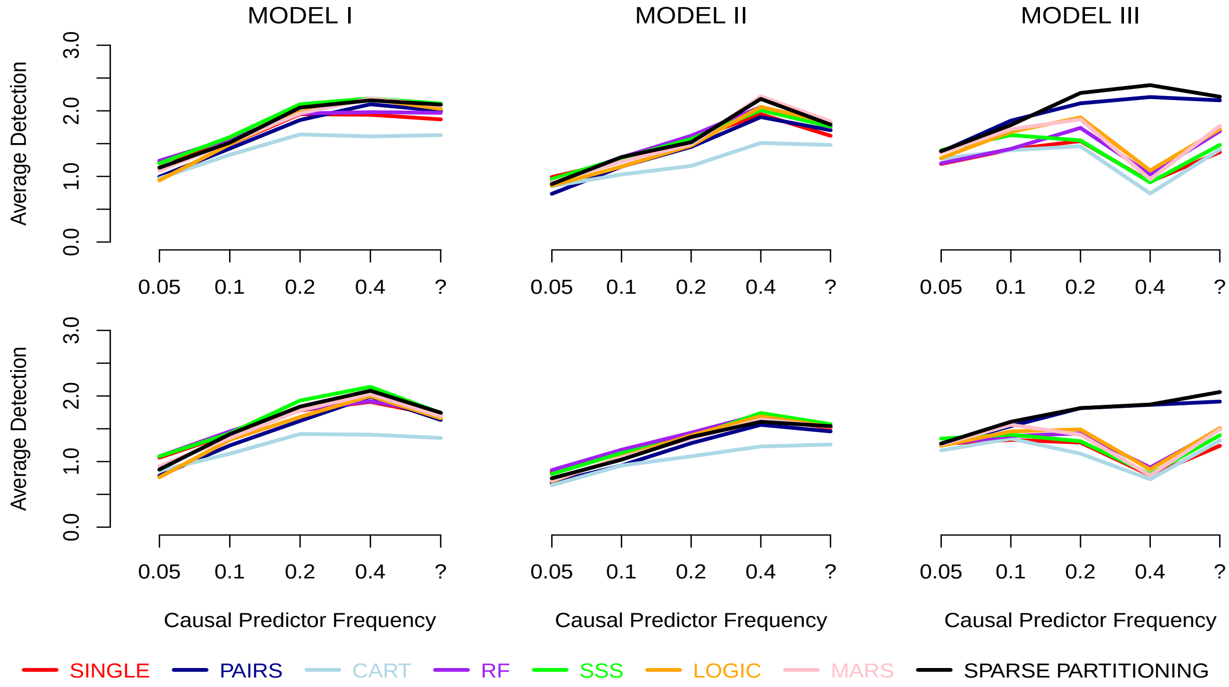

FIG 3. *Results of Study Two: causal predictors unobserved. Top plot,* r^2*=0.9; bottom plot,* r^2*=0.8.*

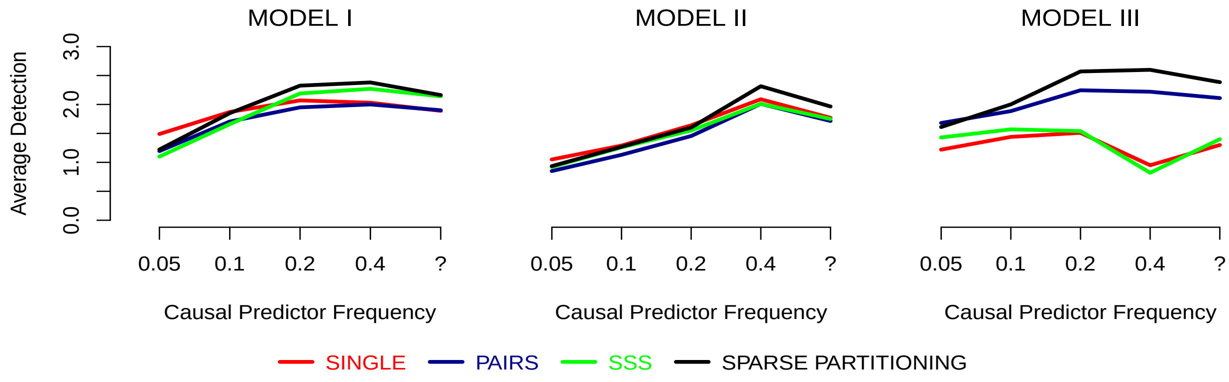

FIG 4. *Results of Study Three: 10% of predictor values missing. The performance of* CART, RF, Logic *and* MARS *could not be compared as they are not designed to handle missing values.*

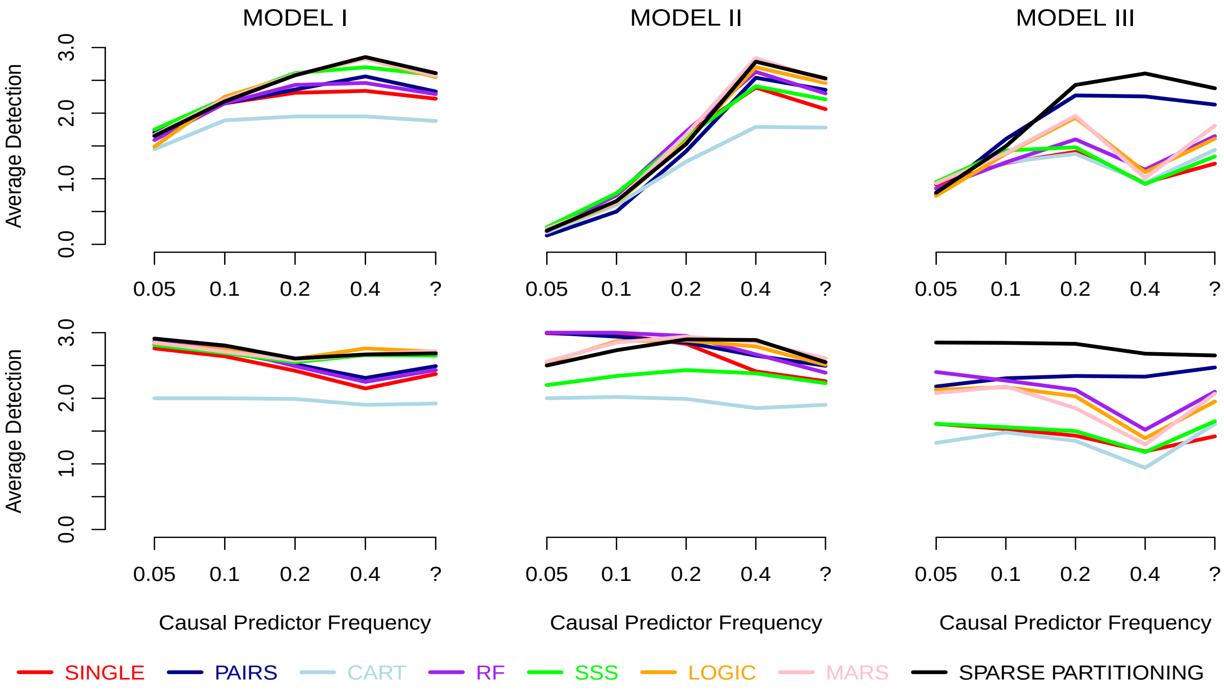

FIG 5. *Results of Study Four: non-normal noise. Top plot, exponential noise; bottom plot, uniform noise.*

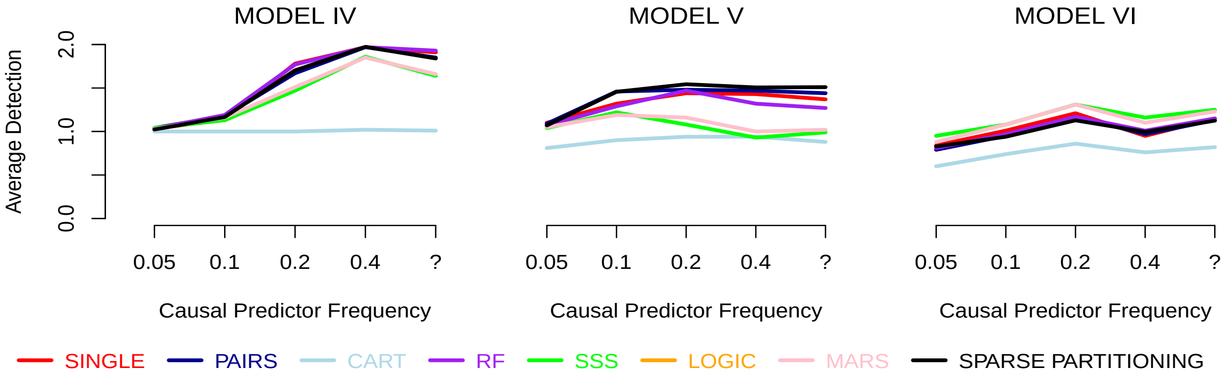

FIG 6. *Results of Study Five: tertiary predictors.*

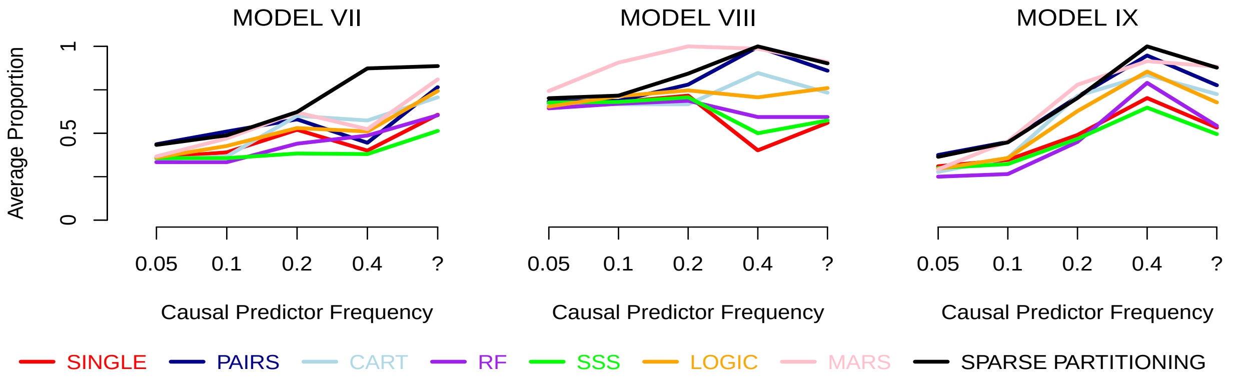

FIG 7. *Results of Study Six: binary response. The number of causal predictors varies between models, so these plots report the proportion of causal predictors correctly identified, rather than the number.*

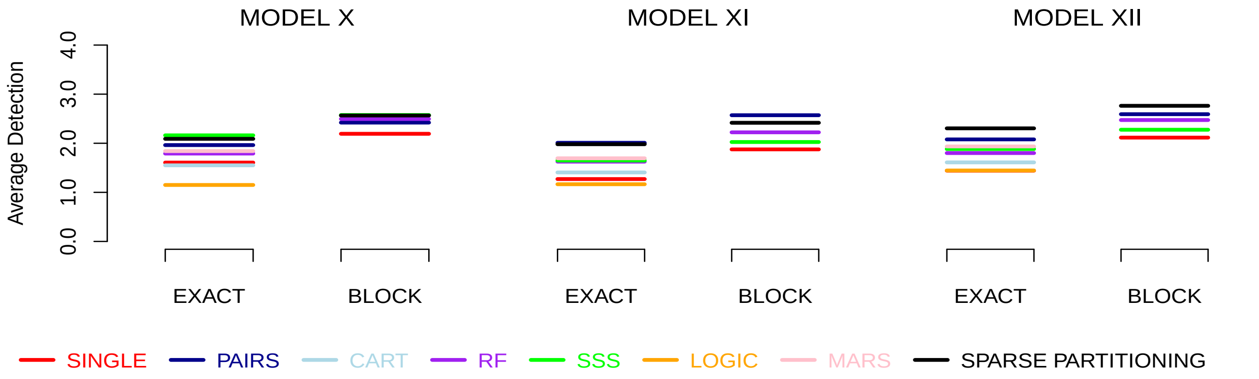

FIG 8. *Results of Study Seven: correlated predictors. Two scoring systems were used, "EXACT" and "BLOCK," details of which are provided in the main text. BLOCK could not be applied to* CART, Logic *and* MARS *so their scores under this system are not recorded.*

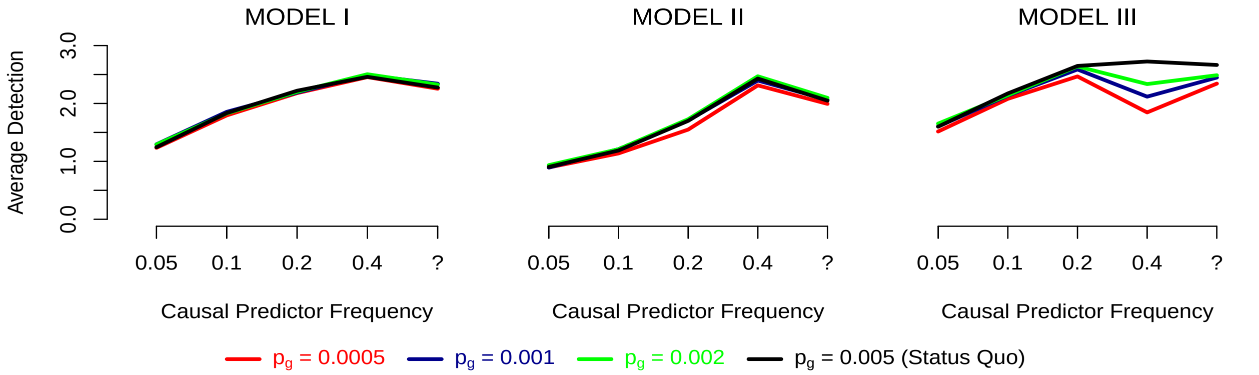

FIG 9. *Results of Study Eight: effect of prior p_g.*

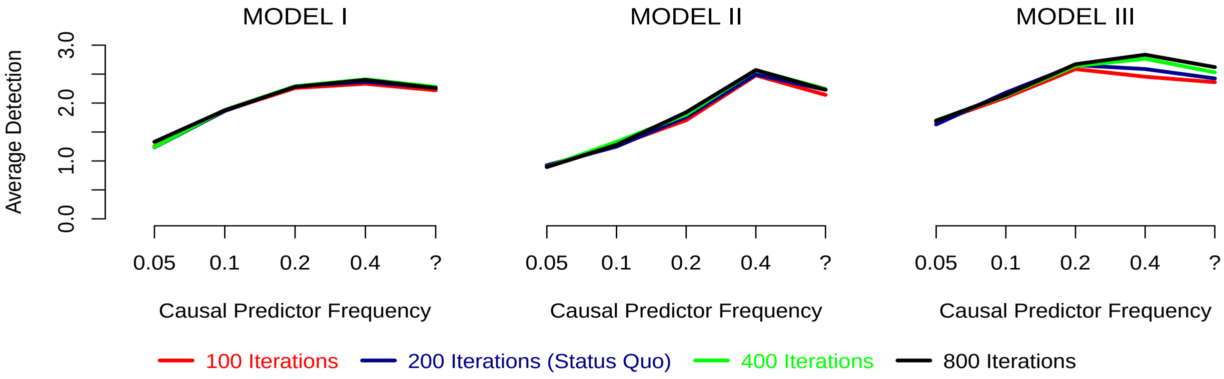

FIG 10. *Results of Study Nine: effect of number of iterations.*

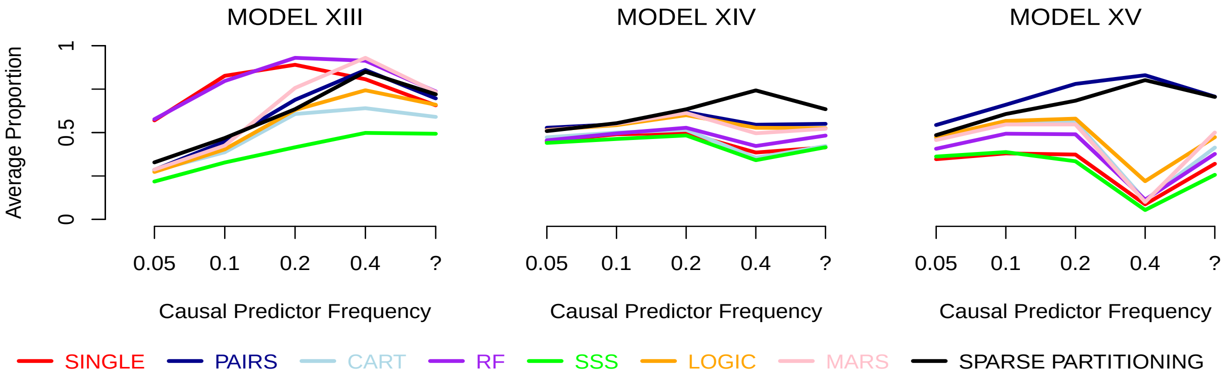

FIG 11. *Results of Study Ten: non-disjoint underlying relationship. The number of causal predictors varies between models, so these plots report the proportion of causal predictors correctly identified, rather than the number.*

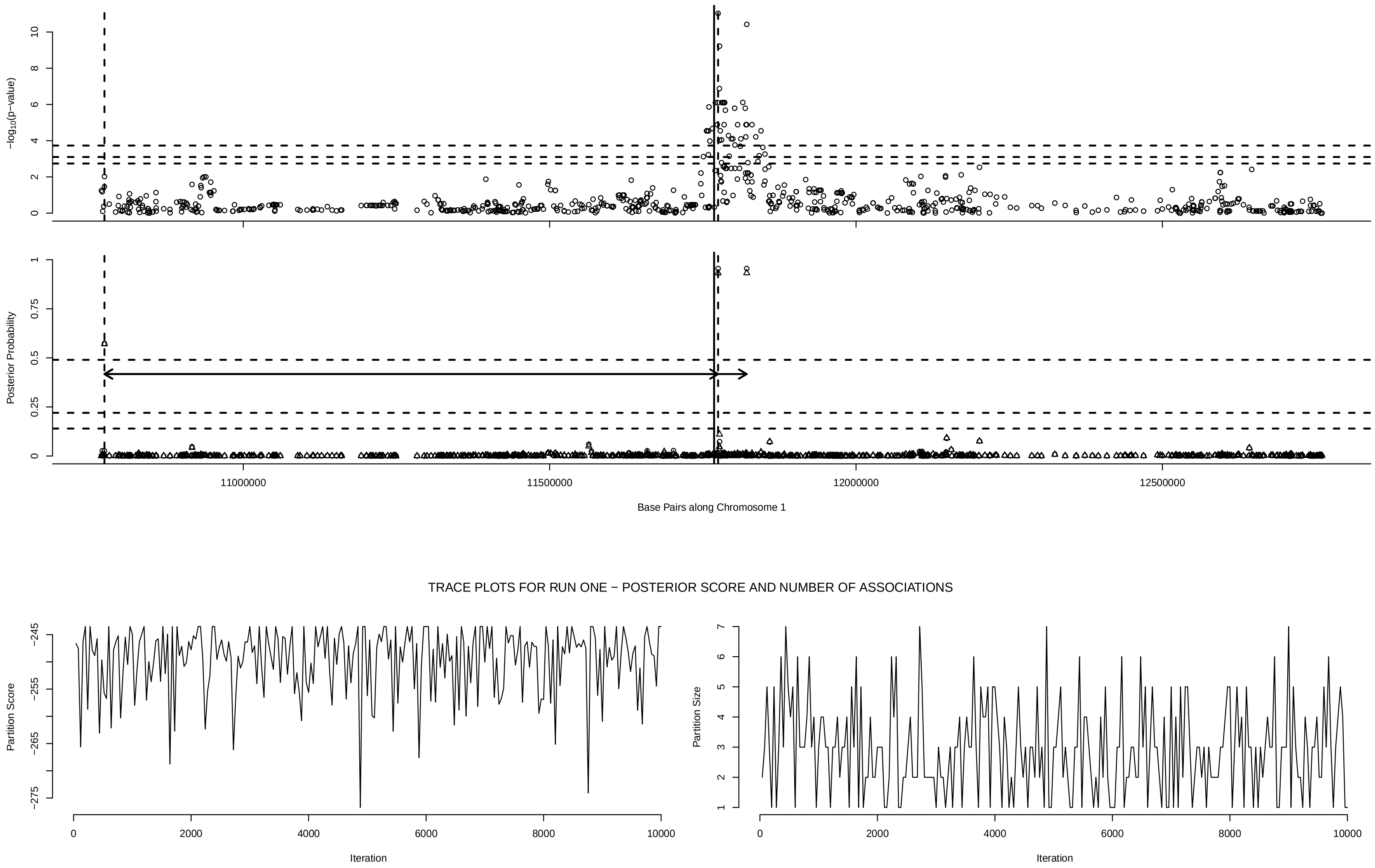

FIG 12. *Analysis of HapMap data. The top plot reports p-values for each SNP, calculated by* Single. *The middle plot reports posterior probabilities of association, returned by* Sparse Partitioning. *The solid vertical line marks the location of the* `MTHFR` *gene, while the dashed vertical lines mark the two SNPs Dr Dimas declared interacting. Additionally, horizontal dashed lines indicate 5, 25 and 50% significance thresholds for the top association of each method, calculated using permutation tests. The pairwise interactions with highest posterior probabilities are represented by the horizontal arrows. The arrow heights indicate the probabilities, while their endpoints mark the pairs of predictors involved. The two plots on the bottom row show posterior scores and number of associations for the iterations visited during the MCMC sampling.*

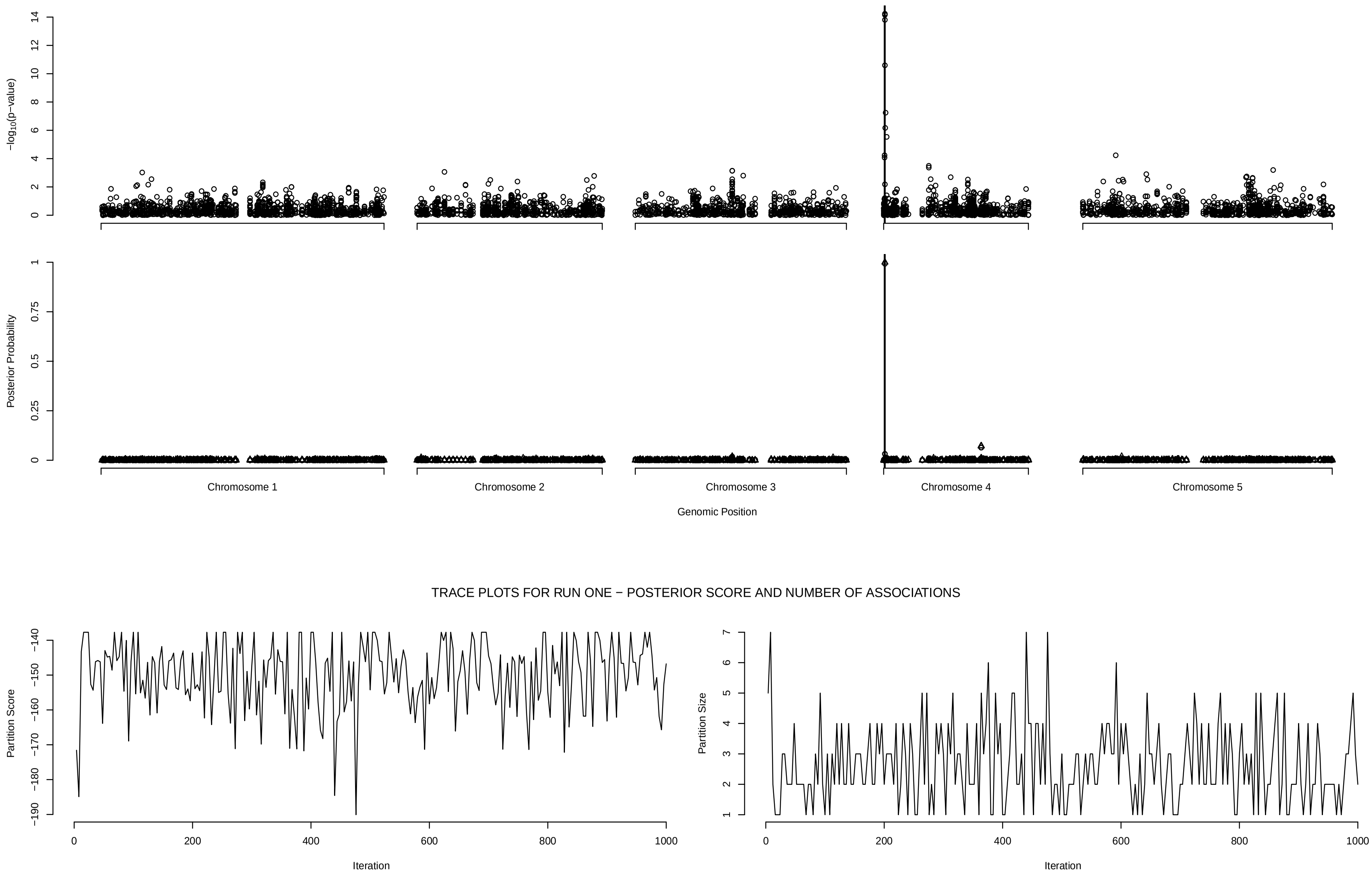

FIG 13. *Analysis of 2010 Project Pilot Data. The top plot reports p-values for each SNP, calculated by* Single. *The middle plot reports posterior probabilities of association, returned by* Sparse Partitioning. *The solid vertical line marks the location of the* ***FRIGIDA*** *gene. The bottom two plots show posterior scores and number of associations for the iterations visited during the MCMC sampling.*

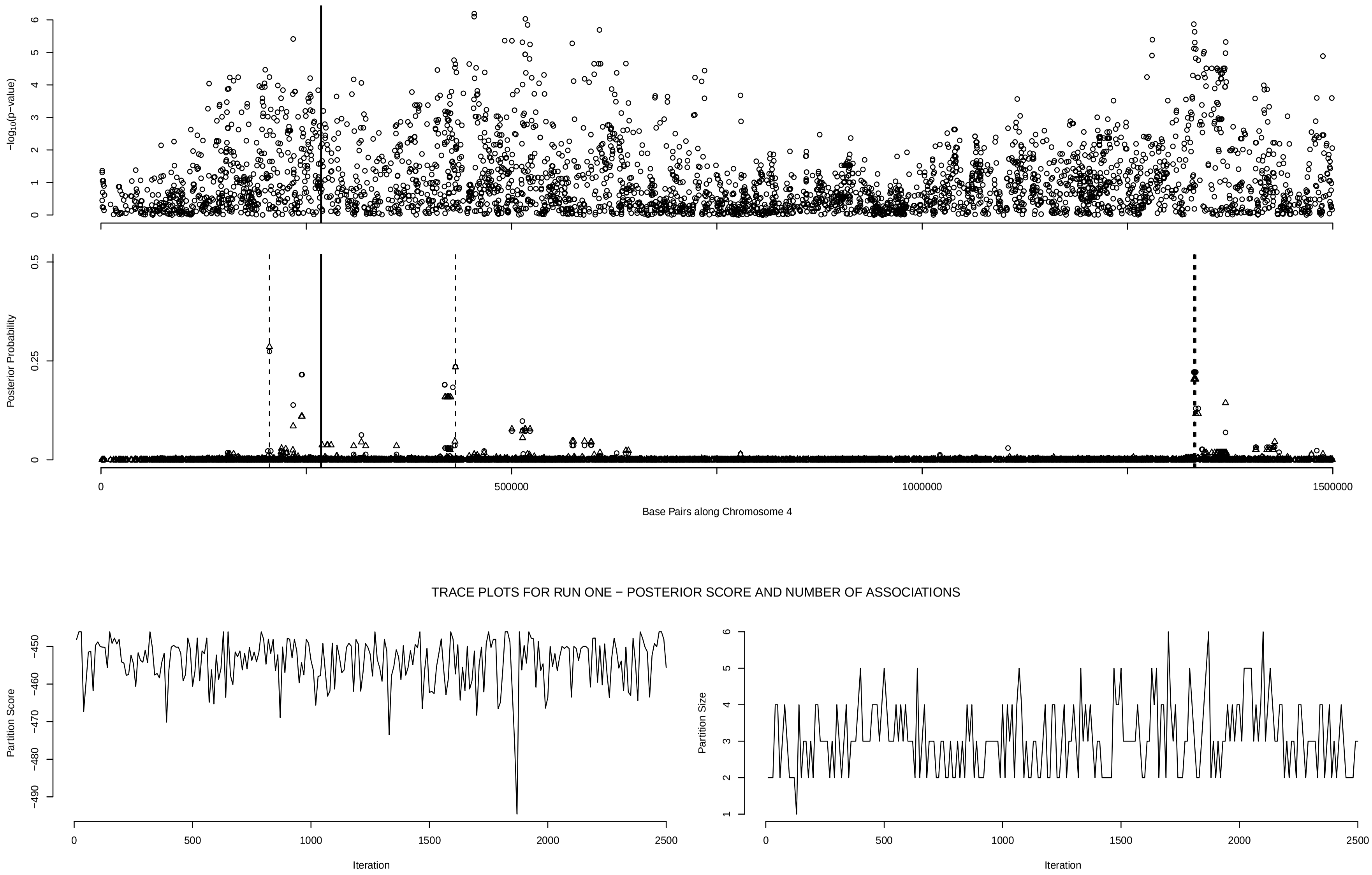

Fig 14. *Analysis of 2010 Project Release 3.04 Data. The top plot reports p-values for each SNP, calculated by* Single. *The middle plot reports posterior probabilities of association, returned by* Sparse Partitioning. *The solid vertical line marks the location of the* ***FRIGIDA*** *gene, while the dashed vertical lines mark the three regions for which* Sparse Partitioning *found most evidence of association. The bottom two plots show posterior scores and number of associations for the iterations visited during the MCMC sampling.*

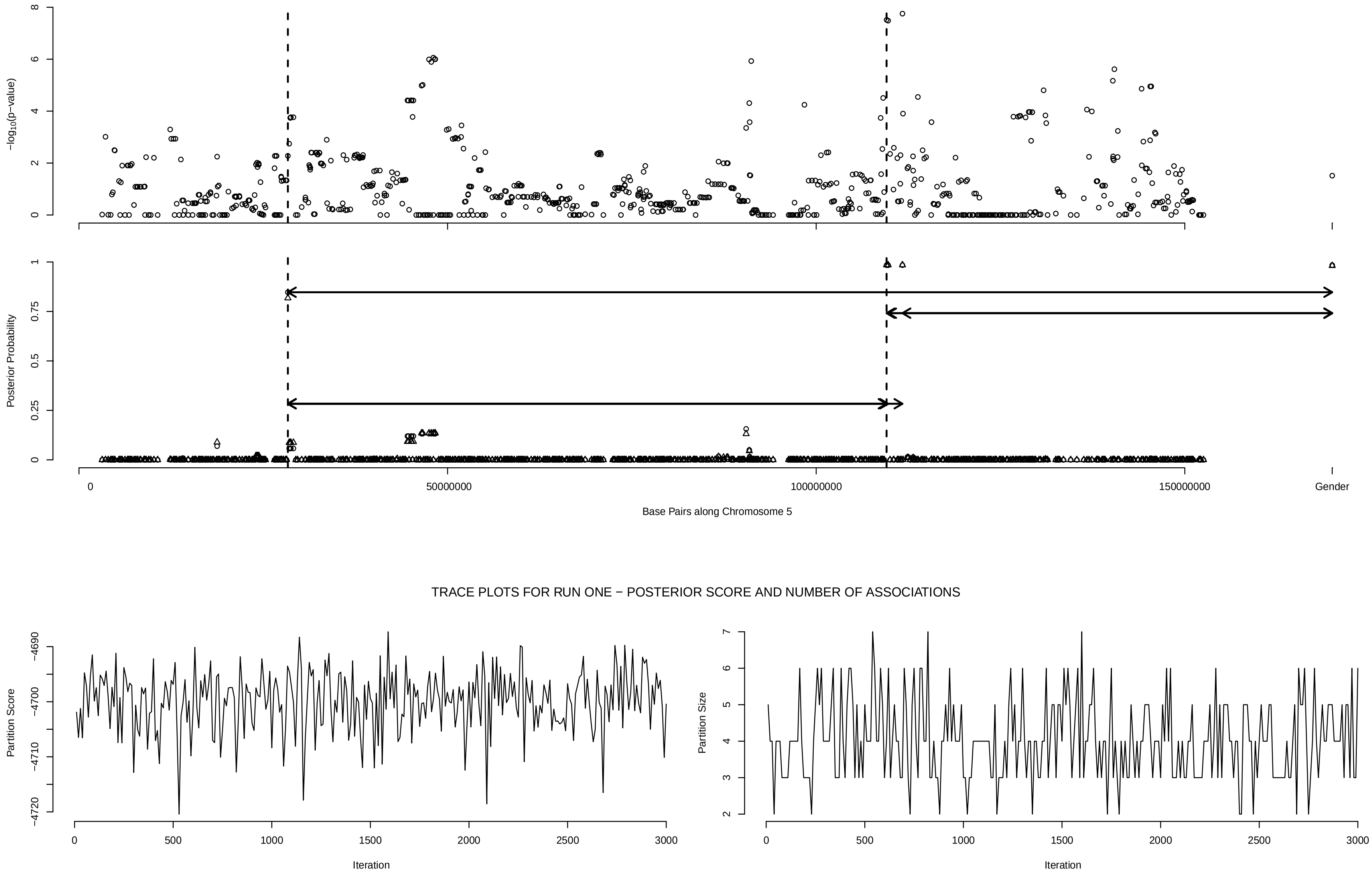

Fig 15. *Analysis of Mouse Data. The top plot reports p-values for each SNP, calculated by* Single*. The middle plot reports posterior probabilities of association, returned by* Sparse Partitioning*. The dashed vertical lines mark the two SNPs for which* Sparse Partitioning *found most evidence of association. The pairwise interactions with highest posterior probabilities are represented by the horizontal arrows. The arrow heights indicate the probabilities, while their endpoints mark the pairs of predictors involved. The bottom two plots show posterior scores and number of associations for the iterations visited during the MCMC sampling.*

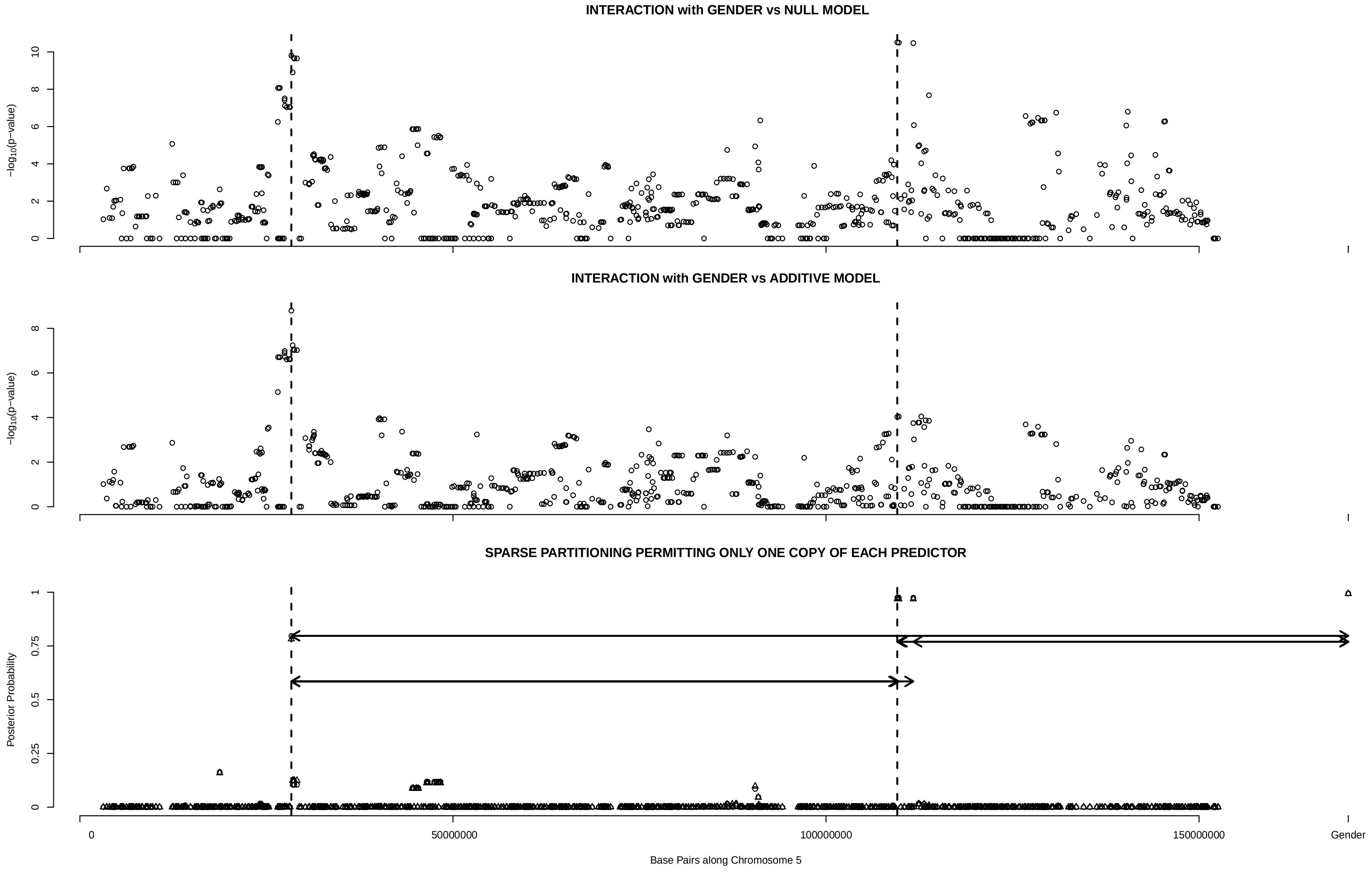

Fig 16. *Analysis of Mouse Data. The top plot shows p-values when comparing the full interaction model of each SNP with gender to the null model of no association. The middle plot shows p-values when comparing the full interaction model of each SNP with gender to the additive model of those two predictors. The dashed vertical lines, carried over from the previous figure, indicate the two SNPs for which* Sparse Partitioning *found most evidence of an association. The bottom plot presents the results of* Sparse Partitioning *when only one copy of each predictor is allowed, demonstrating the effect on the posterior probabilities of interaction when three predictors are forced to feature in only one group of associations.*