



Published in final edited form as:

Genet Epidemiol. 2011 November ; 35(7): 638–649. doi:10.1002/gepi.20613.

Incorporating Model Uncertainty in Detecting Rare Variants: The Bayesian Risk Index

Melanie A. Quintana^{1,*}, Jonine L. Bernstein², Duncan C. Thomas¹, and David V. Conti¹

¹Department of Preventive Medicine, Division of Biostatistics, University of Southern California, Los Angeles, California

²Department of Epidemiology and Biostatistics, Memorial Sloan-Kettering Cancer Center, New York, New York

Abstract

We are interested in investigating the involvement of multiple rare variants within a given region by conducting analyses of individual regions with two goals: (1) to determine if regional rare variation in aggregate is associated with risk; and (2) conditional upon the region being associated, to identify specific genetic variants within the region that are driving the association. In particular, we seek a formal integrated analysis that achieves both of our goals. For rare variants with low minor allele frequencies, there is very little power to statistically test the null hypothesis of equal allele or genotype counts for each variant. Thus, genetic association studies are often limited to detecting association within a subset of the common genetic markers. However, it is very likely that associations exist for the rare variants that may not be captured by the set of common markers. Our framework aims at constructing a risk index based on multiple rare variants within a region. Our analytical strategy is novel in that we use a Bayesian approach to incorporate model uncertainty in the selection of variants to include in the index as well as the direction of the associated effects. Additionally, the approach allows for inference at both the group and variant-specific levels. Using a set of simulations, we show that our methodology has added power over other popular rare variant methods to detect global associations. In addition, we apply the approach to sequence data from the WECARE Study of second primary breast cancers.

Keywords

genetic association studies; Bayesian model uncertainty; Bayes factors; multiplicity correction; sequence analysis; WECARE

INTRODUCTION

In recent years, genetic association studies have proved successful in identifying numerous variants showing strong evidence of associations with various complex diseases [Hindorff et al., 2009]. Most of these studies focus solely on the common-disease common variant (CDCV) hypothesis, since current genotyping panels are limited in their ability to capture rare variation (minor allele frequencies less than 5%). However, despite these successes, the common variants identified are only able to explain a modest amount of the heritability for most complex diseases. This has led researchers to shift more attention to examining the impact of rare variants. This shift was facilitated by the availability of cost-efficient

technologies for complete sequencing of chromosomal regions, as well as the entire genome. Research in this area is often focused on identifying rare variants with a large population attributable risk. For individual rare variants, this corresponds mostly to variants with large effects or high penetrance. In contrast to the CDCV hypothesis, the multiple rare variant (MRV) hypothesis aims to assess how the number of rare variants may impact disease. While motivated in part by biological plausibility, due to power limitations this approach is often the most feasible solution to test whether rare variants contribute to complex disease. Thus, the development of statistical methods for rare variant analyses that are powerful enough to detect individual variants has become essential. In particular, we wish to develop statistical tools that not only answer the question of a global association (are any of the variants associated with the outcome of interest?) but also answer the question of which variants (if any) are driving that association.

Although statistical methods for rare variants have not been studied in as much detail as methods for common variants, there have been an increasing number of developments in this area within recent years. For extensive reviews, see Asimit and Zeggini [2010]; Bansal et al. [2010]; Morris and Zeggini [2010]. The simplest approach to testing rare variants is to assume conditional independence of the variants and investigate each marker individually using standard marginal tests. One major advantage of this approach is that formal marginal inference of the markers is readily available. One must take into consideration multiple testing when determining if any given variant is associated with the outcome of interest. Unfortunately, unlike studies with common variants, the power to detect a true marginal association is extremely limited due to the lower allele frequencies of the markers. Thus, it is not clear that the multiplicity corrections used for common variants can be directly applied to a rare variant analysis.

Alternatives to the marginal approach are multi-marker approaches that model the markers simultaneously. The main advantage of these approaches is that they achieve an increase in power by combining information across markers. A commonly used subset of multi-marker approaches are collapsing methods. Within collapsing methods, a set of variants are combined and their collective frequency in the cases and controls (the rare variant load) is tested. Common examples of collapsing methods include the Combined Multivariate and collapsing method [Li and Leal, 2008], the Cohort allelic sums test [Morgenthaler and Thilly, 2007], and the Weighted Sum test [Madsen and Browning, 2009]. One major drawback of these methods is that they assume that the direction of the effect (protective vs. risk causing) of all variants is the same. In light of this, recent advances have been made to bypass this assumption. In particular, Han and Pan [2010] pre-specify the direction of the effect in a data-driven manner. Also, Neale et al. [2011] have developed the C-alpha test of over-dispersion that is an alternative to the collapsing methods altogether. Their test compares the null hypothesis that there is no effect within the group of variants of interest vs. the alternative hypothesis that there is an increase or decrease in the probability of some of the mutations being found in the affected individuals. The C-alpha test is based on determining if the mixture distribution of the probabilities under the alternative hypothesis has variance and thus bypasses the need to specify the direction of the effect of the variants.

Another major draw back of these more commonly used and newly developed multi-marker methods is that they suffer from a loss of power when the proportion of associated variants to null variants is low since all rare variants (or all variants with a minor allele frequency (MAF) less than a pre-specified threshold) are used to test the hypothesis that an association exists. With this in mind, numerous recent methodological developments have been published, which aim to introduce uncertainty in the subset of variants that are included in the risk index. In particular, Price et al. [2010] introduce a variable threshold method in which all variants with a MAF less than a varying threshold are included in the risk index.

Alternatively, Hoffmann et al. [2010] recently developed a more general model framework that encompasses most of the current collapsing methods by allowing the “weight” of each variant in the risk index to have three components: (1) a continuous component that is a function of the variants MAF, (2) a discrete component that indicates the inclusion or exclusion of the variant in the index, and (3) a discrete component that specifies the direction of the effect. The direction of each effect is pre-specified in a data-driven manner and the inclusion indicator is determined by a step-up procedure. Finally, Bhatia et al. [2010] use a greedy search algorithm to determine a set of variants of pre-specified size that will be included in the risk index. Although these methods do provide some uncertainty in the selection of variants to include in the risk index, they select only one model by using a step-up or greedy algorithm. They also focus mainly on a global test of association by calculating a permutation p -value once the “best” model has been selected by the search function and lack a formal test to determine marginal inference. Thus, if a global association is found within a set of variants, these methods are not able to formally determine which markers are most likely driving the association. This is a major drawback of these approaches since in many cases rare variant analyses will come as a second stage to genome-wide association studies in which a targeted region is investigated with sequencing. Thus, in most cases the researcher will be fairly confident that a global association exists (although this should still be formally tested) and will be most interested in determining which, if any, of the rare variants are driving the association.

The aim of our manuscript is to take advantage of the Bayesian framework and formally incorporate uncertainty via prior probabilities on the inclusion of variants in the risk index and the direction of the effect. In light of this, we have developed a multi-marker Bayesian model uncertainty approach that extends the methodology behind collapsing methods to allow for formal inference at both the global and marginal level. The goal of such an approach is to maintain global power when the proportion of casual variants is low via marker selection while including an implicit multiplicity correction as the number of variants investigated increases. This approach builds upon previous research demonstrating potential advantages to using Bayes model uncertainty approaches for genetic association studies. Conti and Gauderman [2004] used a Bayes model selection strategy to highlight key SNPs and phase terms to identify the best representative SNPs for capturing the genetic variation association with risk. Wilson et al. [2010] demonstrated that a Bayesian model uncertainty approach can lead to an increase in global and marginal power over alternative methods for detecting common variants in candidate gene studies. The idea behind this approach is extended for rare variants so that each model is defined by the inclusion or exclusion of each variant and, if included, the assumed direction of the association. This information is then used to collapse the included variants into a risk index in which the rare variant load is tested. A prior probability is assigned to each model that has an implicit multiplicity correction and the posterior probabilities of each model in the model space are used to calculate formal global and marginal tests of association via Bayes factors (BF) [Kass and Raftery, 1995]. Using a simulation study, we demonstrate that the C-alpha approach of Neale et al. [2011], the comprehensive step-up approach of Hoffmann et al. [2010], and the Bayesian risk index approach developed herein have an increase in power to identify a global association over the commonly used weighted sum test of Madsen and Browning [2009]. The Bayesian risk index approach shows a slight increase in global power over the C-alpha and comprehensive step-up approaches with the most noticeable difference occurring when the proportion of casual variants is low. In addition, the Bayesian risk index enables the detection of true marginal associations given a global association. Finally, we demonstrate the applicability of our novel approach in the WECARE study of bilateral breast cancer risk for rare variants within *BRCA1*.

METHODS

BAYESIAN RISK INDEX MODEL SPECIFICATION

The Bayesian Risk Index methodology uses logistic regression models to relate a subset of the total p variants to a binary outcome variable. Since the methodology is developed within the regression framework, we can also easily adjust for a set of design variables within the analysis. We begin by assuming that we have information on n individuals that includes (1) \mathcal{Y} , a $n \times 1$ vector comprised of the disease status for each individual, (2) \mathbf{G} , a $n \times p$ genotype matrix coded as $G_{ij}=0,1,2$ for the number of copies of the minor allele measured for individual i at SNP j , and (3) \mathbf{Z} , a $n \times q$ dimensional matrix of design variables for each individual (i.e. age, race, etc.). Given this information for each individual, we use Bayesian model uncertainty techniques in which an individual model, denoted by \mathcal{M}_γ , is specified by the p -dimensional vector γ . In particular, γ_j indicates the inclusion of covariate i in model \mathcal{M}_γ and if included the effect of the covariate on the outcome such that $\gamma_j=1$ if the covariate is included as a risk factor and $\gamma_j=-1$ if the covariate is included as a protective factor. A less flexible model can also be investigated where we only allow $\gamma_j=1$. This model is appropriate if one believes that there are not mixed effects (the presence of both protective and risk effects) within the group of rare variants being investigated. We denote the collection of all possible models by \mathcal{M} . Given the subset of variants (of size p_γ) included in model \mathcal{M}_γ and the assumed direction of association for each included variant, we define the risk index \mathbf{X}_γ as

$$\mathbf{X}_\gamma = \mathbf{G} * \gamma,$$

where \mathbf{X}_γ is a vector of length n which gives the risk index for each individual. Then, for each model we use logistic regression to relate the binary outcome variable to the risk index defined by the model \mathcal{M}_γ that is of the form:

$$\text{logit}(\mathcal{Y}=1) = \mu_\gamma,$$

with mean vector $\boldsymbol{\mu}_\gamma = (\mu_{\gamma 1}, \dots, \mu_{\gamma n})^T$, where μ_γ is specified as:

$$\mathcal{M}_\gamma: \mu_\gamma = \alpha_0 + \mathbf{Z}\boldsymbol{\alpha} + \mathbf{X}_\gamma\boldsymbol{\beta}_\gamma.$$

Here, $\boldsymbol{\theta}_\gamma$ is the vector of model-specific parameters $(\alpha_0, \boldsymbol{\alpha}^T, \boldsymbol{\beta}_\gamma)$. These model-specific parameters include the intercept α_0 , vector of design variable coefficients $\boldsymbol{\alpha}$, and the index coefficient $\boldsymbol{\beta}_\gamma$. We notice that this model includes only one effect $\boldsymbol{\beta}_\gamma$ for our group of predictor variables that we refer to as the rare variant load for model \mathcal{M}_γ .

POSTERIOR QUANTITIES OF INTEREST

Within a group of variables we are interested in answering two questions: (1) Globally, is there at least one association within the set of markers? and (2) Given evidence of a global association, what is the most likely marker (or markers) driving the association. Both of these questions can be answered via multilevel posterior probabilities and BFs that are calculated as a function of the posterior probability of each model in the model space.

Model posterior probabilities—The degree to which any model \mathcal{M}_γ in the model space \mathcal{M} is supported by the data is calculated via posterior model probabilities of the form:

$$p(\mathcal{M}_\gamma | \mathcal{Y}) = \frac{p(\mathcal{Y} | \mathcal{M}_\gamma) p(\mathcal{M}_\gamma)}{\sum_{\mathcal{M}_\gamma \in \mathcal{M}} p(\mathcal{Y} | \mathcal{M}_\gamma) p(\mathcal{M}_\gamma)}.$$

Here, $p(\mathcal{M}_\gamma | \mathcal{Y})$ depends on both the marginal likelihood of the model $p(\mathcal{Y} | \mathcal{M}_\gamma)$ and the prior probability of the model $p(\mathcal{M}_\gamma)$. To calculate the marginal likelihood, we use the following approximation:

$$\begin{aligned} p(\mathcal{Y} | \mathcal{M}_\gamma) &= \int p(\mathcal{Y} | \mathcal{M}_\gamma, \theta_\gamma) p(\theta_\gamma) d\theta_\gamma; \\ &\approx p(\mathcal{Y} | \mathcal{M}_\gamma, \hat{\theta}_\gamma). \end{aligned}$$

This approximation corresponds to assuming that all of prior mass of the model-specific parameters, θ_γ is placed on the maximum likelihood estimate (MLE). The MLE of the model-specific parameters and in turn the approximation to the marginal likelihood of each model can easily be calculated using standard statistical packages for generalized linear models.

Given the approximation to the marginal likelihood, we are left to define the prior distribution on the model space $\mathcal{M}_\gamma \in \mathcal{M}$. To do so, we assume that the number of variants in model \mathcal{M}_γ , denoted as p_γ , is distributed as Binomial:

$$p_\gamma | \pi \text{Bin}(p, \pi),$$

where π is the prior inclusion probability of each marker in the analysis. Then conditional upon a variant being included, we will assume a uniform prior on the probability that $\gamma_j=1$ vs. $\gamma_j=-1$. An initial “non-informative” choice for the prior distribution on model size would be to choose $\pi=0.5$ thus giving a uniform prior across the model space. While this prior seems to be noninformative with respect to the model space, it actually can be quite informative with respect to the global prior probability of at least one association as p increases. Thus, to increase the flexibility of our model space prior we place a hyper-prior Beta distribution on the inclusion probability π :

$$\pi \text{Beta}(a, b),$$

with hyper-parameters a and b . This is the Beta-Binomial prior on model size which we will denote as $\text{BB}(a,b)$. In particular, if we set the hyper-parameters to be $\text{BB}(1,p)$, [as introduced in Wilson et al., 2010] we will maintain a constant global prior probability (prior probability that at least one of the variants is associated) of $1/2$ no matter how many variants are included in the analysis. Table I gives several characteristics of the $\text{Bin}(p,1/2)$ and $\text{BB}(1,p)$ priors. With the $\text{BB}(1,p)$ prior, we are assured that the global prior probability of an association will not increase as we increase the number of (possibly redundant or highly correlated) variables within an analysis and the marginal prior of any particular variant being associated will decrease as a sort of multiplicity correction. Given the implicit multiplicity correction with the $\text{BB}(1,p)$ prior, we choose to use this for the remainder of the article.

Under the $\text{BB}(1,p)$ distribution, the prior of any model \mathcal{M}_γ can be calculated as

$$p(\mathcal{M}_\gamma) = p\left(\frac{1}{2}\right)^{p_\gamma} \left[(2p - p_\gamma) \binom{2p}{p_\gamma} \right]^{-1},$$

where the term $(1/2)^{p_\gamma}$ accounts for the uniform prior on $\gamma_j = 1$ vs. $\gamma_j = -1$. This term can be removed if we are only considering models with $\gamma_j = 1$.

Model search—In most studies of interest the total number of variants being analyzed can be quite large and in turn the model space \mathcal{M} quickly becomes innumerable. In particular, when both protective and risk effects are introduced into each model the total number of models in the model space is 3^p . Thus, with $p > 20$ enumeration is infeasible and the normalizing constant in the calculation of the posterior model probabilities must be approximated by summing over all models sampled by some model search algorithm denoted \mathcal{M}_s as follows:

$$p(\mathcal{M}_\gamma | \mathcal{Y}) \approx \frac{p(\mathcal{Y} | \mathcal{M}_\gamma) p(\mathcal{M}_\gamma)}{\sum_{\mathcal{M}_\gamma \in \mathcal{M}_s} p(\mathcal{Y} | \mathcal{M}_\gamma) p(\mathcal{M}_\gamma)}.$$

For our purpose, in cases where $p > 20$ we use a simple Metropolis Hastings (MH) algorithm to sample models from the model space. In the MH algorithm, models are evaluated based on the following fitness function:

$$\text{fitness}(\mathcal{M}_\gamma) = p(\mathcal{Y} | \mathcal{M}_\gamma) p(\mathcal{M}_\gamma).$$

New models are proposed by randomly selecting one variant and changing the status of the variant within the current model. For example, if variant j is randomly selected and is included in the current model with $\gamma_j = -1$ we propose to either remove variant j from the new model all together or to include variant j in the new model with $\gamma_j = 1$. The new model is then accepted with probability $a = \min(1, \text{fitness}(\mathcal{M}_{\text{new}}) / \text{fitness}(\mathcal{M}_{\text{old}}))$ so that the new model is always accepted if the fitness of it is larger than that of the old model and is accepted with a probability of less than 1 if the fitness of the new model is smaller than that of the old.

Global posterior quantities—To determine if a global association exists within the variants of interest we wish to test the following global hypotheses:

- H_A : At least one variant is associated with the outcome of interest
- H_0 : There is no association between the variants of interest and the outcome

To test these hypotheses we can calculate posterior probabilities and BFs. Posterior probabilities quantify the extent to which the data support each of the hypotheses and are calculated as:

$$p(H_A | \mathcal{Y}) = \frac{p(\mathcal{Y} | H_A) p(H_A)}{p(\mathcal{Y} | H_A) p(H_A) + p(\mathcal{Y} | H_0) p(H_0)}$$

and $p(H_0 | \mathcal{Y}) = 1 - p(H_A | \mathcal{Y})$. These quantities are a function of both the marginal likelihood of the hypotheses and the prior distributions placed on the model space and in

turn the global hypotheses. In particular, the global posterior probability of an association can be calculated as the sum of the posterior probabilities of all of the non-null models:

$$p(H_A|\mathcal{Y}) = \sum_{\mathcal{M}_\gamma \neq \mathcal{M}_0} p(\mathcal{M}_\gamma|\mathcal{Y}).$$

The posterior probability of the null hypothesis can then be calculated as the posterior probability of the null model:

$$p(H_0|\mathcal{Y}) = p(\mathcal{M}_0|\mathcal{Y}).$$

Given the posterior probabilities of the hypotheses, we can also calculate the ratio of the weight of evidence for any two hypotheses (H_A vs. H_0) based on BFs:

$$\text{BF}[H_A:H_0] = \frac{p(H_A|\mathcal{Y})}{p(H_0|\mathcal{Y})} \div \frac{p(H_A)}{p(H_0)}.$$

A BF [Kass and Raftery, 1995] compares the posterior odds of any two hypotheses to the prior odds and measures the change of evidence provided by the data for one hypothesis to the other. The Global BF for comparing H_A to H_0 may be simplified to

$$\text{BF}[H_A:H_0] = \sum_{\mathcal{M}_\gamma \in \mathcal{M}} \text{BF}(\mathcal{M}_\gamma:\mathcal{M}_0) p(\mathcal{M}_\gamma|H_A),$$

which is the weighted average of the individual BF for comparing each model in H_A to the null model with weights given by the prior probability of \mathcal{M}_γ conditional on being in H_A , $p(\mathcal{M}_\gamma|H_A)$. However, when enumerability of the model space is not feasible we can replace the sum over the entire model space \mathcal{M} in this calculation with the sum over the models sampled in the model search algorithm which is shown to be a lower bound for the global BF calculated under the entire model space [Wilson et al., 2010]:

$$\text{BF}[H_A:H_0] \leq \sum_{\mathcal{M}_\gamma \in \mathcal{M}_s} \text{BF}(\mathcal{M}_\gamma:\mathcal{M}_0) p(\mathcal{M}_\gamma|H_A).$$

Thus, if the global BF suggests that there is strong evidence of at least one variant being associated in the study then we can be confident that this BF will only increase as we sample more models in the model space.

Marginal posterior quantities—Given that there is posterior evidence of the global hypothesis that at least one of the variables of interest is associated with the outcome, we are further interested in answering the question of which variable (or variables) are most likely driving the association. This question can be answered based on marginal posterior probabilities and BFs. The marginal posterior inclusion probability for any variable G_j can be calculated simply as the sum of the posterior probabilities for every model that includes the variable G_j :

$$p(\gamma_j \neq 0 | \mathcal{Y}) = \sum_{\mathcal{M}_\gamma \in \mathcal{M} : \gamma_j \neq 0} p(\mathcal{M}_\gamma | \mathcal{Y}).$$

The marginal BF is the ratio of the posterior odds of any variable G_j to the prior odds of the same and is defined as:

$$\text{BF}[\gamma_j \neq 0 : \gamma_j = 0] = \frac{p(\gamma_j \neq 0 | \mathcal{Y})}{p(\gamma_j = 0 | \mathcal{Y})} \div \frac{p(\gamma_j \neq 0)}{p(\gamma_j = 0)},$$

where $p(\gamma_j = 0)$ is the prior probability of variable G_j being associated.

Evidence—Jeffreys [1961] presents a descriptive classification of BFs into “grades of evidence” (reproduced in Table II) to assist in their interpretation, which is also reproduced in Kass and Raftery [1995]. Thus, decisions about which hypothesis is more likely can be made based on these grades of evidence.

POWER SIMULATIONS

To explore the power of our methodology vs. other alternative methods and to determine optimal parametrizations ($\gamma_j \in \{0, 1\}$ vs. $\{-1, 0, 1\}$) we use a set of 10,000 simplistic simulations. In particular, we assume that in each simulation there are 10 variants (each with a MAF ranging in $\{0.001, 0.005, 0.01\}$) of interest genotyped for 500 cases and 500 controls. Each variant was simulated independently. Within each simulation, we assume that anywhere between 0 and all 10 of the variants are associated with the binary outcome of interest. If associated, we assume that the variants will act independently and a log additive disease model for each. We also assume that the absolute value of the log odds ratio (OR) of each effect in the simulation is in $\{0.5, 1, 1.5, 2, 2.5, 3\}$ (with a probability of 0.2 of being a protective effect and a probability of 0.8 of being a risk causing effect). This corresponds to ORs of risk effects between 1.6 and 20.

To better understand the power of rare variant methods, when the proportion of causal to null variants is low we also use a set of 1,000 simulation replicates where the nature of the simulations (MAF and assumed log OR) is the same as the small simulations but there are a total of 100 variants in each study. We still assume that the true number of casual variants is between 0 and 10 so that the proportion of casual to null variants is between 0.0 and 0.1.

COMPARISON WITH OTHER METHODS

We compare our Bayesian risk index methodology with the commonly used weighted sum method of Madsen and Browning [2009], the more recently introduced C-alpha test of Neale et al. [2011] and the comprehensive step-up approach of Hoffmann et al. [2010].

- *Weighted sum*: The weighted sum statistic of Madsen and Browning as described in Madsen and Browning [2009]. Here, the rare allele is assumed to be the mutation and a weight is calculated for each variant that is equivalent to the inverse of the estimated standard deviation under the null hypothesis of no association of the total number of mutations in the sample. Using these weights, a genetic score for each individual in the simulation is calculated and the individuals are ranked according to their score. We then calculate the sum of the ranks for affected individuals and compute a standardized score sum and global permutation p -value for each simulation based on k permuted samples. Here, we assume that $k = 1,000$ as recommended in Madsen and Browning [2009].

- *C-alpha*: The C-alpha method of Neale et al. as described in Neale et al. [2011]. Here, we compute a C-alpha test statistic that compares the null hypothesis that there is no association between the group of variants of interest and the affected individuals vs. the alternative hypothesis that there is an increase or decrease in the probability of some of the mutations being found in the affected individuals. This test statistic is based on determining if the mixture distribution of probabilities under the alternative hypothesis has variance. If there is no variance, then there is an indication that the mixture distribution is actually a point mass and we do not reject the null hypothesis. As recommended in Neale et al. [2011], we assume that the probability of observing a variant in the cases, p_0 , is fixed at the known value of 0.5 for the simulations and we also pool the singletons into a single variant. We compute global p -values for each simulation based on the asymptotic distribution of the C-alpha test statistic.
- *Comp. step-up*: The comprehensive step-up approach of Hoffmann, Marini, and Witte as described in Hoffmann et al. [2010]. Here, a weight is calculated for each variant that is composed of a continuous component that is a function of the variants MAF, a discrete component that indicates the inclusion or exclusion of the variant in the index, and a discrete component that specifies the direction of the effect (risk vs. protective). The direction of each effect is pre-specified in a data-driven manner and the inclusion indicator is determined by a step-up procedure. A global permutation p -value is calculated based on the “best” model that is found in the step-up procedure using k permuted samples. Here, we assume that $k = 1,000$ as recommended in Hoffmann et al. [2010].
- *Bayesian risk index*: We compute global and marginal BF for each simulation as described above using the $BB(1,p)$ prior on model size. These posterior quantities are calculated under the model $\gamma_j \in \{-1, 0, 1\}$ and the less flexible $\gamma_j \in \{0, 1\}$ model. For the set of large simulations, we run the basic MH algorithm for 100,000 iterations to sample a set of models from the model space and estimate global and marginal BFs. Convergence of the model search algorithm is determined by inspecting global and marginal BF from two independent runs of the MH algorithm to make sure that the values are similar between the two runs.

WECARE STUDY EXAMPLE

The WECARE (Women's Environmental Cancer and Radiation Epidemiology) Study is a population-based case-control study designed to investigate the joint roles of genetic and environmental risk factors in the etiology of second primary breast cancer, and especially radiation-induced breast cancers [Bernstein et al., 2004]. The study included 705 women with contralateral breast cancer (cases) and 1,398 women with unilateral breast cancer (controls). Two controls were individually matched to each case on birth year, year of first primary diagnosis, race/ethnicity, and reporting registry. For all participants, the complete coding sequences of *BRCA1* (5,589 bp split into 22 coding exons) and *BRCA2* (10,254 bp and 26 coding exons) were screened for variations by denaturing high-performance liquid chromatography, using leukocyte genomic DNA as a template, identifying a large number of both common and rare variants [Begg et al., 2008; Borg et al., 2010; Malone et al., 2010]. Previously, we have shown that when rare sequence variants known to have a deleterious effect are combined, the risk of developing a second primary breast cancer increased over fourfold [Malone et al., 2010]. In more recent work using established bioinformatic prediction tools and hierarchical modeling, we have identified a set of individual rare *BRCA1* and *BRCA2* missense variants likely to be deleterious [Capanu et al., 2011]. To demonstrate the use of the Bayesian risk index methodology with the aim of identifying additional individual rare variants while also investigating the aggregate effect on risk of

developing contralateral breast cancer, below we investigate a subset of 134 rare *BRCA1* variants ($MAF < 0.05$). We restrict this analysis to white non-Hispanic cases ($n_{\text{case}} = 640$) and controls ($n_{\text{control}} = 1,272$).

RESULTS

GLOBAL POWER

We first aim to examine the power to identify a global association (which we refer to as global power) of the Bayesian risk index vs. the alternative methods. To test the global power of the Bayesian risk index, we assume two parametrizations for the analysis: (1) We assume that all effects are in the same direction (protective or risk causing) and $\gamma \in \{0, 1\}$ (denoted as BRI); and (2) We allow for both protective and risk factors using $\gamma \in \{-1, 0, 1\}$ (MixBRI). For each analysis, we use a $BB(1, p)$ prior on the model space. Figures 1 and 2 plot the global false-positive rate (gFPR) vs. global true-positive rate (gTPR) as the global BF threshold varies for the Bayesian risk index method and as the global p -value threshold varies for the weighted sum, C-alpha, and comp. step-up methods. The gFPR and gTPR are calculated across all 10,000 small simulations (10 total variants) in Figure 1 plot A, across a subset of the small simulations that assume all of the effects are in the same direction (referred to as non-mixed simulations) in plot B, a subset of the small simulations with $OR < 10$ in plot C, and a subset of the small simulations with $OR > 10$ in plot D. The gFPR and gTPR are calculated across all 1,000 large simulations (100 total variants) in Figure 2 plot A, across a subset of the nonmixed large simulations in plot B, a subset of the large simulations with $OR < 10$ in plot C, and a subset of the large simulations with $OR > 10$ in plot D. Given a specific global BF or p -value threshold, the gFPR is calculated as the ratio of the number of nonassociated simulation scenarios (none of the variants are associated with the outcome) in which a global association has been detected (global BF is greater than the threshold or global p -value is less than a threshold) vs. the total number of nonassociated simulation scenarios. The gTPR is calculated as the ratio of the number of associated simulation scenarios (at least one of the variants is associated with the outcome) in which a global association has been detected vs. the total number of associated simulations.

Figures 1 and 2 show that the Bayesian risk index, C-alpha, and comp. step-up methods have much greater power to detect an association globally than the weighted sum method in both the small and large simulations. As expected, the difference in power is greatest when the gTPR and gFPR are calculated across simulations with mixed and nonmixed effects (Figs. 1 and 2 plot A) since the Bayesian risk index, C-alpha, and comp. step-up allow for both protective and risk causing effects whereas the weighted sum method assumes that the direction of the effect will be the same across all associated variants. The difference in power is reduced when the quantities are calculated under the subset of nonmixed simulations (Figs. 1 and 2 plot B). The Bayesian risk index, C-alpha, and comp. step-up methods have comparable global power in the small simulations (Fig. 1), while the Bayesian risk index shows a slight increase in global over C-alpha and comp. step-up in the large simulations (Fig. 2) where the proportion of associated variants to null variants is small. This reduction in power for C-alpha when the proportion of associated variants to null variants is small is not surprising since this method does not account for uncertainty in the subset of variants that are included in the analysis. This reduction in power is somewhat surprising for the comp. step-up method since Hoffmann et al. do introduce uncertainty in the subset of variants that are included in the risk index. However, unlike the Bayesian risk index methodology developed herein, Hoffmann et al. determine the inclusion/exclusion of each variant by a step-up procedure and use only the “best” model found by this procedure to calculate the global p -value. Thus, the reduction in power in the comp. step-up method when the proportion of associated variants to null variants is low may reflect the weakness

of using simple step-up procedures for model selection as well as the added benefit of using Bayesian model averaging in testing global hypotheses. Finally, the Bayesian risk index with $\gamma \in \{0, 1\}$ has comparable global power to the more general γ parametrization with $\gamma \in \{-1, 0, 1\}$. The less general parametrization of using $\gamma \in \{0, 1\}$ for the Bayesian risk index explicitly assumes that the direction of the effects of each variant within a specific model is the same. However, both protective and risk causing effects can be accounted for in this less general setting through separate models. Therefore, the comparable global power of the two parametrizations may be due to and reflect the benefit of using Bayesian model averaging techniques to construct global summaries. Table III shows the estimated gTPR across all small simulations. The method-specific threshold (reported next to each method) is determined by holding the gFPR at a constant value of 0.2. These true-positive rates are reported with the minor allele frequencies and the number of true marginal associations varying across the simulations.

MARGINAL POWER

Our next aim is to determine the marginal power for each variant using the Bayesian risk index. To test the marginal power using the Bayesian risk index, we use the same two parametrizations as in the previous section (BRI vs. MixBRI). As before, we use a $BB(1,p)$ prior on the model space. Figure 3 plots the marginal false-positive rate (mFPR) vs. marginal true-positive rate (mTPR) as the marginal BF threshold varies for both parametrizations of the Bayesian risk index method. The mFPR and mTPR are calculated across all 10,000 small simulations (10 total variants) in Figure 3 plot A, across a subset of the small simulations that assume all of the effects are in the same direction (referred to as non-mixed simulations) in B, across all 1,000 large simulations (100 total variants) in plot C, and across nonmixed large simulations in plot D. Given a specific marginal BF threshold mFPR is calculated as the ratio of the number of nonassociated variants in which a marginal association has been detected (marginal BF is greater than the threshold) vs. the total number of nonassociated variants. The mTPR is calculated as the ratio of the number of associated variants in which a marginal association has been detected vs. the total number of associated variants.

Figure 3 plots A and C show that the power to detect a marginal association is higher using the $\gamma \in \{-1, 0, 1\}$ model parametrization (MixBRI) for both the small and large simulations. This is expected since the Bayesian risk index with $\gamma \in \{0, 1\}$ model parametrization assumes that the effects of all the variants included in the index are of the same direction and magnitude. Thus, a variant with a protective effect and another variant with a risk effect will mostly likely not to be included within the same model. This will cause a dilution of the marginal inclusion probabilities and marginal BF. As expected, when we look at only the nonmixed simulations (Fig. 3 plots B and D), the power to detect a marginal association is higher using the $\gamma \in \{0, 1\}$ model parametrization (BRI) due to the unnecessary increase in model space and in our uncertainty with the $\gamma \in \{-1, 0, 1\}$ model parametrization.

WECARE ANALYSIS

For the WECARE example data, we apply the Bayesian risk index methodology on all rare variants ($MAF < 0.5$) in *BRCA1*. Convergence of the MH algorithm is determined based on investigating the marginal BFs calculated under two independent runs of the algorithm. Once convergence is determined, we combine the results from both independent runs of the MH algorithm to calculate the global and marginal BFs for *BRCA1*. From this analysis, the global BFs for *BRCA1* is calculated as 29.3—providing strong evidence that at least one of the rare variants within *BRCA1* is associated with contralateral breast cancer. To determine which variants are most likely driving the association, we plot the rare variants included within the top models for the Bayesian risk index in Figure 4. The figure plots the inclusion

or exclusion of the top 10 variants within *BRCA1* (ordered by marginal BF) within the top 25 models (ordered by posterior model probability). The width of each column is proportional to the corresponding posterior model probability and estimated ORs are reported below each of the top three models. Here, we see that the top model corresponds to that with only variant 13 from exon 5. Also, the corresponding effect for the risk index produced by the top model has an OR of 8.2. We note that all of the top 25 models are comprised of risk indices from anywhere between one and four rare variants across exons 2,3,5,11 and 12. This suggests that it is likely that the proportion of casual to null variants within *BRCA1* is low and other multimarker rare variant methods may lack power in this instance. Also, these methods would not be able to pinpoint the most likely variant driving the association. In particular, when we run the C-alpha and weighted sum approaches on the 134 rare variants in *BRCA1* we calculate modest p -values of 0.0133 and 0.0033, respectively. However, if we were to limit the analyses to look at only the top 10 variants discovered by the Bayesian risk index approach we calculate much smaller p -values of $1.528e^{-22}$ and $4.306e^{-15}$, respectively. While these p -values are not valid since they do not account for the model selection, this reflects the advantage of methods that incorporate uncertainty into which rare variants should enter the analysis.

Table IV provides more information on the top 10 rare variants that are most likely driving the global association within *BRCA1*. In general, we see that the evidence for each variant via the marginal BF is correlated with the proportion of cases that have the variant vs. controls. However, we note the slight difference in the marginal BFs for all the variants that occur in two cases and 0 controls. In part, this reflects the uncertainty in the estimation of each BF. In addition, this also reflects that each marginal BF is calculated conditional upon all other markers in the analysis. Thus, if two variants are correlated or found on the same individual it is unlikely that they will co-occur within the same top model since the added information from the additional variant is reduced. This will cause a dilution in the marginal BF of the less likely marker driving the association. In our example, variant 7 in Exon 3 and variant 2 in Exon 2 show a modest correlation (they are both found on the same individual). Thus, we see in Figure 4 that variant 7 and variant 2 do not co-occur within any of the top 25 models which leads to a dilution in the marginal BF of variant 7. In contrast, since there is no correlation between variants 3, 52, 106, 8, 137, and 112 and any of the other markers within the top 10 variants, the disparity within marginal BF of these variants is most likely due to the sampling variability within the model search algorithm. We also note that due to their low frequency in the population, none of the variants that occur in 2 cases and 0 controls appear in the top 25 models by themselves as a single effect. It is only in the models with MRVs that we are able to detect an association between these variants and contralateral breast cancer. This emphasizes the importance of adopting multimarker methods for rare variant analyses. Finally, we note that these results vary slightly from those reported in Capanu et al. [2011], an analysis using a novel hierarchical modeling approach [Capanu et al., 2008; Capanu and Begg, 2010] and bioinformatic categorizations to assess the role of missense mutations in *BRCA1* and *BRCA2* in the WECARE data. While qualitatively similar in concluding that rare variation within *BRCA1* does impact contralateral breast cancer, to highlight key aspects of our approach, the analysis reported here examines all rare variants within *BRCA1* with a $MAF < 0.05$ and uses a risk index model with stochastic selection with an unconditional likelihood and limited covariate adjustment. Specifically among missense mutations, the top two missense variants identified in the Capanu et al. article, variant 13 in Exon 5 (C61G) and variant 8 in Exon 3 (C44S) are also notable variants in our analysis with BFs of 124.3 and 13.3, respectively. Capanu et al. report several additional missense variants with each having a single occurrence in controls and with support from bioinformatic predictors. These variants were aggregated in our analysis so independent inference is not feasible.

DISCUSSION

The Bayesian risk index is a highly flexible method that allows for uncertainty into which variants are included in the risk index as well as if the variants have a protective or risk effect. The methodology developed herein extends upon current multimarker rare variant methods that provide only global inference by providing formal intuitive inference at both global and marginal levels. Thus, we are able to formally test two hypotheses of interest: (1) Is there evidence of a global association with at least one of the markers of interest; and (2) If there is evidence of a global association, what are the most likely markers driving that association. As demonstrated in the applied WECARE Study example, the Bayesian risk index methodology was able to detect a global association of the rare variants within *BRCA1* and also pinpoint specific variants within the gene that are most likely driving the association. By performing a model search to determine which variants to include and using Bayesian model averaging techniques to calculate posterior quantities of interest, the Bayesian risk index shows an increase in power to detect global associations over other popular methods for rare variant analyses. The increase in power for the Bayesian risk index is most noticeable when the proportion of associated variants to null variants is low. The global power comparison described herein is calculated on a set of small and large simulated data sets that assume conditional independence of the variants within each simulation. To examine the effect of the conditional independence assumption on our global and marginal power results, we also calculated the results under a small set of simulations where the LD structure of the genotypes was taken directly from the *BRCA1* data in the WECARE Study. Since the correlation of most rare variants is negligible, we did not see any difference in the results from the correlated simulations and our simulations that assume conditional independence between the markers. Finally, unlike some of the commonly used methods, the Bayesian risk index has been developed within a generalized linear regression framework making extensions to many different data types feasible, such as quantitative, binary or survival outcomes. Likewise, the regression framework facilitates inclusion of covariates. Again, this has direct implications in the applicability of our method. One such example is in targeted rare variant studies where regions are investigated as followup to SNPs that have been detected in external studies involving common variants. In this case, the detected common SNP can be added to the analysis as a forced covariate and the Bayesian risk index methodology can be used to test if any of the rare variants in the region provide additional associations that are not captured by the common variant.

While we do aim to relax common assumptions of rare variant methods by incorporating uncertainty in many of the model parameters, like any analytical strategy the Bayesian risk index is not assumption free. Since our Bayesian risk index conditions on the observed distribution of alleles, we are making an explicit assumption about the form of the risk index and that this index is related to differences in the phenotype, not fitness. This is mainly a statistically motivated construction with the ability of detecting an association clearly dependent upon the appropriateness of this assumption to the true underlying disease model. By conditioning the risk index to include only rare variants, there is the additional implicit assumption regarding the distribution of the alleles in the population. Here, selection independent of the phenotype may impact the number of observed alleles and thus the variants that are included in the risk index, but should not impact case-control comparisons conditioned on those observed alleles. However, in regions in which selection is acting within a disease susceptibility region there is the potential for an increase in the frequency of disease alleles among cases and thus there may be a corresponding increase in power to detect differences [Garner, 2010; Pritchard, 2001; Pritchard and Cox, 2002].

The model developed herein uses a simple summation risk index. While this simplicity contributes to the gain in power from incorporating a variant selection strategy, extensions

may provide additional gains in power. For example, the approach can be easily extended to incorporate other prior specifications or risk indices. Currently, the marginal likelihood of each model is approximated by the likelihood function evaluated at the MLEs of the model-specific parameters. This corresponds to placing all of the prior mass of the model-specific parameters on the MLE. Although the marginal likelihood may not be available in closed form, it may be advantageous to place a noninformative prior distribution on the model-specific parameters and integrate over this distribution to obtain the marginal likelihood or use an Metropolis-Hastings algorithm to sample from it. Also, our model implicitly assumes that the magnitude of the effect of each of the variants included in the risk index is the same. Therefore, it is of interest to investigate how the incorporation of variant-specific weights to our Bayesian risk index may effect the power of our method. This may include approaches that weight by functions of the MAF (as in the weighted sum approach) or by incorporating biological information. In very recent work, Yi and Zhi [2011] propose a Bayesian analysis of rare variants in which the weights are estimated from the data by placing a prior distribution on both the rare variant load and the weights of each variant. Because Yi and Zhi do not allow for uncertainty into which variants are included in the risk index, their method only provides inference at the global level. Likewise, King et al. [2010] introduce a mixed-effects model where they assume that the variant-specific effects are a function of a wider population of variant effects. In particular, the effect of each variant on the phenotype is a linear function of the overall fitness effect of the variant (which is estimated using the variants observed frequency and other population genetic parameters). However, their method has been developed within the linear regression framework and has not yet been adapted for case-control studies. It is of future interest to investigate the power of Yi and Zhi's method as well as the method of King et al. with respect to the Bayesian risk index and with respect to an extension of our method in which variant-specific weights are incorporated. Furthermore, the current prior on the model space (or on the probability that each variant is included in the risk index) incorporates an implicit multiplicity correction in a way that is not highly informative. This prior can be extended naturally so that the inclusion probability of each variant is not assumed to be identical but rather weighted based on MAF and other external biological information on the variant. One could even potentially adapt the Bayesian model selection framework to alternative methods such as the C-alpha approach gaining power from both the reduction in variants included and from the particular type of association test implemented.

Throughout this article, we have demonstrated the power of the Bayesian risk index methodology with a focus on a candidate gene region. The proposed method can be naturally scaled up to incorporate multiple regions in which a risk index is calculated for each region and added to the regression framework, potentially with a model selection constructed in a hierarchical fashion. In theory, the proposed Bayesian risk index could also be used when the number of variants is larger than the number of individuals since we are estimating only one rare variant load for all of the markers that enter the risk index. Of course, in practice, as the number of variants under study increases the model space will increase exponentially and the computational limits for approaches that depend upon sampling procedures for inference will quickly reach their limits. The computation time of the Bayesian risk index methodology is a linear function of the number of iterations that it takes to reach convergence of the model search algorithm. Thus, in its current form, the Bayesian risk index algorithm took about 8 hr to run the WECARE Study example for 100,000 iterations on a single node of an AMD Opteron CPU running a Linux OS (CentOS 4). In comparison, the comp. step-up, weighted sum and C-alpha algorithms took 354.53, 60.52, and 0.296 sec, respectively, to run on the WECARE Study example. The computation time of the Bayesian risk index algorithm can be speed up significantly by running several smaller independent runs of the model search algorithm across multiple nodes of a CPU. Due to the computational intensity of the Bayesian risk index algorithm, we are currently in

the process of investigating more appropriate model search algorithms with the goal of reducing the computational time substantially and making whole genome sequence analysis feasible using the Bayesian risk index. Finally, functions implemented within the Bayesian risk index approach herein are available in an R package developed by the authors that is available upon request.

Acknowledgments

We thank the WECARE Study Collaborative Group (R01 CA097397 and U01 CA083178) for the example data and Paul Marjoram for many useful discussions.

Contract grant sponsor: NIEHS; Contract grant numbers: R01 ES016813; U01 ES015090; Contract grant sponsor: NHRGI; Contract grant number: U01 HG005927; Contract grant sponsor: WECARE Study Collaborative Group; Contract grant numbers: R01 CA097397; U01 CA083178.

REFERENCES

- Asimit J, Zeggini E. Rare variant association analysis methods for complex traits. *Annu Rev Genet.* 2010; 44:293–308. [PubMed: 21047260]
- Bansal V, Libiger O, Torkamani A, Schork NJ. Statistical analysis strategies for association studies involving rare variants. *Nat Rev Genet.* 2010; 11:773–785. [PubMed: 20940738]
- Begg CB, Haile RW, Borg A, Malone KE, Concannon P, Thomas DC, Langholz B, Bernstein L, Olsen JH, Lynch CF, Anton-Culver H, Capanu M, Liang X, Hummer AJ, Sima C, Bernstein JL. Variation of breast cancer risk among BRCA1/2 carriers. *JAMA.* 2008; 299:194–201. [PubMed: 18182601]
- Bernstein JL, Langholz B, Haile RW, Bernstein L, Thomas DC, Stovall M, Malone KE, Lynch CF, Olsen JH, Anton-Culver H, Shore RE, Boice JD, Berkowitz GS, Gatti RA, Teitelbaum SL, Smith SA, Rosenstein BS, Borresen-Dale A, Concannon P, Thompson WD. Study design: evaluating gene-environment interactions in the etiology of breast cancer- the WECARE study. *Breast Cancer Res.* 2004; 6:R199–R214. [PubMed: 15084244]
- Bhatia G, Bansal V, Harismendy O, Schork NJ, Topol EJ, Frazer K, Bafna V. A covering method for detecting genetic associations between rare variants and common phenotypes. *PLoS Comput Biol.* 2010; 6:e1000954. [PubMed: 20976246]
- Borg A, Haile RW, Malone KE, Capanu M, Diep A, Törngren T, Teraoka S, Begg CB, Thomas DC, Concannon P, Mellekjaer L, Bernstein L, Tellhed L, Xue S, Olson ER, Liang X, Dolle J, Borresen-Dale A-L, Bernstein JL. Characterization of BRCA1 and BRCA2 deleterious mutations and variants of unknown clinical significance in unilateral and bilateral breast cancer: the WECARE study. *Hum Mutat.* 2010; 31:E1200–E1240. [PubMed: 20104584]
- Capanu M, Begg CB. Hierarchical modeling for estimating relative risks of rare genetic variants: properties of the pseudo-likelihood method. *Biometrics.* 2011; 67:371–380. [PubMed: 20707869]
- Capanu M, Orlow I, Berwick M, Hummer AJ, Thomas DC, Begg CB. The use of hierarchical models for estimating relative risks of individual genetic variants: an application to a study of melanoma. *Stat Med.* 2008; 27:1973–1992. [PubMed: 18335566]
- Capanu M, Concannon P, Haile RW, Bernstein L, Malone KE, Lynch CF, Liang X, Teraoka SN, Diep AT, Thomas DC, Bernstein JL, The WECARE Study Collaborative Group, Begg CB. Assessment of rare BRCA1 and BRCA2 variants of unknown significance using hierarchical modeling. *Genet Epidemiol.* 2011; 35:389–397. [PubMed: 21520273]
- Conti DV, Gauderman J. SNPs, haplotypes, and model selection in a candidate gene region: the SIMPLE analysis for multilocus data. *Genet Epidemiol.* 2004; 27:429–442. [PubMed: 15543635]
- Garner C. A statistical method for scanning the genome for regions with rare disease alleles. *Genet Epidemiol.* 2010; 34:386–395. [PubMed: 20568275]
- Han F, Pan W. A data-adaptive sum test for disease association with multiple common or rare variants. *Hum Hered.* 2010; 70:42–54. [PubMed: 20413981]
- Hindorf LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *PNAS.* 2009; 106:9362–9367. [PubMed: 19474294]

- Hoffmann TJ, Marini NJ, Witte JS. Comprehensive approach to analyzing rare genetic variants. *PLoS ONE*. 2010; 5:e13584. [PubMed: 21072163]
- Jeffreys, H. *Theory of Probability*. 3rd edition. Oxford Univ. Press; Oxford: 1961. ISBN 0-19-850368-7
- Kass RE, Raftery AE. Bayes factors. *J Am Stat Assoc*. 1995; 9:773–795.
- King CR, Rathouz PJ, Nicolae DL. An evolutionary framework for association testing in resequencing studies. *PLoS Genet*. 2010; 6:e1001202. [PubMed: 21085648]
- Li B, Leal SM. Methods for detecting associations with rare variants for common disease: application to analysis of sequence data. *Am J Hum Genet*. 2008; 83:311–321. [PubMed: 18691683]
- Madsen E, Browning SR. A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet*. 2009; 5:e1000384. [PubMed: 19214210]
- Malone KE, Begg CB, Haile RW, Borg A, Concannon P, Tellhed L, Xue S, Teraoka S, Bernstein L, Capanu M, Reiner AS, Riedel ER, Thomas DC, Mellekjær L, Lynch CF, Boice JD, Anton-Culver H, Bernstein JL. Population-based study of the risk of second primary contralateral breast cancer associated with carrying a mutation in BRCA1 or BRCA2. *J Clin Oncol*. 2010; 28:2404–2410. [PubMed: 20368571]
- Morgenthaler S, Thilly WG. A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (CAST). *Mutat Res*. 2007; 615:28–56. [PubMed: 17101154]
- Morris AP, Zeggini E. An evaluation of statistical approaches to rare variant analysis in genetic association studies. *Genet Epidemiol*. 2010; 34:188–193. [PubMed: 19810025]
- Neale BM, Rivas MA, Voight BF, Altshuler D, Devlin B, Orho-Melander M, Kathiresan S, Purcell SM, Roeder K, Daly MJ. Testing for an unusual distribution of rare variants. *PLoS Genet*. 2011; 7:e1001322. [PubMed: 21408211]
- Price AL, Kryukov GV, de Bakker PI, Purcell SM, Staples J, Wei L-J, Sunyaev SR. Pooled association tests for rare variants in exon-resequencing studies. *Am J Hum Genet*. 2010; 86:832–838. [PubMed: 20471002]
- Pritchard JK. Are rare variants responsible for susceptibility to complex diseases? *Am J Hum Genet*. 2001; 69:124–137. [PubMed: 11404818]
- Pritchard JK, Cox NJ. The allelic architecture of human disease genes: common disease-common variant ... or not? *Hum Mol Genet*. 2002; 11:2417–2423. [PubMed: 12351577]
- Wilson MA, Iversen ES, Clyde MA, Schmidler SC, Schildkraut JM. Bayesian model search and multilevel inference for SNP association studies. *Ann Appl Stat*. 2010; 4:1342–1364. [PubMed: 21179394]
- Yi N, Zhi D. Bayesian analysis of rare variants in genetic association studies. *Genet Epidemiol*. 2011; 35:57–69. [PubMed: 21181897]

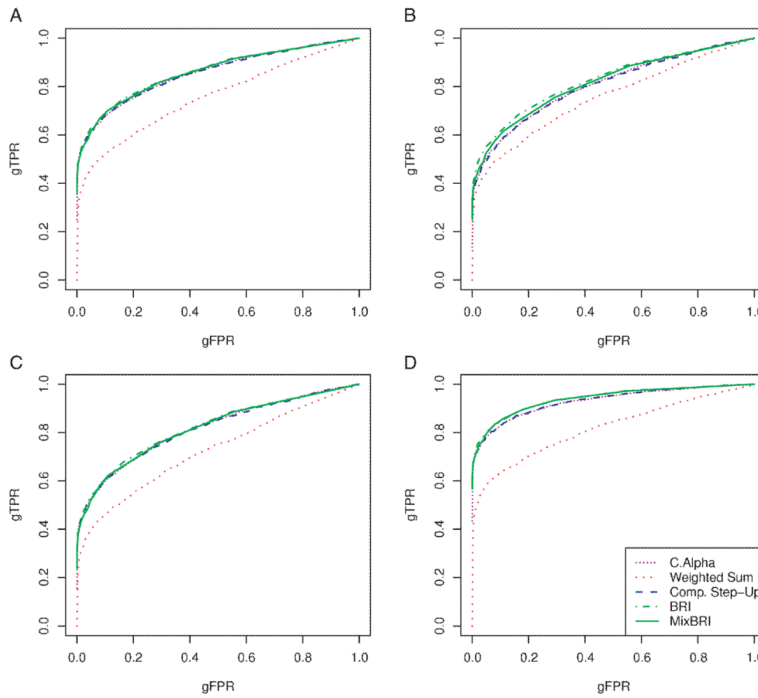


Fig. 1. Global ROC curves for Bayesian risk index vs. competing methods: small simulations (10 total variants). Each plot represents a ROC curve as we vary the Global BF threshold for the Bayesian risk index and the p -value for the weighted sum, C-alpha, and comp. step-up methods. For the Bayesian risk index, the solid line is calculated using the $\gamma \in \{0, 1\}$ model parametrization (BRI) and the dashed line is calculated under the $\gamma \in \{-1, 0, 1\}$ model parametrization (MixBRI). The gTPR and gFPR are calculated under all the small simulations in plot A, a subset of the nonmixed small simulations in plot B, a subset of the small simulations with OR<10 in plot C, and a subset of the small simulations with OR>10 in plot D.

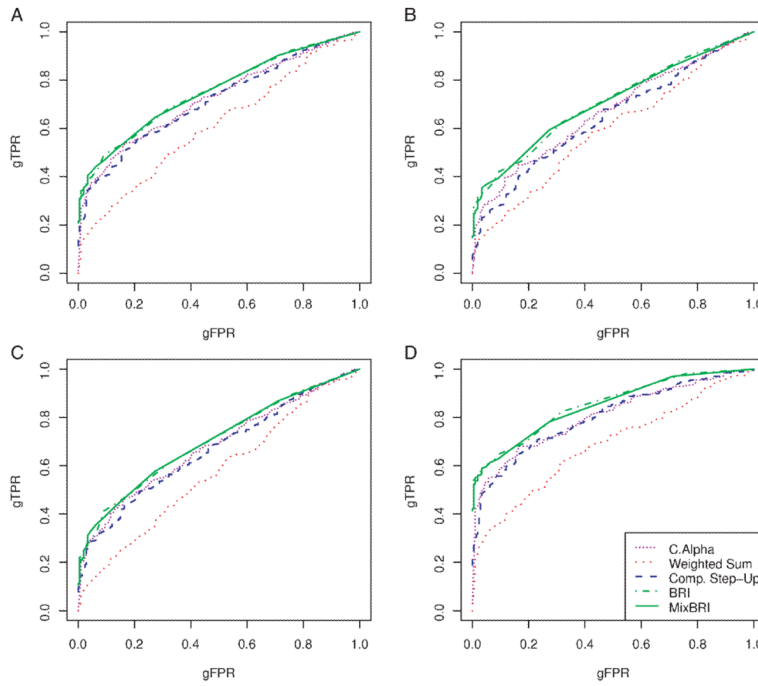


Fig. 2. Global ROC curves for Bayesian risk index vs. competing methods: large simulations. Each plot represents a ROC curve as we vary the Global BF threshold for the Bayesian risk index and the p -value for the weighted sum, C-alpha, and comp. step-up methods. For the Bayesian risk index, the solid line is calculated using the $\gamma \in \{, 0, 1\}$ model parametrization (BRI) and the dashed line is calculated under the $\gamma \in \{-1, 0, 1\}$ model parametrization (MixBRI). The gTPR and gFPR are calculated under all of the large simulations in plot A, a subset of the nonmixed large simulations in plot B, a subset of the large simulations with $OR < 10$ in plot C, and a subset of the large simulations with $OR > 10$ in plot D.

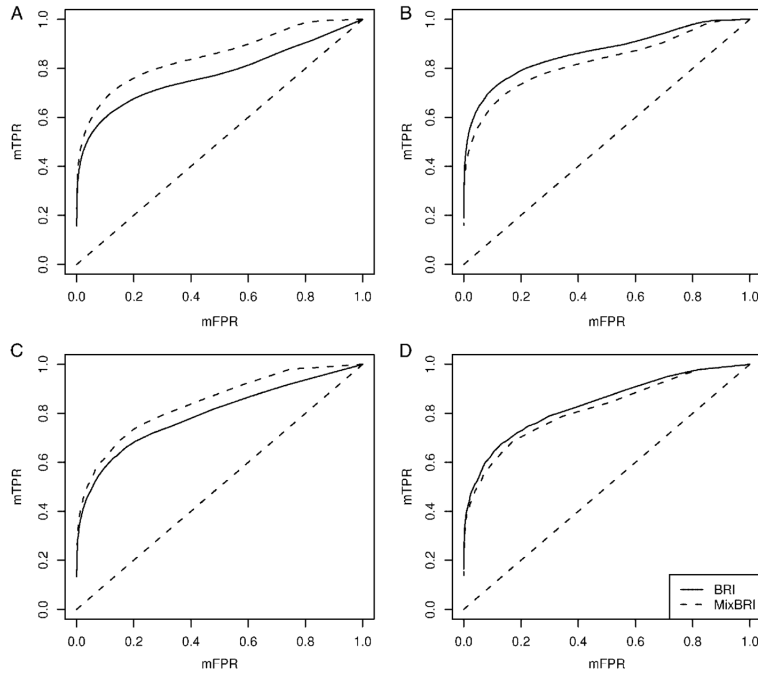


Fig. 3. Marginal ROC curves for Bayesian risk index. Each plot represents a ROC curve as we vary the Global BF threshold for the Bayesian risk index. For the Bayesian risk index, the solid line is calculated using the $\gamma \in \{, 0, 1\}$ model parametrization (BRI) and the dashed line is calculated using the $\gamma \in \{-1, 0, 1\}$ model parametrization (MixBRI). The plots on the left represent all simulations and the plots on the right represent only the simulations with non-mixed effects. The top plots are calculated for the set of small simulations with 10 total variants of interest and the bottom plots are calculated for the set of large simulations with 100 total variants of interest.

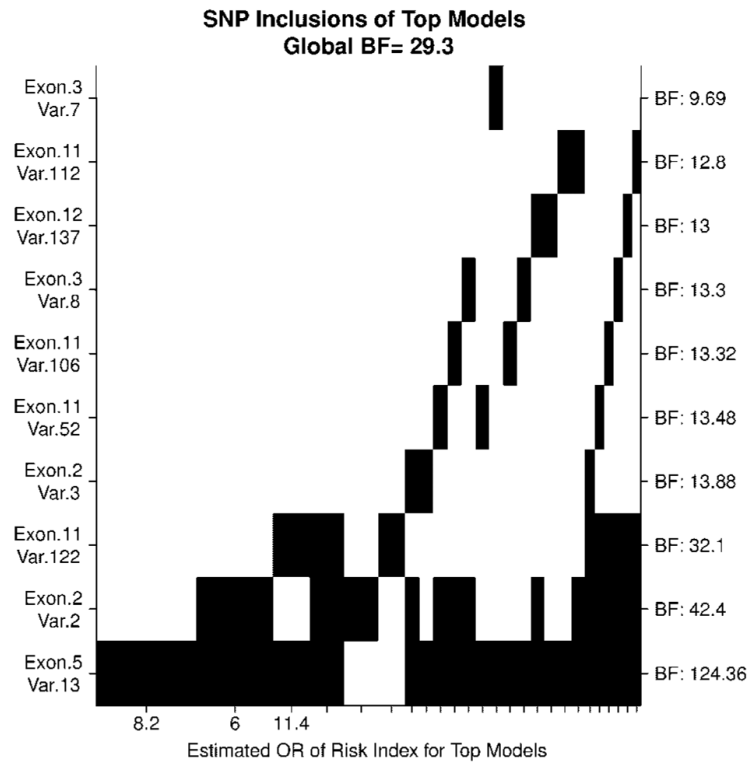


Fig. 4. Top Model Inclusions for Top Variants in *BRCA1*. Image plot of the rare variants included in the top 25 Models. The inclusion/exclusion of the top 10 variants are plotted in the rows and the rows are ordered based on each variants marginal Bayes factor (reported on the right axis). The variants included/excluded within each of the top 25 models are plotted in the columns. These columns have width proportional to and are ordered based on the posterior probability of the corresponding model. Estimated ORs for the model specific rare variant load are reported below each of the top three models.

TABLE IGeneral prior characteristics of the $\text{Bin}(p, 1/2)$ and $\text{BB}(1, p)$ distribution on model size

	Binomial ($p, 1/2$)	Beta-Binomial ($1, p$)
Expected model size	$p/2$	$p/(p+1)$
Global prior assoc.: $p(H_A)$	$1 - 1/2^p$	$1/2$
Marginal prior inclusion: $p(\gamma_j = 0)$	$1/2$	$1/(p+1)$

TABLE II

Jeffrey's grades of evidence [Jeffreys, 1961, p 432]

Grade	$BF(H_A:H_0)$	Evidence against H_0
1	1–3.2	Indeterminate
2	3.2–10	Positive
3	10–31.6	Strong
4	31.6–100	Very strong
5	>100	Decisive

TABLE III

Estimated gTPR given a fixed gFPR of 0.2 for the Bayesian risk index vs. competing methods as a function of MAF and number of true marginal associations (M_{assoc}) in the simulation

MAF: M_{assoc}	0.001			0.005			0.01		
	1-3	4-7	8-10	1-3	4-7	8-10	1-3	4-7	8-10
WeightSum (p -value <0.097)	0.26	0.42	0.58	0.39	0.67	0.84	0.52	0.80	0.90
C-alpha (p -value <0.232)	0.35	0.51	0.70	0.66	0.85	0.93	0.83	0.94	0.97
Comp. Step-Up (p -value <0.215)	0.34	0.53	0.71	0.65	0.86	0.93	0.82	0.94	0.98
BRI (BF>3.6)	0.37	0.54	0.72	0.70	0.86	0.93	0.85	0.94	0.98
MixBRI(BF>3.7)	0.36	0.52	0.70	0.70	0.85	0.93	0.85	0.94	0.97

MAF, minor allele frequency; BF, Bayes factor; gFPR, global false-positive rate; gTPR, global true-positive rate.

TABLE IVInformation on the top 10 variants within *BRCA1* calculated under the Bayesian risk index methodology

Exon: variant	Variant name	Base change	Amino acid change	Deleterious	Effect	Cases/controls	Marg. BF
E5:V13	C61G	T to G	Cys to Gly	Yes	M	8/2	124.4
E2:V2	185delAG	del AG	Stop 39	Yes	T	9/4	42.4
E11:V122	3875del4	del GTCT	Stop 1262	Yes	T	3/0	32.1
E2:V3	188insAG	insAG	Stop 31	Yes	T	2/0	13.88
E11:V52	1675delA	del A	Stop 531	Yes	T	2/0	13.48
E11:V106	E1107X	G to T	Glu to Stop	Yes	T	2/0	13.32
E3:V8	C44S	T to A	Cys to Ser	Yes	M	2/0	13.3
E12:V137	IVS12+37delTG	del TG	NA	No	IVS	2/0	13.0
E11:V112	3600del11	delGAAGATACTAG	Stop 1163	Yes	T	2/0	12.8
E3:V7	K38K	G to A	Lys to Lys	No	S	2/0	9.7

For the type of effect change the possibilities are missense (M), truncation (T), intervening sequence (IVS), or synonymous (S).