UCSF
UC San Francisco Previously Published Works

Title

Joint Association Testing of Common and Rare Genetic Variants Using Hierarchical Modeling

Permalink
https://escholarship.org/uc/item/580246hw

Journal
Genetic Epidemiology, 36(6)

ISSN
0741-0395

Authors
Cardin, Niall J
Mefford, Joel A
Witte, John S

Publication Date
2012-09-01

DOI
10.1002/gepi.21659

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

NIH Public Access
Author Manuscript
Genet Epidemiol. Author manuscript; available in PMC 2015 February 24.

Published in final edited form as:
Genet Epidemiol. 2012 September ; 36(6): 642–651. doi:10.1002/gepi.21659.

Joint Association Testing of Common and Rare Genetic Variants Using Hierarchical Modeling

Niall J. Cardin, **Joel A. Mefford**, and **John S. Witte**[*]
Department of Epidemiology and Biostatistics, Institute for Human Genetics, University of California, San Francisco, California

Abstract

New sequencing technologies provide an opportunity for assessing the impact of rare and common variants on complex diseases. Several methods have been developed for evaluating rare variants, many of which use weighted collapsing to combine rare variants. Some approaches require arbitrary frequency thresholds below which to collapse alleles, and most assume that effect sizes for each collapsed variant are either the same or a function of minor allele frequency. Some methods also further assume that all rare variants are deleterious rather than protective. We expect that such assumptions will not hold in general, and as a result performance of these tests will be adversely affected. We propose a hierarchical model, implemented in the new program CHARM, to detect the joint signal from rare and common variants within a genomic region while properly accounting for linkage disequilibrium between variants. Our model explores the scale, rather than the center of the odds ratio distribution, allowing for both causative and protective effects. We use cross-validation to assess the evidence for association in a region. We use model averaging to widen the range of disease models under which we will have good power. To assess this approach, we simulate data under a range of disease models with effects at common and/or rare variants. Overall, our method had more power than other well-known rare variant approaches; it performed well when either only rare, or only common variants were causal, and better than other approaches when both common and rare variants contributed to disease.

Keywords

cross-validation; multi-SNP; gene based

INTRODUCTION

Complex diseases are likely to arise from both common and rare genetic variants. For the latter, recent work has developed new analysis methods that help address the issue of data sparsity. However, using methods designed only for rare variants may miss important signals. Running separate methods for rare and common variants can lead to problems of

© 2012 Wiley Periodicals, Inc.

[*]Correspondence to: John S. Witte, Department of Epidemiology and Biostatistics, Institute for Human Genetics, Helen Diller Family Comprehensive Cancer Center, University of California, San Francisco, 1450 3rd Street, Room 388, San Francisco, CA 94158-9001. jwitte@ucsf.edu.

Supporting Information is available in the online issue at wileyonlinelibrary.com.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

interpretation, especially when rare variants tested with one method are in linkage disequilibrium (LD) with common variants tested by another.

Many methods exist for finding disease associations with rare variants. Morgenthaler and Thilly [2007] suggest collecting variants (e.g., the variants from the exons and splice sites of genes) and then comparing the counts of cases and controls that carry at least one mutation in that group of variants. Morris and Zeggini [2010] take the sum of the mutations in each individual, rather than the indicator of presence or absence of mutation. When using such approaches, removing all variants above a certain minor allele frequency (MAF) threshold is necessary, so that the signal is not overwhelmed by common variants. However, no frequency threshold is perfect. If the threshold is too high, the signal from rare variants will be swamped by noise from common variants; while if the threshold is too low, true causal variants may be missed.

To overcome this problem, Madsen and Browning [2009] propose collapsing all variants, but weighting rare variants more highly than common variants. This method, designed for continuous traits, assumes that rare variants have stronger effects than common variants. If the causal variants are not rare then this method will have reduced power compared to more agnostic methods. Furthermore, it is unknown how much larger the effect sizes of rare variants will be compared to common variants. If the true effects do not correspond to the weighting scheme used, then the weights may decrease power compared to the unweighted approach.

Alternatively, Price et al. [2010] suggest exploring all frequency thresholds and choosing the threshold that gives the most power. Unfortunately even in this case, the signal from causal rare variants may be drowned out by variants with no effect on disease. To focus specifically on causal variants, Hoffmann et al. [2010] propose collapsing over variants that show signal. The step-up procedure successively adds the variant that most improves the *P*-value of the collapsed score. By summing alleles over a set of variants, all of the methods described so far make the implicit assumption that the effect sizes are equal or have some specific relationship with MAF.

Li and Leal [2008] suggested jointly testing rare and common variants by collapsing variants below an MAF threshold and then testing the collapsed variants and remaining common variants, using Hotelling's *T*-squared test. By thresholding and collapsing rare variants, this method suffers from many of the problems discussed above, though it does allow for evaluating common variants above the rare variant threshold.

Any collapsing method must address the fact that variants are not inherited independently. LD invalidates the assumption of independence used to calculate the null distribution of test statistics. One might account for LD using permutation testing, producing a valid test. However, multiple noncausal variants in LD might mask causal signals.

To address these limitations of existing methods, we propose a novel approach called Cross-Validated Hierarchical Aggregated Regression Method, or CHARM. CHARM simultaneously analyzes all variants within a region, while allowing different effect sizes for each variant and accounting for LD. We use a multiple logistic regression model with a

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

hierarchical structure that captures the joint signal from multiple effects within a region. CHARM avoids the need for collapsing variants by using a likelihood-based approach to assess the overall evidence of association within a region. We use cross-validation to provide a computationally efficient means of comparing our multiple regression with the null model of no genetic association. CHARM allows each variant to have a different effect size but we do not perform variable selection within a genomic region. Instead, we fit a joint model with separate effect sizes for each single nucleotide polymorphism (SNP), drawn from a smooth prior distribution. To capture the considerable uncertainty regarding likely genetic effects, we average over multiple models, i.e. prior distributions of effect sizes. While previous methods have used hierarchical models for genetic association studies, e.g. [Chen and Witte, 2007; Hung et al., 2007; Heron et al., 2011], these methods used the hierarchy as a means of incorporating prior information on the location or likely locations of associations. Here, the hierarchical structure is used as a means of providing a richer model and pooling the evidence for association across multiple SNPs.

We compare CHARM to existing methods using simulated data reflecting a range of disease models. In each of the scenarios, the best performing approach is one of three: CHARM, min-p, or Step Up. Uniquely, CHARM has either the most power, or at least 90% of the power of the best performing method in all simulated scenarios.

METHODS

OUTLINE

Our method jointly tests all of the SNPs in a region for association with a disease or other dichotomous trait. The underlying model is a multiple logistic regression, with a separate odds ratio for each SNP. We place independent prior distributions on the effect size of each SNP, the properties of which are controlled by hyperparameters. To increase flexibility, we average over multiple disease models, and inference is performed by comparing the likelihood under this model to the likelihood under the null. To overcome the computational difficulty in calculations of marginal likelihood, we use a likelihood based on cross-validation.

LOGISTIC REGRESSION

Assume we type *S* SNPs (rare and common) in *N* individuals:

$$\begin{aligned} logit\left(P\left(\mathrm{Y}=1|G,\boldsymbol{\beta}\right)\right) &= \boldsymbol{\beta}_0+G\boldsymbol{\beta} \\ &= \boldsymbol{\beta}_0+G_1\boldsymbol{\beta}_1+G_2\boldsymbol{\beta}_2+\ldots+G_s\boldsymbol{\beta}_s \end{aligned} \quad (1)$$

where *Y* denotes the phenotype, and *G* denotes a matrix of genotypes, G_i being the data at the *i*th SNP, (the G_i are coded additively). Finally, β is a vector of logistic regression coefficients (log odds ratios) for disease, where the *i*th element corresponds to the *i*th SNP.

HIERARCHY

Inference of individual odds ratios can be challenging under such a model. Sparsity at rare variants, and LD at common variants, can lead to great uncertainty in effect estimates. Nevertheless, the joint distribution of odds ratios may indicate strong evidence for

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

association in the region as a whole and we assess this using a hierarchical model (see, e.g. Gelman and Hill, 2007). To assess this, we place a hierarchical prior on the log odds ratios, β, in this model, see Figure 1. We then compare the likelihood (using cross-validation) of the data under the hierarchical model with the likelihood under the null.

We use a prior distribution of effect sizes centered on zero, but with a hierarchical parameter that controls the degree to which effect sizes can vary from zero. In particular:

$$\boldsymbol{\pi}\left(\boldsymbol{\beta}_{\mathbf{i}}\right) \sim \mathbf{F}\left(\text{center}=\mathbf{0}, \quad \text{scale}=\boldsymbol{\delta}\right) \quad (2)$$

for $i \geq 1$. In this paper, the distribution, F , indicates either a Cauchy or normal distribution and we ultimately average over models that use either of these. For normally distributed log odds ratios, the parametrization uses a mean and variance. For the purposes of clarity, we refer to the term scale only in the rest of the paper. The value of β_0 is described in the supplementary materials.

The scale parameter, δ, controls the strength of effects that may be observed in the region of interest. While one might place a prior on the center parameter, we prefer to focus inference on the scale, δ, because controlling the scale provides the flexibility to allow for causative and protective effects. For example, if one SNP has an log odds ratio equal to 0.3 and another equal to −0.3, then both provide evidence of deviation from the null model. In contrast, if one only controls the mean effect size in a region then, given a certain SNP coding, only multiple effects that point in the same direction will provide evidence for the alternative model.

We include a Cauchy prior on effect sizes for the following two overlapping reasons. First, a Cauchy distribution better reflects our expectation that most variants will have modest effect sizes. Indeed, it is implausible, given current genome wide association studies (GWAS) results, that multiple common SNPs would independently have strong effects . We would like to have most effect sizes close to zero while allowing for effect sizes relatively far from zero; a Cauchy distribution fits this description better than a normal distribution. Second, given a causal SNP, we have little idea of what effect sizes to expect, especially at rare SNPs. With much heavier tails, the Cauchy distribution better reflects that uncertainty than a normal distribution.

MODEL AVERAGING

One of the primary motivations for our approach is the uncertainty about the genetic architecture of complex traits. Possibilities for a given region range from clusters of very rare and large-effect variants, through collections of weaker effect variants across all frequencies, to a single common causal variant. It therefore makes sense to account for a range of possibilities when modeling the effect of our genetics on disease.

In the presence of uncertainty, the natural approach is to integrate out our uncertainty. While the full space of disease models, including interactive effects, pathways, etc. is too large a space to explore at this stage, we can still account for a range of possibilities. For computational reasons, we pick a fixed set of plausible disease submodels and then take the

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

average of the likelihoods (equation [8]) from these submodels to assess the evidence for association. Each sub-model consists of a different prior on the effect size of each SNP, and we include all the SNPs from the predefined region in the submodel.

MODEL SPECIFICS

Each submodel has a different prior on the log odds ratios of the SNPs in the region. Given Cauchy (or normally) distributed log odds ratios, the distribution is specified by the scale (or variance) parameter of the distribution. We specify this scale in terms of numbers of SNPs at certain relative risks. This is meant to ease interpretation of these priors on log odds ratios, and to facilitate user specification of new models. An expected number of causal SNPs, *n*, and a relative risk, *r*, are chosen. The scale of the distribution is set to give an expected *n* SNPs with either a relative risk of at least *r*, or less than $\frac{1}{r}$, symmetrically. The values of *n* and *r* used by default are given in Table I.

Under some models (Table I), the effect size priors are a function of MAF. There are many choices for how to achieve this, and it is not possible to derive a “best” method without making assumptions about the unknown disease models that we hope to uncover. We describe one possible approach here. We first note that the main reason why rare variants should have stronger effects is the action of natural selection. However, for common diseases, selection cannot act effectively at very rare SNPs, and the frequency of a causative allele will be dominated by drift. We therefore allow only a moderate increase in the spread of the effect size distribution at rarer SNPs and below a MAF of ~0.002 we do not further increase this.

First, a canonical frequency, f_c, is chosen (this was set to 25%). Variants at this frequency use the scale defined for models with no MAF dependence. Call the scale given to canonical frequency variants s_c, this corresponds to a median absolute relative risk of r_c. Then, the prior is set so that the median absolute relative risk for another variant is

$$1+(r_c-1)\left(\frac{f_c}{f}\right)^{1/2} \quad (3)$$

where *f* is the frequency of that variant. A maximum multiplier, $\left(\frac{f_c}{f}\right)^{1/2}$, of 12 is also imposed due to the argument given above, although this is unlikely to have a substantial effect on results.

For very small regions, this prior on effect sizes would be very wide. As a result, the scale of the prior on effect sizes for any SNP is capped at 0.5, which gives a median absolute odds ratio of 1.6.

INFERENCE

Given the data in a genomic region, we are interested here in evaluating evidence for at least one disease associated variant in this region. This contrasts with the conventional approach of testing each variant for association. Suppose, for now, that we have some fixed value of δ. We want to compare the model described above, M_1, with the null model M_0. Under the null

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

model, the case phenotype, *Y*, is drawn from *N* independent and identical Bernoulli distributions where the probability of being a case, $P(Y = 1)$, is the proportion of cases in the data.

In principle, we could then calculate a Bayes Factor

$$BF=\frac{P\left(\mathrm{Y}|G,M_1\right)}{P\left(\mathrm{Y}|G,M_0\right)} \quad (4)$$

where

$$P\left(\mathrm{Y}|G,M_1\right)=\int P\left(\mathrm{Y}|G,\boldsymbol{\delta},\boldsymbol{\beta}\right)P\left(\boldsymbol{\beta}|\boldsymbol{\delta}\right)d\boldsymbol{\beta} \quad (5)$$

Unfortunately, this becomes computationally difficult or intractable for large numbers of SNPs, even using numerical or stochastic methods such as Markov Chain Monte Carlo (MCMC).

We might also want to place a prior on δ, and integrate over that prior under M_1. However, this also entails a high computational cost. Thus, we explore a range of fixed values, as explained in the section on model averaging.

USING CROSS-VALIDATION

To perform inference, we compare how well the model of genetic effect on disease compares to a model of no genetic effect. One might use maximum likelihood-based methods, but these have the problem that more complex models get higher likelihoods than less complex models. This problem becomes much more severe when multiple SNPs are jointly analyzed. Marginal likelihoods are “honest,” in that model complexity confers no inherent advantage, but these are difficult to compute. There are many methods that can be used to approximate marginal likelihoods; unfortunately most of these do not respond appropriately to changes in the degree of parameter shrinkage, but rather to the number of parameters in the model. Other approaches, such as those based on MCMC give the correct answer in principle, but are unstable and it can be difficult to know when they have converged.

Nevertheless, likelihoods are a powerful tool when performing inference, so we use a likelihood based on cross-validation (see, e.g. Hastie et al., 2009). The motivation is that good models will accurately predict outcomes in new samples, and models that overfit the data will less accurately predict outcomes in new samples. We describe a statistic motivated by Bayes Factors, but which is computationally tractable for large data sets.

We randomly partition the sample into $\mathscr{J}$ groups. We then take each of these groups in turn, calculate the maximum a posteriori estimate of β given δ on those individuals not in that group (training set). Then, we calculate the probability of sampling the case status of the remaining individuals (test set) from the predictions of that model. We repeat this until each group has been picked once, hence every individual has been included in the test set exactly once. The procedure is stochastic, so we perform this procedure multiple times and combine results. Let L_c^i denote the product of these individuals’ predicted sampling probabilities in

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

the *i*th iteration of cross-validation (the *c* subscript denotes that this is a likelihood derived using cross-validation, rather than more familiar maximization or marginalization):

$$log\left(L_c^i\right)=\sum_{j=1}^{\mathscr{J}}log\,P\left(\mathrm{Y}_i^j|G_i^j,M_i^j\right) \tag{6}$$

where $G_{i\,j}$ and $Y_{i\,j}$, respectively, denote the genotypes and outcomes for test set *j*, and $M_{i\,j}$ denotes the model fit acquired from training set *j*. We use $\mathscr{J}=10$ fold cross-validation that is computationally reasonable and allows the model to be fitted on 90% of the data.

We then take the average of the log-likelihoods over *k* replicates of this procedure:

$$log\left(L_c\right)=\frac{1}{k}\sum_{i=1}^{k}log\left(L_c^i\right) \tag{7}$$

which we found performed well.

Although the cross-validated likelihoods, L_c, are distinct from the commonly used marginal and maximized likelihoods, for ease of exposition, we will refer to them as “likelihoods” hereafter. We could have different likelihoods under different submodels, let L_c^m denote the likelihoods for a specific submodel *m*, and let L_c^M denote the likelihood averaged over these submodels, then:

$$log\left(L_c^M\right)=log\left(\frac{1}{\mathscr{M}}\sum_{m=1}^{\mathscr{M}}L_c^m\right). \tag{8}$$

We use these likelihoods to compute a cross-validation-based analog to the Bayes Factor. That is

$$log\left(BF_c\right)=log\left(L_c^M\right)-log\left(L_N\right) \tag{9}$$

where L_N denotes the likelihood under the null model, M_0.

Unfortunately, as this overall procedure does not produce a true Bayes Factor, we cannot perform full Bayesian inference without further research into the properties of this approach. Instead, the BF_c values can be treated as test statistics and permutation methods can then be used to calculate a *P*-value for the BF_c statistic. For whole-genome analyses, or with a large number of regions, *i*, we can be more efficient. By permuting the case-control labels across the whole-genome *j* times, we can produce a set of *i j* approximate Bayes Factors under the null. The Bayes Factors for the original data can then be compared with these to assess the signal. The null distribution of Bayes Factors varies slightly by region size, so greater accuracy can be achieved by stratifying according to region size. In the simulation study section, we assess the performance of CHARM using this approach.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

SIMULATION STUDY

We compared the proposed hierarchical model with four existing approaches using simulated case-control data under a range of disease models. The existing approaches were min-p, Step Up [Hoffmann et al., 2010], variable MAF (VMAF) thresholding [Price et al., 2010], and combined multivariate collapsing (CMC) [Li and Leal, 2008].

THE GENOTYPES

We used a simulated set of 10,000 haplotypes over a stretch of 4.7 mega bases (Mb) of chromosome 17. These data were produced using the program SFS CODE [Hernandez, 2008] using parameters derived from the Seattle SNPs project. All SNPs are simulated, taking into account empirical results about the patterns of diversity as a function of estimated recombination rates, genes, and other genetic annotations (see [Hernandez, 2008]). To simulate 10,000 haplotypes over 4.7 Mb with SFS CODE is computationally intensive, and it is not currently possible to use it to produce whole-genome data on this scale. The 10,000 haplotypes were taken in pairs to produce 5,000 diploid individuals.

Using the UCSC Genome Browser, we downloaded the positions of genes in this region and a nonoverlapping subset was used to define 15 genic regions (including introns). In addition, we defined 23 arbitrary nongenic regions, fixing the number of SNPs in each to represent a range of region sizes and placing them evenly across the 4.7 Mb simulated region.

The majority of SNPs in this data were rare, with 31.4% of variants being singletons or doubletons, and ~80.3% of all variants having an MAF less than 1%. Table II summarizes the number of SNPs at different minor allele frequencies in each of the 38 regions used in the simulation study. The third column gives the total number of SNPs among the 5,000 simulated individuals; the other columns give the average number of remaining SNPs after the cases and controls were sampled, and singletons removed. Removing singletons has little effect on power under our models and is a substantial computational saving. Columns 4–7 give the number of SNPs structured by minor allele frequency. The genes are numbered in the order that they appear in the genome, and ordered by size.

DISEASE MODELS

The disease models used by CHARM allow for relatively straightforward inference; however, reality may be somewhat different. For these simulations, we use a different disease model structure and different specific disease models from those used by default in CHARM. Our disease simulations use a more flexible model, with specific causal SNPs having random effect sizes, and remaining SNPs having zero effect. In total, a set of eight different disease models were used to explore the properties of the different methods (see Table III). These were designed to cover a range of relative contributions from rare and common variants.

The framework is that of logistic regression with additive noninteracting effects. The variants were categorized as *rare* (MAF < 1%), *uncommon* (1% ≤ MAF < 5%), and common (MAF > 5%). In each region, a fixed number of each category of variants were chosen uniformly at random, without replacement, to be causal (see Table III).

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

The baseline probability of disease was 0.01, and the effect sizes for the causal SNPs were chosen according to the MAF categorization (see Table III). Effect sizes were partially stochastic, with an attempt to reflect plausible values given previous studies. For a given disease model, each causal variant has an effect size independently assigned in the following way:

(1) Take the relative risk, *r*, associated with this allele frequency category and disease model from Table III.

(2) Simulate a relative risk of

$$\Gamma\left(\text{mean}=r-1,\ var=\frac{r}{a}\right)+1$$

(where *a* is a predefined constant).

(3) With probability 0.5 take the reciprocal of this effect size, giving it a protective effect.

The values of *a* used were 4, 2.5, and 1.5 for common, uncommon, and rare variants, respectively. One exception to this is disease model *B*, where the relative risks were fixed at exactly 1.5. The purpose of this disease model was to favor simple collapsing tests by using many causal SNPs each with exactly the same effect size.

For each of 38 regions, and eight disease models, 30 replicates were simulated by resampling causative SNPs, effect sizes, and case-control labels. This gives 9,120 data sets under alternative models. Under the null, 100 replicates of each region were simulated, for 3,800 null data sets. In each replicate, 1,000 cases and 1,000 controls were drawn from the population of 5,000 with replacement, according to the probability of case status. We restricted to 30 replicates under the alternative models for computational reasons.

Note that somewhat weak effects have been chosen here, so no method will have close to complete power under these scenarios. This may well be realistic, and is also practically useful as some of the methods require permutation testing, limiting our ability to calculate relative power for strong effects. For example, to discern the difference between *P*-values of 10^{-4} and 10^{-5} would require more than 10^5 permutations. To address this, power was calculated at the 0.01 significance level, and the models were designed to give a power of roughly 0.5 on each data set, to maximize our ability to calculate relative power between the methods.

COMPARISON WITH EXISTING METHODS

The simplest approach to testing for association, even with rare variants, is to test each SNP individually. Thus far, there has been implicit agreement that such an approach would provide little power when the signal is comprised of clusters of rare variants. However, we are unaware of any previous publications that examine this hypothesis, and thus we include min-p to compare its performance with collapsing methods.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

There is a complication with the single SNP approach, in that it provides multiple test statistics for each region. We take the simple approach of using the smallest *P*-value as the statistic for each region. This makes the results of the single SNP directly comparable with collapsing methods. To calculate a single *P*-value, we compare this smallest *P*-value to a null distribution generated by permuting the case-control labels.

Hoffmann et al. [2010] propose a step-wise collapsing method. The SNP, G_i, with the smallest *P*-value is taken as a starting point, giving the score at each individual of $X_j = G_{ji}$. The step-wise procedure then tries adding each remaining SNP to X, the model is fit with this new collapsed variable, and the SNP that gives the smallest *P*-value is added. SNPs are added until no SNP that decreases the *P*-value can be found. We try a version with a uniform weighting on each SNP, and also one weighted by allele frequency using the scheme proposed in [Madsen and Browning, 2009].

The VMAF threshold approach method sums the alleles from rare variants in each individual to create a single score which is then tested for association. The procedure explores all possible allele frequency thresholds and uses permutation to correct for the inflation in test statistics. This is one of the methods from Price et al. [2010], Price also suggests weighting variants according to MAF (see Madsen and Browning, 2009) and we implement both a weighted version and one where a uniform weighting is used.

The CMC method [Li and Leal, 2008] collapses variants below a certain threshold, but then performs a joint test of all remaining common variants and the collapsed signal from the rare variants. Here, we adjust this method slightly to suit our disease model more closely. First, we use an “additive” version, so the collapsed score, X_j, in each individual, j, has the form

$$X_j = \sum_i G_{ji} \quad (10)$$

summed over SNP genotypes G_i. This is in contrast to the indicator variable in the original formulation [Li and Leal, 2008] that denotes presence or absence of a rare variant in the region. CMC uses a Hotelling's T^2 test, and when the genotype matrix is singular this fails to give a test statistic. We attempted to solve this iteratively, by numerically estimating the rank of the matrix, and filtering SNPs at with lower r^2 values, starting from 0.95; unfortunately this did not always succeed. Furthermore, because CMC often fails to run, it was very hard to create an accurate permutation test. Instead, the values of the un-permuted test statistic on a set of null data sets is used, and this empirical null distribution (stratified by region size) was used to calculate power at various significance levels.

For all of the rare variant tests that do not allow for protective effects, we first change the sign of variants with more rare alleles in the controls. This does not invalidate the test as this procedure is also used when creating the empirical null distributions of test statistics.

To increase computational efficiency, we use the empirical null distribution of test statistics for all of the methods from the simulated null data sets. Then, the proportion of test results that are greater than $(1 - \alpha)$ of the results from the matched null data sets are defined to be significant at the α level.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

RESULTS

To assess the power (at α = 0.01) of CHARM relative to other approaches, we ran a simulation study using a range of disease models (see Table III). We stratify results by the size of the region (small, medium, large) to give a total of 24 different scenarios. We compare CHARM to min-p, VMAF thresholding, with and without weighting by MAF [Price et al., 2010], Step Up, with and without weighting by MAF [Hoffmann et al., 2010], and CMC [Li and Leal, 2008] (see Methods for details).

In each of the scenarios, the best performing approach is one of three: CHARM, min-p, or in one case, unweighted Step Up. Overall CHARM has either the highest, or close to the highest power in all simulated scenarios. Absolute power is low by design, so the focus here is on relative power (Fig. 2 and Table IV).

CHARM has a high power relative to the other methods in each of the 24 scenarios. Under models *A* and *B* (Table III), only rare variants (MAF < 1%) are causal and the power of CHARM is at least 95% of that of any other method. Under models *C* to *F* , the causal contribution from uncommon (1% ≤ MAF < 5%) and common (MAF ≥ 5%) variants increases, and CHARM is consistently the most powerful method. Under model *G*, where all causal variants have an MAF ≥ 1% min-p is just as powerful as CHARM. Finally, under model *H*, where all causal variants have an MAF ≥ 5%, min-p is the most powerful, though CHARM has at least 90% of the power of min-p in all cases, even when only a single common variant is causal. CHARM is also less affected by the number of SNPs in a region than the collapsing methods and is always more powerful than collapsing methods for medium and large regions.

Min-p has the most power when only common variants are causal. However, relative power decreases for small regions when rare variants play a substantial role, with 82– 90% of the power of CHARM under models *A* to *F* . Min-p also had more power than the collapsing methods for all scenarios with medium and large regions.

We tried versions of Step Up both weighted by MAF and unweighted. Step Up weighted by MAF had consistently lower power than unweighted Step Up (see Table IV). Un-weighted Step Up has the highest power under model *A*and with small regions (fewer than 30 SNPs). Unweighted Step Up also has the highest overall power among all collapsing methods. However, unweighted Step Up is particularly sensitive to the number of SNPs in a region and has less than 69% of the power of CHARM for large regions (more than 100 SNPs). Also, the power of unweighted Step Up decreases progressively as the contribution from rare variants diminishes, with 21–89% of the power of CHARM under models *G* and *H*.

We also ran VMAF both weighted by MAF and un-weighted. Weighted VMAF showed patterns similar to those of Step Up, though weighted VMAF had a lower power than Step Up overall. Unweighted VMAF performed better under models with no causal rare variants, but power was low when rare variants were causal, with only 3–61% of the power of CHARM under models *A* to *F* . The power of VMAF also decreases more than CHARM as region size increases, for large regions (at least 100 SNPs) power was 3 –76% and 38–57% of that of CHARM for unweighted and weighted VMAF, respectively.

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

CMC did not perform well in this study. The power of CMC was less than 80% of the power of CHARM in all cases and less than 71% for medium regions. When applied to large regions, Hotelling's test failed due to singular matrices for hundreds of data sets and this was not solved by our LD filtering approach (see Methods). The test also reported 77 *P*-values less than 10^{-6} under the null, due to this and the number of data sets with no result we cannot report reliable power estimates for large regions.

To assess the sensitivity of CHARM to the specific models used, we also repeated the above analyses leaving each of the models given in Table I out in turn. Power relative to CHARM with all models included, averaged across all regions and disease models, lay between 0.96 and 0.99. In the worst case, when model ii was not included, the relative power of the reduced model was on average 0.96, with a range for the relative power from 0.88 to 1.02 in analyses of particular genomic regions.

In the current *R* implementation of CHARM, with the simulated data presented here, a single core from a 2.26GHz Quad-Core Intel Xeon processor took approximately 1 min per SNP with regions between 20 and 100 SNPs long. The average time for regions with an average of 43 SNPs was ~ 47 sec per SNP and regions with an average of 300 SNPs took ~ 108 sec per SNP. Thus, under these settings and using the current implementation, a genome of 5 million SNPs and with 2,000 individuals can be analyzed in approximately a week on a cluster of 500 nodes.

DISCUSSION

We introduce a hierarchical modeling method for detecting common and rare genetic associations. In our simulation study, the method always had close to as much, and often more, power than the other approaches evaluated here. Having a generally powerful approach is important. The next generation of genome-wide association studies will investigate rarer variants and we will soon have association scans using whole exome chips and resequencing, and even whole-genome resequencing. CHARM allows us to make powerful inference about associations in the presence of clustering of rare variant signals while maintaining power when associations come from common variants.

The power of min-p was unexpectedly high compared to existing rare variant methods, especially when testing more than 30 SNPs at once. This likely reflects the fact that min-p focuses only on the strongest association, whereas the other rare variant methods collapse over causal and null associations, and may exclude more common causal variants.

However, CHARM did often outperform the min-p approach. This is probably because CHARM accounts for LD between SNPs and multiple effects in a single region. By jointly modeling the SNPs, the method avoids double counting causal SNPs due to correlations with linked markers. However, when multiple causal SNPs truly are present, CHARM can gain power—both by detecting the signal of multiple effects and because accounting for the effect of one causal SNP reduces noise when estimating the effect of others. As an illustration of this, it is worth considering the following two distinct scenarios: (1) three SNPs in LD with only one causal variant, and (2) three SNPs not in LD when all three are

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

causal (with the same effect size). Methods, such as min-p or collapsing methods, do not distinguish between these quite different cases. CHARM will see a considerably bigger signal in the latter case as the data will explain a considerably higher proportion of disease here than with only one causal variant.

As with any multiparameter method, our hierarchical model requires the specification of a reasonable set of tuning parameters. Here, we specified values based on expected effect sizes from previous studies and adjusted these using results from early and simple test data sets. The models mostly use Cauchy distributed log odds ratios; here it was clear from an early stage that Cauchy distributions performed better than normal distributions. However, performance was better still when averaging over a larger set of models, including a contribution from models with normally distributed log odds ratios (data not shown). There may be concerns that CHARM is sensitive to parameter choice, in particular that it may be overly tuned to the setting described here. However, we note that the disease models used in the simulations were constructed differently from the model used in CHARM. Furthermore, CHARM was run using the same settings for different region sizes and a range of disease models, and performed well in all of these cases. This demonstrates that CHARM is not sensitive to being highly tuned for a particular situation.

There are a number of improvements that we hope to make to CHARM, and release in future versions of the software. At present, the method does not automatically account for missing data and even with imputation there may still be significant uncertainty in some individuals, especially at rare SNPs. When the patterns of missing data are confounded with case-control status then almost any test will suffer from bias and inflated test statistics. When the proportion of missing data is low, and is not thought to be confounded with case-control status, then we suggest using the “expected” genotype. The model does not require integer genetic types and this reflects a reasonable approximation to integration over possible genotypes.

CHARM does not yet implement any form of pathway analysis, or explicitly allow models with increased probability/strength of association at predicted functional sites (or e.g. nonsynonymous sites). These are natural developments and due to the model averaging approach these can be implemented, and hence tested for, in a coherent framework. This avoids complications with multiple tests that have similar hypotheses, which could be important when marginally significant findings arise.

Covariates can naturally be taken into account in this framework. Additionally, if there is existing information about the effect of these covariates then this can easily be incorporated by placing priors on their effect sizes. Genetic interactions can, in principle be accounted for, though the computational cost of including all pairwise interactions would, in the present framework, be computationally intractable for regions with much more than ~50 SNPs.

Our approach can be extended from binary phenotypes to continuous phenotypes. The current method uses the probability of predicting exactly the right outcome. With continuous phenotypes, the density function could be used, and the product of densities could be

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

compared to the analogous product under the null model. However, this would then require careful thought as to the best null model, and its relationship with the alternative. For example, a normal error distribution around the sample mean could perform worse than a more flexible noise model under the alternative, even in the absence of a genetic effect.

CHARM assesses the evidence for association between a genomic region and disease. It is possible to run CHARM genome wide by breaking the genome up into multiple regions. However, it is not clear how to best break the genome outside genes. Although in exon resequencing studies, the genome is naturally broken into genes, in whole genome analyses, there is the question of how to define the regions between genes. We suggest an approach of tiling the genome with overlapping regions, e.g. as in [Su et al., 2009]. While this approach complicates interpretation as tests are not independent, this is no more severe than interpreting the familiar nonindependence due to LD of single SNP tests in a standard genome-wide association study.

We do not expect that larger genes are more likely to be associated with disease, or to have larger total effects. Our prior places the same weight on the probability of some effect in each gene. However, this scheme produces tighter priors, hence greater shrinkage, on individual SNPs in larger regions. As we do not know what disease models to expect this decision may turn out to be suboptimal. In principle, one could try multiple approaches and use further model averaging. However, there are many such possibilities, and one must be wary of attempting a complete exploration of such possibilities as the resulting exponential increase in the model space quickly leads to computationally intractable approaches.

In summary, the challenge of testing for associations in the presence of rare variants is upon us. We propose a flexible new method that will allow researchers to test for rare variant associations while maintaining power if common variants have a role in disease. Software for applying this method is available from the authors.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENTS

This work supported by National Institutes of Health grants CA088164 and U01 CA127298. We thank Ryan Hernandez for sharing simulated data.

REFERENCES

Chen GK, Witte JS. Enriching the analysis of genomewide association studies with hierarchical modeling. Am J Hum Genet. 2007; 81(2):397–404. [PubMed: 17668389]

Gelman, A.; Hill, J. Data Analysis Using Regression and Multilevel/Hierarchical Models. Vol. 3. Cambridge University Press; New York: 2007.

Hastie, T.; Tibshirani, R.; Friedman, J.; Franklin, J. The Elements of Statistical Learning: Data Mining, Inference and Prediction. 2nd. edition.. Springer; New York: 2009.

Hernandez RD. A flexible forward simulator for populations subject to selection and demography. Bioinformatics. 2008; 24(23):2786–2787. [PubMed: 18842601]

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Heron EA, O'dushlaine C, Segurado R, Gallagher L, Gill M. Exploration of empirical Bayes hierarchical modeling for the analysis of genome-wide association study data. Biostatistics. 2011; 12(3):445–461. 2011. [PubMed: 21252078]

Hoffmann TJ, Marini NJ, Witte JS. Comprehensive approach to analyzing rare genetic variants. PloS One. 2010; 5(11):124–137.

Hung RJ, Baragatti M, Thomas D, McKay J, Szeszenia-Dabrowska N, Zaridze D, Lissowska J, Rudnai P, Fabianova E, Mates D, Foretova L, Janout V, Bencko V, Chabrier A, Moullan N, Canzian F, Hall J, Boffetta P, Brennan P. Inherited predisposition of lung cancer: a hierarchical modeling approach to dna repair and cell cycle control pathways. 2007. Cancer Epidemiol Biomarkers Prev. 2007; 16(12):2736–2744. [PubMed: 18086781]

Li B, Leal SM. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. Am J Hum Genet. 2008; 83(3):311–321. [PubMed: 18691683]

Madsen BE, Browning SR. A groupwise association test for rare mutations using a weighted sum statistic. PLoS Genet. 2009; 5(2):e1000384. [PubMed: 19214210]

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/Fundam Mol Mech Mutagen. 2007; 615(1-2):28–56.

Morris AP, Zeggini E. An evaluation of statistical approaches to rare variant analysis in genetic association studies. Genet Epidemiol. 2010; 34(2):188–193. [PubMed: 19810025]

Price AL, Kryukov GV, de Bakker PIW, Purcell SM, Staples J, Wei LJ, Sunyaev SR. Pooled association tests for rare variants in exonresequencing studies. Am J Hum Genet. 2010; 86(6):832–838. [PubMed: 20471002]

Su Z, Cardin NJ, The Wellcome Trust Case Control Consortium. Donnelly P, Marchini J. A Bayesian method for detecting and characterizing allelic heterogeneity and boosting signals in genome-wide association studies. Stat Sci. 2009; 24(4):430–450.

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

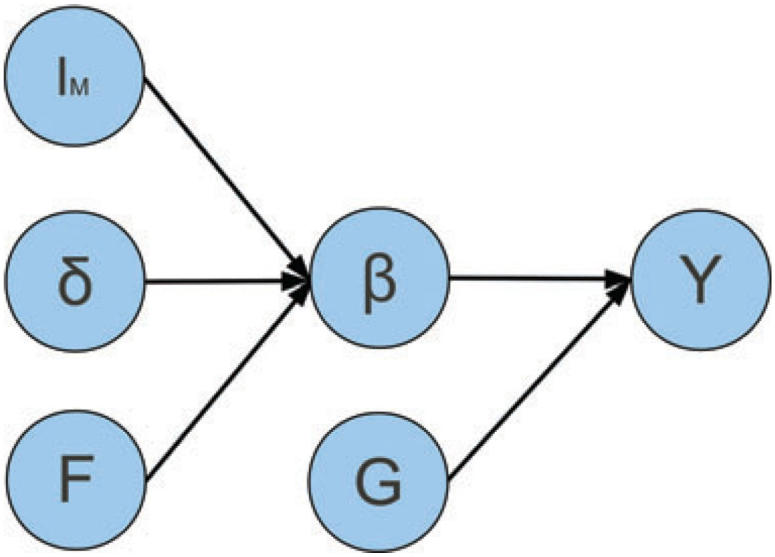

Fig. 1.

Outline of the hierarchical model structure: the middle column represents the logistic regression model. The effect sizes (log odds ratios), β, and the genotypes define the risk of disease for each individual. The parameter δ controls the prior distribution of effect sizes on the SNPs, β. Model averaging is used, so that multiple values of δ are explored. Also we explore whether effect sizes should depend on the MAF of the SNP, I_M, and whether to use a Cauchy or normal distribution, *F*.

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

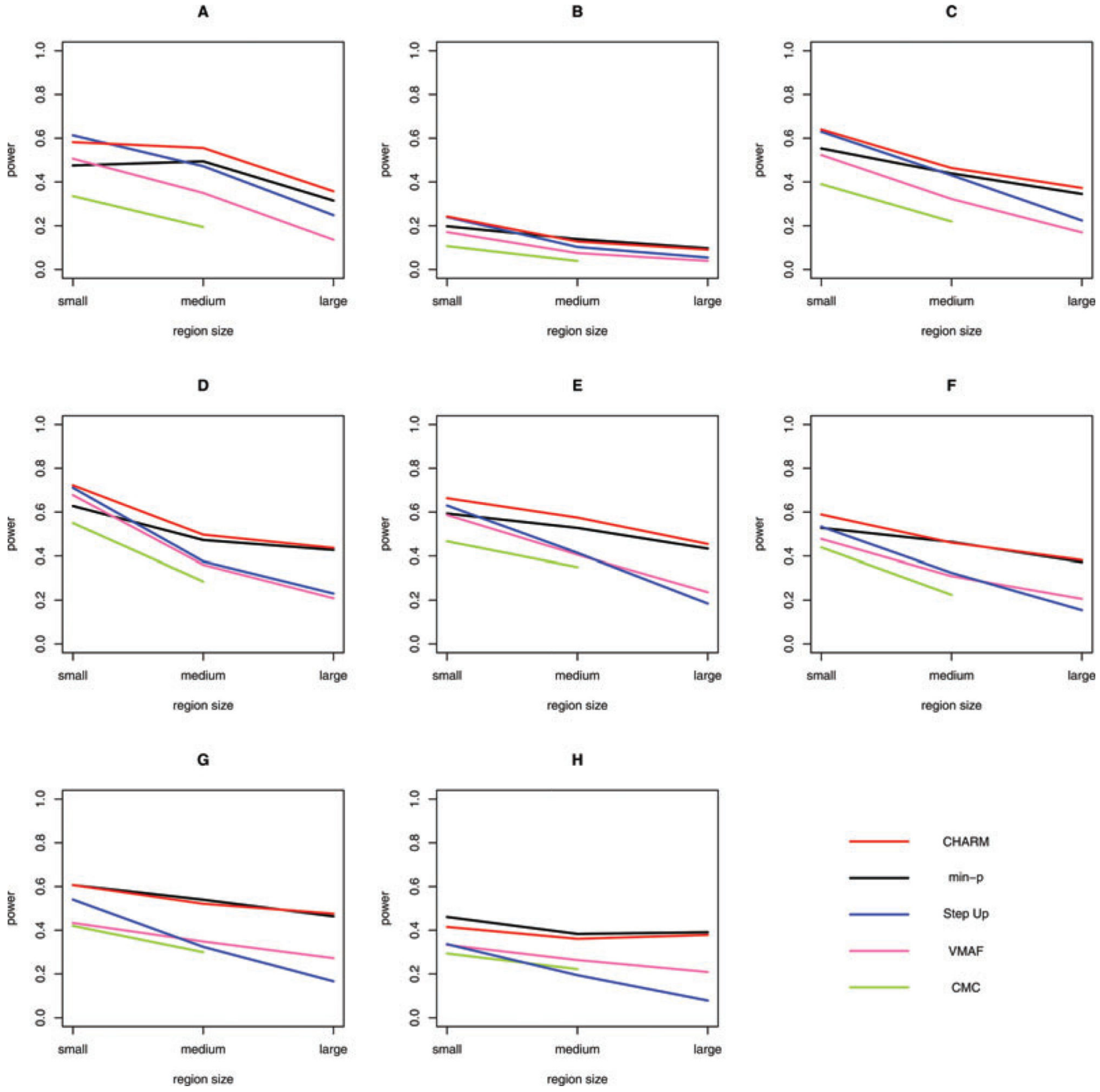

Fig. 2.

Power by region size and disease model: each panel gives the power of five methods under a specific disease model at a significance level of 0.01. The 38 regions of the simulation study were split into three categories, to highlight the different performance of methods with different region sizes. The category *small* denotes regions with fewer than 30 SNPs, *medium* represents regions with 30 100 SNPs, and *large* represents regions with more than 100 SNPs. Models *A* and *B* use only rare causal variants; models *C* through *H*−add progressively more uncommon and common variants, with the contribution from rare variants reduced until, under model *H*, all causal variants are common. See Table III for full disease model

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

details. As noted, the results for CMC were problematic for very large regions, so these are not shown here.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

TABLE I

Details of the seven models that CHARM averages over under the default settings. The models used here are chosen to reflect a reasonable range of effect sizes that we might encounter. For simplicity, each model shown here is given equal weight, though variable model weights are possible in CHARM

Model	n: number causal	*r*: relative risk	Normal or Cauchy	MAF dependent
i	2	1.2	Cauchy	No
ii	4	1.5	Cauchy	No
iii	8	3	Cauchy	No
iv	3	1.5	normal	No
v	6	2	normal	No
vi	2	1.1	Cauchy	Yes
vii	6	1.3	Cauchy	Yes

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

TABLE II

Region type and SNPs at different minor allele frequencies for 38 loci used in simulation study.

Region number	Region type	No. of SNPs total	No. of SNPs used	MAF < 1%	1% ≤ MAF< 5%	MAF ≥ 5%
1	Intergenic	61	15.3	11.8	1.4	2
2	Intergenic	61	8	6.9	0.1	1
3	Intergenic	61	12.7	9.3	3.3	0
4	Intergenic	61	15.4	13.2	0.8	1.4
5	Intergenic	61	13.4	11.3	1.2	0.9
6	Intergenic	101	16.7	11.6	1.1	4
7	Intergenic	101	27.2	19.4	2.9	5
8	intergenic	101	22.1	13.9	4	4.3
9	Intergenic	101	26.6	19.4	3	4.2
10	Intergenic	101	26.6	18	2.4	6.2
11	Intergenic	201	39.9	29.7	3.3	6.8
12	Intergenic	201	45.4	29.4	6	10
13	Intergenic	201	55.6	40.6	7.1	7.9
14	Intergenic	201	44.3	30	2.3	12
15	Intergenic	201	40.7	33.4	5.1	2.2
16	Intergenic	301	86.4	55.3	16.6	14.4
17	Intergenic	301	88.3	63.3	3	22
18	Intergenic	301	76.2	54	8.2	14
19	Intergenic	301	62.8	42.2	12.5	8
20	Intergenic	301	65.2	35	15.2	15
21	Intergenic	601	150.3	86.5	41.1	22.6
22	Intergenic	601	133	88.1	12.4	32.5
23	Intergenic	601	135.4	96	14.5	25
14	Gene	37	12.2	7.2	2.6	2.3
16	Gene	37	9.5	5.4	1	3
20	Gene	47	11.6	7.8	0.7	3.1
4	Gene	90	26.4	20.4	1.3	4.7
25	Gene	106	24.5	17.2	3.2	4.1
11	Gene	259	68.9	47.4	11.3	10.2
24	Gene	339	81.6	52.3	5.2	24.2
5	Gene	484	108.2	65.5	10.7	32.1
13	Gene	531	118.5	74	31.9	12.5
6	Gene	680	115	85.8	17.2	12
23	Gene	712	133.5	88.9	8.5	36
18	Gene	1,069	287.1	178.4	40.6	68.2
21	Gene	1,090	244.8	171.2	21.9	51.7
19	Gene	1,328	302.7	213.6	31.2	57.9
15	Gene	1,401	346.2	230.9	54.7	60.7

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

TABLE III

Description of eight disease models in the simulation study. The columns give the number of causal variants in each allele frequency category, with the target odds ratio in parentheses.

Model	Common MAF ≥ 5% Num (OR)	Uncommon 1% ≤ MAF < 5% Num (OR)	Rare MAF < 1% Num (OR)
A	0	0	15 (3.06)
B[a]	0	0	20(1.51)
C	0	2 (1.35)	13 (2.54)
D	2 (1.09)	3 (1.15)	15 (2.54)
E	1 (1.2)	2 (1.3)	15 (2.28)
F	2 (1.1)	3 (1.15)	8 (2.54)
G	3 (1.15)	3 (1.3)	0
H	3 (1.15)	0	0

[a] Disease model *B* uses a relative risk of exactly 1.5 (hence an odds ratio of 1.51) for each causal variant. For all other models, simulated odds ratios are random and centered on the given values, as described (Methods).

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

TABLE IV

Power of the methods under a specific disease model at a significance level of 0.01. Table III gives the details of the disease models *A* through *H* for small, medium, and large numbers of SNPs. Weighted VMAF and Step Up are represented by VMAF_W and Step Up_W, respectively

Scenario	CHARM	Min-p	Step Up	Step Up_W	VMAF	VMAF_W	CMC
A small	0.58	0.48	0.61	0.55	0.17	0.51	0.36
A medium	0.56	0.49	0.47	0.38	0.12	0.35	0.23
A large	0.36	0.32	0.25	0.21	0.06	0.14	NA
B small	0.24	0.20	0.24	0.17	0.05	0.17	0.12
B medium	0.13	0.14	0.10	0.08	0.02	0.08	0.05
Blarge	0.09	0.10	0.06	0.05	0	0.04	NA
C small	0.64	0.55	0.63	0.58	0.29	0.52	0.42
C medium	0.46	0.44	0.43	0.36	0.14	0.32	0.26
C large	0.37	0.34	0.22	0.18	0.10	0.17	NA
D small	0.72	0.63	0.71	0.68	0.43	0.68	0.58
D medium	0.50	0.47	0.38	0.36	0.26	0.36	0.35
D large	0.44	0.43	0.23	0.18	0.20	0.21	NA
E small	0.66	0.59	0.63	0.58	0.39	0.59	0.50
E medium	0.57	0.53	0.41	0.39	0.31	0.41	0.42
E large	0.46	0.43	0.18	0.17	0.21	0.24	NA
F small	0.59	0.53	0.53	0.42	0.36	0.48	0.47
F medium	0.46	0.46	0.32	0.26	0.27	0.31	0.28
F large	0.38	0.37	0.16	0.14	0.19	0.21	NA
G small	0.61	0.61	0.54	0.49	0.45	0.43	0.45
G medium	0.52	0.54	0.32	0.24	0.39	0.35	0.37
G large	0.48	0.46	0.17	0.10	0.35	0.27	NA
H small	0.42	0.46	0.34	0.32	0.44	0.33	0.31
H medium	0.36	0.38	0.19	0.18	0.32	0.26	0.27
H large	0.38	0.39	0.08	0.07	0.29	0.21	NA