Published in final edited form as:

Genet Epidemiol. 2013 September ; 37(6): 529–538. doi:10.1002/gepi.21736.

Testing for Rare Variant Associations in the Presence of Missing Data

Paul L. Auer¹, Gao Wang², NHLBI Exome Sequencing Project³, and Suzanne M. Leal^{2,*}

¹Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, Washington

²Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, Texas

³NHLBI Exome Sequencing Project authorship list is included in the Supplement

Abstract

For studies of genetically complex diseases, many association methods have been developed to analyze rare variants. When variant calls are missing, naïve implementation of rare variant association (RVA) methods may lead to inflated type I error rates as well as a reduction in power. To overcome these problems, we developed extensions for four commonly used RVA tests. Data from the National Heart Lung and Blood Institute-Exome Sequencing Project were used to demonstrate that missing variant calls can lead to increased false-positive rates and that the extended RVA methods control type I error without reducing power. We suggest a combined strategy of data filtering based on variant and sample level missing genotypes along with implementation of these extended RVA tests.

Keywords

rare variant association studies; next-generation sequencing; complex disease

Introduction

Rare variant association (RVA) studies of complex traits using both whole exome (WE) and whole genome (WG) sequencing data have become feasible with the advent of next-generation sequencing (NGS) technology. As these data have become widely available, many association methods have been developed specifically to analyze rare variants [e.g., those with minor allele frequency (MAF) <1%]. Rather than considering variants individually, as is typically done in array-based genome-wide association studies (GWAS), these methods aggregate variants across a specified region, usually a gene or transcript [Li

© 2013 Wiley Periodicals, Inc.

*Correspondence to: Suzanne M. Leal, Department of Molecular and Human Genetics, Baylor College of Medicine, One Baylor Plaza, 7000D, Houston, TX 77030. sleal@bcm.edu.

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

Web Resources

The URLs for data presented herein are as follows:

NHLBI Exome Sequencing Project Exome Variant Server, <http://evs.gs.washington.edu/EVS/>

Online Mendelian Inheritance in Man (OMIM), <http://www.ncbi.nlm.nih.gov/omim>

and Leal, 2008; Lin and Tang, 2011; Liu and Leal, 2010; Madsen and Browning, 2009; Morris and Zeggini, 2010; Price et al., 2010; Wu et al., 2011], and are often referred to as “aggregate,” “collapsing,” or “burden” tests. In contrast to single variant association tests, aggregate RVA tests encounter unique problems when there are missing variant calls (i.e., missing genotypes). For studies of disease associations with rare genetic variants, there may be a bias in the rate of missing genotypes between cases and controls (i.e., differential missing data). In such instances, RVA tests may suffer from both a decrease in power as well as an inflated type I error rate. In contrast, when individual variants are analyzed, missing genotypes only reduce sample size, leading to a reduction in power but not an increase in type I error.

In the context of WE and WG sequencing studies, differential missing variant calls may occur in a number of ways. Anytime cases and controls are processed differently, there exists the potential for confounding case-control status with certain batch effects. This can be particularly problematic if convenience controls are used from public repositories such as the Database of Genotypes and Phenotypes (dbGAP) [Mailman et al., 2007] or The European Genome-Phenome Archive (EGA) [Leinonen et al., 2011]. Such batch effects include, but are not limited to differences in: input DNA quality, library preparation, exome capture arrays, read length, depth of coverage, and sequencing machines. Although a well-thought-out experimental design would avoid many of these issues, cost constraints and convenience often preclude an optimal design. Accordingly, the statistical methods used to test for RVAs should be reasonably robust to biases imposed by a suboptimal design. Unfortunately, obtaining empirical *P*-values through permutations will not solve the problem of inflated type I error when there are differential missing genotypes. It was therefore necessary to develop RVA methods that adequately control type I error in the presence of differential missing data.

We examined type I error and power in the presence of missing variant calls for four commonly used RVA methods: combined multivariate and collapsing (CMC) [Li and Leal, 2008], weighted sum statistic (WSS) [Lin and Tang, 2011; Madsen and Browning, 2009], variable threshold (VT) [Price et al., 2010], and the burden of rare variants (BRV) which is our modified version of gene- or region-based analysis of variants of intermediate and low (GRANVIL) frequency [Morris and Zeggini, 2010]. When analyzing dichotomous traits (e.g., case-control status), all four aggregate methods properly control type I error when the frequency of missing genotypes is equivalent between cases and controls. However, substantial increases in type I error occur when there is differential missing genotype calls (between cases and controls) at the variant sites being aggregated. The extent of the increase in type I error is dependent on the overall percent of missing data, the number of variants being aggregated, the difference in missing rates between cases and controls, and the type of RVA method that is used. Although type I error can be controlled for all methods by removing any variant site that has a missing genotype, this procedure can lead to a substantial loss of power.

In order to control type I error without sacrificing power, we developed extensions for the CMC, WSS, BRV, and VT RVA methods. The increase in type I error that we observed before extending the RVA methods is due to the fact that all four tests implicitly assume that

a missing genotype is homozygous for the common allele. Rather than make this assumption, we substitute an appropriate probability of the presence of the minor allele for each missing genotype. For all four tests, these probabilities are based on the site specific allele frequencies across all samples, regardless of phenotypic (e.g., case-control) status. This is a specific application of “mean imputation,” which has been extensively studied in missing data problems [Little and Rubin, 2002]. For the BRV method, a “dosage” [Zheng et al., 2011] score is substituted for the missing genotype. The “dosage” for a variant site is obtained by taking the average minor allele count. The “burden score” for each individual is the sum of the dosages (at missing variant sites) plus the sum of the observed minor allele counts. The WSS is extended in a similar fashion with weights derived from each variant’s MAF. For the CMC, all individuals are categorized by whether they carry a rare variant, regardless if heterozygous, homozygous, or compound heterozygous. The extension of the CMC method provides an estimate of the probability that an individual carries a rare variant. This value is simply 1 for individuals that carry a rare variant or 0 for those individuals that have no rare variants and have no missing data and between 0 and 1 for all other individuals with missing data. The VT method, which computes a maximum test statistic taken across all observed variant frequencies, is extended using either the CMC or BRV coding as described above.

Using data from the NHLBI-ESP (National Heart Lung and Blood Institute-Exome Sequencing Project), we conducted extensive simulations and analyses using the four original methods and their extensions (i.e., CMC-M, BRV-M, WSS-M, and VT-M). In the presence of differential missing data, we show that the extended RVA methods effectively control type I error without a loss of power.

Results

Type I Error Simulations

We considered a variety of situations where missing genotype calls could be relevant to RVA testing, in that type I error is inflated or power is reduced. Data were simulated using observed variant frequencies from the NHLBI-ESP set of 3,510 European Americans (EAs). In order to assess the effect of the number of variant sites, two genes [*MC4R* (melanocortin 4 receptor, MIM 155541) and *ALK* (anaplastic lymphoma kinase, MIM 105590)] of different sizes (1.4 kb and 728.8 kb) and number of variant sites (18 and 54) were selected to perform data simulation.

As proof of principle to demonstrate that random missing data do not inflate type I error rates, we generated missing genotypes completely at random with respect to case-control status. For both *MC4R* and *ALK* genes, both the original and extended versions of each of the four tests properly controlled type I error when missing data are random (data not shown). However, when missing data were generated according to case-control status type I error rates were inflated beyond their nominal levels. For instance, in the *ALK* gene, at missing rates of 0% in cases and 20% in controls (i.e., an average missing rate of 10%), the original RVA methods demonstrate inflated type I error rates while the extended methods do not (Fig. 1, Table 1). The results are similar for the *MC4R* gene although the type I error inflation is not as severe (Table 1). The *MC4R* gene contains fewer variant sites with a lower

cumulative allele frequency compared to the *ALK* gene, thus it contains less information to detect an association (in this case, a false-positive association). Type I error rates are still inflated for all the original RVA tests even when missing rates are decreased in controls and increased in cases (Supplementary Fig. S1).

To determine whether a filter on missing genotypes would mitigate some of the observed inflation, individuals missing 80% of their variant calls (within gene) and variants sites missing 10% of their genotype calls were removed. For settings with an overall missing rate of 10%, this filter removed most variant sites from the analysis. For the simulation settings with lower rates of missing calls, this quality control (QC) step helped reduce the inflation in type I error. Although all the extended methods properly controlled type I error (Fig. 1, Supplementary Fig. S1); even after data filtering, the BRV, WSS, and VT still displayed inflation of type I error.

These simulations demonstrate that the extended RVA tests (i.e., CMC-M, BRV-M, WSS-M, and VT-M), along with a filtering strategy that removes samples missing 80% of their variant sites for a specific gene and variants missing 10% of their genotype calls, effectively control type I error rates in the presence of differential missing data.

Power Simulations

We evaluated the statistical power of the extended RVA tests and compared them to the original versions. In order to fairly compare methods, we examined power for those simulation settings that did not lead to inflation in type I error. When missing data were randomly distributed across cases and controls, all methods control type I error (data not shown). When there are no missing data, the extended and original versions of the RVA tests are equivalent and so is the power to detect an association. When data are either not missing or missing completely at random, the extended tests have similar power to the original versions (Supplementary Figs. S2 and S3). For differential missing data, we compared methods when 0% of the cases and 10% of the controls were missing variant calls. As shown in Table 1, the extended versions properly control type I error at the 0.05 level. However, even after filtering, the original methods do not control the type I error. In order to control type I error for the original tests, we removed from analysis every variant missing any genotypes. This ended up removing most of the variants from the analysis and the power for these tests suffered accordingly. From the results displayed in Figure 2, it is clear that the extended RVA tests along with a modest pervariant and per-sample filter provide a powerful alternative to simply removing sites with missing calls and performing RVA analysis using CMC, BRV, WSS, or VT.

Data Analysis

To evaluate the performance of these methods on exome sequence data, we analyzed data from the NHLBI-ESP. We assigned 1,000 EA samples that had been processed using the Agilent array to have “case” status and 1,000 EA samples that were captured using the Roche Nimblegen array to have “control” status. We are interested in evaluating false positives and not detecting true associations. Therefore, instead of analyzing disease phenotypes where true genetic associations may exist, we analyzed a dataset for which there

were differential missing rates between cases and controls that were due to the different capture arrays. We removed related and duplicate samples and performed a standard variant-level QC. Finally, we removed any genes from the analysis that were only captured on one array or were located in processed pseudo-genes, large segmental duplications, or copy number variants. Missing rates between cases and controls are shown in Supplementary Figure S4.

An exome-wide case-control analysis was performed using each of the original and extended tests before and after filtering. In principle, the assigned “phenotypes” should not be associated with any of the genotypes, thus providing an opportunity to observe how well these tests control type I error in an exome sequence dataset. Figure 3 clearly demonstrates the superiority of the extended RVA testing methods in terms of type I error control. For the BRV, CMC, and WSS tests, the extensions control type I error without filtering. The VT-M requires filtering in order to effectively control type I error. For RVA testing, this analysis establishes the need for a combined approach to properly control type I error: implementation of the extended methods (CMC-M, BRV-M, WSS-M, VT-M) as well as enforcing missing data filters at both the sample and variant level.

Methods

Simulation Framework

We based our simulations on the empirical distributions of rare and low frequency ($MAF < 5\%$) nonsynonymous variants (missense and nonsense) found in the NHLBI-ESP Exome Variant Server. Two genes were selected, *MC4R* and *ALK*, whose variant frequency spectrum within the EA population represents a small- and medium-sized gene. *MC4R* contains 18 nonsynonymous variant sites that are observed between 1 and 112 times in the 7,020 EA chromosomes. Within the same EA individuals, the *ALK* gene contains 54 nonsynonymous variant sites observed between 1 and 260 times. MAFs which range from 0.00014 to 0.037 were used to generate genotypes for 2,000 individuals based on the Hardy-Weinberg proportions, assuming independence between variant sites.

To evaluate type I error, we generated phenotypes by assigning case status to 1,000 individuals and control status to an additional 1,000 individuals completely at random. For the simulations evaluating power to detect an association, we considered 25%, 50%, 75%, and 100% of the nonsynonymous variant sites to be causal [i.e., odds ratio (OR) > 1]. Fixed and variable effect models were used to determine the effect size of the causal nonsynonymous variant sites. For the fixed effect model, each causal variant was assigned an $OR = 3.0$. For the variable effect model, variant sites with an allele frequency > 0.01 were assigned an $OR_{\min} = 2$, variant sites with the lowest frequencies (i.e., 0.00014) were assigned an $OR_{\max} = 10$, and all variant sites with intermediate allele frequencies were assigned an OR by interpolation between $OR_{\min}$ and $OR_{\max}$. To generate phenotypes, we drew 100,000 Bernoulli (p_i) trials, where

$$p_i = \frac{\exp\left(\sum_j \log(OR_j) D_j G_{ij}\right)}{1 + \exp\left(\sum_j \log(OR_j) D_j G_{ij}\right)},$$

OR_j is the odds ratio for variant j , D_j is 1 if variant j is causal and 0 otherwise, and G_{ij} is the minor allele count of variant j in the i th sample. A total of 1,000 cases and 1,000 controls were generated for each replicate.

CMC-M: Extension of the CMC Approach

The CMC approach tests the association between phenotype and status as a carrier of a rare variant. Formally, let $\delta_j = 1$ if SNP_j has $MAF \geq T$, and 0 otherwise. For sample i , carrier status is represented with the following variable

$$X_i^{CMC} = I\left[\left\{\sum_j G_{ij} \delta_j\right\} > 0\right].$$

This approach implicitly treats missing genotypes as $G_{ij} = 0$ (i.e., a homozygote for the major allele). Rather than make this assumption, we have extended the CMC to model the probability of carrying a rare variant for subjects with missing genotypes who would otherwise be categorized as noncarriers. To do so, we let $A = \{i = 1, \dots, n, \text{ such that } X_i^{CMC} = 0 \text{ and } G_{ij} \text{ is missing for some } j\}$. This is the set of samples that have at least one missing genotype and are not otherwise considered carriers of the rare allele. Considering the set of SNPs meeting the MAF threshold with missing genotypes for sample i [$B_i = \{j, \text{ such that } G_{ij} \text{ is missing for } SNP_j \text{ in sample } i \text{ and } \delta_j = 1\}$], we then calculate a new independent variable

$$\tilde{X}_i^{CMC} = \begin{cases} X_i^{CMC}, & i \notin A \\ P_i^{CMC}, & i \in A \end{cases}$$

where

$$P_i^{CMC} = 1 - \left[\prod_{j \in B_i} (1 - MAF_j)^2 \right]$$

is the probability that individual i is a carrier of the minor allele at the variant sites missing genotype calls.

BRV-M: Extension of GRANVIL

The GRANVIL aggregate test is similar to the CMC except that carrier status at each variant site is summed across the gene and a ratio is formed where the denominators are the total

number of sites for which a genotype call is available. The BRV is quite similar to GRANVIL except that heterozygous sites contribute a single count while homozygous sites contribute two counts to the total sum of variants within a gene region and no denominator is used. We developed and implemented the BRV, whose power is equivalent or slightly superior to the GRANVIL, for two reasons: (1) the GRANVIL denominator has undesirable properties when there are missing data, in that individuals with rare variants and missing data are given greater weights than individuals with rare variants without missing data and (2) for the BRV, unlike for GRANVIL, variant “dosage” can be incorporated when there are missing data. We call the extended version of the BRV “BRV-M.” The BRV-M aggregate test is constructed as follows.

For each sample i , one calculates $X_i^{BRV} = \sum_j G_{ij} \delta_j$, the sum of minor alleles in sample i . Similar to the CMC, this approach treats missing genotypes as $G_{ij} = 0$. To extend the BRV approach for missing data, one simply sums the average genotypes at SNPs with missing values. Specifically, let $P'_j = 2 \times MAF_j(1 - MAF_j)$ and $P''_j = MAF_j^2$ be the probability of being a heterozygote or a homozygote for the rare allele of SNP_j , then

$$P_i^{BRV} = \sum_{j \in B_i} P'_j + 2 \sum_{j \in B_i} P''_j \text{ and } \tilde{X}_i^{BRV} = X_i^{BRV} + P_i^{BRV}.$$

WSS-M: Extension of the Weighted Approach

In the weighted approach, each variant is weighted by a function of the observed MAF.

The weight for SNP_j is

$$w_j = \frac{1}{\sqrt{nMAF_j(1 - MAF_j)}} \text{ and } X_i^{WSS} = \sum_j G_{ij} w_j.$$

Just as with the CMC and BRV approaches, this method treats $G_{ij} = 0$. To extend this method, let $C_{ij} = 1$ if G_{ij} is nonmissing for sample i and SNP_j , and 0 otherwise. Then

$$\tilde{G}_{ij} = \begin{cases} G_{ij}, & \text{When } C_{ij}=1 \\ \hat{G}_{ij}, & \text{When } C_{ij}=0 \end{cases} \text{ and } \hat{G}_{ij} = P'_j + 2P''_j$$

which is the average genotype for SNP_j . Finally, $\tilde{X}_i^{WSS} = \sum_j \tilde{G}_{ij} w_j$.

VT-M: Extension of the VT Approach

The VT method can be implemented by using either the CMC method of counting carriers, or the BRV method of counting alleles. For simplicity in what follows, we count carriers letting

$$X_{ik}^{VT} = I \left[\left\{ \sum_j G_{ij} \delta_j^k \right\} > 0 \right]$$

be an indicator for whether the i th individual is a rare-variant carrier for the k th MAF threshold where $\delta_j^k = 1$ if SNP_j has $MAF \geq T_k$, and 0 otherwise and T_k is the k th MAF threshold.

As with the other tests, missing values are treated as $G_{ij} = 0$. To extend the VT to deal with missing genotypes, the CMC-M is used for each MAF threshold. The notation from the CMC-M approach is extended to deal with multiple MAF thresholds where $A^k = \{i = 1, \dots, n, \text{ such that } X_{ik}^{VT} = 0 \text{ and } G_{ij} \text{ is missing for some } j, \text{ and the } k\text{th MAF threshold}\}$ is the set of samples that have at least one missing genotype and are not otherwise considered carriers of the rare allele, $B_i^k = \{j, \text{ such that } G_{ij} \text{ is missing for } SNP_j \text{ in sample } i \text{ and } \delta_j^k = 1\}$ is the set of SNPs, meeting the k th MAF threshold, with missing genotypes for sample i , and

$$\tilde{X}_{ik}^{VT} = \begin{cases} X_{ik}^{VT}, & i \notin A^k \\ P_{ik}^{VT}, & i \in A^k \end{cases},$$

where

$$P_{ik}^{VT} = 1 - \left[\prod_{j \in B_i^k} (1 - MAF_j)^2 \right]$$

is the probability that sample i is a carrier of the minor allele at a variant site (meeting the k th MAF threshold) with a missing genotype, and MAF_j is the MAF at SNP_j .

Association Testing

Logistic regression can be used to test for associations with the CMC-M, BRV-M, WSS-M, and VT-M methods. However, in many circumstances, the X variables may be sparse and asymptotics do not hold. To guard against this problem, significance was evaluated empirically for all tests by permuting genotypes. To test for association with the VT and VT-M methods, regression z -scores were calculated at every observed MAF threshold, taking $z\text{-max} = \text{the maximum value of } Z_k \text{ over the } k \text{ MAF thresholds}$. Statistical significance was assessed by permuting phenotypes and re-calculating $z\text{-max}$ at every permutation, allowing $z\text{-max}$ to be obtained at different values of k for every permutation. Just as was performed in Price et al. [2010], we used linear rather than logistic regression for computational speedup in our simulations.

Simulation Settings

For the WSS and VT, all simulated variants were analyzed, i.e., $MAF < 5\%$, while for BRV and CMC, only those variants with MAF of $\geq 1\%$ were analyzed. Type I error was evaluated for BRV, CMC, WSS, VT, BRV-M, CMC-M, WSS-M, and VT-M by generating 10,000 replicates. In order to empirically estimate P -values, permutation was performed with a stopping rule of 1 million iterations or 1,000 test statistics more extreme than the one

observed. Due to lack of power to detect associations, we disregarded replicates where a gene displayed a cumulative MAF < 0.005 or had less than two variant sites.

Type I error was evaluated when 5%, 10%, and 15% of the variant sites are randomly missing variant calls. Additionally, type I error was evaluated in the presence of differential missing data between cases and controls, where 15% of the cases were missing variant calls and 5% of the controls were missing variant calls. We also generated differential missing data for the situation where there is no missing data in cases, but the controls have missing rates of genotypes of 20%. In order to evaluate how effective filters are in controlling type I error, we also re-ran each analysis but this time filtering out individuals in the analyzed gene region who are missing 80% of their variant calls and variant sites missing 10% of their genotypes.

We also examined power to detect associations for tests and scenarios (i.e., filtering on missing variant calls) that did not lead to inflation in type I error. Power was examined both for the CMC, BRV, WSS, and VT methods as well as the extended methods, CMC-M, BRV-M, WSS-M, and VT-M with 0%, 5%, and 10% of the genotypes randomly missing. For differential missing data, we examined power for the CMC-M, BRV-M, WSS-M, and VT-M methods. We compared the power of these four extended methods to the original methods when variant sites missing data were removed from the analysis. Power was evaluated to detect an association for $\alpha = 0.05$, by generating 1,000 replicates each with 1,000 cases and 1,000 controls. For each replicate, the *P*-value was obtained empirically via 1,000 permutations.

Analysis of NHLBI-ESP Data

In order to evaluate the performance of the extended RVA methods to control false-positive rates, EA data from the NHLBI-ESP project were analyzed. Prior to exome sequencing, shotgun libraries were captured for exome enrichment using one of four in-solution capture products. One thousand individuals whose DNA samples had been processed using the Agilent capture array [RefSeq2010V2, 36.5 Mb] were assigned to the “case” group and 1,000 samples processed on the Roche/Nimblegen capture array [SeqCap EZ Human Exome Library v1.0, 32.8 Mb] were assigned to the “control” group. Figure 4 displays the overlap between transcripts and variant sites in common between the two arrays. RVA analysis was performed by analyzing nonsynonymous variant sites in aggregate for each gene with at least two variant sites and a cumulative MAF > 0.005 . Association analysis was performed using the CMC, BRV, WSS, VT, CMC-M, BRV-M, WSS-M, and VT-M methods. For the CMC, CMC-M, BRV, and BRV-M methods, only those nonsynonymous sites with an MAF of $< 1\%$ were analyzed while for the WSS, WSS-M, VT, and VT-M methods, all nonsynonymous sites with an MAF $< 5\%$ were analyzed.

For the variant sites passing QC, a total of 10,375 genes were captured by both arrays, contained at least two variant sites with MAF $< 1\%$, and had a cumulative MAF > 0.005 .

For each of the eight methods, the analysis was performed a second time after enforcing a per-gene filtering strategy that removed individuals missing 80% of variant calls within the analyzed gene region and variant sites missing 10% of their variant calls. Significance was

assessed via permutation with a stopping rule of 1 million iterations or 1,000 test statistics more extreme than the one observed.

Discussion

When analyzing individual SNPs, differential rates of missing genotype calls between cases and controls do not increase the type I error rate. Missing genotype data in array-based GWAS can be indicative of genotyping errors, which if not independent of phenotype status can lead to an increase in the type I error rate. Therefore, it is common practice in GWAS to remove those SNPs missing >5% of their genotypes. Likewise, for rare variant data obtained from NGS technologies, missing genotypes may indicate genotyping errors and it is wise to consider a call-rate based per-variant filter. However, unlike with the analysis of individuals SNPs, RVA analyses may show an increase in the type I error rate if there are differential rates of missing data between cases and controls.

To reduce missing variant calls for WG sequence data, the current best practices from the 1000 Genomes project recommend linkage disequilibrium (LD) based calling for missing genotypes [Genomes Project, 2010]. Although this approach is often applied to low pass WG data, LD-based calling is not feasible for exome sequence data [Do et al., 2012]. Even for WG data, LD-based calling will not work well for very rare variants seen only a few times because these variants are exceedingly difficult to correctly phase [Browning and Browning, 2011]. Tennesen et al. [2012] recently reported that 72% of the variants in the human exome contain three or fewer minor alleles, e.g., singletons, doubletons, or tripletons. Therefore, LD-based methods do not resolve the problem of missing data, because the majority of the variants included in an aggregate RVA test are very rare.

We have observed that the WSS and VT seem to be more heavily influenced by differential missing data, in that they display higher type I error rates than the CMC and BRV approaches. The WSS weights each variant by a function of its MAF such that lower frequency variants receive higher weights. Thus, the WSS test is designed to up-weight null (i.e., nondisease associated) variants where differential missing data causes an imbalance in the observed MAF between cases and controls and a corresponding decrease in the observed overall MAF. On the other hand, since the VT finds the maximum test statistic across MAF thresholds, under the null model, it will maximize over the frequency ranges with the highest levels of differential missing calls. Although the extensions for WSS and VT do aid in the control of type I error when there is differential missing data, there is still a modest inflation in type I error for the VT-M, when the percent of differential missing data is high.

This inflation can be removed by enforcing a per-sample (within gene) and per-variant call-rate filter. Although we filtered by removing samples missing 80% of their variant sites within a region and variant sites missing 10% of their gene calls, in some situations a more stringent filter may be advisable. For example, if it is suspected that low call rate is due to copy number variation in the region, a more stringent per-sample and per-variant filter should be used. Generally, rigorous filters should be considered if one suspects that call rate is correlated with genotyping error. More samples and variants provide higher power to detect associations, but can do so at the cost of inflated type I error rates. Ultimately, the

cutoffs used for call rate based filters are at the investigator's discretion in seeking a balance between sensitivity and specificity.

It is worth noting that there is another class of RVA methods that test for heterogeneity of effect within a genetic region. The sequence kernel association test (SKAT) [Wu et al., 2011] is based on a variance components model. Although conceptually quite different from the mean-shift models we considered, the default implementation of SKAT calls for removing variant sites with >15% missing data and replaces missing genotypes with a dosage based on the observed MAF. This default behavior is essentially equivalent to our proposed extensions (CMC-M, BRV-M, WSS-M, VT-M). Accordingly, we show that SKAT without this correction demonstrates substantially elevated type I error rates in the presence of differential missing rates between cases and controls (Supplementary Figs. S5–S7). Our conclusions regarding the power for CMC-M, BRV-M, WSS-M, and VT-M generalize to SKAT. That is, when missing data are random, the power of the RVA tests with and without a correction for missing data is equivalent for CMC, BRV, WSS, and VT (Supplementary Figs. S2 and S3) and for SKAT (Supplementary Figs. S8 and S9).

Although this article concentrates on correcting for missing variants in case-control data, these methods can easily be applied to quantitative trait (QT) analysis using linear instead of logistic regression. For analysis of QTs either QT values or QT residuals after adjusting for potential confounders can be analyzed; the methods to correct for missing data remain exactly the same as for case-control data.

As was proposed in the original articles [Madsen and Browning, 2009; Price et al., 2006], for the VT-M and WSS-M, *P*-values should be obtained via permutations. Since the data used in RVA tests may be quite sparse, in order to effectively control the type I error rate empirical *P*-values based on permutations should also be obtained for the CMC-M and BRV-M.

We have shown that for RVA analyses, differential missing data may cause substantial increases in type I errors that can lead to spurious associations. We developed extensions of four common RVA tests (CMC-M, BRV-M, WSS-M, and VT-M) and showed that, along with a modest filtering strategy, these extended RVA tests properly control type I error in the presence of differential missing data. Importantly, they do so without sacrificing power compared to their original counterparts.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors wish to acknowledge the support of the National Heart, Lung, and Blood Institute (NHLBI) and the contributions of the research institutions, study investigators, field staff, and study participants in creating this resource for biomedical research. Funding for GO ESP was provided by NHLBI grants RC2 HL-103010 (HeartGO), RC2 HL-102923 (LungGO), and RC2 HL-102924 (WHISP). The exome sequencing was performed through NHLBI grants RC2 HL-02925 (BroadGO) and RC2 HL-102926 (SeattleGO).

References

- Browning SR, Browning BL. Haplotype phasing: existing methods and new developments. *Nat Rev Genet.* 2011; 12(10):703–714. [PubMed: 21921926]
- Do R, Kathiresan S, Abecasis GR. Exome sequencing and complex disease: practical aspects of rare variant association studies. *Hum Mol Genet.* 2012; 21(R1):R1–R9. [PubMed: 22983955]
- Genomes Project C. A map of human genome variation from population-scale sequencing. *Nature.* 2010; 467(7319):1061–1073. [PubMed: 20981092]
- Leinonen R, Akhtar R, Birney E, Bower L, Cerdeno-Tarraga A, Cheng Y, Cleland I, Faruque N, Goodgame N, Gibson R, et al. The European nucleotide archive. *Nucleic Acids Res.* 2011; 39:D28–D31. (Database issue). [PubMed: 20972220]
- 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]
- Lin DY, Tang ZZ. A general framework for detecting disease associations with rare variants in sequencing studies. *Am J Hum Genet.* 2011; 89(3):354–367. [PubMed: 21885029]
- Little, RJ.; Rubin, DB. *Statistical Analysis with Missing Data.* New York: John Wiley and Sons; 2002.
- Liu DJ, Leal SM. A novel adaptive method for the analysis of next-generation sequencing data to detect complex trait associations with rare variants due to gene main effects and interactions. *PLoS Genet.* 2010; 6(10):e1001156. [PubMed: 20976247]
- Madsen BE, Browning SR. A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet.* 2009; 5(2):e1000384. [PubMed: 19214210]
- Mailman MD, Feolo M, Jin Y, Kimura M, Tryka K, Bagoutdinov R, Hao L, Kiang A, Paschall J, Phan L, et al. The NCBI dbGaP database of genotypes and phenotypes. *Nat Genet.* 2007; 39(10):1181–1186. [PubMed: 17898773]
- 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, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet.* 2006; 38(8):904–909. [PubMed: 16862161]
- Price AL, Kryukov GV, de Bakker PI, Purcell SM, Staples J, Wei LJ, Sunyaev SR. Pooled association tests for rare variants in exon-resequencing studies. *Am J Hum Genet.* 2010; 86(6):832–838. [PubMed: 20471002]
- Tennessen JA, Bigham AW, O'Connor TD, Fu W, Kenny EE, Gravel S, McGee S, Do R, Liu X, Jun G, et al. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science.* 2012; 337(6090):64–69. [PubMed: 22604720]
- Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare-variant association testing for sequencing data with the sequence kernel association test. *Am J Hum Genet.* 2011; 89(1):82–93. [PubMed: 21737059]
- Zheng J, Li Y, Abecasis GR, Scheet P. A comparison of approaches to account for uncertainty in analysis of imputed genotypes. *Genet Epidemiol.* 2011; 35(2):102–110. [PubMed: 21254217]

Figure 1.

QQ plots of the four RVA tests (CMC, BRV, WSS, and VT) and their extensions (CMC-M, BRV-M, WSS-M, and VT-M). Data were generated under the null for 1,000 cases and 1,000 controls for the *ALK* gene. For the cases, 0% of the genotypes are missing while for controls 20% of the genotypes are missing. The results are shown for CMC and CMC-M (panel A), BRV and BRV-M (panel B), WSS and WSS-M (panel C), and VT and VT-M (panel D) before and after filtering out those individuals missing 80% of their variant sites for the analyzed gene region and variant sites missing 10% of their variant calls.

Figure 2.

Results from the power study were performed with 0% missing genotypes in cases and 10% missing genotypes in controls. The y-axis displays the power and the x-axis displays the percent of variants that are causal. The results are shown for the *ALK* gene for 1,000 cases and 1,000 controls. The power is displayed for both the fixed effect (OR = 3) and the variable effect (OR_{min} = 2; OR_{max} = 10) models. Results are shown for analysis performed using CMC and CMC-M (panel A), BRV and BRV-M (panel B), WSS and WSS-M (panel C), and VT and VT-M (panel D). Analysis implementing the extended versions was

performed after filtering by removing samples missing 80% of their variant sites (within gene) and variant sites missing 10% of their variant calls. Analyses using the original RVA tests were performed after removing all variant sites missing >0% of their variant calls.

Figure 3.

Results from the exome-wide case-control analysis. Analysis was performed by assigning 1,000 samples for which the Agilent capture array was used as the “case” group and 1,000 samples for which the Nimblegen capture array was implemented as the “control” group. The results are shown for when analysis was performed using the CMC and CMC-M (panel S), BRV and BRV-M (panel N), WSS and WSS-M (panel C), and VT and VT-M (panel D) before and after filtering out those samples missing 80% of their variant sites (within gene) and variant sites missing 10% of their variant calls.

Figure 4.

Overlap between variant sites and transcripts between the two capture arrays that were analyzed. Blue indicates variants and transcripts on the Agilent array, red indicates variants and transcripts on the Nimblegen array. Although there is substantial overlap, there are hundreds of transcripts and thousands of variants that are exclusive to each array.

Table 1

Type I error levels for the four aggregate rare variant tests and their extensions

α -level	Filter	Gene	CMC	CMC-M	BRV	BRV-M	WSS	WSS-M	VT	VT-M
0.05	No	ALK	0.103	0.046	0.101	0.044	0.239	0.043	0.273	0.047
0.005	No	ALK	0.0168	0.0045	0.0168	0.0037	0.0580	0.0046	0.0861	0.0048
0.05	Yes	ALK	0.070	0.039	0.078	0.045	0.143	0.044	0.161	0.044
0.005	Yes	ALK	0.00626	0.0035	0.0118	0.0055	0.0323	0.0048	0.0408	0.0048
0.05	No	MC4R	0.064	0.049	0.067	0.046	0.150	0.045	0.141	0.043
0.005	No	MC4R	0.0109	0.0057	0.0103	0.0045	0.0308	0.0056	0.0331	0.0049
0.05	Yes	MC4R	0.0816	0.042	0.103	0.047	0.113	0.047	0.108	0.043
0.005	Yes	MC4R	0.0191	0.0095	0.0192	0.0017	0.0202	0.0045	0.0220	0.0037

Missing rates were generated for the *ALK* and *MC4R* genes at rates of 0% in cases and 20% in controls. Results both before and after filtering (removing those individuals missing 80% of their variant sites for the analyzed gene region and variant sites missing 10 of their variant calls) are displayed.