Xing et al. BMC Genetics 2014, 15:104
http://www.biomedcentral.com/1471-2156/15/104

BMC Genetics

RESEARCH ARTICLE Open Access

Performance of statistical methods on CHARGE targeted sequencing data

Chuanhua Xing[1*], Josée Dupuis[1,2] and L Adrienne Cupples[1,2*]

Abstract

Background: The CHARGE (Cohorts for Heart and Aging Research in Genomic Epidemiology) Sequencing Project is a national, collaborative effort from 3 studies: Framingham Heart Study (FHS), Cardiovascular Health Study (CHS), and Atherosclerosis Risk in Communities (ARIC). It uses a case-cohort design, whereby a random sample of study participants is enriched with participants in extremes of traits. Although statistical methods are available to investigate the role of rare variants, few have evaluated their performance in a case-cohort design.

Results: We evaluate several methods, including the sequence kernel association test (SKAT), Score-Seq, and weighted (Madsen and Browning) and unweighted burden tests. Using genotypes from the CHARGE targeted-sequencing project for FHS (n = 1096), we simulate phenotypes in a large population for 11 correlated traits and then sample individuals to mimic the CHARGE Sequencing study design. We evaluate type I error and power for 77 targeted regions.

Conclusions: We provide some guidelines on the performance of these aggregate-based tests to detect associations with rare variants when applied to case-cohort study designs, using CHARGE targeted sequencing data. Type I error is conservative when we consider variants with minor allele frequency (MAF) < 1%. Power is generally low, although it is relatively larger for Score-Seq. Greater numbers of causal variants and a greater proportion of variance improve the power, but it tends to be lower in the presence of bi-directionality of effects of causal genotypes, especially for Score-Seq.

Keywords: Case-cohort design, CHARGE targeted sequencing data, Rare variants, Type I error, Power, SKAT, Score-Seq, Madsen and browning, Burden tests

Background

Genome-wide association studies (GWAS) have identified hundreds of disease susceptible loci that harbor common variants, but most are not causal and explain only a small portion of the genetic risk for most diseases. The role of rare variants with minor allele frequency (MAF) < 0.05 has not been comprehensively explored in GWAS, while rare variant associations are believed to play an important role in disease etiology [1-12]. Emerging sequencing technologies allow for the characterization of virtually all of an individual's genetic variation. Hence, motivations for this work are: 1) the shift in measurement of genetic variants away from common variation using genotyping arrays to genotyping or sequencing of rare variants, requiring greater understanding of rare variant methods; and 2) the high cost of sequencing requires careful consideration of efficient study designs. Here we discuss the case-cohort study design for sequencing studies and evaluate the possible limitations of current methods for data collected under this study design.

The Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Sequencing Project is a national, collaborative effort from three studies: Framingham Heart Study (FHS), Cardiovascular Health Study (CHS) and Atherosclerosis Risk in Communities (ARIC). What makes the CHARGE sequencing study different from other studies is its case-cohort design, where a cohort random sample plus selected individuals with extreme values from one or more pre-specified traits are considered for analysis. Such a study design is advantageous when investigators wish to examine multiple traits. One component of the CHARGE targeted sequencing study involves 1096 individuals from FHS, consisting of a cohort random

* Correspondence: chuanhua.xing@gmail.com; adrienne@bu.edu
[1]Department of Biostatistics, Boston University, Boston, MA, USA
[2]Framingham Heart Study, Framingham, MA, USA

sample of 504 study participants from the Offspring Cohort and 592 participants selected from the extremes of 11 traits.

In recent years, many statistical approaches have been developed to jointly analyze multiple rare variants in aggregate-based tests to gain power. But current statistical methods for rare variant association studies rarely consider a case-cohort design, and hence potential bias in estimation and type-I error might be observed in analyses of CHARGE targeted sequencing data. The methods developed to date are generally for studies in which participants are assumed to be independent and a random representation of the general population such as case–control design. Typically, all study participants are considered for analyses in a case-cohort design; for dichotomous traits, participants affected by a specific disease or trait are considered as cases and other participants carrying other diseases or from a random non-diseased sample are considered as controls; for quantitative traits, all participants are used in genetic association studies of a specific disease trait, but potential bias in effect estimates may arise when including selected extreme values. Some participants as "potential risk carriers" for multiple traits can also make the issue even more complex. The uniquely ascertained participants in a case-cohort design with correlated traits form a non-representative dataset and may generate biases.

To address the concerns regarding the case-cohort study design and application of methods for rare variants, we evaluate type I error and power of statistical methods for aggregate-based association tests of rare variants in the case-cohort study design of the CHARGE targeted sequencing project. We examine the statistical performance of commonly used methods, the Sequence Kernel Association Test (SKAT) [13], Score-Seq [14], weighted [15], and unweighted (T1 [16]) burden tests. These methods have been well-studied using simulated data. Although Ladouceur et al., [17] used Sanger sequencing from 1998 individuals for both continuous and binary traits in their power comparison, they did not perform type 1 error comparisons. Our work contributes in the following aspects. (1) We evaluate the statistical performance of several statistical methods that aggregate data in a genomic region on measured CHARGE targeted sequencing data based on a case-cohort design. (2) We evaluate over seventy-seven targeted sequencing regions in CHARGE, representing a wide range of genotype structures. (3) We consider complex, correlated phenotypes. (4) We evaluate both type I error and power, because power is a valid measure only if type I error is properly controlled. We aim to provide some guidelines on the performance of these methods to detect associations with rare variants on CHARGE targeted sequencing data using the case-cohort study design.

Methods

Data

We used genotypes from our CHARGE targeted-sequencing project for the Framingham Heart Study (n = 1096), and we simulated correlated phenotypes to mimic the potential relationship among phenotypes in real data. We generated 11 correlated traits similar to those found in FHS for a very large population of 12,000 people, using the method described by Lumley et al. (http://stattech.wordpress.fos.auckland.ac.nz/files/2012/05/design-paper.pdf). The correlation between traits was induced by an iterative process. The initial trait was generated using a t distribution with 15 degrees of freedom. The number of traits was doubled after each iteration; half were generated by adding the previous traits to a randomly generated t value, and the second half were generated by adding the negative of the previous traits to a randomly generated t value. We generated $2^4 = 16$ traits in this manner, and we selected the first 11 traits for analysis. The correlation among the traits is given in Table 1.

Table 1 Correlation among 11 traits

Traits\Traits	1	2	3	4	5	6	7	8	9	10	11
1	1	0.6	0.2	0.6	−0.2	0.2	0.6	0.2	−0.6	−0.2	0.2
2	0.6	1	0.6	0.2	0.2	−0.2	0.2	0.6	−0.2	−0.6	−0.2
3	0.2	0.6	1	0.6	0.6	0.2	−0.2	0.2	0.2	−0.2	−0.6
4	0.6	0.2	0.6	1	0.2	0.6	0.2	−0.2	−0.2	0.2	−0.2
5	−0.2	0.2	0.6	0.2	1	0.6	0.2	0.6	0.6	0.2	−0.2
6	0.2	−0.2	0.2	0.6	0.6	1	0.6	0.2	0.2	0.6	0.2
7	0.6	0.2	−0.2	0.2	0.2	0.6	1	0.6	−0.2	0.2	0.6
8	0.2	0.6	0.2	−0.2	0.6	0.2	0.6	1	0.2	−0.2	0.2
9	−0.6	−0.2	0.2	−0.2	0.6	0.2	−0.2	0.2	1	0.6	0.2
10	−0.2	−0.6	−0.2	0.2	0.2	0.6	0.2	−0.2	0.6	1	0.6
11	0.2	−0.2	−0.6	−0.2	−0.2	0.2	0.6	0.2	0.2	0.6	1

We considered both positive and negative correlations among traits. There were strong positive correlations between pairs of traits such as traits 5 and 9, traits 6 and 10, and traits 7 and 11. There were also strong negative correlations between pairs of traits such as traits 1 and 9, traits 2 and 10, and traits 3 and 11. We picked some representative traits having a wide range of pairwise correlations to test the performance of the statistical methods. The selected traits included traits 1, 2, 6, 9, and 10. We focus on traits with differing correlations, positive and negative, especially with correlation 0.6 between traits 1 and 2, 0.2 between traits 1 and 6, −0.6 between traits 1 and 9, and −0.2 between traits 1 and 10.

Next, we sampled a subset of individuals from the large population, using the same sampling scheme that was used to select participants for the CHARGE targeted sequencing project. We first selected a random cohort with 504 individuals. We then sampled extremes for each of 11 traits, with participants in the extreme for one trait not eligible for selection for other traits. We chose the top 50 unselected individuals at the extremes for each of 10 traits, and then chose the top 92 to mimic one trait in FHS that had more individuals in the extreme. All individuals, regardless of selection, are analyzed using continuous traits in our case-cohort design.

We randomly assigned the generated phenotypes to genotypes for type I error tests, and denote them as y_0 under the null hypothesis. We generated phenotypes for our power evaluation using the equation

$$y_p = y_0 + \sum_{j=1}^{P} \beta_j G_j, \quad (1)$$

where $\sum_{j=1}^{P} \beta_j G_j$ indicates the additional power generated from P causal SNPs (coefficient β_j for SNP G_j with $j = 1,\ldots P$) and y_0 is generated under the null hypothesis. We randomly selected a portion of rare variants with MAF < 1% as causal variants, and the effect sizes for the causal variants were calculated by 0.4*|log10(MAF)|, following the approach of Wu et al. [13]. Power will increase with the larger the number of causal variants in this aggregate sum and the larger their effect sizes.

We selected causal variants by including all potentially functional variants as annotated by [18], while avoiding a low total number of causal variants in a region. Previous studies have used simulated genotype data and selected a low percentage of causal variants. We, however, used real targeted CHARGE sequencing genotypes, for which we can also obtain some known genetic information to aid in selection. Among SNPs with MAF < 1%, we selected all non-synonymous, stop-gain (non-sense) mutation, and splicing SNPs, because such SNPs have a higher chance to be causal, and we called them high risk variants. The number of such high risk variants for each of the 77 targeted regions in the CHARGE sequencing project varies from 0 to 81. For regions with a low number of causal variants, we selected additional causal variants using the following rules.

1. When the total number of variants for a region was low and less than 10, we selected all variants as causal regardless of whether they are high risk or not. We had 2 such regions.
2. When the number of high risk variants in a region was low and less than 5 and the total number of variants was between 10 and 100, we randomly selected an additional 50% of the variants as causal. We had 22 such regions.
3. When the number of high risk variants in a region was low and less than 5 but the total number was greater than 100, we selected an additional 5% of the non-high risk variants as causal.
4. When the number of high risk variants in a region was greater than 5, we chose all of them as causal.
5. We assigned causal variants to have the same direction of genetic effects for phenotypes using rules 1–4. We also assigned a second set of causal variants to have bi-directional effects on phenotypes using rules 1–4 by setting the first half to have positive effects and the second half to have negative effects on the phenotypes.

Note that rules 2–4 ensure that the number of causal variants in a region is 5 or more. However, removal of variants with a high missing rate (>10%) results in several regions having the number of causal variants < 5. These regions are 1, 9, 12, 18, 39 that have 4 causal variants and region 45 that has 3 causal variants.

Statistical methods description

We chose several representative analysis methods for aggregate tests to compare, including an unweighted burden test [16], weighting of variants by a function of MAF (similar to Madsen and Browning [15], referred to "MB"), SKAT [13], and Score-Seq [14]. Let G_{ij} denote the genotype of the jth variant for the ith person with values of 0, 1, or 2 according to the number of rare alleles for variant j, where $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots P$. Let Y_i denote the trait, and Z_{ik} the k th covariate for participant i, where $k = 1, \ldots M$. We present methods for quantitative traits, but they can readily be extended to dichotomous traits.

1. **Unweighted burden test statistic (T1-Count)** [16]

For each participant, a new variable is defined that counts the number of variants (0/1/2/3…) with MAF < 1%

in a targeted region where that person carries at least 1 rare allele. Association analysis with this new variable (T1-count) and a trait is carried out using linear (for a quantitative trait) or logistic (for a dichotomous trait) regression.

2. **Weighting of Variants by a function of MAF** (similar to Madsen & Browning [15] for binary traits and Xing et al. [1] for continuous traits; labeled MB in this paper)

For each person, a statistic is computed that is a weighted count of that person's rare alleles within a targeted, using weights based on the MAF averaged over the three studies of the CHARGE targeted sequencing-project. The approach gives more weight to rarer variants. We restricted our tests to rare variants with MAF < 1%. For a targeted region, the weighted genotype score is

$$S_i = \sum_{j=1}^{P} \frac{G_{ij}}{\hat{w}_j}, \tag{2}$$

where $\hat{w}_j = \sqrt{n\hat{p}_j\left(1-\hat{p}_j\right)}$, and $\hat{p}_j =$ estimate of the MAF of variant j. Association with this genotype score and the trait of interest can be evaluated using linear or logistic regression.

3. **SKAT statistics** [13]

The Sequence Kernel Association Test (SKAT) assumes the model

$$Y_i = \alpha_0 + \sum_{k=1}^{M} \alpha_k Z_{ik} + \sum_{j=1}^{P} \beta_j G_{ij} + \epsilon_i, \tag{3}$$

where α_0, α_k and β_j are regression parameters. SKAT is a general approach and uses weights computed from the data. For our purposes in testing the null hypothesis H_0: $\boldsymbol{\beta} = 0$, SKAT takes a simple form, where $\boldsymbol{\beta}$ is the vector of all $\beta_j s$. For a given set of weights, the score test can be expressed as

$$Q = \sum_{j=1}^{P} w_j S_j^2, \tag{4}$$

where $S_j = \sum_{i=1}^{n} G_{ij}(y_i - \hat{\mu}_{i0})$, $\hat{\mu}_{i0}$. is the predicted value of y_i from the model when there are no genotypes in the model. We used the Beta distribution for the weights, $w_j \sim Beta(a_1, a_2)$, with the default parameters $a_1 = 1$ d $a_2 = 25$ as suggested by Wu et al. [13]. Association between the trait of interest and the rare variants can be evaluated using the score test, and its significance is computed analytically using a mixture of chi-square distributions.

4. **Score-Seq statistics** [14]

We can relate Y_i to G_i and Z_i using the following linear regression model,

$Y_i = \tau S_i + \gamma^T Z_i + \epsilon_i$, where $\epsilon_i \sim N(0, \sigma^2)$.

Here $S_i = \xi^T G_i$, a scalar from the product of a weighted linear combination of $G_{i1}, \ldots, G_{iP}$ with weights of ξ_j. $\xi = (\xi_1, \ldots, \xi_P)^T$ is a $P \times 1$ vector, $\xi = \beta/\tau$ and τ is a scalar constant, and β is a vector of coefficients for G_i as defined in Equation (3) .

The score statistic and its variance are

$$U = \sum_{i=1}^{n} \left(Y_i - \hat{\gamma}^T Z_i \right) S_i$$

and

$$V = \hat{\sigma}^2 \left\{ \sum_{i=1}^{n} S_i^2 - \left(\sum_{i=1}^{n} S_i Z_i \right)^T \left(\sum_{i=1}^{n} Z_i Z^T{}_i \right)^{-1} \left(\sum_{i=1}^{n} S_i Z_i \right) \right\},$$

where

$$\hat{\gamma} = \left(\sum_{i=1}^{n} Z_i Z_i^T \right)^{-1} \sum_{i=1}^{n} Y_i Z_i,$$

and

$$\hat{\sigma}^2 = n^{-1} \sum_{i=1}^{n} \left(Y_i - \hat{\gamma}^T Z_i \right)^2.$$

der H_0, the test statistic $T = U/V^{1/2}$ has an asymptotic standard normal distribution. Lin et al. [14] also provides permutation-type tests for more accurate p values. We implemented permutation-type tests in this article.

Results and discussion

We applied the four methods to the simulated data for each targeted region using targeted sequence genotype data from 1096 FHS participants. We evaluated type I error and power with 10,000 replicates, at significance levels of $\alpha = 0.001$, 0.01, and 0.05. We restricted our analyses to genetic variants with MAF < 1%. We evaluate the statistical performance of the four approaches under a case-cohort design to provide some guidelines for studies with case-cohort designs.

Type I error

Type I error, estimated by the number of rejections divided by the number of replicates (10,000), for the 77 regions for trait 1 is presented in Figure 1 for $\alpha = 0.01$. Assuming the number of rejections follows a binomial

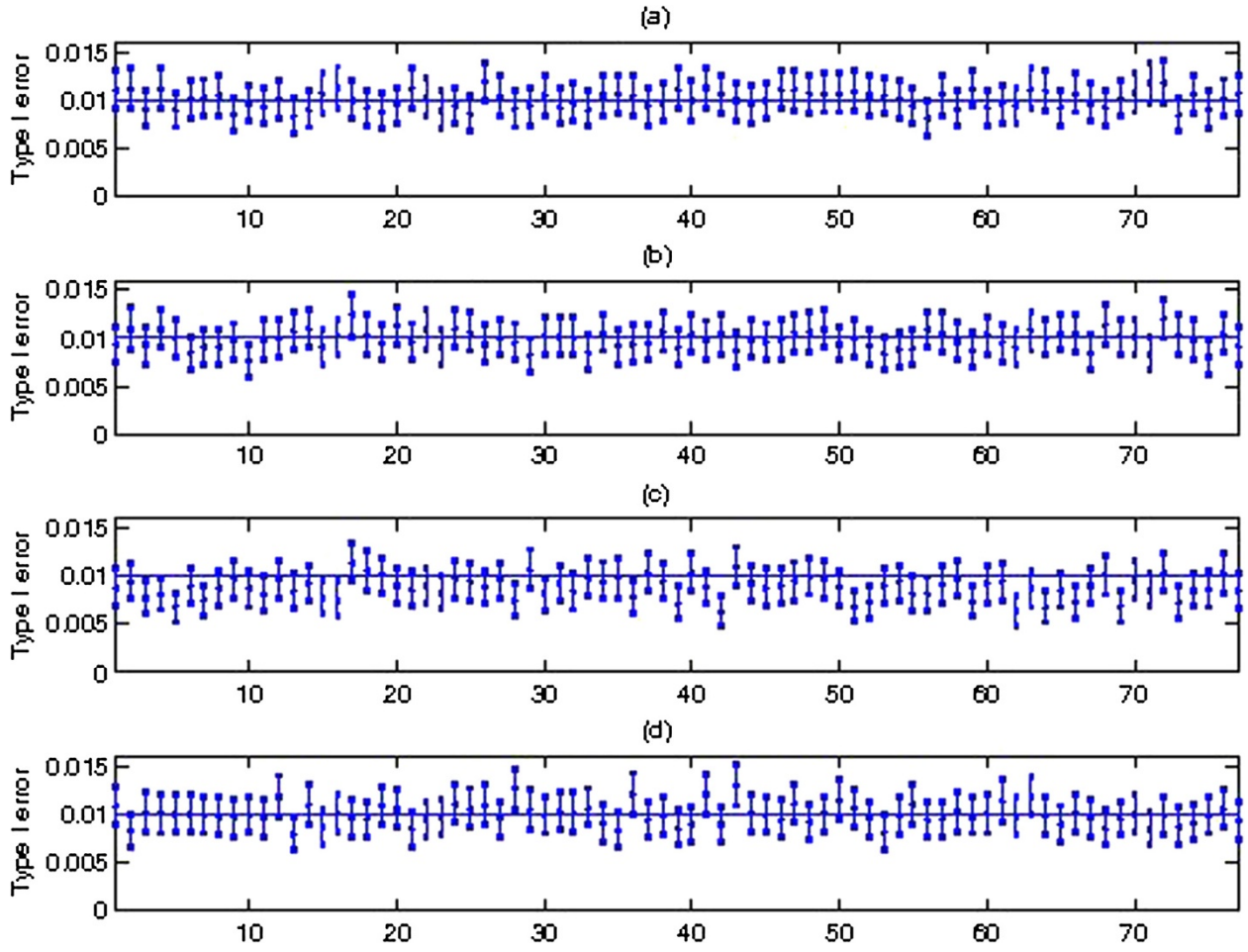

Figure 1 Comparison of Type I Error for trait 1 with alpha = 0.01. (a). Type I error for T1. **(b)**. Type I error for Madson and Browning, 2009. **(c)**. Type I error for SKAT [13]. **(d)**. Type I error for Score-Seq [14].

distribution, we calculated the 95% confidence interval for the type I error, and its bounds are indicated as the two ends for each region in Figure 1. The horizontal solid line indicates the nominal level of $\alpha = 0.01$. When the nominal level is within the 95% CI of type I error, we consider type I error to be properly controlled. We use numbers to indicate regions for simplicity of presentation. The mapping from region numbers to gene names and their chromosomes and positions are given in Additional file 1: Table S1 in the supplementary file.

From Figure 1 (a) for T1, we observe that the nominal significance level of 0.01 is within the 95% CI for most targeted regions. There are a few regions where the type I error is slightly inflated and the lower bound of the 95% CI is close to the nominal level, such as regions 26, 71 and 72. From Figure 1 (b) for MB, the type I error is also within the 95% CI for most targeted regions, with only a few regions having the lower bound of 95% CI close to the nominal level. From Figure 1 (c) for Score-Seq, the nominal level for most regions is within the 95% CI, with a few regions having a lower bound above the nominal level. From Figure 1 (d) for SKAT, no regions have type I error above the nominal level and the type I error tends to be conservative so that there are some regions with the upper bound of the 95% CI below the nominal level. As a result, SKAT has some regions with better controlled type I error than other methods such as region 24 compared to Score-Seq, although SKAT has more regions with conservative type I error. Regions 13 and 73 tend to have type I error close to or lower than the nominal level in T1, Score-Seq and SKAT. A closer look at them indicates that region 13 has 34 variants but region 73 has a large number of 257 variants with MAF < 1%. Overall, no methods have regions with 95% CIs above the nominal level except that SKAT has a few regions with the upper bound of 95% CI lower than the nominal level.

Type I error is also mostly under control when the nominal level alpha is larger at 0.05 or smaller at 0.001 (Additional file 1: Figure S1 in the Supplement). The overall control of type I error across methods is consistent

with previous reports [13-15,19]. Hence, although the case-cohort study design could induce biases by including extremes when using existing statistical methods, the type I error remains under control when applying these approaches to a case-cohort design.

Next, we examine the variation in type I error over traits that are correlated using SKAT (Figure 2 for alpha = 0.01). We omit traits 2 and 9 because their results are similar to traits 1 and 10. Type I error over the different traits is similar and well controlled, except for three regions for trait 6 (regions 7, 30 and 32) with inflated type I error. Further investigation indicates that region 7 has 194 variants, region 30 has 9 variants, and region 32 has 6 variants with MAF < 1%. The smaller number of variants for regions 30 and 32 may increase the risk of having elevated type I error for a region, because we treated each targeted region as a unit to jointly analyze rare variants regardless of the length of a region. However, there are other regions that have fewer variants with MAF < 1% such as region 40 with 3 variants and region 9 with 4 variants, but these two regions have appropriate type I error. More summary statistics, such as the number of variants in each region, can be found from Additional file 1: Table S1 in the Supplement. We also examined type I error over traits when alpha = 0.001 and 0.05 (Additional file 1: Figure S2 in the Supplement). When alpha is smaller with a value of 0.001 or larger with a value of 0.05, the type I error for all traits tends to be conservative, except for the three regions for trait 6. The type I error for the three regions for trait 6 tend to be smaller when alpha is smaller with value of 0.001, but tend to be larger when alpha is large with value of 0.05.

Power

We calculated power for the same four methods, T1, MB, SKAT and Score-Seq, and the powerfor all regions

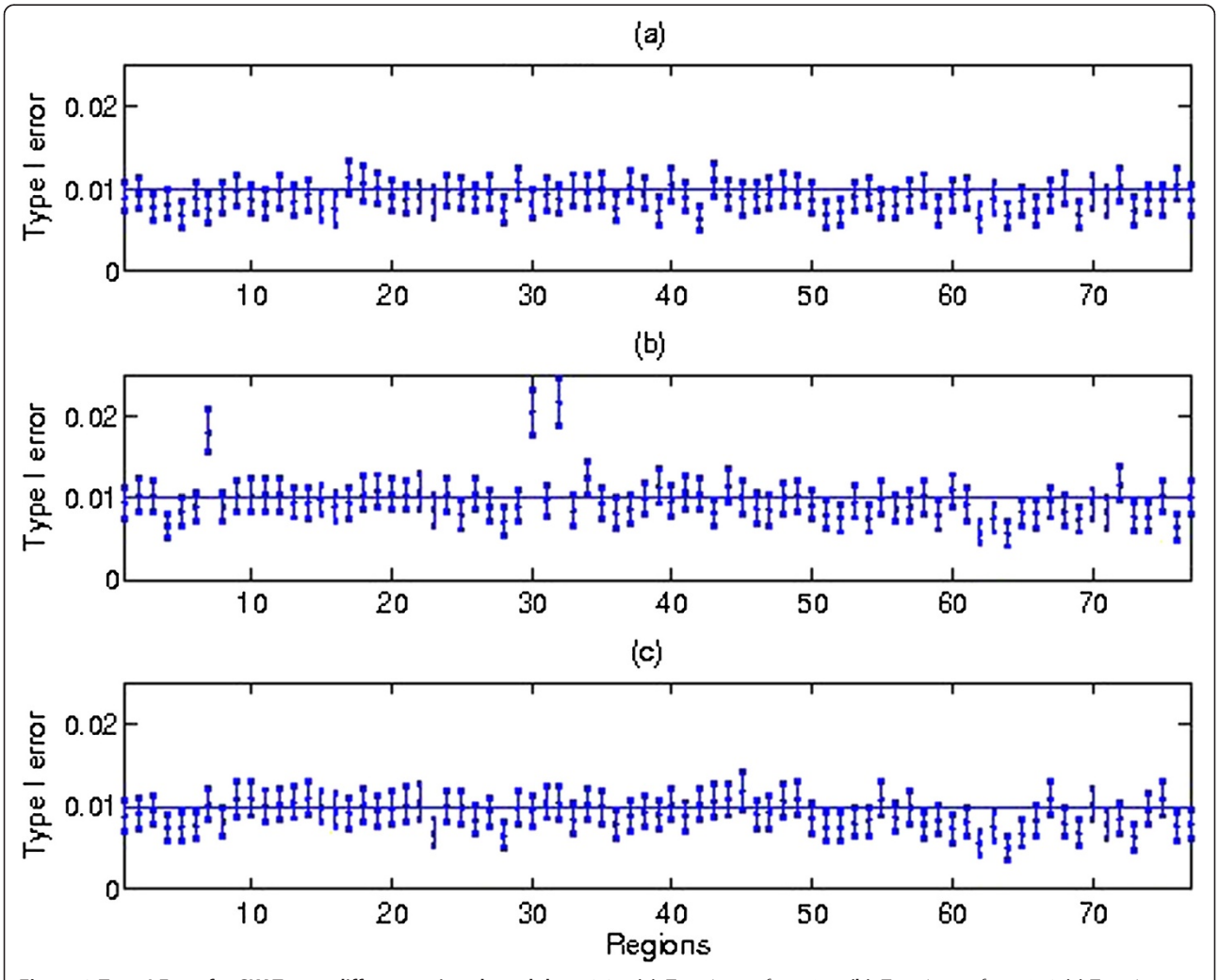

Figure 2 Type I Error for SKAT over different traits when alpha = 0.01. (a). Type I error for trait 1. **(b)**. Type I error for trait 6. **(c)** Type I error for trait 10.

is presented in Figure 3 when all causal variants have the same directionality (the upper half of Figure 3) and when 50% of causal variants have positive effect and another 50% negative effect (the lower half of Figure 3) with alpha = 0.05. We present power multiplied by the alpha/type I error for a fair comparison across methods, with alpha as the nominal significant level. From the upper half of Figure 3, we observe that the power in our sample for 1096 individuals is generally low for detecting association with rare variants when MAF < 1%. This result is expected, because the number of simulated causal variants is generally low for most regions, varying from 4 to 60, while our sample size is modest. Generally, the power for Score-Seq tends to be higher.

We explored different characteristics of regions to investigate the possible explanations for the difference in power over regions. The characteristics included the total number of causal variants, the total allele count, and the proportion of the variance explained by the rare variants (R^2). The summary is in Additional file 1: Table S1 of the Supplement. The power for T1 and MB are consistently low across all regions regardless of characteristics, no matter whether the genetic effects are in the same or in opposite direction.

When the number of causal variants for a region was large, the power tended to be high for Score-Seq and SKAT (Additional file 1: Figure S3 (a) of the Supplement), particularly for Score-Seq, which is consistent with the observation from [14]. For example, regions 44 and 54 have 51 and 60 causal variants, and the power of Score-Seq was 0.258 and 0.360, respectively. When the number of causal variants was modest and greater than 10, the power for Score-Seq was lower, varying from 0.05 to 0.2. When the number of causal variants was low and less than 10, the power was even lower.

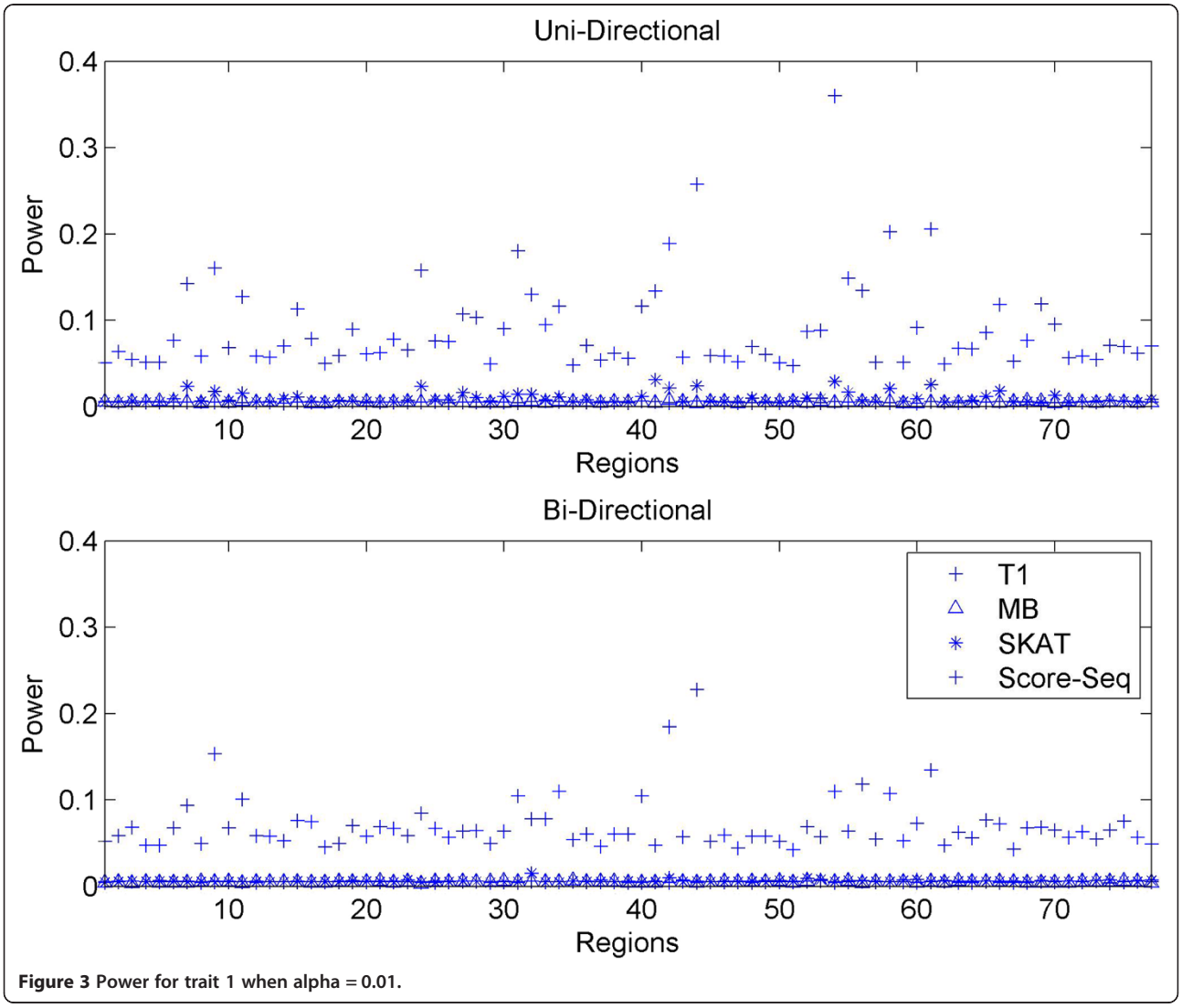

Figure 3 Power for trait 1 when alpha = 0.01.

We also evaluated relevant measures, total allele count and total number of variants. The total allele count was the total number of the minor alleles in all causal variants in a region. If a region had a higher total allele count, the power tended to be high (Additional file 1: Figure S3 (c) of the Supplement). However, not all regions that had higher total allele count had higher power. For example, region 73 had a total allele count of 31, but the power was only 0.006 for SKAT and was 0.055 for Score-Seq when genetic effects.

We further investigated the variation in power across the regions, using the proportion of variance explained by the variants $R^2 = \frac{var\left(\sum_{p=1}^{P}\beta_p G_p\right)}{var(y_p)}$ for each region, where $\sum_{p=1}^{P}\beta_p G_p$ and y_p are from equation (1). The calculated R^2 over regions is given in Figure 4, with the overall mean R^2 equal to 0.0040. Looking at Figures 3 and 4, both methods tended to have larger power when R^2 is larger, particularly Score-Seq with a larger increase (Additional file 1: Figure S3 (d) in the Supplement).

We investigated the power when causal variants have bi-directionality with 50% of variants having a positive effect and 50% of variants having a negative effect. The influence of relevant measures for power is plotted in Additional file 1: Figure S3 in the Supplement, when the causal effects are both uni-directional and bi-directional. The power over 77 regions is plotted in the lower half of Figure 3. The power for SKAT was low, and did not change much compared with the scenarios having the same directionality of genetic effect. The power for Score-Seq however decreased. The other characteristics of the targeted regions including the total number of causal variants and the total allele count did not change the power for SKAT much, compared to the power evaluated when all rare alleles shared the same directionality. This result is expected, as the power for SKAT is closely related to R^2 and is robust to directionality (Figure 4). Further observation indicates that the power for alpha = 0.01 tended to be lower (Additional file 1: Figure S4 in the Supplement). Other factors that may influence the power include sample size and weighting schemes/distributions of effects [14,17].

Conclusion

Many statistical methods have been developed in recent years to evaluate the risk conferred by rare variants in human complex diseases. However, no statistical method has considered the case-cohort design. We evaluated several approaches to access the association between groups of rare variants and a complex quantitative trait, because most FHS traits in the targeted sequencing project were quantitative. We aimed to evaluate several representative statistical methods instead of a comprehensive evaluation of all existing methods and compared their performance on our CHARGE targeted sequencing data. Our work contributes in that we are the first to evaluate both type I error and power using a case-cohort design of observed targeted CHARGE sequencing data. Seventy-seven targeted regions represent a wide range of genotypic characteristics from real sequencing data, and we evaluated the

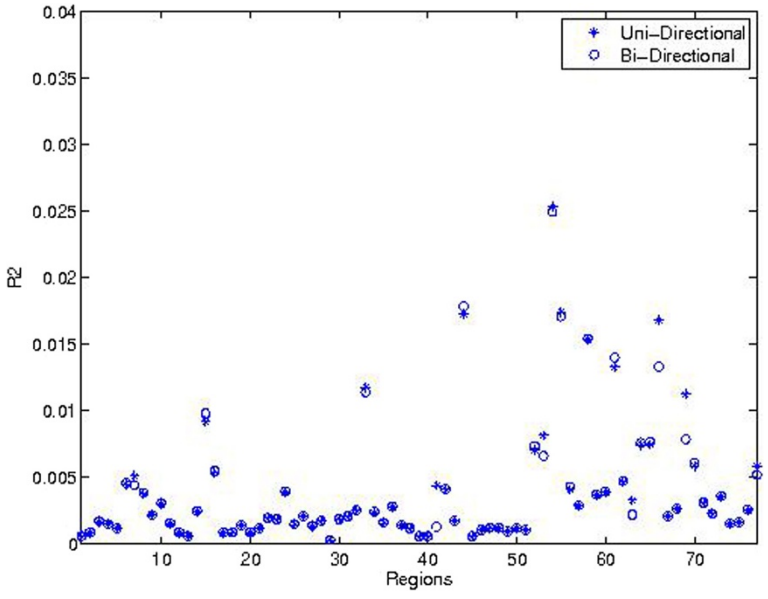

Figure 4 R^2 over target regions.

performance using correlated complex phenotypes by mimicking the sampling scheme used in CHARGE targeted sequencing project.

Type I error in the case-cohort design of CHARGE targeted sequencing data was mostly under control, although type I error for SKAT tended to be conservative. We tested the type I error over correlated traits using SKAT. Most regions for most traits had appropriate type 1 error. These results suggested that correlated complex phenotypes in a case-cohort design may influence the behavior of type I error but not substantially. Power was generally low in our studies no matter whether the effects of causal variants have the same directionality or bi-directionality, consistent with observations in previous studies [13,14,17]. As SKAT tended to have somewhat conservative type I error, power should be evaluated carefully, considering the type I error for each method in a particular region.

We examined different characteristics of targeted regions to explore possible explanations for the difference in power across regions in the case-cohort design based on CHARGE targeted sequencing data. The characteristics that we examined included the total number of causal variants, the total allele count, the proportion of variance explained by the causal variants (R^2), the significance level and directionality. Power for Score-Seq tended to be higher, when the total number of causal variants and the total allele count were larger. Score-Seq tended to have higher power when R^2 is larger. Bi-directionality does not seem to influence power much for SKAT, but lowers the power for Score-Seq. Other characteristics were also investigated in prior reports to determine the influence on power of characteristics such as the ratio of causal variants to total number of variants, effect sizes and sample sizes [13,14,17]. The proportion of the total number of causal variants among the total number of variants in a region is often used when the regions have a similar number of variants [13]. Our studies used effect sizes of the form 0.4*| log10(MAF)| [13] to generate larger effect sizes for rarer variants. Our Targeted Sequencing study in FHS had a fixed sample size of 1096, and hence we did not examine the influence of varying sample sizes. Although our results for power are limited to this sample size, we expect that comparisons across methods will be similar. Our work could also be extended to exome sequencing by applying tests to all variants within the exome of a gene, or within subsets of exons of a gene. Future improvements for statistical methods and a better understanding of the underlying genetic structure may aid in evaluating rare variant association studies.

Availability of supporting data

The data set supporting the results of this article is available in the dbGAP repository, http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000651.v3.p8.

Additional file

Additional file 1: The summary table of target regions and the additional figures.

Competing interests

The authors declare that they have no competing interests.

Authors' contribution

CX conceived the study design, carried out the data processing and the methods evaluation and comparison, and drafted the manuscript. JD conceived the study design, and participated the performance analyses and the writing. LAC conceived the study design, participated the performance analyses, and helped to draft the manuscript. All authors read and approved the final manuscript.

Acknowledgements

Chuanhua Xing's work is supported by 5 RC2 HL102419-02 (Boerwinkle), NIH/NHLBI.

Received: 28 May 2014 Accepted: 22 September 2014
Published online: 03 October 2014

References

1. Xing C, Satten GA, Allen AS: **A weighted accumulation test for associating rare genetic variation with quantitative phenotypes.** *BMC Proc* 2011, **5**(Suppl 9):S6. 10.1186/1753-6561-5-S9-S6.
2. Hoffmann TJ, Marini NJ, Witte JS: **Comprehensive approach to analyzing rare genetic variants.** *PLoS One* 2010, **5**(11):e13584. 10.1371/journal.pone.0013584.
3. Ahituv N, Kavaslar N, Schackwitz W, Ustaszewska A, Martin J, Hebert S, Doelle H, Ersoy B, Kryukov G, Schmidt S, Yosef N, Ruppin E, Sharan R, Vaisse C, Sunyaev S, Dent R, Cohen J, McPherson R, Pennacchio L: **Medical sequencing at the extremes of human body mass.** *Am J Hum Genet* 2007, **80:**779–791.
4. Azzopardi D, Dallosso AR, Eliason K, Hendrickson BC, Jones N, Rawstorne E, Colley J, Moskvina V, Frye C, Sampson JR: **Multiple rare nonsynonymous variants in the adenomatous polyposis coli gene predispose to colorectal adenomas.** *Cancer Res* 2008, **68:**358–363.
5. Brunham LR, Singaraja RR, Hayden MR: **Variations on a gene: Rare and common variants in ABCA1 and their impact on HDL cholesterol levels and atherosclerosis.** *Annu Rev Nutr* 2006, **26:**105–129.
6. Cohen JC, Kiss RS, Pertsemlidis A, Marcel YL, McPherson R, Hobbs HH: **Multiple rare alleles contribute to low plasma levels of HDL cholesterol.** *Science* 2004, **305:**869–872.
7. Cohen JC, Pertsemlidis A, Fahmi S, Esmail S, Vega GL, Grundy SM, Hobbs HH: **Multiple rare variants in NPC1L1 associated with reduced sterol absorption and plasma low-density lipoprotein levels.** *Proc Natl Acad Sci U S A* 2006, **103:**1810–1815.
8. Ji W, Foo JN, O'Roak BJ, Zhao H, Larson MG, Simon DB, Newton-Cheh C, State MW, Levy D, Lifton RP: **Rare independent mutations in renal salt handling genes contribute to blood pressure variation.** *Nat Genet* 2008, **40:**592–599.
9. Nejentsev S, Walker N, Riches D, Egholm M, Todd JA: **Rare variants of IFIH1, a gene implicated in antiviral responses, protect against type 1 diabetes.** *Science* 2009, **324:**387–389.
10. Romeo S, Yin W, Kozlitina J, Pennacchio LA, Boerwinkle E, Hobbs HH, Cohen JC: **Rare loss-of-function mutations in ANGPTL family members contribute to plasma triglyceride levels in humans.** *J Clin Invest* 2009, **119:**70–79.
11. Slatter TL, Jones GT, Williams MJ, van Rij AM, McCormick SP: **Novel rare mutations and promoter haplotypes in ABCA1 contribute to low-HDL-C levels.** *Clin Genet* 2008, **73:**179–184.
12. Walsh T, McClellan JM, McCarthy SE, Addington AM, Pierce SB, Cooper GM, Nord AS, Kusenda M, Malhotra D, Bhandari A: **Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia.** *Science* 2008, **320:**539–543.
13. 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.

14. 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.
15. Madsen B, Browning S: **A groupwise association test for rare mutations using a weighted sum statistic.** *PLoS Genet* 2009, **5**(2):e1000384.
16. Morris A, Zeggini E: **An evaluation of statistical approaches to rare variant analysis in genetic association studies.** *Genet Epidemiol* 2010, **34**(2):188–193.
17. Ladouceur M, Dastani Z, Aulchenko YS, Greenwood CMT, Richards JB: **The empirical power of rare variant association methods: results from Sanger sequencing in 1,998 individuals.** *PLoS Genet* 2012, **8**(2):e1002496.
18. Wang K, Li M, Hakonarson H: **ANNOVAR: Functional annotation of genetic variants from next-generation sequencing data.** *Nucleic Acids Res* 2010, **38:**e164.
19. Basu S, Pan W: **Comparison of statistical tests for disease association with rare variants.** *Genet Epidemiol* 2011, **35**(7):606–619. 10.1002/gepi.20609. Epub 2011 Jul 18.

doi:10.1186/s12863-014-0104-9
Cite this article as: Xing *et al.*: Performance of statistical methods on CHARGE targeted sequencing data. *BMC Genetics* 2014 **15**:104.