Replication Strategies for Rare Variant Complex Trait Association Studies via Next-Generation Sequencing

Dajiang J. Liu[1,2] and Suzanne M. Leal[1,2,*]

There is solid evidence that complex traits can be caused by rare variants. Next-generation sequencing technologies are powerful tools for mapping rare variants. Confirmation of significant findings in stage 1 through replication in an independent stage 2 sample is necessary for association studies. For gene-based mapping of rare variants, two replication strategies are possible: (1) variant-based replication, wherein only variants from nucleotide sites uncovered in stage 1 are genotyped and followed-up and (2) sequence-based replication, wherein the gene region is sequenced in the replication sample and both known and novel variants are tested. The efficiency of the two strategies is dependent on the proportions of causative variants discovered in stage 1 and sequencing/genotyping errors. With rigorous population genetic and phenotypic models, it is demonstrated that sequence-based replication is consistently more powerful. However, the power gain is small (1) for large-scale studies with thousands of individuals, because a large fraction of causative variant sites can be observed and (2) for small- to medium-scale studies with a few hundred samples, because a large proportion of the locus population attributable risk can be explained by the uncovered variants. Therefore, genotyping can be a temporal solution for replicating genetic studies if stage 1 and 2 samples are drawn from the same population. However, sequence-based replication is advantageous if the stage 1 sample is small or novel variants discovery is also of interest. It is shown that currently attainable levels of sequencing error only minimally affect the comparison, and the advantage of sequence-based replication remains.

Introduction

Currently there is worldwide interest in studying the role of rare genetic variants in the etiology of complex traits. A number of studies provide evidence that rare variants are involved in the etiology of complex diseases and quantitative phenotypes.[1–5] Indirect association mapping via tagSNPs is underpowered to detect associations with rare variants resulting from the weak correlations between higher-frequency tagSNPs and rare variants.[6] Instead, direct association mapping through sequencing candidate genes, exomes, or entire genomes needs to be applied, where variants are discovered and tested. With the rapid development of cost-effective next-generation sequencing technologies such as Illumina HiSeq, ABI SOLiD, and Roche 454 as well as target enrichment methods, sequence-based genetic association studies of complex traits have been made possible. For targeting large numbers of genetic regions, hybrid-based methods such as on-array or in-solution capture with NimbleGen or Agilent products have been very beneficial.[7–9] When targeting small genetic regions is of interest, such as in candidate genes, capture methods that use molecular inversion probes are advantageous.[9] Although sequencing only captured genetic regions can be cost and time effective, high sequencing-associated cost is still a concern, especially for sequencing a large number of individuals at high coverage, which is necessary to accurately detect rare variants.

Another constraint of the application of next-generation sequencing to association studies is error rate. Relatively high false variant discovery rates have been reported for short reads technologies even at high coverage, e.g., Illumina Solexa (6.3%) and ABI SOLiD (7.8%).[10] Given these concerns, there is interest in exploring alternative technologies after the variant discovery stage to extract information from targeted genetic regions, such as customized genotyping or the development of an exome genotyping chip. Compared to next-generation sequencing platforms, high throughput genotyping technologies can be more cost effective and less error prone.

In this article, the plausibility of applying customized genotyping and next-generation sequencing in replication studies is explored from a combined genetic epidemiology and population genetics perspective. In order to avoid spurious or false positive findings in association mapping, replicating significant associations discovered in an exploratory sample (stage 1) with an independent data set (stage 2) is an indispensible part of every genetic association study. It has been demonstrated that for the analysis of rare variants, both single marker and multivariate methods used for the analyzing common variants are underpowered because of extreme allelic heterogeneity.[6,11,12] Therefore, for mapping rare variants, gene-based tests are usually performed, where multiple rare variants in a gene region are jointly analyzed. Many gene-based tests have been proposed such as the combined multivariate and collapsing method (CMC),[6] the weight sum statistic (WSS),[12] and the test of aggregated number of rare variants (ANRV).[13] To replicate significant findings in stage 1 studies, two different strategies can be used. As a first strategy, only the variants at the nucleotide sites uncovered from the original sample are followed up. With this strategy, novel nucleotide sites that are present

[1]Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX 77030, USA; [2]Department of Statistics, Rice University, Houston, TX 77005, USA
*Correspondence: sleal@bcm.edu
DOI 10.1016/j.ajhg.2010.10.025.

only in the stage 2 sample will not be incorporated in the replication study. This constitutes a replication in a "strict" sense, i.e., both the gene region and the variants uncovered in the stage 1 sample are followed up in the replication sample. When only variants uncovered in the stage 1 sample are of interest, genotyping is sufficient. We will refer to this replication strategy as "variant-based." An alternative strategy is to follow-up the entire gene region identified in the stage 1 sample. For this design, analysis of the stage 2 sample is not restricted to the nucleotide sites uncovered in stage 1. Variants from novel sites in the replication sample are also assessed for their associations with the phenotype of interest. We will refer to this design as "sequence-based" replication. With this strategy, sequencing the target gene in the stage 2 sample is necessary. The efficiencies of the two proposed strategies are compared.

The power of the two replication strategies is dependent on the percentage of causal variant sites that were uncovered for the gene region in the stage 1 sample. If the stage 1 sample is small, there can be an advantage to sequence-based replication, because many low-frequency variants may not have been observed. However, the difference between variant-based and sequence-based replication strategies will diminish if a majority of causal variants can be uncovered in stage 1. Discovery of SNPs via population-based samples has been addressed previously.[14,15] However, in these studies the population genetic models employed were overly simplistic; they did not incorporate complex human demographic history and purifying selection, which are well-known factors that can affect rare variant site frequency spectrums.[16] A rigorous population genetic model for Africans was used with parameters estimated from sequence data.[17] Together with realistic phenotypic models motivated by complex traits, we investigate the probability of uncovering rare variants in the context of case-control studies and demonstrate their impact on the relative performances of sequence- and variant-based replication.

Additionally, the relative power of the two replication strategies will also be affected by the error rates of next-generation sequencing and customized genotyping technologies that are employed. To assess the impact of sequencing error on the power of rare variant association mapping, a data-based sequencing error model is used. The parameters of the sequencing error model were estimated according to reported false-positive and false-negative variant discovery rates from commonly used next-generation sequencing platforms.[7,10,18,19]

It is demonstrated through extensive simulations that the sequence-based replication is more powerful than variant-based replication for both small- and large-scale studies. In the ideal scenario where sequencing and genotyping are both of perfect quality, for small-scale studies with several hundred cases and controls, a large proportion of variant nucleotide sites will not be uncovered. However, uncovered rare variants in small-scale genetic studies can account for more than 80% of the locus-specific population attributable risk (PAR). Therefore, the power advantage of sequencing can be small. For large-scale studies with thousands of cases and controls, more than 90% of the causative variant nucleotide sites can be uncovered and nearly 100% of the locus PAR can be explained by the uncovered rare variants. As a result, genotyping can be a temporal solution for replicating stage 1 studies if stage 1 and 2 samples are drawn from the same population. Resequencing-based replication studies have an irreplaceable advantage in that novel variants can be discovered. This benefit is more pronounced when the stage 1 sample is small. In the presence of sequencing errors, genotyping errors, and unconverted genotyping assays, the relative performances of two replication strategies remain largely unchanged. We show that the power for sequence-based association mapping is only slightly impacted by currently attainable levels of error rates, for example, false-positive rate/false-negative rates of 6.3%/1%[10] or 10%/5%.[19]

In order to further illustrate the relative performances of sequence-based and variant-based replication, phenotype data on energy metabolism traits and sequence data from the Dallas Heart Study on the *ANGPTL3* (MIM 604774), *ANGPTL4* (MIM 605910), *ANGPTL5* (MIM 607666), and *ANGPTL6* (MIM 609336) genes[4,5] were analyzed with CMC, WSS, and ANRV. The results provide solid support for the simulation experiments.

Material and Methods

It is assumed that for a gene region of length L there are S variant sites in the study population. The major allele at site s is denoted by a^s, whereas the minor allele is labeled A^S. The underlying $L_{\text{-site}}$ genotype of an "individual" i is coded by a vector, i.e., $\vec{X}_i = (x_i^1, \cdots, x_i^S, x_i^{S+1}, \cdots, x_i^L)$. Without loss of generality, sites $1, \cdots, S$ are assumed to be variant sites in the population. Dominant genotype coding is adopted for variant nucleotide sites, i.e.,

$$x_i^s = \begin{cases} 1 & \text{if the genotype at nucleotide site } s \text{ is } A^s a^s, \text{ or } A^s A^s \\ 0 & \text{otherwise} \end{cases}, s = 1, \cdots, S.$$

The genotypes at monomorphic sites are identically coded as 0, i.e., $x_i^s = 0, i = S+1, \cdots, L$.

According to the approach in Li and Leal,[6] the collapsed genotype is introduced with an indicator function $\delta(\bullet)$, i.e.,

$$X_i = \delta\left(\sum_{s=1}^{L} x_i^s > 0\right).$$

The affection status for individual i is encoded by a binary variable Y_i, which takes value 1 if the individual is affected and 0 otherwise.

Probabilistic Model for Sequencing Errors

Because of the presence of sequencing errors, the observed genotype of an "individual" i may be different from the true underlying genotype. The observed genotype from sequence data is given by $\vec{Z}_i = (z_i^1, z_i^2, \cdots, z_i^L)$, where

$$z_i^s = \begin{cases} 1 & \text{if the genotype at nucleotide site } s \text{ is called as } A^s a^s, \text{ or } A^s A^s \\ 0 & \text{otherwise.} \end{cases}$$

The corresponding collapsed genotype Z_i is similarly defined as $Z_i = \delta(\sum_{s=1}^{L} z_i^s > 0)$.

Two types of sequencing error events are given probabilistically.[10] First, a false-positive event is defined as $\{z_i^s = 1, x_i^s = 0\}$, where a nonvariant genotype at nucleotide site s is falsely called a variant. The error rate that corresponds to the false-positive event is defined as the conditional probability $e_{01}^s = P(z_i^s = 1|x_i^s = 0)$. Second, a false-negative event can be defined as $\{z_i^s = 0, x_i^s = 1\}$ where a variant at site s is falsely called homozygous for the reference allele. Its error probability is defined as $e_{10}^s = P(z_i^s = 0|x_i^s = 1)$.

To measure and report sequencing error at rare variant nucleotide sites, it is common to use *false-positive discovery rate* and *false-negative error rate,*[7,10,18,19] i.e.,

$$\widehat{FP} = \frac{\sum_{i,s} \delta(z_i^s = 1, x_i^s = 0)}{\sum_{i,s} \delta(z_i^s = 1)}, \widehat{FN} = \frac{\sum_{i,s} \delta(z_i^s = 0, x_i^s = 1)}{\sum_{i,s} \delta(x_i^s = 1)}. \quad (1)$$

Empirical estimates of false-positive and false-negative rates are usually obtained by comparing next-generation sequencing with less error-prone technologies, e.g., Sanger sequencing or customized genotyping.[10,19]

By using reported false-positive and false-negative rates, base-pair error rates can be calculated as

$$FP = \frac{\sum_{s=1}^{L} e_{01} P(x_i^s = 0)}{\sum_{s=1}^{L} e_{01} P(x_i^s = 0) + (1 - e_{10}) P(x_i^s = 1)}, FN = \frac{\sum_{s=1}^{L} e_{10} P(x_i^s = 1)}{\sum_{s=1}^{L} P(x_i^s = 1)}. \quad (2)$$

The observed number of carriers of rare variants at site s in cases and controls are defined, respectively, as m_A^s and m_U^s. For a sample with R_A cases and R_U controls, m_A^s, m_U^s marginally follow binomial distribution based upon the above sequencing error model, i.e.,$m_A^s = \sum_i \delta(x_i^s = 1, Y_i = 1) \sim Binom(R_A, p_A^s)$, $m_U^s = \sum_i \delta(x_i^s = 1, Y_i = 0) \sim Binom(R_U, p_U^s)$, where parameters $\vec{p}_A = (p_A^s)_{s=1,\cdots,L}$, $\vec{p}_U = (p_U^s)_{s=1,\cdots,L}$ are given by

$$\begin{cases} p_A^s = P(z_i^s = 1 \mid Y_i = 1) = (1 - e_{10}) \\ \quad \times P(x_i^s = 1 \mid Y_i = 1) + e_{01} \times P(x_i^s = 0 \mid Y_i = 1) \\ p_U^s = P(z_i^s = 1 \mid Y_i = 0) = (1 - e_{10}) \\ \quad \times P(x_i^s = 1 \mid Y_i = 0) + e_{01} \times P(x_i^s = 0 \mid Y_i = 0) \end{cases}, s = 1, \cdots, L.$$

Models of Genotyping Errors

It is assumed that a set K of rare variant sites are uncovered in the stage 1 sample. Rare variants from sites K are genotyped and followed up in the stage 2 replication sample. Although the accuracy for commercially available genotyping array is high, the error rate for customized genotyping is not negligible.[20] Additionally, assays on customized probes may have a low conversion rate.

The observed locus genotype from genotyping data is denoted by $\vec{W}_i = (w_i^1, w_i^2, \cdots, w_i^L)$, where

$$w_i^s = \begin{cases} 1 \text{ if the genotype at nucleotide site } s \text{ is called as } A^s a^s, \\ \quad \text{or } A^s A^s \\ 0 \qquad\qquad \text{otherwise} \end{cases}, s \in K.$$

The corresponding collapsed genotype W_i is similarly defined. For a converted assay, the genotyping error is traditionally measured as sample error rate (SER),[21,22] i.e., $SER_s = P(w_i^s = 1, x_i^s = 0|I_C^s = 1) + P(w_i^s = 0, x_i^s = 1|I_C^s = 1), s \in K$, where I_C^s is an indicator for an assay being successful (e.g., converted with genotype calls generated) at nucleotide site s. Similar to sequencing errors, the genotyping error rates at converted probes are defined as:

$$f_{01}^s = P(w_i^s = 1 \mid x_i^s = 0, I_C^s = 1),\ f_{10}^s = P(w_i^s = 0 \mid x_i^s = 1, I_C^s = 1), S \in K. \quad (4)$$

To facilitate comparisons of the two replication strategies, an error ratio ER is introduced to measure the relative error rates for two types of sequencing and customized genotyping errors. It is assumed that $ER = \sum_{s \in K} f_{10}^s / \sum_{s=1}^{L} e_{10}^s = \sum_{s \in K} f_{01}^s / \sum_{s=1}^{L} e_{01}^s$. When the two replication strategies are compared in the presence of imperfect technologies, two error ratios are used, i.e., $ER = 1$ or $ER = 0.5$. The rate of success for a given assay at site s, i.e., $P(I_C^s = 1)$, is assumed to be 90%.

For a genotyped sample with R_A cases and R_U controls, the observed counts of carriers of variants at each nucleotide site are denoted by $\vec{N}_A = (n_A^s)_{s \in K}$, $\vec{N}_U = (n_U^s)_{s \in K}$. The counts at nucleotide site s follow a binomial distribution marginally, i.e., $n_A^s \sim Binom(R_A, q_A^s)$, and $n_U^s \sim Binom(R_U, q_U^s)$. The parameters $\vec{q}_A = (q_A^s)_{s \in K}$, $\vec{q}_U = (q_U^s)_{s \in K}$ are provided by

$$\begin{cases} q_A^s = P(z_i^s = 1 \mid Y_i = 1) = (1 - f_{10}^s) \times P(I_C^s = 1) \\ \quad \times P(x_i^s = 1 \mid Y_i = 1) + f_{01}^s \times P(I_C^s = 1) \times P(x_i^s = 0 \mid Y_i = 1) \\ q_U^s = P(z_i^s = 1 \mid Y_i = 0) = (1 - f_{10}^s) \times P(I_C^s = 1) \\ \quad \times P(x_i^s = 1 \mid Y_i = 0) + f_{01}^s \times P(I_C^s = 1) \times P(x_i^s = 0 \mid Y_i = 0) \end{cases}, s \in K. \quad (5)$$

Power Calculation for Sequence-Based and Variant-Based Replication

Several test statistics for mapping rare variants have been proposed, such as CMC,[6] WSS,[12] and ANRV.[13] The ANRV test was developed to detect associations with quantitative trait, but can be easily generalized to the study of binary traits. These tests have been shown to be more powerful than multivariate methods such as Hotelling T^2. Here, CMC, WSS, and ANRV are used in the comparisons of variant-based and sequence-based replication. The CMC has a closed form exact distribution, which makes it computationally efficient for candidate gene, exome-, and genome-wide studies. The power of the permutation-based WSS and ANRV methods were evaluated for small-scale candidate gene studies. For all the scenarios evaluated, WSS and ANRV are more powerful than CMC, but the comparisons for the two replication strategies are largely unaffected by the choice of the test statistics.

For both sequence- and variant-based replication with CMC, the association tests in both the stage 1 and stage 2 studies are implemented via Fisher exact test, which compares rare variant carrier frequencies between cases and controls. When the WSS and ANRV statistic are used, to guarantee that type I error is well controlled, p values are estimated empirically based upon 2000 permutations for each replicate. Because of the computation intensity of estimating small empirical p values, the WSS and ANRV were not used for the evaluation of power to replicate large-scale studies.

The test statistics used for the stage 1 and sequence-based stage 2 studies are denoted by $T^{S1}$and T^{seq}, respectively. The power of successfully replicating a true significant association from the stage 1 study is investigated, i.e., $P_{H^A}(|T^{seq}| > z_{1-\alpha_{S2}/2} || T^{S1}| > z_{1-\alpha_{S1}/2})$, where α_{S1} and α_{S2} are significance levels used for stage 1 and the replication study. Because the statistics are conditionally independent given the parameters $\vec{p}_A$ and $\vec{p}_U$, the following equation must be satisfied,

$$P_{H^A}\left(|T^{seq}| > z_{1-\alpha_{S2}/2}, |T^{S1}| > z_{1-\alpha_{S1}/2} \,|\, \vec{p}^{\,A}, \vec{p}^{\,U}\right) = P_{H^A}\left(|T^{seq}| > z_{1-\alpha_{S2}/2} \,|\, \vec{p}^{\,A}, \vec{p}^{\,U}\right) P_{H^A}\left(|T^{S1}| > z_{1-\alpha_{S1}/2} \,|\, \vec{p}^{\,A}, \vec{p}^{\,U}\right). \quad (6)$$

The power for variant-based replication is given by $P_{H^A}(|T^{var}| > z_{1-\alpha_{S2}/2} | |T^{S1}| > z_{1-\alpha_{S1}/2})$ but unlike for sequence-based replication, the test statistics T^{var}, T^S are not conditionally independent. Under the alternative hypothesis, the distribution of T^{var} depends on K, which is the set of rare variant sites uncovered in stage 1. Because it is impossible to enumerate the parameter space of $(\vec{p}_A, \vec{p}_U, \vec{q}_A, \vec{q}_U, K)$, an efficient Monte Carlo algorithm was developed to calculate replication power for both sequence-based and variant-based strategies (see Appendix).

For notational convenience, the ratio of total frequencies of uncovered rare variants to the total frequencies of all locus rare variants (including those that are not uncovered) is denoted by $f_{MAF} = \sum_{s\in K} P(x_i^s = 1) / \sum_{s=1}^{S} P(x_i^s = 1)$.

In addition, the ratio

$$f_{PAR} = \sum_{s\in K\cap C} P(x_i^s = 1 \,|\, Y_i = 1) / \sum_{s\in C} P(x_i^s = 1 \,|\, Y_i = 1), \quad (7)$$

represents the proportion of locus PAR that can be explained by the uncovered causal variants. The probability $\sum_{s\in C} P(x_i^s = 1|Y_i = 1)$ is asymptotically equivalent to the epidemiological definition of PAR, which is the reduction of disease incidence rate that would be observed if the population were unexposed, i.e., if there were no carriers of causative variants.

The power comparisons were performed for both the small- and large-scale genetic studies. In order to have sufficient power to detect associations, 250 cases/250 controls or 500 cases/500 controls were used for both the stage 1 and 2 samples in a small-scale study. For the scenario of a large-scale study, 2000 cases/2000 controls, as well as 3500 cases/3500 controls were examined. For small-scale studies, examples are given with significance levels $\alpha_{S1} = \alpha_{S2} = 0.05$. The commonly accepted exome-wide significance level $\alpha_{S1} = \alpha_{S2} = 2.5 \times 10^{-6}$ is used for large-scale genetic studies, which is based upon Bonferroni corrections for testing 20,000 genes. The empirical power for each scenario was estimated with 10,000 replicates.

Simulations of Complex Demographic Models and Selections

To compare relative efficiencies of sequence-based and variant-based replication strategies, population genetic data were generated with forward time simulations.[23] Genetic data for the African population were generated. The parameters for demographic changes and selections were estimated in Boyko et al.[17] The demographic change for the African population is described with a two-epoch instant change model. Purifying selection was also simulated, with u and $2u$ being the selective disadvantage of heterozygous and homozygous new mutations. Scaled fitness effect $\gamma = 2N_{curr}u$ (where N_{curr} is the current effective population size) is assumed to follow a gamma distribution, i.e.,

$$\gamma = -r, r \sim \frac{b^a}{\Gamma(a)} r^{a-1} \exp(-br), \text{ where } a = 0.184, b = 8200.$$

The model was shown to be parsimonious and fit the data well. A mutation rate of $\mu_S = 1.8 \times 10^{-8}$ per nucleotide site per generation is assumed. Because the average length for human gene-coding region is 1500 base pairs (bp) long,[24,25] $L = 1500$ bps was used in the simulation to specify the locus-scaled mutation rate. Based upon the above parameter specification, 100 sets of rare variant site frequencies were generated. As suggested by Kryukov et al.,[26] only nonsynonymous (NS) variants were used in the analysis in order to increase the signal to noise ratio and reduce the negative impact of nonfunctional variants on power.

Generations of Phenotypic Model

Phenotypic effects of rare NS variants are assumed independent of their fitness.[24] Fifty percent of the rare NS variants (with MAF $\leq$ 0.01) are randomly picked to be causal and affect the binary phenotype of interest. Based upon surveys of multifactorial diseases,[27] two types of phenotypic models were considered. For the first type of model, the genetic effects of causal variants are inversely correlated with their MAFs. It is assumed that causal variants with the smallest (or largest) MAFs (i.e., p_{min} or p_{max}) have largest (or smallest) log odds ratio (log-OR) of β_{max} (or β_{min}), respectively. For a causal variant with MAF p_i, the log-OR follows the interpolation relation: $\beta_i = \beta_{max} + (\beta_{max} - \beta_{min})/(p_{max} - p_{min}) \times (p_i - p_{min})$, $i \in C$. The ORs for causal variants thus satisfy an exponential relationship with their MAFs. A choice of $\beta_{max} = \log(10)$, $\beta_{min} = \log(2)$ was used. For the second type of model, each causal variant has equal disease odds, which is given by $\beta_i = \log(3)$, $i \in C$. Under both types of models, the affection status for an individual with multisite genotype $\vec{X}$ is assigned by the following model:

$$P(Y_i = 1 \,|\, \vec{X}) = \frac{\exp\left(\beta_0 + \sum_{i\in C} \beta_s x_i^s\right)}{1 + \exp\left(\beta_0 + \sum_{i\in C} \beta_s x_i^s\right)}. \quad (8)$$

A baseline penetrance of 0.01 is assumed, which gives $\beta_0 = \log(0.01/(1 - 0.01))$.

Applications to the Dallas Heart Study Sequence Data

In order to illustrate the relative efficiency of sequence-based versus variant-based replication strategies, a data set from DHS was analyzed. The data set is a multiethnic population based sample (1830 African Americans [AA], 601 Hispanics [H], 1045 European Americans [EA], and 75 individuals from other ethnic groups) from Dallas County residents whose lipids and glucose metabolism have been characterized and recorded.[28,29] In order to investigate how sequence variations in *ANGTPL3, 4, 5*, and *6* influence energy metabolism in humans, coding regions of the four genes were sequenced via DNA samples obtained from 3551 participants in DHS.[4] A total of 348 nucleotide sites of sequence variations were uncovered in the four genes. Most of them are rare and 86% of them have MAFs $< 1\%$.[4] Nine phenotypes were measured and tested for their associations with rare genetic variants, i.e., body mass index (BMI), diastolic blood pressure (DiasBP), systolic blood pressure (SysBP), total cholesterol level (TCL), low-density lipoprotein (LDL), high-density lipoprotein (HDL), triglyceride (TG), very-low-density lipoprotein (VLDL), and glucose. As a first analysis, to mimic the scenario of stage 1 study, individuals with quantitative trait values in the top and bottom 10% of the phenotypic distributions were used to form a case-control data set. Individuals with intermediate quantitative trait values, i.e., in the range of 10%–35% and 65%–90%, were used as a replication sample. Sequence-based and variant-based replications were compared for the replication data set.

Among the identified significant results, the association between TG level and rare variants in the *ANGPTL4* gene was supported by in vitro functional studies and was replicated with an independent data set.[4,5] It is highly likely to be a true association. Therefore, a second experiment was performed to estimate the

Table 1. Discovery of Rare Variants in Small- and Large-Scale Genetic Studies

Number of Cases/Controls in Stage 1 and 2 Samples	Proportion of Rare Variant Sites Uncovered[a]		Proportion[a]	
	All	Causal	Locus PAR Explained by Uncovered Causal Rare Variants	Causal Variant Sites among All Uncovered Rare Variant Sites
Variable Effect Phenotypic Model				
250/250[b]	0.432	0.599	0.917	0.686
500/500[b]	0.524	0.687	0.950	0.645
2000/2000[c]	0.728	0.887	0.992	0.599
3500/3500[c]	0.808	0.943	0.996	0.572
Fixed Effect Phenotypic Model				
250/250[b]	0.369	0.468	0.937	0.629
500/500[b]	0.455	0.547	0.960	0.591
2000/2000[c]	0.664	0.757	0.993	0.559
3500/3500[c]	0.751	0.827	0.995	0.538

[a] Rare variant data were simulated via African rare variant site frequency spectrums, and case control data sets were generated with variable and fixed effect phenotypic models. A total of 10,000 replicates were generated and each reported value within the table was obtained by averaging over replicates where significant stage 1 results were obtained.
[b] Small-scale study: $\alpha_{S1} = 0.05$.
[c] Large-scale study: $\alpha_{S1} = 2.5 \times 10^{-6}$.

power for replicating the association between TG and rare variants in the *ANGPTL4* gene. Individuals with TG levels in the range of top 35% and bottom 35% were used to form a case-control "cohort." 50% of the cases and 50% of the controls from the "cohort" were randomly selected as the data set for the stage 1 study. The remaining 50% of cases and controls are used as the stage 2 replication sample. The process was repeated 1000 times, and for each replicate, sequence-based and variant-based replication was performed. The fraction of significant stage 1 studies that were successfully replicated in stage 2 was reported for the comparison of two different replication strategies. Association tests with CMC, WSS, and ANRV were performed.

Results

Discovery Rate of Rare Variant Sites and Frequencies

Rare variant discovery rates were compared under the assumption that sequencing data are of perfect quality (Table 1). When sequencing is not perfect, the fractions of uncovered rare variants will be lowered by false-negative rate and additionally a portion of observed variants can be false positives.

When a variable effect model is used, relatively low proportions of variant nucleotide sites are uncovered for small-scale studies. For example, in a sample of 250 cases/250 controls, 43.2% of all variant nucleotide sites and 59.9% of causal variant nucleotide sites are uncovered. On the other hand, a fairly large portion of locus PAR (91.7%) can be explained by the uncovered variants. When a fixed effect model is employed, the results are very similar (Table 1). A slightly lower portion of variants can be uncovered but the uncovered variants explain a higher fraction of locus PAR.

When a sample of 500 cases/500 controls was analyzed, a higher proportion of variant sites are uncovered; however, considerable fractions of rare variant sites in the population are still not observed in the sample for both fixed and variable effect models. For example, when the fixed genetic effect model is used, only 45.5% of causal variant sites are uncovered.

For an exome-wide significance level $\alpha_{S1} = 2.5 \times 10^{-6}$, a much larger sample size is necessary to obtain sufficient power to detect significant associations.[26] Under the variable genetic effect model, when a sample of 2000 cases and 2000 controls was analyzed, for a gene region that attains exome-wide significance, a much larger fraction (72.8%) of rare variant nucleotide sites are present in the data set, and nearly all (88.7%) causal nucleotide sites are uncovered. These uncovered variants explain nearly 100% of the locus PAR. Therefore, in principle, when a large stage 1 sample is analyzed, the advantage of sequencing for novel SNP discoveries diminishes as long as the stage 2 samples are drawn from the same population. Similar results hold if a fixed effect model is assumed for the binary phenotype.

Because affected individuals are enriched in a case-control sample, nucleotide sites containing causal variants have a much higher probability of being uncovered than noncausal variant sites. For example, if a fixed effect model is assumed, 62.9% of the sites uncovered are causal variant sites for a sample of 250 cases and 250 controls. This fraction is much higher than the proportions of causal variant sites in the general population (50%).

Power Comparisons for Sequence-Based and Variant-Based Replication Strategies

The power was compared for sequence-based and variant-based replication under different combinations of false-positive/false-negative variant discovery rates, genotyping assay success rates, and error rates.

Table 2. Power Comparisons of Sequencing-Based and Variant-Based Replication under Variable Effect Model

Number of Cases/Controls in Stage 1 and 2 Samples	Rates[a]				Power for Replication[b]	
	False Positive	False Negative	Assay Succes	Error Ratio	Sequence-Based	Variant-Based
250/250[c]	0	0	1	1	0.542	0.507
	1%	4%	0.9	0.5	0.521	0.458
				1		0.452
	6.3%	1%	0.9	0.5	0.520	0.466
				1		0.461
	10%	5%	0.9	0.5	0.503	0.459
				1		0.447
500/500[c]	0	0	1	1	0.731	0.708
	1%	4%	0.9	0.5	0.719	0.675
				1		0.667
	6.3%	1%	0.9	0.5	0.718	0.674
				1		0.674
	10%	5%	0.9	0.5	0.701	0.672
				1		0.661
2000/2000[d]	0	0	1	1	0.827	0.825
	1%	4%	0.9	0.5	0.816	0.780
				1		0.766
	6.3%	1%	0.9	0.5	0.814	0.781
				1		0.769
	10%	5%	0.9	0.5	0.802	0.781
				1		0.755
3500/3500[d]	0	0	1	1	0.899	0.898
	1%	4%	0.9	0.5	0.893	0.870
				1		0.863
	6.3%	1%	0.9	0.5	0.893	0.868
				1		0.865
	10%	5%	0.9	0.5	0.886	0.874
				1		0.858

[a] Significance levels for small-scale study: $\alpha_{S1} = 0.05$ and $\alpha_{S2} = 0.05$.
[b] Significance levels for large-scale study: $\alpha_{S1} = 2.5 \times 10^{-6}$ and $\alpha_{S2} = 2.5 \times 10^{-6}$.
[c] The impact of different combinations of false-positive/false-negative rate, assay success rate, and genotyping and sequencing error rate ratio on the replication power is examined.
[d] The power was empirically estimated based upon 10,000 replicates.

In the ideal scenario where both sequencing and customized genotyping qualities are perfect, the power for sequence-based and variant-based replication is jointly affected by the sample size, the proportions of rare variants uncovered and the fractions of uncovered rare variant sites that contain causal variants. For most of the examined scenarios, the power of sequence-based replication is consistently higher than variant-based replication when CMC is used for analysis. For example, under the variable effect model (Table 2), for a sample size of 250 cases and 250 controls, the power for sequence-based replication is 54.2% while the power of variant-based replication is 50.7%. For large-scale genetic studies, the power hardly differs between sequence- and variant-based replication. This is because a large proportion of variant sites are uncovered in the stage 1 sample, and the uncovered variants account for nearly 100% of the locus PAR. For example, for a gene that attains exome-wide significance in a sample of 2000 cases and 2000 controls, the power for sequence- and variant-based replication are, respectively, 82.7% and 82.5%.

Table 3. Power Comparisons of Sequence-Based and Variant-Based Replication under Fixed Effect Model

Number of Cases/Controls in Stage 1 and 2 Samples	Rates[a]				Power for Replication[b]	
	False Positive	False Negative	Assay Success	Error Ratio	Sequence-Based	Variant-Based
250/250[c]	0	0	1	1	0.446	0.437
	1%	4%	0.9	0.5	0.432	0.392
				1		0.386
	10%	1%	0.9	0.5	0.429	0.399
				1		0.390
	6.3%	5%	0.9	0.5	0.410	0.390
				1		0.378
500/500[c]	0	0	1	1	0.666	0.658
	1%	4%	0.9	0.5	0.650	0.619
				1		0.607
	6.3%	1%	0.9	0.5	0.652	0.623
				1		0.613
	10%	5%	0.9	0.5	0.632	0.619
				1		0.600
2000/2000[d]	0	0	1	1	0.765	0.767
	1%	4%	0.9	0.5	0.747	0.703
				1		0.689
	6.3%	1%	0.9	0.5	0.746	0.705
				1		0.694
	10%	5%	0.9	0.5	0.724	0.700
				1		0.669
3500/3500[d]	0	0	1	1	0.875	0.878
	1%	4%	0.9	0.5	0.872	0.841
				1		0.834
	6.3%	1%	0.9	0.5	0.870	0.845
				1		0.835
	10%	5%	0.9	0.5	0.856	0.843
				1		0.825

[a] Significance levels for small-scale study: $\alpha_{S1} = 0.05$ and $\alpha_{S2} = 0.05$.
[b] Significance levels for large-scale study: $\alpha_{S1} = 2.5 \times 10^{-6}$ and $\alpha_{S2} = 2.5 \times 10^{-6}$.
[c] The impact of different combinations of false-positive/false-negative rate, assay success rate, and genotyping and sequencing error rate ratio on the replication power is examined.
[d] The power was empirically estimated based upon 10,000 replicates.

The power for sequence- and variant-based replication is negatively impacted by sequencing and genotyping errors. The impact of sequencing error is small. If the fixed effect model is assumed (Table 3), for a sample size of 250 cases and 250 controls, the power of sequenced-based replication is 44.6% in the absence of sequencing errors; when a false-positive rate of 10% and a false-negative rate of 5% are assumed, the power drops to 41.0%. Although a lower error rate is assumed for customized genotyping, the advantage of sequence-based replication remains. For instance, in this scenario, for a genotyping call rate of 90% and an error rate ratio of 0.5, the power for variant-based replication is 39.0%.

Comparisons of two replication strategies were also made when WSS and ANRV were used for analysis of both the stage 1 and 2 data sets (Table S1 available online). Although the power is consistently higher for the WSS and the ANRV than for the CMC, the relative performances for sequence-based and variant-based replication are similar in most situations. One noticeable difference is that

Table 4. Analyses of Sequence Data from the *ANGPTL3, 4, 5,* and *6* Genes

Trait	p Values			Proportion	Ratio	Number of Rare Variants Observed	
	Stage 1 Analysis[a] (CMC/WSS/ANRV)	Sequence-Based Replication[b] (CMC/WSS/ANRV)	Variant-Based Replication[b] (CMC/WSS/ANRV)	Nucleotide Sites Uncovered in Stage 1	Rare Variant Freq in Stage 1 Sample/ Rare Variant Freq. in Entire Sample	Sequence-Based Replication	Variant-Based Replication
ANGPTL3							
TCL	0.028/0.022/0.033	0.522/0.493/0.534	0.726/0.724/0.845	0.30	0.87	46/51	39/40
ANGPTL4							
LDL	0.021/0.025/0.022	0.272/0.218/0.194	0.508/0.473/0.387	0.35	0.94	78/62	70/60
TG	0.027/0.022/0.009	0.025/0.016/0.019	0.039/0.028/0.027	0.26	0.92	77/51	69/46
VLDL	0.037/0.024/0.011	0.031/0.020/0.024	0.031/0.023/0.025	0.26	0.92	75/51	69/46
ANGPTL5							
BMI	0.029/0.021/0.033	0.464/0.407/0.670	0.451/0.423/0.661	0.5	0.95	67/71	63/67
HDL	0.025/0.022/0.017	1.0/0.959/0.729	0.772/0.760/0.947	0.5	0.95	63/66	61/60
ANGPTL6							
BMI	0.001/0.001/0.001	0.909/0.874/0.823	0.794/0.774/0.729	0.21	0.78	42/40	33/30

[a] For each phenotype analyzed, individuals with QTVs from the top and bottom 10% were used as a stage 1 sample.
[b] Individuals with QTVs in the range of 10%–35% and 65%–90% were used as the replication sample.

sequencing error tends to have a slightly more negative impact on power for studies via the WSS. The power for variant-based replication can be even higher than sequence-based replication when WSS is implemented for data analysis. For example, for a sample of 250 cases and 250 under the fixed effect model, where the false-positive rate is 10%, the false-negative rate is 5%, and error rate ratio $ER = 0.5$, the power for sequence-based replication (53.3%) is even lower than that for variant-based replication (54.7%).

Applications to the Dallas Heart Study Data

For the first analysis, the stage 1 and 2 data from the *ANGPTL3, 4, 5*, and *6* genes are analyzed. Although a small sample size (individuals with trait values in the top and bottom 10%) was used for the stage 1 study, multiple (novel) associations were detected with the CMC and WSS (Table 4), i.e., (1) TCL with *ANGPTL3* ($p_{CMC} = 0.028$, $p_{WSS} = 0.022$, $p_{ANRV} = 0.033$), (2) LDL with *ANGPTL4* ($p_{CMC} = 0.021$, $p_{WSS} = 0.025$, $p_{ANRV} = 0.022$), (3) TG with *ANGPTL4* ($p_{CMC} = 0.027$, $p_{WSS} = 0.022$, $p_{ANRV} = 0.009$), (4) VLDL with *ANGPTL4* ($p_{CMC} = 0.037$, $p_{WSS} = 0.024$, $p_{ANRV} = 0.011$, (5) BMI with *ANGPTL5* ($p_{CMC} = 0.029$, $p_{WSS} = 0.021$, $p_{ANRV} = 0.033$), (6) HDL with *ANGPTL5* ($p_{CMC} = 0.025$, $p_{WSS} = 0.022$, $p_{ANRV} = 0.017$), and (7) BMI with *ANGPTL6* ($p_{CMC} = 0.001$, $p_{WSS} = 0.001$, $p_{ANRV} = 0.001$). Among these, the association between BMI and *ANGPTL6* is significant even after Bonferroni correction for testing multiple genotypes and phenotypes. For most of the analyses, approximately 25%–40% of the nucleotide sites observed in the entire DHS sample are also observed in stage 1. The stage 2 replication sample consists of individuals with less extreme quantitative trait values. To ensure that the power of the stage 2 sample is adequate, the stage 2 samples are chosen to be larger than the stage 1 sample size. Two of the seven identified associations in the stage 1 sample were successfully replicated by both sequence- and variant-based replication, i.e., associations between TG and *ANGPTL4* as well as between VLDL and *ANGPTL4*. For both associations, sequence-based replication has slightly smaller p values.

For the second analysis, the empirical power for replicating the validated association between TG and rare variants in *ANGPTL4* gene was compared. When the CMC is used, the empirically estimated power for sequence-based and variant-based replication is 65.3% and 62.7%, respectively. The power for sequence-based replication is only slightly better. This is very compatible with observations from simulated data. When the WSS and ANRV are used, estimated power is greater but the relative performances (69.3% versus 67.0% for WSS and 68.2% versus 64.4% for ANRV) are concordant.

Discussion

In this article, sequence-based replication and variant-based replication for complex trait rare variant association studies were compared with a rigorous population genetic framework. It is demonstrated that in the ideal scenario where sequencing and genotyping are both of perfect quality, sequence-based replication is consistently more powerful. However, because the uncovered variants can account for a large proportion of locus PAR even for stage 1 studies with only a few hundred samples, the advantage in power can be very small if stage 1 and stage 2 samples are drawn from the same population. The power of sequence- and variant- based replication studies is negatively impacted by sequencing and genotyping errors. For currently

attainable levels of sequencing errors, the impact is minimal, and the advantage of using sequence-based replication studies remains.

It has been found previously that rare variants tend to be population specific.[27] Many studies have suggested that disease-associated variants in different populations can have very different frequencies. For example, the E40K variant in *ANGPTL4* gene was shown to be associated with TG levels. The MAF for E40K is approximately 3% in European-Americans but is very rare in African-Americans and Hispanics.[4] These differences can be observed in even more closely related populations, for example, rare variants in *CFTR* (MIM 602421), *BRCA1* (MIM 113705), and *BRCA2* (MIM 600185) genes have higher frequencies in the Ashkenazi Jewish population compared to other European Jewish populations such as Sephardic Jews and also to non-Jewish populations.[30,31] Population-specific diversity of variant frequencies and sites is believed to be more pronounced for rare variants than for common variants because rare variants tend to be younger and occur more recently in human history.[27] When stage 2 samples are drawn from a different population than the stage 1 samples, the variant-based replication studies may be at a disadvantage. Given that the demographic and selection models incorporating complex migration and admixtures are limited,[17] simulation studies for variant discovery with multiethnic samples still remain to be explored. Evaluating the benefits and drawbacks of replication studies with samples from different populations will be very important.

This article also provides a model for incorporating sequencing error uncertainties into downstream association analysis. Some of the error rates discussed in this article (e.g., $FP = 6.3\%$, $FN = 1\%$) are attained when a saturated coverage depth is used. With the maturation of next-generation sequencing technologies, as well as the development of more sophisticated genotype calling algorithms, such as using pooled population samples,[32] even lower rates should be attainable in the near future. For currently attainable levels of sequencing errors, their impact on the power of rare variant association mapping is minimal.

The WSS is more sensitive to sequencing errors than the CMC and ANRV, because it assigns higher weights to lower-frequency variants.[12] Sequencing error can create false-positive variant sites that have very low allele frequencies.[32] Therefore, in some scenarios with higher sequencing error rates, when analysis is performed with WSS, sequence-based replication can be less powerful than variant-based replication.

Although the error model for Sanger sequencing is well known, the error model for next-generation sequencing has not been extensively evaluated.[19,33,34] Because of the paucity of information on error rate estimation, our error model assumes equal error rates across different nucleotide sites. This is certainly an oversimplification. In practice, for different frequency bins, different false-positive and false-negative discovery rates can be expected. The proposed error model can be refined and applied to specific frequency bins when corresponding false-positive and false-negative rate estimates become available. Various studies have shown that nonrandom systematic errors exist and cannot be ignored.[10,32] The systematic errors can be dependent on the genetic context of the variants. However, given that the main interest lies in gene-based association mapping, modeling error rate variation across different nucleotide sites may not be a necessity because only their overall impact in the gene region needs to be assessed. In particular, when the CMC method is used, the power is affected only by total number of errors, but not by the nucleotide sites where those errors occur. The error rates used in our model can be taken as locus averages. When comparing variant-based replication, genotyping error rates are assumed to be equal or lower than sequencing error rates. It can be argued that this is sensible for two reasons. First, genotyping technology is more mature than sequencing, so it tends to have a lower error rate per base pair. Second, because customized genotyping is performed only at nucleotide sites with putative polymorphisms, it is less error prone than sequencing where SNP discovery and genotype calling are performed simultaneously.

Population genetic data were generated through forward time simulations. Both demographic change and purifying selections are known to be important factors affecting rare variant site frequency spectrums. Therefore, they are both modeled and incorporated in the simulations. Two types of phenotypic models were considered. According to surveys on multifactorial disorders, most of the uncovered disease-causative rare variants have ORs between 2 and 4.[27] Although results are shown only for $OR = 3$, $OR = 2$ and $OR = 4$ were also investigated, and the conclusions remain the same. On the other hand, variable effect models also have empirical support. It has been observed that lower-frequency rare genetic variants tend to have larger disease odds compared to more frequent variants.[11,27] There is also evidence that highly penetrant rare genetic variants may be involved in the etiology of complex traits.[35,36] Because a majority of rare variants have very low frequencies, when $OR_{max} = 10$, $OR_{min} = 2$ is used, most of the uncovered rare variants have ORs ≤ 4. The results of comparisons of replication study designs remain valid and robust under both types of phenotypic models.

In the examples discussed in this article, two different significance levels are used in stage 1 ($\alpha_{S1} = 0.05$, $\alpha_{S1} = 2.5 \times 10^{-6}$). These significance levels are chosen for illustrative purposes. In practice, the significant levels used are dependent on the effective number of tests that can be performed. Currently for exome data where analysis is performed on a gene-by-gene basis, it is recommended to use an α level of 2.5×10^{-6}. This significance level is based on the Bonferroni correction for testing 20,000 genes. Because there is little linkage disequilibrium between rare variants in different genes, a Bonferroni correction will

not be overly conservative. If the analysis is not only performed on the gene level but pathway analysis is also performed, a more stringent α level is necessary. The choices for stage 2 significance levels are also for illustrative purposes. If gene(s) are found to be associated with a trait of interest with an α level that adequately controls the FWER in stage 1, it is not necessary to use the same stringent α level to replicate the association. The appropriate significance level is determined by the number of tests performed in stage 2.

In order to proceed to replication, the stage 1 study must be sufficiently powered to detect associations. For the examples given in this article, the smallest sample size shown is 250 cases/250 controls. Examples with smaller sample sizes are not shown because if a realistic complex trait model is used they will not be adequately powered. Given the cost of whole-exome sequencing studies, some existing studies use very small sample sizes. These studies are mostly for exploratory purposes and targeted at Mendelian disorders.[7,8,37–39] They will be extremely underpowered for detecting associations with rare variants involved in complex disease etiologies.

Sequencing has an irreplaceable advantage over genotyping, which is to discover novel genetic variants. The human population has experienced complex patterns of demographic expansion and purifying selection.[4,16] Large numbers of very rare variant nucleotide sites exist. Based upon observations from our extensive simulations and real data, for moderate-sized stage 1 studies, only a limited proportion of rare variant nucleotide sites can be uncovered. Identifying and cataloging rare variants themselves can be of great importance in genetic studies. Novel rare causal variants that are uncovered will help enhance the understanding of the genetic architecture of complex traits. They can also be useful for risk prediction and personalized medicine. And therefore, even if a gene is implicated in disease etiology via variant-based replication, it can still be beneficial to sequence the region in the stage 2 sample in order to uncover potential novel causal variants. For large-scale genetic studies with thousands of cases and controls the yield of sequencing, the stage 2 sample can be low, because the majority of the disease-causative variants may have been identified in stage 1.

For both stage 1 and 2 studies, it is important to be able to control for population substructure/admixture. If data from genome-wide association studies (GWAS) are available for the stage 2 sample, it can be used to control for population substructure/admixture. If GWAS data are not available, customized genotyping can be advantageous to targeted sequencing in that it can be used to genotype additional unlinked markers to control for population substructure/admixture.

With the rapid large-scale application of next-generation sequencing, understandings of genetic etiologies of rare variants will advance to an unprecedented level. Replications of significant findings will be an indispensible part of every genetic study. Sequence-based replication for both small-scale and large-scale genetic studies is advantageous and will eventually be affordable and widely applied. In the meantime, variant-based replication can be a temporal cost-effective solution for replications of genetic studies and will greatly accelerate the process of identifying disease-causative variants.

Appendix A

Algorithm 1: An Efficient Algorithm for Estimating Power for Gene-Based Replication

(1) Randomly pick one set of generated variant site frequencies and select 50% of the rare variant sites as causal, which effects the phenotype of interest.
(2) According to the chosen haplotype pool, the set of causal variant sites, and error model, the variant site carrier frequencies in cases and controls, i.e., $\vec{p}_A, \vec{p}_U$, are determined by formula (3). A stage 1 data set with O_A cases and O_U controls and replication data set with R_A cases and R_U controls are generated according to $\vec{p}_A$ and $\vec{p}_U$.
(3) Repeat steps (1) and (2) N times, and for each iteration i, the test statistics T_i^{S1}, T_i^{seq} are computed for the stage 1 data set and the replication data set.
(4) The power of replicating a significant association via gene-based replication, i.e., $P(|T^{seq}| > z_{1-\alpha_{S2}/2} || T^{S1}| > z_{1-\alpha_{S1}/2})$, can be approximated by

$$\widehat{P}\left(|T^{seq}| > z_{1-\alpha_{S2}/2} \,|\, |T^{S1}| > z_{1-\alpha_{S1}/2}\right) = \frac{1/N \sum_i \delta\left(|T_i^{S1}| > z_{1-\alpha_{S1}/2},\, |T_i^{seq}| \rangle z_{1-\alpha_{S2}/2}\right)}{1/N \sum_i \delta\left(|T_i^{S1}| > z_{1-\alpha_{S1}/2}\right)},$$

where the numerator and denominator are consistent estimators for $P(|T^{S1}| > z_{1-\alpha_{S1}/2}, |T^{seq}| > z_{1-\alpha_{S2}/2})$ and $P(|T^{S1}| > z_{1-\alpha_{S1}/2})$, respectively.

Algorithm 2: A Similar Algorithm for Estimating Power for Variant-Based Replication Is Given Below

(1) Randomly pick one set of generated variant site frequencies and select 50% of the variant sites as causal, which affects the phenotype of interest.
(2) According to the chosen haplotype pool, the set of causal variant sites, and error model, the variants site carrier frequencies in cases and controls $\vec{p}_A, \vec{p}_U$ are determined by formula (3). A stage 1 data set with O_A cases and O_U controls is generated according to $\vec{p}_A$ and $\vec{p}_U$. The set of variant nucleotide sites in the stage 1 sample is denoted by K. A corresponding replication data set of R_A cases and R_U controls is generated based upon $\vec{q}_A = (q_A^s)_{s\in K}$, $\vec{q}_U = (q_U^s)_{s\in K}$ given by formula (5).
(3) Repeat steps (1) and (2) N times, and for each iteration i, the test statistics $T_i^{S1}, T_i^{\text{var}}$ are computed for the stage 1 data set and the replication data set.

(4) The power of replicating a significant association via variant-based replication, i.e., $P(|T^{\text{var}}| > z_{1-\alpha_{S2}/2} || T^{S1}| > z_{1-\alpha_{S1}/2})$, can be approximated by

$$\widehat{P}\left(|T^{\text{var}}| > z_{1-\alpha_{S2}/2} \,\middle|\, |T^{S1}| > z_{1-\alpha_{S1}/2}\right) = \frac{1/N\sum_i \delta\left(|T_i^{S1}| > z_{1-\alpha_{S1}/2}, |T_i^{\text{var}}| > z_{1-\alpha_{S2}/2}\right)}{1/N\sum_i \delta\left(|T_i^{S1}| > z_{1-\alpha_{S1}/2}\right)},$$

where the numerator and denominator are, respectively, consistent estimators for $P(|T^{S1}| > z_{1-\alpha_{S1}/2}, |T^{\text{var}}| > z_{1-\alpha_{S2}/2})$ and $P(|T^{S1}| > z_{1-\alpha_{S1}/2})$.

Supplemental Data

Supplemental Data include one table and can be found with this article online at http://www.cell.com/AJHG/.

Acknowledgments

This research is supported by National Institutes of Health grant R01-DC03594 and 1RC2HL102926 (S.M.L.). D.J.L. is partially supported by a training fellowship from the Keck Center Pharmacoinformatics Training Program of the Gulf Coast Consortia (NIH Grant No. 5 R90 DK071505-04). We would like to thank Drs. Jonathan Cohen and Helen Hobbs for providing us with data from the Dallas Heart Study on the *ANGTPL* family genes, which was supported by National Institute of Health grant RL1HL092550 (J. Cohen). The authors would also like to thank Drs. John W. Belmont, Hua Chen, Xiang Qin, and Fuli Yu for illuminating discussions on this work. Computation for this research was supported in part by the Shared University Grid at Rice funded by NSF under Grant EIA-0216467, and a partnership between Rice University, Sun Microsystems, and Sigma Solutions, Inc.

Received: July 16, 2010
Revised: October 8, 2010
Accepted: October 26, 2010
Published online: December 2, 2010

Web Resources

The URL for data presented herein is as follows:

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

References

1. Cohen, J.C., Kiss, R.S., Pertsemlidis, A., Marcel, Y.L., McPherson, R., and Hobbs, H.H. (2004). Multiple rare alleles contribute to low plasma levels of HDL cholesterol. Science *305*, 869–872.
2. Cohen, J.C., Pertsemlidis, A., Fahmi, S., Esmail, S., Vega, G.L., Grundy, S.M., and Hobbs, H.H. (2006). Multiple rare variants in NPC1L1 associated with reduced sterol absorption and plasma low-density lipoprotein levels. Proc. Natl. Acad. Sci. USA *103*, 1810–1815.
3. Ji, W., Foo, J.N., O'Roak, B.J., Zhao, H., Larson, M.G., Simon, D.B., Newton-Cheh, C., State, M.W., Levy, D., and Lifton, R.P. (2008). Rare independent mutations in renal salt handling genes contribute to blood pressure variation. Nat. Genet. *40*, 592–599.
4. Romeo, S., Pennacchio, L.A., Fu, Y., Boerwinkle, E., Tybjaerg-Hansen, A., Hobbs, H.H., and Cohen, J.C. (2007). Population-based resequencing of ANGPTL4 uncovers variations that reduce triglycerides and increase HDL. Nat. Genet. *39*, 513–516.
5. Romeo, S., Yin, W., Kozlitina, J., Pennacchio, L.A., Boerwinkle, E., Hobbs, H.H., and Cohen, J.C. (2009). Rare loss-of-function mutations in ANGPTL family members contribute to plasma triglyceride levels in humans. J. Clin. Invest. *119*, 70–79.
6. Li, B., and Leal, S.M. (2008). Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. Am. J. Hum. Genet. *83*, 311–321.
7. Ng, S.B., Buckingham, K.J., Lee, C., Bigham, A.W., Tabor, H.K., Dent, K.M., Huff, C.D., Shannon, P.T., Jabs, E.W., Nickerson, D.A., et al. (2010). Exome sequencing identifies the cause of a mendelian disorder. Nat. Genet. *42*, 30–35.
8. Ng, S.B., Turner, E.H., Robertson, P.D., Flygare, S.D., Bigham, A.W., Lee, C., Shaffer, T., Wong, M., Bhattacharjee, A., Eichler, E.E., et al. (2009). Targeted capture and massively parallel sequencing of 12 human exomes. Nature *461*, 272–276.
9. Mamanova, L., Coffey, A.J., Scott, C.E., Kozarewa, I., Turner, E.H., Kumar, A., Howard, E., Shendure, J., and Turner, D.J. (2010). Target-enrichment strategies for next-generation sequencing. Nat. Methods *7*, 111–118.
10. Harismendy, O., Ng, P.C., Strausberg, R.L., Wang, X., Stockwell, T.B., Beeson, K.Y., Schork, N.J., Murray, S.S., Topol, E.J., Levy, S., and Frazer, K.A. (2009). Evaluation of next generation sequencing platforms for population targeted sequencing studies. Genome Biol. *10*, R32.
11. Gorlov, I.P., Gorlova, O.Y., Sunyaev, S.R., Spitz, M.R., and Amos, C.I. (2008). Shifting paradigm of association studies: value of rare single-nucleotide polymorphisms. Am. J. Hum. Genet. *82*, 100–112.
12. Madsen, B.E., and Browning, S.R. (2009). A groupwise association test for rare mutations using a weighted sum statistic. PLoS Genet. *5*, e1000384.
13. Morris, A.P., and Zeggini, E. (2010). An evaluation of statistical approaches to rare variant analysis in genetic association studies. Genet. Epidemiol. *34*, 188–193.
14. Ionita-Laza, I., Lange, C., and M Laird, N. (2009). Estimating the number of unseen variants in the human genome. Proc. Natl. Acad. Sci. USA *106*, 5008–5013.
15. Li, B., and Leal, S.M. (2009). Discovery of rare variants via sequencing: implications for the design of complex trait association studies. PLoS Genet. *5*, e1000481.
16. Nielsen, R., Hellmann, I., Hubisz, M., Bustamante, C., and Clark, A.G. (2007). Recent and ongoing selection in the human genome. Nat. Rev. Genet. *8*, 857–868.
17. Boyko, A.R., Williamson, S.H., Indap, A.R., Degenhardt, J.D., Hernandez, R.D., Lohmueller, K.E., Adams, M.D., Schmidt, S., Sninsky, J.J., Sunyaev, S.R., et al. (2008). Assessing the evolutionary impact of amino acid mutations in the human genome. PLoS Genet. *4*, e1000083.
18. Lin, D.Y., and Zeng, D. (2010). Meta-analysis of genome-wide association studies: no efficiency gain in using individual participant data. Genet. Epidemiol. *34*, 60–66.
19. Shen, Y., Wan, Z., Coarfa, C., Drabek, R., Chen, L., Ostrowski, E.A., Liu, Y., Weinstock, G.M., Wheeler, D.A., Gibbs, R.A., and Yu, F.A. (2010). A SNP discovery method to assess variant allele probability from next-generation resequencing data. Genome Res. *20*, 273–280.
20. Scheet, P., and Stephens, M. (2008). Linkage disequilibrium-based quality control for large-scale genetic studies. PLoS Genet. *4*, e1000147.

21. Leal, S.M. (2005). Detection of genotyping errors and pseudo-SNPs via deviations from Hardy-Weinberg equilibrium. Genet. Epidemiol. *29*, 204–214.
22. Douglas, J.A., Boehnke, M., and Lange, K. (2000). A multipoint method for detecting genotyping errors and mutations in sibling-pair linkage data. Am. J. Hum. Genet. *66*, 1287–1297.
23. Hernandez, R.D. (2008). A flexible forward simulator for populations subject to selection and demography. Bioinformatics *24*, 2786–2787.
24. Pritchard, J.K. (2001). Are rare variants responsible for susceptibility to complex diseases? Am. J. Hum. Genet. *69*, 124–137.
25. Eyre-Walker, A., and Keightley, P.D. (1999). High genomic deleterious mutation rates in hominids. Nature *397*, 344–347.
26. Kryukov, G.V., Shpunt, A., Stamatoyannopoulos, J.A., and Sunyaev, S.R. (2009). Power of deep, all-exon resequencing for discovery of human trait genes. Proc. Natl. Acad. Sci. USA *106*, 3871–3876.
27. Bodmer, W., and Bonilla, C. (2008). Common and rare variants in multifactorial susceptibility to common diseases. Nat. Genet. *40*, 695–701.
28. Browning, J.D., Szczepaniak, L.S., Dobbins, R., Nuremberg, P., Horton, J.D., Cohen, J.C., Grundy, S.M., and Hobbs, H.H. (2004). Prevalence of hepatic steatosis in an urban population in the United States: impact of ethnicity. Hepatology *40*, 1387–1395.
29. Victor, R.G., Haley, R.W., Willett, D.L., Peshock, R.M., Vaeth, P.C., Leonard, D., Basit, M., Cooper, R.S., Iannacchione, V.G., Visscher, W.A., et al; Dallas Heart Study Investigators. (2004). The Dallas Heart Study: a population-based probability sample for the multidisciplinary study of ethnic differences in cardiovascular health. Am. J. Cardiol. *93*, 1473–1480.
30. Kerem, B., Chiba-Falek, O., and Kerem, E. (1997). Cystic fibrosis in Jews: frequency and mutation distribution. Genet. Test. *1*, 35–39.
31. King, M.C., Rowell, S., and Love, S.M. (1993). Inherited breast and ovarian cancer. What are the risks? What are the choices? JAMA *269*, 1975–1980.
32. Bansal, V., Harismendy, O., Tewhey, R., Murray, S.S., Schork, N.J., Topol, E.J., and Frazer, K.A. (2010). Accurate detection and genotyping of SNPs utilizing population sequencing data. Genome Res. *20*, 537–545.
33. Ewing, B., and Green, P. (1998). Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res. *8*, 186–194.
34. Ewing, B., Hillier, L., Wendl, M.C., and Green, P. (1998). Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res. *8*, 175–185.
35. Daly, A.K., Donaldson, P.T., Bhatnagar, P., Shen, Y., Pe'er, I., Floratos, A., Daly, M.J., Goldstein, D.B., John, S., Nelson, M.R., et al; DILIGEN Study; International SAE Consortium. (2009). HLA-B*5701 genotype is a major determinant of drug-induced liver injury due to flucloxacillin. Nat. Genet. *41*, 816–819.
36. Mitsui, J., Mizuta, I., Toyoda, A., Ashida, R., Takahashi, Y., Goto, J., Fukuda, Y., Date, H., Iwata, A., Yamamoto, M., et al. (2009). Mutations for Gaucher disease confer high susceptibility to Parkinson disease. Arch. Neurol. *66*, 571–576.
37. Ng, S.B., Bigham, A.W., Buckingham, K.J., Hannibal, M.C., McMillin, M.J., Gildersleeve, H.I., Beck, A.E., Tabor, H.K., Cooper, G.M., Mefford, H.C., et al. (2010). Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. Nat. Genet. *42*, 790–793.
38. Gilissen, C., Arts, H.H., Hoischen, A., Spruijt, L., Mans, D.A., Arts, P., van Lier, B., Steehouwer, M., van Reeuwijk, J., Kant, S.G., et al. (2010). Exome sequencing identifies WDR35 variants involved in Sensenbrenner syndrome. Am. J. Hum. Genet. *87*, 418–423.
39. Bilgüvar, K., Oztürk, A.K., Louvi, A., Kwan, K.Y., Choi, M., Tatli, B., Yalnizoğlu, D., Tüysüz, B., Cağlayan, A.O., Gökben, S., et al. (2010). Whole-exome sequencing identifies recessive WDR62 mutations in severe brain malformations. Nature *467*, 207–210.