Two-Stage Design of Sequencing Studies for Testing Association with Rare Variants

Fan Yang Duncan C. Thomas

Department of Preventive Medicine, University of Southern California, Los Angeles, Calif., USA

Key Words

Next-generation sequencing • Multiple rare variants • Burden indices • Study design

Abstract

Multiple rare variants have been suggested as accounting for some of the associations with common single nucleotide polymorphisms identified in genome-wide association studies or possibly some of the as yet undiscovered heritability. We consider the power of various approaches to designing substudies aimed at using next-generation sequencing technologies to discover novel variants and to select some subsets that are possibly causal for genotyping in the original case-control study and testing for association using various weighted sum indices. We find that the selection of variants based on the statistical significance of the case-control difference in the subsample yields good power for testing rare variant indices in the main study, and that multivariate models including both the summary index of rare variants and the associated common single nucleotide polymorphisms can distinguish which is the causal factor. By simulation, we explore the effects of varying the size of the discovery subsample, choice of index, and true causal model.

Copyright © 2011 S. Karger AG, Basel

Introduction

The current crop of genome-wide association scans (GWAS) for various complex diseases has yielded a number of important insights. First, the relative risks for most replicated associations have been small, typically less than 1.5 [1]. Given the high degree of linkage disequilibrium between all known common variants and those on commercial genotyping chips, it has always been understood that most discovered single nucleotide polymorphism (SNP) associations are not likely to be directly causal, but merely surrogates for truly causal variants in the region, whose effects could be stronger than the induced associations discovered. Second, even in the aggregate, the overall effect of all discovered associations explains only a small portion of the estimated heritability for most complex traits. Various hypotheses about the remaining dark matter have been proposed [2, 3], one of which is multiple rare variants that are not well tagged by HapMap SNPs [4–7]. Because of their rarity and the penalty being even larger for multiple comparisons than for common SNPs, detecting the effects of any single such variant is nearly impossible, but the aggregate effect of multiple variants within a region may be detectable [8]. Resequencing of an associated region to discover novel

variants that could be subject to such an analysis is thus a necessary first step in this process [9]. Because of the high cost of sequencing, even with NextGen technologies [10, 11], it is generally feasible to do this only on a targeted subsample. In this paper, we consider the design of such studies with the aim of selecting some subsets of variants for direct genotyping in the original GWAS samples and testing regional association with rare variants.

Various approaches have been suggested for testing association with multiple rare variants. Morgenthaler and Thilly [12] described a cohort allelic sums test (CAST), which contrasts the frequency of individuals carrying any mutant alleles in sequenced exons between cases and controls. Li and Leal [13] proposed the combined multivariate and collapsing (CMC) method that collapses rare variants using an indicator variable for the presence of any rare alleles. This set of rare variants can be tested simultaneously in a multivariate framework with other groups of variants collapsed in the same fashion, where the groups are formed according to predefined criteria such as allele frequencies. Such an approach has been applied, for example, to variants of unknown significance in candidate genes for breast cancer [14]. Madsen and Browning [15] suggested refining this approach by weighting alleles inversely by the population frequency. Li and Leal [16] considered the design of resequencing studies and pointed out the serious inflation of type I error rates than can arise if the same sample is used to select variants and test association with the selected variants. This problem can be simply avoided by using disjoint samples for variant discovery and testing, but the optimal design of such studies in terms of the allocation of sample sizes and sampling schemes between discovery and testing phases, depth of sequencing, criteria for selecting variants for genotyping in the main study, and analysis methods remains unresolved.

We considered study designs in which random samples of cases and controls would be fully sequenced in a region carrying one or more causal variants, typically regions identified by an association with one or more common SNPs. Based on univariate analyses of the association of every polymorphic site with disease status in these samples, a subset of potentially causal variants would then be genotyped in a much larger, independent case-control study and their association with risk tested using various types of summary indices. We considered several criteria for selecting discovered variants for genotyping and several forms of summary indices, and we studied the effects of varying sample sizes for discovery and testing with each as well as the dependence of the optimal design and

analysis on the true distribution of allele frequencies of causal alleles and their relative risks. As a benchmark for comparison with feasible indices, we used similarly constructed summary indices that included all truly causal variants and excluded all truly non-causal variants. By simulation, we demonstrate that our tests of regional association have the nominal type I error rate and compared methods on the basis of their power to detect regional association. Finally, for each instance where regional association with the summary index was deemed significant, we assessed the false discovery and false non-discovery rates for the specific variants included in the index.

Methods

Simulations

Populations and Samples

A haplotype pool of size 20,000 for a 100-kb region was simulated using coalescent simulator *ms* [17], in which the scaled mutation rate $\theta = 4N_0\mu$ was set to 100 and the scaled crossover rate $\rho = 4N_0r$ was set to 50 to mimic a human population. Two independent case-control samples were drawn from the haplotype pool: a discovery panel for variant detection via sequencing; and a larger main study sample in which a regional association test was conducted. To generate a case-control sample, genotypes of each individual were formed by sampling a random pair of haplotypes with replacement from the pool, and disease status was determined by a specified disease model. For 1,000 Genomes Project controls, 1,000 individuals were randomly sampled from the same haplotype pool where the cases were sampled, without determining their disease status.

Disease Model

A logistic model with an additive coding of genotypes was used to determine binary disease status

$$\text{logit}[\Pr(Y_i = 1|G_i)] = \beta_0 + \sum_{j=1}^n \beta_j G_{ij} \quad (1)$$

where G_{ij} is the number of mutations in individual i at variant j . The probability of variant j being causal was calculated as

$$\text{logit}[\Pr(\beta_j \neq 0)] = \pi_0 + \pi_1 \ln(q_j/\bar{q}) \quad (2)$$

where q_j is the population minor allele frequency (MAF) for variant j , $\bar{q}$ is a scaling constant set to 0.01, and the parameters π_0 and π_1 determine the overall proportion of causal variants and their allele frequency spectrum relative to all variants. When β_j is set to be non-zero in equ. 2, its value, interpreted as the genotypic relative risk (RR) on a log scale, was simulated as

$$\beta_j = \varepsilon_j \exp(\alpha_0 + \alpha_1 \ln(q_j/\bar{q})) \quad (3)$$

where ε was sampled from either a Normal(0,1) distribution to account for the existence of both deleterious and protective minor alleles – denoted the mixed causal variants (MCV) model – or a

truncated Normal(0,1) distribution that has only positive values – the deleterious causal variants (DCV) model. The parameters π_1 and α_1 that specify the dependence on MAF were both set to non-zero values in the MAF-dependent model and both set to 0 in the MAF-independent model. For simplicity, we describe the distribution of effect sizes by the 95th percentile of the genotypic RR for a variant with MAF = 0.01 (primarily controlled by α_0), denoted, e.g., as $RR_{95} = 2$ or 3. The corresponding α_1 was set in such a way that the 95th percentile of the genotypic RR for a variant with the lowest possible MAF (1/20,000) is 20 or 30. We present results for proportions of all rare variants being causal ranging from 1 to 5% in the belief that most non-coding variants are likely to be neutral with respect to the trait of interest.

Analyses

Variant Selection

An indicator variable D was assigned to each variant to account for the direction of its apparent effect in the discovery panel: $D = +1$ for an apparently deleterious variant and $D = -1$ for an apparently protective variant. Then a subset of variants identified in the discovery panel was selected by various criteria, among which two types of thresholds were particularly focused upon: (1) p value thresholds, where a marginal two-sided p value for each variant was calculated and those with p values smaller than various thresholds were included; (2) frequency thresholds, where MAF for each variant was calculated using the entire discovery panel and those with MAF lower than various thresholds were included. All indices include variants with $D = +1$, some also include those with $D = -1$. For comparison purpose, we also considered indices that include all discovered variants, or variants found only in cases.

Test Indices

The summary genetic indices for individual i were calculated as

$$I_j = \sum_{j \in S} D_j w_j G_{ij} \quad (4)$$

where S is the selected subset of variants, and w_j is a weight, either the inverse of the estimated binomial standard deviation $1/\sqrt{n_j q_j (1 - q_j)}$ in the Madsen and Browning weighted sum method, or 1 for an unweighted index. G_{ij} is the genotype for individual i at variant j using additive coding, obtained by genotyping variant set S in an independent main study sample. Regional association was then assessed by Armitage's trend test [18] for each variant selection criterion and each w_j , respectively. This procedure was repeated for 1,000 simulated datasets for a range of parameter sets, and power was estimated as the proportion of significant replicates.

Common Variant Signals

Since many association studies involve sequencing to investigate candidate regions around GWAS hits, we simulated associations with common variants by scanning all variants with MAF ≥ 0.05 using the main study sample, retaining replicates with at least one hit with a p value $< 0.05/N$, where N is the number of common variants tested. Power calculated from these replicates is denoted 'conditional power'. We also computed power over all replicates, with or without common variant signals, simply denoted 'power'. The purpose is to extend the two-stage analytical

scheme into the context of exome sequencing studies, where sequenced individuals in each stratum (cases or controls for example) are randomly separated into two disjoint samples, one for variant selection, and the other for testing.

Adjustment for Common Variants

A question for the multiple rare variants hypothesis is whether the SNP signals detected by a GWAS can be explained by synthetic associations with multiple rare variants [19]. The reverse question is if a summary index is used to test a group of rare variants and shows significant association, can this signal be actually due to causal effects of common SNPs? To answer these questions, we performed simulations under models where either a single common SNP or multiple rare variants were the causal factors. All rare variants (MAF < 0.05 in controls) discovered in 200 cases and 200 controls were collapsed into a weighted sum index and tested for association in 1,000 cases and 1,000 controls. The test was adjusted for all SNPs that were significant in single marker tests with Bonferroni correction in a multiple logistic regression. The power of such a test is denoted 'adjusted power'.

Permutation Test

We also considered a one-stage design, where variant selection and tests for association were performed using the same sample. For this design, the sampling distribution of the t statistic testing a summary index was obtained through permutation. For each random permutation of case-control labels, we selected a subset S of variants for the particular criterion, calculated the index for each individual, and computed the t statistic for the permutation. For N permutations, the p value was assessed as $(n + 1)/(N + 1)$, where n is the number of permutations that generated more extreme t statistics. Ties were resolved randomly. $N = 1,000$ was used for each test in our simulation study.

Results

Description of the Simulated Data

The majority of the simulated alleles were rare, specifically 28.1% were common (MAF > 0.05), 15.7% were uncommon ($0.01 < \text{MAF} \leq 0.05$), and 56.2% were rare (MAF ≤ 0.01) across 10,000 haplotype pools. The corresponding numbers for the combined case-control samples were 33.6% (MAF > 0.05), 18.8% ($0.01 < \text{MAF} \leq 0.05$), and 47.6% (MAF ≤ 0.01). The distributions of all discovered variants and of discovered causal variants were almost identical for MAF-dependent and MAF-independent models, whether or not at least one common variant signal was required. For MAF-dependent models, however, the discovered causal variants were dominated by rare variants. For example, discovered causal variants were distributed as 8.7% (MAF > 0.05), 9.7% ($0.01 < \text{MAF} \leq 0.05$), and 81.6% (MAF ≤ 0.01) for 10,000 replicates. When at least one common variant signal is required, as arises in a follow-up of a GWAS hit, these numbers were

Table 1. Power of regional association using unweighted and Madsen and Browning weighted sum indices

Index	MAF-dependent model		MAF-independent model	
	DCV	MCV	DCV	MCV
Unweighted	0.263	0.167	0.778	0.417
Weighted	0.320	0.176	0.844	0.452

Power for 1,000 cases and 1,000 controls. The two sum indices include all variants discovered in an independent sample with 200 cases and 200 controls. 1,000 independent replicates with 2.5% causal variants and $RR_{95} = 2.0$ were simulated for each model.

DCV = Deleterious causal variant model; MCV = mixed (deleterious and protective) causal variant model.

Table 2. Power of regional association using the Madsen and Browning weighted sum index collapsing rare variants under common disease-common variants assumption and common disease-rare variants assumption

Causal variants	Model	RR_{95}	Unconditional power	Unadjusted conditional power	Adjusted conditional power
SNP	MAF-dependent	2	0.056	0.120	0.046
		3	0.093	0.156	0.055
	MAF-independent	2	0.105	0.165	0.052
		3	0.173	0.209	0.051
MRV	MAF-dependent	2	0.147	0.295	0.184
		3	0.310	0.489	0.480
	MAF-independent	2	0.286	0.564	0.275
		3	0.522	0.662	0.415

200 cases and 200 controls were used for discovery. An independent sample with 1,000 cases and 1,000 controls was used for testing.

SNP = Single nucleotide polymorphism; MRV = multiple rare variant.

11.9% ($MAF > 0.05$), 10.0% ($0.01 < MAF \leq 0.05$), and 78.1% ($MAF \leq 0.01$).

The statistical power to detect regional association using unweighted or weighted sum index collapsing all discovered variants was highly dependent on whether the probability of a variant being causal and their distribution of effect sizes depend upon the MAF and whether there are protective variants (table 1). The Madsen and Browning weighting scheme generally improves power, but this advantage is not as pronounced in MCV models as in DCV models using this method because it does not handle deleterious and protective variants differently.

Test of Summary Rare Variant Index with Adjustment for Common SNPs

We first demonstrated that our regional association tests have the correct type I error rate using a ‘global-null’ model where no causal variants were simulated

(data not shown). Then we simulated a ‘pseudo-null’ model for testing the aggregate of rare variants where a single common SNP (population $MAF \geq 0.05$) is designated as the only causal variant in the region and examined the type I error rate for testing association with rare variants. Table 2 shows that testing a weighted index collapsing all discovered rare variants in an independent sample has an inflated type I error under this pseudo-null model. The inflation becomes more serious if the region is required to have common variant signals (e.g. for a follow-up of GWAS hits). Thus, false positive associations with aggregate indices of rare variants can be induced by common SNPs in linkage disequilibrium with them. Adjusting for the common SNPs that show region-wide significance can control this inflation, as illustrated in the last column of table 2, this remains true even for the pseudo-null model with more common causal SNPs and larger marginal effect sizes (data not

Fig. 1. Power of regional association using various thresholds and causal models: **a, b** MAF-dependent DCV models, **c, d** MAF-dependent MCV models. **a, c** p value thresholds, **b, d** frequency thresholds. All analyses were carried out with 200 cases and 200 controls in the discovery panel and with 2.5% (C0.025) or 5% (C0.05) causal variants and $RR_{95} = 2.0$ (E2) or 3.0 (E3), when frequency = 0.01.

shown). Since we made no assumptions about the frequency of the underlying causal variants, both common and rare variants could be causal in the following simulations, and therefore summary indices could include both common and rare variants, depending on variant selection criteria. Tests of these indices have correct type I error rates (at $\alpha = 0.05$) under the true null model where no causal variants are simulated.

Variant Selection Using a Discovery Panel

The motivation for including a subset of variants in a summary index is two-fold: (1) to increase statistical power by improving the signal-to-noise ratio and (2) to minimize the genotyping cost for the main study by ex-

cluding apparently irrelevant variants. Under models in which both rare and common variants can be disease causing, we considered various p value thresholds and frequency thresholds, with the restriction that the frequency in cases should be higher than in controls. For indices based on frequency thresholds, power consistently increased with higher frequency thresholds, but eventually the gain became marginal, as rare variants were simulated to have higher probability of being causal in MAF-dependent models (fig. 1b, d). Thus, if a fixed frequency threshold (e.g. 5%) is used in studies that focus on rare variants, the statistical power of the study is suboptimal. p value thresholds generally outperformed frequency thresholds (fig. 1a, c), while using fewer variants.

Table 3. The influence of discovery sample size on variant discovery rates and statistical power using various indices

RR ₉₅	Percentage causal	Discovery sample size	Dis%	Causal%	Tp0.1	Tp0.2	Tq0.05	Tq0.5
2	2.5	100	62.8	31.8	0.209	0.216	0.164	0.223
		200	69.4	41.3	0.288	0.285	0.190	0.282
		500	77.8	56.6	0.392	0.387	0.264	0.357
	5	100	62.8	32.3	0.406	0.426	0.340	0.443
		200	69.3	42.1	0.511	0.547	0.424	0.546
		500	78.0	57.1	0.669	0.660	0.519	0.621
3	2.5	100	62.8	32.7	0.487	0.500	0.373	0.489
		200	69.4	42.1	0.521	0.532	0.375	0.498
		500	77.8	57.4	0.700	0.654	0.474	0.617
	5	100	62.9	33.5	0.712	0.727	0.572	0.721
		200	69.5	42.9	0.853	0.851	0.724	0.836
		500	78.0	58.0	0.925	0.917	0.805	0.900

Dis% = Proportion of all variants discovered; Causal% = proportion of causal variants discovered; Tpx = p value threshold x; Tqx = frequency threshold x.

Table 4. Cost-equivalent designs using study or 1,000 Genomes Project controls for discovery

Controls	Causal model	RR ₉₅	Power Tp0.1	Variants used*	Power Tq0.5	Variants used*
Study controls	MAF-dependent	2	0.259	78,961	0.271	332,477
		3	0.586	90,186	0.536	335,403
	MAF-independent	2	0.950	138,535	0.955	342,716
		3	0.997	165,618	0.993	345,581
1,000 Genomes Project controls	MAF-dependent	2	0.363	118,146	0.333	386,487
		3	0.720	147,114	0.622	412,410
	MAF-independent	2	0.991	195,931	0.986	402,164
		3	0.992	205,306	0.992	386,908

Tp0.1 = p value threshold 0.1; Tq0.5 = frequency threshold 0.5. * Total number of variants used across 1,000 replicates within a 100-kb region.

Since power as a function of p value cut-offs levels off rapidly when the p value is >0.1, one can achieve good power using less than 20% of the discovered variants by applying a fixed 0.1 p value threshold. A possible disadvantage of the p value threshold is that it does not necessarily prioritize rare variants, while common variants are likely to have already been discovered and tested.

Under models based on the multiple rare variants (MRV) hypothesis, where no common variants (MAF > 0.05) are causal, we observed that a frequency threshold

of 5% achieved the highest statistical power (data not shown), as expected; the power then decreased as a greater frequency threshold beyond 5% was applied. The power curve for p value thresholds plateaued beyond 0.01, a similar pattern as in the general models above. Although p value thresholds achieve slightly lower power than a fixed frequency threshold of 5%, their performance is at least comparable to other frequency thresholds and is consistent across different scenarios.

Table 5. Improvement in power by including potentially protective minor alleles in models with simulated protective variants

Causal model	<i>D</i>	Tp0.1	Tp0.2	Tq0.05	Tq0.5
<i>MCV</i>					
MAF-dependent	+1 only	0.384	0.380	0.243	0.366
	+1 and -1	0.452	0.425	0.320	0.407
MAF-independent	+1 only	0.970	0.972	0.847	0.968
	+1 and -1	0.985	0.981	0.951	0.977
<i>DCV</i>					
MAF-dependent	+1 only	0.521	0.532	0.375	0.498
	+1 and -1	0.499	0.495	0.393	0.477
MAF-independent	+1 only	0.991	0.991	0.940	0.992
	+1 and -1	0.990	0.991	0.968	0.991

200 cases and 200 controls were used in the discovery panel. A MAF-dependent model was used with $RR_{95} = 3$, and 2.5% of the variants being causal.

D = Indicator for whether direction of association is accounted for in the model; Tpx = p value threshold x; Tqx = frequency threshold x.

Sample Sizes for Discovery and Testing Samples

We considered three aspects in selecting an appropriate sample size for the discovery panel (table 3). The first is the proportion of all variants (causal or not) that are discovered, approximately 63, 69, 78% for sample sizes 100, 200, 500. These numbers were fairly constant across simulation model parameter settings. The second aspect is the proportion of causal variants discovered, which increased slightly with higher effect sizes and more causal variants. The third aspect is the power for detecting regional association. For a main study sample size of 1,000 cases and 1,000 controls, we found that a small discovery panel is useful only when a considerable proportion of variants are causal and they have large effect sizes. For example, at least 500 cases and 500 controls are needed for discovery to achieve 70% power, when 2.5% of all variants are causal and RR_{95} is 2.0–3.0.

Cost-Equivalent Design Comparisons

We compared two cost-equivalent designs for obtaining control data. The first uses study controls, the second uses 1,000 Genomes Project data, assuming they represent the source population of cases; here, only 2*N* cases are sequenced for variant discovery instead of *N* cases and *N* study controls. The results show that using 400 cases and 1,000 Genomes Project controls for discovery has higher power than using 200 cases and 200 controls for discovery (table 4). However, the former requires a few

more variants to be genotyped in the association testing stage (but typically less than twice as many), increasing genotyping cost somewhat. The relative cost-efficiency of the two designs thus would depend upon the relative costs of sequencing and genotyping.

When instead of cost there is a constraint on the number of case samples available, one could vary the allocation into two subsamples, one for discovery and one for testing. Under MAF-dependent models, we found that allocating 40% (DCV) or 50% (MCV) of the sample for variant discovery and selection while reserving the remaining for association testing generally achieved the best power (fig. 2). Results under MAF-independent models exhibited a distinct pattern, showing that power levels off as long as the discovery sample size is within a reasonable range (100 to 900 out of 1,000). An exception is the method using an index that includes *all* discovered variants, for which smaller discovery samples are preferable across models. Again, the most cost-efficient choice would depend upon the relative costs of sequencing and genotyping.

Including Protective Minor Alleles

Sequencing is generally directed at discovery of rare deleterious variants, as associations with rare protective variants are more difficult to detect. However, some variants may be truly protective and sequencing studies will inevitably identify some as apparently protective. We found that the existence of truly protective alleles reduces the statistical power for detecting an association when using a simple summary index that counts only apparently deleterious alleles (table 5). For example, using an index that only includes apparently deleterious variants, we had 52.1% power under a simulated DCV model but only 38.4% under the corresponding MCV model (where 50% of causal variants are protective). Taking potentially protective alleles into account by including the *D* factor in the index consistently improved power in MCV models across all numbers of causal variants and effect sizes. For example, this method increases the probability of detecting the association from 38.4 to 45.2% in the previous situation. However, almost twice as many variants will need to be genotyped, although one can certainly employ different thresholds for potentially deleterious and protective alleles in order to reduce cost. When there truly are no protective alleles, using this method will not improve power and could even compromise it; for example, adding protective alleles to the index in the DCV model described above reduces power from 52.1 to 49.9%.

Fig. 2. Power of regional association using 5 representative weighted indices: p value threshold 0.1 (Tp0.1), p value threshold 0.2 (Tp0.2), frequency threshold 0.05 (Tq0.05), frequency threshold 0.5 (Tq0.5), and all discovered variants, with various ratios of discovery to association-testing sample sizes. The total sam-

ple size is fixed at 1,000 cases and 1,000 controls. These analyses were carried out for a MAF-dependent DCV model (a), a MAF-dependent MCV model (b), a MAF-independent DCV model (c), and a MAF-independent MCV model (d) with 2.5% causal variants and $RR_{95} = 3.0$.

Efficiency of Indices for Distinguishing Causal from Non-Causal Variants

We tabulated the false discovery rate (FDR) and false non-discovery rate (FNDR) for various thresholds and causal models (table 6), where the FDR was estimated by the proportion of all included variants that are non-causal, and the FNDR was estimated by the proportion of all excluded variants that are causal. Although the majority of included variants are non-causal (high FDR), nevertheless we have shown above that these indices yield good power to detect regional association. Also tabulated are sensitivity and specificity, estimated by the proportion of causal variants that are included and the proportion of

non-causal variants that are not included, respectively. The relatively low sensitivity and low causal:non-causal ratio (1:39 for table 6) jointly explain the universally high FDR. Both p value and frequency thresholds achieve lower FDR, lower FNDR, and higher specificity than indices that include all discovered variants. A frequency threshold at 0.05 has overall better performance in terms of these four measures than other criteria in MAF-dependent models but not in MAF-independent models. This reflects the fact that more rare variants were simulated to be causal in MAF-dependent models, suggesting that using a frequency threshold enriches causal variants only if causality is correlated with MAF.

Table 6. FDR and FNDR using various variant selection criteria

Model	RR ₉₅	Measure	Tp0.1	Tp0.2	Tq0.05	Tq0.5	All
MAF-dependent	2	FDR	0.975	0.972	0.970	0.980	0.986
		FNDR	0.024	0.023	0.023	0.026	0.046
		Sensitivity	0.080	0.207	0.221	0.262	0.413
		Specificity	0.923	0.823	0.820	0.680	0.299
	3	FDR	0.969	0.969	0.968	0.978	0.985
		FNDR	0.023	0.022	0.022	0.025	0.045
		Sensitivity	0.110	0.243	0.245	0.290	0.421
		Specificity	0.914	0.816	0.820	0.680	0.300
MAF-independent	2	FDR	0.954	0.961	0.968	0.965	0.974
		FNDR	0.022	0.021	0.024	0.021	0.025
		Sensitivity	0.242	0.347	0.222	0.446	0.698
		Specificity	0.870	0.779	0.825	0.676	0.310
	3	FDR	0.954	0.960	0.966	0.964	0.974
		FNDR	0.021	0.020	0.023	0.020	0.024
		Sensitivity	0.287	0.384	0.241	0.465	0.700
		Specificity	0.847	0.763	0.825	0.678	0.310

The discovery panel consists of 200 cases and 200 controls.

FDR = Proportion of all variants included in the index that are truly non-causal; FNDR1 = proportion of all variants not included in the index that are truly causal; Sensitivity = proportion of truly causal variants that are included; Specificity = proportion of truly non-causal variants that are not included; Tpx = p value threshold x; Tqx = frequency threshold x.

Table 7. Power for testing association using one-stage and two-stage designs

Model	RR ₉₅	One-stage Tp0.01	Two-stage Tp0.1
MAF-dependent	2	0.255	0.263
	3	0.576	0.546
MAF-independent	2	0.929	0.965
	3	0.993	0.994

The highest power of regional association under DCV models, achieved using p value threshold 0.01 in the one-stage design and 0.1 in the two-stage design, is shown for comparison.

The one-stage analysis is conducted using 1,000 cases and 1,000 controls, and permutation test is used to assess significance. For the two-stage analysis, 400 cases and 400 controls are used for variant selection, while an independent sample of 600 cases and 600 controls are used for testing association.

Tpx = p value threshold x.

Two-Stage versus One-Stage Designs

The two-stage design achieves at least comparable power to a one-stage design with a permutation test (table 7). Since various p value thresholds and frequency thresholds behave differently in the two designs, we show only the thresholds that achieve the highest power for either design along with the criteria that achieve that optimum. For example, with 400 cases and 400 controls for

discovery and 600 cases and 600 controls for testing, the two-stage design achieves the best power at a p value threshold of 0.1, whereas a one-stage design with 1,000 cases and 1,000 controls is optimized at a p value threshold of 0.01. In this comparison, the one-stage design outperforms the two-stage design only in MAF-dependent models with high marginal effect sizes (RR₉₅ = 3.0). Therefore, the two-stage design proves to be not only cost-

efficient in terms of reducing sequencing costs, but is also computationally more efficient when a permutation test is needed to control type I errors. Since an analytical framework that selects variants in part of the sample and tests constructed indices in the remaining part of the sample applies even when both subsamples are sequenced, this method has the potential to be a general analytical framework without requiring a two-stage design.

Discussion

Our simulation studies have demonstrated that these simple two-stage designs provide unbiased tests of regional association with multiple rare variants and a reasonable balance of true and false discoveries for specific variants identified in this way that would be candidates for further replication or molecular characterization studies. We also found that sample sizes of 100–500 subjects were sufficient to yield good power for discovering a substantial portion of causal variants for inclusion in these summary indices, depending upon the overall risk attributable to rare variants. If rare variants are truly more likely than common variants to be causally associated or to have larger risks, then weighted indices like that proposed by Madsen and Browning [15] provide more powerful tests than unweighted tests, while there is little loss of power if the probability of being causal or the relative risks are unrelated to allele frequency. For rare diseases, there appears to be little to be gained by random sampling of controls if comparable general population data (e.g. the 1,000 Genomes Project [20]) are available and no strong environmental risk factor(s) is suspected, but further work is needed to assess whether this remains true if cases derive from populations that are not well represented in the shared control data or if the disease is common. In the case that a strong environmental factor(s) is present for the trait, using matched controls rather than general population controls is expected to substantially improve power. From a practical perspective, the differences that arise from different sequencing procedures and data preprocessing steps between case data and shared control data need to be carefully dealt with.

Although some have argued that a high proportion of rare coding variants could be deleterious [5, 21] and the simulations of Li and Leal [13, 16] have assumed proportions of causal variants starting at 20%, we followed instead the lead of Dickson et al. [19], who considered models with only 1–9 causal variants in a 10-Mb region.

Permutation testing is one solution to the type I error inflation problem that arises when phenotype information is involved in variant discovery and/or summary index construction. One merit of our two-stage design over a one-stage design is that it overcomes the need for permutation test while retaining correct type I error, which significantly reduces the computational burden. We found our two-stage design achieves comparable power to currently available statistical tools, although both designs have yet to be optimized with more sophisticated variant selection methods and sampling schemes. It also has been shown that the naïve permutation test does not guarantee an unbiased test when there is population stratification. Further investigation of methods for controlling for population stratification issue in a one-stage framework would be helpful.

A promising extension of the two-stage method is a joint analysis of both the discovery panel and the genotyped sample, which increases sample sizes for selected variants and leverages available information in both phases. Skol et al. [22] have shown that the joint analysis method achieves higher power than treating the second-phase data as replication in a framework where a subset of SNPs selected from the first-phase genotyping study is genotyped in an independent second-phase sample.

In on-going work, we are exploring the use of two-phase case-control designs [23, 24] that would allow reuse of a subsample of cases and controls from an existing study for both sequencing and testing using a joint analysis of the full sequence data on the subset and only SNP data on the parent study. More importantly, such designs permit subsampling jointly on disease status and the associated SNPs in a given region from the original GWAS analysis by taking the biased sampling into account in the analysis [25]. Thus, rather than simply using random samples of cases and controls, the selection of samples for sequencing could be targeted to the individuals most likely to be informative for each associated region separately. If gene-environment interactions are considered, one might also stratify on environmental risk factors. Such designs could use either imputation of the untyped variants in the main study using full sequence data from the subsample or from the 1,000 Genomes Project, or could re-genotype selected variants in the main study as considered here; the relative cost-efficiency of these alternatives is still being explored. Appropriate analysis for two-phase case-control samples can overcome the biases from using the same data twice and from biased sampling, but not the bias noted by Li and Leal [16] from using the same data both

for constructing a summary index and for testing association with it.

One might also wish to stratify on family history. Additional information to support a judgment of causality could then be obtained by studying the co-segregation of the disease within families with specific rare variants. Family-based designs would also allow selection of distantly related case pairs for sequencing in families showing evidence of linkage to better identify likely causal variants.

Alternative methods for aggregating rare variants could be integrated into this framework. The variable threshold method described by Price et al. [26] and the step-up approach described by Hoffman et al. [27] could be used for selecting an optimal subset of variants from a discovery panel. The coding scheme of protective variants proposed by Han and Pan [28] results in an equivalent index to the one using a $(-1)^D$ term that we introduced here.

More sophisticated analysis methods based on Bayesian hierarchical models are also possible. Rather than weighting all variants equally or using weights that depend on a priori fixed functions of allele frequency or other attributes (e.g. coding or not), weights could be estimated empirically by assuming either a simple exchangeable model, a mixture of null and causal distributions, or regression models on prior covariates [29, 30]. For example, one could include a latent variable taking the values +1, 0, or -1 for deleterious, null, or protective associations, using MCMC methods to average over the probability distributions of these weights [31].

The analysis of raw sequence data itself poses challenging difficulties in alignment and allele calling. Depending upon the depth of sequencing used, there can be substantial uncertainty in allele calls that should also be

taken into account in any analysis. We have not attempted to simulate raw sequence fragment data here, basing our simulations instead on the true haplotypes generated by the *ms* program. For variant discovery purposes, Ionita-Laza and Laird [32] showed that lower coverage with more samples is preferable to higher coverage with fewer samples for a fixed study cost. Trade-offs between sample size, depth of sequencing, and region sizes are amongst the issues still under investigation in our ongoing simulations.

DNA pooling methods attempt to capture more sequence information with the same cost. Wang et al. [33] developed a test for single variant in pooled data, accounting for the influence of high sequencing error rates. Kim et al. [34] proposed a likelihood ratio statistic and observed in simulation that prioritizing pool size rather than depth of coverage achieved higher power.

In summary, appropriate design and analysis of resequencing studies can provide a valid and cost-efficient way of testing for the existence of rare variants that could account for associations with SNPs discovered in a GWAS. Once such potentially causal variants have been detected and their aggregate association with disease tested, much remains to be done to replicate the associations of the specific variants in other samples and characterize their effects in functional studies. Meanwhile, further research on the optimal design and analysis of such studies would be useful.

Acknowledgments

This work was supported by NIH grants ES15090 and GM 069890. The authors are grateful to Drs. Paul Marjoram and David Conti for many helpful discussions.

References

- Hindorf LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA: Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci* 2009;106:9362–9367.
- Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorf LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, Cho JH, Guttacher AE, Kong A, Kruglyak L, Mardis E, Rotimi CN, Slatkin M, Valle D, Whittemore AS, Boehnke M, Clark AG, Eichler EE, Gibson G, Haines JL, Mackay TF, McCarroll SA, Visscher PM: Finding the missing heritability of complex diseases. *Nature* 2009;461:747–753.
- Maier B: Personal genomes: the case of the missing heritability. *Nature* 2008;456:18–21.
- Bodmer W, Bonilla C: Common and rare variants in multifactorial susceptibility to common diseases. *Nat Genet* 2008;40:695–701.
- Kryukov GV, Pennacchio LA, Sunyaev SR: Most rare missense alleles are deleterious in humans: implications for complex disease and association studies. *Am J Hum Genet* 2007;80:727–739.
- Iyengar SK, Elston RC: The genetic basis of complex traits: rare variants or ‘common gene, common disease’? *Methods Mol Biol* 2007;376:71–84.
- Fearnhead NS, Wilding JL, Winney B, Tonks S, Bartlett S, Bicknell DC, Tomlinson IP, Mortensen NJ, Bodmer WF: Multiple rare variants in different genes account for multifactorial inherited susceptibility to colorectal adenomas. *Proc Natl Acad Sci* 2004;101:15992–15997.
- Hoggart CJ, Clark TG, De Iorio M, Whittaker JC, Balding DJ: Genome-wide significance for dense SNP and resequencing data. *Genet Epidemiol* 2008;32:179–185.
- Kryukov GV, Shpunt A, Stamatoyannopoulos JA, Sunyaev SR: Power of deep, all-exon resequencing for discovery of human trait genes. *Proc Natl Acad Sci* 2009;106:3871–3876.

- 10 Droege M, Hill B: The Genome Sequencer FLX System – longer reads, more applications, straight forward bioinformatics and more complete data sets. *J Biotechnol* 2008; 136:3–10.
- 11 Bentley DR: Whole-genome re-sequencing. *Curr Opin Genet Dev* 2006;16:545–552.
- 12 Morgenthaler S, Thilly WG: A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (CAST). *Mutat Res* 2007;615:28–56.
- 13 Li B, Leal SM: Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am J Hum Genet* 2008;83:311–321.
- 14 Johnson N, Fletcher O, Palles C, Rudd M, Webb E, Sellick G, Dos Santos Silva I, McCormack V, Gibson L, Fraser A, Leonard A, Gilham C, Tavtigian SV, Ashworth A, Houlston R, Peto J: Counting potentially functional variants in *brca1*, *brca2* and *atm* predicts breast cancer susceptibility. *Hum Mol Genet* 2007;16:1051–1057.
- 15 Madsen BE, Browning SR: A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet* 2009; 5:e1000384.
- 16 Li B, Leal SM: Discovery of rare variants via sequencing: implications for the design of complex trait association studies. *PLoS Genet* 2009;5:e1000481.
- 17 Hudson RR: Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* 2002;18:337–338.
- 18 Armitage P: *Statistical Methods in Medical Research*. Oxford, Blackwell Scientific Publications, 1971.
- 19 Dickson SP, Wang K, Krantz I, Hakonarson H, Goldstein DB: Rare variants create synthetic genome-wide associations. *PLoS Biol* 2010;8:e1000294.
- 20 Via M, Gignoux C, Burchard EG: The 1000 Genomes Project: new opportunities for research and social challenges. *Genome Med* 2010;2:3.
- 21 Pritchard JK: Are rare variants responsible for susceptibility to complex diseases? *Am J Hum Genet* 2001;69:124–137.
- 22 Skol AD, Scott LJ, Abecasis GR, Boehnke M: Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. *Nat Genet* 2006;38: 209–213.
- 23 Breslow NE, Lumley T, Ballantyne CM, Chambless LE, Kulich M: Using the whole cohort in the analysis of case-cohort data. *Am J Epidemiol* 2009;169:1398–1405.
- 24 Breslow NE, Chatterjee N: Design and analysis of two-phase studies with binary outcome applied to Wilms tumour prognosis. *Appl Statist* 1999;48:457–468.
- 25 Thomas DC, Casey G, Conti DV, Haile RW, Lewinger JP, Stram DO: Methodological issues in multistage genome-wide association studies. *Statist Sci* 2009;24:414–429.
- 26 Price AL, Kryukov GV, de Bakker PI, Purcell SM, Staples J, Wei LJ, Sunyaev SR: Pooled association tests for rare variants in exon-resequencing studies. *Am J Hum Genet* 2010;86: 832–838.
- 27 Hoffmann T, Marini N, Witte J: Comprehensive approach to analyzing rare variants. *PLoS One* 2010;5:e13584.
- 28 Han F, Pan W: A data-adaptive sum test for disease association with multiple common or rare variants. *Hum Hered* 2010;70:42–54.
- 29 Lewinger JP, Conti DV, Baurley JW, Triche TJ, Thomas DC: Hierarchical Bayes prioritization of marker associations from a genome-wide association scan for further investigation. *Genet Epidemiol* 2007;31:871–882.
- 30 Chen GK, Witte JS: Enriching the analysis of genomewide association studies with hierarchical modeling. *Am J Hum Genet* 2007;81: 397–404.
- 31 Conti D, Gauderman W: SNPs, haplotypes, and model selection in a candidate gene region: the simple analysis of multilocus data. *Genet Epidemiol* 2004;27:429–441.
- 32 Ionita-Laza I, Laird NM: On the optimal design of genetic variant discovery studies. *Stat Appl Genet and Mol Biol* 2010;9:Article 33.
- 33 Wang T, Lin CY, Rohan TE, Ye K: Resequencing of pooled DNA for detecting disease associations with rare variants. *Genet Epidemiol* 2010;34:492–501.
- 34 Kim SY, Li Y, Guo Y, Li R, Holmkvist J, Hansen T, Pedersen O, Wang J, Nielsen R: Design of association studies with pooled or unpooled next-generation sequencing data. *Genet Epidemiol* 2010;34:479–491.