Hindawi Publishing Corporation
Journal of Probability and Statistics
Volume 2012, Article ID 524724, 19 pages
doi:10.1155/2012/524724

Research Article

Design and Statistical Analysis of Pooled Next Generation Sequencing for Rare Variants

Tao Wang,[1] Chang-Yun Lin,[2] Yuanhao Zhang,[3] Ruofeng Wen,[3] and Kenny Ye[1]

[1] *Department of Epidemiology and Population Health, Albert Einstein College of Medicine, New York, NY 10461, USA*
[2] *Department of Applied Mathematics and Institute of Statistics, National Chung Hsing University, Taichung 402, Taiwan*
[3] *Department of Applied Mathematics and Statistics, Stony Brook University, New York, NY 11794, USA*

Correspondence should be addressed to Tao Wang, tao.wang@einstein.yu.edu

Received 23 March 2012; Accepted 6 June 2012

Academic Editor: Wei T. Pan

Next generation sequencing (NGS) is a revolutionary technology for biomedical research. One highly cost-efficient application of NGS is to detect disease association based on pooled DNA samples. However, several key issues need to be addressed for pooled NGS. One of them is the high sequencing error rate and its high variability across genomic positions and experiment runs, which, if not well considered in the experimental design and analysis, could lead to either inflated false positive rates or loss in statistical power. Another important issue is how to test association of a group of rare variants. To address the first issue, we proposed a new blocked pooling design in which multiple pools of DNA samples from cases and controls are sequenced together on same NGS functional units. To address the second issue, we proposed a testing procedure that does not require individual genotypes but by taking advantage of multiple DNA pools. Through a simulation study, we demonstrated that our approach provides a good control of the type I error rate, and yields satisfactory power compared to the test-based on individual genotypes. Our results also provide guidelines for designing an efficient pooled.

1. Introduction

An understanding of the role of genetic variants in human diseases provides valuable insights into the etiology of diseases. Next generation sequencing (NGS), also known as massively parallel sequencing, is a revolutionary technology for biomedical research [1]. The production of large numbers of low-cost reads makes NGS useful for many applications.

One of the most important applications is to identify DNA variants responsible for human diseases [2]. However, it is still extremely expensive and time consuming to sequence individual genomes of a large number of individuals required to achieve reasonable statistical power for identifying disease variants of common diseases [3, 4]. The yields of a typical single run of NGS are very high (e.g., more than 20 billion bases can be obtained routinely for illumina genome analyzer (GA) II)). Indeed, the throughput of the smallest functional unit, for example, a single "lane" of a sequencer can generate data amounting to many thousand-fold coverage for a small target region of interest, which is far greater than what is needed for genotyping one individual as the individual genotype at a specific position is expected to be accurately called at about 15–30 fold coverage. In this case, it is much more efficient to simultaneously sequence multiple targeted regions of many individuals.

To maximize the power of next generation sequencer, one technology that allows sequencing multiple DNA samples together is bar-coding. Bar-coding ligates the DNA fragments of each sample to a short, sample-specific DNA sequence, and then sequences these DNA fragments from multiple subjects in one single sequencing run [5]. However, the cost on sequencing template preparation could be high for bar-coding if the construction of libraries and bar-coding must be applied to each individual before sequencing. Another approach called "DNA sudoku" is designed to ascertain rare variants by assigning each individual to multiple DNA pools, and decoding the identity of rare variants based on a certain pooling scheme [6–8]. Because the number of pools required for "DNA sudoku" to infer the identity of a rare variant could be much smaller than the number of subjects, the cost on the sequencing and preparation of templates is reduced.

For screening disease association of genetic variants, an alternative approach simply sequences pooled DNA samples of cases and controls, respectively. The idea of this approach is based on comparing the estimated allele frequencies between cases and controls without actually inferring individual genotypes. Compared to bar-coding or "DNA Sudoku", the pooling approach is even more cost- and time-efficient and hence well-suited for screening disease variants. The savings on cost and time come from two sources. The first is that estimating the allele frequency requires much less depth of coverage per individual than that required for calling the genotype of each individual [7]. The second is the reduced efforts in library preparation for a large number of DNA samples. The pooling approach was proposed earlier for high throughput SNP arrays [9–11], but it was not widely accepted as SNP array technology does not provide the required accuracy of the estimate of the allele frequency in the pooled sample. NGS technology, however, has been demonstrated that it could provide an accurate estimate of the allele frequency, as shown by recent studies [12–14]. Another advantage of pooled sequencing is that it has the potential to detect rare variants which may explain the "missing" heritability that are not identified by current array technologies.

To identify disease association by pooled NGS, several key issues have to be addressed by the study design or the analytic approach. NGS has relatively high levels of base-calling errors, which are highly specific to genomic positions as well as experimental runs [15]. Although the average error rate is likely less than 1% after filtering out bases with low quality, the consequence of sequencing error is still not negligible in pooled sequencing, particularly because sequencing errors and the bases of a true variant could confound each other in a large DNA pool. As such, simply applying some filters to eliminate sequencing errors may lead to falsely remove bases of the true variant. Instead of trying to eliminate sequencing errors, this issue could be addressed by using efficient experiment designs and appropriate analytic approaches so that the locus-specific sequencing error rate can be accurately estimated and incorporated into statistical models of testing disease association. It was shown by us and others that disease association can be validly and efficiently examined when the sequencing error parameters can be correctly specified [16, 17]. However, current statistical approaches often assume that the sequencing error parameters in the statistical models are known [16] or able to be estimated by using an internal control, such as a segment of plasmid DNA, in the pool [13, 17]. Because of the high variation in the error rates between genomic positions as well as different runs/lanes of NGS instruments, it is not adequate to use the average error rate for adjusting for the bias, potentially causing either inflated type I error rates or loss in statistical power [16, 17]. How to accurately estimate position- and lane-specific error rates relies on efficient pooled sequencing designs.

One major advantage of the sequencing technology over the SNP array is that it can ascertain novel rare variants that are not present on the array panels. However, it is well known that the power of testing association of rare variants individually is very limited due to the low occurrence of the rare alleles. To improve the statistical power, many statistical approaches have been proposed to simultaneously test a group of rare variants over recent years. Among them, the "collapsing" approach defines a score for each individual by the unweighted or weighted sum of the rare variant alleles of multiple positions in the targeted region. This approach essentially increases the "allele frequency" by pooling multiple variants, and hence improves the power, but the power of this approach relies on the assumption that all rare variant alleles have effects in the same direction [18–23]. To avoid such an assumption that is often not realistic, other approaches such as the test statistic based on the genomic distance and *C*-alpha test were also proposed [24, 25]. Nevertheless, these approaches all require individual genotypes for accounting for the linkage disequilibrium (LD) among multiple variants. Because LD information is largely lost in pooled sequencing, how to test disease association of a group of rare variants is still an open question.

In this paper, we proposed blocked pooling design combining bar-coding and pooling sequencing, along with a new multivariant testing procedure, for testing disease association of rare variants. We conducted a simulation study to examine the performance of the new approach under various situations.

2. Methods

2.1. Blocked Pooling Design

Sequencing error is the major concern of pooled sequencing because it has a significant impact on the validity and efficiency of testing disease association [17]. Because sequencing errors and bases of true variant alleles confound each other in a single DNA pool, it is often too difficult to differentiate them to obtain an accurate estimate of the sequencing error rate and the allele frequency. Nevertheless, if the sequencing error rates across multiple pools are consistent, a more accurate estimate may be obtained by combining data of multiple DNA pools. To understand sequencing error rates across multiple pools, we have conducted a study that used the GA II system to sequence the pooled mitochondria DNA (mtDNA) from 20 subjects, whose mtDNA had been sequenced previously using Sanger dideoxy sequencing on an ABI3730XL [15]. The pooled mtDNA samples were multiplexed by bar-coding at 2 pools per lane and replicated in another lane on a different flow cells. Using the results of Sanger sequencing as the reference, the data suggested that locus-specific base-calling error rates are quite consistent between two pools multiplexed in one lane, but vary between two lanes in different flow cells. In addition, sequencing error rates across genomic positions have a significant variation. Although the majority of positions have an error rate lower than 1%, it can be as high as 20%, suggesting that the use of the average error rate of all genomic positions to account for sequencing error is not adequate in testing disease association, even when such an error rate is estimated from a segment of plasmid DNA as the internal control [13, 16, 17].

Based on data from the pooled mtDNA sequencing study, we propose to combine pooling and bar-coding approaches to sequence multiple DNA pools of cases and controls in one lane with each pool indexed. This experiment design can be looked upon as blocked design, which is known to improve the statistical validity as well as power, in particular, when a large variability between blocks (here lanes) is present [26]. With multiple pools indexed in one lane, the sequencing errors are largely consistent across multiple indexed pools, while bases of the true variant alleles may vary because different numbers of alleles are likely sampled in different pools. The idea of blocked pooling design is that each pool can serve as the control of other pools in the same block to eliminate effects of sequencing errors, and eventually improve the validity and efficiency of testing disease association. Furthermore, an unbalanced pooling design (different sizes of pools) could be considered to obtain an even more accurate estimation of the sequencing error rate. For example, a pool with a single individual and another pool with a large number of individuals can be multiplexed in one lane. In this design, the pool with one individual serves as the control for accurately estimating the sequencing error rate, while the pool with a large number of individuals provides the data for accurately estimating the allele frequency. The pool with a small number of individuals could provide a more accurate estimate of the sequencing error rate because of the large difference between the allele frequency (e.g., 0, 0.5, or 1 for one individual) and the sequencing error rate. In the ideal situation in which there are no sequencing errors, the balanced pooling design provides the most efficient estimate of the allele frequency because of the consistent depth of coverage for each individual. However, in the presence of sequencing errors it is necessary to balance between estimating the allele frequency and estimating the sequencing error rate to obtain an optimal association result. We empirically evaluated the importance of parameters of blocked designs in terms of the bias and standard error (SE) of the estimate of the sequencing error rate and the allele frequency.

2.2. Estimating the Sequencing Error Rate and Testing Association of Single Variants

For a case-control study, let the phenotype of a subject be denoted by $i = 1, 0$ for cases or controls, respectively. We are interested in the question of whether the variant allele is associated with disease. Let θ_i be the allele frequency of the group i. The statistical hypothesis of association can be tested by examining if cases have a different frequency of the variant allele from controls, which could be written as $H_0 : \theta_1 = \theta_0$ versus $H_1 : \theta_1 \neq \theta_0$.

Let n_{ij} be the total number of chromosomes and let v_{ij} be the number of the variant alleles at a locus of interest for the jth pool of ith group. For a pooled sequencing, v_{ij} is unknown and has to be estimated from sequencing reads. We assume that cases and controls are assigned in L_1 and L_0 pools, respectively, indexed in a single sequencing lane. After sequencing, m_{ij} sequencing bases at the locus are observed, and x_{ij} out of m_{ij} bases report the variant allele for the jth pool. To estimate the sequencing error rate (e) and the allele frequency (θ), we consider a simple EM algorithm, given by

(0) Initial $\theta^{(0)}$ and $e^{(0)}$,

(1) E step

$$w_j = p(v_j \mid x_j) = \frac{\binom{n_j}{v_j}\theta^{(0)v_i}\left(1-\theta^{(0)}\right)^{n_j-v_j}\binom{m_j}{x_j}\mathcal{A}}{\sum_0^{n_j}\binom{n_j}{v_j}\theta^{(0)v_j}\left(1-\theta^{(0)}\right)^{n_j-v_j}\binom{m_j}{x_j}\mathcal{A}}, \tag{2.1}$$

where $\mathcal{A}$ donates
$[v_j/n_j(1-e^{(0)}) + (1-v_j/n_j)e^{(0)}]^{x_j}\{1-[v_j/n_j(1-e^{(0)}) + (1-v_j/n_j)e^{(0)}]\}^{m_j-x_j}$,

(2) M step

$$\theta^{(1)} = \frac{\left[\sum_{j=1}^{L}\sum_{v_j=0}^{n_j}(w_j v_j)\right]}{\sum_{j=1}^{L} n_j}, \tag{2.2}$$

$$e^{(1)} = \frac{\sum_{j=1}^{L}\sum_{v_j=0}^{n_j} w_j\left(x_j - v_j m_j/n_j\right)}{\sum_{j=1}^{L} m_j}, \tag{2.3}$$

(3) Iteratively update θ and e until converge.

For testing disease association of a rare variant, we have proposed a simple testing procedure based on a parametric bootstrap (PB), which is defined by the following steps:

(1) estimating the sequencing error rate ($\hat{e}$) and allele frequency ($\hat{\theta}$) of DNA pools under the null hypothesis by the above EM algorithm;

(2) calculating the test statistic $T = \sum(n_{1j}/\sum n_{1j})x_{1j}/m_{1j} - \sum(n_{0j}/\sum n_{0j})x_{0j}/m_{0j}$;

(3) sampling $\tilde{x}_i = (\tilde{x}_{01},\ldots,\tilde{x}_{11},\ldots)$ and calculating the test statistic $\tilde{T}$. First, the number of the variant alleles for each pool is sampled from $Binom(n_{ij},\hat{\theta})$; $\tilde{x}_{ij}$ is then sampled from $Binom(m_{ij}, v_{ij}/n_{ij}(1-\hat{e}) + (1-v_{ij}/n_{ij})\hat{e})$; and lastly $\tilde{T}$ is calculated based on $\tilde{x}_i$;

(4) replicating (3) many times and estimating the P value by the proportion of $|\tilde{T}| > |T|$.

2.3. Testing Association of Multiple Rare Variants

Because the statistical power to detect disease association of rare variants individually is often limited, it is useful to jointly test association of a group of rare variants, for example, rare variants in an exon or a gene. Our test statistic is based on P values of individual variants. Let $p_r(r = 1,\ldots,R)$ be the P value for variant r. The test statistic is defined by

$$z = \frac{a^T Z}{\sqrt{a^T a}}, \tag{2.4}$$

where $Z = (Z_1, \dots, Z_R)^T$ in which the element $Z_r = \Phi^{-1}(p_r)$ is the corresponding upper-tail Z score transformed from the P value, and $a = (a_1, \dots, a_R)^T$ in which a_r is the weight given to variant r. Because the functional information of each variant is usually not available, a reasonable approach is to give equal weights to all variants because it is not prejudiced about which variants are expected to be more relevant to disease. Of note, this test statistic is in spirit close to many test statistics based on individual genotypes, such as the VEGAS statistic [27], Empirical Bayesian score statistic proposed by Goeman (2006) [28], the statistic based on the genomic distance [29], the logistic kernel machine based test statistic [30, 31], as well as C-alpha test [24].

When multiple rare variants are in linkage equilibrium (no correlation), the statistic follows a standard normal distribution. The question is that, when multiple rare variants are in LD, the P value cannot be obtained based on a standard distribution. The permutation procedure randomly shuffling the disease status is often used to account for the correlation among genetic variants. However, such a procedure requires individual genotypes that are not available in pooled sequencing. Instead, we can take a Monte Carlo approach by simulating the test statistics of individual variants under the null hypothesis from multivariate normal distribution to evaluate the P value. In this approach, we simulate the multivariate normally distributed vector with mean 0 and covariance Σ, the $R \times R$ matrix of pair-wise correlations. To do this, we use the Cholesky decomposition: a vector of R independent, standard normally distributed random variables is first generated; then it is multiplied by the Cholesky decomposition matrix of Σ. The simulated test statistic is calculated based on the multivariate normally distributed vector. A large number of multivariate normal vectors are simulated, and the empirical P value is defined by the proportion of simulated test statistics that exceed the observed test statistic .

The statistical challenge is how to estimate the covariance matrix Σ without individual genotypes available. By treating single pools as the sample unit, we estimate the covariance matrix based on the number of variant alleles of pools, instead of the number of alleles of individuals. One option is the standard unbiased empirical covariance matrix $\widehat{S}$ with entries defined as $s_{rr'} = (1/(L_0 + L_1 - 1)) \sum_{j=1}^{L_0+L_1} (\widehat{v}_{jr} - \overline{v}_r)(\widehat{v}_{jr'} - \overline{v}_{r'})$, which $\widehat{v}_{jr}$ is the estimated number of variant alleles of the rth variant in the jth pool. However, this unbiased estimate is known to be inefficient, particularly because the number of the pools is often relatively small. Because rare mutations usually occur on different haplotypes within a target region [32], and therefore their correlations are often low. This motivated us to use an empirical Bayesian shrinkage estimate of the covariance, which may provide better balance between efficiency and bias [33]. The proposed shrinkage estimate is in the following form:

$$S^* = \lambda I_{R\times R} + (1 - \lambda)\widehat{S}, \tag{2.5}$$

where $\lambda = (\sum_{r \neq r'} v\widehat{a}r(s_{rr'}) + \sum_r v\widehat{a}r(s_{rr'}))/(\sum_{r \neq r'} s_{rr'}^2 + \sum_r (s_{rr'} - 1)^2)$ is the shrinkage intensity. The idea of this empirical Bayesian estimate is that, when the data do not provide evidence of correlation of variants, the estimate is shrunk toward an identity matrix, the possibly efficient estimator under the assumption of independency of variants. Of note, this estimate is essentially equivalent to that proposed by Schafer [34].

2.4. Simulations

We conducted a simulation study to examine the impact of varied parameters of pooled designs on the estimation of the sequencing error rate and the allele frequency as well as the test of disease association in terms of validity and efficiency. For each replicate, the pooled sequencing reads of each pool were simulated in the following two-steps: the individual genotypes were first generated under Hardy-Weinberg equilibrium; the sequencing reads of each pool were then generated independently. The sample size was set at 500 cases and 500 controls; individuals were included in different numbers of DNA pools under either the balanced or the unbalanced designs. For the unbalanced designs, one half of the pools included single subjects, and the remaining individuals were evenly assigned to the other half of pools. We set the numbers of reads were consistent cross pools. The type I error rate and power will be evaluated by the proportion of replicates having a P value that is less than a significant level of 0.05. For each simulated situation, the process was repeated for 1,000 replicates.

The performance of the PB test was examined for testing single variants for different types of designs under different allele frequencies (1% and 5%), sequencing error rates (0.5% and 1%), depths of coverage per chromosome (5, 10, and 20×), and numbers of pools (2, 10 and 40). To evaluate the type I error rate, we simulated the sequencing reads under the null hypothesis of no association, in which circumstance the cases and controls have the same allele frequency. For comparison, we also considered a Naïve Fisher's (FN) exact test that is based on the estimated allele frequency without taking sequencing errors into account, Fisher's exact test based on the estimated allele frequency with taking sequencing error into account (FE), and Fisher's exact test based on the true individual genotypes (FT). For the FN test, the number of variant alleles is directly estimated by the proportion of reads that report the variant allele. For the FE test, the number of the variant alleles is based on the allele frequency estimated by the EM algorithm; and the FT test assumes that the genotype of each individual is known and hence the number of the variant alleles can be simply counted. To evaluate the power, we fixed the allele frequency in controls, but allowed the allele frequency in cases to vary in order to yield different effect sizes.

The performance of the PB test was then examined for testing multiple variants. Different numbers of variants and varied correlations were considered. To simulate correlated variants, a set of variables was sampled from multivariate normal distribution with mean 0 and covariance Σ, which had equal pair-wise correlations ($\rho = 0$ or 0.5). The haplotype was generated by dichotomizing the normal variables based on the allele frequencies of cases and controls. The genotypes in each DNA pool were randomly sampled from a large number of haplotypes, and reads for each variant were then sampled independently. We examined multivariant tests based on three different estimates of covariance matrix [the unbiased empirical covariance estimate (E), the independent matrix (I), and the shrinkage estimate (S)] and compared them to the single-variant test with Bonferroni correction (min$P$).

3. Results

3.1. Estimating the Sequencing Error Rate and the Allele Frequency

Table 1 presents results for estimating the sequencing error rate and the allele frequency for balanced and unbalanced pooled sequencing designs under different sequencing depths of coverage, numbers of pools, allele frequencies, and sequencing error rates. As expected, the unbalanced pooling design had smaller bias of the estimate of the sequencing error rate than the balanced design. For example, when the allele frequency and the sequencing error rate were both 1% and the number of pools was 10, the bias of the unbalanced design was <0.0001,

Table 1: Estimated bias and standard error (SE) of the sequencing error rate and the allele frequency for the unbalanced and balanced designs under different sequencing depths of coverage, numbers of pools, allele frequencies, and sequencing error rates.

Depths of coverage	Pool number	θ	e	Unbalanced design				Balanced design			
				Bias (e)	SE (e)	Bias (θ)	SE (θ)	Bias (e)	SE (e)	Bias (θ)	SE (θ)
5×	2	0.01	0.005	0.0000	0.0015	0.0002	0.0029	−0.0011	0.0007	0.0011	0.0019
			0.01	−0.0001	0.0020	0.0001	0.0034	−0.0063	0.0014	0.0062	0.0025
		0.05	0.005	−0.0002	0.0014	0.0002	0.0046	−0.0011	0.0056	0.0010	0.0062
			0.01	−0.0002	0.0019	0.0002	0.0051	−0.0066	0.0054	0.0061	0.0060
	10	0.01	0.005	−0.0001	0.0014	0.0003	0.0028	−0.0009	0.0024	0.0011	0.0029
			0.01	0.0000	0.0021	0.0003	0.0034	−0.0074	0.0027	0.0078	0.0033
		0.05	0.005	0.0002	0.0014	−0.0004	0.0047	0.0084	0.0116	−0.0076	0.0108
			0.01	0.0000	0.0020	−0.0001	0.0051	0.0031	0.0127	−0.0032	0.0119
	40	0.01	0.005	0.0002	0.0014	−0.0001	0.0027	0.0007	0.0017	−0.0003	0.0017
			0.01	0.0001	0.0020	0.0001	0.0034	−0.0005	0.0033	0.0010	0.0038
		0.05	0.005	0.0019	0.0015	−0.0018	0.0046	0.0098	0.0063	−0.0078	0.0058
			0.01	0.0009	0.0020	−0.0005	0.0048	0.0090	0.0079	−0.0073	0.0072
10×	2	0.01	0.005	−0.0001	0.0010	0.0002	0.0020	−0.0006	0.0002	0.0006	0.0012
			0.01	−0.0001	0.0014	0.0001	0.0025	−0.0056	0.0001	0.0055	0.0014
		0.05	0.005	−0.0001	0.0010	0.0000	0.0033	0.0002	0.0035	−0.0002	0.0038
			0.01	−0.0001	0.0014	0.0004	0.0035	−0.0049	0.0033	0.0045	0.0039
	10	0.01	0.005	0.0000	0.0010	0.0000	0.0020	−0.0003	0.0017	0.0004	0.0019
			0.01	0.0000	0.0014	0.0000	0.0024	−0.0074	0.0023	0.0076	0.0029
		0.05	0.005	0.0000	0.0010	0.0001	0.0032	0.0041	0.0093	−0.0037	0.0087
			0.01	−0.0001	0.0014	0.0001	0.0037	−0.0010	0.0101	0.0009	0.0093
	40	0.01	0.005	0.0000	0.0009	0.0001	0.0017	0.0006	0.0010	−0.0001	0.0008
			0.01	0.0000	0.0014	0.0002	0.0023	0.0002	0.0015	0.0001	0.0011
		0.05	0.005	0.0006	0.0011	−0.0004	0.0032	0.0035	0.0021	−0.0021	0.0024
			0.01	0.0003	0.0016	0.0000	0.0035	0.0025	0.0032	−0.0015	0.0032
20×	2	0.01	0.005	−0.0001	0.0006	0.0001	0.0014	−0.0003	0.0001	0.0003	0.0009
			0.01	−0.0001	0.0010	0.0001	0.0017	−0.0053	0.0000	0.0052	0.0010
		0.05	0.005	0.0000	0.0007	0.0001	0.0024	0.0000	0.0023	0.0000	0.0027
			0.01	−0.0001	0.0010	0.0001	0.0025	−0.0051	0.0014	0.0047	0.0021
	10	0.01	0.005	−0.0001	0.0007	0.0001	0.0014	−0.0002	0.0011	0.0003	0.0011
			0.01	0.0000	0.0010	0.0000	0.0017	−0.0074	0.0026	0.0076	0.0032
		0.05	0.005	0.0000	0.0007	0.0000	0.0023	0.0003	0.0034	−0.0002	0.0035
			0.01	0.0000	0.0010	0.0001	0.0026	−0.0047	0.0042	0.0043	0.0042
	40	0.01	0.005	0.0001	0.0006	0.0000	0.0010	0.0004	0.0007	0.0000	0.0004
			0.01	0.0000	0.0009	0.0001	0.0013	0.0002	0.0009	0.0000	0.0005
		0.05	0.005	0.0002	0.0008	−0.0002	0.0022	0.0019	0.0013	−0.0006	0.0014
			0.01	0.0002	0.0011	−0.0001	0.0024	0.0009	0.0019	−0.0003	0.0017

while the bias of the balanced design was −0.0074. In addition, the SE of the sequencing error rate of the unbalanced design was comparable to that of the balanced design. Interestingly, the bias of the allele frequency of the unbalanced design was also smaller than the balanced design and their SEs were comparable. Surprisingly, the bias and SE of both the sequencing error rate and the allele frequency were not significantly improved by increasing the number of pools from 2 to 40. As expected, the bias and SE of the sequencing error rate and the allele frequency tended to decrease with an increasing sequencing coverage.

3.2. Testing for Single Variants

3.2.1. Type I Error Rate

The empirical type I error rate at a significance level of 0.05 is shown in Table 2. In general, the FT test tended to be overconservative when the allele frequency was low. When the depth of coverage is relatively low (5×), the FE test often had a very poor control of type I error rate for both the unbalanced and balanced designs, partially because the variance of the estimate of the number of variant alleles is not negligible due to the low depth of coverage. The FN test was either overliberal or overconservative because it ignores both the sequencing error and the variation of the estimate of the number of variant alleles. Table 2 indicates that the type I error rate of the PB test was consistently close to the nominal level of 0.05 for the unbalanced design, while it could be either liberal or conservative for the balanced design, which is likely due to that the balanced design could not provide an accurate estimate of the sequencing error rate and the allele frequency under low depths of coverage. With an increased depth of sequencing coverage (10× and 20×), the FE test had an improved control of the Type I error rate for the unbalanced design, while it was still a little conservative for the balanced design. The FN test tended to be more conservative for both the balanced and unbalanced designs with an increasing depth of coverage. The PB test consistently kept a good control of the type I error for the unbalanced design.

3.2.2. Power

We only evaluated the power of the PB test for the unbalanced design because the balanced design did not provide a good control of the type I error rate. As the reference, the FT test that assumes individual genotypes are observed was compared.

Figure 1 shows the empirical power of the PB test for testing association of single variants. Because of the confounding effect of the sequencing error as well as the uncertainty of the estimate of the number of variants allele in a pool, the PB test was generally less powerful than the FT test. However, the loss in power was reduced with a decreasing sequencing error rate or an increasing sequencing depth of coverage. The power of the PB test was not significantly different between various numbers of pools, in particular for the numbers of pools were 10 and 40. The difference in power between the PB test and the FT test seemed more obvious for a more common variant, which could be due to the conservativeness of the FT test for testing relatively rare variants. The results of two versions of Fisher’s exact test based on the estimated number of the variant alleles were not presented here, because they generally have a poor control of the type I error rate. Nevertheless, after adjusting for the inflated type I rate they tend to be less powerful than the proposed PB test,

Table 2: Type I error rates at a level of 5% for the PB test and Fisher's exact tests under various depths of coverage, numbers of pools, allele frequencies, and the error rates for testing association. Sample size was set at 500 cases and 500 controls.

Depth of coverage	Pool number	θ	e	Unbalanced design				Balanced design			
				PB	FT	FN	FE	PB	FT	FN	FE
5×	2	0.01	0.005	0.046	0.033	0.041	0.096	0.041	0.036	0.027	0.017
			0.01	0.044	0.041	0.024	0.16	0.013	0.028	0.008	0.004
		0.05	0.005	0.053	0.041	0.072	0.086	0.059	0.037	0.049	0.113
			0.01	0.047	0.044	0.062	0.091	0.059	0.042	0.052	0.092
	10	0.01	0.005	0.048	0.026	0.04	0.111	0.056	0.039	0.025	0.065
			0.01	0.057	0.039	0.028	0.152	0.02	0.026	0.012	0.033
		0.05	0.005	0.051	0.037	0.066	0.083	0.09	0.035	0.052	0.302
			0.01	0.046	0.033	0.06	0.093	0.076	0.05	0.048	0.317
	40	0.01	0.005	0.05	0.029	0.032	0.082	0.049	0.028	0.019	0.044
			0.01	0.049	0.03	0.02	0.122	0.072	0.05	0.017	0.133
		0.05	0.005	0.054	0.032	0.065	0.073	0.07	0.044	0.047	0.043
			0.01	0.055	0.032	0.065	0.098	0.083	0.033	0.05	0.088
10×	2	0.01	0.005	0.045	0.033	0.02	0.066	0.051	0.034	0.01	0.018
			0.01	0.04	0.037	0.012	0.107	0.032	0.037	0.011	0.013
		0.05	0.005	0.055	0.042	0.056	0.063	0.057	0.053	0.03	0.055
			0.01	0.039	0.042	0.034	0.056	0.04	0.045	0.033	0.051
	10	0.01	0.005	0.05	0.032	0.016	0.072	0.061	0.045	0.021	0.062
			0.01	0.045	0.037	0.016	0.098	0.012	0.031	0.006	0.032
		0.05	0.005	0.051	0.046	0.052	0.067	0.066	0.042	0.044	0.168
			0.01	0.048	0.039	0.04	0.071	0.059	0.042	0.03	0.181
	40	0.01	0.005	0.043	0.036	0.023	0.051	0.048	0.026	0.012	0.034
			0.01	0.05	0.032	0.01	0.079	0.053	0.026	0.003	0.035
		0.05	0.005	0.041	0.042	0.04	0.049	0.06	0.049	0.039	0.045
			0.01	0.051	0.047	0.041	0.065	0.053	0.033	0.031	0.046
20×	2	0.01	0.005	0.06	0.024	0.018	0.054	0.058	0.033	0.015	0.037
			0.01	0.036	0.034	0.008	0.049	0.014	0.031	0.003	0.006
		0.05	0.005	0.043	0.036	0.039	0.043	0.045	0.035	0.032	0.036
			0.01	0.054	0.039	0.043	0.054	0.037	0.04	0.026	0.035
	10	0.01	0.005	0.045	0.025	0.014	0.047	0.061	0.03	0.015	0.051
			0.01	0.05	0.04	0.004	0.058	0.011	0.026	0.004	0.034
		0.05	0.005	0.043	0.035	0.034	0.041	0.055	0.038	0.036	0.05
			0.01	0.06	0.041	0.036	0.064	0.048	0.038	0.029	0.051
	40	0.01	0.005	0.051	0.023	0.012	0.035	0.042	0.036	0.013	0.034
			0.01	0.054	0.033	0.002	0.052	0.039	0.022	0.002	0.023
		0.05	0.005	0.052	0.033	0.045	0.049	0.061	0.046	0.038	0.036
			0.01	0.052	0.04	0.034	0.056	0.052	0.048	0.039	0.043

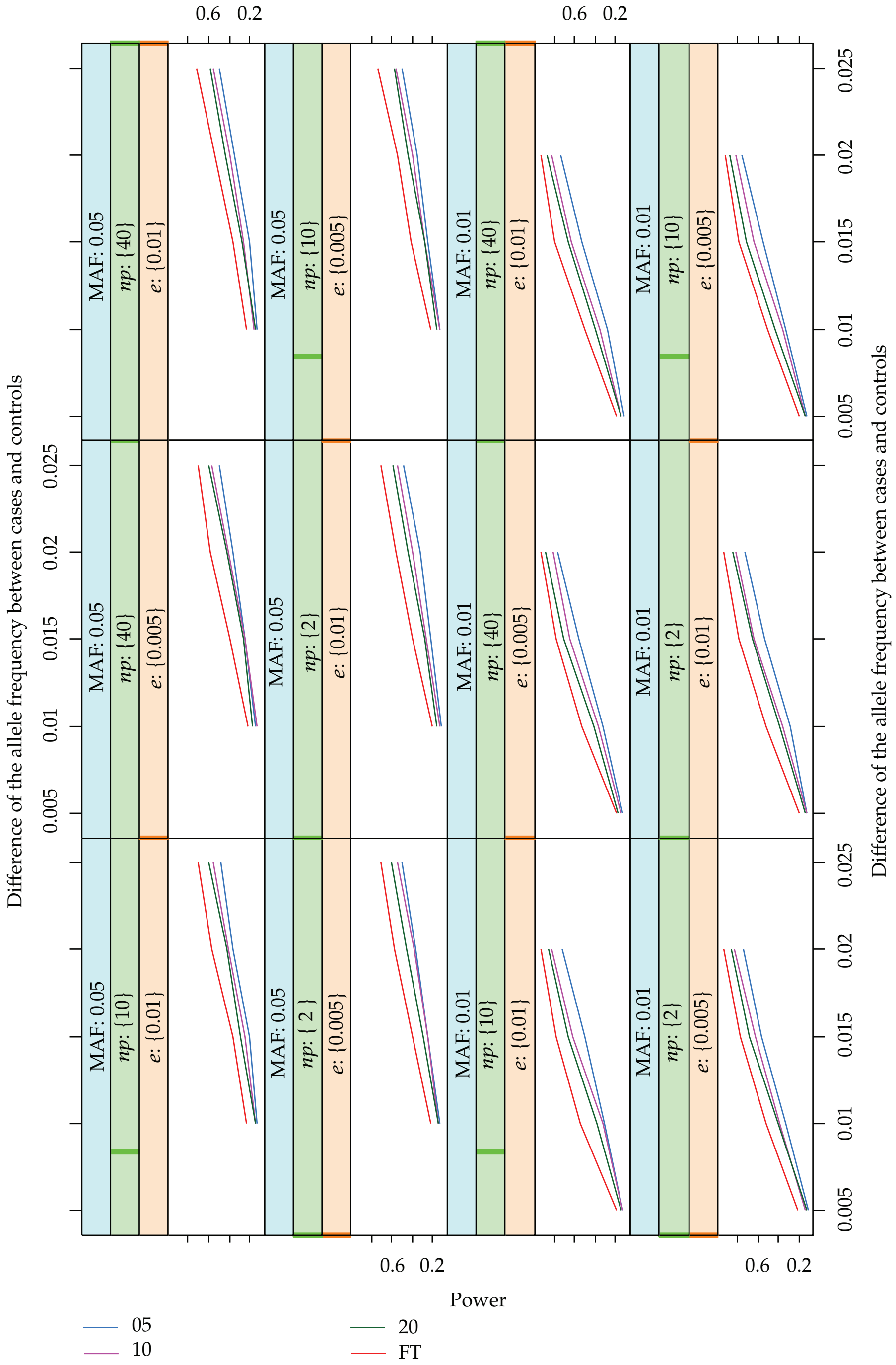


Figure 1: Empirical power at the a level of 5% for the parametric bootstrap (PB) test as a function of the difference of the allele frequency between cases and controls under various sequencing error rates, numbers of pools, and depths of sequencing coverage for testing association. Sample size was set at 500 cases and 500 controls. The minor allele frequencies (MAF) of controls were set at 0.01 and 0.5, sequencing error rates (e) were set at 0.005 and 0.01; and numbers of pools were set at 2, 10, and 40. Lines with different colors indicate the power of the PB test under different depths of coverage, which are compared to that of the Fisher's exact test (read line) based on the true individual genotypes (FT).

in particular for rare variants, because of the tendency of conservativeness of Fisher’s exact test itself in particular for rare variants (data not shown).

3.3. Testing for Multiple Variants

3.3.1. Type I Error Rate

The empirical type I error rate at a significant level of 0.05 for testing association of multiple rare variants is shown in Table 3. The multi-variant PB test based on the empirical unbiased estimate of the covariance had the worst performance, it was too liberal when multiple rare variants were in linkage equilibrium, while it was overconservative when variants were in LD. This was more obvious when the sequencing error rate was high (1%). As expected, the test based on an *identity* covariance matrix had a good control of the type I error rate when multiple variants were uncorrelated, but it tended to be liberal when variants were in LD. The single-variant test based on Bonferroni correction was consistently conservative when variants were in either LD or linkage equilibrium. Compared to other tests, the multivariants PB test based on a shrinkage estimate had the best performance. The results were similar for different numbers of pools for an unbalanced design. As expected, the type I error rate was improved for the test based on the empirical estimate of the covariance with an increasing number of pools. The PB test based on the shrinkage estimate kept a good control of the type I error rate.

3.3.2. Power

Figure 2 shows the empirical power of different tests for testing association of multiple variants under various numbers of pools, numbers of variants, sequencing error rates, depths of sequencing coverage, and correlation structures. In general, the single-variant test with Bonferroni correction had the worst performance in terms of power, which may be due to two reasons: first, it does not make use of the accumulated effects from all variants; second, it has a conservative type I error rate. Among different multi-variant tests, the test based on the unbiased estimate of the covariance was consistently less powerful than the other two tests, even though it had a liberal type I error rate when variants were in LD (data not shown). The power of the tests based on a shrinkage estimate and an identity covariance matrix was comparable when variants are in linkage equilibrium (Figure 2(a)), but the identity covariance matrix seemed slightly more powerful than the shrinkage estimate in particular when the variants were in LD, which may be due to the fact that the test based on an identity covariance matrix had a liberal type I error rate in this case (data not shown).

4. Discussion

In this paper, we addressed two important questions of testing disease association of rare variants by pooled sequencing. One critical issue is that the sequencing error rate is high and has a significant variability across genomic positions. Ignoring the position-specific sequencing error could lead to a biased estimate of the allele frequency, and eventually a biased association result that can be either conservative or liberal, which was shown in our simulations. Another important issue is that the pooling procedure introduces an extra

Table 3: Type I error rates at a level of 5% for multivariant tests under various allele frequencies, error rates, depths of coverage, and numbers of variants for testing association. Sample size was set at 500 cases and 500 controls. Simulations were based on the unbalanced design with 10 and 20 pools.

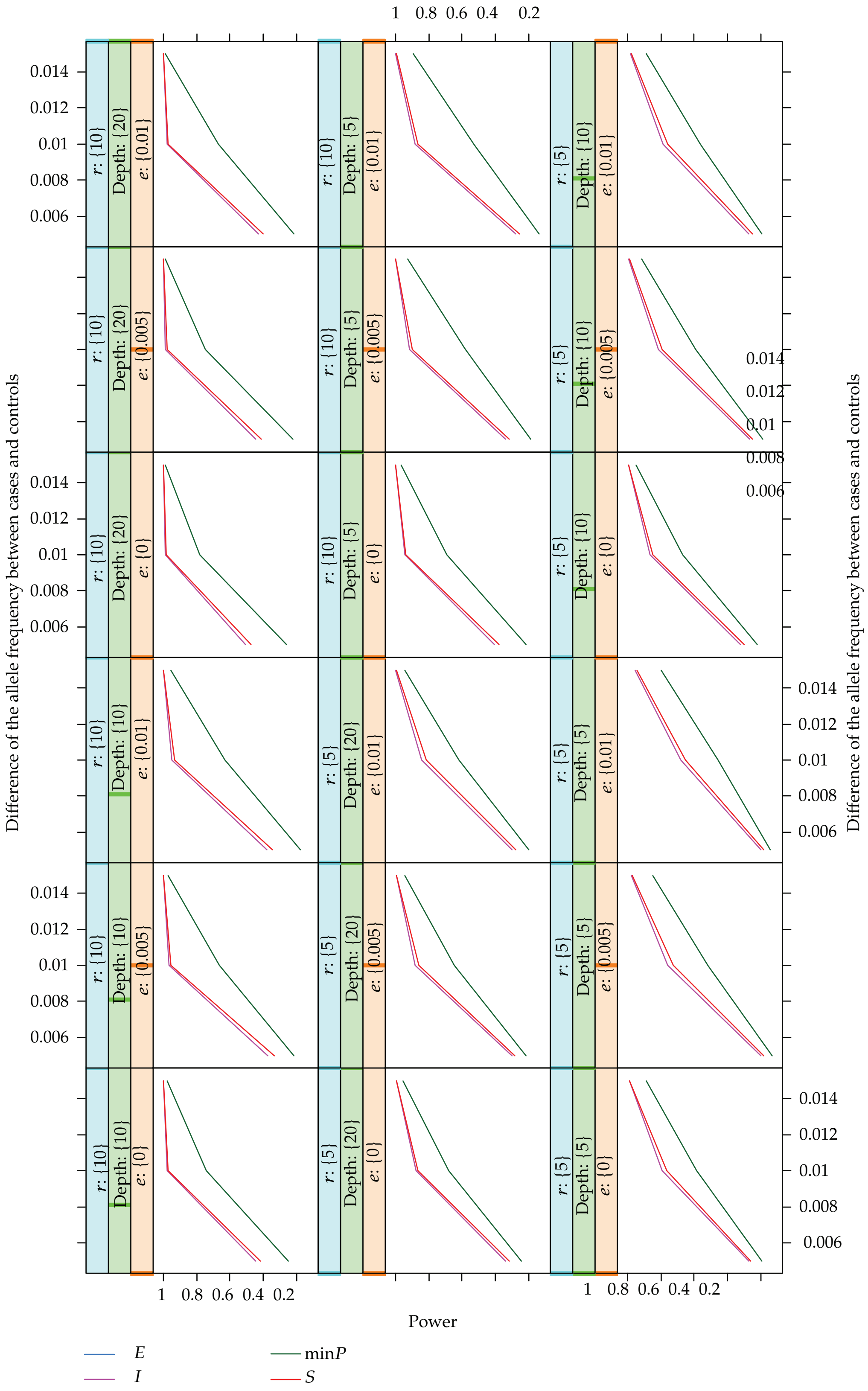
Number of pool	Number of variant	Depth	θ	e	Uncorrelated				Correlated ($\rho = 0.5$)			
					E	I	S	min	E	I	S	min
10	5	5×	0.01	0.005	0.058	0.038	0.039	0.037	0.051	0.067	0.062	0.036
				0.01	0.051	0.045	0.045	0.034	0.046	0.05	0.046	0.035
			0.05	0.005	0.062	0.052	0.058	0.048	0.028	0.046	0.038	0.04
				0.01	0.072	0.042	0.044	0.042	0.062	0.034	0.07	0.056
		10×	0.01	0.005	0.065	0.053	0.056	0.035	0.033	0.038	0.035	0.038
				0.01	0.057	0.043	0.044	0.042	0.037	0.053	0.043	0.04
			0.05	0.005	0.066	0.048	0.048	0.046	0.03	0.06	0.04	0.046
				0.01	0.07	0.052	0.052	0.06	0.032	0.072	0.052	0.052
		20×	0.01	0.005	0.063	0.049	0.052	0.039	0.039	0.058	0.045	0.035
				0.01	0.069	0.049	0.049	0.036	0.032	0.045	0.042	0.041
			0.05	0.005	0.054	0.042	0.042	0.042	0.038	0.066	0.048	0.064
				0.01	0.076	0.06	0.06	0.042	0.018	0.05	0.03	0.032
	10	5×	0.01	0.005	0.071	0.05	0.049	0.03	0.027	0.049	0.045	0.036
				0.01	0.064	0.051	0.052	0.023	0.03	0.049	0.043	0.036
			0.05	0.005	0.066	0.048	0.048	0.066	0.008	0.046	0.028	0.058
				0.01	0.068	0.044	0.048	0.036	0.02	0.082	0.062	0.05
		10×	0.01	0.005	0.063	0.058	0.058	0.04	0.024	0.063	0.05	0.045
				0.01	0.062	0.05	0.049	0.036	0.02	0.049	0.039	0.043
			0.05	0.005	0.064	0.054	0.052	0.048	0.014	0.056	0.044	0.044
				0.01	0.082	0.068	0.066	0.05	0.016	0.056	0.032	0.04
		20×	0.01	0.005	0.067	0.048	0.049	0.041	0.021	0.06	0.041	0.049
				0.01	0.067	0.046	0.045	0.038	0.022	0.053	0.04	0.026
			0.05	0.005	0.066	0.044	0.046	0.046	0.008	0.06	0.04	0.05
				0.01	0.098	0.078	0.084	0.044	0.008	0.068	0.038	0.044
20	5	5×	0.01	0.005	0.058	0.053	0.052	0.048	0.037	0.055	0.05	0.039
				0.01	0.052	0.048	0.05	0.049	0.028	0.044	0.037	0.041
			0.05	0.005	0.057	0.047	0.048	0.061	0.027	0.058	0.046	0.053
				0.01	0.058	0.051	0.047	0.044	0.026	0.055	0.044	0.047
		10×	0.01	0.005	0.052	0.043	0.045	0.031	0.039	0.06	0.052	0.058
				0.01	0.05	0.041	0.041	0.054	0.029	0.046	0.043	0.037
			0.05	0.005	0.048	0.038	0.04	0.044	0.025	0.055	0.045	0.054
				0.01	0.051	0.042	0.043	0.044	0.025	0.063	0.042	0.051
		20×	0.01	0.005	0.058	0.051	0.05	0.035	0.034	0.053	0.046	0.042
				0.01	0.051	0.04	0.038	0.035	0.026	0.043	0.041	0.041
			0.05	0.005	0.051	0.049	0.05	0.042	0.022	0.056	0.037	0.047
				0.01	0.054	0.045	0.041	0.042	0.02	0.055	0.034	0.044
	10	5×	0.01	0.005	0.068	0.06	0.058	0.024	0.03	0.058	0.046	0.03
				0.01	0.07	0.046	0.048	0.044	0.024	0.048	0.044	0.034
			0.05	0.005	0.08	0.06	0.65	0.052	0.012	0.062	0.04	0.038
				0.01	0.054	0.048	0.046	0.036	0.025	0.066	0.046	0.038
		10×	0.01	0.005	0.064	0.048	0.05	0.038	0.02	0.046	0.032	0.036
				0.01	0.076	0.05	0.054	0.048	0.02	0.068	0.054	0.034
			0.05	0.005	0.068	0.052	0.05	0.044	0.008	0.072	0.038	0.032
				0.01	0.08	0.048	0.052	0.05	0.01	0.07	0.038	0.05

Table 3: Continued.

Number of pool	Number of variant	Depth	θ	e	Uncorrelated				Correlated ($\rho = 0.5$)			
					E	I	S	min	E	I	S	min
		20×	0.01	0.005	0.056	0.054	0.052	0.042	0.022	0.04	0.034	0.036
				0.01	0.072	0.06	0.06	0.04	0.028	0.064	0.052	0.028
			0.05	0.005	0.074	0.054	0.06	0.038	0.026	0.09	0.054	0.038
				0.01	0.044	0.038	0.038	0.046	0.014	0.074	0.034	0.034

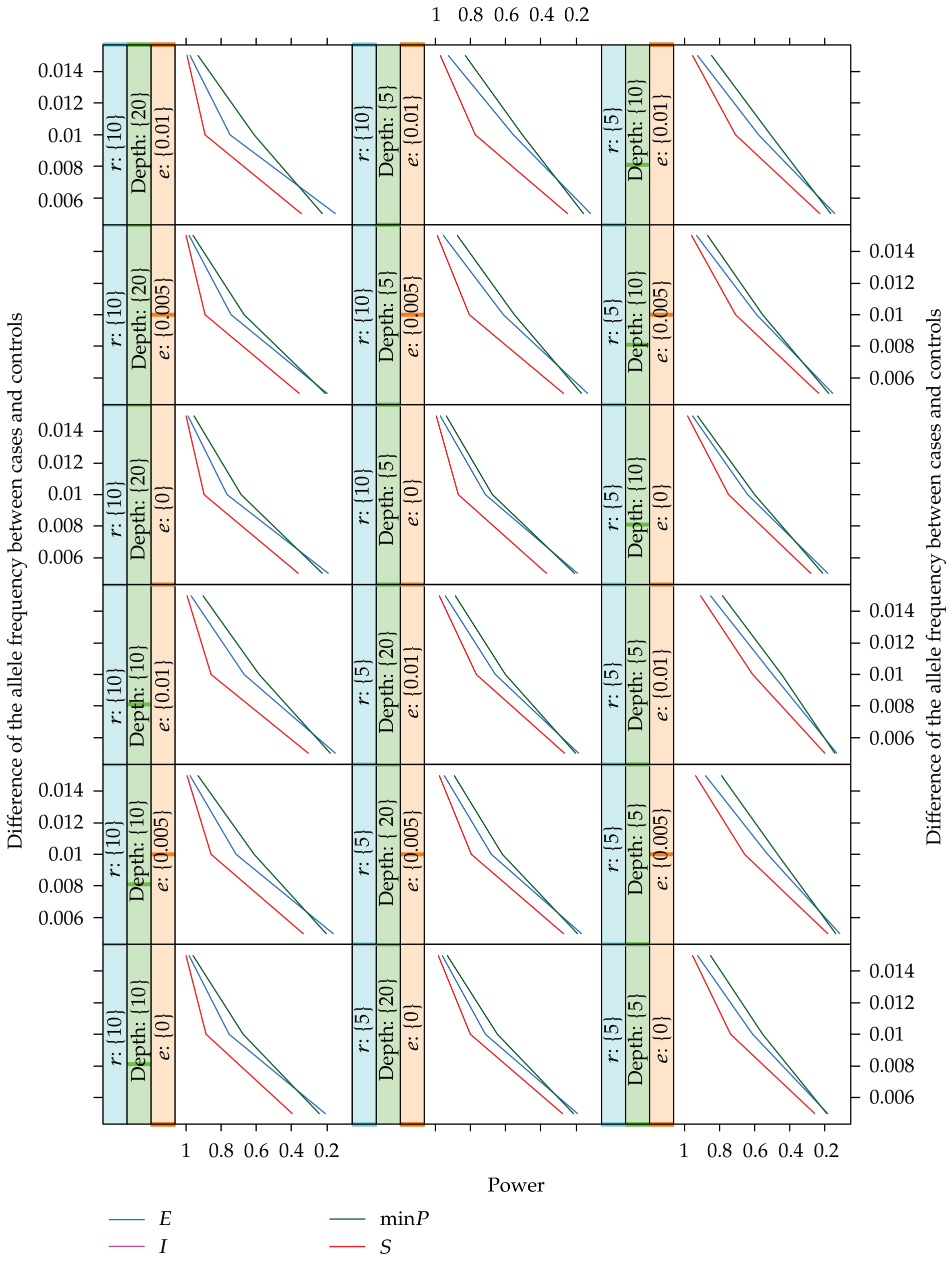
variance of the estimated number of variant alleles in a pool. Ignoring the uncertainty of the number of variant alleles could result in an inflated type I error rate, in particular in the case that the sequencing depth of coverage is low. This problem was indicated by the simulation results of the FE test which is directly based on the estimated number of variant alleles. To tackle these two questions, we proposed to use blocked pooling design to efficiently estimate the position-specific sequencing error rate and the allele frequency, along with a parameter bootstrap testing procedure to account for the extra variance of the estimate of the number of variant alleles in a pool.

We have proposed blocked pooling design to address the above two questions. Although blocked design in this paper was discussed based on lanes of flow cell, the similar idea could be extended to flow cells to take into account two sources of variation: variation between lanes within a flow cell and variation between flowcells. Based on blocked pooling design, an EM algorithm was used for estimating the position-specific sequencing error rate by making use of data from multiple pools. We examined the bias and standard error of the estimate of the sequencing error rates of different pooling designs under various situations through simulations. Intuitively, the EM algorithm should have a better performance when the number of pools is large and the number of individuals in a single pool is small because of the large difference between the minimal allele frequency of a pool and the sequencing error rate. As the result, we found the unbalanced design in which one half of pools included single individuals could provide a much more accurate estimate of the sequencing error rate as well as the allele frequency, while it does not sacrifice much on the variance of these estimates. Previously, we found that misspecification sequencing error has much more important impact on the statistical power than other parameters of pooled sequencing, for example, the depth of coverage and the number of pools [17]. Because the unbalanced design could provide more accurate estimates of the error rate and the allele frequency, the proposed PB test based on the unbalanced design not only consistently maintained a good control of the Type I error rate, but also provided higher power than the balanced design under various situations, even when the depth of coverage was low (5×). For balanced design, however, the proposed PB test tended to be anticonservative for low coverage data. As such, we suggest that the unbalanced blocked design, rather than the more commonly-used balance design, should be used in practice. Before a pooled sequencing study, it may be a good strategy to perform a simulation study to obtain the optimal unbalanced design based on the size of sequencing region and total depth of coverage. Under our simulated situations, for the given number of subjects, depth of coverage and type of design, the number of pools ranging from 10 to 40 did not significantly improve the estimate of the sequencing error rate and the allele frequency, and hence it was not a significant parameter for the statistical power. This result could be important, because it suggested that the pooled sequencing can be very cost-effective by including a small number of large pools with many individuals and small



(a)

Figure 2: Continued.



(b)

Figure 2: Empirical power at a level of 5% as a function of the difference of the allele frequency between cases and controls for the proposed PB test based on various estimates of the covariance matrix for testing multiple rare variants under the unbalanced design. The allele frequency of controls was 0.01; the sample size was set at 500 cases and 500 controls; the error rates (e) were set at 0, 0.005, and 0.01; the depths of coverage were set at 10× and 20×, and the numbers of pools (r) were set at 5 and 10. Figure (a) shows the power of the PB test based on an identity covariance matrix (I), the PB test based on the shrinkage estimate of the covariance matrix (S) and the single-variant test with Bonferroni correction the number of variants for independent variants (min$P$). Figure (b) shows the power of the PB test based on the empirical estimate of the covariance matrix (E), the PB test based on the shrinkage estimate of the covariance matrix (S) and the single-variant test with Bonferroni correction (min$P$) the number of variants for independent variants.

pools with single individuals in an unbalanced design, which is able to achieve adequate power.

As a single rare variant is likely to have a low marginal effect on disease risk, particularly in the presence of genetic heterogeneity, it is beneficial to jointly test a group of rare variants in a functional unit, such as genes or pathways. We extended the PB method for multiple rare variants. As with other multivariant tests based on individual genotypes, the multivariant PB test is designed for situations in which many rare variants present in the target region. Because our multivariant test is defined by the sum of Z scores transformed from single P values, it does not rely on the assumption on the direction of effects. Even if the effects of rare allele are uniformly in one direction, such as increasing risk, the proposed test can easily incorporate such information by using one-sided single P values to define the test statistic. Its another advantage is that the power is not primarily driven by more common variants when variants with different allele frequencies present in the target region. Because individual genotypes are not available in pooled sequencing, permutation testing is not an option for accurate significance estimation in scenarios where LD is present. We proposed a Monte Carlo approach by simulating the null distribution of the test statistic based on the estimate covariance between variants. The validity and efficiency of this approach rely on how well the covariance can be estimated. Because of the limit number of pools, the test based on the empirical unbiased covariance estimate did not have a good control of the type I error rate and often led to loss in power. However, the test based on the shrinkage estimate could provide a more satisfactory control on the type I error rate. Yet, it maintained comparable power to the test based on the unknown true covariance. One concern of the proposed approach is that the simulation procedure may lead to significant computational time for large-scale sequencing-based studies. To reduce computational burden, more effective approaches could also be obtained based on the shrinkage estimate of the covariance matrix [35].

The test procedure relies on several assumptions for the different steps of resequencing. The first step of resequencing is typically pulldown of the target genomic region and amplification. We assumed that the targeted genomic regions of subjects in a pool are amplified independently with an equal probability. One concern about this is the presence of heterogeneity in DNA amount in a pool. In this case, individuals are not evenly represented in the pool, and hence the assumption of the resampling approach that alleles of different subjects are drawn with the same probability is not valid. Indeed, the presence of heterogeneity in DNA amount was found to inflate the variance of the test statistic and hence lead to an inflated type I error rate (data not shown). However, if multiple independent markers (≥ 30) are sequenced, it may be possible to use an approach similar to the genomic control to adjust for the inflated variance [36, 37].

In summary, our results suggest that pooled next-generation sequencing with the unbalance blocked design and the appropriate analytic approach could be a valid and cost-effective tool for screening the association of rare variants with diseases. Compared with individual sequencing, it is beneficial in terms of the reduction in cost and time but does not sacrifice much in statistical efficiency.

Acknowledgment

T. Wang was supported in part by the CTSA Grant UL1 RR025750 and KL2 RR025749 and TL1 RR025748 from the National Center for Research Resources (NCRR), a component of the

National Institutes of Health (NIH) and NIH roadmap for Medical Research, R21HG006150 from National Human Genome Research Institute (NHGRI). The codes written in R for the proposed PB test is available by email to Dr. Tao Wang (tao.wang@einstein.yu.edu).

References

[1] J. Shendure and H. Ji, "Next-generation DNA sequencing," *Nature Biotechnology*, vol. 26, no. 10, pp. 1135–1145, 2008.

[2] M. L. Metzker, "Sequencing technologies the next generation," *Nature Reviews Genetics*, vol. 11, no. 1, pp. 31–46, 2010.

[3] W. Bodmer and C. Bonilla, "Common and rare variants in multifactorial susceptibility to common diseases," *Nature Genetics*, vol. 40, no. 6, pp. 695–701, 2008.

[4] D. R. Bentley, "Whole-genome re-sequencing," *Current Opinion in Genetics & Development*, vol. 16, pp. 545–552, 2006.

[5] D. W. Craig, J. V. Pearson, S. Szelinger et al., "Identification of genetic variants using bar-coded multiplexed sequencing," *Nature Methods*, vol. 5, no. 10, pp. 887–893, 2008.

[6] Y. Erlich, K. Chang, A. Gordon et al., "DNA Sudoku - Harnessing high-throughput sequencing for multiplexed specimen analysis," *Genome Research*, vol. 19, no. 7, pp. 1243–1253, 2009.

[7] A. Futschik and C. Schlötterer, "The next generation of molecular markers from massively parallel sequencing of pooled DNA samples," *Genetics*, vol. 186, no. 1, pp. 207–218, 2010.

[8] N. Shental, A. Amir, and O. Zuk, "Identification of rare alleles and their carriers using compressed se(que)nsing," *Nucleic Acids Research*, vol. 38, no. 19, Article ID gkq675, p. e179, 2010.

[9] T. Ito, S. Chiku, E. Inoue et al., "Estimation of haplotype frequencies, linkage-disequilibrium measures, and combination of haplotype copies in each pool by use of pooled DNA data," *The American Journal of Human Genetics*, vol. 72, no. 2, pp. 384–398, 2003.

[10] S. H. Shaw, M. M. Carrasquillo, C. Kashuk, E. G. Puffenberger, and A. Chakravarti, "Allele frequency distributions in pooled DNA samples: applications to mapping complex disease genes," *Genome Research*, vol. 8, no. 2, pp. 111–123, 1998.

[11] D. Zeng and D. Y. Lin, "Estimating Haplotype-disease associations with pooled genotype data," *Genetic Epidemiology*, vol. 28, no. 1, pp. 70–82, 2005.

[12] S. E. Calvo, E. J. Tucker, A. G. Compton et al., "High-throughput, pooled sequencing identifies mutations in NUBPL and FOXRED1 in human complex i deficiency," *Nature Genetics*, vol. 42, no. 10, pp. 851–858, 2010.

[13] T. E. Druley, F. L. M. Vallania, D. J. Wegner et al., "Quantification of rare allelic variants from pooled genomic DNA," *Nature Methods*, vol. 6, no. 4, pp. 263–265, 2009.

[14] S. Nejentsev, N. Walker, D. Riches, M. Egholm, and J. A. Todd, "Rare variants of IFIH1, a gene implicated in antiviral responses, protect against type 1 diabetes," *Science*, vol. 324, no. 5925, pp. 387–389, 2009.

[15] T. Wang, K. Pradhan, L. J. Wong, K. Ye, and T. E. Rohan, "Estimating allele frequency from next-generation sequencing of pooled mitochondrial DNA samples," *Frontiers in Genetics*, vol. 2, article 51, 2011.

[16] S. Y. Kim, Y. Li, Y. Guo et al., "Design of association studies with pooled or un-pooled next-generation sequencing data," *Genetic Epidemiology*, vol. 34, no. 5, pp. 479–491, 2010.

[17] T. Wang, C. Y. Lin, T. E. Rohan, and K. Ye, "Resequencing of pooled DNA for detecting disease associations with rare variants," *Genetic Epidemiology*, vol. 34, no. 5, pp. 492–501, 2010.

[18] J. C. Cohen, R. S. Kiss, A. Pertsemlidis, Y. L. Marcel, R. McPherson, and H. H. Hobbs, "Multiple rare alleles contribute to low plasma levels of HDL cholesterol," *Science*, vol. 305, no. 5685, pp. 869–872, 2004.

[19] S. Morgenthaler and W. G. Thilly, "A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: A cohort allelic sums test (CAST)," *Mutation Research - Fundamental and Molecular Mechanisms of Mutagenesis*, vol. 615, no. 1-2, pp. 28–56, 2007.

[20] B. Li and S. M. Leal, "Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data," *The American Journal of Human Genetics*, vol. 83, no. 3, pp. 311–321, 2008.

[21] B. E. Madsen and S. R. Browning, "A groupwise association test for rare mutations using a weighted sum statistic," *PLoS Genetics*, vol. 5, no. 2, Article ID e1000384, 2009.

[22] A. P. Morris and E. Zeggini, "An evaluation of statistical approaches to rare variant analysis in genetic association studies," *Genetic Epidemiology*, vol. 34, no. 2, pp. 188–193, 2010.

[23] A. L. Price, G. V. Kryukov, P. I. W. de Bakker et al., "Pooled association tests for rare variants in exon-resequencing studies," *The American Journal of Human Genetics*, vol. 86, no. 6, pp. 832–838, 2010.

[24] B. M. Neale, M. A. Rivas, B. F. Voight et al., "Testing for an unusual distribution of rare variants," *PLoS Genetics*, vol. 7, no. 3, Article ID e1001322, 2011.

[25] W. Pan and X. Shen, "Adaptive tests for association analysis of rare variants," *Genetic Epidemiology*, vol. 35, no. 5, pp. 381–388, 2011.

[26] B. Hunter, Ed., *Statistics for Experimenters*, Wiley, 1987.

[27] J. Z. Liu, A. F. McRae, D. R. Nyholt et al., "A versatile gene-based test for genome-wide association studies," *The American Journal of Human Genetics*, vol. 87, no. 1, pp. 139–145, 2010.

[28] J. J. Goeman, S. A. van de Geer, and H. C. van Houwelingen, "Testing against a high dimensional alternative," *Journal of the Royal Statistical Society B*, vol. 68, no. 3, pp. 477–493, 2006.

[29] J. Wessel and N. J. Schork, "Generalized genomic distance-based regression methodology for multilocus association analysis," *The American Journal of Human Genetics*, vol. 79, no. 5, pp. 792–806, 2006.

[30] L. C. Kwee, D. Liu, X. Lin, D. Ghosh, and M. P. Epstein, "A powerful and flexible multilocus association test for quantitative traits," *The American Journal of Human Genetics*, vol. 82, no. 2, pp. 386–397, 2008.

[31] M. C. Wu, P. Kraft, M. P. Epstein et al., "Powerful SNP-set analysis for case-control genome-wide association studies," *The American Journal of Human Genetics*, vol. 86, no. 6, pp. 929–942, 2010.

[32] J. K. Pritchard, "Are rare variants responsible for susceptibility to complex diseases?" *The American Journal of Human Genetics*, vol. 69, no. 1, pp. 124–137, 2001.

[33] S. Greenland, "Methods for epidemiologic analyses of multiple exposures: a review and comparative study of maximum-likelihood, preliminary-testing, and Empirical-Bayes regression," *Statistics in Medicine*, vol. 12, no. 8, pp. 717–736, 1993.

[34] J. Schäfer and K. Strimmer, "A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics," *Statistical Applications in Genetics and Molecular Biology*, vol. 4, article 32, 2005.

[35] K. N. Conneely and M. Boehnke, "So many correlated tests, so little time! Rapid adjustment of P values for multiple correlated tests," *The American Journal of Human Genetics*, vol. 81, pp. 1158–1168, 2007.

[36] B. Devlin and K. Roeder, "Genomic control for association studies," *Biometrics*, vol. 55, no. 4, pp. 997–1004, 1999.

[37] B. Devlin, S. A. Bacanu, and K. Roeder, "Genomic control to the extreme," *Nature Genetics*, vol. 36, no. 11, pp. 1129–1131, 2004, Author reply p. 31.