Published in final edited form as:

Genet Epidemiol. 2010 July ; 34(5): 492–501. doi:10.1002/gepi.20502.

Resequencing of Pooled DNA for Detecting Disease Associations with Rare Variants

Tao Wang, Chang-Yun Lin, Thomas E Rohan, and Kenny Ye

Department of Epidemiology and Population Health, Albert Einstein College of Medicine, Bronx, New York

Abstract

A combination of common and rare variants is thought to contribute to genetic susceptibility to complex diseases. Recently, next-generation sequencers have greatly lowered sequencing costs, providing an opportunity to identify rare disease variants in large genetic epidemiology studies. At present, it is still expensive and time-consuming to resequence large numbers of individual genomes. However, given that next generation sequencing technology can provide accurate estimates of allele frequencies from pooled DNA samples, it is possible to detect associations of rare variants using pooled DNA sequencing. Current statistical approaches to the analysis of associations with rare variants are not designed for use with pooled next-generation sequencing data. Hence, they may not be optimal in terms of both validity and power. Therefore, we propose here a new statistical procedure to analyze the output of pooled sequencing data. The test statistic can be computed rapidly, making it feasible to test the association of a large number of variants with disease. By simulation, we compare this approach to Fisher's exact test based either on pooled or individual genotypic data. Our results demonstrate that the proposed method provides good control of the Type I error rate, while yielding substantially higher power than Fisher's exact test using pooled genotypic data for testing rare variants, and has similar or higher power than that of Fisher's exact test using individual genotypic data. Our results also provide guidelines on how various parameters of the pooled sequencing design affect the efficiency of detecting associations.

Keywords

next-generation sequencing; rare variant; DNA pooling; case-control study; genetic association

Introduction

With the availability of high density SNP arrays, a number of common genetic variants related to risk of human diseases have been successfully identified in recent genome-wide association studies (GWAS). Such studies take an indirect linkage disequilibrium (LD) mapping strategy that uses tagging SNPs as a proxy for the interval containing the causal variant. Through screening a sufficiently large number of SNP markers, GWAS studies can

cover almost completely the common variants of the human genome. However, results of recent GWAS, aiming to discover all or most disease-associated common variants, explain a relatively small proportion of the heritability of most diseases, e.g. Type II diabetes and cancer [2007; Saxena, et al. 2007; Scott, et al. 2007; Sladek, et al. 2007; Steinthorsdottir, et al. 2007; Zeggini, et al. 2007]. One possibility for this may be that a substantial fraction of the heritability of common diseases is contributed by rarer variants that are not detectable by GWAS, partially because of the weak LD between rare causal variants and tag-SNP [Iyengar, et al. 2004; Manolio, et al. 2009].

To capture rare genetic variants of candidate genomic regions involved in common human diseases, an ideal approach is to directly sequence all the samples in a study [Bodmer and Bonilla 2008]. Recent technological advances in high-throughput sequencing, so-called next generation or massively parallel sequencing, has greatly reduced the time and cost of sequencing. At present, three platforms for massively parallel DNA sequencing read production, including the Roche/454 FLX (<http://www.454.com/enabling-technology/the-system.asp>) (Margulies *et al.* 2005), the Illumina/Solexa Genome Analyzer (<http://www.illumina.com/pages.ilmn?ID=203>) [Bentley 2006] and the Applied Biosystems SOLiD™ System (http://marketing.appliedbiosystems.com/images/Product/Solid_Knowledge/flash/102207/solid.html), are widely used. Although it is now feasible to sequence large numbers of individual genomes, as in the 1,000 Genomes Project [Hayden 2008], it is still extremely expensive and time consuming for large genetic epidemiology studies, which usually require inclusion of thousands of samples to achieve reasonable statistical power. However, when such efforts focus on targeted regions, pooled sequencing, i.e. sequence a large number of individuals together on a single sequence run, is an appealing alternative that can dramatically reduce the cost and time [Shaw, et al. 1998]. In such cases, the sequencing throughput required per individual is much less than what is provided by a single run, and hence it is feasible to sequence multiple individuals together. The idea is to sequence pooled DNA samples with cases and controls in separated pools, and to detect associations by comparing the allele frequencies between cases and controls. This strategy was proposed earlier for high throughput SNP arrays [Ito, et al. 2003; Shaw, et al. 1998; Zeng and Lin 2005], but it was not widely accepted as SNP array technology does not provide accurate estimates of the allele frequencies in the pooled samples. Next generation sequencing technology, however, might provide more accurate estimates of allele frequency, as shown by recent studies [Druley, et al. 2009; Nejentsev, et al. 2009]. Recently, one study adopted this strategy using 454 sequencing technology and identified associations of rare variants with insulin-dependent diabetes mellitus (IDDM) [Nejentsev, et al. 2009].

The main objective of this article is to investigate the statistical validity and power of detecting associations with rare variants by sequencing pooled DNA samples, and to investigate the relative importance of different design parameters for such studies. To analyze the data, in addition to the naïve Fisher's exact test, which is based on the estimated allele counts and hence does not consider variation in sequencing results, we developed a statistical method that directly uses the sequencing reads to take into account its variation. More importantly, our method also considers sequencing error (the base calling error), which can be as high as 1% in some current sequencing methods, and by doing so, our

method greatly improves the power of detecting disease-associated rare variants. Our simulation results demonstrate that sequencing pooled DNA samples is a cost-effective and efficient alternative in case-control studies that seek to identify rare variants associated with common human diseases. Compared to the naïve Fisher's exact test, the proposed new statistical test better controls the Type I error rate, and has substantially higher power for detecting disease-associated rare variants. The power of this new test is comparable to or even higher than that of Fisher's exact test with the assumption that the true individual genotype is known.

Method

Consider a case-control study. The phenotype of a subject is denoted by $i = (1,0)$ for cases or controls, respectively. To discover rare variants and test their association with disease, we consider a sequencing procedure by pooling a normalized amount of DNA from multiple subjects with the same phenotype status. In this way, we can estimate and compare allele frequency in DNA pools of cases and controls. We assume that the targeted genomic regions of subjects in a pool are amplified independently and the region of interest is randomly sheared into short fragments. These short fragments are then randomly selected for sequencing, giving sequencing reads that cover a fixed length of segment of each selected fragment. The length of the reads ranges from 30 to 200 base pairs depending on the sequencing platforms. Let n_{ij} be the total number of chromosomes in the j^{th} pool of phenotypic group i , and z_{ij} be the unknown number of rare alleles at a specific locus of interest. After resequencing, a total number of m_{ij} sequencing reads at the loci are observed, and x_{ij} out of m_{ij} reads report the rare variant. We denote the random vector $\mathbf{Z}_i = (Z_{i1}, \dots)$ and $\mathbf{X}_i = (X_{i1}, \dots)$.

We are interested in the question of whether this locus is associated with the disease. Let θ_i be the frequency of the minor allele at this locus for group i . The test of the association can be written as:

$$H_0: \theta_1 = \theta_0 \text{ vs. } H_1: \theta_1 \neq \theta_0 \quad (1)$$

For a small θ_i , i.e. the variant is rare, a commonly mis-used method is Fisher's exact test. To analyze pooled sequencing data, a naïve approach is to apply Fisher's exact test based on the estimated counts of the rare variant for cases and controls [Nejentsev, et al. 2009]. In this approach, the allele frequencies of cases and controls are calculated from the observed numbers of total reads and from the number of reads reporting the variant; and the number

of variants carried by each phenotypic group is estimated by $\hat{z}_i = \hat{\theta}_i n_i = \frac{\sum x_{ij}}{\sum m_{ij}} \times n_i$, where $\hat{\theta}_i$ is the estimated allele frequency for group i . The data are then summarized in a 2×2 contingency table (shown in table 1) and Fisher's exact test is applied to calculate the p-value.

For a two-sided alternative hypothesis, the p-value is obtained by summing the probabilities for all possible 2×2 tables with the same row and column margins that have probabilities less than or equal to that of the observed table.

Although this test is simple to implement, it treats the estimated numbers of the rare variant as if they were observed without considering the uncertainty of such estimates. Thus, it may have an inflated Type I error rate. Secondly, the sampling scheme of Fisher's exact test is based on the hypergeometric distribution, which in principal requires both the column and row marginal totals of a 2×2 table are fixed, i.e. both the sample size (the number of cases and controls) and the number of variants and non-variants are fixed in Table 1. The number of variants and non-variants are usually not fixed in a genetic case-control study. Fisher's exact test used in this way can become very conservative [Upton 1982]. Lastly and most importantly, next-generation sequencing has a relatively high rate of base-calling error and from sequence reads of pooled samples, it is difficult to distinguish true rare variants from such errors. For a rare variant whose frequency is not much higher than the error rate, the power to detect its association with a disease would be very low without adjusting for such error in the statistical method.

To address the above concerns of the naïve approach, we propose a different statistical test. To better illustrate our idea, we first present the method under a situation in which no base-calling error occurs in the sequencing.

Our test statistic is based on the case-control difference in the proportions of the sequencing reads reporting the variant. Without base-calling error, this proportion is an estimate of the allele frequency. For l_i pools used for phenotypic group i , a weighted average estimate

across pools is $\sum_j \frac{w_{ij}}{\sum_j w_{ij}} \frac{X_{ij}}{m_{ij}}$, where the weight w_{ij} is inversely proportional to the variance

of $\frac{X_{ij}}{m_{ij}}$, approximated by $(1/m_{ij} + 1/n_{ij})\theta$ (Appendix 1). Consider the difference in such

estimates between the cases and the controls $\sum_j \frac{w_{1j}}{\sum_j w_{1j}} X_{1j}/m_{1j} - \sum_j \frac{w_{0j}}{\sum_j w_{0j}} X_{0j}/m_{0j}$, whose variance $V_{\hat{\mathbf{x}}}$ can be approximated by

$\frac{1}{\sum_j [(m_{1j}n_{1j})/(m_{1j}+n_{1j})]} \hat{\theta}_1 + \frac{1}{\sum_j [(m_{0j}n_{0j})/(m_{0j}+n_{0j})]} \hat{\theta}_0$ (see Appendix 1). Under the null

hypothesis, let $\hat{\theta} = \sum_i \sum_j \frac{w_{ij}}{\sum_j w_{ij}} \frac{X_{ij}}{m_{ij}}$. This leads to the variance estimate $V_{\hat{\mathbf{x}}}$. We define a test statistic for (1)

$$T_{\mathbf{x}_1, \mathbf{x}_0} = \frac{\sum_j \frac{w_{1j}}{\sum_j w_{1j}} X_{1j}/m_{1j} - \sum_j \frac{w_{0j}}{\sum_j w_{0j}} X_{0j}/m_{0j}}{\sqrt{\hat{V}_{\mathbf{x}}}}. \quad (2)$$

The two-sided p-value for testing (1) is defined by the probability of observing an absolute value of the statistic $T_{\mathbf{x}_1, \mathbf{x}_0}$ that is equal to or larger than the absolute value of the observed $T_{\mathbf{x}_1, \mathbf{x}_0}$ under the null hypothesis, i.e.

$$\begin{aligned}
P(T_{\mathbf{x}_1, \mathbf{x}_0} \geq T_{\mathbf{x}_1, \mathbf{x}_0} | H_0) \\
&= E_{\mathbf{x}_1, \mathbf{x}_2} [I(|T_{\mathbf{x}_1, \mathbf{x}_2}| \geq |T_{\mathbf{x}_1, \mathbf{x}_0}|) | H_0] \\
&= \sum_{x_{11}=0}^{m_{11}} \cdots \sum_{x_{1l_1}=0}^{m_{1l_1}} \sum_{x_{01}=0}^{m_{01}} \cdots \sum_{x_{0l_0}=0}^{m_{0l_0}} Pr(X_{11}, \dots, X_{1l_1} | \theta_1) Pr(X_{01}, \dots, X_{0l_0} | \theta_0) I(|T_{\mathbf{x}_1, \mathbf{x}_2}| \geq |T_{\mathbf{x}_1, \mathbf{x}_0}|)
\end{aligned} \quad (3)$$

where θ_i is the unknown allele frequency of group i , and under the null hypothesis (1), $\theta_1 = \theta_0$. Under the assumption of Hardy-Weinberg equilibrium we assume that the number of the variants in a pool, Z_{ij} , follows a binomial distribution, the rate of occurrence of events being θ_i . Given the number of Z_{ij} , we assume that the number of reads reporting the variant, X_{ij} , follows a binomial distribution, with a success rate of Z_{ij}/n_{ij} . Hence,

$$\begin{aligned}
P(X_{i1}, \dots, X_{il_i} | \theta_i) \\
&= \prod_{j=1}^{l_i} \sum_{Z_{ij}=0}^{n_{ij}} P(X_{ij} | Z_{ij}) P(Z_{ij} | \theta_i) \\
&= \prod_{j=1}^{l_i} \sum_{Z_{ij}=0}^{n_{ij}} \binom{n_{ij}}{x_{ij}} (Z_{ij}/n_{ij})^{x_{ij}} (1 - Z_{ij}/n_{ij})^{(n_{ij}-x_{ij})} \binom{n_{ij}}{Z_{ij}} \theta_i^{Z_{ij}} (1 - \theta_i)^{(n_{ij}-Z_{ij})}
\end{aligned} \quad (4)$$

Substituting (4) in (3), we can see that the p-value in (3) depends on the unknown parameter θ . Following the approaches of Liddell (1978) and Storer and Kim (1990), we can calculate the p-value by using the estimated $\hat{\theta}$ [Liddell 1976; Storer and Kim 1990] That is,

$$\begin{aligned}
P(T_{\mathbf{x}_1, \mathbf{x}_0} \geq T_{\mathbf{x}_1, \mathbf{x}_0} | H_0) \\
&= \sum_{x_{11}=0}^{m_{11}} \cdots \sum_{x_{1l_1}=0}^{m_{1l_1}} \sum_{x_{01}=0}^{m_{01}} \cdots \sum_{x_{0l_0}=0}^{m_{0l_0}} Pr(X_{11}, \dots, X_{1l_1} | \hat{\theta}) Pr(X_{01}, \dots, X_{0l_0} | \hat{\theta}) I(|T_{\mathbf{x}_1, \mathbf{x}_2}| \geq |T_{\mathbf{x}_1, \mathbf{x}_0}|).
\end{aligned} \quad (5)$$

Next, we will consider a test statistic that can adjust for the base-calling error. In sequencing an individual genome, with multiple reads at a locus, one is able to distinguish a base-calling error from a true variant with a sufficiently large number of reads, as the latter is expected to have a probability of 0.5 if the locus is heterozygous. In the pooled DNA samples of a large number of individuals, it is almost impossible to distinguish a true rare variant from the base-calling error because the frequency of the rare variant carried by multiple subjects is likely to be quite close to, and even lower than, the sequencing error rate. The bias introduced by such error in estimating θ is not trivial. Because the reference allele is much more common than the rare variant for a given locus, false positive reads that occur at non-variant bases are likely to occur much more often than false negative reads that occur at variant bases for a rare variant, so that the frequency of a rare allele will tend to be over-estimated in both cases and controls. Although the power of the test statistic can be reduced by either false-positive or false negative misclassification, false positive sequence errors have a greater effect on power as they affect a greater number of alleles [Quade, et al. 1980]. In order to have a more efficient test procedure, it is necessary to take the sequencing error into account.

To quantify the sequencing error rate in pooled sequencing, one approach is to include a control DNA sequence in each pool [Druley, et al. 2009]. In this way, it is possible to obtain the empirical distribution of the sequencing error of individual pools. Let $v, v \in (A, C, G, T)$ be the variant and $r, r \in (A, C, G, T)$ be the reference allele. We define $\tau_{1,ij} = Pr(v|r, i, j)$ to be the false positive error rate, i.e., the probability of reporting a variant given the reference base, and $\tau_{0,ij} = Pr(v|v, i, j)$ to be the true positive rate, i.e., the probability of reporting a variant given the variant base. Both the false positive and true positive rates can be estimated for each pool by the proportion of reads reporting the variant for the reference base and the variant base, respectively. Then, equation (4) may be rewritten as

$$P(X_{i1}, \dots, X_{il_i} | \theta_i, \tau_1, \tau_0) = \prod_{j=1}^{l_i} \sum_{Z_{ij}=0}^{n_{ij}} \binom{n_{ij}}{x_{ij}} \left[\frac{Z_{ij}}{n_{ij}} \tau_{0,ij} + \frac{n_{ij} - Z_{ij}}{n_{ij}} \tau_{1,ij} \right]^{x_{ij}} \left[\frac{Z_{ij}}{n_{ij}} (1 - \tau_{0,ij}) + \frac{n_{ij} - Z_{ij}}{n_{ij}} (1 - \tau_{1,ij}) \right]^{(n_{ij} - x_{ij})} \times \binom{n_{ij}}{Z_{ij}} \theta_i^{Z_{ij}} (1 - \theta_i)^{(n_{ij} - Z_{ij})} \quad (6)$$

Replacing (4) by (6) in (5), we can obtain the p-value after adjusting for the sequencing error.

Direct calculation of (5) may be computationally intensive. We note that a parametric bootstrap (PB) procedure can be used to determine the p value [Krishnamoorthy and Thomson 2004]. In our simulations, we used a simple but fast PB procedure defined by the following five steps:

1. Estimating the sequencing error rate from the control sequence. The number of reads that correctly report the base of the variant $v \in (A, C, G, T)$ in the control sequence is $m_{v|v}$ and the number of reads that mistakenly report a base of the non-

variant $nv \in (A, C, G, T, N)$ is $m_{nv|v}$, in a specific pool, $\hat{\tau}_{0,ij} = \frac{m_{v|v}}{m_{v|v} + m_{nv|v}}$.

Similarly, $\hat{\tau}_{1,ij} = \frac{m_{v|r}}{m_{v|r} + m_{nv|r}}$, in which $m_{v|r}$ and $m_{nv|r}$ are the number of reads that report a variant or a non-variant for a reference base in the control sequence.

2. Estimating the allele frequency $\hat{\theta}$ under the null hypothesis. For large pools and rare variants, we consider a weighted-average estimate across single pools, where the

weight is proportional to $\frac{n_{ij} m_{ij}}{n_{ij} + m_{ij}}$ and the allele frequency of a single pool is

estimated by $\hat{\theta}_{ij} = \sum_{Z_{ij}=0}^{n_{ij}} \pi_{ij} Z_{ij} / n_{ij}$, where

$$\pi_{ij} = \frac{\binom{n_{ij}}{x_{ij}} \left[\frac{Z_{ij}}{n_{ij}} \hat{\tau}_{0,ij} + \frac{n_{ij} - Z_{ij}}{n_{ij}} \hat{\tau}_{1,ij} \right]^{x_{ij}} \left[\frac{Z_{ij}}{n_{ij}} (1 - \hat{\tau}_{0,ij}) + \frac{n_{ij} - Z_{ij}}{n_{ij}} (1 - \hat{\tau}_{1,ij}) \right]^{(n_{ij} - x_{ij})}}{\sum_{Z_{ij}=0}^{n_{ij}} \binom{n_{ij}}{x_{ij}} \left[\frac{Z_{ij}}{n_{ij}} \hat{\tau}_{0,ij} + \frac{n_{ij} - Z_{ij}}{n_{ij}} \hat{\tau}_{1,ij} \right]^{x_{ij}} \left[\frac{Z_{ij}}{n_{ij}} (1 - \hat{\tau}_{0,ij}) + \frac{n_{ij} - Z_{ij}}{n_{ij}} (1 - \hat{\tau}_{1,ij}) \right]^{(n_{ij} - x_{ij})}} \quad (\text{Appendix 2}).$$

3. Calculating the test statistic $T_{\mathbf{x}_1, \mathbf{x}_0} = \frac{\sum_j \frac{w_{1j}}{\sum w_{1j}} X_{1j} / m_{1j} - \sum_j \frac{w_{0j}}{\sum w_{0j}} X_{0j} / m_{0j}}{\sqrt{\hat{\mathbf{V}}_x}}$, in which $w_{ij} = \frac{n_{ij} m_{ij}}{(\hat{\tau}_{0,ij} - \hat{\tau}_{1,ij}) n_{ij} \hat{\theta} + \hat{\tau}_{1,ij} n_{ij} + (\hat{\tau}_{0,ij} - \hat{\tau}_{1,ij})^2 m_{ij} \hat{\theta}}$ and $\hat{\mathbf{V}}_x = \frac{1}{\sum (m_{1j} n_{1j}) / [m_{1j} \hat{\theta} (\tau_{0,1j} - \tau_{1,1j})^2 + n_{1j} (\tau_{0,1j} \hat{\theta} - \tau_{1,1j} \hat{\theta} + \tau_{1,1j})]} + \frac{1}{\sum (m_{0j} n_{0j}) / [m_{0j} \hat{\theta} (\tau_{0,0j} - \tau_{1,0j})^2 + n_{0j} (\tau_{0,0j} \hat{\theta} - \tau_{1,0j} \hat{\theta} + \tau_{1,0j})]}$ (See Appendix 1)
4. Sampling $\mathbf{x}_i^* = (x_{i1}^*, \dots)$ from $Pr(X_{i1}, \dots, X_{it_i} | \hat{\theta}, \hat{\tau}_{1,ij}, \hat{\tau}_{0,ij})$ and calculating the statistic $T_{\mathbf{x}_1^*, \mathbf{x}_0^*}$.
5. Replicating (4) r times and estimating the p-value by the proportion of $|T_{\mathbf{x}_1^*, \mathbf{x}_0^*}| \geq |T_{\mathbf{x}_1, \mathbf{x}_0}|$.

Results

We evaluated the performance of the proposed PB test and the naïve Fisher's exact test (Fisher_n) under various situations by simulation. In addition, we added a comparison with Fisher's exact test using the true individual genotypes (Fisher_g) as a reference. In our simulations, the p-value of the proposed PB test was evaluated on 1,000 bootstrap samples and Fisher's exact test was performed using the function of *fisher.test* implemented in the R package (<http://www.r-project.org/>). 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 by independently sampling each allele from the pool with replacement. The sample size was set at 500 cases and 500 controls; and individuals of a phenotypic group were included in one single pool. This process was repeated for 5,000 replicates. The Type I error rate and power were evaluated by the proportion of replicates having a p-value ≤ 0.05 .

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. We considered different allele frequencies ranging from 0.5% (rare) to 20% (common), and sequence coverage ranging from low (2×) to high (30×). In our simulation, coverage of individual pools was allowed to vary following the density of the Gamma distribution with shape parameters $\alpha=6.3$ and $\beta=\text{average coverage}/\alpha$ [Sarin, et al. 2008]. To allow for errors in sequencing, we considered a wide range of the base calling error rates from 0.5%–2% and stipulated that they be consistent across pools.

To evaluate the power of different tests in detecting associations with rare variants, we fixed the allele frequency in controls at 0.5% and 1%, but allowed the allele frequency in cases to vary in order to yield different effect sizes. We considered the influence of three parameters: the sequencing coverage, the base calling error rate, as well as the precision of the estimate of the sequencing error rate, which is determined by the number of bases included in the control sequence as well as coverage per base.

Type I error rate

The empirical Type I error rate at a significance level of 0.05 is shown in Figure 1 and 2. Figure 1 indicates that the Type I error rate of the proposed PB test was consistent under various allele frequencies, coverage, and sequencing error rates, and maintained good control of the nominal significance level consistently in all simulated scenarios. However, both Fisher's exact tests, as expected, tended to become over-conservative when the allele frequency was low. The test Fisher_n was also influenced by the sequencing coverage as it tended to inflate the Type I error when the sequence coverage was low. However, this was a less serious problem when testing rare variants ($\theta < 0.01$), because this effect was offset by the intrinsic conservativeness of Fisher's exact test. Moreover, the Type I error rate of Fisher_n was severely impacted by sequencing errors. It became very conservative as the sequencing error rate increased.

Figure 2 shows the Type I error rate of the proposed PB test as a function of the number of bases (variant and reference) included in the control sequence for estimating the sequencing error rate. The simulation results indicated that the estimate of the sequencing error rate did not inflate the Type I error rate of the proposed test. However, when the number of bases included in the control sequence was low (e.g., 100), the proposed PB test became slightly conservative in detecting associations of rare variants ($\theta=0.005$ or 0.01). Nevertheless, the PB test was consistently closer to the nominal significance level than the naïve Fisher's exact test. The control sequence had no impact on either of the Fisher's exact tests because they do not use any information from the control sequence.

Power

Figures 3 shows the empirical power of the three tests under different sequencing error rates and coverage. Under ideal situations, without any sequencing error, the two tests based on the pooled sequencing output (Fisher_n and PB) have similar power based on individual genotypic data (Fisher_g). Interestingly, the proposed PB test had even slightly higher power than Fisher_g, in particular, when allele frequencies were low in both cases and controls and coverage was high (30×). This reflects the tendency of Fisher's exact test to be conservative. With increasing sequencing error rate, Fisher_n lost significant power, while the PB test maintained comparable power to Fisher_g. Comparing the power of the PB test under different sequencing coverage scenarios (across panels), we can see that low sequencing coverage (5×) slightly reduced the power, in particular when the sequencing error rate was high (2%), suggesting that the reduction in power may be due to the higher variation in the allele frequency estimated from low coverage data.

Figure 4 shows the influence of variation in the estimate of the sequencing error rate on the empirical power of the PB test. Under the ideal situation of no error in the sequencing procedure, the number of variant and reference bases included in the control sequence did not affect the power of the PB test. However, it was necessary to have enough bases in the control sequence to maintain the optimal power of the PB test in the presence of high levels of sequencing error. When a small number of bases (e.g. 100) were included in the control sequence for estimating the error rates, there was a slight loss of power of the PB test.

We also simulated two pools for each phenotypic group, which revealed similar results to those described above (data not shown).

Discussion

In this paper we addressed the question of how to test the association of rare variants with phenotype using pooled next-generation sequencing data, as the naïve Fisher's exact test does not always perform well. As shown in our simulation, this test is over-conservative when the minor allele frequency is low, but more importantly, it inflates Type I error when the sequence coverage is low. There is a trade-off between sequencing coverage versus the size of targeted regions and the number of individuals within a pool. For a fixed budget, to use this test one must reduce targeted regions or limit the number of individuals within a pool to achieve the coverage required to maintain the Type I error rate due to cost-throughput constraints [Smith, et al. 2008]. To maintain good control of the Type I error rate at low sequence coverage while retaining satisfactory power, we proposed a new test that is based on modeling variation in the reads of the pooled sequencing procedure. 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 and the region of interest is randomly sheared into short fragments, which are then sequenced. We investigated the performance of the proposed testing procedure by simulation, which showed that this procedure can maintain good control of the Type I error rate even for low coverage data. Moreover, it can achieve much higher power than the naïve Fisher's exact test and have similar or even higher power than Fisher's exact test based on the true individual genotypic data for testing rare variants. Even though the computation of the exact p-values of this test statistic can be intensive, we showed that p-values can be approximately estimated by a parametric bootstrap approach in a relatively easy and fast manner, which makes it feasible to screen a large number of variants from the sequencing data.

A major obstacle to the use of pooled second-generation sequencing in searching for rare disease variants is the relatively high error rates. As showed in our simulation, the power of the naïve Fisher's exact test can be severely impaired by sequencing errors. It has been reported that the sequencing error rate per read using the Illumina sequencer can vary over a wide range from <0.05% to 5%, depending on the position in the sequence read (cycle number), and that sequencing errors are present at a substantially lower frequency in the first 10 reads than in later reads [Druley, et al. 2009]. Hence, it is possible to alleviate the impact of sequencing errors by using fewer reads that are more accurate, but it therefore sacrifices sequencing coverage. Nevertheless, as only a small number of rare variants are expected to be carried by pooled samples, the power for testing rare variants can still be largely influenced by even a low level of sequencing error rate because of the relatively high noise/signal ratio. Our simulation showed that the power of pooled sequencing can be further improved by adding a control sequence. To simplify our simulations, we ignored variation in the error rates across cycle numbers as such variation is likely to be small when only the most accurate cycles are used. However, the proposed procedure can easily be extended to take this variation into account - the error rates and the allele frequency can be estimated

separately for specific cycle numbers and $\hat{\theta}$ can then be calculated similarly by averaging across cycles.

The proposed approach is based on the idea of comparing the allele frequency of a single locus between cases and controls. As a single rare variant is likely to have a low marginal effect on disease risk, particularly in the presence of genetic heterogeneity, it may also be beneficial to jointly test a group of rare variants according to function, such as genes or pathways. Our approach is also applicable to detection of a group of rare variants by testing if they are more enriched in one phenotype group, e.g., if there is any difference of the average (or weighted average) of the allele frequencies of a group of variants between phenotypic groups [Li and Leal 2008; Madsen and Browning 2009]. However, such an approach is based on the assumption that the effects of multiple variants have the same direction, which may not always be true. For testing the joint association of a group of rare variants that may have opposing effects, most of the current statistical methods, such as [Hotelling's test](#) [Xiong, et al. 2002], need individual genotypic information. It is possible to extract individual genotypic information from the pooled sequencing by using a different experimental design or technology, such as the overlapping pool design [Prabhu and Pe'er 2009] or Barcoding [Craig, et al. 2008; Cronn, et al. 2008]. However, the above two approaches involve more complex assays or designs and hence may be difficult to apply in large genetic epidemiology studies. Currently, we are working on the development of a statistical multiple-locus method designed for analyzing pooled next-generation sequence data.

In summary, our results raise the possibility of using pooled next-generation sequencing as a cost-effective tool for screening the association of rare variants with disease. Compared with individual sequencing, it is beneficial in terms of the reduction in cost and time, but has similar statistical efficiency.

Acknowledgments

T.W. 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, and P60 DK020541.

Appendices

1: The derivation of T_{x_1, x_0}

We assume that Z_{ij} and X_{ij} follow a binomial distribution, respectively. In using large pools to detect rare variants, they may be approximated by Poisson distribution. Then,

$$\begin{aligned} \text{Var}(X_{ij}) &= \text{Var}[E(X_{ij}|Z_{ij})] + E[\text{Var}(X_{ij}|Z_{ij})] \\ &= \text{Var}(Z_{ij}m_{ij}/n_{ij}) + E(Z_{ij}m_{ij}/n_{ij}) \\ &= (m_{ij} + m_{ij}^2/n_{ij})\theta_i \text{Var}(X_{ij}/m_{ij}) \\ &= (1/m_{ij} \\ &\quad + 1/n_{ij})\theta_i. \end{aligned}$$

For multiple pools of a phenotypic group, let $w_{ij} = \left(\frac{m_{ij} + n_{ij}}{m_{ij} n_{ij}} \right) \theta_i$.

Because X_{ij} are independent across pools,

$$\text{Var}\left(\sum \frac{w_{ij}}{\sum w_{ij}} X_{ij} / m_{ij}\right) = \sum \text{Var}\left(\frac{w_{ij}}{\sum w_{ij}} X_{ij} / m_{ij}\right) = \sum \frac{w_{ij}^2}{(\sum w_{ij})^2} \text{Var}(X_{ij} / m_{ij}) = \frac{\theta_i}{\sum [(m_{ij} n_{ij}) / (m_{ij} + n_{ij})]}$$

and

$$\text{Var}\left[\sum_{j=1}^{l_1} w_{1j} X_{1j} / m_{1j} - \sum_{j=1}^{l_0} w_{0j} X_{0j} / m_{0j}\right] = \frac{1}{\sum [(m_{1j} n_{1j}) / (m_{1j} + n_{1j})]} \theta_1 + \frac{1}{\sum [(m_{0j} n_{0j}) / (m_{0j} + n_{0j})]} \theta_0.$$

Similarly, in the presence of sequencing errors,

$$\text{Var}(X_{ij} / m_{ij}) = [(\tau_{0,ij} - \tau_{1,ij}) \theta_i + \tau_{1,ij}] / m_{ij} + (\tau_{0,ij} - \tau_{1,ij})^2 \theta_i / n_{ij}$$

$$\begin{aligned} \text{Let } w_{ij} &= \frac{n_{ij} m_{ij}}{m_{ij} (\tau_{0,ij} - \tau_{1,ij})^2 \theta_i + n_{ij} (\tau_{0,ij} \theta_i - \tau_{1,ij} \theta_i + \tau_{1,ij})} \\ \text{Var}\left[\sum_j \frac{w_{1j}}{\sum w_{1j}} X_{1j} / m_{1j} - \sum_j \frac{w_{0j}}{\sum w_{0j}} X_{0j} / m_{0j}\right] &= \frac{1}{\sum [(m_{1j} n_{1j}) / (m_{1j} \theta_1 (\tau_{0,1j} - \tau_{1,1j})^2 + n_{1j} (\tau_{0,1j} \theta_1 - \tau_{1,1j} \theta_1 + \tau_{1,1j}))]} \\ &+ \frac{1}{\sum [(m_{0j} n_{0j}) / (m_{0j} \theta_0 (\tau_{0,0j} - \tau_{1,0j})^2 + n_{0j} (\tau_{0,0j} \theta_0 + \tau_{1,0j} \theta_0 + \tau_{1,0j}))]} \end{aligned}$$

With the estimated $\hat{\theta}$, $\hat{\tau}_{1,ij}$ and $\hat{\tau}_{0,ij}$ under the null hypothesis, the test statistic is

$$\mathbf{T}_{\mathbf{x}_1, \mathbf{x}_2} = \frac{\sum \frac{w_{1j}}{\sum w_{1j}} X_{1j} / m_{1j} - \sum \frac{w_{0j}}{\sum w_{0j}} X_{0j} / m_{0j}}{\sqrt{\hat{V}_x}}.$$

2: The estimate of the allele frequency

If Z_{ij} is known, the allele frequency for each pool can be estimated by the maximum

likelihood method, i.e., $\hat{\theta}_{ij} = \frac{Z_{ij}}{n_{ij}}$. Because Z_{ij} is missing, we use the mean of Z_{ij}/n_{ij} given X_{ij} to estimate the allele frequency, i.e., $\hat{\theta}_{ij} = E_{Z_{ij}|X_{ij}}[Z_{ij}/n_{ij}]$. By using an uninformative prior on Z_{ij} , the probability of Z_{ij} given X_{ij} is

$$\pi_{ij} = Pr \left(Z_{ij} \middle| X_{ij} \right) = \frac{\binom{m_{ij}}{x_{ij}} \left[\frac{Z_{ij}}{n_{ij}} \hat{\tau}_{0,ij} + \frac{n_{ij} - Z_{ij}}{n_{ij}} \hat{\tau}_{1,ij} \right]^{x_{ij}} \left[\frac{Z_{ij}}{n_{ij}} (1 - \hat{\tau}_{0,ij}) + \frac{n_{ij} - Z_{ij}}{n_{ij}} (1 - \hat{\tau}_{1,ij}) \right]^{(m_{ij} - x_{ij})}}{\sum_{Z_{ij}=0}^{n_{ij}} \binom{m_{ij}}{x_{ij}} \left[\frac{Z_{ij}}{n_{ij}} \hat{\tau}_{0,ij} + \frac{n_{ij} - Z_{ij}}{n_{ij}} \hat{\tau}_{1,ij} \right]^{x_{ij}} \left[\frac{Z_{ij}}{n_{ij}} (1 - \hat{\tau}_{0,ij}) + \frac{n_{ij} - Z_{ij}}{n_{ij}} (1 - \hat{\tau}_{1,ij}) \right]^{(m_{ij} - x_{ij})}}$$

Therefore,

$$\hat{\theta}_{ij} = \sum_{Z_{ij}=0}^{n_{ij}} \pi_{ij} [Z_{ij} / n_{ij}].$$

Reference

1. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*. 2007; 447(7145):661–678. [PubMed: 17554300]
2. Bentley DR. Whole-genome re-sequencing. *Curr Opin Genet Dev*. 2006; 16(6):545–552. [PubMed: 17055251]
3. Bodmer W, Bonilla C. Common and rare variants in multifactorial susceptibility to common diseases. *Nature Genetics*. 2008; 40(6):695–701. [PubMed: 18509313]
4. Craig DW, Pearson JV, Szelinger S, Sekar A, Redman M, Corneveaux JJ, Pawlowski TL, Laub T, Nunn G, Stephan DA, et al. Identification of genetic variants using bar-coded multiplexed sequencing. *Nature Methods*. 2008; 5(10):887–893. [PubMed: 18794863]
5. Cronn R, Liston A, Parks M, Gernandt DS, Shen R, Mockler T. Multiplex sequencing of plant chloroplast genomes using Solexa sequencing-by-synthesis technology. *Nucleic Acids Research*. 2008; 36(19)
6. Druley TE, Vallania FLM, Wegner DJ, Varley KE, Knowles OL, Bonds JA, Robison SW, Doniger SW, Hamvas A, Cole FS, et al. Quantification of rare allelic variants from pooled genomic DNA. *Nature Methods*. 2009; 6(4):263–265. [PubMed: 19252504]
7. Hayden EC. International genome project launched. *Nature*. 2008; 451(7177):378–379. [PubMed: 18216809]
8. Ito T, Chiku S, Inoue E, Tomita M, Morisaki T, Morisaki H, Kamatani N. Estimation of haplotype frequencies, linkage-disequilibrium measures, and combination of haplotype copies in each pool by use of pooled DNA data. *American Journal of Human Genetics*. 2003; 72(2):384–398. [PubMed: 12533787]
9. Iyengar SK, Song DH, Klein BEK, Klein R, Schick JH, Humphrey J, Millard C, Liptak R, Russo K, Jun G, et al. Dissection of genomewide-scan data in extended families reveals a major locus and oligogenic susceptibility for age-related macular degeneration. *American Journal of Human Genetics*. 2004; 74(1):20–39. [PubMed: 14691731]
10. Krishnamoorthy K, Thomson J. A more powerful test for comparing two Poisson means. *Journal of Statistical Planning and Inference*. 2004; 119(1):23–35.
11. Li BS, Leal SM. Methods for detecting associations with rare variants for common diseases: Application to analysis of sequence data. *American Journal of Human Genetics*. 2008; 83(3):311–321. [PubMed: 18691683]
12. Liddell D. Practical tests of 2×2 contingency tables. *The Statistician*. 1976; 25:295–304.
13. Madsen BE, Browning SR. A Groupwise Association Test for Rare Mutations Using a Weighted Sum Statistic. *Plos Genetics*. 2009; 5(2)
14. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, et al. Finding the missing heritability of complex diseases. *Nature*. 2009; 461(7265):747–753. [PubMed: 19812666]

15. Nejentsev S, Walker N, Riches D, Egholm M, Todd JA. Rare variants of IFIH1, a gene implicated in antiviral responses, protect against type 1 diabetes. *Science*. 2009; 324(5925):387–389. [PubMed: 19264985]
16. Prabhu S, Pe'er I. Overlapping pools for high-throughput targeted resequencing. *Genome Research*. 2009; 19(7):1254–1261. [PubMed: 19447964]
17. Quade D, Lachenbruch PA, Whaley FS, McClish DK, Haley RW. Effects of misclassifications on statistical inferences in epidemiology. *Am J Epidemiol*. 1980; 111(5):503–515. [PubMed: 7377195]
18. Sarin S, Prabhu S, O'Meara MM, Pe'er I, Hobert O. *Caenorhabditis elegans* mutant allele identification by whole-genome sequencing. *Nature Methods*. 2008; 5(10):865–867. [PubMed: 18677319]
19. Saxena R, Voight BF, Lyssenko V, Burtt NP, de Bakker PI, Chen H, Roix JJ, Kathiresan S, Hirschhorn JN, Daly MJ, et al. Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science*. 2007; 316(5829):1331–1336. [PubMed: 17463246]
20. Scott LJ, Mohlke KL, Bonnycastle LL, Willer CJ, Li Y, Duren WL, Erdos MR, Stringham HM, Chines PS, Jackson AU, et al. A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. *Science*. 2007; 316(5829):1341–1345. [PubMed: 17463248]
21. Shaw SH, Carrasquillo MM, Kashuk C, Puffenberger EG, Chakravarti A. Allele frequency distributions in pooled DNA samples: applications to mapping complex disease genes. *Genome Research*. 1998; 8(2):111–123. [PubMed: 9477339]
22. Sladek R, Rocheleau G, Rung J, Dina C, Shen L, Serre D, Boutin P, Vincent D, Belisle A, Hadjadj S, et al. A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature*. 2007; 445(7130):881–885. [PubMed: 17293876]
23. Smith DR, Quinlan AR, Peckham HE, Makowsky K, Tao W, Woolf B, Shen L, Donahue WF, Tusneem N, Stromberg MP, et al. Rapid whole-genome mutational profiling using next-generation sequencing technologies. *Genome Research*. 2008; 18(10):1638–1642. [PubMed: 18775913]
24. Steinthorsdottir V, Thorleifsson G, Reynisdottir I, Benediktsson R, Jonsdottir T, Walters GB, Styrkarsdottir U, Gretarsdottir S, Emilsson V, Ghosh S, et al. A variant in CDKAL1 influences insulin response and risk of type 2 diabetes. *Nature Genetics*. 2007; 39(6):770–775. [PubMed: 17460697]
25. Storer BE, Kim C. Exact Properties of Some Exact Test Statistics for Comparing Two Binomial Proportions. *Journal of the American Statistical Association*. 1990; 85(409):146–155.
26. Upton GJG. A comparison of tests for the 2×2 comparative trial. *Journal of the Royal Statistical Society. Series A (General)*. 1982; 145:86–105.
27. Xiong M, Zhao J, Boerwinkle E. Generalized T2 test for genome association studies. *American Journal of Human Genetics*. 2002; 70(5):1257–1268. [PubMed: 11923914]
28. Zeggini E, Weedon MN, Lindgren CM, Frayling TM, Elliott KS, Lango H, Timpson NJ, Perry JR, Rayner NW, Freathy RM, et al. Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. *Science*. 2007; 316(5829):1336–1341. [PubMed: 17463249]
29. Zeng D, Lin DY. Estimating haplotype-disease associations with pooled genotype data. *Genet Epidemiol*. 2005; 28(1):70–82. [PubMed: 15558554]

Figure 1.

Type I error rates at the α level of 5% for the proposed test and Fisher's exact tests as a function of the allele frequency (θ), coverage, and the error rate for testing association. Sample size was set at 500 cases and 500 controls. The allele frequencies were set at 0.5%, 1%, 5% and 20%, the coverage ranged from 2 to 30 fold, and both the false positive and false negative error rates were set at 0%, 0.5%, 1% and 2%. A control sequence with 1,000 variant and 1,000 reference bases was used to estimate the error rates.

Figure 2.

Type I error rate of tests as a function of the number of bases (variant and reference) in the control sequence for estimating the sequencing error rate. Sample size was set at 500 cases and 500 controls. The allele frequencies (θ) were set at 0.5%, 1%, 5% and 20%, the coverage was fixed at 30 fold, and both the false positive and false negative error rates were set at 0%, 0.5%, 1% and 2%.

Figure 3.

Efficiency of pooled sequencing in detecting the association of rare variants compared to individual genotyping under various sequencing coverage and sequencing error rates. The sample size was 500 cases and 500 controls. The sequencing error rate (false positive and

false negative rates) was set at 0%, 0.5%, 1% and 2%, the coverage was at 5×, 15× and 30× per base, and 1,000 variant or reference bases were simulated in the control sequence. The allele frequency of controls was set at 0.5% (A) and 1% (B). The allele frequency of cases (θ) varied, as shown on the x-axis.

Figure 4.

The influence of the number of bases included in the control sequence on the power of the PB test. The sample size was 500 cases and 500 controls. The allele frequency of controls (θ_0) was set at 0.5% and 1% and the allele frequency of cases (θ_1) varied, as shown on the x-axis. The average coverage per base of a pool was set at 30 fold. The error rates (false positive and false negative rates) were set at 0%, 0.5%, 1% and 2%. The same number of variant and reference bases (100, 1,000 and 10,000) was simulated in the control sequence.

Table 1

Contingency table of the estimated numbers of the variant for case-control sample

	Cases	Controls	Total
Variant	$\hat{z}_1$	$\hat{z}_0$	$\hat{z}$
Non-variant	$n_1 - \hat{z}_1$	$n_0 - \hat{z}_0$	$n - \hat{z}$
Total	n_1	n_0	n