Evaluation of Methods to Analyze SNP-based Association Studies in a DNA-Pooling Experiment with Preferential Amplification

Chul Ahn[1] **Kyusang Lee**[2]
[1]*Department of Medicine, University of Texas Medical School, Houston, TX, USA*
[2]*Samsung Advanced Institute of Technology, Giheung, Gyunggi-Do, Korea*
Email : chul.w.ahn@uth.tmc.edu, kyusang.lee@samsung.com

ABSTRACT: Genetic association case-control studies using DNA pools are efficient ways of detecting association between a marker allele and disease status. DNA pooling is an efficient screening method for locating susceptibility genes associated with the disease. However, DNA pooling is efficient only when allele frequency estimation is done precisely and accurately. Through the evaluation of empirical type I errors and empirical powers by simulation, we will evaluate the methods that correct for preferential amplification of nucleotides when estimating the allele frequency of single-nucleotide polymorphisms.

1 INTRODUCION

Complex diseases are likely to be influenced by the factors such as genetic heterogeneity, phenocopies, incomplete penetrance, interactions between genotype and environment, and multilocus effects. Unlike diseases with simple Mendelian inheritance such as cystic fibrosis, the application of linkage analysis to complex diseases has been much less successful.

Genetic association studies using a set of SNPs that covers the human genome densely would be very expensive, and beyond the reach of most laboratories even though the cost of large-scale single-nucleotide polymorphisms (SNP) determination dropped dramatically in recent years. As a result, the development of innovative study designs that reduce the cost is warranted.

Haplotype-tagging SNPs and DNA pooling have potential to reduce the cost barriers of the genetic association study. Botstein and Risch [1] provide an excellent review for the comparison of genome-wide haplotype map-based versus sequenced-based strategies. The use of haplotype-tagging SNPs is a map-based approach while DNA pooling is a sequenced-based approach. A comprehensive map-based approach may require genotyping 10-fold more SNPs than a sequence-based approach (500,000-1,000,000 for map-based versus 50,000-100,000 for sequenced-based). Botstein and Risch [1] suggest a sequence-based approach for the initial stage of a major program aimed at genome-wide association studies. Bansal *et al.* [2] demonstrate the potential of a sequence-based DNA pooling techniques and their associated technologies as an initial screen in the search for genetic association. Ahn *et al.* [3] and Sham *et al.* [4] give a review of DNA pooling.

A common observation in DNA pooling is that the two alleles at a polymorphic SNP locus are not amplified in equal amounts in heterozygous individuals. The aim of this paper is to investigate the impact of preferential amplification in DNA pooling. Simulation study is conducted to evaluate the methods that correct for preferential amplification of nucleotides when estimating the allele frequency of single-nucleotide polymorphisms.

2 METHODS

The population allele frequency can be estimated directly by counting the number of alleles from representative samples in individual genotyping experiments. However, direct counting of the number of alleles cannot be done in DNA-pooling experiments since only the peak intensities are measured.

First, in DNA pooling, heterozygous individuals are collected to estimate the coefficient of preferential amplification. The two peak intensities (A and B) corresponding to two polymorphic alleles at the SNP locus are determined for each heterozygous individual. Hoogendoorn *et al.* [5] estimates the preferential amplification as the arithmetic mean of ratios of two peak intensities using the individual genotyping data from heterozygous individuals. This approach has been used by many investigators [6-10].

Since the ratio of two peak intensities often show a skewed distribution, we can use the log-transformation to reduce skewness and variance, and then take the arithmetic mean of log-transformation. Yang *et al.* [11] estimates the bias of the preferential amplification factor of Hoogendoorn et al. [5], and estimate the sample size for the first stage using the sample size formula widely used in the clinical trials literature.

Second, DNA from cases and controls are pooled to estimate allele frequencies separately from cases and controls. The same genotyping principle is used as in the first stage. Peak intensities will be obtained to estimate allele frequencies from DNA pools. From DNA pools, we obtain two allele-specific estimates, H_A and H_B corresponding to the quantitative estimates of the products representing alleles A and B. From the pool, we obtain the estimated frequency of allele A as $p_A = H_A / (H_A + k H_B)$, where k is the estimated coefficient of preferential amplification obtained from the first stage. The main aim of a DNA pooling study [2-4] is to screen potential SNP markers associated with a disease locus. To achieve the aim, the chi-square test is conducted.

$$\chi^2 = (\hat{p}_A^{Case} - \hat{p}_A^{Control})^2 \Big/ V(\hat{p}_A^{Case} - \hat{p}_A^{Control}),$$

where the variance of the estimated allele frequency is given as

$$V(\hat{p}_A) = p_A(1-p_A)/(2N) + V(e)$$
$$+ p_A{}^2(1-p_A)^2 CV(k),$$

and CV(k)=SE(k)/k, the coefficient of variation of k.

The above variance formula shows three potential sources of error. One is due to sampling a finite number of individuals from a population, another is due to estimating the preferential amplification coefficient (k), and the other is due to a pool specific measurement error. The above statistic follows a chi-square distribution with 1 degree of freedom asymptotically. Chi-square tests are conducted to identify markers for individual genotyping.

Third, individual genotyping will be done for the selected markers with the large allele-frequency differences at the second stage. Once the individual genotyping is done, markers significantly associated with the disease are identified through statistical or mathematical models. Even though complex diseases are generally caused by multiple genetic variations, most available association methods are based on the assumption that a single genetic variation is primarily responsible for the disease under study. Only a few approaches consider interactions of multiple genes and environmental factors in identifying susceptibility loci for complex disease [12-17].

Hoh *et al.* [12] develops a novel test procedure, called a set association approach, to identify genetic variation responsible for complex diseases when multiple genes are involved. This approach is appropriate for many study designs, such as case-control, trio and extended families. The method uses a score statistic that is weighted by the allele contribution to a Hardy-Weinberg equilibrium measurement. All alleles are jointly estimated and the minimal p-value identifies the combination of alleles, across genes that appear to act in concert to alter the risk of disease. They applied the set association approach to a real restenosis data set and could identify several SNPs of interest that were in linkage disequilibrium with susceptibility locus for restenosis, and the re-blockage of the coronary after treatment. Zee *et al.* [13] uses this approach to define a panel of contributory genes in instant restenosis. Hoh *et al.* [11] did not evaluate the empirical the type I errors and empirical powers of the set association approach. Hao *et al.* [14] systematically evaluates the performances of multiple SNP association test in terms of power and accuracy in capturing the real disease SNPs. Hao *et al.* [2004] shows that the inclusion of Hardy-Weinberg Disequilibrium (HWD) reduces the power through simulation. They demonstrate that the test procedure could capture the SNPs associated with disease fairly successfully.

DNA pooling can be used as an efficient and sensitive method of screening numerous markers to identify a subset of markers for more detailed studies. Botstein and Risch [1] suggests a sequence-based approach such as DNA pooling for the initial stage of a major program aimed at genome-wide association studies. Bansal *et al.* [2] demonstrates the potential of a sequence-based DNA pooling techniques and their associated technologies as an initial screen in the search for genetic association.

3 SIMULATION

Simulation studies are conducted to estimate preferential amplification factors and assess the performance of preferential amplification factors. We also assess the impact of preferential amplification factors on pooling-based association test.

In the first stage, we set the number of heterozygous individuals to 10 and 20, and set the true coefficient of preferential amplification (k) to 0.5 and 1. Bivariate normal distribution was used to generate the peak intensities with the correlation of the two peak intensities as 0.5, and the coefficient of variation taking values of 0.1 and 0.3. The coefficient of preferential amplification was estimated using the average of the ratios of two peak intensities (k_R), using the geometric mean of ratios (k_G), and using the bias-corrected adjustments (k_Y)of Yang et al. (2005).
In the second stage, the numbers of cases and controls are fixed as 250 subjects. The population frequencies for cases and controls are set as ($p_{Control}$=0.35, p_{Case}=0.5) and ($p_{Control}$=0.35, p_{Case}=0.35). The sample frequencies for controls and cases are generated from normal distributions N($p_{Control}$,.V($p_{Control}|k$)) and N(p_{Case} ,.V($p_{Case}|k$)).

For each set of parameter combination, 10,000 simulations are conducted.

4 RESULTS

We examine the impact of coefficient of preferential amplification on empirical type I errors and empirical powers. Table 1 shows the empirical type I errors.

σ_e	k	*Unadjusted*	k_R	k_G	k_Y
0.01	0.5	0.061	0.045	0.051	0.048
	1	0.049	0.046	0.050	0.045
0.03	0.5	0.122	0.043	0.048	0.046
	1	0.052	0.045	0.046	0.043

Table 1. Empirical Type I errors for 250 cases and 250 controls with 5% significance levels.

Table 1 shows the empirical type I errors for different preferential amplification methods. When the adjustments are made for preferential amplification, the empirical type I errors are close to the nominal significance levels. The adjustment based on geometric mean and bias-correction yields slightly closer to nominal levels than the adjustment based on arithmetic mean of ratios. When there is no adjustment for preferential amplification, the empirical type I errors are larger than the nominal significance level. So, the empirical powers are not shown for empirical powers for unadjusted preferential amplification.

σ_e	k	k_R	k_G	k_Y
0.01	0.5	0.854	0.872	0.865
	1	0.848	0.861	0.830
0.03	0.5	0.725	0.746	0.735
	1	0.728	0.730	0.724

Table 2. Empirical powers for 250 cases and 250 controls.

Table 2 shows very similar empirical powers for three adjusted approaches. Three adjusted approaches for preferential amplification yield very similar empirical type I errors and powers for various parameter combinations. Simulation results of empirical type I errors show that it is essential to make adjustments for preferential amplification. The simulation results show that the adjustment of the preferential amplification yields a more accurate and reliable results. DNA-pooling-based association studies conducted without any consideration of SNP-specific correction factors for preferential amplification result in unacceptable rate.

In addition, we conducted the simulation study using the replication of pools. We set the number of replicated pools as 4 and 8. For the replicated pools, we estimate the intraclass correlation coefficient (ICC) from a nested design of population samples and replicated pools within samples using Analysis of Variance (ANOVA) technique. Then, we inflate the variance using ICC. The results for this are not shown here.

5 DISCUSSION

Allele frequency estimation through DNA pooling provides a fast, cheap, reliable way of testing individual genetic markers for association of complex disease if experiment is performed carefully [2-4]. However, preferential amplification needs to be incorporated into analysis to avid a severe bias in allele frequency estimation and consequent reduction in the power of the association study since preferential amplification of nucleotides frequently occurs in DNA pooling studies.

Simulation results show that the empirical type I errors are close to the nominal significance levels with the adjustments for preferential amplification. The adjustment based on geometric mean and bias-correction yields slightly closer to nominal levels than the adjustment based on arithmetic mean of ratios. The empirical type I errors are larger than the nominal significance level f no adjustment is made for preferential amplification. Three adjusted approaches for preferential amplification yield very similar empirical type I errors and powers for various parameter combinations. Simulation results of empirical type I errors show that it is essential to make adjustments for preferential amplification. Robustness of these adjustments needs to be studied further for different parameter distributions and different parameter combinations. The simulation results show that the adjustment of the preferential amplification yields a more accurate and reliable results. DNA pooling technique is ideal for screening a large number of markers for associations although positive results will require confirmation through individual genotyping. Considerable savings can be achieved concerning DNA, cost and labor through the use of DNA pooling.

REFERENCES

[1] Botstein D, Risch N. Discovering genotypes underlying human phenotypes: past successes for Mendelian disease, future approaches for complex disease. *Nature Genetics* 33:Suppl:228-237,, 2003.

[2] Bansal A, Boom D, Krammerer S, Honisch C, Adam G, Cantor C, Kleyn P, Braun A. Association testing by DNA pooling: An effective initial screen. *Proc. Natl. Acad. Sci. USA* 99:16871-16874, 2002.

[3] Ahn C, King T, Lee K, Kang S. DNA pooling as a tool for case-control association studies of complex traits. *Genomics & Informatics* 3(1): 1-7, 2005.

[4] Sham P, Bader JS, Craig I, O'Donovan M, Owen M. DNA pooling: a tool for large scale association studies. *Nature Reviews. Genetics* 3: 862-871, 2002.

[5] Hoogendoorn B, Norton N, Jirov G, Williams N, Hamshere N, Spurlock G, Austin J, Stephens M, Buckland P, Owen M, O'Donovan M. Cheap, accurate and rapid allele frequency estimation of single nucleotide polymorphisms by primer extension and DHPLC in DNA pools. *Human Genetics* 107:488-493, 2000.

[6] Le Hellard S, Ballereau S, Visscher P. Torrance H, Pinson J, Morris S, Thomson M, Semple C, Muir W, Blackwood D, Porteous D, Evans K. SNP genotyping on pooled DNAs: comparison of genotyping technologies and a semi automated method for data storage and analysis. *Nucleic Acids Research* 30(15):e74., 2002.

[7] Norton N, Williams HJ, Williams NM, Spurlock G, Zammit S, Jones G, Jones S, Owen R, O'Donovan MC, Owen MJ. Mutation screening of the Homer gene family and association analysis in schizophrenia. *American Journal of Medical Genetics* 120B: 18-21, 2003.

[8] Norton N, Williams N, O'Donovan M, Owen M. DNA pooling as a tool for large-scale association studies in complex traits. *Annals of Medicine* 36:146-152, 2004.

[9] Visscher PM, Le Hellard S. Simple method to analyze SNP-based association studies using DNA pools. *Genetic Epidemiology* 24:291-296, 2003.

[10] Mohlke KL, Erdos MR, Scott LJ, Fingerlin TE, Jackson AU, Silander K, Hollstein P, Boehnke M, Collins FS. High-throughput screening for evidence of association by using mass spectrometry genotyping on DNA pools. *Proc. Natl. Acad. Sci. USA* 99:16928-16933, 2002.

[11] Yang HC, Pan CC, Lu R, Fann C. New adjustment factors and sample size calculation in a DNA-pooling experiment with preferential amplification. *Genetics* 169:399-410, 2005.

[12] Hoh J, Wille A, Ott J. Trimming, weighting, and grouping SNPs in human case-control association studies. *Genome Research* 11:2115-2119, 2001.

[13] Zee RY, Hoh J, Cheng S, Reynolds R, Grow MA, Silbergleit A, Walker K, Steiner L, Zangenberg G, Fernandez-Ortiz A, Mayaca C, Pinto E, Fernandez-Cruz A, Ott J, Lindpainter K. Multi-locus interactions predict risk for post-PTCA restenosis: an approach to the genetic analysis of common complex disease. *Pharmacogenomics Journal* 2:197-201, 2002.

[14] Hao K, Xu X, Laid N, Wang X, Xu X. (2004) Power estimation of multiple SNP association test of case-control study and application. *Genetic Epidemiology* 26:22-30.

[15] Kim S, Zhang K, Sun F. Detecting susceptibility genes in case-control studies using set association. *BMC Genetics* 4:Suppl 1: S9, 2003.

[16] Nelson MR, Kardia SL, Ferell RE, Sing CF. A combinatorial partitioning method to identify multilocus genotypic partitions that predict quantitative trait variation. *Genome Research* 11:458-470, 2001.

[17] Ritchie MD, Hahn LW, Roodi N, Bailey LR, Dupont WD, Pari FF, Moore JH. Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *American Journal of Human Genetics* 69:138-147, 2001.