New adjustment factors and sample size calculation in a DNA-pooling experiment with preferential amplification

Hsin-Chou Yang\textsuperscript{a}, Chia-Ching Pan\textsuperscript{b}, Richard C.Y. Lu\textsuperscript{c} and Cathy S. J. Fann\textsuperscript{a,b}

\textsuperscript{a}Institute of Biomedical Sciences, Academia Sinica, Taipei, Taiwan 115
\textsuperscript{b}Institute of Public Health, Yang-Ming University, Taipei, Taiwan 112
\textsuperscript{c}National Genotyping Center, Academia Sinica, Taipei, Taiwan 115
Calibration of preferential amplification

Keywords: DNA pooling, unequal allelic detection, allele frequency estimation, disease-gene mapping, sample size.

Corresponding author: Cathy S. J. Fann. Institute of Biomedical Sciences, Academia Sinica, 128, Academia Road, Section 2 Nankang, Taipei, Taiwan 115.

csjfann@ibms.sinica.edu.tw (E-mail)
886-2-27823047 (Fax)
886-2-27899144 (Tel.)
ABSTRACT

In the post-genome era, disease gene mapping using dense genetic markers has become an important tool for dissecting complex inheritable diseases. Locating disease susceptibility genes using DNA-pooling experiments is a potentially economical alternative to those involving individual genotyping. The foundation of a successful DNA-pooling association test is a precise and accurate estimation of allele frequency. In this paper, we propose two new adjustment methods that correct for preferential amplification of nucleotides when estimating the allele frequency of single nucleotide polymorphisms. We also discuss the effect of sample size when calibrating unequal allelic amplification. We conducted simulation studies to assess the performance of different adjustment procedures and found that our proposed adjustments are more reliable with respect to the estimation bias and root mean square error compared with the current approach. The improved performance not only improves the accuracy and precision of allele frequency estimations but leads to more powerful disease gene mapping.
INTRODUCTION

Locating disease susceptibility genes is an important topic in the post-genome era. To detect small genetic effect for many susceptible genes, a plausible etiological model for complex traits, many large case-control studies have been launched. Advances in biological techniques have made available thousands of single nucleotide polymorphisms (SNPs) for disease gene mapping. The availability of these dense markers vastly improves the power of association tests and increases the resolution of gene mapping in candidate region research and genome scanning studies.

Conventional case-control association studies are popular for disease gene mapping using individual genotyping data. However, analyses of large samples are often impractical due to the expense of individual genotyping. In this regard, DNA-pooling experiments may represent an economical alternative. As the name implies, DNA pooling involves the mixing of genomic DNA from many different individuals. Allele frequencies for each SNP marker in the pooled DNA are measured using the same principles that apply to genotyping.

A complete DNA-pooling experiment consists primarily of three stages. The first stage is a pilot study in which heterozygous individuals are collected to estimate the coefficient of preferential amplification (CPA). The coefficient is subsequently used to correct the estimates of allele frequencies in the second stage. In the second stage, DNA-pooling experiments are conducted for a large number of SNPs, and pooling association tests are carried out to screen for potential genetic markers. Only a small proportion of markers selected from the second stage experiments are included in the third stage in which all individuals are genotyped to confirm the validity of the markers selected from the second stage. As a consequence of the preliminary
screen in the second stage, the number of SNPs in the third stage is drastically reduced, thereby lowering genotyping costs.

DNA pooling is an efficient screening method for locating disease susceptibility genes (Bansal et al. 2002; Sham et al. 2002). However, this cost-saving alternative is efficient only if the estimation of allele frequency is accurate and precise. Biased or unreliable estimation of allele frequencies can lead to spurious results in association studies. Variation in the data from a DNA-pooling study may arise from several different sources, such as pool formation, polymerase chain reaction (PCR) amplification, allele frequency measurement and other uncontrollable experimental errors (Barratt et al. 2002; Visscher and Le Hellard 2003). Importantly, preferential amplification is a natural chemical attribute of PCR; it arises both from heterogeneous nucleotide incorporation during primer extension and differential efficiency of nucleotide detection during DNA quantification (Sham et al. 2002). These factors perturb the measurement of the intensity of different nucleotides and, consequently, the estimation of allele frequency. Under such a research background, we have focused on the impact of preferential amplification and propose new adjustment methods to rectify the problems inherent in the process. We also discuss the issue of sample size when correcting for unequal allelic amplification.

RESEARCH METHODS

Data and notation:

First, we discuss the design of the pilot stage. Consider that the two SNP-containing alleles are denoted $A$ and $a$, where allele $A$ is of interest. Given $n_{\text{total}}$ samples randomly drawn from a target population, individual genotyping results show that there are $n_{\text{heter}}$ heterozygous
individuals and \( n_{\text{homo}} \) homozygous individuals in the sample, i.e., \( n_{\text{total}} = n_{\text{heter}} + n_{\text{homo}} \). The pair of peak intensities for each heterozygous individual is determined (e.g., from MALDI-TOF spectrometry) as the area under the nucleotide mapping curve. The readings for alleles \( A \) and \( a \) are denoted \( \{H^1(j) = [h^1_A(j), h^1_a(j)], j = 1, \ldots, n_{\text{heter}}\} \). These bivariate vectors are used to quantify the magnitude of preferential amplification.

In the second stage of the screening experiment, genomic DNA from \( m \) different individuals is pooled. Applying the same genotyping principle used in the pilot study, we obtain a reading of the peak intensities for allele typing in a DNA-pooling experiment. The reading is the summary measure of this pool composed of genomic DNA from \( m \) individuals and is defined as \( H^p = [h^p_A, h^p_a] \). These data are used to estimate the allele frequency.

Let the population allele frequency of allele \( A \) be \( p_A \), the main parameter of interest. We define CPA \( \kappa \) as a measure of the peak intensities of allele \( A \) relative to allele \( a \), \( \kappa = \mu_A / \mu_a \), where \( \mu_A \) and \( \mu_a \) denote the average peak intensities of alleles \( A \) and \( a \) in the population. In other words, \( \kappa \) is the relative magnitude of the averaged amplified intensities of two different nucleotides and is an unknown calibration parameter that serves as an adjustment factor for allele frequency estimation. For \( \kappa > 1 \), the first nucleotide tends to be amplified more than the second; for \( \kappa < 1 \), the second nucleotide is likely to be amplified less than the first; for \( \kappa = 1 \), equal amplification is likely. The following sections introduce the statistical model/estimation of CPA, the estimation of population allele frequency, and association tests.

**Statistical model and estimation of CPA:**
The population allele frequency $p_A$ is defined as $p_A = \frac{N_A}{N_A + N_a}$, where $N_A$ and $N_a$ denote the number of alleles $A$ and $a$ in the population. In individual genotyping experiments, the population allele frequency can be estimated by directly counting the number of alleles from representative samples. The direct counting approach does not apply to DNA-pooling experiments because only the peak intensities are measured.

The relationship between peak intensity and allele frequency is the kernel of allele frequency estimation in a DNA-pooling study. The bridge is CPA, which connects both of them as indicated in the following expression

$$p_A = \frac{\kappa \times N_A}{\kappa \times N_A + N_a} = \frac{\mu_A N_a}{\mu_A N_A + \mu_a N_a} = \frac{H_A}{H_A + H_a} \quad \text{if} \quad \kappa = 1,$$

where $H_A$ and $H_a$ denote accumulated peak intensities of alleles $A$ and $a$, respectively. The allele frequency can be estimated by calculating the proportion of the peak intensities. However, if the amplification rate varies depending on the specific nucleotide at the SNP, then parameter $\kappa$ must be estimated and considered in the estimation of allele frequency. Below, we discuss the procedure for estimating preferential amplification using individual genotyping data from heterozygous individuals.

Suppose that there are $n_{\text{heter}}$ independent heterozygous individuals in the individual genotyping pilot study. Let the intensities of the two peaks for the $j$th heterozygous individual be $h_A^j(j)$ and $h_a^j(j)$, $j = 1, \ldots, n_{\text{heter}}$. The two-dimensional peak intensities $\{H^1(j) = [h_A^j(j), h_a^j(j)], j = 1, \ldots, n_{\text{heter}}\}$ are assumed to follow a bivariate distribution $G(\mu_A, \mu_a, \sigma_A^2, \sigma_a^2, \rho)$, where $(\mu_A, \mu_a)$ are the population means of the peak intensities for alleles
\(A\) and \(a\), \((\sigma_A^2, \sigma_a^2)\) are the variances of the peak intensities for alleles \(A\) and \(a\), and \(\rho\) denotes the correlation of the two intensities.

Under this model, we propose two measures to estimate CPA and compare them with the previous adjustment method for unequal amplification proposed by Hoogendoorn et al. (2000). Previously, their adjustment factor is defined as the arithmetic mean of ratios, i.e.,

\[
\hat{k}_H = n_{\text{heter}}^{-1} \times \sum_{j=1}^{n_{\text{max}}} \left[ \frac{h_A(j)}{h_a(j)} \right].
\]

This pioneering approach has been adopted by many of researchers in cases of preferential amplification (Le Hellard et al. 2002; Mohlke et al. 2002; Werner et al. 2002). The advantage of this method is very simple in concept and calculation.

Our first proposed adjustment reduces the bias in Hoogendoorn’s method using a bias-reduction technique and can be represented as follows:

\[
\hat{k}_U = \hat{k}_H + \frac{n_{\text{heter}}}{n_{\text{heter}} - 1} \left( \frac{\overline{h}_A}{\overline{h}_a} - \hat{k}_H \right),
\]

where \(\overline{h}_A = n_{\text{heter}}^{-1} \sum_{j=1}^{n_{\text{max}}} h_A(j)\) and \(\overline{h}_a = n_{\text{heter}}^{-1} \sum_{j=1}^{n_{\text{max}}} h_a(j)\). The difference between \(\hat{k}_H\) and \(\hat{k}_U\) in Equation (3) is the estimated bias of Hoogendoorn’s method. A detailed derivation is presented in Appendix A. The ratio of a pair of peak intensities often exhibits a skew distribution and log transformation is often considered to reduce the skewness and variability. Therefore, our second proposed adjustment factor is the geometric mean of ratios

\[
\hat{k}_G = n_{\text{max}} \left( \prod_{j=1}^{n_{\text{max}}} \frac{h_A(j)}{h_a(j)} \right)^{1/n_{\text{max}}},
\]

This factor can be regarded as an exponential function whose power is the arithmetic mean of log-scale ratio.
The standard error of each adjustment measure reflects sampling variability and is critical for the association test in the next stage. Because the number of heterozygous individuals might be small, and an exact statistical distribution of these adjustment measures is difficult to derive, a bootstrapping procedure (Efron and Tibshirani 1993) is recommended to estimate the standard errors. Original data are used to estimate the hyper-parameter by a moment-based or likelihood-based approach to obtain the empirical distribution \( G(\mu_A, \sigma^2_A, \sigma^2_a, \hat{\rho}) \). Pseudo-samples are generated using resampling from the empirical distribution with replacement. Suppose the number of bootstrap replications is \( B \). Each adjustment method in Equations (3) or (4) is applied to the samples to obtain the corresponding estimates \((\hat{\kappa}_1, \cdots, \hat{\kappa}_B)\). Hence, the standard error of the adjustment measure can be calculated by taking the sample standard deviation of the bootstrap estimates

\[
\hat{\sigma}_\kappa = \left[ \sum_{b=1}^{B} (\hat{\kappa}_b - \bar{\kappa})^2 / (B-1) \right]^{1/2}, \tag{5}
\]

where \( \bar{\kappa} = \sum_{b=1}^{B} \hat{\kappa}_b / B \).

**Estimation of allele frequency and test of allelic association:**

In this section, we discuss the estimation of allele frequency when preferential amplification is involved. The genomic DNA from all cases is mixed together in a pool and that of controls is mixed in the other pool. The pairs of peak intensities in control and case groups are denoted by \( H^{p,\text{control}} = [h_A^{p,\text{control}}, h_a^{p,\text{control}}] \) and \( H^{p,\text{case}} = [h_A^{p,\text{case}}, h_a^{p,\text{case}}] \), respectively.

If there is no preferential amplification, then the coefficient \( \kappa \) will be approximately one, and hence no adjustment is necessary. The allele frequencies of allele \( A \) in control and case groups can be estimated directly by calculating the proportion of peak intensities as follows:
If $\kappa$ is larger than one, then allele $A$ tends to be amplified more than allele $a$, and vice versa. In these two cases, the scales of the two intensities differ, and the population-level relative proportion of the two amplified abilities is simply $\kappa$. To adjust for non-equivalent allelic amplification, the method proposed by Hoogendoorn et al. (2000) increased the suppressed intensity by multiplying the CPA. This transformation procedure standardizes the two intensities in scale. At the population level, substituting $\tilde{H}_a = \kappa H_a$ for $H_a$ makes the equality on the left side of Equation (1) hold even for $\kappa \neq 1$; at the sample level, $\tilde{h}_a = \kappa \times \hat{h}_a$ and $\tilde{h}_a = \kappa \times \hat{h}_a$ are used to adjust for unequal amplification. The adjusted terms that improve the estimated allele frequency in Equation (6) for control and case groups are as follows:

$$\hat{p}_A^{\text{control}} = \frac{\hat{h}_a}{\hat{h}_a^{\text{control}} + \hat{h}_a^{\text{control}}} \quad \text{and} \quad \hat{p}_A^{\text{case}} = \frac{\hat{h}_a}{\hat{h}_a^{\text{case}} + \hat{h}_a^{\text{case}}}.$$  \hspace{1cm} (6)

Because the allele frequency estimator is a function of CPA, it varies with the adjustment factor $\hat{\kappa}$. The performances should be evaluated. In the Simulation Study section below, we discuss how simulation studies assess the performance of these adjustment factors.

Screening potential SNP markers associated with a disease locus is the main purpose of a DNA-pooling study (Bansal et al. 2002; Sham et al. 2002). This can be achieved using the pooling-based association test (Visscher and Le Hellard 2003)

$$\chi^2 = \frac{(\hat{p}_A^{\text{case}} - \hat{p}_A^{\text{control}})^2}{V(\hat{p}_A^{\text{case}} - \hat{p}_A^{\text{control}})},$$  \hspace{1cm} (8)

where

$$V(\hat{p}_A^{\text{case}} - \hat{p}_A^{\text{control}}) = \frac{\sigma_a^2}{2n_{\text{case}}} \left(\frac{\hat{p}_a^{\text{case}} p_a}{2n_{\text{case}}} + \frac{\hat{p}_a^{\text{control}} p_a}{2n_{\text{control}}} + V(\hat{\kappa}) \left(\frac{\hat{p}_a^{\text{case}} p_a}{2n_{\text{case}}} - \frac{\hat{p}_a^{\text{control}} p_a}{2n_{\text{control}}}\right)^2 + 2\hat{\kappa}^2\right)$$  \hspace{1cm} (9)
and $\sigma^2$ denote the experimental variation. The sampling distribution of the test statistic follows a chi-squared distribution with one degree of freedom asymptotically. The first two terms after the equality in Equation (9) are the variance components due to sampling variation; the third term is resulted from the adjustment variation of preferential amplification; the fourth term is the experimental variation from several different sources, such as pool formation. All of the parameters in Equation (9) are unknown and therefore must be estimated.

Parameter $\kappa$ is estimated by our proposed method in Equation (3) or Equation (4); variance $V(\hat{\kappa})$ is estimated by the proposed bootstrap variance in Equation (5); the allele frequencies are estimated using Equation (7). Finally, the experimental variance can be estimated by calculating the mean square errors (Barratt et al. 2002) or using the restricted maximum likelihood method (Downes et al. 2004) based on a hierarchical experimental design.

Estimation of CPA affects both the denominator and the numerator of the test statistic in Equation (8) simultaneously. The impact of CPA on the denominator is explicit in Equation (9). CPA affects the numerator by way of allele frequency estimates. Based on the adjusted allele frequency defined in Equation (7), the expectation of difference between the estimated allele frequencies in case and control groups is zero under null hypothesis (no association). If the adjusted allele frequency in Equation (7) is replaced by the unadjusted allele frequency in Equation (6) for constructing test statistic, the zero expectation may not hold true under null hypothesis. Upon algebraic manipulation, we have proved that the discrepancy between case and control groups with regard to CPA-adjusted and unadjusted allele frequencies (i.e. the difference between the CPA-adjusted case and control group allele frequencies minus the difference between the unadjusted case and control group allele frequencies) is
\[ \delta = \frac{(\hat{k} - 1)(h_j^{p,\text{case}} h_i^{p,\text{control}} - h_j^{p,\text{control}} h_i^{p,\text{case}})(\hat{k}h_j^{p,\text{case}} h_i^{p,\text{control}} - h_j^{p,\text{case}} h_i^{p,\text{control}})}{(h_j^{p,\text{case}} + h_i^{p,\text{case}})(h_j^{p,\text{control}} + \hat{k}h_i^{p,\text{case}})(h_j^{p,\text{control}} + h_i^{p,\text{control}})(h_j^{p,\text{control}} + \hat{k}h_i^{p,\text{control}})}. \]

The numerator represents three cases in which there is no effect of adjustment: (1) No preferential amplification; (2) No difference in allele frequency between case and control groups; (3) The sum of the case group allele frequency with (without) adjustment and the control group allele frequency without (with) adjustment is equal to one.

**Sample size in the pilot study:**

In the pilot study, the peak intensities of heterozygous individuals are needed to estimate the CPA. An immediate question is how many heterozygous individuals are required to obtain a precise estimate of CPA. Proceeding in the context of confidence intervals, we calculate the sample size under risk $\alpha$ and a specified absolute error $\xi$ as follows:

\[ n_{\text{heter}} = \left[ \left( t_{1-\frac{\alpha}{2}} p(1-p) CV_r \right) \cdot \frac{1}{\xi^2} \right]^2, \text{ if } P\{ | p(\hat{k}) - p | < \xi \} \geq 1 - \alpha, \]  
\[
(10)
\]

where $CV_r^2 = V(r_j)/k^2$ and $r_j = h_i^l(j)/h_j^l(j), j = 1, \ldots, n_{\text{heter}}$. Equation (10) is derived based on Hoogendoorn’s method, and the details are shown in Appendix B. From our simulation study (discussed below), we find that our proposed method gives a lower standard error for the variance compared with Hoogendoorn’s method. Hence, the sample size in Equation (10) is regarded as the upper bound for our proposed estimators. Under $\alpha = 0.05$ and $\xi = 0.05, 0.075$ and 0.10, the relationships between the sample size and different parameters are shown in Figure 1. The results show that sample size correlates positively with $CV_r$ and is inversely proportional to $\xi$. The symmetry and highest points of the sample size curve occur concurrently when the allele frequency is 0.5.

(Figure 1 is about here)
Sample sizes for additional genotyping must be evaluated to achieve the required number of heterozygous individuals derived from Equation (10). This aspect depends on the design of the genotyping experiment, the genetic background of SNP markers and population characteristics. In a sequential genotyping experiment, \( n_{\text{total}} \) is a random variable that follows a negative binomial distribution with successful probability (probability of heterozygote) \( p_{H} = p_{A} \). However, in large-scale genotyping experiments, individuals are genotyped simultaneously, not sequentially. Under this circumstance, \( n_{\text{total}} \) is pre-specified and \( n_{\text{heter}} \) is a random variable from a binomial distribution with successful probability \( p_{H} = p_{A} \).

The theory based on the assumption of individual homogeneity is sometimes too stringent. Heterogeneity among individuals may be due to various individual covariates or unobserved attribution. If this potential factor is ignored, the genotyping efforts will be underestimated. A possible remedy is to assume a random-effect model. A beta distribution \( \text{beta}(\theta, \tau) \) with the corresponding mean \( \theta/ (\theta + \tau) \) and coefficient of variation (CV) \( \{\tau/ [\theta(\theta + \tau + 1)]\}^{1/2} \) can be used to model random allele frequency (RAF) and random genotype frequency (RGF), respectively. The corresponding marginal distributions of random variables \( n_{\text{total}} \) and \( n_{\text{heter}} \) with respect to the sequential genotyping experiment and large-scale genotyping experiment can be obtained by integrating out the hyper-parameters in the beta distribution as follows:

\[
f(n_{\text{total}}) = \begin{cases} 
\frac{n_{\text{total}} - 1}{n_{\text{heter}} - 1} \frac{2^{n_{\text{heter}}}}{B(\theta, \tau)} \sum_{y=0}^{n_{\text{homo}}} \binom{n_{\text{homo}}}{y} (-1)^y 2^y B(n_{\text{heter}} + y + \theta, n_{\text{heter}} + y + \tau), & \text{for RAF}, \\
\frac{n_{\text{total}} - 1}{n_{\text{heter}} - 1} \frac{B(\theta + n_{\text{heter}} + \tau + n_{\text{homo}})}{B(\theta, \tau)}, & \text{for RGF}, 
\end{cases}
\]
\[
f(n_{\text{heter}}) = \begin{cases} 
\left( \frac{n_{\text{total}}}{n_{\text{heter}}} \right)^{2^n_{\text{base}}} \frac{\sum_{y=0}^{n_{\text{homo}}} \binom{n_{\text{homo}}}{y}}{B(\theta, \tau)} (-1)^y 2^y B(n_{\text{heter}} + y + \theta, n_{\text{heter}} + y + \tau), \text{ for RAF,} \\
\left( \frac{n_{\text{total}}}{n_{\text{heter}}} \right) B(\theta + n_{\text{heter}}, \tau + n_{\text{homo}}) B(\theta, \tau), \text{ for RGF,} 
\end{cases}
\]

where \( B(., .) \) is the conventional beta function. A detailed derivation is presented in Appendix C.

The relationships between sample size and the observation probability under different genotyping strategies are shown in Figure 2 and Figure 3, respectively. Only the results based on genotype frequency are shown. If hyper-parameters in the beta distribution are unity, then random-effect model reduces to the special case in which no individual heterogeneity exists.

Figure 2 shows the distribution of \( n_{\text{total}} \) with a genotype frequency range of 0.05–0.5 in increments of 0.05. The pattern reveals the positive correlation between \( P(N \leq n_{\text{total}}) \) and \( n_{\text{total}} \).

A heterozygous genotype frequency of a SNP of greater than 0.15 almost guarantees that eight heterozygous individuals can be observed after genotyping 96 individuals (Figure 2 A); the probability of observing 16 heterozygous individuals is higher than 0.8 (Figure 2 B). Figure 2 (C) and (D) present the results under the condition of individual random effects. A lower \( E(p_{\text{da}}) \) corresponds to a less polymorphic case, and therefore genotyping requires many more individuals. The coefficient of variation \( CV(p_{\text{da}}) \) affects the curvature of different lines. In general, a smaller \( CV(p_{\text{da}}) \) yields a steeper slope; in other words, the marginal increase in cumulative probability (corresponding to an increase in \( n_{\text{total}} \)) is larger when \( CV(p_{\text{da}}) \) is smaller.

The required number of genotyped individuals can be obtained using a pre-specified probability from these figures.

(Figure 2 is about here)
Figure 3 shows the probability distribution of $n_{\text{heter}}$. Figures 3 (A) and (B) present the cases for $n_{\text{total}} = 48$ and $n_{\text{total}} = 96$ in the absence of individual heterogeneity, and Figures 3 (C) and (D) present the case when individual heterogeneity is present. The genotype frequencies range is 0.05–0.5 in increments of 0.05. In general, the correlation between $n_{\text{heter}}$ and $P(N \geq n_{\text{heter}})$ is negative. If 96 individuals are genotyped, Figure 3 (B) and (D) show that, even with individual heterogeneity, eight heterozygous individuals can be observed in most cases except for the non-polymorphic one. A similar pattern is evident in Figure 3 (A) and (C) for the case $n_{\text{total}} = 48$, but the probability $P(N \geq n_{\text{heter}})$ decreases. The index of heterogeneity $CV(p_{\text{heter}})$ affects the pattern of the probability of identifying heterozygous individuals. Figure 3 (C) and (D) show the curves with a smaller $CV(p_{\text{heter}})$ have a sharper reduction whereas those with a large $CV(p_{\text{heter}})$ have a gentler slope.

(Figure 3 is about here)

In this section, we focused exclusively on sample size in a pilot study of a DNA-pooling study. Discussions concerning sample size in the second stage can be found in Barratt et al. (2002).

SIMULATION STUDY

Procedures:

We carried out simulation studies to assess both the performance of different adjustment factors for estimating CPA in the first stage and the consequential impact on the pooling-based association test in the second stage.
For the first stage, the simulation procedures can be outlined as follows: (1) *Specify the simulation conditions*: Consider that the number of heterozygous individuals $n_{inter}$ ranges from 8 to 40 with an increment of 16, and the true CPA $\kappa$ is set to 0.5, 1 and 2. (2) *Generate peak intensity data*: Because a gamma distribution can cover many different random patterns, we considered a bivariate gamma distribution of the peak intensities in the simulation study. The parameters for the bivariate gamma distribution were set to yield CVs of 0.1 or 0.3 for the peak intensities, and the correlation of the pair of peak intensities was 0.5. (3) *Estimate the adjustment factor and hyper-parameters*: Based on the data from Step 2, we calculated the adjustment factor $\kappa^{(s)}$ and estimated the hyper-parameters of the gamma distribution using the moment method, where the superscript was the simulation index. (4) *Calculate bootstrap standard error*: Bootstrapping data from the empirical gamma distribution $\text{Gamma}(\hat{\mu}, \hat{\sigma}, \hat{\sigma}^2, \hat{\rho})$ for $B$ times were used to estimate the CPA $(\hat{\kappa}_1^{(s)}, \ldots, \hat{\kappa}_B^{(s)})$. The bootstrap standard error of the adjustment factor was obtained by calculating sample standard deviation over the $B$ estimates, $\hat{\sigma}_\kappa^2 = \left[ \frac{1}{B} \sum_{b=1}^{B} (\hat{\kappa}_b^{(s)} - \overline{\kappa}^{(s)})^2 \right]^{1/2}$, where $\overline{\kappa}^{(s)} = \frac{1}{B} \sum_{b=1}^{B} \hat{\kappa}_b^{(s)} / B$. (5) *Calculate the estimation bias, standard error and root mean square error*: Steps 2 to 4 were repeated for $S$ simulation replications. The estimation bias and standard error were calculated using $\text{BIAS} = (\sum_{s=1}^{S} \hat{\kappa}^{(s)} / S) - \kappa$ and $\text{SE} = (\sum_{s=1}^{S} \hat{\sigma}_\kappa^{(s)} / S)$, respectively. The root mean square error (RMSE) was $\text{RMSE} = (\text{BIAS}^2 + \text{SE}^2)^{1/2}$.

In the second stage, we simulated case-control association tests using the following procedures: (1) *Specify the simulation conditions*: The sample sizes were $n_{case} = n_{control} = 500$. The population allele frequencies in case and control groups were $(p_{case} = 0.25, p_{control} = 0.25)$ and $(p_{case} = 0.25, p_{control} = 0.15)$ for the calculation of type I error and power respectively. The
standard deviation of experimental error was set to $\sigma_E = 0.02$ or $\sigma_E = 0.05$. (2) **Generate allele frequency**: The sample frequencies in the case group were generated from a normal distribution $N(p_{\text{case}}, \text{Var}(\hat{p}_{\text{case}} | \hat{\kappa}))$, and a similar approach was applied to the control group. (3) **Summarize the test results**: Based on the simulated data, we calculated the type I error when the case and control groups had the same true allele frequency; we calculated the power when the groups had different true allele frequencies.

A total of 12 simulation conditions were considered and were arranged in the following order: $(\sigma_E, CV, n_{\text{het}}) = \{(0.02, 0.1, 8), (0.02, 0.1, 24), (0.02, 0.1, 40), (0.02, 0.3, 8), (0.02, 0.3, 24), (0.02, 0.3, 40), (0.05, 0.1, 8), (0.05, 0.1, 24), (0.05, 0.1, 40), (0.05, 0.3, 8), (0.05, 0.3, 24), (0.05, 0.3, 40)\}$. All simulations were carried out using 200 simulation replications and 500 bootstrap replications.

**Results:**

In the simulation studies, we explored the impact of several elements on the adjustment factor and association test. These elements included the structure of peak intensity, degree of preferential amplification, experimental error and sample size. The simulation results of $\kappa = 0.5$, $\kappa = 1$ and $\kappa = 2$ are summarized in Tables 1, 2 and 3, respectively.

(Tables 1, 2 and 3 are about here)

With regard to the performance of adjustment factors, we found several meaningful patterns. The estimated CPA is affected by the CV of the peak intensities. As CV increases, i.e., in the case of large variability of peak intensities, there is a concomitant rise in the estimation bias, standard error and RMSE of $\hat{\kappa}$. Relative to CV, the mean or variance of the peak intensities alone is insufficient to explain the changes in performance of adjustment factors. As more
heterozygous individuals were collected in the pilot study, the standard error and RMSE of $\hat{\kappa}$ were reduced; however, the effect on the bias of $\hat{\kappa}$ was not obvious.

Regarding the performance of the association test, we found that the increase in experimental variation dramatically reduced the statistical power of the association test. Generally speaking, from $\sigma_e = 0.02$ to $\sigma_e = 0.05$, the power decreases from 0.85 to 0.30. Moreover, the type I error correlates positively with the CV of the peak intensities. In most cases, the type I error range was 0.03–0.07. Although larger numbers of heterozygous individuals are useful to reduce the RMSE of the adjustment factor, the efficacy of the association test depends on the degree of experimental error. When $\sigma_e = 0.02$, the reduction in the variation of $\hat{\kappa}$ improves the power; when $\sigma_e = 0.05$, the improvement in the adjustment is neutralized by an increase in the experimental error. The same idea applies to the CV of peak intensity.

In general, the proposed adjustment measures yielded better performance than Hoogendoorn’s method with respect to the estimation of $\kappa$ and the association test. In all simulation trials, we found that the two proposed adjustment factors yielded a smaller bias, standard error and RMSE compared with Hoogendoorn’s method. Given a pre-specified test size, $\hat{\kappa}_u$ yielded the highest power among the three adjustment methods with regard to relatively small experimental errors ($\sigma_e = 0.02$); $\hat{\kappa}_g$ yielded the highest power in cases of larger experimental errors ($\sigma_e = 0.05$).

ANALYSIS OF A LABORATORY EXAMPLE

We conducted a DNA-pooling study to illustrate the efficacy of the proposed adjustment methods and to facilitate a comparison with Hoogendoorn’s method (Hoogendoorn et al. 2000).
Using normal control samples that we collected previously, 95 individuals were randomly chosen and genotyped individually. There are six SNPs in total. The peak intensities of heterozygous individuals were used to calculate the various estimates of CPA. Later, the genomic DNA of 30 individuals randomly selected from the 95 individuals was pooled to estimate allele frequency. The true minor allele frequency of the 30-individual population was attained based on individual genotyping data using the allele counting approach. The results are shown in Table 4 (column 2).

In the DNA-pooling experiment, each individual’s genomic DNA was diluted to 12.5 ng/µl and quantified using the PicoGreen assay (Molecular Probes). Equimolar amounts of genomic DNA from the 30 individuals were then pooled. PCR amplification and primer extension reactions were performed using an ABI 9700 system (AME Bioscience). The peak intensities of alleles were measured using a MALDI-TOF spectrometer (Sequenom) based on wavelet technology. Hence, the unadjusted allele frequencies for the six SNPs could be estimated based on Equation (6), and the results are shown in Table 4 (column 3).

Hoogendoorn’s adjustment and our two proposed methods were applied to this data set. For the six SNPs, the estimates of CPA were as follows: based on Hoogendoorn’s method, 1.259, 0.765, 0.662, 0.873, 1.771, 2.288; based on the geometric-mean method, 1.252, 0.687, 0.634, 0.859, 1.749, 2.265; and based on bias-reduction method, 1.221, 0.674, 0.632, 0.851, 1.788, 2.320. The adjusted allele frequency estimates, based on Equation (7), are shown in Table 4 in columns 4 (Hoogendoorn’s), 5 (geometric-mean) and 6 (bias-reduction).

To summarize the findings in this analysis of laboratory data, we found that it is essential to adjust for preferential amplification. This adjustment reduced the estimation bias of the allele frequencies except for the second SNP in this data set. In this case, the serious underestimate of
allele frequency might have arisen from uncontrollable experimental variations, such as overestimation of the extended primer or an effect of DNA quality on SNP variance (Werner et al. 2002). For this specific SNP, the adjustment procedure yielded only limited improvement.

In most cases in our study, Hoogendoorn’s method reduced the discrepancy between the allele frequencies estimated from the individual genotyping and pooling experiments. Our proposed adjustments reduced the error further than did Hoogendoorn’s method. Moreover, our proposed methods yielded a smaller variation in the CPA compared with Hoogendoorn’s method, and in turn our methods gave a smaller variation in allele frequency estimation. Overall, our results demonstrate that the proposed adjustments provide a more accurate and reliable estimation of allele frequency for this data set.

DISCUSSION

Preferential amplification of nucleotides occurs frequently in DNA-pooling studies. Therefore, it is critical to adjust this interference presented in two nucleotides in the same SNP so as to avoid a severe bias in the allele frequency estimation and a consequent reduction in the power of the association test. In the present work, we propose two adjustment methods that improve on the Hoogendoorn’s adjustment (Hoogendoorn et al. 2000). The performance was evaluated by simulation studies. The new methods yield not only lower bias, standard error and RMSE in allele frequency estimation, but also better statistical power for genetic association mapping under the given test size.

In our method, type I error is usually controlled well except when the CV of the peak intensity is high and \( \kappa > 1 \). Also, the performance of \( \hat{\kappa} \) is apparently not symmetric when \( \kappa = 1 \). In general, a smaller CPA yields a correspondingly smaller RMSE. Regarding the instances of
\( \kappa = 2 \) and \( \kappa = 0.5 \), the latter gives better performance with respect to RMSE, power and type I error. Hence, we suggest that the preferentially amplified allele be placed in the denominator when calculating the adjustment factor (see Equation (3) and Equation (4)).

In addition to our two new proposed adjustment factors, we also investigated several other methods, including the median-based measure, harmonic-mean-based measure and some modified ratio estimators (Beale 1962; Tin 1965). Although some of these methods yielded a better estimate of CPA than Hoogendoorn’s method during the simulation study (Yang et al. 2003), they are not superior to our proposed adjustment factors in this paper.

We investigated the role of sample size during the pilot stage of the DNA-pooling study. The use of a large number of heterozygous individuals reduces the RMSE of the adjustment factor. However, 24 heterozygous individuals should be sufficient to obtain a good adjustment factor. To explore the sample size requirement, we considered both fixed-effect and random-effect models for different scenarios. In general, random-effect models yield a larger sample size than fixed-effect models when the same set of parameters is used. For example, given that the genotype frequency in population is 0.45 and eight heterozygous individuals are required, 21 individuals need to be genotyped under the fixed-effect model (i.e., \( CV(p_{da}) = 0 \)); 24 and 34 individuals need to be genotyped under \( CV(p_{da}) = 0.25 \) and 0.5 in the random-effect model respectively. This is so because the random-effect models take into consideration variations among individuals, resulting in an increase in total variation that requires a larger sample size in order to compensate. Ignoring the heterogeneity may lead to a serious underestimation of sample sizes.

DNA-pooling studies can achieve the objective of screening for important genetic variants at a reasonable cost. However, the unavoidable drawback is that genotypic information
and individual features are lost once genomic DNA is mixed. Although some advanced studies have attempted to reconstruct the lost information (Ito et al. 2003), a bottleneck still exists due to the mass of genotype and haplotype combinations in the pool. Stringent limitations must be satisfied for small pools to reduce the number of combinatorial calculations. To date, DNA-pooling experiments have been used primarily as a screening technique rather than a legitimate replacement for individual genotyping studies.

Use of DNA pooling is a potentially cost-effective alternative to individual genotyping. Association testing in DNA pools is an efficient method for screening important genetic markers, and have been applied successfully to many practical applications and found significant conclusions (Carmi et al. 1995; Bansal et al. 2002). Our new approaches provide valid adjustments and further improve the conventional method, thereby the reliable allele frequency estimation and powerful association tests. These advantages enhance greatly the applicability of DNA-pooling experiment.

ACKNOWLEDGEMENTS

We would like to thank Dr. Jer-Yuan Wu and the National Genotyping Center at Academia Sinica for genotyping support. We appreciate the two anonymous reviewers for providing insightful suggestions and comments, which have greatly enhanced the presentation of this paper. This research was supported by a National Science Council grant (NSC 92-3112-B-001-014) and an Academia Sinica grant (93IBMS2PP-C) of Taiwan.
APPENDIX A

Properties of the proposed adjustment factor:

Suppose that the population size of heterozygous individuals is \(N_{\text{heter}}\) and \(n_{\text{heter}}\) samples are randomly drawn from the population. Let \(r(j) = h^1_a(j)/h^1_a(j)\). Hence, the bias of Hoogendoorn’s measure can be calculated using the following:

\[
\hat{\lambda} = E(\hat{\kappa}_{hi}) - \kappa = E(n_{\text{heter}}^{-1} \sum_{j=1}^{n_{\text{heter}}} [h^1_a(j)/h^1_a(j)]) - \kappa
\]

\[
\hat{\lambda} = N_{\text{heter}}^{-1} \sum_{j=1}^{n_{\text{heter}}} \left[ h^1_a(j)/h^1_a(j) \right] - H_A/H_a
\]

\[
= (N_{\text{heter}} \cdot H_a)^{-1} \sum_{j=1}^{n_{\text{heter}}} \left[ r(j) \cdot H_a \right] - \sum_{j=1}^{n_{\text{heter}}} \left[ r(j)h^1_a(j) \right] / H_a
\]

\[
= (N_{\text{heter}} \cdot H_a)^{-1} \times \sum_{j=1}^{n_{\text{heter}}} \left[ r(j) \cdot H_a \right] - (N_{\text{heter}} \cdot H_a)^{-1} \times n_{\text{heter}} \sum_{j=1}^{n_{\text{heter}}} \left[ r(j)h^1_a(j) \right]
\]

\[
= (N_{\text{heter}} \cdot \mu_a)^{-1} \sum_{j=1}^{n_{\text{heter}}} r(j) \cdot [\mu_a - h^1_a(j)].
\]

We used the following estimated bias

\[
\hat{\lambda} = (\hat{h}^1_a)^{-1} \times (n_{\text{heter}} - 1)^{-1} \times n_{\text{heter}} \times (\hat{\kappa} \cdot \hat{h}^1_a - \hat{h}^1_a)
\]

to correct the bias in Hoogendoorn’s method. This estimator is based on an infinite population issue relative to the consideration of finite population in Hartley and Ross (1954).

APPENDIX B

Sample size calculation in a pilot study:

Suppose that the estimated allele frequency \(\hat{p}\) is a differentiable function of \(\hat{\kappa}\) and \(h^p_a/h^p_a\) at a point \((\kappa, H_A/H_a)\). By a first-order Taylor expansion of two variables \(\hat{\kappa}\) and \(h^p_a/h^p_a\) at \((\kappa, H_A/H_a)\), we know that
\[ \hat{p} = p - \frac{1}{2} \left( \frac{p^2}{\left( H_A / H_a \right)} \right) \cdot (\hat{k} - \kappa) + p^2 \cdot \left( \frac{k}{\left( H_A / H_a \right)} \right)^2 \cdot \left( \frac{h_A^p / h_a^p}{H_A / H_a} - \frac{H_A / H_a}{H_A / H_a} \right) + R_i, \]

where \( \eta = (\hat{k} - \kappa, h_A^p / h_a^p - H_A / H_a) \) and satisfies \( R_i / \| \eta \| \to 0 \) as \( \| \eta \| \to 0 \). The approximate mean and variance can be calculated as follows:

\[
E(\hat{p}) \approx p \quad \text{and} \quad \text{Var}(\hat{p}) \approx p^2 (1 - p)^2 \frac{\kappa^2 p^4}{\left( H_A / H_a \right)^2} \frac{CV^2_{h_A^p / h_a^p}},
\]

where \( CV^2_{h_A^p / h_a^p} = V(\hat{k}) / \kappa^2 \) and \( CV^2_{h_A^p / h_a^p} = V(h_A^p / h_a^p) / (H_A / H_a)^2 \). The pure variation due to the adjustment of preferential amplification can be obtained by setting \( h_A^p / h_a^p = H_A / H_a \), resulting in the following variance

\[
\text{Var}(\varepsilon_{\kappa}) \approx p^2 (1 - p)^2 CV^2_{\kappa}.
\]

If \( \hat{k} = \kappa \), then we obtain the pure variance due to the measurement of the ratio of the peak intensities is \( \text{Var}(\hat{p} \mid \hat{k} = \kappa) \approx \frac{\kappa^2 p^4}{\left( H_A / H_a \right)^2} \frac{CV^2_{h_A^p / h_a^p}}{\kappa}. \)

To calculate the required samples size, the variance in Equation (A1) must be rewritten as a function of \( n_{\text{heter}} \). If the adjustment of Hoogendoorn et al. (2000) is applied, then the variance in Equation (A1) can be represented as

\[
\text{Var}(\varepsilon_{\kappa}) \approx p^2 (1 - p)^2 \frac{CV^2_{r_j}}{n_{\text{heter}}},
\]

where \( CV^2_{r_j} = V(r_j) / \kappa^2 \) and \( r_j = h_A^p(j) / h_a^p(j), j = 1, \ldots, n_{\text{heter}} \). Under risk \( \alpha \) and the specified absolute error \( \xi \), the required number of heterozygous individuals is

\[
n_{\text{heter}} = \left[ \left( h_a^p p(1 - p) CV_{r_j} \cdot \xi^{-1} \right) \right]^2,
\]

where \( P(\mid p(\hat{k}) - p < \xi) \geq 1 - \alpha \).
APPENDIX C

The marginal distributions of sample sizes under different genotyping experiments:

Firstly, under sequential genotyping experiments, \( n_{\text{total}} \) is a random variable that follows a negative binomial distribution with successful probability \( p_{HI} \) (probability of heterozygote). Under random allele frequency model (RAF), we assume allele frequency \( p_A \) follows from a beta distribution \( \text{beta}(\theta, \tau) \). Hence, the marginal distribution of \( n_{\text{total}} \) can be derived as follows:

\[
f(n_{\text{total}}) = \int \left( \begin{pmatrix} n_{\text{total}} - 1 \\ n_{\text{heter}} - 1 \end{pmatrix} \right) (2p_Ap_a)^{n_{\text{homo}}} (1 - 2p_Ap_a)^{n_{\text{homo}}} \times \frac{(p_A)^{\theta - 1}(1 - p_A)^{\tau - 1}}{B(\theta, \tau)} dp_A
\]

\[
= \left( \begin{pmatrix} n_{\text{total}} - 1 \\ n_{\text{heter}} - 1 \end{pmatrix} \right) \frac{2^{n_{\text{homo}}}}{B(\theta, \tau)} \int_0^1 (p_A)^{n_{\text{homo}} + \theta - 1}(1 - p_A)^{n_{\text{homo}} + \tau - 1} \sum_{y=0}^{n_{\text{homo}}} \binom{n_{\text{heter}}}{y} (-1)^y (2p_Ap_a)^y dp_A
\]

Under random allele genotype model (RGF), we assume genotype frequency \( p_{AA} \) follows from a beta distribution \( \text{beta}(\theta, \tau) \). The marginal distribution of \( n_{\text{total}} \) is

\[
f(n_{\text{total}}) = \int \left( \begin{pmatrix} n_{\text{total}} - 1 \\ n_{\text{heter}} - 1 \end{pmatrix} \right) (p_{AA})^{n_{\text{homo}}} (1 - p_{AA})^{n_{\text{homo}}} \times \frac{(p_A)^{\theta - 1}(1 - p_A)^{\tau - 1}}{B(\theta, \tau)} dp_A
\]

\[
= \left( \begin{pmatrix} n_{\text{total}} - 1 \\ n_{\text{heter}} - 1 \end{pmatrix} \right) \frac{1}{B(\theta, \tau)} \int_0^1 (p_A)^{n_{\text{homo}} + \theta - 1}(1 - p_A)^{n_{\text{homo}} + \tau - 1} dp_A
\]

Secondly, under large-scale genotyping experiments, \( n_{\text{heter}} \) is a random variable that follows a binomial distribution with successful probability \( p_{HI} \). Under beta-binomial random allele frequency model, the marginal distribution of \( n_{\text{heter}} \) is

\[
f(n_{\text{heter}}) = \int \left( \begin{pmatrix} n_{\text{total}} \\ n_{\text{heter}} \end{pmatrix} \right) (2p_Ap_a)^{n_{\text{homo}}} (1 - 2p_Ap_a)^{n_{\text{homo}}} \times \frac{(p_A)^{\theta - 1}(1 - p_A)^{\tau - 1}}{B(\theta, \tau)} dp_A
$\tau = \sum_{y=0}^{\tau_{\text{homo}}} \left( \begin{array}{c} n_{\text{heter}} \\ n_{\text{heter}} \end{array} \right) (-1)^y 2^y B(n_{\text{heter}} + y + \theta, n_{\text{heter}} + y + \tau).$

Under beta-binomial random genotype frequency model, the marginal distribution of $n_{\text{heter}}$ is

$$f(n_{\text{heter}}) = \int \left( \begin{array}{c} n_{\text{total}} \\ n_{\text{heter}} \end{array} \right) (p_{aa})^{n_{aa}} (1 - p_{aa})^{n_{a\overline{a}}} \left( \frac{p_{aa}}{\theta} (1 - p_{aa})^{\tau - 1} \right) dp_{aa} = \left( \begin{array}{c} n_{\text{total}} \\ n_{\text{heter}} \end{array} \right) \frac{B(\theta + n_{\text{heter}} + \tau + n_{\text{homo}})}{B(\theta, \tau)}.$$

LITERATURE CITED


Table 1. Comparison of three adjustment factors under $\kappa = 0.5$

<table>
<thead>
<tr>
<th>Trial</th>
<th>Estimator</th>
<th>BIAS</th>
<th>SE</th>
<th>RMSE</th>
<th>Type I error</th>
<th>Power</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\hat{\kappa}_H$</td>
<td>0.0038</td>
<td>0.0103</td>
<td>0.0110</td>
<td>0.0350</td>
<td>0.8850</td>
</tr>
<tr>
<td>1</td>
<td>$\hat{\kappa}_G$</td>
<td>0.0013</td>
<td>0.0102</td>
<td>0.0103</td>
<td>0.0500</td>
<td>0.8650</td>
</tr>
<tr>
<td></td>
<td>$\hat{\kappa}_U$</td>
<td>0.0007</td>
<td>0.0101</td>
<td>0.0102</td>
<td>0.0350</td>
<td>0.9000 **</td>
</tr>
<tr>
<td></td>
<td>$\hat{\kappa}_H$</td>
<td>0.0016</td>
<td>0.0057</td>
<td>0.0060</td>
<td>0.0350</td>
<td>0.8850</td>
</tr>
<tr>
<td>2</td>
<td>$\hat{\kappa}_G$</td>
<td>-0.0008</td>
<td>0.0057</td>
<td>0.0058</td>
<td>0.0500</td>
<td>0.8650</td>
</tr>
<tr>
<td></td>
<td>$\hat{\kappa}_U$</td>
<td>-0.0010</td>
<td>0.0057</td>
<td>0.0058</td>
<td>0.0350</td>
<td>0.9000 **</td>
</tr>
<tr>
<td></td>
<td>$\hat{\kappa}_H$</td>
<td>0.0023</td>
<td>0.0044</td>
<td>0.0050</td>
<td>0.0350</td>
<td>0.8850</td>
</tr>
<tr>
<td>3</td>
<td>$\hat{\kappa}_G$</td>
<td>-0.0002</td>
<td>0.0044</td>
<td>0.0044</td>
<td>0.0500</td>
<td>0.8650</td>
</tr>
<tr>
<td></td>
<td>$\hat{\kappa}_U$</td>
<td>-0.0002</td>
<td>0.0044</td>
<td>0.0044</td>
<td>0.0350</td>
<td>0.9000 **</td>
</tr>
<tr>
<td></td>
<td>$\hat{\kappa}_H$</td>
<td>0.0328</td>
<td>0.0340</td>
<td>0.0473</td>
<td>0.0550</td>
<td>0.8750</td>
</tr>
<tr>
<td>4</td>
<td>$\hat{\kappa}_G$</td>
<td>0.0077</td>
<td>0.0296</td>
<td>0.0306</td>
<td>0.0550</td>
<td>0.8500</td>
</tr>
<tr>
<td></td>
<td>$\hat{\kappa}_U$</td>
<td>0.0015</td>
<td>0.0293</td>
<td>0.0294</td>
<td>0.0600</td>
<td>0.8750 **</td>
</tr>
<tr>
<td></td>
<td>$\hat{\kappa}_H$</td>
<td>0.0275</td>
<td>0.0187</td>
<td>0.0333</td>
<td>0.0450</td>
<td>0.8750</td>
</tr>
<tr>
<td>5</td>
<td>$\hat{\kappa}_G$</td>
<td>0.0027</td>
<td>0.0165</td>
<td>0.0167</td>
<td>0.0500</td>
<td>0.8550</td>
</tr>
<tr>
<td></td>
<td>$\hat{\kappa}_U$</td>
<td>0.0006</td>
<td>0.0167</td>
<td>0.0167</td>
<td>0.0350</td>
<td>0.8950 **</td>
</tr>
<tr>
<td></td>
<td>$\hat{\kappa}_H$</td>
<td>0.0247</td>
<td>0.0141</td>
<td>0.0284</td>
<td>0.0400</td>
<td>0.8800</td>
</tr>
<tr>
<td>6</td>
<td>$\hat{\kappa}_G$</td>
<td>0.0008</td>
<td>0.0125</td>
<td>0.0125</td>
<td>0.0500</td>
<td>0.8600</td>
</tr>
<tr>
<td></td>
<td>$\hat{\kappa}_U$</td>
<td>-0.0001</td>
<td>0.0127</td>
<td>0.0127</td>
<td>0.0350</td>
<td>0.9000 **</td>
</tr>
<tr>
<td></td>
<td>$\hat{\kappa}_H$</td>
<td>$\hat{\kappa}_G$</td>
<td>$\hat{\kappa}_U$</td>
<td>$\hat{\kappa}_H$</td>
<td>$\hat{\kappa}_G$</td>
<td>$\hat{\kappa}_U$</td>
</tr>
<tr>
<td>---</td>
<td>-----------------</td>
<td>-----------------</td>
<td>-----------------</td>
<td>-----------------</td>
<td>-----------------</td>
<td>-----------------</td>
</tr>
<tr>
<td>7</td>
<td>0.0030</td>
<td>0.0006</td>
<td>0.0001</td>
<td>0.0029</td>
<td>0.0004</td>
<td>0.0004</td>
</tr>
<tr>
<td></td>
<td>0.0100</td>
<td>0.0099</td>
<td>0.0099</td>
<td>0.0057</td>
<td>0.0057</td>
<td>0.0057</td>
</tr>
<tr>
<td></td>
<td>0.0104</td>
<td>0.0100</td>
<td>0.0099         *</td>
<td>0.0064</td>
<td>0.0057</td>
<td>0.0057</td>
</tr>
<tr>
<td></td>
<td>0.0650</td>
<td>0.0650</td>
<td>0.0400         *</td>
<td>0.0650</td>
<td>0.0650</td>
<td>0.0650</td>
</tr>
<tr>
<td></td>
<td>0.2950</td>
<td>0.3600</td>
<td>0.3150</td>
<td>0.2950</td>
<td>0.3600</td>
<td>0.3150</td>
</tr>
<tr>
<td>8</td>
<td>0.0029</td>
<td>0.0004</td>
<td>0.0004</td>
<td>0.0212</td>
<td>-0.0011</td>
<td>-0.0056</td>
</tr>
<tr>
<td></td>
<td>0.0057</td>
<td>0.0057</td>
<td>0.0057</td>
<td>0.0303</td>
<td>0.0269</td>
<td>0.0270</td>
</tr>
<tr>
<td></td>
<td>0.0064</td>
<td>0.0057</td>
<td>0.0057         *</td>
<td>0.0370</td>
<td>0.0269         *</td>
<td>0.0275</td>
</tr>
<tr>
<td></td>
<td>0.0650</td>
<td>0.0650</td>
<td>0.0400         *</td>
<td>0.0700</td>
<td>0.0650</td>
<td>0.0400</td>
</tr>
<tr>
<td></td>
<td>0.2950</td>
<td>0.3600</td>
<td>0.3150</td>
<td>0.3000</td>
<td>0.3600</td>
<td>0.3200</td>
</tr>
<tr>
<td>9</td>
<td>0.0035</td>
<td>0.0010</td>
<td>0.0010</td>
<td>0.0251</td>
<td>0.0004</td>
<td>-0.0022</td>
</tr>
<tr>
<td></td>
<td>0.0045</td>
<td>0.0044</td>
<td>0.0044</td>
<td>0.0183</td>
<td>0.0162</td>
<td>0.0126</td>
</tr>
<tr>
<td></td>
<td>0.0057</td>
<td>0.0046</td>
<td>0.0046         *</td>
<td>0.0311</td>
<td>0.0162         *</td>
<td>0.0125</td>
</tr>
<tr>
<td></td>
<td>0.0650</td>
<td>0.0650</td>
<td>0.0400         *</td>
<td>0.0700</td>
<td>0.0650</td>
<td>0.0400</td>
</tr>
<tr>
<td></td>
<td>0.2950</td>
<td>0.3600</td>
<td>0.3150</td>
<td>0.3000</td>
<td>0.3600</td>
<td>0.3200</td>
</tr>
</tbody>
</table>

*Denotes the estimator with minimum RMSE among three estimators

**Denotes the estimator with maximum power among three estimators
Table 2. Comparison of three adjustment factors under $\kappa = 1$

<table>
<thead>
<tr>
<th>Trial</th>
<th>Estimator</th>
<th>BIAS</th>
<th>SE</th>
<th>RMSE</th>
<th>Type I error</th>
<th>Power</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$\hat{\kappa}_H$</td>
<td>0.0057</td>
<td>0.0343</td>
<td>0.0348</td>
<td>0.0440</td>
<td>0.8680</td>
</tr>
<tr>
<td></td>
<td>$\hat{\kappa}_G$</td>
<td>0.0013</td>
<td>0.0339</td>
<td>0.0339</td>
<td>0.0460</td>
<td>0.8420</td>
</tr>
<tr>
<td></td>
<td>$\hat{\kappa}_U$</td>
<td>0.0004</td>
<td>0.0338</td>
<td>0.0338 *</td>
<td>0.0400</td>
<td>0.8960 **</td>
</tr>
<tr>
<td>2</td>
<td>$\hat{\kappa}_H$</td>
<td>0.0023</td>
<td>0.0203</td>
<td>0.0204</td>
<td>0.0350</td>
<td>0.8850</td>
</tr>
<tr>
<td></td>
<td>$\hat{\kappa}_G$</td>
<td>-0.0025</td>
<td>0.0201</td>
<td>0.0202</td>
<td>0.0500</td>
<td>0.8650</td>
</tr>
<tr>
<td></td>
<td>$\hat{\kappa}_U$</td>
<td>-0.0027</td>
<td>0.0199</td>
<td>0.0201 *</td>
<td>0.0350</td>
<td>0.9000 **</td>
</tr>
<tr>
<td>3</td>
<td>$\hat{\kappa}_H$</td>
<td>0.0038</td>
<td>0.0161</td>
<td>0.0165</td>
<td>0.0350</td>
<td>0.8850</td>
</tr>
<tr>
<td></td>
<td>$\hat{\kappa}_G$</td>
<td>-0.0011</td>
<td>0.0159</td>
<td>0.0159</td>
<td>0.0500</td>
<td>0.8650</td>
</tr>
<tr>
<td></td>
<td>$\hat{\kappa}_U$</td>
<td>-0.0013</td>
<td>0.0158</td>
<td>0.0159 *</td>
<td>0.0350</td>
<td>0.9000 **</td>
</tr>
<tr>
<td>4</td>
<td>$\hat{\kappa}_H$</td>
<td>0.0380</td>
<td>0.0910</td>
<td>0.1074</td>
<td>0.0750</td>
<td>0.8600</td>
</tr>
<tr>
<td></td>
<td>$\hat{\kappa}_G$</td>
<td>0.0079</td>
<td>0.0789</td>
<td>0.0793</td>
<td>0.1000</td>
<td>0.8350</td>
</tr>
<tr>
<td></td>
<td>$\hat{\kappa}_U$</td>
<td>0.0030</td>
<td>0.0758</td>
<td>0.0758 *</td>
<td>0.0850</td>
<td>0.8700 **</td>
</tr>
<tr>
<td>5</td>
<td>$\hat{\kappa}_H$</td>
<td>0.0421</td>
<td>0.0711</td>
<td>0.0826</td>
<td>0.0550</td>
<td>0.8650</td>
</tr>
<tr>
<td></td>
<td>$\hat{\kappa}_G$</td>
<td>-0.0042</td>
<td>0.0619</td>
<td>0.0620</td>
<td>0.0700</td>
<td>0.8500</td>
</tr>
<tr>
<td></td>
<td>$\hat{\kappa}_U$</td>
<td>-0.0044</td>
<td>0.0599</td>
<td>0.0601 *</td>
<td>0.0550</td>
<td>0.8800 **</td>
</tr>
<tr>
<td>6</td>
<td>$\hat{\kappa}_H$</td>
<td>0.0522</td>
<td>0.0566</td>
<td>0.0770</td>
<td>0.0450</td>
<td>0.8750</td>
</tr>
<tr>
<td></td>
<td>$\hat{\kappa}_G$</td>
<td>0.0031</td>
<td>0.0492</td>
<td>0.0493</td>
<td>0.0600</td>
<td>0.8500</td>
</tr>
<tr>
<td></td>
<td>$\hat{\kappa}_U$</td>
<td>0.0021</td>
<td>0.0477</td>
<td>0.0478 *</td>
<td>0.0500</td>
<td>0.8900 **</td>
</tr>
<tr>
<td></td>
<td>( \hat{\kappa}_H )</td>
<td>( \hat{\kappa}_G )</td>
<td>( \hat{\kappa}_U )</td>
<td>( \hat{\kappa}_H )</td>
<td>( \hat{\kappa}_G )</td>
<td>( \hat{\kappa}_U )</td>
</tr>
<tr>
<td>----</td>
<td>----------------</td>
<td>----------------</td>
<td>----------------</td>
<td>----------------</td>
<td>----------------</td>
<td>----------------</td>
</tr>
<tr>
<td>7</td>
<td>0.0067 0.0650 0.2950</td>
<td>0.0024 0.0332 0.0333 0.0650 0.3600 **</td>
<td>0.0015 0.0331 0.0332 * 0.0400 0.3150</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>0.0048 0.0203 0.0208 0.0650 0.2950</td>
<td>0.0001 0.0202 0.0202 0.0650 0.3600 **</td>
<td>-0.0001 0.0201 0.0201 * 0.0400 0.3150</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>0.0055 0.0161 0.0170 0.0650 0.2950</td>
<td>0.0005 0.0159 0.0159 0.0650 0.3600 **</td>
<td>0.0003 0.0158 0.0158 * 0.0400 0.3150</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>0.0569 0.1264 0.1386 0.0800 0.2750</td>
<td>0.0110 0.1100 0.1105 0.0790 0.2770</td>
<td>0.0029 0.1047 0.1048 * 0.0680 0.2970 **</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>0.0555 0.0737 0.0922 0.0700 0.3000</td>
<td>0.0066 0.0643 0.0646 0.0650 0.3650 **</td>
<td>0.0037 0.0618 0.0619 * 0.0400 0.3200</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>0.0499 0.0575 0.0761 0.0700 0.2950</td>
<td>-0.0003 0.0499 0.0499 0.0650 0.3600 **</td>
<td>-0.0010 0.0483 0.0483 * 0.0400 0.3200</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* denotes the estimator with minimum RMSE among three estimators

** denotes the estimator with maximum power among three estimators
Table 3. Comparison of three adjustment factors under $\kappa = 2$

<table>
<thead>
<tr>
<th>Trial</th>
<th>Estimator</th>
<th>BIAS</th>
<th>SE</th>
<th>RMSE</th>
<th>Type I error</th>
<th>Power</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$\hat{\kappa}_H$</td>
<td>0.0079</td>
<td>0.1596</td>
<td>0.1598</td>
<td>0.0600</td>
<td>0.8550</td>
</tr>
<tr>
<td></td>
<td>$\hat{\kappa}_G$</td>
<td>-0.0015</td>
<td>0.1564</td>
<td>0.1564</td>
<td>0.0950</td>
<td>0.8400</td>
</tr>
<tr>
<td></td>
<td>$\hat{\kappa}_U$</td>
<td>-0.0025</td>
<td>0.1546</td>
<td>0.1546 *</td>
<td>0.0850</td>
<td>0.8700 **</td>
</tr>
<tr>
<td>2</td>
<td>$\hat{\kappa}_H$</td>
<td>0.0072</td>
<td>0.0939</td>
<td>0.0941</td>
<td>0.0450</td>
<td>0.8750</td>
</tr>
<tr>
<td></td>
<td>$\hat{\kappa}_G$</td>
<td>-0.0026</td>
<td>0.0916</td>
<td>0.0916</td>
<td>0.0550</td>
<td>0.8500</td>
</tr>
<tr>
<td></td>
<td>$\hat{\kappa}_U$</td>
<td>-0.0031</td>
<td>0.0906</td>
<td>0.0906 *</td>
<td>0.0450</td>
<td>0.8850 **</td>
</tr>
<tr>
<td>3</td>
<td>$\hat{\kappa}_H$</td>
<td>0.0119</td>
<td>0.0728</td>
<td>0.0738</td>
<td>0.0450</td>
<td>0.8800</td>
</tr>
<tr>
<td></td>
<td>$\hat{\kappa}_G$</td>
<td>0.0016</td>
<td>0.0711</td>
<td>0.0704</td>
<td>0.0500</td>
<td>0.8550</td>
</tr>
<tr>
<td></td>
<td>$\hat{\kappa}_U$</td>
<td>0.0016</td>
<td>0.0703</td>
<td>0.0703 *</td>
<td>0.0400</td>
<td>0.8950 **</td>
</tr>
<tr>
<td>4</td>
<td>$\hat{\kappa}_H$</td>
<td>0.1003</td>
<td>0.4724</td>
<td>0.4830</td>
<td>0.2820</td>
<td>0.7240</td>
</tr>
<tr>
<td></td>
<td>$\hat{\kappa}_G$</td>
<td>0.0073</td>
<td>0.4423</td>
<td>0.4423</td>
<td>0.2620</td>
<td>0.7400</td>
</tr>
<tr>
<td></td>
<td>$\hat{\kappa}_U$</td>
<td>-0.0023</td>
<td>0.4322</td>
<td>0.4322 *</td>
<td>0.2840</td>
<td>0.7840 **</td>
</tr>
<tr>
<td>5</td>
<td>$\hat{\kappa}_H$</td>
<td>0.1079</td>
<td>0.2851</td>
<td>0.3048</td>
<td>0.1350</td>
<td>0.8150</td>
</tr>
<tr>
<td></td>
<td>$\hat{\kappa}_G$</td>
<td>0.0109</td>
<td>0.2638</td>
<td>0.2640</td>
<td>0.1750</td>
<td>0.8150</td>
</tr>
<tr>
<td></td>
<td>$\hat{\kappa}_U$</td>
<td>0.0064</td>
<td>0.2563</td>
<td>0.2564 *</td>
<td>0.1650</td>
<td>0.8350 **</td>
</tr>
<tr>
<td>6</td>
<td>$\hat{\kappa}_H$</td>
<td>0.0989</td>
<td>0.2232</td>
<td>0.2441</td>
<td>0.0900</td>
<td>0.8400</td>
</tr>
<tr>
<td></td>
<td>$\hat{\kappa}_G$</td>
<td>0.0099</td>
<td>0.2037</td>
<td>0.2037</td>
<td>0.1300</td>
<td>0.8250</td>
</tr>
<tr>
<td></td>
<td>$\hat{\kappa}_U$</td>
<td>-0.0011</td>
<td>0.1968</td>
<td>0.1968 *</td>
<td>0.1100</td>
<td>0.8500 **</td>
</tr>
<tr>
<td></td>
<td>$\hat{\kappa}_H$</td>
<td>$\hat{\kappa}_G$</td>
<td>$\hat{\kappa}_U$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>-----------------</td>
<td>-----------------</td>
<td>-----------------</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>0.0153 0.1630 0.1637 0.0700 0.3000</td>
<td>0.0060 0.1591 0.1592 0.0650 0.3650</td>
<td>0.0057 0.1578 0.1579 0.0400 0.3150</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>0.0097 0.0930 0.0935 0.0700 0.2950</td>
<td>-0.0000 0.0914 0.0914 0.0650 0.3600</td>
<td>-0.0003 0.0906 0.0906 0.0400 0.3150</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>0.0072 0.0708 0.0722 0.0650 0.2950</td>
<td>-0.0024 0.0703 0.0704 0.0650 0.3600</td>
<td>-0.0026 0.0697 0.0697 0.0400 0.3150</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>0.0850 0.4721 0.4797 0.1150 0.2350</td>
<td>-0.0168 0.4393 0.4396 0.1100 0.2950</td>
<td>-0.0326 0.4233 0.4245 0.1300 0.2800</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>0.1128 0.2882 0.3095 0.0650 0.2800</td>
<td>0.0131 0.2672 0.2675 0.0850 0.3550</td>
<td>0.0086 0.2581 0.2583 0.0450 0.3300</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>0.1004 0.2231 0.2446 0.0640 0.3120</td>
<td>0.0042 0.2041 0.2042 0.0640 0.3460</td>
<td>0.0019 0.1976 0.1976 0.0036 0.3160</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* denotes the estimator with minimum RMSE among three estimators

** denotes the estimator with maximum power among three estimators
Table 4. Comparisons of estimated allele frequencies from individual genotyping, the pooling experiment without adjustment, and the pooling experiments using different adjustments

<table>
<thead>
<tr>
<th>SNP</th>
<th>Individual genotyping</th>
<th>Without adjustment</th>
<th>Adjustment $\hat{\kappa}_H$</th>
<th>Adjustment $\hat{\kappa}_G$</th>
<th>Adjustment $\hat{\kappa}_U$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.067</td>
<td>0.058</td>
<td>0.073</td>
<td>0.072</td>
<td>0.070</td>
</tr>
<tr>
<td>2</td>
<td>0.217</td>
<td>0.034</td>
<td>0.040</td>
<td>0.050</td>
<td>0.051</td>
</tr>
<tr>
<td>3</td>
<td>0.233</td>
<td>0.163</td>
<td>0.227</td>
<td>0.235</td>
<td>0.236</td>
</tr>
<tr>
<td>4</td>
<td>0.250</td>
<td>0.357</td>
<td>0.327</td>
<td>0.323</td>
<td>0.321</td>
</tr>
<tr>
<td>5</td>
<td>0.383</td>
<td>0.183</td>
<td>0.283</td>
<td>0.281</td>
<td>0.285</td>
</tr>
<tr>
<td>6</td>
<td>0.433</td>
<td>0.194</td>
<td>0.355</td>
<td>0.353</td>
<td>0.359</td>
</tr>
</tbody>
</table>
Figure 1. The number of heterozygous individuals under different conditions.

(A) $\xi = 0.05$

(B) $\xi = 0.075$

(C) $\xi = 0.10$
Figure 2. Distribution of the number of genotyped individuals required to attain the required number of heterozygous individuals, assuming that the heterozygous genotype frequency is beta distributed. (A) Constant frequency and $n_{\text{heter}} = 8$; (B) Constant frequency and $n_{\text{heter}} = 16$; (C) Random effect and $n_{\text{heter}} = 8$; (D) Random effect and $n_{\text{heter}} = 16$. 

(A) 

(B) 

(C) 

(D)
Figure 3. Distribution of the number of heterozygous individuals after genotyping a fixed number of individuals $n_{total}$, assuming that the heterozygous genotype frequency is beta distributed. (A) Constant frequency and $n_{total} = 48$; (B) Constant frequency and $n_{total} = 96$; (C) Random effect and $n_{total} = 48$; (D) Random effect and $n_{total} = 96$. 

(A) 

(B) 

(C) 

(D)