Published in final edited form as:

Ann Hum Genet. 2008 May ; 72(Pt 3): 375–387. doi:10.1111/j.1469-1809.2007.00419.x.

Optimal two-stage design for case-control association analysis incorporating genotyping errors

Yijun Zuo^{1,†}, Guohua Zou^{2,3,†}, Jiexun Wang^{2,4}, Hongyu Zhao⁵, and Hua Liang^{3,*}

¹Department of Statistics and Probability, Michigan State University, East Lansing, MI 48824, USA

²Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100080, China

³Department of Biostatistics and Computational Biology, University of Rochester, Rochester, NY 14642, USA

⁴Department of Applied Mathematics and Physics, Graduate School of Informatics, Kyoto University, Kyoto 606-8501, Japan

⁵Department of Epidemiology and Public Health, Yale University School of Medicine, New Haven, CT 06520, USA

Abstract

Two-stage design is a cost effective approach for identifying disease genes in genetic studies and it has received much attention recently. In general, there are two types of two-stage designs that differ on the methods and samples used to measure allele frequencies in the first stage: (1) Individual genotyping is used in the first stage; (2) DNA pooling is used in the first stage. In this paper, we focus on the latter. Zuo et al. (2006) investigated statistical power of such a design, among other things, but the cost of study was not taken into account. The purpose of this paper is to study the optimal design under the given overall cost. We investigate how to allocate the resources to the two stages. Note that in addition to the measurement errors associated with DNA pooling, genotyping errors are also unavoidable with individual genotyping. Therefore, we discuss the optimal design combining genotyping errors associated with individual genotyping. The joint statistical distributions of test statistics in the first and second stages are derived. For a fixed cost, our results show that the optimal design requires no additional samples in the second stage but only that the samples in the first stage be re-used. When the second stage uses an entirely independent sample, however, the optimal design under a given cost depends on the population allele frequency and allele frequency difference between the case and control groups. For the current genotyping costs, we can roughly allocate 1/3 to 1/2 of the total sample size to the first stage for screening.

Keywords

DNA pooling; genotyping errors; individual genotyping; measurement errors; power; two-stage design

*Corresponding author: Department of Biostatistics and Computational Biology, University of Rochester, 601 Elmwood Avenue, Box 630, Rochester, NY 14642, U.S.A., hua_liang@urmc.rochester.edu.

[†]Equal contribution

Introduction

In recent years, case-control association studies, in comparison to linkage analysis, have enjoyed much popularity as a potentially more effective strategy to identify disease susceptibility genes (Risch 2000). Association studies can be conducted by using either the candidate gene approach or the genome-wide association approach. Candidate-gene association studies aim to identify genes/markers that show different distributions between the case and control groups motivated through biological knowledge and hypothesis. But currently our knowledge for biological mechanism is rather limited, and it is extremely difficult to predict which specific genes play a role in disease etiology. Therefore, candidate-gene association studies will likely miss important genes. However, with the completion of the human genome sequencing, the identification of millions of genetic variations, and the advances in genotyping technologies, genome-wide association studies have become a viable and potentially powerful approach to identifying disease genes. Such studies allow a comprehensive scan in the whole genome and so increase the possibility of detecting disease associated genes. Tens or hundreds of thousands of markers, however, are required for these studies. On the other hand, other factors, such as disease heterogeneity and gene-environment interactions, make disease gene identification a very challenging task, and so hundreds or even thousands of subjects may be needed to achieve sufficient statistical power. As a result, the total study cost would rise rapidly in spite of great improvements of genotyping technologies.

A cost-effective approach for detecting disease genes is to use a two-stage design, which has received much attention recently. There are two types of the two-stage designs depending on the methods/samples used to measure allele frequencies in the first stage: (1) individual genotyping is used in the first stage (Satagopan et al. 2002, 2004; Satagopan and Elston 2003; Lin 2006; Skol et al. 2006; Wang et al. 2006), hereafter referenced as Two-Stage Design for Individual Genotyping (TSD-IG); (2) DNA pooling is used in the first stage (Zuo et al. 2006), hereafter referenced as Two-Stage Design for DNA Pooling (TSD-P). In this paper, we will focus on the TSD-P. Zuo et al. (2006) investigated statistical power of such a design, among other things, but the study cost was not taken into account. The purpose of this paper is to study the optimal TSD-P with a focus on the overall study cost. We investigate how to allocate the resources to the two stages. The corresponding issues for the TSD-IG were considered by Satagopan et al. (2002) and Satagopan and Elston (2003).

Note that an accurate measurement of single-nucleotide polymorphism (SNP) frequencies in a sample is crucially important to successful association studies. Although individual genotyping accuracy has been greatly improved over the past decade, genotyping errors still inevitably occur (Buetow 1991; Shields et al. 1991; Akey et al. 2001; Gordon et al. 2001; Gordon et al. 2002). Many researchers have showed that genotyping errors can have large effects on the precision of allele frequency estimation and power of tests and so on (Gordon et al. 1999; Rice and Holmans 2003; Ott 2004; Zou and Zhao 2004). Therefore, we will discuss the optimal two-stage design incorporating genotyping errors associated with individual genotyping, in addition to the measurement errors associated with DNA pooling (see, for example, Barratt et al. 2002; Jawaid et al. 2002) considered in Zuo et al. (2006) for the TSD-P. Note that genotyping errors associated with individual genotyping were not taken into account by Satagopan et al. (2002) or Satagopan and Elston (2003) for the TSD-IG.

This article is organized as follows. We will first derive the formulas for calculating the power of the TSD-P incorporating genotyping errors, and then discuss the methods for finding the optimal designs under a given cost. Based on these results, we will present our numerical results. The technical details are provided in the Appendix.

Methods

We consider two alleles, A and a , at a candidate marker, whose frequencies are p and $q = 1 - p$, respectively. For simplicity, we consider a case-control study with n cases and n controls. Let X_i denote the number of allele A carried by the i th individual in the case group, and Y_i is similarly defined for the i th individual in the control group. Assuming Hardy-Weinberg Equilibrium (HWE), each X_i and Y_i has a value of 2, 1, 0 with respective probabilities p^2 , $2pq$ and q^2 under the null hypothesis of no association between the candidate marker and disease. When the candidate marker is associated with disease, we assume that the penetrances are f_2 for genotype AA , f_1 for genotype Aa , and f_0 for genotype aa . Note that these two alleles may be true functional alleles or may be in LD with true functional alleles. Under this genetic model, the probability of an individual having k copies of A among the cases, $m_k = P(X_i = k)$, and that among the controls, $m'_k = P(Y_i = k)$, are given by

$$\begin{aligned} m_0 &= \frac{q^2 f_0}{p^2 f_2 + 2pq f_1 + q^2 f_0}, \\ m_1 &= \frac{2pq f_1}{p^2 f_2 + 2pq f_1 + q^2 f_0}, \\ m_2 &= \frac{p^2 f_2}{p^2 f_2 + 2pq f_1 + q^2 f_0}, \\ m'_0 &= \frac{q^2 (1-f_0)}{p^2 (1-f_2) + 2pq (1-f_1) + q^2 (1-f_0)}, \\ m'_1 &= \frac{2pq (1-f_1)}{p^2 (1-f_2) + 2pq (1-f_1) + q^2 (1-f_0)}, \\ m'_2 &= \frac{p^2 (1-f_2)}{p^2 (1-f_2) + 2pq (1-f_1) + q^2 (1-f_0)}. \end{aligned}$$

One-stage designs

For convenience and useful reference, we first briefly summarize the test statistics and power expressions based on a one-stage design using either individual genotyping or DNA pooling. See Zou and Zhao (2004) or Zuo et al. (2006) for details.

(a) Individual genotyping—For individual genotyping, we assume that genotyping errors are introduced independently to each allele, and the error rate from the true allele A to the erroneous allele a is e_1 and from the true allele a to the erroneous allele A is e_2 . A more complete model for genotyping errors may be based on genotypes instead of alleles (Mote and Anderson 1965; Kang et al. 2004a; Kang et al. 2004b). Let n_A^* and n_U^* denote the observed numbers of allele A in the case group and control group, respectively. The statistic to test the association between the candidate marker and disease is

$$t_{ind} = \frac{(n_A^* - n_U^*) / (2n)}{\sqrt{\widehat{p}^* (1 - \widehat{p}^*) / n}},$$

where $\widehat{p}^* = (n_A^* + n_U^*) / (4n)$.

Like Risch and Teng (1998) and Satagopan and Elston (2003), we consider a one-sided test. With a significance level of α , the power of the test statistic t_{ind} is

$$\Phi \left(\frac{-z_\alpha \sqrt{\widehat{p}^* (1 - \widehat{p}^*) / (1 - e_1 - e_2)} + \sqrt{n} \mu^*}{\sigma^*} \right), \quad (1)$$

where Φ is the cumulative standard normal distribution function, z_α is the upper 100 α th percentile of the standard normal distribution, $\widehat{p}^* = (p_A^* + p_U^*) / 2$, $\mu^* = (p_A - p_U) / (1 - e_1 - e_2)$,

$$\sigma^{*2} = \frac{[p_A^*(1-p_A^*) - p_{A12}^*/4] + [p_U^*(1-p_U^*) - p_{U12}^*/4]}{(1-e_1-e_2)^2},$$

and

$$\begin{aligned} p_A^* &= p_{A11}^* + p_{A12}^*/2, \\ p_U^* &= p_{U11}^* + p_{U12}^*/2, \end{aligned}$$

with

$$\begin{aligned} p_{A11}^* &= (1-e_1)^2 m_2 + (1-e_1)e_2 m_1 + e_2^2 m_0, \\ p_{A12}^* &= 2(1-e_1)e_1 m_2 + [(1-e_1)(1-e_2) + e_1 e_2] m_1 + 2(1-e_2)e_2 m_0, \end{aligned}$$

and p_{U11}^* and p_{U12}^* are similarly defined for the control group and their values can be calculated by replacing m_i with m'_i ($i=0, 1, 2$) in the above formulas for p_{A11}^* and p_{A12}^* .

(b) DNA pooling—For DNA pooling, we consider m pools of cases and m pools of controls each having size s such that $n=ms$ and assume the following models relating the observed frequencies to the true frequencies of allele A in the sample:

$$\begin{aligned} \widehat{p}_{Ai}^{pool} &= p_{Ai} + u_i, \\ \widehat{p}_{Ui}^{pool} &= p_{Ui} + v_i, \end{aligned}$$

where p_{Ai} and p_{Ui} denote the true frequencies of allele A in the i th case group and control group, respectively, u_i and v_i are deviations with mean 0 and variance ε^2 and are assumed to be independent and normally distributed. Define

$$\widehat{p}_A^{pool} = \frac{1}{m} \sum_{i=1}^m \widehat{p}_{Ai}^{pool},$$

and

$$\widehat{p}_U^{pool} = \frac{1}{m} \sum_{i=1}^m \widehat{p}_{Ui}^{pool}.$$

The following statistic can be used to test genetic association based on DNA pooling data:

$$t_{pool} = \frac{\widehat{p}_A^{pool} - \widehat{p}_U^{pool}}{\sqrt{\frac{\widehat{p}_{pool}(1-\widehat{p}_{pool})}{n} + \frac{2\varepsilon^2}{m}}},$$

$$\text{where } \widehat{p}_{pool} = \frac{1}{2} (\widehat{p}_A^{pool} + \widehat{p}_U^{pool}).$$

If we use a one-sided test and a significance level of α , then the power of the test statistic is t_{pool} is

$$\Phi \left(\frac{-z_{\alpha} \sqrt{\frac{\tilde{p}(1-\tilde{p})}{n} + \frac{2\varepsilon^2}{m}} + \mu}{\sqrt{\frac{\sigma^2}{n} + \frac{2\varepsilon^2}{m}}} \right),$$

where $\tilde{p} = \mu/2 + m'_2 + m'_1/2$, $\mu = m_2 + \frac{1}{2}m_1 - m'_2 - \frac{1}{2}m'_1$, and $\sigma^2 = \frac{1}{4}[4m_2 + m_1 - (2m_2 + m_1)^2 + 4m'_2 + m'_1 - (2m'_2 + m'_1)^2]$.

Two-stage designs

We consider the following two-stage design. In the first stage, we use pooled DNA data to test all M (a known number) markers, and assume the number of statistically significant markers is M_1 which is unknown and needs to be estimated. In the second stage, we use individual genotype data to test the M_1 promising markers. In practical genetic epidemiological studies, the sample in the first stage can be re-used in the second stage, and this introduces some correlation between the two test statistics, t_{pool} and t_{ind} . In Zuo et al. (2006), this two-stage scheme was called *the two-stage dependent design*. On the other hand, we may use two separate samples in the two stages with one sample used for screening and the other used for confirmation. In this scenario, the two test statistics, t_{pool} and t_{ind} are independent. Zuo et al. (2006) called such a two-stage scheme *the two-stage independent design*. In the following, we first consider the situation of the two-stage dependent design.

The joint distribution of test statistics in the first and second stages for the two-stage dependent design

In the first stage, i.e., the DNA pooling stage, we consider m pools of cases and m pools of controls each having size s such that $n=ms$. In the second stage, i.e., the individual genotyping stage based on a selected set of markers, in addition to the n cases and n controls used in the pooling stage, we also consider an additional case sample of size n_a and an additional control sample of size n_a .

Under the null hypothesis H_0 , we can write approximately,

$$t_{pool} = \frac{\sqrt{\frac{pq}{n}} \xi_0 + w}{\sqrt{\frac{pq}{n} + \frac{2\varepsilon^2}{m}}},$$

and

$$t_{ind} = \sqrt{\frac{n}{n+n_a}} \cdot \xi_0^* + \sqrt{\frac{n_a}{n+n_a}} \cdot \eta_0^*,$$

where $(\xi_0, \xi_0^*)^T \sim N(0, \Sigma_0)$ (T means transpose) with

$$\Sigma_0 = \begin{pmatrix} 1 & \frac{(1-e_1-e_2)\sqrt{pq}}{\sqrt{p^*(1-p^*)}} \\ \frac{(1-e_1-e_2)\sqrt{pq}}{\sqrt{p^*(1-p^*)}} & 1 \end{pmatrix},$$

$$p^* = (1 - e_1)p + e_2q, \eta_0 \sim N(0, 1)$$

$$, w = \bar{u} - \bar{v} \sim N\left(0, \frac{2\varepsilon^2}{m}\right)$$

$$, \bar{u} = \frac{1}{m} \sum_{i=1}^m u_i, \bar{v}$$

and $= \frac{1}{m} \sum_{i=1}^m v_i$, and $(\xi_0, \xi_0^*)^T, \eta_0^*$ and w are mutually independent (see the Appendix for the strict proof under the alternative hypothesis H_1). So under the null hypothesis of no association, $(t_{pool}, t_{ind})^T$ has an approximate joint bivariate normal distribution $N(0, \Sigma_0^*)$, where

$$\Sigma_0^* = \begin{pmatrix} 1 & \frac{(1-e_1-e_2)pq}{\sqrt{(n+n_u)p^*(1-p^*)} \cdot \sqrt{pq/n+2\varepsilon^2/m}} \\ \frac{(1-e_1-e_2)pq}{\sqrt{(n+n_u)p^*(1-p^*)} \cdot \sqrt{pq/n+2\varepsilon^2/m}} & 1 \end{pmatrix}.$$

This distribution involves the parameters p, e_1, e_2 , and ε , which are either known or can be easily estimated (Bansal et al. 2002; Rice and Holmans 2003; Zou and Zhao 2003; Gordon et al. 2004; Lai et al. 2007; Tintle et al. 2007). Thus, the determination of the critical values for the two-stage dependent design is available.

Under the genetic model discussed above, we can write approximately,

$$t_{pool} = \frac{\sqrt{\frac{\sigma^2}{n}} \xi_1 + w}{\sqrt{\frac{\tilde{p}(1-\tilde{p})}{n} + \frac{2\varepsilon^2}{m}}},$$

and

$$t_{ind} = \frac{\sqrt{\frac{n}{n+n_u}} \sigma^* \cdot \xi_1^* + \sqrt{\frac{n_u}{n+n_u}} \sigma^* \cdot \eta_1^*}{\sqrt{\tilde{p}^*(1-\tilde{p}^*)}},$$

where $(\xi_1, \xi_1^*)^T \sim N(\sqrt{n}\tilde{\mu}, \Sigma_1)$ with

$$\tilde{\mu} = (\mu/\sigma, \mu^*(1 - e_1 - e_2)/\sigma^*)^T, \\ \Sigma_1 = \begin{pmatrix} 1 & \frac{\sigma_{12}^*}{4\sigma\sigma^*} \\ \frac{\sigma_{12}^*}{4\sigma\sigma^*} & (1 - e_1 - e_2)^2 \end{pmatrix},$$

and

$$\begin{aligned}\sigma_{12}^* &= \sigma_{A12}^* + \sigma_{U12}^*, \\ \sigma_{A12}^* &= 4m_2(1 - e_1) + m_1(1 - e_1 + e_2) - (2m_2 + m_1)(2p_{A11}^* + p_{A12}^*), \\ \sigma_{U12}^* &= 4m'_2(1 - e_1) + m'_1(1 - e_1 + e_2) - (2m'_2 + m'_1)(2p_{U11}^* + p_{U12}^*),\end{aligned}$$

and $\eta_1^* \sim N(\sqrt{n_a} \mu^*(1 - e_1 - e_2)/\sigma^*, (1 - e_1 - e_2)^2)$, and $(\xi_1, \xi_1^*)^T, \eta_1^*$ and w are mutually independent. Thus, under the alternative hypothesis H_1 , $(t_{pool}, t_{ind})^T$ has an approximate joint bivariate normal distribution $N(\tilde{\mu}^*, \Sigma_1^*)$ (see the Appendix for the details of proof), where

$$\tilde{\mu}^* = \left(\frac{\mu}{\sqrt{\tilde{p}(1 - \tilde{p})/n + 2\varepsilon^2/m}}, \frac{\sqrt{n + n_a} \mu^*(1 - e_1 - e_2)}{\sqrt{\tilde{p}^*(1 - \tilde{p}^*)}} \right)^T,$$

and

$$\Sigma_1^* = \begin{pmatrix} \frac{\sigma^2/n + 2\varepsilon^2/m}{\tilde{p}(1 - \tilde{p})/n + 2\varepsilon^2/m} & \frac{\sigma_{12}^*}{4\sqrt{(n + n_a)\tilde{p}^*(1 - \tilde{p}^*)} \cdot \sqrt{\tilde{p}(1 - \tilde{p})/n + 2\varepsilon^2/m}} \\ \frac{\sigma_{12}^*}{4\sqrt{(n + n_a)\tilde{p}^*(1 - \tilde{p}^*)} \cdot \sqrt{\tilde{p}(1 - \tilde{p})/n + 2\varepsilon^2/m}} & \frac{\sigma_{12}^*}{\frac{\sigma^{*2}(1 - e_1 - e_2)^2}{\tilde{p}^*(1 - \tilde{p}^*)}} \end{pmatrix}.$$

This distribution involves the seven parameters $p, e_1, e_2, \varepsilon, f_2, f_1$, and f_0 . The first four are either known or can be readily estimated as in the situation of H_0 . The last three, however, are generally unknown and not so easy to estimate.

Power for two-stage designs

For the given sample size n in the first stage, if we assume that one of the significance level α_1 and power $1 - \beta_1$ of the first stage is known, then we can determine a critical value k_1 by solving $\alpha_1 = P(t_{pool} > k_1 | H_0)$ or $1 - \beta_1 = P(t_{pool} > k_1 | H_1)$. Thus, for the overall significance level α and additional sample size n_a , we can determine the critical value k_2 in the second stage by solving

$$\alpha/M = P(t_{pool} > k_1, t_{ind} > k_2 | H_0) = \int_{k_1}^{\infty} \int_{k_2}^{\infty} h_0(x, y) dx dy,$$

where $h_0(x, y)$ is the density function of $(t_{pool}, t_{ind})^T$ under H_0 , which is given by

$$h_0(x, y) = \frac{1}{2\pi \sqrt{|\Sigma_0^*|}} \exp \left\{ -\frac{1}{2} (x, y) \Sigma_0^{*-1} \begin{pmatrix} x \\ y \end{pmatrix} \right\}$$

with $|\Sigma_0^*|$ being the determinant of the matrix Σ_0^* , and Σ_0^{*-1} being the inverse of Σ_0^* . Here for M markers, we have used Bonferroni correction for each marker test.

The power of the two-stage design for detecting a disease associated marker is then given by

$$1 - \beta = P(t_{pool} > k_1, t_{ind} > k_2 | H_1) \\ = \int_{k_1}^{\infty} \int_{k_2}^{\infty} h_1(x, y) dx dy,$$

where $h_1(x, y)$ is the density function of $(t_{pool}, t_{ind})^T$ under H_1 , which is given by

$$h_1(x, y) = \frac{1}{2\pi \sqrt{|\Sigma_1^*|}} \exp \left\{ -\frac{1}{2} ((x, y) - \tilde{\mu}^{*T}) \Sigma_1^{*-1} \begin{pmatrix} x \\ y \end{pmatrix} - \tilde{\mu}^* \right\}.$$

For the two-stage independent design, the Type I error rate and statistical power are simply the products of those in both stages, respectively. That is,

$$P(t_{pool} > k_1, t_{ind} > k_2 | H_0) = P(t_{pool} > k_1 | H_0) \cdot P(t_{ind} > k_2 | H_0),$$

and

$$P(t_{pool} > k_1, t_{ind} > k_2 | H_1) = P(t_{pool} > k_1 | H_1) \cdot P(t_{ind} > k_2 | H_1).$$

The optimal choice of the parameters in two-stage design for a fixed cost

In the above discussion, the sample sizes used in both stages are fixed in the design which implies that the cost is not our main concern. If the total cost is the major constraint which is often the situation in practical studies, how do we optimize the two-stage design? That is, given the total cost and overall significance level, how do we choose the sample size in the first stage for screening, the sample size in the second stage for confirmation, the significance levels in the two stages, and the power in the first stage so that the power of the two-stage design is maximized?

Now let C be the total cost, C_1 be the cost of recruiting an individual, C_{pool} be the cost of measuring allele frequency at a single marker for a DNA pool, C_{ind} be the cost of genotyping a single marker for an individual, and C_0 be the other cost associated with a study, such as infrastructure cost that does not scale with the study size. Then we have

$$C = C_0 + C_1 \cdot 2(n + n_a) + C_{pool} \cdot 2mM + C_{ind} \cdot 2(n + n_a)M_1 \quad (2)$$

for the two-stage dependent design, and

$$C = C_0 + C_1 \cdot 2(n + n_a) + C_{pool} \cdot 2mM + C_{ind} \cdot 2n_a M_1 \quad (3)$$

for the two-stage independent design with the sample size of n_a in the second stage, where M is the number of entire markers, and M_1 is the number of markers selected from the first

stage as before. As in Satagopan and Elston (2003) and Satagopan et al. (2004), the expected number of markers selected from the first stage is

$$\overline{M}_1 = (M - K)\alpha_1 + K(1 - \beta_1),$$

where K is the number of disease associated markers which is unknown. Thus, the cost functions in (2) and (3) can be replaced by

$$\begin{aligned} C = & C_0 + C_1 \cdot 2(n \\ & + n_a) + C_{pool} \cdot 2mM \\ & + C_{ind} \cdot 2(n \\ & + n_a)[(M \\ & - K)\alpha_1 \\ & + K(1 - \beta_1)], \end{aligned} \quad (4)$$

and

$$\begin{aligned} C = & C_0 + C_1 \cdot 2(n \\ & + n_a) + C_{pool} \cdot 2mM \\ & + C_{ind} \cdot 2n_a[(M - K)\alpha_1 \\ & + K(1 - \beta_1)], \end{aligned} \quad (5)$$

respectively.

Denote the critical values in the first and second stages by k_1 and k_2 , respectively. Then we have the following constraints on the unknown parameters α_1 and k_1 , and $1 - \beta_1$, k_1 and n :

$$\alpha_1 = 1 - \Phi(k_1), \quad (6)$$

and

$$1 - \beta_1 = \Phi\left(\frac{-k_1 \sqrt{\frac{\tilde{p}(1-\tilde{p})}{n} + \frac{2\varepsilon^2}{m}} + \mu}{\sqrt{\frac{\sigma^2}{n} + \frac{2\varepsilon^2}{m}}}\right). \quad (7)$$

Further, for an overall fixed significance level α ,

$$\alpha/M = \int_{k_1}^{\infty} \int_{k_2}^{\infty} h_0(x, y) dx dy \quad (8)$$

for the two-stage dependent design, and

$$\alpha/M = \alpha_1 \cdot [1 - \Phi(k_2)] \quad (9)$$

for the two-stage independent design.

Our goal is to maximize the statistical power

$$1 - \beta = \int_{k_1}^{\infty} \int_{k_2}^{\infty} h_1(x, y) dx dy$$

for the two-stage dependent design, or

$$1 - \beta = (1 - \beta_1) \cdot \Phi \left(\frac{-k_2 \sqrt{\tilde{p}^*(1 - \tilde{p}^*)} / (1 - e_1 - e_2) + \sqrt{n_a} \mu^*}{\sigma^*} \right)$$

for the two-stage independent design both as the functions of the unknown parameters ($n, n_a, k_1, k_2, \alpha_1, \beta_1$) for a given cost C and overall significance level α .

The calculation procedure of optimal parameters

The above two maximization problems, denoted as P_1 and P_2 for the two-stage dependent design and two-stage independent design, respectively, are the Mixed Integer Nonlinear Programming (MINLP) with discrete and continuous variables, nonlinearity in the objective functions and constraints, and hence difficult to solve. In order to obtain their global solutions, we adopt a simple and intuitive method. We illustrate our procedure of calculation using the two-stage dependent design as follows (The situation of the two-stage independent design is similar):

Step 1: We convert the original problem (P_1) with the parameters $n, n_a, k_1, k_2, \alpha_1$ and β_1 into the problem (denoted by P_1') with the parameters n and β_1 by the constraints (4), (6)–(8);

Step 2: We divide the problem (P_1') into n nonlinear optimization sub-problems $\{P_1^{(1)}, \dots, P_1^{(n)}\}$ with the parameter β_1 ;

Step 3: We solve each sub-problem ($P_1^{(i)}$) by using the existing function “fminbnd” in Matlab software, which finds minimum of single-variable function in a fixed interval, and write down its minimal objective value $f^{*(i)}$ and its corresponding solution $\beta_1^{*(i)}$;

Step 4: We obtain the global minimum of the objective function $f^* = \min \{f^{*(i)}, i = 1, 2, \dots, n\}$ and the corresponding global solution $\beta_1^* = \arg \min_{1 \leq i \leq n} \{f^{*(i)}\}$.

Results

To investigate the statistical power of the two-stage design in the presence of genotyping errors, we set the sample size in the first stage to be $n = 500$, and the supplemental sample size in the second stage to be $n_a = 500$. Note that the main purpose in the first stage is to screen for the truly associated markers, and so we hope that the probability of the truly associated markers being included is large. Therefore, as in Zuo et al. (2006), we set the power to be 95% in the pooling stage. The significance level of the two-stage design for a single marker test is taken to be $\alpha / M = 5 \times 10^{-8}$, a level suggested by Risch and Merikangas (1996) for the genome-wide association studies. Similar to the situation in the absence of genotyping errors, the power of the two-stage design in the presence of genotyping errors depends on the genetic model and population allele frequency through almost only the population allele frequency and allele frequency difference between the cases and controls (data not shown). So for specificity, we consider the multiplicative model (i.e., $f_1^2 = f_2 f_0$) and set f_0 as 0.01 in the following calculations.

Table 1 gives the statistical power of the two-stage dependent design that a disease-associated marker is detected under various error rates. It is easily seen that for almost all the situations, the impact of genotyping errors on detecting a disease-associated marker with smaller allele frequency difference between the cases and controls is larger. That is, to detect a disease-associated marker with large allele frequency difference between the cases and controls, the impact of genotyping errors is small. This observation is consistent with the results of Gordon and Finch (2005), who pointed out that one way to minimize the effect of genotyping errors is to design studies with higher power initially (see Figure 4 of their paper). The impact of unequal error rates from one allele to another allele is between those of equal error rates. Similar patterns are observed for the two-stage independent design (Table 2). As in the situation of no genotyping errors in the second stage (Zuo et al. 2006), the effect of the measurement error rate ε on the power is not serious, especially for the two-stage dependent design (data not shown). The same phenomenon is observed when we use multiple pools: the use of multiple pools will not significantly increase the power of the two-stage design, especially for the two-stage dependent design (data not shown). This can be expected because the benefit of using multiple pools is to reduce the measurement errors which have no large effect on the power especially for the two-stage dependent design as mentioned above. In Table 1 and Table 2, the power in the first stage is set as 0.95. We have also considered the effect of various $1 - \beta_1$. In general, the effect of $1 - \beta_1$ on the power of the two-stage design depends on the population allele frequency p and the allele frequency difference between the cases and controls $p_A - p_U$. In most situations, higher power can be obtained by increasing $1 - \beta_1$, especially for the big values of $p_A - p_U$ (data not shown). As for the optimal value of $1 - \beta_1$, it can be found by maximizing the power of the two-stage design for the given cost or minimizing the cost for the given power.

Comparing the two-stage dependent and two-stage independent designs, we see from Table 1 and Table 2 that the two-stage dependent design has much higher power, of course, at the price of extra genotyping in the second stage. On the other hand, if we use the sample size of $n_a = 1000$ in the second stage for the two-stage independent design which is the same as that for the two-stage dependent design, then the two-stage independent design will have higher power (data not shown). Clearly, in this situation, the extra sample collection cost is needed for the two-stage independent design. Noting that the extra genotyping cost for the two-stage dependent design is generally much lower than the extra sample collection cost for the two-stage independent design, the two-stage dependent design would generally have higher power. Of course, a fair comparison should base on the same total cost (see below).

To see the optimal choice of the parameters in the two-stage design for a fixed cost, we assume the total number of markers is $M = 10^6$, the number of the truly disease-associated markers is $K = 1$, and the number of pool pairs is $m = 1$. Further, we let the total cost $C = 2 \times 10^6$ (Unit: US\$), the cost of recruiting an individual $C_1 = 2000$, the cost of genotyping a single marker for an individual $C_{ind} = 0.02$, the cost of measuring allele frequency at a single marker for a DNA pool $C_{pool} = 0.03$, the other cost $C_0 = 0$, the genotyping error rates $e_1 = e_2 = 0.001$, and the measurement error rate $\varepsilon = 0.01$. The calculation results for various population allele frequencies and allele frequency differences between the cases and controls are summarized in Table 3 for the two-stage dependent design and Table 4 for the two-stage independent design. It is clear from Table 3 that under the given total cost, to obtain the optimal design with the highest power, no additional sample is substantially needed at the second stage for the two-stage dependent design. This means that for the two-stage dependent design, all individuals should be used at both stages. We have also considered other cost settings, e.g. $C = 5 \times 10^5$, $C_1 = 200$, $C_0 = 0$, $C_{ind} = 0.02$, $C_{pool} = 0.03$, and also confirmed our findings (Table 5). For the two-stage independent design, the optimal allocation of sample size depends on the cost and the allele frequency difference between the case and control groups. Roughly, for the current

genotyping costs and error rates, we recommend to allocate 1/3 to 1/2 of the total sample size to the first stage for the two-stage independent design.

Unlike the situation where the sample sizes are fixed in the two stages (which implies that the cost is not our main concern), when the cost is given, both the measurement error rate ϵ and the number of pool pairs m can have large effect on the power of the two-stage design. For example, for the two-stage dependent design, Table 5 shows that the increase of ϵ from 0.01 to 0.03 substantially reduces the power. This seems to be in contradiction with our previous observation when the sample sizes are fixed in the two stages. In fact, as Zuo et al. (2006) pointed out, for a two-stage design, measurement errors have a large impact only on the first stage. This means that the number of the promising markers M_1 will be seriously affected by ϵ . Now, M_1 is an important part in the cost function C , it is therefore not difficult to understand why the effect of ϵ on the power could be large. Such an effect would become small when multiple pools are used (data not shown). As for the effect of multiple pools on the power, it is observed that using multiple pools often leads to reduced power (Table 3 and Table 6) which looks like unusual. An intuitive explanation is that the use of multiple pools will greatly increase the cost of measuring allele frequency in the first stage, and so reduce the sample sizes at both stages, although it can reduce the measurement error rate ϵ .

For a fixed cost, it is now clear from Table 3 and Table 4 that the two-stage dependent design has higher power than the two-stage independent design. This is in fact not surprising because all of the sample information is used in the second stage for the two-stage dependent design but not for the two-stage independent design.

It should be noted that the error model for individual genotyping we considered here is a simple allele error model. Such a model means that the loci in HWE before the errors are introduced will still be in HWE after the errors are introduced. However, departure from HWE is often used as a way to detect genotyping errors (Hosking et al. 2004; Leal 2005; Cox and Kraft 2006). So in the following, we investigate the robustness of our methods to the departure from HWE through simulations. We consider a general error model (Kang et al. 2004a; Kang et al. 2004b) which is on genotypes but not on alleles and given in Table 7. Some special situations of the model can be found in Sobel et al. (2002) and Douglas et al. (2002). In our simulations, we consider two groups of the values of error rates: “ $\epsilon_{21} = 0.01$, $\epsilon_{20} = 0.02$, $\epsilon_{12} = 0.03$, $\epsilon_{10} = 0.02$, $\epsilon_{02} = 0.01$, $\epsilon_{01} = 0.02$ ” and “ $\epsilon_{21} = 0.1$, $\epsilon_{20} = 0.2$, $\epsilon_{12} = 0.3$, $\epsilon_{10} = 0.2$, $\epsilon_{02} = 0.1$, $\epsilon_{01} = 0.2$ ”. We first generate a population with the size of 5×10^6 assuming the population allele frequency $p = 0.2$ under HWE, then we use the penetrances f_2, f_1 and f_0 to simulate the disease status of each individual and thus we obtain the case subpopulation and control subpopulation. We randomly draw $n = n_a = 500$ cases from the case subpopulation and $n = n_a = 500$ controls from the control subpopulation. Using the general genotyping error model in Table 7, we generate the genotype data with errors for the samples from which we can estimate the allele error rates e_1 and e_2 . By making use of our test statistics, the empirical Type I error rate and power can be calculated. Note that to see whether our method leads to a nominal level, a very large number of simulations is needed if we set $\alpha / M = 5 \times 10^{-8}$. So we consider the nominal level of $\alpha / M = 0.05$ instead. The empirical Type I error rate is the proportion of significant replicates out of the total number of replicates under H_0 . The empirical power is the proportion of significant replicates out of the total number of replicates under H_1 . Based on 10,000 replicates, our results for the empirical Type I error rate and empirical power are summarized in Table 8 when the two-stage dependent design is used (The results are similar for the two-stage independent design). It can be seen that the empirical Type I error rates and empirical power are generally close to the nominal level of 0.05 and the theoretical power calculated by our asymptotic normal formulas, respectively. While we also observed a slight increase in Type I error rate for extremely large genotyping error rates, we comment that, in practice most SNP

genotyping error rates are below 0.005. So the results of Table 8 indicate that when genotyping error rates are this low, our methods are robust to deviations from HWE.

Discussion

In this paper, we have investigated the optimization of the TSD-P incorporating genotyping errors in genetic association studies. For the same total cost, it was observed that the two-stage dependent design is more powerful than the two-stage independent design. Further, under the fixed cost, for the two-stage dependent design, the optimal design that leads to the highest power is to use the whole sample in both stages. This contrasts with the TSD-IG in Skol et al. (2006) in which it is best to allocate 1/3 to 1/2 of the sample to the first stage. An intuitive explanation is that pooling costs are the same regardless of how many individuals are in the pool and hence more individuals can be pooled together in the first stage for the TSD-P. For the two-stage independent design, the optimal design under a given cost depends on the population allele frequency and allele frequency difference between the case and control groups. For the current genotyping costs and error rates, we can roughly allocate 1/3 to 1/2 of the total sample size to the first stage for screening.

The impact of genotyping errors on the two-stage design has also been considered. For both the two-stage dependent and independent designs, we found that the impact of genotyping errors on detecting a disease-associated marker with smaller allele frequency difference between the cases and controls is larger. Also, the impact of unequal genotyping error rates from one allele to another allele is between those of equal error rates. When the sample sizes at both stages are fixed, i.e., the cost is not a main factor we concern, the effects of both the measurement errors associated with pooled DNA and the number of pool pairs on the power of the two-stage design are not large, especially for the two-stage dependent design. When the cost is our main concern, however, both the measurement errors and the number of pool pairs can have large effects on the power of the two-stage design: the increase of measurement errors can substantially decrease the power, and the use of multiple pools often leads to reduced power. Our suggestions are therefore that when the cost is given, we should reduce the measurement errors in the first stage but need not to form too many pools.

Although the error model we considered is a simple allele error model for individual genotyping, our simulations show that our methods are generally robust to the departure from HWE. On the other hand, we have used the two different error models for the DNA pooling and individual genotyping technologies. An interesting relationship between them can be built which is available from the authors upon request.

In this article, we have optimized the two-stage design based on a given cost. Clearly, optimizing the two-stage design for a given power is also an interesting topic. The related problem is our ongoing work.

Throughout the paper, we have considered genetic association tests between diseases and genes on the basis of single-marker and used the Bonferroni correction to deal with the testing problem of multiple markers. But Bonferroni method is known to be conservative, especially when the markers under study are in LD (Malley et al. 2002), because in this situation, it would ignore the correlations among the loci and this may lead to detect only those loci with large marginal genetic effect but not those with small ones. On the other hand, multiple loci may contribute to disease susceptibility through complex interactions. Therefore, how to develop effective methods to test multiple markers simultaneously warrants further research.

In practical genetic epidemiological studies, missing genotype data are unavoidable. For example, the missing rates for individual genotyping are generally between 5 to 10% across

different platforms. The research on the two-stage designs in the presence of missing data is no doubt an interesting topic for future research.

In the above two-stage independent design, the second stage data is used alone. This is a traditional method which does not make full use of the data information from the first stage. Recently, for the TSD-IG, Skol et al. (2006) proposed a joint analysis strategy which combines the full data information from the first stage in the second stage analysis and so leads to higher power to detect genetic association. For the two-stage independent design considered here, how to conduct similar joint analysis is a future research topic.

Acknowledgments

The authors are grateful to the two reviewers for their constructive comments and suggestions which led to substantial improvements of the original paper. This work was supported in part by grants DMS0234078 from the National Science Foundation (to Y. Zuo), Nos. 70625004 and 70221001 from the National Natural Science Foundation of China (to G. Zou), GM59507 from the National Institutes of Health (to H. Zhao), and AI62247-01 and AI59773 from the National Institute of Allergy and Infectious Diseases (to H. Liang).

References

1. Akey JM, Zhang K, Xiong M, Doris P, Jin L. The effect that genotyping errors have on the robustness of common linkage-disequilibrium measures. *Am J Hum Genet* 2001;68:1447–1456. [PubMed: 11359212]
2. Bansal A, van den Boom D, Kammerer S, Honisch C, Adam G, Cantor CR, Kleyn P, Braun A. Association testing by DNA pooling: an effective initial screen. *Proc Natl Acad Sci USA* 2002;99:16871–16874. [PubMed: 12475937]
3. Barratt BJ, Payne F, Rance HE, Nuland S, Todd JA, Clayton DG. Identification of the sources of error in allele frequency estimations from pooled DNA indicated an optimal experimental design. *Ann Hum Genet* 2002;66:393–405. [PubMed: 12485472]
4. Buetow KH. Influence of aberrant observations on high-resolution linkage analysis outcomes. *Am J Hum Genet* 1991;49:985–994. [PubMed: 1928104]
5. Cox DG, Kraft P. Quantification of the power of Hardy-Weinberg equilibrium testing to detect genotyping error. *Hum Hered* 2006;61:10–14. [PubMed: 16514241]
6. Douglas JA, Skol AD, Boehnke M. Probability of detection of genotyping errors and mutations as inheritance inconsistencies in nuclear- family data. *Am J Hum Genet* 2002;70:487–495. [PubMed: 11791214]
7. Gordon D, Finch SJ. Factors affecting statistical power in the detection of genetic association. *J Clin Invest* 2005;115:1408–1418. [PubMed: 15931375]
8. Gordon D, Finch SJ, Nothnagel M, Ott J. Power and sample size calculations for case-control genetic association tests when errors are present: application to single nucleotide polymorphisms. *Hum Hered* 2002;54:22–33. [PubMed: 12446984]
9. Gordon D, Heath SC, Liu X, Ott J. A transmission/disequilibrium test that allows for genotyping errors in the analysis of single-nucleotide polymorphism data. *Am J Hum Genet* 2001;69:371–380. [PubMed: 11443542]
10. Gordon D, Matise TC, Heath SC, Ott J. Power loss for multiallelic transmission/disequilibrium test when errors introduced: GAW11 simulated data. *Genet Epidemiol* 1999;17:S587–S592. [PubMed: 10597497]
11. Gordon D, Yang Y, Haynes C, Finch SJ, Mendell NR, Brown AM, Haroutunian V. Increasing power for tests of genetic association in the presence of phenotype and/or genotype error by use of double-sampling. *Stat Appl Genet Mol Biol* 2004;3 Article 26.
12. Hosking L, Lumsden S, Lewis K, Yeo A, McCarthy L, Bansal A, Riley J, Purvis I, Xu CF. Detection of genotyping errors by Hardy-Weinberg equilibrium testing. *Eur J Hum Genet* 2004;12:395–399. [PubMed: 14872201]
13. Jawaid A, Bader JS, Purcell S, Cherny SS, Sham P. Optimal selection strategies for QTL mapping using pooled DNA samples. *Eur J Hum Genet* 2002;10:125–132. [PubMed: 11938443]

14. Kang SJ, Finch SJ, Haynes C, Gordon D. Quantifying the percent increase in minimum sample size for SNP genotyping errors in genetic model-based association studies. *Hum Hered* 2004a;58:139–144. [PubMed: 15812170]
15. Kang SJ, Gordon D, Finch SJ. What SNP genotyping errors are most costly for genetic association studies? *Genet Epidemiol* 2004b;26:132–141. [PubMed: 14748013]
16. Lai R, Zhang H, Yang Y. Repeated measurement sampling in genetic association analysis with genotyping errors. *Genet Epidemiol* 2007;31:143–153. [PubMed: 17187401]
17. Leal SM. Detection of genotyping errors and pseudo-SNPs via deviations from Hardy-Weinberg equilibrium. *Genet Epidemiol* 2005;29:204–214. [PubMed: 16080207]
18. Lin DY. Evaluating statistical significance in two-stage genomewide association studies. *Am J Hum Genet* 2006;78:505–509. [PubMed: 16408254]
19. Malley JD, Naiman DQ, Bailey-Wilson JE. A comprehensive method for genome scans. *Hum Hered* 2002;54:174–185. [PubMed: 12771550]
20. Mote VL, Anderson RL. An investigation of the effect of misclassification on the properties of chi square-tests in the analysis of categorical data. *Biometrika* 1965;52:95–109. [PubMed: 14341284]
21. Ott J. Issues in Association Analysis: Error Control in Case-Control Association Studies for Disease Gene Discovery. *Hum Hered* 2004;58:171–174. [PubMed: 15812174]
22. Rice KM, Holmans P. Allowing for genotyping error in analysis of unmatched cases and controls. *Ann Hum Genet* 2003;67:165–174. [PubMed: 12675691]
23. Risch N. Searching for genetic determinants in the new millennium. *Nature* 2000;405:847–856. [PubMed: 10866211]
24. Risch N, Merikangas K. The future of genetic studies of complex human diseases. *Science* 1996;273:1516–1517. [PubMed: 8801636]
25. Risch N, Teng J. The relative power of family-based and case-control designs for linkage disequilibrium studies of complex human diseases I. DNA pooling. *Genome Res* 1998;8:1273–1288. [PubMed: 9872982]
26. Satagopan JM, Elston RC. Optimal two-stage genotyping in population-based association studies. *Genet Epidemiol* 2003;25:149–157. [PubMed: 12916023]
27. Satagopan JM, Venkatraman ES, Begg CB. Two-stage designs for gene-disease association studies with sample size constraints. *Biometrics* 2004;60:589–597. [PubMed: 15339280]
28. Satagopan JM, Verbel DA, Venkatraman ES, Offit KE, Begg CB. Two-stage designs for gene-disease association studies. *Biometrics* 2002;58:163–170. [PubMed: 11890312]
29. Shields DC, Collins A, Buetow KH, Morton NE. Error filtration, interference, and the human linkage map. *Proc Natl Acad Sci USA* 1991;88:6501–6505. [PubMed: 1862077]
30. Skol AD, Scott LJ, Abecasis GR, Boehnke M. Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. *Nat Genet* 2006;38:209–213. [PubMed: 16415888]
31. Sobel E, Papp JC, Lange K. Detection and integration of genotyping errors in statistical genetics. *Am J Hum Genet* 2002;70:496–508. [PubMed: 11791215]
32. Tintle NL, Gordon D, McMahon FJ, Finch SJ. Using duplicate genotyped data in genetic analyses: testing association and estimating error rates. *Stat Appl Genet Mol Biol* 2007;6 Article 4.
33. Wang H, Thomas DC, Pe'er I, Stram DO. Optimal two-stage genotyping designs for genome-wide association scans. *Genet Epidemiol* 2006;30:356–368. [PubMed: 16607626]
34. Zou G, Zhao H. Haplotype frequency estimation in the presence of genotyping errors. *Hum Hered* 2003;56:131–138. [PubMed: 14614247]
35. Zou G, Zhao H. The impacts of errors in individual genotyping and DNA pooling on association studies. *Genet Epidemiol* 2004;26:1–10. [PubMed: 14691952]
36. Zuo Y, Zou G, Zhao H. Two-stage designs in case-control association analysis. *Genetics* 2006;173:1747–1760. [PubMed: 16624925]

Appendix

The proof of approximate normality for the test statistic $(t_{\text{pool}}, t_{\text{ind}})^T$ of the two-stage dependent design

Clearly, we need only to consider the situation under the alternative hypothesis. Let X_i (Y_i) denote the number of allele A carried by the i th individual in the case (control) group, $i = 1, \dots, n$. When the n cases (controls) are partitioned into m pools each having the size of s , we let X_{ij} (Y_{ij}) denote the number of allele A carried by the j th individual in the i th case (control) pool, $j = 1, \dots, s$; $i = 1, \dots, m$. Also, let X_i^* (Y_i^*) denote the observed number of allele A for the i th individual in the case (control) group when individual genotyping is used. Then

$$p_{Ai} = \sum_{j=1}^s X_{ij}/(2s), \quad p_{Ui} = \sum_{j=1}^s Y_{ij}/(2s), \quad n_A = \sum_{i=1}^{n+n_u} X_i^* \\ , \text{ and } n_U = \sum_{i=1}^{n+n_u} Y_i^* . \text{ Therefore,}$$

$$t_{\text{pool}} = \frac{\frac{1}{2ms} \sum_{i=1}^m \sum_{j=1}^s (X_{ij} - Y_{ij}) + (\bar{u} - \bar{v})}{\sqrt{\frac{\bar{p}_{\text{pool}}(1-\bar{p}_{\text{pool}})}{n} + \frac{2e^2}{m}}} \\ = \frac{\sqrt{\frac{e^2}{n}} \cdot \frac{1}{2s} \sum_{i=1}^n (X_i - Y_i) / \sqrt{\frac{e^2}{n} + w}}{\sqrt{\frac{\bar{p}_{\text{pool}}(1-\bar{p}_{\text{pool}})}{n} + \frac{2e^2}{m}}} \\ \equiv \frac{\sqrt{\frac{e^2}{n}} \cdot \sqrt{n} U_n + w}{\sqrt{\frac{\bar{p}_{\text{pool}}(1-\bar{p}_{\text{pool}})}{n} + \frac{2e^2}{m}}}, \quad (\text{say}) \quad (\text{A.1})$$

and

$$t_{\text{ind}} = \frac{\sqrt{\frac{n}{n+n_u}} \sigma^* \cdot \frac{1}{2n} \sum_{i=1}^n (X_i^* - Y_i^*) / \frac{\sigma^*}{\sqrt{n}} + \sqrt{\frac{n_u}{n+n_u}} \sigma^* \cdot \frac{1}{2n_u} \sum_{i=n+1}^{n+n_u} (X_i^* - Y_i^*) / \frac{\sigma^*}{\sqrt{n_u}}}{\sqrt{\bar{p}^*(1-\bar{p}^*)}} \\ \equiv \frac{\sqrt{\frac{n}{n+n_u}} \sigma^* \cdot \sqrt{n} V_n^{(1)} + \sqrt{\frac{n_u}{n+n_u}} \sigma^* \cdot \sqrt{n_u} V_{n_u}^{(2)}}{\sqrt{\bar{p}^*(1-\bar{p}^*)}}. \quad (\text{say}) \quad (\text{A.2})$$

We suppose that $n/(n+n_u) \rightarrow \ell_1 (< \infty)$. When ε is a non-zero constant, the proof is easy. In the following, we assume that $\sqrt{n}\varepsilon \rightarrow \ell_2 (< \infty)$. Denote

$$\zeta_i^{(1)} = \frac{X_i - Y_i}{2\sigma}, \quad \zeta_i^{(2)} = \frac{X_i^* - Y_i^*}{2\sigma^*} \\ , w_i = u_i - v_i .$$

$$U_n = \frac{1}{n} \sum_{i=1}^n \zeta_i^{(1)}, V_n^{(1)}$$

Then $= \frac{1}{n} \sum_{i=1}^n \zeta_i^{(2)}, V_{n_a}^{(2)} = \frac{1}{n_a} \sum_{i=n+1}^{n+n_a} \zeta_i^{(2)}$, and $w = \frac{1}{n} \sum_{i=1}^n w_i$. Clearly, $(\zeta_i^{(1)}, \zeta_i^{(2)})^T$ ($i = 1, \dots, n, \dots$) are i.i.d. random vectors, whose mean is $\tilde{\mu} = (\mu / \sigma, \mu^* (1 - e_1 - e_2) / \sigma^*)^T$, and variance-covariance matrix is

$$\Sigma_1 = \begin{pmatrix} 1 & \frac{\sigma_{12}^*}{4\sigma\sigma^*} \\ \frac{\sigma_{12}^*}{4\sigma\sigma^*} & (1 - e_1 - e_2)^2 \end{pmatrix}.$$

Thus, from the central limit theorem, we obtain

$$\sqrt{n}((U_n, V_n^{(1)})^T - \tilde{\mu}) \xrightarrow{d} N(0, \Sigma_1),$$

where $\xrightarrow{d}$ means convergence in distribution. Further, we have

$$\begin{aligned} \sqrt{n_a} (V_{n_a}^{(2)} - \frac{\mu^* (1 - e_1 - e_2)}{\sigma^*}) &\xrightarrow{d} N(0 \\ &, (1 - e_1 - e_2)^2). \end{aligned}$$

Now, denote

$$W_{n,n_a} = (\sqrt{n}(U_n - \mu/\sigma), \sqrt{n}[V_n^{(1)} - \mu^*(1 - e_1 - e_2)/\sigma^*], \sqrt{n_a}[V_{n_a}^{(2)} - \mu^*(1 - e_1 - e_2)/\sigma^*], w/\varepsilon)^T,$$

$$\theta_{n,n_a} = (\sqrt{n}\mu/\sigma, \sqrt{n}\mu^*(1 - e_1 - e_2)/\sigma^*, \sqrt{n_a}\mu^*(1 - e_1 - e_2)/\sigma^*, 0)^T,$$

and

$$B_{n,n_a} = \begin{pmatrix} \frac{\sqrt{\sigma^2/n}}{\sqrt{\hat{p}_{pool}(1-\hat{p}_{pool})/n+2\varepsilon^2/m}} & 0 & 0 & \frac{\varepsilon}{\sqrt{\hat{p}_{pool}(1-\hat{p}_{pool})/n+2\varepsilon^2/m}} \\ 0 & \frac{\sqrt{n/(n+n_a)}\sigma^*}{\sqrt{\hat{p}^*(1-\hat{p}^*)}} & \frac{\sqrt{n_a/(n+n_a)}\sigma^*}{\sqrt{\hat{p}^*(1-\hat{p}^*)}} & 0 \end{pmatrix}.$$

Then from (A.1) and (A.2), we have

$$\begin{aligned} (t_{pool}, t_{ind})^T &= B_{n,n_a} W_{n,n_a} \\ &+ B_{n,n_a} \theta_{n,n_a}. \end{aligned} \quad (\text{A.3})$$

Note that $\widehat{p}_{pool} \xrightarrow{p.} \tilde{p}$ and $\widehat{p}^* \xrightarrow{p.} \tilde{p}^*$, where $\xrightarrow{p.}$ means convergence in probability. So the limit of B_{n,n_a} in probability is

$$B = \begin{pmatrix} \frac{\sigma}{\sqrt{\tilde{p}(1-\tilde{p})+2\ell_2^2/m}} & 0 & 0 & \frac{\ell_2}{\sqrt{\tilde{p}(1-\tilde{p})+2\ell_2^2/m}} \\ 0 & \frac{\sqrt{\ell_1}\sigma^*}{\sqrt{\tilde{p}^*(1-\tilde{p}^*)}} & \frac{\sqrt{1-\ell_1}\sigma^*}{\sqrt{\tilde{p}^*(1-\tilde{p}^*)}} & 0 \end{pmatrix}.$$

From (A.3), when $n + n_a \rightarrow \infty$, we obtain

$$\begin{pmatrix} t_{pool}, t_{ind} \end{pmatrix}^T - \widehat{\mu}_{n,n_a} \xrightarrow{d.} N(0, \Sigma_1^{**}), \quad (\text{A.4})$$

where

$$\begin{aligned} \widehat{\mu}_{n,n_a} &= B_{n,n_a} \theta_{n,n_a} \\ &= \left(\frac{\mu}{\sqrt{\widehat{p}_{pool}(1-\widehat{p}_{pool})/n+2\ell_2^2/m}}, \frac{\sqrt{n+n_a}\mu^*(1-e_1-e_2)}{\sqrt{\widehat{p}^*(1-\widehat{p}^*)}} \right)^T, \end{aligned}$$

and

$$\begin{aligned} \Sigma_1^{**} &= B \begin{pmatrix} \Sigma_1 & 0 & 0 \\ 0 & (1-e_1-e_2)^2 & 0 \\ 0 & 0 & 2/m \end{pmatrix} B^T \\ &= \begin{pmatrix} \frac{\sigma^2+2\ell_2^2/m}{\tilde{p}(1-\tilde{p})+2\ell_2^2/m} & \frac{\sqrt{\ell_1}\sigma_{12}^*}{4\sqrt{\tilde{p}^*(1-\tilde{p}^*)}\sqrt{\tilde{p}(1-\tilde{p})+2\ell_2^2/m}} \\ \frac{\sqrt{\ell_1}\sigma_{12}^*}{4\sqrt{\tilde{p}^*(1-\tilde{p}^*)}\sqrt{\tilde{p}(1-\tilde{p})+2\ell_2^2/m}} & \frac{\sigma^{*2}(1-e_1-e_2)^2}{\tilde{p}^*(1-\tilde{p}^*)} \end{pmatrix}. \end{aligned}$$

Finally, it is readily seen that $\widehat{\mu}_{n,n_a} = \tilde{\mu}^* + o_p(\sqrt{n+n_a})$, and $\Sigma_1^* \rightarrow \Sigma_1^{**}$. Combining these and (A.4), we see that under the alternative hypothesis H_1 , $(t_{pool}, t_{ind})^T$ approximately follows a joint bivariate normal distribution $N(\tilde{\mu}^*, \Sigma_1^*)$.

Table 1

The power of the two-stage **dependent** design for fixed population allele frequency and allele frequency difference between the case and control groups in the presence of genotyping errors

	$p_A - p_U = 0.03$	$p_A - p_U = 0.05$	$p_A - p_U = 0.07$	$p_A - p_U = 0.10$
$[e_1, e_2]$	$p = 0.05$			
[0, 0]	0.070	0.744	0.948	0.950
[0.01, 0.01]	0.039	0.613	0.943	0.950
[0.03, 0.03]	0.013	0.369	0.902	0.950
[0.01, 0.03]	0.015	0.395	0.910	0.950
$[e_1, e_2]$	$p = 0.20$			
[0, 0]	0.001	0.062	0.458	0.938
[0.01, 0.01]	0.001	0.050	0.403	0.930
[0.03, 0.03]	6.52×10^{-4}	0.032	0.302	0.902
[0.01, 0.03]	7.26×10^{-4}	0.036	0.327	0.911
$[e_1, e_2]$	$p = 0.70$			
[0, 0]	6.24×10^{-4}	0.037	0.374	0.936
[0.01, 0.01]	5.21×10^{-4}	0.030	0.327	0.927
[0.03, 0.03]	3.64×10^{-4}	0.020	0.244	0.895
[0.01, 0.03]	4.69×10^{-4}	0.027	0.302	0.920

* The power in the pooling stage is $1 - \beta_1 = 95\%$, and the significance level for the two-stage design is $\alpha / M = 5 \times 10^{-8}$.

** The sample size in the first stage is 500, the supplemental sample size in the second stage is also 500, the measurement error rate is $\epsilon = 0.01$, and the number of pool pairs is $m = 1$.

Table 2

The power of the two-stage **independent** design for fixed population allele frequency and allele frequency difference between the case and control groups in the presence of genotyping errors

	$p_A - p_U = 0.03$	$p_A - p_U = 0.05$	$p_A - p_U = 0.07$	$p_A - p_U = 0.10$
$[e_1, e_2]$	$p = 0.05$			
[0, 0]	0.006	0.229	0.820	0.950
[0.01, 0.01]	0.004	0.153	0.734	0.950
[0.03, 0.03]	0.001	0.069	0.539	0.948
[0.01, 0.03]	0.001	0.076	0.564	0.949
$[e_1, e_2]$	$p = 0.20$			
[0, 0]	1.44×10^{-4}	0.007	0.112	0.773
[0.01, 0.01]	1.20×10^{-4}	0.006	0.094	0.735
[0.03, 0.03]	8.34×10^{-5}	0.004	0.065	0.651
[0.01, 0.03]	9.10×10^{-5}	0.004	0.072	0.675
$[e_1, e_2]$	$p = 0.70$			
[0, 0]	7.79×10^{-5}	0.004	0.081	0.764
[0.01, 0.01]	6.74×10^{-5}	0.003	0.069	0.726
[0.03, 0.03]	5.05×10^{-5}	0.002	0.049	0.642
[0.01, 0.03]	6.19×10^{-5}	0.003	0.063	0.705

* The power in the pooling stage is $1 - \beta_1 = 95\%$, and the significance level for the two-stage design is $\alpha / M = 5 \times 10^{-8}$.

** The sample sizes in the first and second stages are both 500, the measurement error rate is $\varepsilon = 0.01$, and the number of pool pairs is $m = 1$.

Table 3

The optimal power of the two-stage **dependent** design for a fixed total cost of $C = 2 \times 10^6$ when the number of pool pairs is $m = 1$

	$p_A - p_U = 0.05$				$p_A - p_U = 0.10$			
	$p = 0.05$	$p = 0.20$	$p = 0.70$		$p = 0.05$	$p = 0.20$	$p = 0.70$	
Power	0.113	0.003	0.002		0.972	0.398	0.386	
n	478	456	484		473	484	483	
n_a	0	27	0		0	0	0	
k_1	2.975	3.420	3.585		2.803	3.532	3.344	
k_2	5.217	5.307	5.317		5.246	5.306	5.323	
β_1	0.610	0.911	0.941		0.011	0.274	0.219	

* The significance level for the two-stage design is $\alpha / M = 5 \times 10^{-8}$, $C_0 = 0$, $C_1 = 2000$, $C_{ind} = 0.02$, $C_{pool} = 0.03$, $M = 10^6$, $K = 1$, and the error rates are $e_1 = e_2 = 0.001$ and $\varepsilon = 0.01$.

Table 4

The optimal power of the two-stage **independent** design for a fixed total cost of $C = 2 \times 10^6$ when the number of pool pairs is $m = 1$

	$p_A - p_U = 0.05$			$p_A - p_U = 0.10$		
	$p = 0.05$	$p = 0.20$	$p = 0.70$	$p = 0.05$	$p = 0.20$	$p = 0.70$
Power	0.050	0.002	0.001	0.826	0.232	0.226
n	124	140	146	169	162	163
n_a	339	335	330	294	312	311
k_1	2.484	2.749	2.779	2.433	2.694	2.693
k_2	4.323	4.148	4.128	4.354	4.186	4.187
β_1	0.748	0.924	0.936	0.099	0.490	0.491

* The significance level for the two-stage design is $\alpha / M = 5 \times 10^{-8}$, $C_0 = 0$, $C_1 = 2000$, $C_{ind} = 0.02$, $C_{pool} = 0.03$, $M = 10^6$, $K = 1$, and the error rates are $e_1 = e_2 = 0.001$ and $\varepsilon = 0.01$.

Table 5

The optimal power of the two-stage **dependent** design for a fixed total cost of $C = 5 \times 10^5$ when the number of pool pairs is $m = 1$

	$p_A - p_U = 0.05$			$p_A - p_U = 0.10$		
	$p = 0.05$	$p = 0.20$	$p = 0.70$	$p = 0.05$	$p = 0.20$	$p = 0.70$
Power	0.501 (0.067)	0.074 (0.012)	0.046 (0.008)	0.999 (0.545)	0.971 (0.362)	0.969 (0.356)
n	950 (681)	1110 (946)	1116 (975)	828 (379)	1053 (667)	1056 (674)
n_a	0 (0)	1 (1)	0 (0)	0 (0)	0 (0)	0 (0)
k_1	2.862 (2.463)	3.389 (2.856)	3.428 (2.912)	2.662 (2.047)	3.115 (2.445)	3.125 (2.454)
k_2	5.090 (4.798)	5.216 (4.831)	5.257 (4.875)	5.170 (5.051)	5.267 (5.065)	5.296 (5.118)
β_1	0.434 (0.906)	0.771 (0.959)	0.805 (0.964)	0.001 (0.432)	0.021 (0.600)	0.022 (0.603)

* The significance level for the two-stage design is $\alpha / M = 5 \times 10^{-8}$, $C_0 = 0$, $C_1 = 200$, $C_{ind} = C_{pool} = 0.02$, $M = 10^6$, $K = 1$, and the error rates are $e_1 = e_2 = 0.001$ and $\varepsilon = 0.01$ (The values in bracket correspond to $\varepsilon = 0.03$).

Table 6

The optimal power of the two-stage **dependent** design for a fixed total cost of $C = 2 \times 10^6$ when the number of pool pairs is $m = 10$

	$p_A - p_U = 0.05$				$p_A - p_U = 0.10$			
	$p = 0.05$	$p = 0.20$	$p = 0.70$		$p = 0.05$	$p = 0.20$	$p = 0.70$	
Power	0.038	0.001	0.001		0.830	0.154	0.152	
n	349	342	343		349	349	349	
n_a	0	0	0		0	0	0	
k_1	4.860	2.829	2.872		4.872	4.978	4.225	
k_2	5.245	5.266	5.255		5.240	5.312	5.311	
β_1	0.930	0.744	0.798		0.120	0.770	0.500	

* The significance level for the two-stage design is $\alpha / M = 5 \times 10^{-8}$, $C_0 = 0$, $C_1 = 2000$, $C_{ind} = 0.02$, $C_{pool} = 0.03$, $M = 10^6$, $K = 1$, and the error rates are $e_1 = e_2 = 0.001$ and $\varepsilon = 0.01$.

Table 7

A general genotyping error model (Kang et al. 2004a, Kang et al. 2004b)

Observed genotype	True genotype at the candidate marker		
	<i>AA</i>	<i>Aa</i>	<i>aa</i>
<i>AA</i>	$1 - \varepsilon_{21} - \varepsilon_{20}$	ε_{12}	ε_{02}
<i>Aa</i>	ε_{21}	$1 - \varepsilon_{12} - \varepsilon_{10}$	ε_{01}
<i>aa</i>	ε_{20}	ε_{10}	$1 - \varepsilon_{01} - \varepsilon_{02}$

Table 8

Empirical Type I error rate for the prevalence of 0.05 and empirical power for the penetrances $f_2 = 0.0226, f_1 = f_0 = 0.0113$ for the two-stage **dependent** design in the presence of genotyping errors

Genotyping error rate	$\varepsilon_{21} = 0.01, \varepsilon_{20} = 0.02, \varepsilon_{12} = 0.03, \varepsilon_{10} = 0.02, \varepsilon_{02} = 0.01, \varepsilon_{01} = 0.02$	$\varepsilon_{21} = 0.1, \varepsilon_{20} = 0.2, \varepsilon_{12} = 0.3, \varepsilon_{10} = 0.2, \varepsilon_{02} = 0.1, \varepsilon_{01} = 0.2$
Empirical Type I error rate (nominal level)	0.049 (0.050)	0.077 (0.050)
Empirical power (theoretical power)	0.712 (0.715)	0.351 (0.349)

* The theoretical power in the pooling stage is $1 - \beta_1 = 95\%$, the population allele frequency is 0.2, the sample size in the first stage is 500, the supplemental sample size in the second stage is also 500, the measurement error rate is $\varepsilon = 0.01$, and the number of pool pairs is $m = 1$.