A DNA Pooling Strategy for Family-Based Association Studies

Wen-Chung Lee

Graduate Institute of Epidemiology, College of Public Health, National Taiwan University

Abstract

Genome-wide association scans for disease susceptibility genes of complex diseases require genotyping on a massive scale. A DNA pooling strategy for family-based association studies is described, which is robust to population stratification biases and to errors in pooling. It can achieve a statistical efficiency of 0.95 with ~1 of 8 or fewer genotyping efforts, and an efficiency of 0.90 with ~1 of 16 or fewer efforts compared with individual genotyping. The pooling method described in this article provides a tradeoff between genotyping efforts and subject recruitment efforts. (Cancer Epidemiol Biomarkers Prev 2005;14(4):958–62)

Introduction

A genome-wide association study to search for the disease susceptibility gene(s) of a complex disease requires genotyping on a massive scale, in terms either of the number of markers that have to be scanned or of the number of individuals that have to be recruited (1-3). A practical way to reduce genotyping efforts is to carry out analyses on pools made up of DNA from many individuals rather than on individual samples (4-9).

Previous studies on DNA pooling considered the simplest case-control design where the cases and the unrelated controls were sampled from a homogeneous population in Hardy-Weinberg equilibrium (6-9). However, family-based association designs have become popular in the recent decade in part because they are robust to population stratification biases (10-15). The designs recruit relatives of the cases as the control subjects. One can distinguish between two lines of relatives (15): line I, the parents or, if the parents are missing, the unaffected siblings; and line II, the spouse or, if the spouse is missing, the offspring. For early-onset diseases, the sample of line I relatives should be simple to collect, whereas for late-onset diseases, one could easily turn to line II relatives instead. Furthermore, the information from both lines of relatives (if available) can be integrated to improve the study power (15). (Note that the relatives here include those related to the affected cases genetically and those only in law. However, we still refer to the design as a ''family''-based study, because these subjects are coming from the same family literally).

Risch and Teng (4) have constructed statistical tests for case-parents and case-siblings DNA pooling studies. However, the tests are valid only if the study population is in Hardy-Weinberg equilibrium. This defeats one purpose of a family-based study, which is meant to be robust to population structure. Furthermore, they assumed the allele frequency measurement for a DNA pool to be 100% accurate, which is not possible with current technologies. In this article, a robust and efficient DNA pooling strategy for family-based association study is described.

DNA Pooling Strategy

Assume that there is only one affected case in a (nuclear) family. Let two numbers (x and y) distinguish the various family configurations. The x is the number of line I relatives (parents or unaffected siblings) an affected case has (parents are treated as if $x = \infty$). The y is the number of line II relatives (spouse or offspring) an affected case has (spouse is treated as if $y = \infty$). Assume that we form a total of J ($j = 1, \ldots, J$) ''pooling sets'' in a study. Each pooling set contains a certain number of families exclusively of the same family configuration. Let (x_j, y_j) represent the family configuration of the jth set, and n_j, the number of families in the jth set. (The problem of how many families should be put into a pooling set will be discussed later.)

At the jth pooling set, the affected cases are pooled into a single ''case pool''. The allele frequency of the pool is measured using quantitative PCR. The result is denoted as C_j. Note that here and hereinafter, we do not attempt to correct for unequal allelic amplification often associated with a quantitative PCR (6, 7). If parents are available ($x_j = \infty$), the fathers in this jth set are pooled into a single ''father pool'', and the mothers, a single ''mother pool''. The measured allele frequency of the father(mother) pool is denoted as $F_j(M_j)$. Then we calculate for all the families in the jth pooling set, D_j^{I} = total allele counts for the transmitted genotypes (the affected cases) − total allele counts for the nontransmitted genotypes = $2n_jC_j - 2n_j(F_j + M_j - C_j) = 4n_j[C_j - (F_j + M_j)/2]$. If parents are not available but unaffected siblings are ($0 < x_j < \infty$), we form a total of x_j ''sibling pools'' ($k = 1, \ldots, x_j$), with the kth pool containing each and every one of the kth eldest unaffected siblings of the families in the jth set. The measured allele frequency for the kth sibling pool in the jth set is denoted as B_{jk}. And we calculate for all the families in the jth pooling set, D_j^{I} = total allele counts for the transmitted genotypes (the affected cases) − total allele counts for the ''imputed nontransmitted'' (see ref. 15)

$$D_j^{\mathrm{I}} = 2n_jC_j - 2n_j(2/x_j \cdot \sum_{k=1}^{x_j} B_{jk} - C_j) = 4n_j(C_j - 1/x_j \cdot \sum_{k=1}^{x_j} B_{jk}).$$

Next, let us turn to the line II relatives. If spouses are available ($y_j = \infty$), the spouses are pooled into a single ''spouse pool''. The measured allele frequency of the spouse pool is denoted as S_j. We calculate for all the families in the jth pooling set, D_j^{II} = total allele counts for the affected cases − total allele

Received 7/7/04; revised 11/11/04; accepted 11/23/04.

Grant support: National Science Council, Republic of China.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked advertisement in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Requests for reprints: Wen-Chung Lee, No. 1, Jen-Ai Road, 1st Section, Taipei, Taiwan. Phone: 886-2-23511955; Fax: 886-2-23511955. E-mail: wenchung@ha.mc.ntu.edu.tw

counts for the spouses = $2n_j(C_j - S_j)$. If spouses are not available but offspring are ($0 < y_j < \infty$), we form a total of y_j ''offspring pools'' ($l = 1, \ldots, y_j$), with the lth pool containing each and every one of the lth eldest offspring of the families in the jth set. The measured allele frequency for the lth offspring pool in the jth set is denoted as O_{jl}. And we calculate for all the families in the jth pooling set, D_j^{II} = total allele counts for the affected cases − total allele counts for the ''imputed spouses'' (see ref. 13, 15)

$$D_j^{II} = 2n_jC_j - 2n_j(2/y_j \cdot \sum_{l=1}^{y_j} O_{jl} - C_j) = 4n_j(C_j - 1/y_j \cdot \sum_{l=1}^{y_j} O_{jl}).$$

Expected Total Allele Differences

If the study population is a homogeneous population or is a stratified population but mating is restrictive to subjects in the same stratum (13), the probability distribution of the allele frequency for the case pool under the null will be exactly the same as those for the father pool, the mother pool, the spouse pool, and each and every one of the sibling and the offspring pools in the same pooling set. Therefore, the expected values of D_j^I and D_j^{II} are both zero under the null hypothesis of no genetic association, no matter how ''biased'' a quantitative PCR can be (a complex error structure that is asymmetric for the alleles in a pool, or an allelic amplification that is a nonlinear function of allele frequency, etc.). By contrast, a ''greedy'' strategy that puts all line I relatives and line II relatives into a ''grand control'' pool does not have this property. A grand control pool so formed will contain more subjects than the corresponding case pool. Consequently, even under the null hypothesis, the distributions of the allele frequencies will not be the same for the case pool and the grand control pool in a pooling set, due to the law of large numbers.

To better illustrate the point, Table 1 presents the probability distributions for the various DNA pools in a hypothetical pooling set, which contains two cases and their parents recruited from a Hardy-Weinberg population with allele frequency of 0.4. The quantitative PCR was assumed to have a very complex error structure (column 2 of Table 1) in that it will amplify an allele if that allele wins a majority in a pool. Moreover, the amplification is not a linear function of the allele frequency and is not symmetric for the two alleles. It can be seen that the expected allele frequencies measured by this grossly biased PCR are exactly the same (0.3885) for the case pool, the father pool, and the mother pool (although the values are not equal to the true value of 0.4). By contrast, a grand control pool formed by putting the father and the mother pools together has a different expected value (0.3587) by the same PCR.

Table 1. Probability distributions and expected allele frequencies for the various DNA pools in a hypothetical pooling set, which contains two cases and their parents recruited from a Hardy-Weinberg population with allele frequency of 0.4

Allele frequency		Probability			
Actual	Measured	Case pool	Father pool	Mother pool	Parents pool
0.000	0.000	0.1296	0.1296	0.1296	0.0168
0.125	0.050	0.0000	0.0000	0.0000	0.0896
0.250	0.150	0.3456	0.3456	0.3456	0.2090
0.375	0.200	0.0000	0.0000	0.0000	0.2787
0.500	0.500	0.3456	0.3456	0.3456	0.2322
0.625	0.850	0.0000	0.0000	0.0000	0.1239
0.750	0.900	0.1536	0.1536	0.1536	0.0413
0.875	0.990	0.0000	0.0000	0.0000	0.0079
1.000	1.000	0.0256	0.0256	0.0256	0.0007
Expected allele frequency					
Actual		0.4000	0.4000	0.4000	0.4000
Measured		0.3885	0.3885	0.3885	0.3587

Disequilibrium Test for Pooled Data

Because $E(D_j^I) = E(D_j^{II}) = 0$, we have $E(w_j^I D_j^I + w_j^{II} D_j^{II}) = 0$ for arbitrary coefficients w_j^I and w_j^{II} under the null (condition 1). Furthermore, because subjects from the same family will and only will appear in one pooling set and the measurements of the various DNA pools are done independently, the $w_j^I D_j^I + w_j^{II} D_j^{II}$ for $j = 1, \ldots, J$ are independent to one another (condition 2). A disequilibrium test for pooled data is constructed below

$$Z^2 = \left[\sum_{j=1}^{J}(w_j^I D_j^I + w_j^{II} D_j^{II})\right]^2 \Big/ \sum_{j=1}^{J}(w_j^I D_j^I + w_j^{II} D_j^{II})^2.$$

For large J (e.g., $J > 30$), Z^2 is distributed as a one degree-of-freedom χ^2 distribution under the null, for whatever values chosen for the w_j^I and w_j^{II} (as long as they do not both assume the value of zero). Note that the conditions 1 and 2 stated above are sufficient to ensure Z^2 to be a valid test for genetic association. It does not matter whether or not the PCR is unbiased or has a simple error structure (e.g., can be described using a single parameter) and whether or not the population under study is in Hardy-Weinberg equilibrium.

In this article, the same coefficients previously proposed for individually genotyped data (15) are used for pooled data, i.e.,

$$w_j^I = \frac{x_j/(x_j+1) - x_jy_j/[2(x_j+1)(y_j+1)]}{1 - x_jy_j/[4(x_j+1)(y_j+1)]}$$

and

$$w_j^{II} = \frac{y_j/(y_j+1) - x_jy_j/[2(x_j+1)(y_j+1)]}{1 - x_jy_j/[4(x_j+1)(y_j+1)]}.$$

These coefficients are optimal under the null hypothesis and in a Hardy-Weinberg population. Note that the sum of the two coefficients,

$$w_j^I + w_j^{II} = \frac{x_j/(x_j+1) + y_j/(y_j+1) - x_jy_j/[(x_j+1)(y_j+1)]}{1 - x_jy_j/[4(x_j+1)(y_j+1)]},$$

is in general not equal to one. Rather, it reflects the statistical efficiency for family configuration (x_j, y_j), relative to case-parents data (∞, 0).

Statistical Efficiency

Assume that the study population is a random-mating population in Hardy-Weinberg equilibrium with allele frequency of P. Lee (15) showed that the variance of $w_j^I D_j^I + w_j^{II} D_j^{II}$ under the null hypothesis is

$$V_j^{\text{individual}} = 4n_jP(1-P)(w_j^I + w_j^{II}),$$

if the allele counts are based on individual genotyping (assuming no genotyping error). With pooled genotyping, the variance is composed of two terms (i.e., $V_j^{\text{pooled}} = V_j^{\text{individual}} + V_j^{\text{quantitative PCR}}$). To calculate $V_j^{\text{quantitative PCR}}$, we assume

that the measurement error of the quantitative PCR in estimating the allele frequency of a DNA pool is a constant σ, irrespective of the actual allele frequency. As an example, in a pooling set with affected cases together with their fathers, their mothers, and their three offspring (∞, 3), we have

$$\begin{aligned} w_j^{\mathrm{I}} D_j^{\mathrm{I}} + w_j^{\mathrm{II}} D_j^{\mathrm{II}} &= 4n_j w_j^{\mathrm{I}} \cdot [C_j - (F_j + M_j)/2] + 4n_j w_j^{\mathrm{II}} \cdot [C_j - 1/3 \\ &\quad \cdot (O_{j1} + O_{j2} + O_{j3})] \\ &= 4n_j (w_j^{\mathrm{I}} + w_j^{\mathrm{II}}) \cdot C_j - 2n_j w_j^{\mathrm{I}} \cdot (F_j + M_j) - 4n_j w_j^{\mathrm{II}}/3 \\ &\quad \cdot (O_{j1} + O_{j2} + O_{j3}). \end{aligned}$$

And the variance associated with the quantitative PCR is

$$\begin{aligned} &16n_j^2 (w_j^{\mathrm{I}} + w_j^{\mathrm{II}})^2 \cdot \sigma^2 + 4n_j^2 w_j^{\mathrm{I}} w_j^{\mathrm{I}} \cdot (\sigma^2 + \sigma^2) \\ &+16/9 \cdot n_j^2 w_j^{\mathrm{II}} w_j^{\mathrm{II}} \cdot (\sigma^2 + \sigma^2 + \sigma^2). \end{aligned}$$

For a general pooling set with family configuration (x_j, y_j), it is easy to show that

$$\begin{aligned} V_j^{\text{quantitative PCR}} &= 16n_j^2 \sigma^2 \{[w_j^{\mathrm{I}} + (1 - I_{(y_j=\infty)}/2) w_j^{\mathrm{II}}]^2 \\ &\quad +(I_{(x_j=\infty)}/2 + I_{(x_j>0)}/x_j) w_j^{\mathrm{I}} w_j^{\mathrm{I}} \\ &\quad +(I_{(y_j=\infty)}/4 + I_{(y_j>0)}/y_j) w_j^{\mathrm{II}} w_j^{\mathrm{II}}\}, \end{aligned}$$

where the $I_{(\text{statement})}$ is an indicator function (value = 1, if the statement is true; value = 0, otherwise). The statistical efficiency of the pooled study relative to individual genotyping can thus be approximated by $\Sigma_j V_j^{\text{individual}} / \Sigma_j V_j^{\text{pooled}}$. To achieve a statistical efficiency of c ($0 < c < 1$), the number of families in a pooling set can be kept below a certain limit, depending on the family configuration of the set, i.e.,

$$\begin{aligned} n_j &< \frac{1-c}{c} \times \frac{P(1-P)}{4\sigma^2} \\ &\quad \times \frac{w_j^{\mathrm{I}} + w_j^{\mathrm{II}}}{[w_j^{\mathrm{I}} + (1 - I_{(y_j=\infty)}/2) w_j^{\mathrm{II}}]^2 + (I_{(x_j=\infty)}/2 + I_{(x_j>0)}/x_j) w_j^{\mathrm{I}} w_j^{\mathrm{I}}} \\ &\quad +(I_{(y_j=\infty)}/4 + I_{(y_j>0)}/y_j) w_j^{\mathrm{II}} w_j^{\mathrm{II}} \end{aligned}.$$

Note that the above calculations were carried out under the null hypothesis, with a constant σ, and in a homogeneous population in Hardy-Weinberg equilibrium. A more relevant comparison of the methods should be made with respect to the power under a specific alternative hypothesis, for a more complex but realistic error structure, and in a structured population. Yet, the simple and elegant formulas presented above provide a concrete guideline for forming DNA pools. Adhering to the guideline will optimize the study efficiency at least nearly, for a gene with a weak effect, using a PCR with error nearly constant, and in a population deviating from Hardy-Weinberg equilibrium not too much.

Assuming σ = 0.01 (corresponding to the method of mass spectrometry; ref. 6), we found that the pooling strategy can achieve a statistical efficiency of 0.95 with ~1 of 8 or fewer genotyping efforts (Table 2), and a statistical efficiency of 0.90 with ~1 of 16 or fewer efforts (Table 3), compared with individual genotyping. The reduction in genotyping efforts is greater when the frequency of the minor allele is higher, and smaller when it is lower. In a Hardy-Weinberg population, a case-control study with equal number of cases and controls is equivalent to a case-spouse study in terms of statistical efficiency. Tables 2 and 3 (right upper corners) indicate that a pooled case-control study (and a pooled case-spouse study) has the greatest reduction in genotyping efforts (a statistical efficiency of 0.95 with ~1 of 20 or fewer genotyping efforts and a statistical efficiency of 0.90 with ~1 of 50 or fewer efforts). However, it should be pointed out again that a

Table 2. Maximum number of families that can be pooled together (% total numbers of genotyping) for various family configurations (*x* and *y*) to achieve a statistical efficiency of 0.95 as compared to individual genotyping, assuming the measurement error associated with a pooled genotyping is 0.01

No. line I relatives (x)	No. line II relatives (y)					
	0	1	2	3	4	∞ (Spouse)
Frequency of the minor allele = 0.1						
0	—	11.8 (8.4)	11.8 (8.4)	11.8 (8.4)	11.8 (8.4)	23.7 (4.2)
1	11.8 (8.4)	9.9 (10.1)	9.7 (10.3)	9.7 (10.3)	9.7 (10.3)	17.5 (5.7)
2	11.8 (8.4)	9.7 (10.3)	9.5 (10.6)	9.4 (10.6)	9.4 (10.7)	16.1 (6.2)
3	11.8 (8.4)	9.7 (10.3)	9.4 (10.6)	9.3 (10.7)	9.3 (10.8)	15.6 (6.4)
4	11.8 (8.4)	9.7 (10.3)	9.4 (10.7)	9.3 (10.8)	9.2 (10.9)	15.3 (6.5)
∞ (Parents)	7.9 (12.7)	7.7 (13.0)	7.7 (12.9)	7.7 (12.9)	7.8 (12.9)	11.8 (8.4)
Frequency of the minor allele = 0.3						
0	—	27.6 (3.6)	27.6 (3.6)	27.6 (3.6)	27.6 (3.6)	55.3 (1.8)
1	27.6 (3.6)	23.0 (4.3)	22.7 (4.4)	22.6 (4.4)	22.6 (4.4)	40.7 (2.5)
2	27.6 (3.6)	22.7 (4.4)	22.1 (4.5)	21.9 (4.6)	21.9 (4.6)	37.7 (2.7)
3	27.6 (3.6)	22.6 (4.4)	21.9 (4.6)	21.7 (4.6)	21.6 (4.6)	36.4 (2.7)
4	27.6 (3.6)	22.6 (4.4)	21.9 (4.6)	21.6 (4.6)	21.5 (4.7)	35.7 (2.8)
∞ (Parents)	18.4 (5.4)	18.0 (5.6)	18.0 (5.5)	18.1 (5.5)	18.1 (5.5)	27.6 (3.6)
Frequency of the minor allele = 0.5						
0	—	32.9 (3.0)	32.9 (3.0)	32.9 (3.0)	32.9 (3.0)	65.8 (1.5)
1	32.9 (3.0)	27.4 (3.6)	27.0 (3.7)	26.9 (3.7)	26.9 (3.7)	48.5 (2.1)
2	32.9 (3.0)	27.0 (3.7)	26.3 (3.8)	26.1 (3.8)	26.0 (3.8)	44.9 (2.2)
3	32.9 (3.0)	26.9 (3.7)	26.1 (3.8)	25.8 (3.9)	25.7 (3.9)	43.3 (2.3)
4	32.9 (3.0)	26.9 (3.7)	26.0 (3.8)	25.7 (3.9)	25.6 (3.9)	42.4 (2.4)
∞ (Parents)	21.9 (4.6)	21.4 (4.7)	21.5 (4.7)	21.5 (4.6)	21.6 (4.6)	32.9 (3.0)

Table 3. Maximum number of families that can be pooled together (% total numbers of genotyping) for various family configurations (*x* and *y*) to achieve a statistical efficiency of 0.90 as compared to individual genotyping, assuming the measurement error associated with a pooled genotyping is 0.01

No. line I relatives (x)	No. line II relatives (y)					
	0	1	2	3	4	∞ (Spouse)
Frequency of the minor allele = 0.1						
0	—	25.0 (4.0)	25.0 (4.0)	25.0 (4.0)	25.0 (4.0)	50.0 (2.0)
1	25.0 (4.0)	20.8 (4.8)	20.5 (4.9)	20.5 (4.9)	20.5 (4.9)	36.8 (2.7)
2	25.0 (4.0)	20.5 (4.9)	20.0 (5.0)	19.8 (5.0)	19.8 (5.1)	34.1 (2.9)
3	25.0 (4.0)	20.5 (4.9)	19.8 (5.0)	19.6 (5.1)	19.6 (5.1)	32.9 (3.0)
4	25.0 (4.0)	20.5 (4.9)	19.8 (5.1)	19.6 (5.1)	19.4 (5.1)	32.3 (3.1)
∞ (Parents)	16.7 (6.0)	16.3 (6.1)	16.3 (6.1)	16.4 (6.1)	16.4 (6.1)	25.0 (4.0)
Frequency of the minor allele = 0.3						
0	—	58.3 (1.7)	58.3 (1.7)	58.3 (1.7)	58.3 (1.7)	116.7 (0.9)
1	58.3 (1.7)	48.6 (2.1)	47.9 (2.1)	47.7 (2.1)	47.7 (2.1)	86.0 (1.2)
2	58.3 (1.7)	47.9 (2.1)	46.7 (2.1)	46.3 (2.2)	46.2 (2.2)	79.5 (1.3)
3	58.3 (1.7)	47.7 (2.1)	46.3 (2.2)	45.8 (2.2)	45.6 (2.2)	76.8 (1.3)
4	58.3 (1.7)	47.7 (2.1)	46.2 (2.2)	45.6 (2.2)	45.4 (2.2)	75.3 (1.3)
∞ (Parents)	38.9 (2.6)	38.0 (2.6)	38.0 (2.6)	38.2 (2.6)	38.3 (2.6)	58.3 (1.7)
Frequency of the minor allele = 0.5						
0	—	69.4 (1.4)	69.4 (1.4)	69.4 (1.4)	69.4 (1.4)	138.9 (0.7)
1	69.4 (1.4)	57.9 (1.7)	57.0 (1.8)	56.8 (1.8)	56.8 (1.8)	102.3 (1.0)
2	69.4 (1.4)	57.0 (1.8)	55.6 (1.8)	55.1 (1.8)	55.0 (1.8)	94.7 (1.1)
3	69.4 (1.4)	56.8 (1.8)	55.1 (1.8)	54.6 (1.8)	54.3 (1.8)	91.4 (1.1)
4	69.4 (1.4)	56.8 (1.8)	55.0 (1.8)	54.3 (1.8)	54.0 (1.9)	89.6 (1.1)
∞ (Parents)	46.3 (2.2)	45.2 (2.2)	45.3 (2.2)	45.4 (2.2)	45.5 (2.2)	69.4 (1.4)

simple case-control study, individual genotyping and pooled genotyping alike, offers no protection for population stratification biases.

Discussion

For most complex diseases such as noninsulin-dependent diabetes, cardiovascular diseases, Alzheimer's disease, and many forms of cancers, the incidence rates are very low in the population, so that in the great majority of families, there is at most one affected case. In real practice, one does occasionally encounter families with multiple affected siblings, or with affected subjects across multiple generations, etc. Such complex family configurations cannot be described using the (x, y) system in this article and it is difficult to design a pooling strategy for them. However, because such families represent a very small portion of the data, subjects in these families can be individually typed without too much inconvenience.

Previously, DNA pooling is usually embedded in a two-stage procedure (5, 8, 9). The first-stage is a DNA pooling study, which scans across the genome for markers that have different frequencies in the pools of cases and controls. The markers identified by the first-stage study are then followed up by a second-stage individual-typing study. Without the follow up, DNA pooling alone will produce an excess of false-positive results, because (i) correction of the measurement errors of DNA pools and the unequal allelic amplification (6, 7) might not be sufficiently accurate and (ii) estimation of the allele frequencies in DNA pools is valid only if Hardy-Weinberg equilibrium holds. The method in this article will maintain the nominal α level exactly even for an error-prone PCR and in a stratified population. This is a property not shared by previous pooling methods (4-9). One may instead argue that some inflation of the type I error rate in the first-stage pooling study is tolerable, or even welcome, because this would identify more markers for the second-stage individual genotyping and hence decrease the overall type II error rate. This is a misconception however. In fact, one can raise the nominal α level of the present method to a level comparable with the inflated α level of the previous pooling methods and achieve the same or even lower level of the overall type II error rate.

As shown in this article, the proposed DNA pooling method can achieve a very high statistical efficiency relative to individual typing. This means that if one can recruit just a few more families for study (e.g., 5-10% more), he/she can rightfully resort to DNA pooling and do away with the costly and laborious individual follow-ups entirely. This single-stage pooling study needs 1 of 8 to 1 of 16 or fewer genotyping efforts compared with an individual-typing study. It should be noted that, by trading more subjects for fewer genotypings, neither the type I nor the type II error rate has to be compromised using the DNA pooling strategy in this article. With the rapid progress in biotechnology, the pooled genotyping is expected in the future to become even more accurate and thus have even higher statistical efficiency. This means that, compared with those described in Tables 2 and 3, more families can be put together and fewer genotyping efforts will be required. On the other hand, it may be that today's costly and laborious individual genotyping will become an easy matter in the future, leaving no cause for a DNA pooling study. Before that day really comes, the pooling strategy presented in this article provides a tradeoff between genotyping efforts and subject recruitment efforts.

References

1. Risch N, Merikangas K. The future of genetic studies of complex human diseases. Science 1996;273:1516–7.
2. Khoury MJ, Yang Q. The future of genetic studies of complex human diseases: an epidemiologic perspective. Epidemiology 1998;9:350–4.
3. Knapp M. A note on power approximation for the transmission/disequilibrium test. Am J Hum Genet 1999;64:1177–85.
4. 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;8:1273–88.
5. Sham P. DNA pooling: a tool for large-scale association studies. Nat Rev Genet 2002;3:862–71.
6. LeHellard S, Ballereau SJ, Visscher PM, et al. SNP genotyping on pooled DNAs: comparison of genotyping technologies and a semi

automated method for data storage and analysis. Nucleic Acids Res 2002;30:e74.
7. Visscher PM, LeHellard S. Simple method to analyze SNP-based association studies using DNA pools. Genet Epidemiol 2003;24:291–6.
8. Zou G, Zhao H. The impacts of errors in individual genotyping and DNA pooling on association studies. Genet Epidemiol 2004;26:1–10.
9. Konig IR, Ziegler A. Analysis of SNPs in pooled DNA: a decision theoretic model. Genet Epidemiol 2004;26:31–43.
10. Spielman RS, McGinnis RE, Ewens WJ. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). Am J Hum Genet 1993;52:506–16.
11. Ewens WJ, Spielman RS. The transmission/disequilibrium test: history, subdivision and admixture. Am J Hum Genet 1995;57:455–64.
12. Weinberg CR, Umbach DM. Choosing a retrospective design to assess joint genetic and environmental contributions to risk. Am J Epidemiol 2000; 152:197–203.
13. Lee W-C. Genetic association studies of adult-onset diseases using the case-spouse and case-offspring designs. Am J Epidemiol 2003;158:1023–32.
14. Weinberg CR. Invited commentary: making the most of genotype asymmetries. Am J Epidemiol 2003;158:1033–5.
15. Lee W-C. Lee responds to ''making the most of genotype asymmetries''. Am J Epidemiol 2003;158:1036–8.