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 gene(s) of a complex disease 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 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 the $j$th 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 $F_j$. Then we calculate for all the families in the $j$th 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 $k$th pool containing each and every one of the $k$th eldest unaffected siblings of the families in the $j$th set. The measured allele frequency for the $k$th sibling pool in the $j$th set is denoted as $B_{jk}$. And we calculate for all the families in the $j$th 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) $D_j^I = 2n_jC_j - 2n_j(2/xy \cdot \sum_{k=1}^{x_j} B_{jk} - C_j) = 4n_j(C_j - 1/xy \cdot \sum_{k=1}^{x_j} B_{jk})$.

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 $S$. We calculate for all the families in the $j$th pooling set, $S_j = \text{total allele counts for the affected cases} - \text{total allele counts for the spouses}$.
counts for the spouses = 2n(Cj − Sj). If spouses are not available but offspring are (0 < yj < ∞), we form a total of yj “offspring pools” (l = 1, . . . , yj), with the lth pool containing each and every one of the lth eldest offspring of the families in the lth set. The measured allele frequency for the lth offspring pool in the lth set is denoted as Ojl. And we calculate for all the families in the lth pooling set, Dj = total allele counts for the affected cases − total allele counts for the “imputed spouses” (see ref. 13, 15)

\[ \text{Dj} = 2n(Cj - 2n(2/yj \cdot \sum_{j=1}^{yj} Ojl - Cj) = 4n(Cj - 1/yj \cdot \sum_{j=1}^{yj} Ojl). \]

**Expected Total Allele Differences**

If the study population is a homogenous 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 Dj and Dl 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 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.

<table>
<thead>
<tr>
<th>Allele frequency</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actual</td>
<td>measured</td>
</tr>
<tr>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>0.125</td>
<td>0.050</td>
</tr>
<tr>
<td>0.250</td>
<td>0.150</td>
</tr>
<tr>
<td>0.375</td>
<td>0.200</td>
</tr>
<tr>
<td>0.500</td>
<td>0.050</td>
</tr>
<tr>
<td>0.625</td>
<td>0.000</td>
</tr>
<tr>
<td>0.750</td>
<td>0.000</td>
</tr>
<tr>
<td>0.875</td>
<td>0.000</td>
</tr>
<tr>
<td>1.000</td>
<td>1.000</td>
</tr>
</tbody>
</table>

*Expected allele frequency*:

*Actual* = measured = 0.4000, 0.4000, 0.3885, 0.3885

*Measured* = 0.4000, 0.4000, 0.3885, 0.3885

**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**

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.

**Dissequilibrium 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 dissequilibrium test for pooled data is constructed below

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

For large \( J \) (e.g., \( J > 30 \)), \( Z^2 \) is distributed as a one-degree-of-freedom \( \chi^2 \) 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.,

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

and

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

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

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

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 (\( \infty, 0 \)).

**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

\[ \text{V}_{\text{individual}} = 4n_P(1 - P)(w_j^I + w_j^{II}), \]

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., \( \text{V}_{\text{pooled}} = \text{V}_{\text{individual}} + \text{V}_{\text{quantitative PCR}} \)). To calculate \( \text{V}_{\text{quantitative PCR}} \), we assume
that the measurement error of the quantitative PCR in estimating the allele frequency of a DNA pool is a constant \( \sigma \), 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

\[
\begin{align*}
\frac{w_j^D}{D_j^H} + \frac{w_j^H}{D_j^H} &= 4n_j w_j^I \cdot [C_j - (F_j + M_j)/2] + 4n_j w_j^I \cdot [C_j - 1/3] \\
&\quad \cdot (O_1 + O_2 + O_3)] \\
&= 4n_j (w_j^I + w_j^H) \cdot C_j - 2n_j w_j^I \cdot (F_j + M_j) - 4n_j w_j^H/3 \\
&\quad \cdot (O_1 + O_2 + O_3).
\end{align*}
\]

And the variance associated with the quantitative PCR is

\[
16n_j (w_j^I + w_j^H)^2 \cdot \sigma^2 + 4n_j w_j^I w_j^H \cdot (\sigma^2 + \sigma^2) \\
+16/9 \cdot n_j^2 w_j^I w_j^H \cdot (\sigma^2 + \sigma^2).
\]

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

\[
V_j^{\text{quantitative PCR}} = 16n_j \sigma^2 \left\{ [w_j^I + (1 - I_{(y_j<∞)})/2]w_j^I ]^2 \\
+ (I_{(y_j<∞)} + (I_{(y_j>0)})/x_j)w_j^I w_j^H \\
+ (I_{(y_j<∞)} + (I_{(y_j>0)})/y_j)w_j^I w_j^H \right\},
\]

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 \(\Sigma \psi_j^2\) individual \(/\Sigma \psi_j^2\) 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.,

\[
\frac{n_j}{1 - c} \leq \frac{P(1 - P)}{4\sigma^2} \times \\
\frac{w_j^I + w_j^H}{w_j^I + (1 - I_{(y_j<∞)})/2}w_j^I ]^2 + (I_{(y_j<∞)} + (I_{(y_j>0)})/x_j)w_j^I w_j^H \\
+ (I_{(y_j<∞)} + (I_{(y_j>0)})/y_j)w_j^I w_j^H.
\]

Note that the above calculations were carried out under the null hypothesis, with a constant \(\sigma\), 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 \(\sigma = 0.01\) (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 \((x, y)\) 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

<table>
<thead>
<tr>
<th>No. line I relatives ((x))</th>
<th>No. line II relatives ((y))</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>∞ (Spouse)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency of the minor allele = 0.1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>11.8 (8.4)</td>
<td>11.8 (8.4)</td>
<td>11.8 (8.4)</td>
<td>11.8 (8.4)</td>
<td>11.8 (8.4)</td>
<td>23.7 (4.2)</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>9.9 (10.1)</td>
<td>9.7 (10.3)</td>
<td>9.7 (10.3)</td>
<td>9.7 (10.3)</td>
<td>9.7 (10.3)</td>
<td>17.5 (5.7)</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>9.7 (10.6)</td>
<td>9.5 (10.6)</td>
<td>9.4 (10.6)</td>
<td>9.4 (10.6)</td>
<td>9.4 (10.6)</td>
<td>16.1 (6.2)</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>9.4 (10.6)</td>
<td>9.3 (10.7)</td>
<td>9.3 (10.7)</td>
<td>9.3 (10.7)</td>
<td>9.3 (10.7)</td>
<td>15.6 (6.4)</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>9.3 (10.7)</td>
<td>9.2 (10.9)</td>
<td>9.2 (10.9)</td>
<td>9.2 (10.9)</td>
<td>9.2 (10.9)</td>
<td>15.3 (6.5)</td>
<td></td>
</tr>
<tr>
<td>∞ (Parents)</td>
<td>7.9 (12.7)</td>
<td>7.7 (13.0)</td>
<td>7.7 (12.9)</td>
<td>7.7 (12.9)</td>
<td>7.8 (12.9)</td>
<td>11.8 (8.4)</td>
<td></td>
</tr>
<tr>
<td>Frequency of the minor allele = 0.3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>27.6 (3.6)</td>
<td>27.6 (3.6)</td>
<td>27.6 (3.6)</td>
<td>27.6 (3.6)</td>
<td>27.6 (3.6)</td>
<td>55.3 (1.8)</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>23.0 (4.3)</td>
<td>22.7 (4.4)</td>
<td>22.6 (4.4)</td>
<td>22.6 (4.4)</td>
<td>22.6 (4.4)</td>
<td>40.7 (2.5)</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>22.7 (4.4)</td>
<td>22.1 (4.5)</td>
<td>21.9 (4.6)</td>
<td>21.9 (4.6)</td>
<td>21.9 (4.6)</td>
<td>37.7 (2.7)</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>22.6 (4.4)</td>
<td>21.9 (4.6)</td>
<td>21.7 (4.6)</td>
<td>21.7 (4.6)</td>
<td>21.7 (4.6)</td>
<td>36.4 (2.7)</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>22.6 (4.4)</td>
<td>21.9 (4.6)</td>
<td>21.6 (4.6)</td>
<td>21.6 (4.6)</td>
<td>21.6 (4.6)</td>
<td>35.7 (2.8)</td>
<td></td>
</tr>
<tr>
<td>∞ (Parents)</td>
<td>18.4 (5.4)</td>
<td>18.0 (5.6)</td>
<td>18.0 (5.5)</td>
<td>18.1 (5.5)</td>
<td>18.1 (5.5)</td>
<td>27.6 (3.6)</td>
<td></td>
</tr>
<tr>
<td>Frequency of the minor allele = 0.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>32.9 (3.0)</td>
<td>32.9 (3.0)</td>
<td>32.9 (3.0)</td>
<td>32.9 (3.0)</td>
<td>32.9 (3.0)</td>
<td>65.8 (1.5)</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>32.9 (3.0)</td>
<td>27.4 (3.6)</td>
<td>27.0 (3.7)</td>
<td>26.9 (3.7)</td>
<td>26.9 (3.7)</td>
<td>48.5 (2.1)</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>32.9 (3.0)</td>
<td>27.0 (3.7)</td>
<td>26.3 (3.8)</td>
<td>26.1 (3.8)</td>
<td>26.0 (3.8)</td>
<td>44.9 (2.2)</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>32.9 (3.0)</td>
<td>26.9 (3.7)</td>
<td>26.1 (3.8)</td>
<td>25.8 (3.9)</td>
<td>25.7 (3.9)</td>
<td>43.3 (2.3)</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>32.9 (3.0)</td>
<td>26.9 (3.7)</td>
<td>26.0 (3.8)</td>
<td>25.7 (3.9)</td>
<td>25.6 (3.9)</td>
<td>42.4 (2.4)</td>
<td></td>
</tr>
<tr>
<td>∞ (Parents)</td>
<td>21.9 (4.6)</td>
<td>21.1 (4.7)</td>
<td>21.5 (4.6)</td>
<td>21.5 (4.6)</td>
<td>21.6 (4.6)</td>
<td>32.9 (3.0)</td>
<td></td>
</tr>
</tbody>
</table>
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 been identified by the first-stage study are then followed 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 level exactly even for the type I and II error rate. This is a misconception however. In fact, one can raise the nominal level of the present method to a level comparable with the inflated \(\alpha\) 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**

6. LeHillard S, Ballereau SJ, Vischer PM, et al. SNP genotyping on pooled DNAs: comparison of genotyping technologies and a semi-
A DNA Pooling Strategy for Family-Based Association Studies

Wen-Chung Lee


Updated version
Access the most recent version of this article at:
http://cebp.aacrjournals.org/content/14/4/958

Cited articles
This article cites 15 articles, 2 of which you can access for free at:
http://cebp.aacrjournals.org/content/14/4/958.full#ref-list-1

E-mail alerts
Sign up to receive free email-alerts related to this article or journal.

Reprints and Subscriptions
To order reprints of this article or to subscribe to the journal, contact the AACR Publications Department at pubs@aacr.org.

Permissions
To request permission to re-use all or part of this article, contact the AACR Publications Department at permissions@aacr.org.