

On the Use of DNA Pooling to Estimate Haplotype Frequencies

Shuang Wang,¹ Kenneth K. Kidd,² Hongyu Zhao^{1,2*}

¹Department of Epidemiology and Public Health, Yale University School of Medicine, New Haven, Connecticut

²Department of Genetics, Yale University School of Medicine, New Haven, Connecticut

Genome-wide association studies may be necessary to identify genes underlying certain complex diseases. Because such studies can be extremely expensive, DNA pooling has been introduced, as it may greatly reduce the genotyping burden. Parallel to DNA pooling developments, the importance of haplotypes in genetic studies has been amply demonstrated in the literature. However, DNA pooling of a large number of samples may lose haplotype information among tightly linked genetic markers. Here, we examine the cost-effectiveness of DNA pooling in the estimation of haplotype frequencies from population data. When the maximum likelihood estimates of haplotype frequencies are obtained from pooled samples, we compare the overall cost of the study, including both DNA collection and marker genotyping, between the individual genotyping strategy and the DNA pooling strategy. We find that the DNA pooling of two individuals can be more cost-effective than individual genotypings, especially when a large number of haplotype systems are studied. *Genet Epidemiol* 24:74–82, 2003. © 2003 Wiley-Liss, Inc.

Key words: haplotypes; genotypes; DNA pooling; expectation-maximization algorithm

Grant sponsor: National Institute of Health, Grant numbers: GM59507, GM57672.

*Correspondence to: Hongyu Zhao, Ph.D., Department of Epidemiology and Public Health, Yale University School of Medicine, 60 College St., New Haven, CT 06520-8034. E-mail: hongyu.zhao@yale.edu

Received for publication 8 May 2002; Revision accepted 25 July 2002

Published online in Wiley InterScience (www.interscience.wiley.com).

DOI: 10.1002/gepi.10195

INTRODUCTION

Most common disorders do not display simple patterns of Mendelian inheritance and are believed to be the results of multiple interacting contributing loci. Because the contribution of any specific gene in a complex trait may be small, association studies using a large number of genetic markers have been proposed as a viable alternative to the more traditional linkage studies [Risch and Merikangas, 1996; Risch, 2000]. Various estimates of the number of markers needed in a genome-wide association study have been discussed in the literature [Kruglyak, 1999; Wang et al., 1998]. Even when only candidate genes are typed, the number of genes potentially involved in the disease etiology can be large. Therefore, consider a study involving hundreds or thousands of individuals: it is necessary to develop efficient methods for genotyping a large number of markers among a large number of individuals to make the genome scan (or even the candidate gene search) feasible. One approach that can greatly reduce genotyping efforts is DNA pooling.

The idea of using pooled DNA samples to reduce the genotyping burden was first introduced by Arnheim et al. [1985] in the context of case-control studies. In that study, pooled DNA samples from insulin-dependent diabetes mellitus (IDDM) patients were compared with pooled DNA samples from randomly selected controls. The authors argued that some individual polymorphic fragments correlated with IDDM susceptibility loci were increased in the pooled sample of patients, in comparison with the pooled sample of controls [Arnheim et al., 1985]. Using this approach, they detected specific DR and DQ alleles elevated in the IDDM population. Patek et al. [1993] developed a method that determines the population allele frequencies at loci by quantitative analysis of DNA amplified from pooled samples. The authors argued that the distribution of polymerase chain reaction (PCR) products obtained from the alleles present in the pooled samples directly corresponded to the allele frequency in the population. Furthermore, they found that the allele frequencies determined by quantitative analysis of PCR products from pooled DNA

samples agreed quite well with the allele frequencies determined by analyzing individual samples. Shaw et al. [1998] quantitatively tested the accuracy of estimated allele frequencies in comparison with that estimated by direct genotyping, which showed that not only can the allele frequencies estimated from DNA pools be quantitative and accurate, but also the results are reproducible in that the replicate-to-replicate variation is small.

The DNA pooling strategy has been proposed for linkage studies [Churchill et al., 1993; Darvasi and Soller, 1994], association studies [Daniels et al., 1998; Shaw et al., 1998; Risch and Teng, 1998], and physical-mapping studies [Barillot et al., 1991; Bruno et al., 1995]. This strategy has been shown to be quite effective in identifying disease-causing loci with Mendelian founder mutations [Cami et al., 1995; Barcellos et al., 1997; Amos et al., 2000] and complex diseases [Arnheim et al., 1985].

More recently, Pfeiffer et al. [2002] discussed efficient methods of analyzing pooled DNA samples with qualitative assays to estimate the joint prevalence of variants at two or more loci. Some qualitative assays can only determine whether certain genotypes are present in a pooled sample, whereas other qualitative assays can only determine whether a particular allele is present in the pool.

With the availability of hundreds of thousands of single-nucleotide polymorphisms (SNPs), haplotype analysis methods have become increasingly important for linkage-disequilibrium assessment and association studies of candidate genes. Haplotype methods rely on phase information from the individuals under study, which can be established by genotyping family members of each study subject to infer parental chromosomes, or by laboratory techniques [e.g., Michalatos-Beloin et al., 1996]. However, these two approaches may not be feasible, as the first requires sampling and genotyping of relatives which may be not available, and the second is usually technically demanding and cost-prohibitive. Therefore, several statistical methods for haplotype inference and/or estimating haplotype frequencies from genotype data have been proposed, including a sequential haplotype inference algorithm [Clark, 1990], the expectation-maximization (EM) algorithm [Excoffier and Slatkin, 1995; Hawley and Kidd, 1995; Long et al., 1995], and coalescent model-based methods [Stephens et al., 2001]. Haplotype frequency estimates via the EM algorithm have been found to be quite accurate,

both in simulation studies [Fallin and Schork, 2000] and with molecularly haplotyped data [Tishkoff et al., 2000], and comparisons have been made for haplotype reconstructions between the EM algorithm-based method and the coalescent model-based method [Stephens et al., 2001; Zhang et al., 2001].

Despite the usefulness of the DNA pooling strategy, its application in haplotype analysis has only been explored in the literature for qualitative assays [Pfeiffer et al., 2002]. Although DNA pools of a large number of individuals will likely result in the loss of haplotype information, if haplotype information can be well-recovered from certain DNA pooling designs, there can be substantial savings in genotyping costs, especially when a large number of genes are studied. Here, we consider the cost-effectiveness of obtaining accurate haplotype frequency estimates from individually genotyped data and pooled DNA samples, with the objective of identifying a cost-effective approach to estimating haplotype frequencies from population-based samples. We conclude that the DNA pooling strategy can be a more cost-effective alternative for haplotype frequency estimates.

METHODS

HAPLOTYPE FREQUENCY ESTIMATES AND THEIR VARIANCES

We consider maximum likelihood estimates of haplotype frequencies from both individually typed markers and pooled DNA samples. We refer to typing results as phenotypes. Although for diploid individuals, one usually assumes strict codominant inheritance, so that the phenotype and genotype have a 1:1 mapping, null alleles, known to occur for some systems typed using PCRs, cause exceptions. In the case of two SNPs (with alleles A and a at the first marker and alleles B and b at the second marker), for individually typed markers, there are a total of nine possible genotypes not considering phase, i.e., (AA, BB), (AA, Bb), (AA, bb), (Aa, BB), (Aa, Bb), (Aa, bb), (aa, BB), (aa, Bb), and (aa, bb). For pooled samples with two individuals in each pool, there are a total of 25 possible typing results across these two markers, as the number of copies of a marker allele can vary from 0–4 at each marker. Similarly, for pooled samples with three individuals, there are 49 possible typing results. The log-likelihood

of the observed set of pooled samples is

$$\log L = \sum_{i=1}^m \{n_i \log(P_i)\},$$

where m is the number of distinct marker phenotypes observed, n_i is the number of pools with the i th observed marker phenotype, where we have $\sum_{i=1}^m n_i = N$, the total number of pools, and P_i is the probability for the i th marker phenotype. Under a given set of haplotype frequencies and Hardy-Weinberg equilibrium, P_i can be calculated as the function of haplotype frequencies. For example, in the case of two markers, P_i can be calculated as a function of three haplotype frequencies: P_{AB} , P_{Ab} , and P_{aB} , as there is the constraint of $P_{ab} = 1 - P_{AB} - P_{Ab} - P_{aB}$, i.e.,

$$\log L(P_{AB}, P_{Ab}, P_{aB}) = \sum_{i=1}^m \{n_i \log[P_i(P_{AB}, P_{Ab}, P_{aB})]\},$$

where $n_i = N \times P_i$, N is the total number of pools, $m=9$ for individual typings, $m=25$ for pooled samples with two individuals, and $m=49$ for pooled samples with three individuals. Based on this likelihood function, haplotype frequencies can be estimated by maximum likelihood estimates, using the EM algorithm (Appendix).

The accuracy for estimates can be derived from the Fisher information matrix. The Fisher information matrix was calculated based on the above likelihood, using the software package Mathematica (Wolfram Research, Inc.). From the Fisher information matrix [Wolfram, 1999], we can obtain the asymptotic variance of the haplotype frequency estimates of P_{AB} , P_{Ab} , P_{aB} . The asymptotic variance of haplotype frequency P_{ab} can be obtained via:

$$\begin{aligned} \text{Var}(P_{ab}) &= \text{Var}(P_{AB}) + \text{Var}(P_{Ab}) + \text{Var}(P_{aB}) \\ &+ 2^*[\text{Cov}(P_{AB}, P_{Ab}) + \text{Cov}(P_{AB}, P_{aB}) \\ &+ \text{Cov}(P_{Ab}, P_{aB})]. \end{aligned}$$

The approach is the same when there are three or more SNPs.

To examine the adequacy of the asymptotic variance estimates from the Fisher information matrix, we performed simulation experiments to compare the variance of the estimated haplotype frequencies from repeated simulations to the asymptotic variance estimates. For a given genetic system (defined by number of markers and haplotype frequencies), we first calculated the

probability of each phenotype under a given typing strategy (individuals or DNA pools). For each simulated data set from a multinomial distribution, we used the EM algorithm to estimate haplotype frequencies, which maximize the probability of obtaining the observed marker phenotypes. In our analysis, we assume Hardy-Weinberg equilibrium. The empirical variances of haplotype frequency estimates were then calculated and compared to the asymptotic variances estimated from the Fisher information matrix.

GENETIC SYSTEMS STUDIED

We consider the case of two SNPs, in which the haplotype frequencies are determined by the allele frequencies at the two markers, P_A and P_B , and the linkage disequilibrium between these two markers, D' , defined as standardized D , the classic disequilibrium coefficient defined for two loci as the frequency of the haplotype minus the expected frequency that is the product of the frequencies of the individual alleles, to range from -1 to $+1$. More specifically,

$$D' = \begin{cases} \frac{P_{AB} - P_A P_B}{\min(P_A P_b, P_a P_B)}, & P_{AB} - P_A P_B < 0 \\ \frac{P_{AB} - P_A P_B}{\min(P_A P_b, P_a P_B)}, & P_{AB} - P_A P_B > 0 \end{cases}$$

All four haplotype frequencies can be derived from these three parameters. For example, when $D' > 0$,

$$\begin{aligned} P_{AB} &= P_A P_B + \min(P_A P_b, P_a P_B) \times D'; \\ P_{Ab} &= P_A P_b - \min(P_A P_b, P_a P_B) \times D'; \\ P_{aB} &= P_a P_B - \min(P_A P_b, P_a P_B) \times D'; \\ P_{ab} &= P_a P_b + \min(P_A P_b, P_a P_B) \times D'; \end{aligned}$$

We varied the allele frequencies for locus A and B from 0.1, to 0.3, to 0.5, producing six distinguishable allele frequency combinations at the two SNPs. The D' was varied over -1 , -0.75 , -0.5 , -0.25 , 0 , 0.25 , 0.5 , 0.75 , to 1 . We also considered three marker cases through empirical population genetics data. However, as the results are less interpretable and the complexity in haplotype analysis increases rapidly with the number of loci investigated, we do not summarize our results here. Interested readers can request the details of the simulation setups and results from us.

COST-EFFECTIVENESS COMPARISONS

We consider two designs to yield the same level of accuracy if the largest standard error among all haplotype frequency estimates is the same between the two designs. When we set a common goal, e.g., the maximum standard error is below 1%, i.e.,

$$\max\{se(P_{AB}), se(P_{Ab}), se(P_{aB}), se(P_{ab})\} \leq 0.01$$

in the case of two SNPs, we can calculate the required number of pools to achieve such accuracy through the asymptotic variance from the inverse of the Fisher information matrix. Let $N_{pool,k}$ denote the number of pools needed for DNA pools of k individuals, where $k=1$ corresponds to individual typings. The total number of subjects needed is then

$$N_k = N_{pool,k} \times k.$$

Let $\text{cost}_{\text{sampling}}$ denote the cost of collecting DNA from one individual, and $\text{cost}_{\text{typing}}$ denote the cost of genotyping per pool for up to two SNPs. Then, the total cost for DNA pools of k individuals is:

$$\begin{aligned} & \text{cost}_{\text{sampling}} \times N_k + \text{cost}_{\text{typing}} \times N_k/k \\ & = \text{cost}_{\text{typing}} \times N_{pool,k} \times (rk + 1), \end{aligned}$$

where $r = \text{cost}_{\text{sampling}} / \text{cost}_{\text{typing}}$. We assume the cost of typing DNA is the same for each DNA sample. Therefore, by fixing the typing cost and changing the ratio between the sampling cost and typing cost, we can evaluate cost-effectiveness of different designs as cost ratios.

We make the assumption that it is possible to determine unambiguously the number of copies of each allele in a pool, i.e., for pools of two subjects, 0–4 alleles are distinguishable, and for pools of three subjects, 0–6 alleles are distinguishable.

RESULTS

COMPARISONS BETWEEN EMPIRICAL VARIANCES AND ASYMPTOTIC VARIANCES

The assessments of the approximation of asymptotic variances for haplotype frequency estimates from the Fisher information matrix are summarized in Table I for two sets of haplotype frequencies. In the comparisons, we fix the total number of subjects at 600, which means that $N_{pool,1}=600$, $N_{pool,2}=300$, and $N_{pool,3}=200$ for pools with size $k=1, 2$, and 3. The results show that for sample sizes of this range, the asymptotic variances provide excellent approximations to the true variances of haplotype frequency estimates. Therefore, we use the Fisher information matrix to estimate the required sample sizes to achieve a given accuracy for haplotype frequency estimates.

TWO SNPS

For various marginal allele frequencies and linkage disequilibrium among the markers, the number of DNA pools needed so that the maximum standard error for haplotype frequency estimates is 0.01 is summarized in Table II. The results show that, for most cases, the number of DNA pools with two individuals is slightly greater than one half of the number of pools needed for individual typings. The number of pools with three individuals exceeds one third of the number of individual typings. In most cases, the total number of subjects needed is the greatest for DNA pools with three individuals and the least for typing of individuals. When $D' = -1$ and 1, all three designs require the same number of subjects at all allele frequency combinations.

TABLE I. Mean and standard errors of haplotype frequencies from simulations and from Fisher's information matrix for pools, as a function of pool size k , and true generating haplotype frequencies for a sample of 600 individuals

Case	Pool size k	Mean estimate (\hat{P})				Standard error of \hat{P} from simulation				Standard error of \hat{P} from Fisher's information			
		P_{AB}	P_{Ab}	P_{aB}	P_{ab}	P_{AB}	P_{Ab}	P_{aB}	P_{ab}	P_{AB}	P_{Ab}	P_{aB}	P_{ab}
1	True=	0.02	0.08	0.18	0.72								
	1	0.020	0.081	0.179	0.720	0.005	0.008	0.012	0.013	0.005	0.009	0.012	0.013
	2	0.020	0.081	0.179	0.721	0.007	0.010	0.012	0.015	0.007	0.010	0.013	0.014
	3	0.021	0.079	0.181	0.719	0.007	0.010	0.013	0.015	0.009	0.011	0.014	0.015
2	True=	0.06	0.04	0.14	0.76								
	1	0.060	0.040	0.140	0.760	0.008	0.006	0.010	0.013	0.007	0.006	0.010	0.013
	2	0.059	0.041	0.141	0.759	0.007	0.007	0.012	0.013	0.008	0.007	0.011	0.013
	3	0.059	0.041	0.140	0.760	0.010	0.008	0.011	0.013	0.009	0.008	0.012	0.014

TABLE II. Number of pools needed, together with number of subjects needed in parentheses, to obtain haplotype frequency standard errors less than 0.01

Allele frequency (A, B)	D'								
	-1	-0.75	-0.5	-0.25	0	0.25	0.5	0.75	1
Pool with size k=1									
(0.1, 0.1)	800 (800)	820 (820)	822 (822)	817 (817)	810 (810)	723 (723)	631 (631)	541 (541)	450 (450)
(0.1, 0.3)	1,200 (1,200)	1,239 (1,239)	1,260 (1,260)	1,266 (1,266)	1,260 (1,260)	1,222 (1,222)	1,167 (1,167)	1,108 (1,108)	1,050 (1,050)
(0.1, 0.5)	1,250 (1,250)	1,291 (1,291)	1,325 (1,325)	1,347 (1,347)	1,350 (1,350)	1,347 (1,347)	1,325 (1,325)	1,291 (1,291)	1,250 (1,250)
(0.3, 0.3)	1,200 (1,200)	1,298 (1,298)	1,387 (1,387)	1,448 (1,448)	1,470 (1,470)	1,399 (1,399)	1,273 (1,273)	1,157 (1,157)	1,050 (1,050)
(0.3, 0.5)	1,250 (1,250)	1,303 (1,303)	1,365 (1,365)	1,412 (1,412)	1,400 (1,400)	1,412 (1,412)	1,365 (1,365)	1,303 (1,303)	1,250 (1,250)
(0.5, 0.5)	1,250 (1,250)	1,253 (1,253)	1,273 (1,273)	1,306 (1,306)	1,250 (1,250)	1,306 (1,306)	1,273 (1,273)	1,253 (1,253)	1,250 (1,250)
Pool with size k=2									
(0.1, 0.1)	400 (800)	450 (900)	451 (902)	449 (898)	446 (892)	391 (782)	328 (656)	273 (546)	225 (450)
(0.1, 0.3)	600 (1,200)	696 (1,392)	719 (1,438)	726 (1,452)	725 (1,450)	699 (1,398)	649 (1,298)	588 (1,176)	525 (1,050)
(0.1, 0.5)	625 (1,250)	711 (1,422)	759 (1,518)	783 (1,566)	788 (1,576)	783 (1,566)	759 (1,518)	711 (1,422)	625 (1,250)
(0.3, 0.3)	600 (1,200)	774 (1,548)	879 (1,758)	937 (1,874)	956 (1,912)	881 (1,762)	723 (1,446)	597 (1,194)	525 (1,050)
(0.3, 0.5)	625 (1,250)	737 (1,474)	863 (1,726)	949 (1,898)	963 (1,926)	949 (1,898)	863 (1,726)	737 (1,474)	625 (1,250)
(0.5, 0.5)	625 (1,250)	654 (1,308)	767 (1,534)	916 (1,832)	938 (1,876)	916 (1,832)	767 (1,534)	654 (1,308)	625 (1,250)
Pool with size k=3									
(0.1, 0.1)	267 (800)	327 (981)	328 (984)	327 (981)	324 (972)	282 (846)	229 (687)	184 (552)	150 (450)
(0.1, 0.3)	400 (1,200)	522 (1,566)	540 (1,620)	547 (1,641)	546 (1,638)	525 (1,575)	482 (1,446)	421 (1,263)	350 (1,050)
(0.1, 0.5)	417 (1,250)	531 (1,593)	574 (1,722)	595 (1,785)	600 (1,800)	595 (1,785)	574 (1,722)	531 (1,593)	417 (1,250)
(0.3, 0.3)	400 (1,200)	617 (1,851)	713 (2,139)	767 (2,301)	784 (2,352)	710 (2,130)	549 (1,647)	412 (1,236)	350 (1,050)
(0.3, 0.5)	417 (1,250)	563 (1,689)	702 (2,106)	795 (2,385)	817 (2,451)	795 (2,385)	702 (2,106)	563 (1,689)	417 (1,250)
(0.5, 0.5)	417 (1,250)	457 (1,371)	609 (1,827)	785 (2,355)	834 (2,502)	785 (2,355)	609 (1,827)	457 (1,371)	417 (1,250)

To compare the total cost of pooling with one, two, or three individuals, the cost of typing one DNA sample is fixed at one unit. The ranges of the total costs for six combinations of allele frequencies are obtained by changing the ratio of the sampling cost to the genotyping cost. The results from several allele frequency combinations are summarized in Figures 1–3. We note that when the genotyping cost is much higher than the sampling cost, the DNA pooling strategy is much more cost-effective than the individual genotyping method. When the genotyping cost is high, although there may be substantial

cost savings by using DNA pools with two individuals vs. individual genotypes, there are not much further savings from pools with three individuals. As the ratio of the sampling cost to the genotyping cost increases, i.e., when it is much more expensive to sample subjects from the population than to genotype DNA samples, the individual genotyping method becomes more cost-effective than the DNA pooling strategy, which is consistent with what we expect. When both SNP allele frequencies are common, the pooling strategy is not cost-effective over most cost ratios.

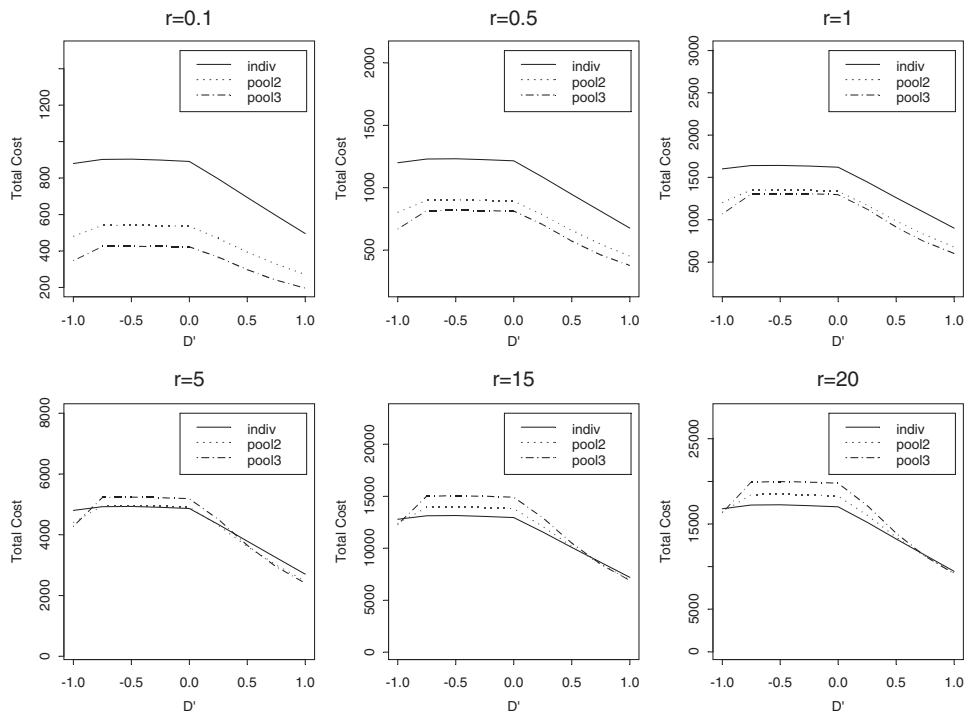


Fig. 1. Total cost at different r , ratio of sampling cost to typing cost with $(P_A, P_B)=(0.1, 0.1)$. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com].

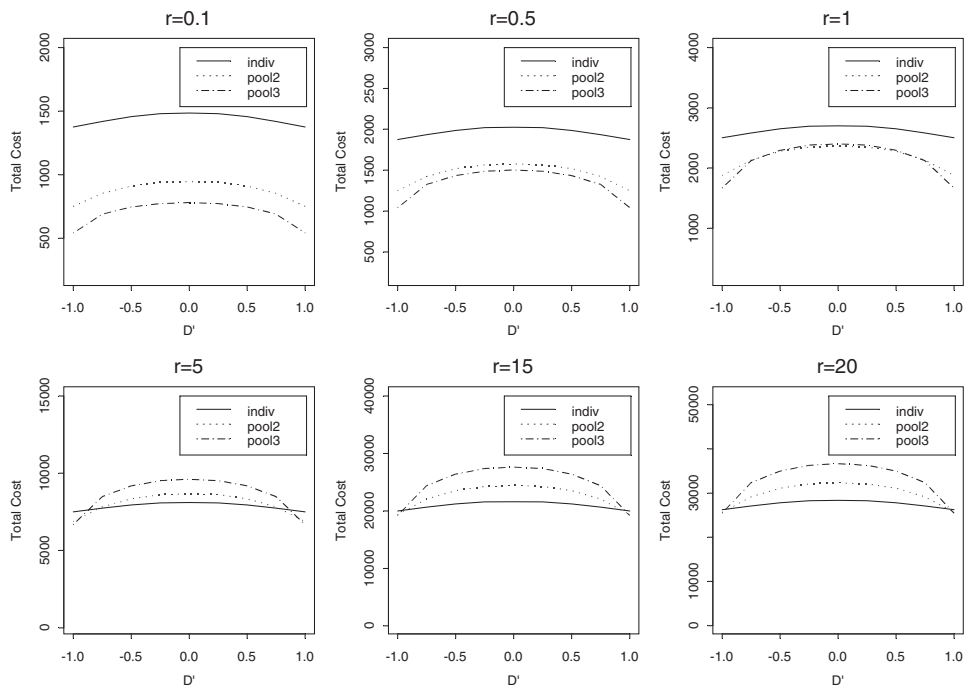


Fig. 2. Total cost at different r , ratio of sampling cost to typing cost with $(P_A, P_B)=(0.1, 0.5)$. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com].

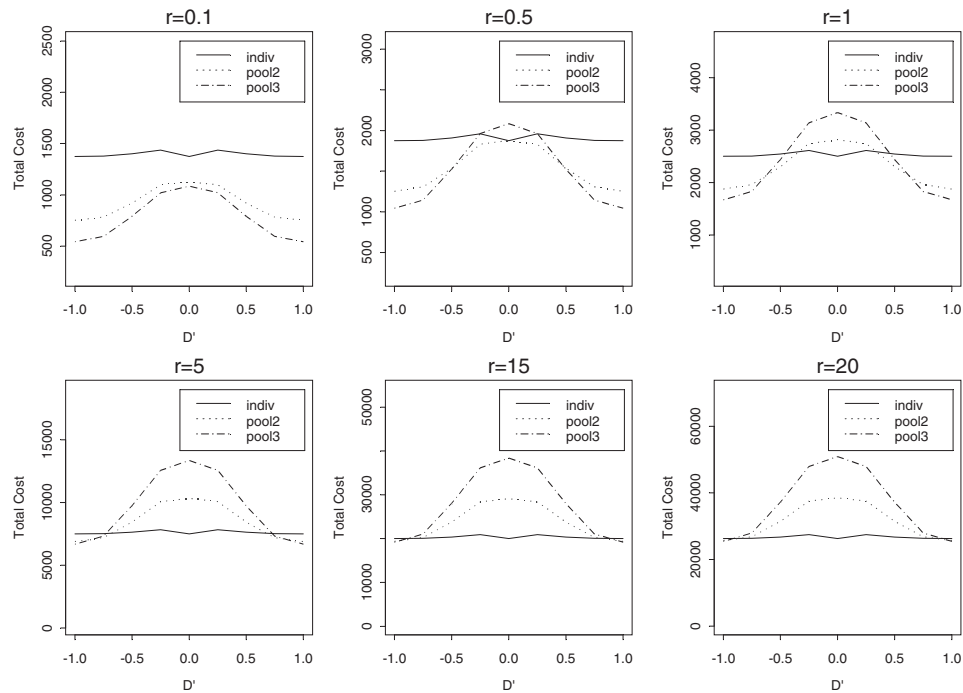


Fig. 3. Total cost at different r , ratio of sampling cost to typing cost with $(P_A, P_B) = (0.5, 0.5)$. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com].

DISCUSSION

We investigated the cost-effectiveness of the DNA pooling strategy vs. the individual genotyping method in the context of estimating haplotype frequencies from population data. The asymptotic variances from the Fisher information matrix are used to obtain the required sample sizes to achieve a prespecified accuracy. Simulations show that, for the sample sizes considered in this article, the asymptotic variances provide excellent approximation to the variances of haplotype frequency estimates.

Our results show that, in most cases, the DNA pooling strategy is more cost-effective when the genotyping cost is higher than or the same as the sampling cost, whereas the individual genotyping method is more cost-effective when the sampling cost is much higher than the genotyping cost. We can also study the changing point of r , which is the ratio of the sampling cost to the genotyping cost, at which the cost efficiency reverses. For the two SNP cases, it changes with the allele frequency combinations, with a much larger r when the rare allele frequency is low. When the rare allele has a

higher frequency, the individual genotyping method is more cost-effective than the DNA pooling strategy when r is relatively small.

One interesting finding from our study is that, for haplotypes with two SNPs, there is not much improvement from pooling two individuals to pooling three individuals, but the cost savings of pooling two individuals can be substantial compared to individual genotyping, especially when the sampling cost is not significantly higher than the genotyping cost. This is one reason that we did not consider pooling with more than three individuals. Another reason is that the typing results may be less interpretable when four or more individuals are typed in a single pool.

For population-based studies, although there may be considerable cost to establish the cell lines in the first place, such cost may be only a fraction of the genotyping cost if a large number of markers are genotyped. In this case, our results show that collecting more samples in the beginning and using DNA pools with two individuals is a more cost-effective strategy for estimating haplotype frequencies from population samples. We did not take into account the limiting factor in the use of the pooling strategy, namely, the

specificity and sensitivity of the quantitative assays, by which we mean the probability of detecting the number of variants at a particular locus, given the number of variants at that locus. When the pooling size is less than or equal to 5 for detecting microsatellite genotypes, adequate sensitivity can be ensured [Coolbaugh-Murphy et al., 1999]. For qualitative assays, Pfeiffer et al. [2002] discussed the influence of sensitivity and specificity on the efficiency of joint allele frequencies and linkage disequilibrium estimation. They found that there may be up to 90% reduction in assays for pool size $k=10$. Even for $k=2$, the number of required assays could be reduced by 50%. Their results are consistent with our findings, in that DNA pooling can be a useful strategy for haplotype analysis.

In our study, we did not investigate how to categorize subjects into pools. Factors such as disease status, gender, or other clinical variables can be used for categorization. For case-control studies, case individuals and control individuals may be pooled separately into DNA pools, and haplotype frequencies can be compared between cases and controls using these DNA pooled samples. The relative efficiency and the cost-effectiveness of different strategies will be reported in future studies.

APPENDIX: EM ALGORITHM TO ESTIMATE HAPLOTYPE FREQUENCIES FROM POOLED DNA SAMPLES

Using the same notation as in the text, let m denote the number of distinct marker phenotypes observed, and n_i denote the number of pools with the i th observed marker phenotype. The EM algorithm proceeds as follows.

1. Step 1: Start with a set of initial haplotype frequency estimates.
2. Step 2 (E-step): For each marker phenotype i , let s_i denote the number of haplotype sets compatible with the marker phenotypes. For example, for a two-sample pool with marker phenotype (AAAa, BBBb), there are two sets of haplotypes consistent with this marker phenotype, namely, (AB, AB, AB, ab) and (AB, AB, Ab, aB). Let P_{ij} denote the probability of observing the j th haplotype set, where $j = 1, \dots, s_i$. The expected number of haplotype h in the overall sample is

calculated as

$$\hat{n}_h = \sum_{i=1}^m \sum_{j=1}^{s_i} \frac{P_{ij} \times X(h)_{ij}}{\sum_{l=1}^{s_i} P_{il}}.$$

Where $X(h)_{ij}$ is the number of haplotype h in the j th haplotype set for the i th marker phenotype. For example, for haplotype set (AB, AB, AB, ab), the number of haplotype AB is three, whereas the number is two for haplotype set (AB, AB, Ab, aB).

3. Step 3 (M-step): Update the haplotype frequencies by

$$\hat{P}_h = \hat{n}_h / T,$$

where T is the total number of chromosomes in the sample. For example, for N two-sample pools, $T=4N$.

4. Step 4: Repeat the E-step and the M-step until convergence.

ACKNOWLEDGMENTS

We thank two reviewers and Dr. Schaid for their constructive comments on the manuscript.

REFERENCES

- Amos CI, Frazier ML, Wang WF. 2000. DNA pooling in mutation detection with reference to sequence analysis. *Am J Hum Genet* 66:1689–92.
- Arnheim N, Strange C, Erlich H. 1985. Use of pooled DNA samples to detect linkage disequilibrium of polymorphic restriction fragments and human disease: studies of the HLA class II loci. *Proc Natl Acad Sci USA* 82:6970–74.
- Barcellos LF, Klitz W, Field LL, Tobias R, Bowcock AM, Wilson R, Nelson MP, Nagatomi J, Thomson G. 1997. Association mapping of disease loci by use of a pooled DNA genomic screen. *Am J Hum Genet* 61:734–47.
- Barillot E, Lacroix B, Cohen D. 1991. Theoretical analysis of library screening using a N-dimensional pooling strategy. *Nucleic Acids Res* 19:6241–47.
- Bruno WJ, Knill E, Balding DJ, Bruce DC, Doggett NA, Sawhill WW, Stallings RL, Whittaker CC, Torney DC. 1995. Efficient pooling designs for library screening. *Genomics* 26:21–30.
- Cami R, Rokhlina T, Kwitek-Black AE, Elbedour K, Nishimura D, Stone EM, Sheffield VC. 1995. Use of DNA pooling strategy to identify a human obesity syndrome locus on chromosome 15. *Hum Mol Genet* 4:9–13.
- Churchill GA, Giovannoni JJ, Tanksley SD. 1993. Pooled-sampling makes high-resolution mapping practical with DNA markers. *Proc Natl Acad Sci USA* 90:16–20.
- Clark AG. 1990. Inference of haplotypes from PCR-amplified samples of diploid populations. *Mol Biol Evol* 7:111–22.
- Coolbaugh-Murphy M, Maleki A, Strong L, Lynch P, Frazier M, Monckton D, Brown B, Siciliano MJ. 1999. Microsatellite instability (MSI) in vitro vs. in vivo? *Am J Hum Genet [Suppl]*. 65:123.

- Daniels J, Holmans P, Williams N, Turic D, McGuffin P, Plomin R, Owne MJ. 1998. A simple method for analyzing microsatellite allele image patterns generated from DNA pools and its application to allelic association studies. *Am J Hum Genet* 62:1189–97.
- Darvasi A, Soller M. 1994. Selective DNA pooling for determination of linkage between a molecular marker and a quantitative trait locus. *Genetics* 138:1365–73.
- Excoffier L, Slatkin M. 1995. Maximization-likelihood estimation of molecular haplotype frequencies in a diploid population. *Mol Biol Evol* 12:921–7.
- Fallin D, Schork NJ. 2000. Accuracy of haplotype frequency estimation for biallelic loci, via the expectation-maximization algorithm for unphased diploid genotype data. *Am J Hum Genet* 67:947–59.
- Hawley M, Kidd K. 1995. HAPLO: a program using the EM algorithm to estimate the frequencies of multi-site haplotypes. *J Hered* 86:409–11.
- Kruglyak K. 1999. Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nature Genet* 22:139–44.
- Long J, Williams R, Urbanek M. 1995. An E-M algorithm and testing strategy for multiple-locus haplotypes. *Am J Hum Genet* 56:799–810.
- Michalatos-Beloin S, Tishkoff SA, Bentley KL, Kidd KK, Ruano G. 1996. Molecular haplotyping of genetic markers 10kb apart by allele-specific long-range PCR. *Nucleic Acids Res* 24:4841–3.
- Pacek P, Sajantila A, Syvanen AC. 1993. Determination of allele frequencies at loci with length polymorphism by quantitative analysis of DNA amplifies from pooled samples. *PCR Methods Appl* 2:313–7.
- Pfeiffer RM, Rutter JL, Gail MH, Struewing J, Gastwirth JL. 2002. Efficiency of DNA pooling to estimate joint allele frequencies and measure linkage disequilibrium. *Genet Epidemiol* 22:94–102.
- Risch N. 2000. Searching for genetic determinants in the new millennium. *Nature* 405:847–56.
- Risch N, Merikangas K. 1996. The future of genetic studies of complex human diseases. *Science* 273:1516–7.
- Risch N, Teng J. 1998. The relative power of family-based and case-control designs for linkage disequilibrium studies of complex human diseases I. DNA pooling. *Genome Res* 8:1273–88.
- Shaw SH, Carrasquillo MM, Kashuk C, Puffenberger EG, Chakravarti A. 1998. Allele frequency distribution in pooled DNA samples: application to mapping complex disease genes. *Genome Res* 8:111–23.
- Stephens M, Smith NJ, Donnelly P. 2001. A new statistical method for haplotype reconstruction from population data. *Am J Hum Genet* 68:978–89.
- Tishkoff SA, Pakstis AJ, Ruano G, Kidd KK. 2000. The accuracy of statistical methods for estimation of haplotype frequencies: an example from the CD4 locus. *Am J Hum Genet* 67:518–22.
- Wang DG, Fan JB, Siao CJ, Berno A, Young P, Sapolsky R, Ghandour G, Perkins N, Winchester E, Spencer J, Kruglyak L, Stein L, Hsie L, Topaloglou T, Hubbell E, Robinson E, Mittmann M, Morris MS, Shen N, Kilburn D, Rioux J, Nusbaum C, Rozen S, Hudson TJ, Lipshutz R, Chee M, Lander ES. 1998. Large-scale identification, mapping and genotyping of single-nucleotide polymorphism in the human genome. *Science* 280:1077–82.
- Wolfram S. 1999. *The Mathematica Book*. Fourth edition. Cambridge: Cambridge University Press.
- Zhang S, Pakstis AJ, Kidd KK, Zhao H. 2001. Comparisons of two methods for haplotype reconstruction and haplotype frequency estimation from population data. *Am J Hum Genet* 69:906–12.