

Efficiency of DNA Pooling to Estimate Joint Allele Frequencies and Measure Linkage Disequilibrium

Ruth M. Pfeiffer,^{1*} Joni L. Rutter,¹ Mitchell H. Gail,¹ Jeffery Struewing,¹ and Joseph L. Gastwirth³

¹*Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, Maryland*

²*Department of Statistics, George Washington University, Washington, DC*

Pooling DNA samples can yield efficient estimates of the prevalence of genetic variants. We extend methods of analyzing pooled DNA samples to estimate the joint prevalence of variants at two or more loci. If one has a sample from the general population, one can adapt the method for joint prevalence estimation to estimate allele frequencies and D , the measure of linkage disequilibrium. The parameter D is fundamental in population genetics and in determining the power of association studies. In addition, joint allelic prevalences can be used in case-control studies to estimate the relative risks of disease from joint exposures to the genetic variants. Our methods allow for imperfect assay sensitivity and specificity. The expected savings in numbers of assays required when pooling is utilized compared to individual testing are quantified. *Genet. Epidemiol.* 22:94–102, 2002. © 2002 Wiley-Liss, Inc.

Key words: joint prevalence estimation; risk modification; population genetics; association studies

INTRODUCTION

Quantitative assays on pooled DNA can, in principle, determine the allele distribution at a particular locus, resulting in very efficient tests for association between disease and given alleles [Barcellos et al., 1997; Breen et al., 1999; Collins et al.,

Contract grant sponsor: NSF; Contract grant number: SBR-9807731.

*Correspondence to: Ruth Pfeiffer, National Cancer Institute, Division of Cancer Epidemiology and Genetics, 6120 Executive Blvd., EPS/8030, Bethesda, MD 20892-7244. E-mail: pfeiffer@mail.nih.gov

Received for publication 14 February 2001; revision accepted 14 May 2001

© 2002 Wiley-Liss, Inc.

2000; Germer et al., 2000; Shaw et al., 1998; Risch and Teng, 1998]. Some qualitative assays can determine only which genotypes are present in a pooled sample, while other qualitative assays can determine only whether a particular allele is present in the pool (carrier status). In this paper, we discuss efficient methods of analyzing pooled DNA samples with qualitative assays to estimate the joint prevalence of variants at two or more loci.

If one has a sample from the general population, one can adapt the method for joint prevalence estimation to estimate allele frequencies and D , the coefficient of allelic association, which is a key determinant of the power of association studies based on linkage disequilibrium. Indeed, empirical studies [Dunning et al., 2000] and modeling of population evolution [Kruglyak, 1999] have been used to estimate D as a function of distance between markers. Kruglyak used this information to determine how dense a genetic map of single nucleotide polymorphisms (SNPs) should be for association studies. In case-control studies, one needs to estimate the joint distribution of the variants separately in cases and controls to understand the relative risks of disease from joint exposures to the genetic variants.

Hughes-Oliver and Rosenberger [2000] presented methods based on an adaptive two-stage design for detecting the proportion of individuals with multiple traits of interest. In contrast to these two-stage procedures, we use a single-stage pooling design to keep laboratory protocols simple and allow for errors in the tests.

JOINT CARRIER PREVALENCE ESTIMATION

To study qualitative assays for carrier status, we denote the two different loci by A and B . Let the carrier status indicator $C_A = 1$ denote the event that a person carries at least one copy of the allele of interest at locus A , and let $C_A = 0$ denote the complementary event. Define the events $C_B = 1$ and $C_B = 0$ analogously. Let $\pi_{ij} = P(C_A = i, C_B = j)$ for $i, j = 0, 1$. We call the alleles of interest the “variants.” Then $\pi_{00}(\pi_{11})$ denotes the probability that an individual has no (both) variants, $\pi_{10}(\pi_{01})$ the probability that an individual has a variant at locus $A(B)$ but not at locus $B(A)$. Let N denote the total number of subjects in the sample, and k the size of each of the m pools, so $km = N$. Let $T_{iA}(T_{iB})$ be one if the i^{th} pool tests positive for a variant at locus $A(B)$ and zero otherwise. Let $S_{iA}(S_{iB})$ be one if there is at least one subject with a variant at locus $A(B)$ in the i^{th} pool and zero otherwise. The sensitivity of the test for a given pool size for locus A is then defined as $\eta_A = P(T_{iA} = 1 | S_{iA} = 1)$, and the specificity is $\phi_A = P(T_{iA} = 0 | S_{iA} = 0)$. We define η_B and ϕ_B analogously for locus B . In addition, we assume that T_{iA} depends only on S_{iA} , and T_{iB} depends only on S_{iB} , so that

$$P(T_{iA} = t_A, T_{iB} = t_B | S_{iA}, S_{iB}) = P(T_{iA} = t_A | S_{iA})P(T_{iB} = t_B | S_{iB}).$$

We base our inference on: $X_{11} = \sum_{i=1}^m T_{iA}T_{iB}$, the number of pools that test positive for variants at both loci, $X_{10} = \sum_{i=1}^m T_{iA}(1-T_{iB})$, the number of pools that test positive for the variant at locus A , but not at locus B , $X_{01} = \sum_{i=1}^m (1-T_{iA})T_{iB}$, and $X_{00} = m - (X_{11} + X_{10} + X_{01})$. The vector $(X_{11}, X_{10}, X_{01}, X_{00})$ has a multinomial distribution with index $m = \sum_{ij} X_{ij}$ and probabilities p_{ij} for $i = 0, 1$ and $j = 0, 1$. p_{11} denotes the probability that a pool tests positive for variants at loci A and B , p_{10} is the probability that a pool tests

positive for a variant at locus A but not at B , and p_{01} and p_{00} are defined similarly. The p_{ij} are found by the law of total probability, for example, $p_{11} = \sum_{S_{iA}, S_{iB}} P(T_{iA}T_{iB} = 1 | S_{iA}S_{iB})P(S_{iA}S_{iB})$. Using the conditional independence assumption and letting $q_{ij} = P(S_{iA} = i, S_{iB} = j)$,

$$\begin{aligned} p_{11} &= \eta_A \eta_B q_{11} + \eta_A (1 - \phi_B) q_{10} + \eta_B (1 - \phi_A) q_{01} + (1 - \phi_A)(1 - \phi_B) q_{00}, \\ p_{10} &= \eta_A (1 - \eta_B) q_{11} + \eta_A \phi_B q_{10} + (1 - \phi_A)(1 - \eta_B) q_{01} + (1 - \phi_A) \phi_B q_{00}, \\ p_{01} &= (1 - \eta_A) \eta_B q_{11} + (1 - \eta_A)(1 - \phi_B) q_{10} + \phi_A \eta_B q_{01} + \phi_A (1 - \phi_B) q_{00}, \\ p_{00} &= 1 - p_{11} - p_{10} - p_{01}. \end{aligned} \quad (1)$$

As the individuals in each pool are assumed to be independent, the q 's depend on the individual probabilities $(\pi_{11}, \pi_{10}, \pi_{01}, \pi_{00})$ through the relationships

$$\begin{aligned} q_{00} &= P(\text{no variants at A or B in the pool}) = \pi_{00}^k, \\ q_{10} &= P(\geq 1 \text{ variant at A, no variants at B}) \\ &= \sum_{i=1}^k \frac{k!}{i!(k-i)!} \pi_{10}^i \pi_{00}^{k-i} = (\pi_{10} + \pi_{00})^k - \pi_{00}^k, \\ q_{01} &= P(\geq 1 \text{ variant at B, no variants at A}) \\ &= (\pi_{01} + \pi_{00})^k - \pi_{00}^k, \\ q_{11} &= 1 - q_{10} - q_{01} - q_{00} = 1 + \pi_{00}^k - (\pi_{10} + \pi_{00})^k - (\pi_{01} + \pi_{00})^k. \end{aligned}$$

The log likelihood for the observations $(X_{11}, X_{10}, X_{01}, X_{00})$ is essentially

$$\begin{aligned} \log P(X_{11}, X_{10}, X_{01}, X_{00}, \boldsymbol{\pi}) &\propto X_{11} \log p_{11} + X_{10} \log p_{10} \\ &\quad + X_{01} \log p_{01} + X_{00} \log p_{00}. \end{aligned} \quad (2)$$

Assuming that the sensitivity and specificity parameters are known, we can solve the score equations obtained by differentiating equation (2) with respect to $\pi_{11}, \pi_{10}, \pi_{01}$, to obtain the maximum likelihood estimates (MLEs) $\hat{\pi}_{11}, \hat{\pi}_{10}, \hat{\pi}_{01}$. The asymptotic distribution of the estimated is multivariate normal, and the asymptotic covariance matrix is found from the inverse of the Fisher information.

In the following numerical studies, we used high levels of sensitivity and specificity, which are reasonable for pool sizes of 10 or smaller [Krook et al, 1992; Chen and Zarbl, 1997]. These parameters might not be appropriate for direct sequencing or single-stranded conformational polymorphism analysis [Amos et al., 2000].

For the first simulated example, we assumed that both variants are rare, and chose $(\pi_{10}, \pi_{01}, \pi_{00}, \pi_{11}) = (0.02, 0.01, 0.965, 0.005)$. For each fixed sample size $N = 1,000$ or 500 , and pool sizes of $k = 2, 5$, and 10 , we generated 500 independent

multinomial samples with parameter π and N , and each such sample was grouped into $m = N/k$ pools. The MATLAB [The Mathworks Inc., 1999] random number generator *unifrnd* was used to generate the multinomial counts. The entries in Table I are the average MLE estimates $\hat{\pi}$ from these simulations, and the average estimated standard deviations of the components of $\hat{\pi}$ obtained from the Fisher information.

For perfect tests, the MLE estimator $\hat{\pi}$ is nearly unbiased in every case except $N = 500$ and $k = 10$, for which $\hat{\pi}_{10}$ has a 20%, and $\hat{\pi}_{11}$ a 40% upward bias. Gastwirth and Hammick [1989] noted a similar bias when estimating the prevalence of a single trait with pool sizes above 10. The small sample bias reflects the fact that only $m =$

TABLE I. Average Estimates, Average Estimated Standard Errors and RSE, the Ratio of Standard Errors From Unpooled Testing to Standard Errors for Pooled Testing, as a Function of Pool Size, k , and Number of Pools, m

N	k	Saving (%)	$\hat{\pi}$				std error ($\hat{\pi}$)				RSE			
Perfect tests														
Example 1: $\pi_{10}, \pi_{01}, \pi_{00}, \pi_{11} = 0.02, 0.01, 0.965, 0.005$														
1,000	1	0	0.020	0.010	0.965	0.005	0.005	0.003	0.006	0.002	1.000	1.000	1.000	1.000
	2	50	0.020	0.010	0.965	0.005	0.005	0.004	0.006	0.002	0.915	0.886	0.844	1.000
	5	80	0.020	0.010	0.965	0.005	0.005	0.003	0.006	0.002	0.915	0.912	0.857	0.917
	10	90	0.020	0.010	0.965	0.005	0.005	0.004	0.006	0.003	0.843	0.861	0.844	0.846
500	1	0	0.020	0.010	0.966	0.005	0.006	0.004	0.080	0.003	1.000	1.000	1.000	1.000
	2	50	0.020	0.010	0.965	0.005	0.011	0.005	0.014	0.004	0.750	0.827	0.609	0.914
	10	90	0.024	0.010	0.959	0.007	0.027	0.006	0.037	0.015	0.233	0.782	0.209	0.212
Specificity: $\phi_A = \phi_B = .99$, Sensitivity: $\eta_A = \eta_B = .99$														
1,000	1	0	0.020	0.010	0.965	0.005	0.005	0.003	0.007	0.004	1.000	1.000	1.000	1.000
	2	50	0.020	0.010	0.965	0.005	0.005	0.004	0.007	0.004	0.849	0.564	0.836	0.958
	5	80	0.018	0.009	0.964	0.009	0.008	0.006	0.009	0.011	0.570	0.400	0.597	0.211
	10	90	0.017	0.008	0.959	0.015	0.010	0.005	0.021	0.026	0.570	0.400	0.597	0.211
500	1	0	0.020	0.010	0.965	0.005	0.008	0.005	0.010	0.005	1.000	1.000	1.000	1.000
	2	50	0.020	0.010	0.965	0.005	0.007	0.005	0.009	0.003	0.459	0.449	0.253	0.086
	5	80	0.017	0.008	0.959	0.015	0.010	0.005	0.021	0.026	0.849	0.537	0.830	0.941
	10	90	0.018	0.008	0.956	0.018	0.013	0.007	0.027	0.031	0.569	0.460	0.308	0.108
Perfect tests														
Example 2: $\pi_{10}, \pi_{01}, \pi_{00}, \pi_{11} = 0.077, 0.073, 0.835, 0.015$														
1,000	1	0	0.077	0.073	0.835	0.015	0.009	0.008	0.011	0.004	1.000	1.000	1.000	1.000
	2	50	0.076	0.073	0.835	0.015	0.009	0.009	0.012	0.005	0.944	0.922	0.933	0.760
	5	80	0.077	0.073	0.834	0.015	0.011	0.010	0.015	0.007	0.773	0.830	0.747	0.543
	10	90	0.077	0.073	0.834	0.015	0.016	0.014	0.019	0.010	0.531	0.593	0.590	0.380
500	1	0	0.077	0.073	0.836	0.015	0.011	0.012	0.016	0.006	1.000	1.000	1.000	1.000
	2	50	0.077	0.072	0.835	0.015	0.013	0.012	0.017	0.007	0.846	1.000	0.941	0.857
	5	80	0.077	0.073	0.833	0.016	0.015	0.015	0.021	0.011	0.745	0.807	0.742	0.500
	10	90	0.077	0.073	0.833	0.016	0.021	0.021	0.027	0.014	0.542	0.576	0.577	0.392
Specificity: $\phi_A = \phi_B = .99$, sensitivity: $\eta_A = \eta_B = .99$														
1,000	1	0	0.076	0.073	0.835	0.015	0.009	0.008	0.013	0.005	1.000	1.000	1.000	1.000
	2	50	0.077	0.073	0.835	0.015	0.010	0.010	0.014	0.005	0.850	0.820	0.892	0.992
	5	80	0.078	0.073	0.835	0.015	0.012	0.011	0.015	0.007	0.708	0.745	0.833	0.728
	10	90	0.081	0.077	0.823	0.018	0.018	0.022	0.027	0.015	0.472	0.372	0.463	0.340
500	1	0	0.077	0.074	0.834	0.015	0.013	0.012	0.018	0.007	1.000	1.000	1.000	1.000
	2	50	0.076	0.072	0.837	0.015	0.013	0.014	0.019	0.008	0.992	0.878	0.931	0.837
	5	80	0.080	0.075	0.828	0.017	0.017	0.017	0.022	0.011	0.758	0.723	0.804	0.609
	10	90	0.084	0.078	0.818	0.021	0.023	0.025	0.027	0.019	0.560	0.492	0.655	0.352

50 pools contribute information with $N = 500$ and $k = 10$. For each fixed N , there is no perceptible loss in precision as one moves from $k = 2$ to $k = 5$ (see average standard errors in Table I) but some decrease in precision for $k = 10$. Thus, for perfect assays, $\hat{\pi}$ yields nearly unbiased estimates with good precision for $k = 2$ and 5 and, in most cases, for $k = 10$, with corresponding reductions in numbers of required assays of 50, 80, and 90%. We also computed the ratio of the theoretical standard error of the estimator based on individual testing to that based on pooled data, RSE. For $N = 1,000$, there is little loss in precision from pooling measured by RSE, even for $k = 10$ (Table I).

If the assays have sensitivity $\eta_A = \eta_B = 0.99$ and specificity $\phi_A = \phi_B = 0.99$, $\hat{\pi}$ remains unbiased for $k = 2$, and there is little loss of precision compared to the case of perfect assays. For $k = 5$ or 10, however $\hat{\pi}_{11}$ is upwardly biased by 40% or more, with the bias increasing as N and m decrease. The standard errors of $\hat{\pi}$ in the presence of imperfect assays are also appreciably larger than for perfect assays for $k = 5$ or 10. Moreover, in the presence of measurement error, the RSE values are appreciably below 1.0 in many cases, indicating a considerable loss of precision compared to individual testing.

A second example illustrates the case of higher prevalences of each variant, $\pi_{11} = .015$, $\pi_{10} = .077$, and $\pi_{01} = .073$ (Table I). With perfect tests, there is little evidence of bias in $\hat{\pi}$ even for $N = 500$ and $k = 10$. Even in the presence of errors, the bias in $\hat{\pi}$ is small in every case except $N = 500$ and $k = 10$, for which π_{11} is overestimated by 36%. Thus, the bias is much less prominent with larger values of π_{10} , π_{01} , and π_{11} than in the case of two rare variants. Pooling results in a widening of confidence intervals, especially for $\hat{\pi}_{11}$, as one moves from $k = 2$ to $k = 10$. This loss of precision is more pronounced in the presence of imperfect sensitivity and specificity, but imperfect testing does not result in the extensive loss of precision from pooling seen for the case of two rare alleles. For $k = 2$, the loss of efficiency compared to individual testing is comparable to the rare allele case for imperfect as well as perfect testing, but for $k = 5$ and $k = 10$, the loss in efficiency can approach 30% for $k = 5$ and 60% for $k = 10$.

ESTIMATION OF LINKAGE DISEQUILIBRIUM

We show how to adapt these procedures to estimate the allele frequencies for biallelic loci A and B and the linkage disequilibrium coefficient, D , from a random sample from a population. Denote the wild type alleles by a and b , respectively, and the variant alleles by A and B . The linkage disequilibrium coefficient is $D = P(AB) - P(A)P(B)$, where $P(AB)$ is the probability that variants A and B appear on the same haplotype, and $P(A)$ and $P(B)$ are the corresponding marginal probabilities. If one has genotypes from parents as well as offspring, one can determine the haplotype of offspring and estimate D directly from haplotype frequencies. Absent family data, one can infer haplotypes from random samples of individuals with known genotypes under the assumption of random mating.

The carrier type $C_A = 1$ is composed of the two genotypes AA and Aa , as we do not observe if an individual is a homozygote or a heterozygote for A . Let p denote the allele frequency of a and r the allele frequency of b . Under random mating, the genotype frequencies can be computed from haplotype probabilities as $P(aabb) =$

$(pr + D)^2$, $P(aaBb) = 2(pr + D)[p(1 - r) - D]$, for example. A complete list can be found in Khoury et al. [1993] (table 8-5, page 257) with D defined as above. Note that $P(AaBb) = 2(pr + D)[(1 - p)(1 - r) + D] + 2[p(1 - r) - D][(1 - p)r - D]$ is composed of two parts, the probability of the double heterozygote in repulsion and the probability of the double heterozygote in coupling.

Noting that $\pi_{00} = P(aabb)$, $\pi_{10} = P(Aabb) + P(AAbb)$, $\pi_{01} = P(aaBb) + P(aaBB)$ and $\pi_{11} = P(AaBb) + P(AAbB) + P(AABB) + P(aABB)$, we can reparameterize the π 's in terms of the three parameters p , r , and D to obtain $\pi_{00} = (pr + D)^2$, $\pi_{10} = r^2 - (pr + D)^2$, and $\pi_{01} = p^2 - (pr + D)^2$.

Maximizing the log likelihood (2) for p , r , and D yields the estimates for the allele frequencies and D . Once p , r , and D have been estimated, one can use the previous formulas and others in Khoury et al. [1993, table 8-5] to estimate joint genotype frequencies.

If the laboratory technique used allows determination of exactly which alleles are present in the pools rather than just whether a particular variant is present at each locus, then the pooled data can be used to estimate genotype probabilities and allele frequencies without an assumption of random mating (Appendix). To estimate D , however, the random mating assumption is still needed.

To illustrate our method, we set $r = .75$, $p = .9$, and $D = 0$ or $D = 0.05$. This choice of allele frequencies and $D = 0.05$ corresponds to potentially informative markers for association studies [Kruglyak, 1999]. The underlying joint distribution of the carrier status is $\pi_{10} = 0.1069$, $\pi_{01} = 0.3544$, $\pi_{00} = 0.4556$, $\pi_{11} = 0.0831$ for $D = 0$ and $\pi_{10} = 0.0369$, $\pi_{01} = 0.2844$, $\pi_{00} = 0.5256$, $\pi_{11} = 0.1531$ for $D = 0.05$.

Table II shows the average estimates of r , p , and D based on simulated pooled and individual data ($k = 1$) with perfect assays. Each simulation used 500 replications. The wild type allele frequencies p and r were estimated without appreciable bias in all cases, even with only $m = 100$ pools of size $k = 5$. Pools of size 5 do yield somewhat less precise estimates of p and r than pools of sizes $k = 1$ or 2, which yield comparable precision.

For $D = 0$, estimates of D are only very slightly biased for $k = 2$ and $k = 5$, and there is modest loss in precision, compared to $\hat{\pi}$ for $k = 2$, but \hat{D}_5 is much less precise than \hat{D}_1 . For $D = 0.05$, $m = 100$, and a pool size of $k = 5$, $\hat{D}_5 = 0.045$ underestimates the true D by 10%, but for larger numbers of pools, even $k = 5$ yields nearly unbiased estimates of D .

DISCUSSION

We extend pooling methods to estimate the joint probabilities of genetic variants at two loci, while taking into account the specificity and sensitivity of the assays, which we assume known. When both variants are rare, $\hat{\pi}$ has little bias and good precision for perfect tests, even for $k = 10$ (90% reduction in assays). In the presence of assay measurement error, however, the upward bias in $\hat{\pi}$ is noticeable for $k = 5$ or 10, and the loss in precision from using $\hat{\pi}$ can be appreciable, compared to individual testing. It may be advisable to restrict pools to small sizes to avoid bias and loss in precision in the presence of measurement error. Even if $k = 2$, however, the number of required assays would be cut by 50%.

Problems of small sample bias with pooled data, even in the presence of mea-

TABLE II. Average Estimates of the allele Frequencies $r = 0.75$, $p = 0.9$, and the Disequilibrium Parameter $D = 0$ and $D = 0.05$ and Average Estimated Standard Errors for Various Pool Size k , and Numbers of Pools, m *

N	k	m	\hat{r}_k	\hat{p}_k	\hat{D}_k	std error ($\hat{r}_k, \hat{p}_k, \hat{D}_k$)		
<i>r = 0.75, p = 0.9, D = 0</i>								
10,000	1	10,000	0.7501	0.9000	0.0006	0.0034	0.0021	0.0008
	2	5,000	0.7498	0.9001	0.0010	0.0040	0.0022	0.0015
	5	2,000	0.7496	0.8999	0.0031	0.0071	0.0029	0.0045
1,000	1	1,000	0.7497	0.8995	0.0017	0.0106	0.0068	0.0025
	2	500	0.7501	0.9001	0.0033	0.0122	0.0074	0.0047
	5	200	0.7336	0.8852	0.0090	0.1020	0.1133	0.0127
500	1	500	0.7508	0.8993	0.0024	0.0146	0.0105	0.0035
	2	250	0.7497	0.9003	0.0044	0.0170	0.0104	0.0065
	5	100	0.7506	0.8999	0.0146	0.0299	0.0133	0.0180
<i>r = 0.75, p = 0.9, D = 0.05</i>								
10,000	1	10,000	0.7500	0.9000	0.0500	0.0033	0.0021	0.0015
	2	5,000	0.7498	0.9001	0.0199	0.0039	0.0025	0.0019
	5	2,000	0.7493	0.8996	0.0498	0.0099	0.0040	0.0065
1,000	1	1,000	0.7500	0.8998	0.0500	0.0105	0.0068	0.0046
	2	500	0.7501	0.9000	0.0497	0.0117	0.0073	0.0058
	5	200	0.7470	0.8993	0.0469	0.0233	0.0086	0.0159
500	1	500	0.7507	0.9000	0.0495	0.0153	0.0100	0.0067
	2	250	0.7479	0.8992	0.0497	0.0180	0.0100	0.0080
	5	100	0.7455	0.8993	0.0454	0.0444	0.0125	0.0180

*Results for $k = 1$, unpooled data, are shown for comparison.

surement error, are less severe when the variants are somewhat more common. Likewise, pooling does not result in severe loss of precision in the presence of measurement error in Example 2 (Table I).

We also show how to use pooled data on carrier status for biallelic loci to estimate the linkage disequilibrium coefficient and allele frequencies under the assumption of random mating. Pooling gives good results when both wild type allele frequencies are high, e.g., 0.75 or higher, and provides additional privacy protection [Gastwirth and Hammick, 1989]. In this setting, estimates based on pools of size $k = 5$ with $N = 500$ lead to nearly unbiased estimates of D but substantial loss of precision compared to $k = 1$. Nonetheless, these sample sizes would have good power to reject $H_0 : D = 0$, when $D = 0.05$, as indicated by the standard errors in Table II.

If the assay allows one to determine exactly which alleles are present in the pooled sample (see Appendix), then one can estimate genotypes and allele frequencies without an assumption of random mating. Indeed, such pooled data could be used to test the Hardy-Weinberg assumption at each locus. In order to estimate D , however, the assumption of random mating is needed, even with assays that allow one to determine exactly which alleles are present in a pool.

An important potential application of pooling methods is to case-control studies of candidate loci in which one wants to assess joint effects of two variants at different loci on disease risk. The pooling approach has some limitations, however, when using logistic regression to control for potential confounders [Weinberg and Umbach, 1999]. If potential confounders can be controlled for by stratification, unbiased estimates of joint relative risk from variants at two loci can be obtained separately within strata.

ACKNOWLEDGMENTS

Professor Gastwirth's research was partly supported by NSF grant SBR-9807731. We thank Lynn Goldin and the referees for helpful remarks and comments.

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.
- 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* 61:734–747.
- Breen G, Sham P, Li T, Shaw D, Collier DA, St Clair D. 1999. Accuracy and sensitivity of DNA pooling with microsatellite repeats using capillary electrophoresis. *Mol Cell Probes* 13:359–65.
- Chen Z-Y, Zarbl H. 1997. A non-radioactive, allele specific polymerase chain reaction for reproducible detection of rare mutations in large amounts of genomic DNA: application to human K-ras. *Anal Biochem* 244:191–4.
- Collins HE, Li HZ, Inda SE, Anderson J, Laiho K, Tuomilehto J, Seldin MF. 2000. A simple and accurate method for determination of microsatellite total allele content differences between DNA pools. *Hum Genet* 106:218–26.
- Dunning AM, Durocher F, Healy CS, Teare MD, McBride SE, Carlomagno F, Xu CF, Dawson E, Rhodes S, Ueda S, Lai E, Luben RN, Van Rensburg EJ, Mannermaa A, Kataja V, Rennart G, Dunham I, Purvis I, Easton D, Ponder BAJ. 2000. *Am J Hum Genet* 67:1544–54.
- Gastwirth JL, Hammick PA. 1989. Estimation of the prevalence of a rare disease, preserving the anonymity of the subjects by grouped testing: application to estimating the prevalence of AIDS antibodies in blood donors. *J Stat Plan Infer* 22:15–27.
- Germer S, Holland MJ, Higuchi R. 2000. High-throughput SNP allele-frequency determination in pooled DNA samples by kinetic PCR. *Genome Res* 10:258–66.
- Hughes-Oliver JM, Rosenberger WF. 2000. Efficient estimation of the prevalence of multiple rare traits. *Biometrika* 87:315–27.
- Khoury MJ, Beaty TH, Cohen BH. 1993. *Fundamentals of genetic epidemiology*. New York: Oxford University Press.
- Krook A, Stratton IM, O'Rahilly S. 1992. Rapid and simultaneous detection of multiple mutations by pooled and multiplex single nucleotide primer extension: application to the study of insulin-responsive glucose transporter and insulin receptor mutations in non-insulin-dependent diabetes. *Hum Mol Genet* 1:391–5.
- Kruglyak L. 1999. Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nat Genet* 22:139–44.
- The Mathworks Inc. 1999. *Statistics toolbox user's guide*.
- 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* 8:1273–1288.
- Shaw SH, Carrasquillo MM, Kashuk C, Puffenberger EG, Chakravarti A. 1998. Allele frequency distributions in pooled DNA samples: applications to mapping complex disease genes. *Genome Res* 8:111–23.
- Weinberg CR, Umbach DM. 1999. Using pooled exposure assessment to improve efficiency in case-control studies. *Biometrics* 55:718–26.

APPENDIX

Some laboratory techniques allow one to determine exactly which alleles are present in a pool rather than just whether particular variants are present. The methods derived in Joint Carrier Prevalence Estimation for carrier data can be modified for genotype measurements to estimate joint genotype frequencies for two biallelic loci *A* and *B*.

There are nine possible genotypes for an individual, with frequencies π_{aabb} , π_{aAbb} , π_{AAAb} , π_{aaBb} , π_{aaBB} , π_{aAbB} , π_{AaBb} , and π_{AABB} . After pooling samples of multiple individuals, the probabilities that various alleles are detected in a pool are p_{aabb} , p_{aAbb} , p_{AAAb} , p_{aaBb} , p_{aaBB} , p_{aAbB} , p_{AaBb} , and p_{AABB} . For example, p_{aAbb} is the probability that a pool tests positive for alleles a and A for the first, and for allele b for the second locus. Assuming that the tests have perfect sensitivity and specificity and a pool size of k , we have $p_{aabb} = \pi_{aabb}^k$, $p_{AAAb} = \pi_{AAAb}^k$, $p_{aaBb} = \pi_{aaBb}^k$, $p_{AaBb} = \pi_{AaBb}^k$, $p_{aAbb} = (\pi_{aabb} + \pi_{aaBB} + \pi_{aAbB})^k - \pi_{aaBB}^k - \pi_{aAbB}^k$, $p_{AAAb} = (\pi_{AAAb} + \pi_{AaBb} + \pi_{AAAb})^k - \pi_{AaBb}^k - \pi_{AAAb}^k$, $p_{aaBB} = (\pi_{aaBB} + \pi_{aAbB} + \pi_{aabb})^k - \pi_{aAbB}^k - \pi_{aabb}^k$, $p_{AaBb} = (\pi_{AaBb} + \pi_{aAbB} + \pi_{aabb})^k - \pi_{aAbB}^k - \pi_{aabb}^k$, $p_{aAbb} = 1 - p_{aabb} - p_{aAbB} - p_{AAAb} - p_{aaBb} - p_{aaBB} - p_{AaBb} - p_{AABB}$. These formulas can be adapted for imperfect testing. The likelihood function for the joint genotype prevalences for m pools of size k is proportional to

$$\log P(x, \pi) \propto x_{aabb} \log p_{aabb} + x_{AAAb} \log p_{AAAb} + x_{aaBb} \log p_{aaBb} + x_{AaBb} \log p_{AaBb} + x_{aAbb} \log p_{aAbb} + x_{AAAb} \log p_{AAAb} + x_{aaBb} \log p_{aaBb} + x_{AaBb} \log p_{AaBb} + x_{aAbb} \log p_{aAbb}.$$

x_{aAbb} denotes the number of pools that test positive for all four alleles, A , a , B , and b , for example. Differentiating $\log P(x, \pi)$ with respect to the π 's yields the score equations.