DOI: 10.1534/genetics.105.053702

Evaluation of DNA Pooling for the Estimation of Microsatellite Allele Frequencies: A Case Study Using Striped Bass (*Morone saxatilis*)

Garrick T. Skalski,*,[1] Charlene R. Couch,[†] Amber F. Garber,[†] Bruce S. Weir[‡] and Craig V. Sullivan[†]

**Department of Ecology and Evolutionary Biology, University of Kansas, Lawrence, Kansas 66045, [†]Department of Zoology, North Carolina State University, Raleigh, North Carolina 27695-7617 and [‡]Department of Biostatistics, University of Washington, Seattle, Washington 98195-7232*

Manuscript received November 18, 2005
Accepted for publication March 17, 2006

ABSTRACT

Using striped bass (*Morone saxatilis*) and six multiplexed microsatellite markers, we evaluated procedures for estimating allele frequencies by pooling DNA from multiple individuals, a method suggested as cost-effective relative to individual genotyping. Using moment-based estimators, we estimated allele frequencies in experimental DNA pools and found that the three primary laboratory steps, DNA quantitation and pooling, PCR amplification, and electrophoresis, accounted for 23, 48, and 29%, respectively, of the technical variance of estimates in pools containing DNA from 2–24 individuals. Exact allele-frequency estimates could be made for pools of sizes 2–8, depending on the locus, by using an integer-valued estimator. Larger pools of size 12 and 24 tended to yield biased estimates; however, replicates of these estimates detected allele frequency differences among pools with different allelic compositions. We also derive an unbiased estimator of Hardy–Weinberg disequilibrium coefficients that uses multiple DNA pools and analyze the cost-efficiency of DNA pooling. DNA pooling yields the most potential cost savings when a large number of loci are employed using a large number of individuals, a situation becoming increasingly common as microsatellite loci are developed in increasing numbers of taxa.

MICROSATELLITE loci are important in population and quantitative genetic analyses, and the number of known loci is increasing for a wide variety of taxa (Zane *et al.* 2002). Increasing the number of loci in a study typically improves the performance of statistical procedures (*e.g.*, Weir and Cockerham 1984; Waples 1989), and genomewide analyses of microsatellite loci will require many loci (*e.g.*, Lee *et al.* 2004). However, increasing the number of loci also increases laboratory costs, sometimes prohibitively. Accordingly, DNA pooling has been proposed as a means of reducing laboratory costs in population and quantitative genetic studies. Furthermore, DNA pools can also arise naturally in polyploid organisms, microbial samples, and forensic analyses.

Laboratory DNA pooling is carried out by combining DNA from two or more individuals prior to implementing the polymerase chain reaction (PCR) and electrophoresis. The resulting data typically take the form of some measure of copy number (*e.g.*, fluorescence intensity) for each fragment size (allele) in the DNA pool. Quantitative genetic studies in human and agricultural systems have sought to find an association between a trait and a marker by pooling DNA from individuals that share a common phenotype and/or ancestry, a technique sometimes called "bulked segregant analysis." Methods for comparing microsatellite DNA pools among experimental populations vary from more qualitative approaches such as counting alleles (Hillel *et al.* 2003) or comparing measures of relative fluorescence intensity (Daniels *et al.* 1998; Moritz *et al.* 2003; Lee *et al.* 2004) to more quantitative approaches such as attempting to estimate allele frequencies within a pool (Perlin *et al.* 1995; Barcellos *et al.* 1997; Band and Ron 1998; Lipkin *et al.* 1998, 2002; Rutyer-Spira *et al.* 1998; Kirov *et al.* 2000; Mosig *et al.* 2001; Schnack *et al.* 2004). More generally, DNA pooling might be applied in any genetic study relying on allele frequency estimation, including marker screening and association mapping, analyses of population differentiation, and analyses of effective population size.

The appeal of DNA pooling relative to individual genotyping arises from its potential for cost savings when evaluating many loci. The cost of genotyping n individuals is

$$C_I = nc_I + l n(c_A + c_E),$$

where l is the number of independent PCRs (*e.g.*, loci, or multiplexed sets), and c_I, c_A, and c_E are the unit costs of DNA isolation, PCR amplification, and electrophoresis,

[1]*Corresponding author:* Department of Ecology and Evolutionary Biology, 7026 Haworth Hall, 1200 Sunnyside Ave., University of Kansas, Lawrence, KS 66045. E-mail: skalski@ku.edu

respectively. The cost of estimating allele frequencies in a single pool of m individuals is

$$C_P = m(c_I + c_Q) + l(c_A + c_E),$$

where c_Q is the unit cost of DNA quantitation and pooling. DNA pooling incurs the cost of DNA quantitation and pooling, but saves on PCR amplification and electrophoresis since a pool can be used repeatedly for different multiplexes (*e.g.*, Kirov *et al.* 2000; Schnack *et al.* 2004). If the same number of individuals is used in a single pool so that $m = n$, then the percentage of savings from DNA pooling is

$$100\left(1 - \frac{C_P}{C_I}\right)\% = 100\left(1 - \frac{n(c_I + c_Q) + l(c_A + c_E)}{nc_I + ln(c_A + c_E)}\right)\%.$$

As the number of loci employed increases, the percentage of savings approaches $100(1 - 1/n)\%$. Hence, in principle, pools of 2 individuals can yield cost savings of 50% if a sufficiently large number of loci are employed. Pools of 100 individuals can potentially result in 99% cost savings given a sufficient number of loci. However, for a particular situation, the cost savings will depend on several details, such as the unit costs, the size of the pools employed, and the statistical properties of allele-frequency estimates based on these DNA pools.

While previous studies in human and agricultural systems have demonstrated applications of DNA pooling, the statistical features of allele-frequency estimates based on DNA pools are not well characterized. Understanding the statistical consequences of DNA pooling is a prerequisite to assessing its methodological potential and cost-effectiveness. Accordingly, we extend previous work by conducting a case study using six loci in two three-locus multiplexes in striped bass (*Morone saxatilis*). In particular, we (i) derive two kinds of allele frequency estimators based on DNA pools; (ii) estimate the variance components of allele-frequency estimates associated with the three primary laboratory steps, DNA quantitation and pooling, PCR amplification, and electrophoresis; (iii) evaluate allele-frequency estimates over a range of pool sizes and types and two different protocols; (iv) derive an estimator of the Hardy–Weinberg disequilibrium coefficient using DNA pools; and (v) use our estimates of technical variance as a function of pool size in a cost analysis of DNA pooling relative to individual genotyping.

MATERIALS AND METHODS

Pooling experiments: We conducted two experiments (experiments 1 and 2) to evaluate microsatellite allele frequency estimation using DNA pooling with six striped bass loci in two three-locus multiplexes [multiplex 1, *SB91* (Roy *et al.* 2000), *SB6* (García de León *et al.* 1995), and *SB108* (I. Wirgin, personal communication); multiplex 2, *AT150-2#4, AG25-1#1* (Brown *et al.* 2003), and *MSM1067* (Couch *et al.* 2006)]. To explore different allele frequency distributions, we used DNA isolated from 12 individuals (appendix a). Each locus had three to six alleles and allele size ranges varied from 8 to 44 bp (Figure 1, appendix a).

In experiment 1, we estimated the variance components of allele-frequency estimates attributable to the three laboratory steps used for DNA pooling: (i) DNA quantitation and pooling, (ii) PCR amplification, and (iii) electrophoresis. For multiplex 1, individuals 1–6 were used to construct pools of sizes 2, 3, 6, 12, and 24 having, respectively, 3, 2, 1, 3, and 3 different allelic compositions (12 total different pool types; appendix a). Individuals 7–12 were used to construct the same kinds of pools for multiplex 2 (appendix a). When necessary, different pool types used DNA from the same individual more than once. To estimate variance components, each pool type was created in triplicate, each pool was PCR amplified in duplicate, and each amplification was electrophoresed in duplicate for a total of 12 observations per pool type. In experiment 1, the total DNA concentration in the pools was diluted to 1 ng/μl (~1000 haploid genomic copies; 1 pg of DNA contains ~1 haploid copy of the striped bass genome) (Hardie and Hebert 2004), 30 cycles of PCR were performed, and DNA samples were pooled on the basis of one replicate quantitation. Motivated by our results from experiment 1, we conducted a two-factor PCR experiment to examine the roles of initial amounts of DNA and the number of PCR cycles in determining patterns of fluorescence intensity. Using multiplex 1 with DNA from individual 1, the initial DNA amount was cross-classified with the number of PCR cycles (24 and 26 cycles with 2.5, 5, and 10 ng of initial DNA and 28 and 30 cycles with 1.25, 2.5, and 5 ng of initial DNA using eight replicates of each treatment combination). In experiment 2, we modified our procedures on the basis of the results in experiment 1 and the PCR experiment and created pools of sizes 4, 6, and 8 having 3, 2, and 3 different allelic compositions, respectively (appendix a). Each pool type was replicated four or six times (for a total of 12 replicates per pool size), PCR amplified, and electrophoresed. In experiment 2, 24 cycles of PCR were performed, the DNA concentration in pools was diluted to contain 1 ng/μl (1000 haploid genomic copies) of DNA per individual, and DNA samples were pooled on the basis of three replicate quantitations.

DNA isolation, quantitation, and pool construction: DNA was isolated from finclips (sampled from a captive striped bass family and stored for 2 years in 70% ethanol) using a modified commercial protocol (PUREGENE DNA purification kit, RNase A step omitted; Gentra Systems, Research Triangle Park, NC), hydrated in TE buffer (10 mm Tris, 0.1 mm EDTA, pH 8.0), and quantified using the PicoGreen assay (Molecular Probes, Eugene, OR) in a Turner Quantech fluorometer (Barnstead Thermolyne, Dubuque, IA). We isolated an average of 1291 ± 54.13 ng (mean ± SE) of DNA per milligram of preserved fin tissue. Prior to pool construction, samples were diluted to a DNA concentration of ~10 ng/μl. DNA concentration was then estimated as the mean of three to eight replicate quantitations for each sample from our group of 12 individuals (the relative standard error was 1–1.5%; all standard curves had $r^2 > 0.99$). Different pool types were constructed using the genotypes of these 12 individuals (appendix a). The empirical distribution of quantitation residuals (~40 residuals) was used to model the laboratory error introduced by quantitation (residual equals the estimated DNA concentration from one quantitation of a sample minus the mean DNA concentration calculated over replicate quantitations of the same sample). Simulated quantitations were constructed by adding randomly drawn residuals to the

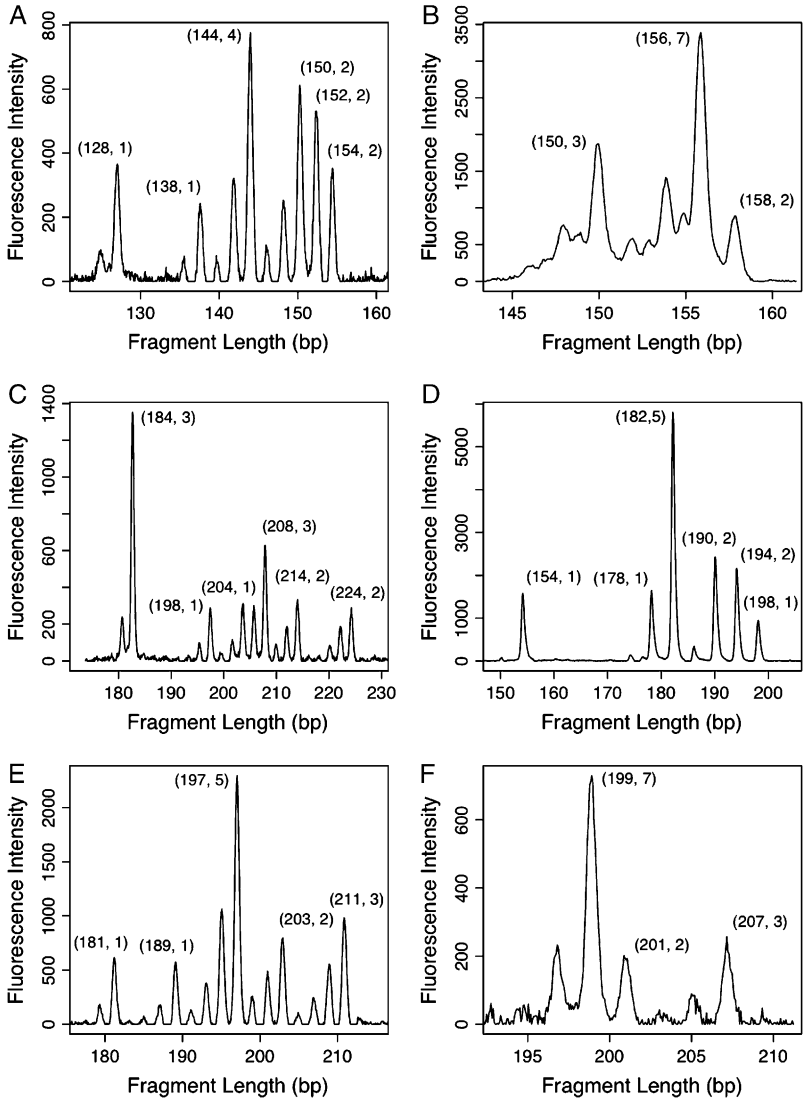

FIGURE 1.—Fluorescence intensity profiles of DNA pools composed of six individual striped bass for each of the six microsatellite loci used in our study. For each allele's fluorescence peak, labels denote the fragment length (in base pairs) and the number of copies of the allele in the pool. Other peaks are stutter peaks. (A) Dinucleotide locus *SB91*. (B) Dinucleotide locus *AT150-2#4*. (C) Dinucleotide locus *SB6*. (D) Tetranucleotide locus *AG25-1#1*. (E) Dinucleotide locus *SB108*. (F) Dinucleotide locus *MSM1067*.

mean estimated DNA concentration for each individual, and aliquot volumes based on these simulated quantitations were used to construct pools. Separate sets of residuals were generated for multiplexes 1 and 2 in experiments 1 and 2. This procedure allowed us to simulate the laboratory process of DNA pooling using a minimum of DNA isolations and quantitations.

PCR amplification: PCR amplification was conducted in multiplex using 1 μl of template in a 12-μl reaction volume with 1.5 mM $MgCl_2$, 0.2 mM dNTPs, 0.5 μM forward primer (Integrated DNA Technologies), 0.5 μM reverse primer (5′ labeled with 6FAM, PET, or NED fluorophores; Applied Biosystems, Foster City, CA), and 0.5 units of HotStarTaq DNA polymerase (QIAGEN, Valencia, CA). Thermal-cycling conditions consisted of an initial denaturation at 95° for 15 min; followed by 30 (experiment 1) or 24 (experiment 2) cycles of 94° for 30 sec, 49° (multiplex 1) or 58° (multiplex 2) for 30 sec, and 72° for 40 sec; and with a final extension at 72° for 5 min. We used the same amplification protocols and reagent quantities for individual and pooled samples.

Electrophoresis: Amplification products were purified by gel filtration (PERFORMA DTR 96-well plate; Edge Biosystems), and then 0.70 μl of product was combined with formamide (Applied Biosystems) and GeneScan 500 LIZ size standard (Applied Biosystems), denatured at 95° for 5 min, and separated by capillary electrophoresis for fluorometric quantitation on an ABI PRISM 3700 DNA analyzer (Applied Biosystems). Raw data files containing fluorescence intensities were read using a script written in MATLAB (The Mathworks). Fluorescence peaks associated with amplified fragments were detected by an algorithm that searches for localized regions of high peak intensity relative to baseline noise, a Gaussian model of peak shape was statistically fit to the peak, and peak height was calculated from the fitted model (MATLAB scripts are available from the authors). Examples of fluorescence data are shown in Figure 1 for pools of six individuals for each of the six loci. Each data file contains $\sim 10^4$ fluorescence intensity values per locus plus $\sim 10^4$ values for the size standard for a total of $\sim 4 \times 10^4$ data values per sample file. We employed peak height instead of peak area because peak height is less vulnerable to errors introduced by overlapping peaks, and because peak height is proportional to the number of molecules of a given fragment length under a diffusion model of electrophoresis transport.

Statistical analysis: Microsatellite allele frequencies are estimated from DNA pools using the fluorescent intensities associated with each allele's peak height (Figure 1). To complicate matters, the PCR amplification process introduces

stutter peaks and differential amplification errors that are reflected in the fluorescence intensities. A stutter peak occurs when an allele of repeat length i produces PCR products of length $j \neq i$, where j is typically $<i$ (Figure 1). Differential amplification is the process by which shorter alleles amplify with higher efficiency than longer alleles, and thus two alleles that are present in equal quantities (*e.g.*, as in a single heterozygote individual) can produce different fluorescence intensities. However, statistical modeling of stutter peaks and differential amplification permits estimation of allele frequencies from DNA pools. Using a simple statistical model, the allele counts in a pool of size m can be estimated using

$$\hat{\mathbf{x}} = 2m\left(\frac{\hat{\mathbf{R}}^{-1}\mathbf{y}}{|\hat{\mathbf{R}}^{-1}\mathbf{y}|}\right),$$

where $\hat{\mathbf{x}}$ is the vector of estimated allele counts, $\mathbf{y}$ is the vector of the observed fluorescent peak intensities (fluorescent peak height or area; we employed peak height), $\hat{\mathbf{R}}^{-1}$ is the inverse of an estimate of a parameter matrix $\mathbf{R}$ that models stutter and differential amplification, and $|\ |$ denotes the operation of summing vector elements (appendix b). The matrix $\hat{\mathbf{R}}$ is obtained using fluorescence data from individual genotypes. In appendix b, we show that the estimator $\hat{\mathbf{x}}$ depends only on the relative fluorescence values, $\mathbf{y}/|\mathbf{y}|$, and that the matrix $\mathbf{R}$ can be estimated using only relative fluorescence values. The need for only relative fluorescence intensities is fortunate, because, in our experience, absolute fluorescence intensities are quite variable.

The estimator $\hat{\mathbf{x}}$ contains real numbers, whereas allele counts are nonnegative integers. Accordingly, we introduce an integer-valued estimator, $\hat{\mathbf{x}}_\mathrm{I}$, that is the value of $\mathbf{u}$ that minimizes $(\mathbf{u}-\hat{\mathbf{x}})'(\mathbf{u}-\hat{\mathbf{x}})$ subject to the constraint that $\mathbf{u}$ contains nonnegative integers, and $|\mathbf{u}| = 2m$ (the $'$ denotes matrix transposition). We refer to $\hat{\mathbf{x}}$ as the "real-valued estimator" and to $\hat{\mathbf{x}}_\mathrm{I}$ as the "integer-valued estimator." We estimated allele frequencies in pools in experiments 1 and 2 using both the real- and integer-valued estimators and evaluated their statistical properties, including their technical bias (estimated minus expected value) and variance (variance of replicate estimates). During variance components estimation in experiment 1, we focused on the role of pool size. Hence, we describe the relative contributions of quantitation and pooling, PCR amplification, and electrophoresis to total variation by averaging over all alleles, loci, and pool types for each pool size. Variance components in experiment 1 were estimated by maximum likelihood (Searle *et al.* 1992), using the real-valued estimator for each allele and locus separately. We used the real-valued estimator to estimate variance components and not the integer-valued estimator because the act of rounding in the latter may obscure variance components. Variance components were averaged over all alleles within each locus–pool type combination, and then proportional variance components were averaged over all loci–pool type combinations within each pool size to control for differences in total variance among loci.

The statistical properties of allele-frequency estimates may be allele specific. Using maximum-likelihood estimates of the bias and variance of the real-valued estimator for each locus and allele as our dependent variables, we tested for allele-specific effects within each locus using a one-way ANOVA with allele as the factor. To further understand allele-specific effects, we used a general linear model to address how locus, allele frequency, and the presence/absence of stutter peaks (a stutter effect is scored as present for an allele of length i if a stutter peak exists, generated by an allele of length $j \neq i$ that overlaps the fluorescence peak associated with allele i; otherwise the stutter effect is scored as absent) affect the bias and variance of real-valued estimates. Finally, for the PCR experiment we measured the magnitude of differential amplification (as the ratio of the fluorescent peak height of the shorter allele to that of the longer allele) and stutter peaks (as the ratio of the fluorescent peak height of the stutter fragments of length $i-1$ to the peak height of the fragments of length i, where i denotes the longer allele in a heterozygote) for each locus in multiplex 1, using individual 1's heterozygous genotypes. We analyzed these measures of differential amplification and stutter peaks, using multiple linear regression with cycle number and initial DNA amount as independent variables.

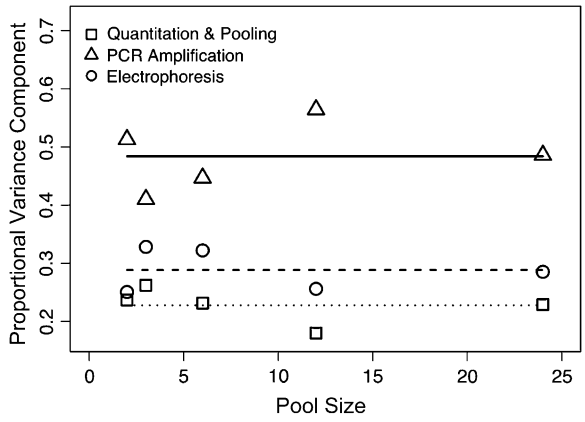

Figure 2.—Proportional variance components of allele-frequency estimates computed from DNA pools as a function of pool size. Variance components are attributable to the three primary laboratory steps: DNA quantitation and pooling, PCR amplification, and electrophoresis. Proportional variance components have been averaged over alleles and loci within each pool size. Lines denote averages over all pool sizes for each variance component.

RESULTS

Experiment 1—variance components: The estimates of laboratory variance components are similar across pool sizes (Figure 2). Averaging over pool sizes, DNA quantitation and pooling, PCR amplification, and electrophoresis accounted for 23, 48, and 29% of the observed variation in allele-frequency estimates based on the real-valued estimator. Any laboratory step may introduce a large error into an allele-frequency estimate, and these proportional variance components represent averages. Nonetheless, these results suggest that optimizing the PCR amplification step offers the most room for improvement in terms of reducing variance in allele-frequency estimates based on DNA pools. These variance component estimates also should apply to individual genotyping when the quantitation and pooling variance is set to zero, since, for example, diploid individual genotypes are pools containing two genomic copies in equal quantities. Thus, PCR amplification and electrophoresis should account for ~62% [62% = 48%/(48% + 29%)] and 38% [38% = 29%/(48% + 29%)], respectively, of the variation in fluorescence intensity in individual genotyping results.

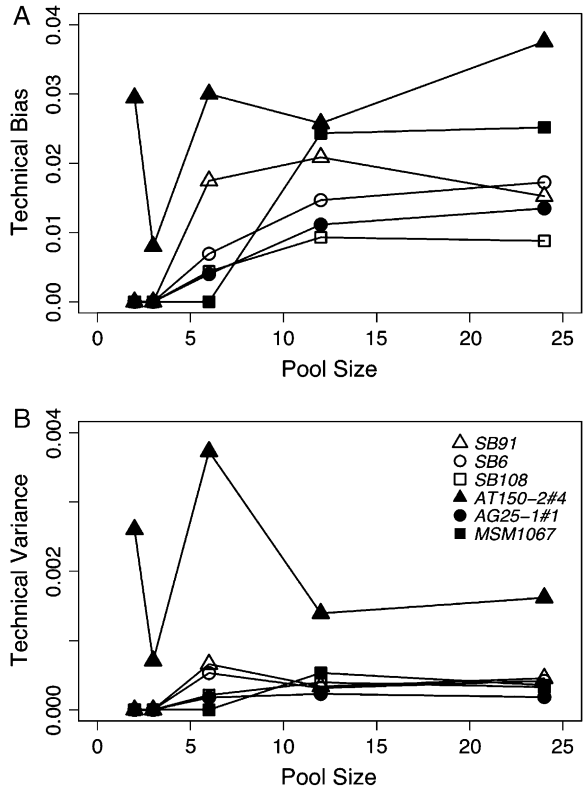

FIGURE 3.—Absolute value of technical bias and technical variance of allele-frequency estimates based on DNA pools for each of six loci as a function of pool size. Bias and variance estimates are averaged over alleles within loci. (A) Bias of the integer-valued estimator. (B) Technical variance of the integer-valued estimator.

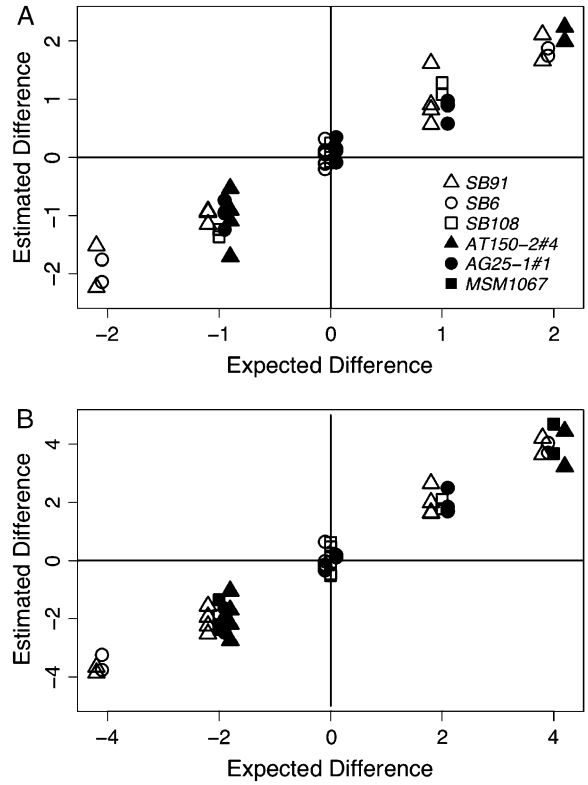

FIGURE 4.—Estimated and expected differences among two allele frequency distributions relative to a third in larger pools using the real-valued estimator. (A) Pools of size 12 for each of six loci. (B) Pools of size 24 for each of six loci. Symbols have been offset in the *x*-direction for clarity.

Experiment 1—allele-frequency estimates: When we used the integer-valued estimator to estimate allele frequencies, all of the loci exhibited similar levels of technical bias and variance with the exception of locus *AT150-2#4*, which exhibited higher bias and variance than the other loci (Figure 3). For small pools of sizes 2 and 3, all of the loci except for *AT150-2#4* yielded exact estimates for all replicates. The locus *MSM1067* yielded exact estimates for all replicates for pools of sizes 2, 3, and 6. For most loci, the bias and variance began to increase above zero at pools of size 6 and then showed little difference between larger pools of sizes 12 and 24. Bias and variance in the real-valued estimator were constant across all pool sizes (data not shown), and these values were very similar to those exhibited by the integer-valued estimator in pools of sizes 12 and 24. These results indicate that technical bias and variance are essentially constant across pool sizes except for small pool sizes when real-valued estimates can be rounded to their expected integer values using the integer-valued estimator.

The previous results assess the properties of the estimators in terms of absolute accuracy and precision. In many actual applications (*e.g.*, DANIELS *et al.* 1998; MORITZ *et al.* 2003; LEE *et al.* 2004), however, the actual value of population allele frequencies is not relevant; instead the main concern is the relative frequencies of alleles between two or more populations. In this situation bias may not be a concern, and we can ask how well larger DNA pools detect differences in allelic composition between populations. For pools of sizes 12 and 24 for which we created three pool types, we calculated the mean estimated difference in allele counts (using the real-valued estimator) between two of the pool types relative to the third and plotted it against the expected difference in pool types (Figure 4). There was a strong correspondence between estimated and expected differences, and, most importantly, estimates were positive or negative when their true values were positive or negative, respectively. These results suggest that larger pools, even in the presence of bias, can be used to detect relative differences in allele-frequency composition among populations.

The above results summarize averages taken over alleles. A one-way ANOVA indicated significant allele-specific effects on the bias and variance of estimates for most loci (bias, $P < 0.05$ for all loci except *AG25-1#1*; variance, $P < 0.05$ for all loci except *SB6* and *MSM1067*). On average, allele-frequency and stutter effects explained little of the among-allele variation in bias

($P > 0.05$), with the effect of allele frequency being somewhat locus dependent (allele frequency by locus interaction, $F_{5,160} = 7.24$, $P < 0.001$). Bias increased with increasing allele frequency for loci *AT150-2#4* and *SB6* and was independent of allele frequency for the other loci. In contrast, the variance for all loci was affected by allele frequency ($F_{1,160} = 16.52$, $P < 0.001$), stutter effects ($F_{1,160} = 13.15$, $P < 0.001$), and their interaction ($F_{1,160} = 10.05$, $P = 0.002$). The presence of stutter effects increased the variance. The variance increased with increasing allele frequency, and allele frequency had a slightly stronger effect on the variance in the absence of stutter effects. Finally, bias and variance were strongly locus dependent (bias, $F_{5,160} = 5.35$, $P < 0.001$; variance, $F_{5,160} = 7.72$, $P < 0.001$), as indicated in Figure 2.

PCR experiment and experiment 2—allele-frequency estimates in small pools: Motivated by the magnitude of the variance component attributable to the PCR amplification and the role of stutter effects in affecting allele-frequency estimates, we examined by experiment the effects of initial amount of DNA and cycle number on differential amplification and stutter peaks using multiplex 1. The magnitude of both differential amplification and stutter peaks increased with increasing cycle number in a graded manner across all loci (linear coefficients for all loci: $P < 0.001$). Similarly, the magnitude of differential amplification increased with increasing initial amount of DNA for all loci (linear coefficients for all loci: $P < 0.001$). In contrast, the effect of DNA amount on stutter was more ambiguous, causing increased stutter in SB91 ($F_{1,87} = 10.55$, $P = 0.002$), marginally significant increased stutter in SB6 ($F_{1,87} = 3.56$, $P = 0.063$), and decreased stutter in SB108 ($F_{1,87} = 18.23$, $P < 0.001$). On average, the effect of cycle number on differential amplification and stutter peaks was stronger than the effect of increasing the initial amount of DNA; hence, PCR products from a 24-cycle reaction starting with 10 ng of DNA yielded less differential amplification and stutter than a 30-cycle reaction starting with 1.25 ng of DNA.

On the basis of results from experiment 1 and the PCR experiment, we further explored the estimation of allele frequencies using the integer-valued estimator with small pools with two aims in mind. First, we wanted to assess the extent to which technical variance could be reduced relative to experiment 1 by using 24 PCR cycles instead of 30 and constructing pools on the basis of three replicate quantitations instead of one. We hypothesized that reducing the number of PCR cycles could reduce the technical variance via the kind of reduction in differential amplification and stutter peaks that we observed in the PCR experiment. Given a reduction in technical variance, our second aim was to estimate with increased replication (12 independent replicates per pool size) the extent to which allele frequencies can be estimated without error using the integer-valued estimator. We focused on pools of sizes 4, 6, and 8, since the integer-valued estimator began to show nonzero technical bias and variance for most loci at pools of size 6 in experiment 1.

In experiment 2, all loci except *AT150-2#4* produced exact estimates for all replicates in pools of size 4 (*AT150-2#4* yielded exact estimates in all but one replicate), four loci (*SB91*, *AG25-1#1*, *SB108*, and *MSM1067*) produced exact estimates for all replicates in pools of size 6, and three loci (*AG25-1#1*, *SB108*, and *MSM1067*) produced exact estimates for all replicates in pools of size 8. Hence, depending on the locus, allele frequencies can be estimated exactly for pools of sizes 4, 6, and 8 using three replicate quantitations per contribution to a pool, 24 cycles of PCR, and the integer-valued estimator. Focusing on pools of size 6 since they were constructed in both experiments 1 and 2, the average variance of allele frequency estimates over all loci decreased from 8.8×10^{-4} in experiment 1 to 2.0×10^{-4} in experiment 2, a significant decrease of ~77% ($F_{5,11} = 4.40$, $P = 0.019$).

Estimation of Hardy–Weinberg disequilibrium coefficient using DNA pools: The pooling of genes over individuals in DNA pools implies that information concerning within-individual covariances among alleles, such as Hardy–Weinberg disequilibrium coefficients, is lost during pooling. However, some information on among-allele covariances is retained during pooling. The Hardy–Weinberg disequilibrium coefficient can be estimated if allele frequencies from more than one pool are available. For the multiallelic case, the Hardy–Weinberg disequilibrium coefficient for alleles a and b is

$$D_{ab} = p_a p_b - \frac{p_{ab}}{2},$$

where p_{ab}, p_a, and p_b are genotype and respective gene frequencies for alleles a and b (Weir 1996). For simplicity, we consider the situation where allele frequencies are estimated exactly using small pools. Letting $\hat{p}_{ai}$ and $\hat{p}_{bi}$ be the observed frequencies of alleles a and b in pool i, D_{ab} can be estimated from r replicate pools of size m using

$$\hat{D}_{ab} = \frac{2m-1}{r(r-1)}\left(\sum_{i=1}^{r}\sum_{i\neq i'}\hat{p}_{ai}\hat{p}_{bi'}\right) - \frac{2m}{r}\left(\sum_{i=1}^{r}\hat{p}_{ai}\hat{p}_{bi}\right),$$

an estimator that is unbiased (appendix c). Simulations suggest that the variance of this estimator increases with increasing pool size and decreases with increasing numbers of replicate pools. These results show that some information regarding Hardy–Weinberg equilibrium can be recovered from pooled DNA, and this estimator could be used with individual genotype data (which would also be used to estimate **R**) to provide information on Hardy–Weinberg disequilibrium. Further, given estimates of the disequilibrium coefficient over all pairs of alleles, population genotype frequencies

can be estimated. We have considered the idealized case of zero technical bias and variance; thus, the effect that errors in allele frequency estimation from DNA pools may have on these estimates should be examined in specific cases prior to implementing these procedures.

Cost analysis: Whether DNA pooling should be implemented, beyond its feasibility, lies in its cost-effectiveness. To evaluate cost-effectiveness, estimates of technical variance as a function of pool size are needed because the experimental design decision is to choose the best pool size from the set of possible pool sizes (including pools of size 1 that correspond to individual genotyping). Knowledge of the relative costs of laboratory procedures is also required to evaluate cost-effectiveness. Experiments 1 and 2 provide estimates of the technical variance for different pools sizes and allow us to calculate cost-effectiveness. In the following calculations, we assume that the technical bias is zero, because small pools are used, or that bias is not a concern because only relative differences among population allele frequencies are of interest. We calculate, for a given pool size, the percentage of cost savings realized by DNA pooling that achieves the same statistical precision as genotyping n individuals. Using the cost equations presented earlier and assuming individual genotyping of n individuals for an allele having frequency p, r replicate pools of size m yield a percentage of cost savings of

$$100\left(1-\frac{C_P}{C_I}\right)\% = 100\left\{1-\left[\frac{2\sigma_T^2}{p(1-p)}+\frac{1}{m}\right]\left[\frac{m(c_I+c_Q)+l(c_A+c_E)}{c_I+l(c_A+c_E)}\right]\right\}\%, \quad (1)$$

where σ_T^2 is the technical variance; l is the number of separate PCRs (*e.g.*, multiplexes); and c_I, c_Q, c_A, and c_E are the relative unit costs (*e.g.*, per individual) of DNA isolation, quantitation and pooling, PCR amplification, and electrophoresis, respectively (appendix d).

Three qualitative points can be illustrated, using a graphical example based on our laboratory costs (estimated labor plus reagent costs) and Equation 1, and assuming, on the basis of our data, that the technical variance is constant as a function of pool size. First, the percentage of savings increases with both pool size and the number of PCRs (Figure 5; the parameters σ_T^2, p, c_I, c_Q, c_A, and c_E are held fixed in Equation 1 and Figure 5, while pool size and number of PCRs are varied). Most of the savings is acquired as pool size initially increases, and there are diminishing returns with subsequent increases in pool size. The same pattern holds for increasing the number of PCRs, with most of the savings being realized after 10 PCRs in our example. These diminishing returns suggest that small pools can be almost as cost-effective as large pools.

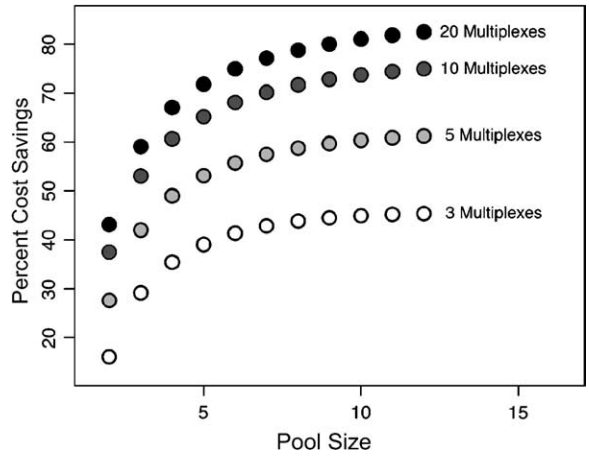

Figure 5.—Percentage of cost savings of DNA pooling relative to individual genotyping for different numbers of independent PCRs (*e.g.*, multiplexes) as a function of pool size using Equation 1 with $\sigma_T^2 = 0.0005$, $p = 0.1$, $c_I = 3$, $c_Q = 3.6$, $c_A = 1.05$, and $c_E = 3.25$.

Second, the percentage of cost savings reaches a maximum value (appendix d) when the pool size is given by

$$m^* = \sqrt{\left(\frac{p(1-p)}{2\sigma_T^2}\right)\left(\frac{l(c_A+c_E)}{(c_I+c_Q)}\right)}.$$

However, while an optimum pool size exists, it is not much more cost-effective than many other pool sizes because the cost savings curves tend to be flat near the optimum pool size (optimum pool size is not shown in Figure 5 because the curves are very flat for larger pool sizes). Finally, Equation 1 shows that the percentage of cost savings is independent of n, where n (the number of individuals genotyped) determines the desired level of statistical precision. Hence, absolute cost savings will be proportional to Equation 1. The qualitative features of the cost savings curves depicted in our example are relatively insensitive to realistic variations in the statistical and cost parameters. Changing the allele frequency p has a small effect, as does varying the technical variance between 0 and 0.001, a realistic range based on our experiments. Of course, the relative costs associated with DNA pooling will be laboratory specific, depending on the costs and procedures of individual genotyping, available equipment, reagents used, institution-specific fees, the capacity for multiplexing, and labor costs. Hence, laboratories can employ Equation 1 and the associated ideas using costs that match their situation. We have considered only recurring costs associated with DNA pooling. For some laboratories, there can be substantial preliminary costs, including optimization of laboratory procedures, the acquisition of appropriate software for analyzing pooled data, and the purchasing of equipment for DNA quantitation. Laboratory-specific

cost analyses should consider whether these costs can be recovered through subsequent savings in recurring costs. DNA pooling will be most cost-effective when recurring costs are large, a situation that arises when many individuals are used with many loci.

DISCUSSION

DNA pooling has been proposed as a cost-effective means to estimate microsatellite allele frequencies in studies using many loci. As microsatellite markers are developed in growing numbers in many species, the cost of analyzing many loci using many individuals will become increasingly relevant. Using six multiplexed microsatellite loci in striped bass, we evaluated the statistical properties of procedures that estimate allele frequencies using DNA pools and used these results to assess the feasibility and cost-effectiveness of DNA pooling.

We provide a statistical and theoretical basis for estimating microsatellite allele frequencies using DNA pools by deriving a moment-based estimator. The procedures implemented in previous studies (Perlin *et al.* 1995; Barcellos *et al.* 1997; Band and Ron 1998; Lipkin *et al.* 1998, 2002; Kirov *et al.* 2000; Mosig *et al.* 2001; Schnack *et al.* 2004) have used some version of this estimator, but a probabilistic basis for the estimator has not been presented. Our results show explicitly that the estimator depends only on the relative fluorescence intensities associated with each allele, explaining why allele frequency estimation from DNA pools is tractable empirically, even if absolute fluorescence intensities are quite variable.

Our results suggest that the technical bias, technical variance, and proportional variance components are approximately constant as a function of pool size, except when the integer-valued estimator was employed for small pools, in which case the estimates of bias and variance were zero for all but one locus. Allele frequency estimates from larger pools can be biased, but such estimates are still able to detect relative differences in allele frequencies among DNA pools. Previous microsatellite DNA pooling studies have focused on pool sizes of $\sim \geq 20$ and information is lacking concerning estimators over a number of pool sizes and types, laboratory variance components, and integer-valued estimators. The comparison of quantitative results among different studies is complicated by differences in the properties of microsatellite loci and methodologies used in various studies. Nonetheless, our estimates of technical bias and variance as a function of pool size are consistent in terms of magnitude and pattern with the findings of other studies for which this kind of comparison can be made in an approximate way (Band and Ron 1998; Lipkin *et al.* 1998, 2002). Our results also demonstrate that allele frequency estimation using DNA pools is feasible using a simple DNA isolation protocol with ethanol-preserved finclips, a convenient tissue sampling procedure in many fishes. Further research on the effect of pooling methodology and allele size range and diversity on estimation procedures as well as the contrasting properties of di-, tri-, and tetranucleotide markers in DNA pools is needed.

We suspect that the unbiased estimation of allele frequencies from large DNA pools could be difficult to achieve for many loci, and that the difficulty increases with increasing pool size. Because $\mathbf{R}$ is estimated rather than known exactly, even if relative fluorescent intensities are measured without error (which is not the case), there will be some bias between the estimated allele counts, $\hat{\mathbf{x}}$, and the expected allele counts, $\mathbf{x}$, given by

$$\hat{\mathbf{x}} - \mathbf{x} = \left(\frac{\hat{\mathbf{R}}^{-1}\mathbf{y}}{|\hat{\mathbf{R}}^{-1}\mathbf{y}|} - \frac{\mathbf{R}^{-1}\mathbf{y}}{|\mathbf{R}^{-1}\mathbf{y}|} \right) 2m,$$

a quantity that is proportional to pool size m. Hence, as m increases, the bias in allele counts increases. Once the bias becomes $>\frac{1}{2}$, then the integer-valued estimator will not round to the true value. The above equation also predicts that bias will be constant across pool sizes, which is the pattern that we observed empirically. Thus, as m increases, $\mathbf{R}$ must be estimated with increasing precision to yield unbiased estimates. In our present study, we found that unbiasedness vanishes between pool sizes of $\sim$6 and 12, depending on the locus. Additional research on the best methods for estimating $\mathbf{R}$ is needed (Schnack *et al.* 2004).

The results of experiment 1 suggest that the DNA quantitation and pooling, PCR amplification, and electrophoresis steps account for $\sim$23, 48, and 29%, respectively, of the total technical variance of allele frequency estimates based on DNA pools. These results imply that optimizing the PCR amplification step offers the most potential, in terms of variance reduction, to minimize the technical variance. Our PCR experiment showed that differential amplification and stutter peaks decrease with decreasing cycle number, findings consistent with related theoretical and empirical work on the PCR process (Lai and Sun 2003; Shinde *et al.* 2003). In experiment 2, we found that the technical variance could be reduced relative to experiment 1 by reducing cycle number and increasing the number of replicate quantitations. The results from our experiments suggest that microsatellite allele frequencies can be estimated exactly, or very nearly exactly, for pools of sizes 2–8, depending on the locus. Employing replicate quantitations and a reduction in PCR cycle number in concert with an integer-valued estimator may be a means for improving DNA pooling protocols.

We also investigated the estimation of Hardy–Weinberg disequilibrium coefficients from DNA pools by deriving

an unbiased estimator of Hardy–Weinberg disequilibrium coefficients for multiple alleles, using allele frequencies estimated from multiple DNA pools. The estimator also can be used to recover genotype frequencies using only information on allele frequencies. In these analyses, we assumed that allele frequencies were estimated without error using DNA pools; however, the effects of violations of this assumption on the performance of the estimator should be explored. The estimation of Hardy–Weinberg disequilibrium coefficients is similar in concept to the estimation of linkage disequilibrium coefficients and haplotype estimation, a topic that has been explored in association studies (Wang *et al.* 2003; Yang *et al.* 2003). The effects of allele frequency estimation using DNA pools on the estimation of other quantities, such as effective population size and measures of population differentiation, also warrant investigation.

The utility of DNA pooling depends upon both its feasibility and its cost-effectiveness. Our study shows that DNA pooling is feasible for both small (sizes 2–8) and larger pools (sizes 12 and 24). Small pools offer the possibility of unbiased and even exact estimates. Larger pools can yield biased estimates, but may remain useful for assessing relative differences in allele frequencies among populations, the main approach of previous pooling studies (Barcellos *et al.* 1997; Band and Ron 1998; Lipkin *et al.* 1998, 2002; Kirov *et al.* 2000; Mosig *et al.* 2001; Schnack *et al.* 2004). We used our estimates of technical variance to assess the cost-effectiveness of DNA pooling relative to individual genotyping. Analyses with cost estimates from our laboratory suggest that most of the increase in relative cost savings with increasing pool size can be realized with smaller pools. Multiple small pools are cost-effective relative to fewer large pools because averaging many replicate pools acts to decrease the technical variance. The cost-effectiveness of small pools coupled with the possibility of estimating both allele frequencies and Hardy–Weinberg disequilibrium coefficients using an integer-valued estimator adds to the appeal of small pools. The performance of these methods across a large number of loci remains to be investigated, but we expect that some loci, such as the tetranucleotide locus *AG25-1#1*, will work well in DNA pools, while others, such the dinucleotide locus *SB6*, with its large range in allele sizes and accompanying differential amplification, will prove more problematic. Costs will be situation specific, and the cost equations that we provided will allow laboratories to conduct individualized cost analyses. High-throughput software that implements the estimation procedures will be required. Because diploid individuals are the special case of pools containing two genomic copies, software for analyzing DNA pools should be similar to existing packages employed in individual genotyping (*e.g.*, Pálsson *et al.* 1999). DNA pooling should be most cost-effective when many individuals are typed at many loci, a situation that will become increasingly common as microsatellite loci are developed in increasing numbers in many different species.

Microsatellite allele frequency estimation using DNA pools can potentially be applied to a variety of quantitative and population genetic situations, including any situation where the allele frequency in a population is the fundamental quantity of interest. Allelic association studies, analyses of spatial population structure and effective population size, and the estimation of genetic diversity can potentially be addressed with DNA pooling. Aspects of marker development, such as the screening of many loci for polymorphisms and for sex linkage (Lee *et al.* 2004), should be amenable to DNA pooling. The possibility of pooling certain tissues directly, such as pooling eggs, blood samples, or groups of larvae, also merits consideration. Finally, more natural DNA pools such as polyploid, single-celled organisms and forensic DNA mixtures arise in the course of many investigations, and the procedures of DNA pooling might be applied in these situations.

We thank Chris Smith for assistance with programming, the North Carolina State University Genome Research Laboratory for assistance with fragment analysis, and two anonymous reviewers for comments on an earlier draft of this manuscript. This research was supported by grants from the National Institutes of Health (NIH-ES 07329) and the National Science Foundation (NSF-DEB 03-43761).

LITERATURE CITED

Band, M., and M. Ron, 1998 Determination of allele frequency from DNA pools using bovine trinucleotide microsatellites. Anim. Biotechnol. **9:** 35–45.

Barcellos, L. F., W. Klitz, L. L. Field, R. Tobias, A. M. Bowcock *et al.*, 1997 Association mapping of disease loci, by use of a pooled DNA genomic screen. Am. J. Hum. Genet. **61:** 734–747.

Brown, K. M., G. A. Baltazar, B. N. Weinstein and M. B. Hamilton, 2003 Isolation and characterization of nuclear microsatellite loci in the anadromous marine fish *Morone saxatilis*. Mol. Ecol. Notes **3:** 414–416.

Couch, C. R., A. F. Garber, C. E. Rexroad, III, J. M. Abrams, J. A. Stannard *et al.*, 2006 Isolation and characterization of 149 novel microsatellite DNA markers for striped bass, *Morone saxatilis*, and cross-species amplification in white bass, *M. chrysops*, and their hybrid. Mol. Ecol. Notes (in press).

Daniels, J., P. Holmans, N. Williams, D. Turic, P. McGuffin *et al.*, 1998 A simple method for analyzing allele image patterns generated from DNA pools and its application to allelic association studies. Am. J. Hum. Genet. **62:** 1189–1197.

García de León, F. J., J. F. Dallas, B. Chatain, M. Canonne, J. J. Versini *et al.*, 1995 Development and use of microsatellite markers in sea bass, *Dicentrarchus labrax* (Linnaeus, 1758) (Perciformes: Serranidae). Mol. Mar. Biol. Biotechnol. **4:** 62–68.

Hardie, D. C., and P. D. N. Hebert, 2004 Genome-size evolution in fishes. Can. J. Fish. Aquat. Sci. **61:** 1636–1646.

Hillel, J., M. A. M. Groenen, M. Tixier-Boichard, A. B. Korol, L. David *et al.*, 2003 Biodiversity of 52 chicken populations assessed by microsatellite typing of DNA pools. Genet. Sel. Evol. **35:** 533–557.

Kirov, G., N. Williams, P. Sham, N. Craddock and M. J. Owen, 2000 Pooled genotyping of microsatellite markers in parent-offspring trios. Genome Res. **10:** 105–115.

Lai, Y., and F. Sun, 2003 Microsatellite mutations during the polymerase chain reaction: mean field approximations and their applications. J. Theor. Biol. **224:** 127–137.

Lee, B.-Y., G. Hulata and T. D. Kocher, 2004 Two unlinked loci controlling the sex of blue tilapia (*Oreochromis aureus*). Heredity **92:** 543–549.

Lipkin, E., M. O. Mosig, A. Darvasi, E. Ezra, A. Shalom *et al.*, 1998 Quantitative trait locus mapping in dairy cattle by means of selective milk DNA pooling using dinucleotide microsatellite markers: analysis of milk protein percentage. Genetics **149:** 1557–1567.

Lipkin, E., J. Fulton, H. Cheng, N. Yonash and M. Soller, 2002 Quantitative trait locus mapping in chickens by selective DNA pooling with dinucleotide microsatellite markers by using purified DNA and fresh or frozen blood cells as applied to marker-assisted selection. Poultry Sci. **81:** 283–292.

Moritz, R. F. A., H. Scharpenberg, H. M. G. Lattorff and P. Neumann, 2003 A technical note for using microsatellite DNA analyses in haploid male DNA pools of social hymenoptera. Insectes Soc. **50:** 398–400.

Mosig, M. O., E. Lipkin, G. Khutoreskaya, E. Tchourzyna, M. Soller *et al.*, 2001 A whole genome scan for quantitative trait loci affecting milk protein percentage in Israeli-Holstein cattle, by means of selective milk DNA pooling in a daughter design, using an adjusted false discovery rate. Genetics **157:** 1683–1698.

Pálsson, B., F. Pálsson, M. Perlin, H. Gudbjartsson, K. Stefánsson *et al.*, 1999 Using quality measures to facilitate allele calling in high-throughput genotyping. Genome Res. **9:** 1002–1012.

Perlin, M. W., G. Lancia and S. Ng, 1995 Toward fully automated genotyping: genotyping microsatellite markers by deconvolution. Am. J. Hum. Genet. **57:** 1199–1210.

Roy, N. K., L. Maceda and I. Wirgin, 2000 Isolation of microsatellites in striped bass *Morone saxatilis* (Teleostei) and their preliminary use in population identification. Mol. Ecol. **9:** 817–829.

Rutyer-Spira, C. P., A. J. C. de Groof, J. J. van der Poel, J. Herbergs, J. Masabanda *et al.*, 1998 The HMGI-C gene is a likely candidate for the autosomal dwarf locus in the chicken. J. Hered. **89:** 295–300.

Schnack, H. G., S. C. Bakker, R. van't Slot, B. M. Groot, R. J. Sinke *et al.*, 2004 Accurate determination of microsatellite allele frequencies in pooled DNA samples. Eur. J. Hum. Genet. **12:** 925–934.

Searle, S. R., G. Casella and C. E. McCulloch, 1992 *Variance Components.* John Wiley & Sons, New York.

Shinde, D., Y. Lai, F. Sun and N. Arnheim, 2003 *Taq* DNA polymerase slippage mutation rates measured by PCR and quasi-likelihood analysis: $(CA/GT)_n$ and $(A/T)_n$ microsatellites. Nucleic Acids Res. **31:** 974–980.

Wang, S., K. K. Kidd and H. Zhao, 2003 On the use of DNA pooling to estimate haplotype frequencies. Genet. Epidemiol. **24:** 74–82.

Waples, R. S., 1989 A generalized approach for estimating effective population size from temporal changes in allele frequency. Genetics **121:** 379–391.

Weir, B. S., 1996 *Genetic Data Analysis II.* Sinauer Associates, Sunderland, MA.

Weir, B. S., and C. C. Cockerham, 1984 Estimating F-statistics for the analysis of population structure. Evolution **38:** 1358–1370.

Yang, Y., J. Zhang, J. Hoh, F. Matsuda, P. Xu *et al.*, 2003 Efficiency of single-nucleotide polymorphism haplotype estimation from pooled DNA. Proc. Natl. Acad. Sci. USA **100:** 7225–7230.

Zane, L., L. Bargelloni and T. Patarnello, 2002 Strategies for microsatellite isolation: a review. Mol. Ecol. **11:** 1–6.

Communicating editor: M. Nordborg

APPENDIX A: INDIVIDUAL GENOTYPES AND POOL COMPOSITIONS

TABLE A1

Individual genotypes

	Individual genotypes[a]											
Locus, motif, repeat length	1	2	3	4	5	6	7	8	9	10	11	12
SB91, GT/CT, 56	128	138	144	144	150	152	138	138	152	152	144	152
	144	144	150	152	154	154	144	144	154	154	152	154
SB6, GT, 24	184	184	204	184	208	208	184	208	184	208	184	184
	214	224	208	198	224	214	204	224	224	224	198	214
SB108, TG, 38	197	197	181	189	203	197	203	181	197	197	189	197
	211	203	197	197	211	211	211	211	197	211	197	197
AT150-2#4, GT, 36	150	156	150	156	156	156	150	156	150	150	156	156
	156	156	156	158	156	156	156	156	156	156	158	158
AG25-1#1, CTTT, 44	182	178	178	182	154	194	154	178	182	182	182	182
	198	194	182	194	182	198	182	194	190	190	194	198
MSM1067, CA/GT, 50	199	199	199	199	201	199	199	199	199	199	199	199
	207	199	201	207	201	207	199	201	201	207	207	207

Genotypes of the 12 individuals used in the experiments at each of the six microsatellite loci are shown. The repeat motif and length (in base pairs) of the sequence from which each locus was derived are given.

[a] Data in top row of each section are for shorter alleles and data in bottom row of each section are for longer alleles.

TABLE A2

Design of experimental pools

	Individual											
	Multiplex 1						Multiplex 2					
Pool size	1	2	3	4	5	6	7	8	9	10	11	12
						Experiment 1						
2	1	0	0	1	0	0	1	0	0	1	0	0
2	0	1	0	0	1	0	0	1	0	0	1	0
2	0	0	1	0	0	1	0	0	1	0	0	1
3	1	0	1	0	1	0	1	0	1	0	1	0
3	0	1	0	1	0	1	0	1	0	1	0	1
6	1	1	1	1	1	1	1	1	1	1	1	1
12	2	2	2	2	2	2	2	2	2	2	2	2
12	3	3	2	2	1	1	3	3	2	2	1	1
12	1	1	2	2	3	3	1	1	2	2	3	3
24	4	4	4	4	4	4	4	4	4	4	4	4
24	6	6	4	4	2	2	6	6	4	4	2	2
24	2	2	4	4	6	6	2	2	4	4	6	6
						Experiment 2						
4	1	1	1	1	0	0	0	0	0	0	0	0
4	0	0	0	0	1	1	1	1	0	0	0	0
4	0	0	0	0	0	0	0	0	1	1	1	1
6	1	1	1	1	1	1	0	0	0	0	0	0
6	0	0	0	0	0	0	1	1	1	1	1	1
8	1	1	1	1	1	1	1	1	0	0	0	0
8	1	1	1	1	0	0	0	0	1	1	1	1
8	0	0	0	0	1	1	1	1	1	1	1	1

The number of times an individual's genotypes contributed to a pool is given for each pool size in experiments 1 and 2 for multiplexes 1 and 2, thereby defining the allelic composition of each pool. Note that in experiment 1 multiplex 1 used only individuals 1–6 and multiplex 2 used only individuals 7 to 12.

APPENDIX B: ALLELE-FREQUENCY ESTIMATION WITH DNA POOLS

We consider a simple linear model of the process by which template DNA is amplified into large numbers of locus-specific fragments and subsequently fluorescently detected following electrophoresis. For a given locus, let $\mathbf{x}$ denote the vector of allele counts, where the alleles are ordered in $\mathbf{x}$ by increasing fragment length. Hence, the length of $\mathbf{x}$ is the number of alleles, the ith element of $\mathbf{x}$ corresponds to the ith longest allele, each element of $\mathbf{x}$ is the number of copies of a particular allele, and the sum of the elements of $\mathbf{x}$ is the total number of gene copies. For pools of diploid individuals, the sum of the elements of $\mathbf{x}$, $|\mathbf{x}|$, is $2m$, where m is the number of individuals in the pool. Let $\mathbf{y}$ denote the observed vector of fluorescence intensities (calculated as fluorescent peak heights or peak areas) associated with each of the alleles in $\mathbf{x}$. A simple stochastic model of the dependence of $\mathbf{y}$ on $\mathbf{x}$ is the linear model

$$\mathbf{y} = \Lambda\mathbf{x} + \varepsilon,$$

where $\mathbf{\Lambda}$ is a nonnegative square matrix of parameters describing the amplification and fluorescent detection of the alleles in $\mathbf{x}$ that results in the observed fluorescence intensities in $\mathbf{y}$, and $\boldsymbol{\epsilon}$ is a vector of stochastic errors introduced in the laboratory having expectation $E(\boldsymbol{\epsilon}) = 0$. In general, the elements of $\mathbf{\Lambda}$ will depend upon the details of the laboratory protocol, including the properties of the microsatellite locus as well as the PCR and electrophoresis conditions. The off-diagonal parameters in $\mathbf{\Lambda}$ model the stutter effects, so that λ_{ij} is the stutter peak produced at position i by allele j. The relative values of the diagonal elements of $\mathbf{\Lambda}$ determine differential amplification, so that $\lambda_{ii}/\lambda_{jj}$ measures the amplification of allele i relative to allele j. All of the diagonal elements in $\mathbf{\Lambda}$ will be positive, but many of the off-diagonal elements will be zero because each allele's stutter peaks usually affect only the next few smaller alleles (Figure 1).

Thus, if $\mathbf{y}$ and $\mathbf{\Lambda}$ are known or estimated, then $\mathbf{x}$ can be estimated. We show that $\mathbf{x}$ can be estimated using only the relative fluorescence intensities. We can define the matrix of relative amplification and fluorescence detection

parameters as $\mathbf{R} = \mathbf{\Lambda}/\lambda^*$, where λ^* is any nonzero element of $\mathbf{\Lambda}$, and note that $\mathbf{x} = \mathbf{q}2m$, where $\mathbf{q}$ is the vector of allele frequencies. Accordingly,

$$E(\mathbf{y}) = \lambda^*\mathbf{R}\mathbf{x} = \lambda^*\mathbf{R}\mathbf{q}2m \quad \text{and} \quad E(|\,\mathbf{y}\,|) = |\,\lambda^*\mathbf{R}\mathbf{q}2m\,| = \lambda^*\,|\,\mathbf{R}\mathbf{q}\,|\,2m,$$

implying that the ratio of these expectations is

$$\frac{E(\mathbf{y})}{E(|\,\mathbf{y}\,|)} = \frac{\mathbf{R}\mathbf{q}}{|\,\mathbf{R}\mathbf{q}\,|}.$$

Hence,

$$\mathbf{R}^{-1}\frac{E(\mathbf{y})}{E(|\,\mathbf{y}\,|)} = \mathbf{R}^{-1}\frac{\mathbf{R}\mathbf{q}}{|\,\mathbf{R}\mathbf{q}\,|} = \frac{\mathbf{q}}{|\,\mathbf{R}\mathbf{q}\,|} \quad \text{and} \quad \left|\mathbf{R}^{-1}\frac{E(\mathbf{y})}{E(|\,\mathbf{y}\,|)}\right| = \left|\frac{\mathbf{q}}{|\,\mathbf{R}\mathbf{q}\,|}\right| = \frac{1}{|\,\mathbf{R}\mathbf{q}\,|}.$$

Thus, $\mathbf{x}$ can be calculated via

$$2m\left(\mathbf{R}^{-1}\frac{E(\mathbf{y})}{E(|\,\mathbf{y}\,|)}\right)\Big/\left(\left|\mathbf{R}^{-1}\frac{E(\mathbf{y})}{E(|\,\mathbf{y}\,|)}\right|\right) = 2m\left(\frac{\mathbf{q}}{|\,\mathbf{R}\mathbf{q}\,|}\right)\Big/\left(\left|\frac{1}{|\,\mathbf{R}\mathbf{q}\,|}\right|\right) = 2m\mathbf{q} = \mathbf{x},$$

which, upon replacing expectations with observed values and using an estimate of $\mathbf{R}$, denoted $\hat{\mathbf{R}}$, leads to the moment-based estimator of $\mathbf{x}$,

$$\hat{\mathbf{x}} = 2m\left(\hat{\mathbf{R}}^{-1}\frac{\mathbf{y}}{|\,\mathbf{y}\,|}\right)\Big/\left(\left|\hat{\mathbf{R}}^{-1}\frac{\mathbf{y}}{|\,\mathbf{y}\,|}\right|\right) = 2m\left(\frac{\hat{\mathbf{R}}^{-1}\mathbf{y}}{|\,\hat{\mathbf{R}}^{-1}\mathbf{y}\,|}\right). \tag{B1}$$

The fluorescence intensities in $\mathbf{y}$ are absolute values, but Equation B1 shows, explicitly, that the estimation of allele counts depends only on the relative fluorescence intensities in a DNA pool, $\mathbf{y}/|\mathbf{y}|$. Moreover, the elements of $\mathbf{R}$ are parameters that are scaled relative to one of the elements of $\mathbf{\Lambda}$; thus, only relative fluorescence intensities are required to estimate $\mathbf{R}$. This result is important for two reasons: first, it provides a theoretical basis for the procedure of estimating allele counts from DNA pools, and second, it shows that, in a simple model, relative fluorescence intensities contain all of the information needed for this calculation.

Previous studies that estimate microsatellite allele frequencies from DNA pools either used an estimation procedure that is equivalent to Equation B1 (Barcellos *et al.* 1997; Kirov *et al.* 2000) or employed a simplification thereof by imposing constraints on the structure of $\mathbf{R}$ (Perlin *et al.* 1995; Band and Ron 1998; Lipkin *et al.* 1998, 2002; Mosig *et al.* 2001; Schnack *et al.* 2004). We estimated the matrix $\mathbf{R}$ by nonlinear least-squares regression via the model

$$\frac{y_i}{|\,\mathbf{y}\,|} = \frac{(\mathbf{R}\mathbf{x})_i}{|\,\mathbf{R}\mathbf{x}\,|} + e,$$

where the y_i's are the dependent variables measured from individuals of known genotype (12–24 individual samples per multiplex, except for locus *SB6* in experiment 2 when 24 independent pools of known allelic content were used) so that $\mathbf{x}$ is known, $(\mathbf{z})_i$ denotes the ith element of a vector $\mathbf{z}$, and e is random, uncorrelated error. Typically, estimates of $\mathbf{R}$ have nonzero diagonal elements, some nonzero upper-diagonal elements, and all other elements are zero (see Perlin *et al.* 1995; Barcellos *et al.* 1997).

APPENDIX C: ESTIMATION OF HARDY–WEINBERG DISEQUILIBRIUM COEFFICIENTS FROM DNA POOLS

Let $\hat{p}_{ai}$ and $\hat{p}_{bi}$ be the observed frequencies of alleles a and b in pool i. Using the definition of D_{ab}, the expected values of the products of these frequencies within and among r pools of size m are

$$E(\hat{P}) = E\left(\sum_{i=1}^{r}\hat{p}_{ai}\hat{p}_{bi}\right) = r\left[\left(1 - \frac{1}{2m}\right)p_a p_b - \frac{D_{ab}}{2m}\right]$$

and

$$E(\hat{C}) = E\left(\sum_{i=1}^{r}\sum_{i\neq i'}\hat{p}_{ai}\hat{p}_{bi'}\right) = r(r-1)p_a p_b.$$

Taking an appropriate linear combination of $\hat{P}$ and $\hat{C}$ yields an unbiased estimator of D_{ab},

$$\hat{D}_{ab} = \frac{2m-1}{r(r-1)}\hat{C} - \frac{2m}{r}\hat{P}.$$

Simulations suggest that the variance of $\hat{D}_{ab}$ increases with increasing m and decreases with increasing r.

APPENDIX D: COST EQUATIONS FOR DNA POOLING

For individual genotyping of n individuals, the statistical variance of the allele frequency estimate is

$$\frac{p(1-p)}{2n},$$

where p is the population allele frequency. We assume that the equivalent variance of an estimate calculated from r pools of size m is the sum of the technical variance and the statistical variance, yielding

$$\frac{1}{r}\left[\sigma_T^2 + \frac{p(1-p)}{2m}\right],$$

where σ_T^2 is the technical variance. Equating these expressions and solving for r yields the number of replicate pools of size m that must be constructed to obtain the same level of precision as genotyping n individuals:

$$r = \frac{2n}{p(1-p)}\left[\sigma_T^2 + \frac{p(1-p)}{2m}\right].$$

To calculate the percentage of cost savings of DNA pooling relative to individual genotyping, this expression for r is used in the cost-savings equation presented in the Introduction, yielding

$$100\left(1 - \frac{C_P}{C_I}\right)\% = 100\left[1 - \frac{rm(c_I + c_Q) + rl(c_A + c_E)}{nc_I + ln(c_A + c_E)}\right]\% = 100\left\{1 - \left[\frac{2\sigma_T^2}{p(1-p)} + \frac{1}{m}\right]\left[\frac{m(c_I + c_Q) + l(c_A + c_E)}{c_I + l(c_A + c_E)}\right]\right\}\%.$$

Differentiation leads to the pool size that maximizes percentage of cost savings,

$$m^* = \sqrt{\left(\frac{p(1-p)}{2\sigma_T^2}\right)\left(\frac{l(c_A + c_E)}{(c_I + c_Q)}\right)}.$$