Microsatellite analysis of pooled *Schistosoma mansoni* DNA: an approach for studies of parasite populations

L. K. SILVA[1,2,3], S. LIU[1] *and* R. E. BLANTON[1]*

[1] *Center for Global Health and Diseases, 2103 Cornell Road, Case University, Cleveland, OH 44106-7286, USA*
[2] *Centro Universitário da Bahia (FIB), R. Xingu, 179, STIEP, Salvador, BA 41770-130, Brazil*
[3] *União Metropolitana de Educação e Cultura (UNIME), Av. Luís Tarquínio Pontes, 600, Centro, Lauro de Freitas, BA 42700-000, Brazil*

(*Received 1 April 2005; revised 18 June and 24 August 2005; accepted 25 August 2005; first published online 28 October 2005*)

SUMMARY

Human parasites are often distributed in metapopulations, which makes random sampling for genetic epidemiology difficult. The typical approach to sampling *Schistosoma mansoni* involves laboratory passage to obtain individual worms with small sample size and selection bias as a consequence. By contrast, the naturally pooled samples from egg output in stool or urine directly represent the genetic composition of current populations. To test whether pooled samples could be used to estimate population allele frequencies, DNA from individual cloned parasites was pooled and amplified by PCR for 7 microsatellites. By polyacrylamide gel analysis, the relative band intensities of the products from the major alleles in the pooled samples differed by 0–6% from the summed intensities of the individual clones (mean = 2·1% ± 2·1% S.D.). The number of PCR cycles (25–40) did not influence the accuracy of the estimate. Varying the frequency of 1 allele in pooled samples from 32 to 69% likewise did not affect accuracy. Allele frequency estimates from aggregate samples such as eggs will be a better foundation for studies of parasite population dynamics as well as the basis for large-scale association studies of host and parasite characteristics.

Key words: allele frequency, PCR, pooled samples, genetics, metapopulation.

INTRODUCTION

Sampling is an essential component for any epidemiological study. Schistosomes, like many parasites, are difficult to assay due to their distribution in their hosts. By contrast, it is relatively easy to obtain a representative sample from an evenly distributed population with unrestricted movement. It is also relatively easy to sample large discrete organisms. Many human parasites, however, are segregated as populations within individual hosts and have more than one kind of host each with its own distinct biology and distributions. They are also too small to easily isolate individual organisms or are located at sites that are difficult to sample.

The sample size has a large effect on the power of a study to detect or refute a difference between populations. As Jarne and Theron observed ''... one should realize that studying too few individuals in a species with very large natural ranges could lead to inappropriate inferences about population biology'' (Jarne and Theron, 2001). Time and money often limit the ability to take large samples for studies of population genetics. In natural populations, there are very limited places where a sample can be obtained, especially in human infections. Even in laboratory-based studies where adult worms can be perfused from the portal system, it can be costly to individually genotype hundreds of worms. For most population indices such as the between population fixation index (F_{ST}) and the effective population size (N_e), the discrete genotype data are, in the end, combined as mean allele frequencies and population allelic variances. In statistical genetics, one technique used to develop preliminary results and generate hypotheses at a reduced cost has been to analyse DNA pooled from a large number of individuals. Genotyping pooled DNA using PCR-based approaches has been validated in the context of genetic association studies (Daniels *et al.* 1998; Shaw *et al.* 1998; Sham *et al.* 2002; Collins *et al.* 2000) with a typical error rate of 1–5%. It has rarely been applied to studies of population genetics. For *S. mansoni*, one potential sampling scheme would be to estimate population allele frequencies directly from the total population of worms in a single infected animal. A similar approach might be considered for the analysis of human or animal infections by genotyping the egg output in the stool or urine in the case of *S. haematobium*. This sample is necessarily composed of a pool of individuals. The main assumption in a pooling study is that all of the alleles amplify in proportion to their individual frequency in the study population. In order to apply this approach to the population

* Corresponding author: Center for Global Health and Diseases, Case Western Reserve University, 2103 Cornell Rd, Wolstein Research Building, Cleveland, Ohio 44106-7286, USA. Tel: +216 368 4814. Fax: +216 368 4825. E-mail: reb6@case.edu

doi:10.1017/S0031182005009066 Printed in the United Kingdom

genetics of *S. mansoni*, proportional allele amplification needs to be demonstrated. Using 7 new microsatellite markers and genomic DNA from *S. mansoni* clones, we explored conditions under which marker amplification of DNA pooled from cloned parasites would provide accurate estimates of population allele frequencies.

MATERIALS AND METHODS

S. mansoni *clones*

The *S. mansoni* strain at Case University (CASE) was originally derived from the Naval Medical Research Institute (NMRI) strain in 1961. The CASE strain has been maintained by passage in snails (*Biomphalaria glabrata*) and CF1 mice since that time. Asexual reproduction takes place in the snail intermediate host, thus, infection of snails with a single miracidium produces a clonal population of cercariae. This is the mammalian infective stage and will produce a clonal population of adult worms in mice. Single miracidia were obtained by serial dilutions and confirmed by visual inspection. These were then used to infect snails. CF1 mice were infected with cercariae from only 1 infected snail. The resulting adult worms were examined to determine that they consisted of the same sex. All infections were single-sex and were clonal by subsequent microsatellite analyses. In addition, DNA from approximately 250 outbred adult worms from the CASE strain were obtained by perfusion of the portal system of infected mice and was used for microsatellite amplification.

DNA extraction

DNA was extracted from adult worms using a standard proteinase K : phenol : chloroform organic extraction protocol (Sambrook and Russell, 2001). DNA concentration from all clones was estimated by spectrophotometric quantitation (GeneQuant Pro RNA/DNA Calculator, Amersham, Piscataway, NJ, USA) and aliquots with a standard concentration of 10 ng/μl were prepared. Equal volumes of each were used to pool samples for analysis.

Microsatellite identification and primer selection

All *S. mansoni* cDNA and expression sequence tags available to GenBank in November 1998 were screened for repetitive motifs using the program RepeatMasker (Smit, 2003) or by a BLAST (Altschul *et al*. 1990) search of the NCBI database for simple repeats. Selection criteria for use were: (1) presence of uninterrupted tri- or tetranucleotide repeating elements, (2) presence of at least 6 repeating units in the database, (3) amplification of only 1 or 2 bands of equal intensity corresponding to the allele (single-copy locus) and (4) presence of polymorphisms in the laboratory strain of *S. mansoni*. A total of 21 *S. mansoni* microsatellite loci designated SMMS 1 to 21 fit the first 2 criteria. Primers were designed from the flanking sequences to produce a 100–350 bp amplicon and to have an optimal annealing temperature near 50 °C. Seven loci that were polymorphic for the CASE strain were used in this study: SMMS2, SMMS3, SMMS13, SMMS16, SMMS17, SMMS18 and SMMS21 (Table 1). Alleles at each locus will be referred to by their size in base pairs as determined on polyacrylamide gels.

Genotyping

PCR reactions were performed in a total volume of 25 μl with 50 ng of genomic DNA, 0·8 μM of each primer; 2 units of Taq DNA polymerase (New England Biolabs, Beverley, MA, USA); 1× Taq buffer with 2 mM $MgSO_4$; and 0·4 mM each dNTP. The reaction was carried out on an MJ Research PTC 200 thermocycler (MJ Research Inc., Reno, NV, USA) using the same conditions for each primer. After an initial denaturation step at 94 °C for 3 min, cycling continued at 94 °C for 1 min, 50 °C for 1 min, and extension at 72 °C for 2 min for 25–40 cycles. The final extension was at 72 °C for 10 min. The PCR products were separated by 12% PAGE and visualized with ethidium bromide. Band sizes were determined by comparison with a 100 bp DNA ladder (New England Biolabs, Beverly, MA, USA). The gels were digitized using a gel documentation system (Doc-It System, UVP, CA, USA). Each experiment included a negative control in which water was substituted for DNA (Fig. 1).

Allele isolation and sequence

For sequence analysis, a DNA band corresponding to a single allele was cut from the gel and eluted by incubation in 0·5 M ammonium acetate and 1 mM EDTA overnight at 37 °C. After centrifugation, DNA was ethanol precipitated, suspended in water (Sambrook and Russell, 2001) and sequenced by a standard ABI protocol (Fig. 2).

Creation of DNA pools

Only spectrophotometric measurements and simple pipette techniques were employed in pool formation, so that DNA quantification was not as accurate as possible. Using these measurements, equal amounts of DNA from 12 parasite clones were combined, and this pool of 12 was amplified. We therefore compared the relative intensity of the amplified bands from pooled DNA to the summed intensity of each cloned parasite as well as allele counts. To explore the effect of variation in pool composition on estimates of allele intensity, 9 pools using different

https://doi.org/10.1017/S0031182005009066 Published online by Cambridge University Press

Table 1. Microsatellite primers and characteristics

Microsatellite locus	Accession number	Repeat sequence	Number of alleles†	Primer pairs (5′–3′)	Polarity	Expected size*
SMMS 2	AI067617	(CAAA)6	2	GAAGGTCATTATATTCGTC	+	239
				GTTGAAATCTATACAG	–	
SMMS 3	AI067567	(TAA)12	4	GGTCAACAGCAATATCAGC	+	192
				GATCATCTTCATGACGTCG	–	
SMMS 13	AI395184	(ATT)7	2	GGCGAAGACGACGGAGAAG	+	189
				GTAATGTATAAATAGGG	–	
SMMS 16	AF325694	(TTA)11	3	CACCCATTGTCTTAAAACC	+	231
				GATGTCACACCCTC	–	
SMMS 17	M85304	(AAT)8	2	CATTTCCCATCTTCAAC	+	292
				CTAAAGCTGGGCACC	–	
SMMS 18	BF936409	(AAT)13	3	CACCTCAACACCTATG	+	224
				GTTGGAAACACATTGGGC	–	
SMMS 21	AI110905	(TTA)10	2	GGTTGTCTGTCGTCCCC	+	187
				GGTACTAGTGGTTGAATAC	–	

† Allele number for the CASE strain of *S. mansoni*.
* Based on GenBank sequence.

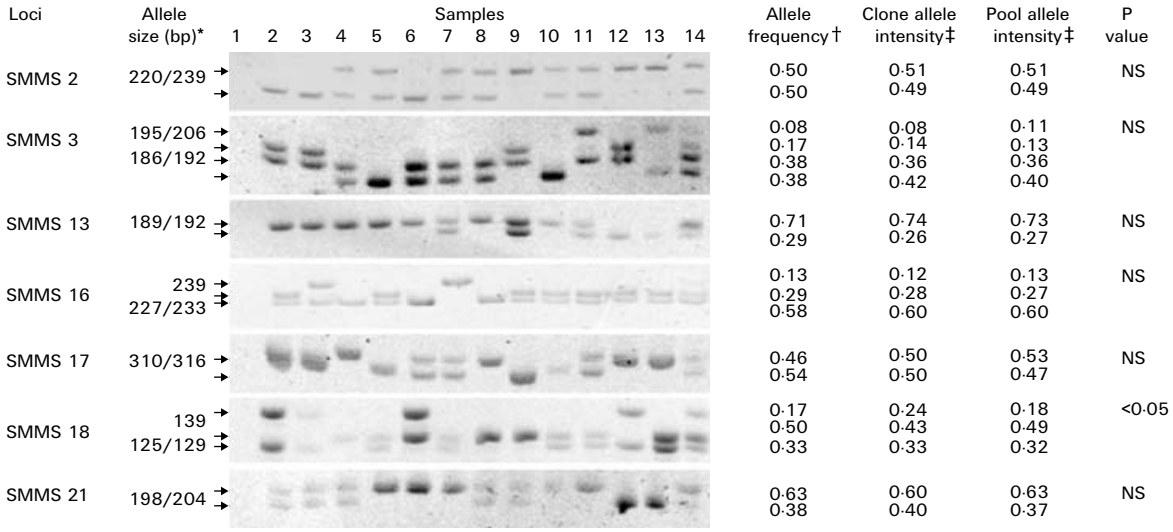

Fig. 1. Genotype of individual *Schistosoma mansoni* clones. Alleles were separated in a 12% polyacrylamide gel. Lane 1, negative control; Lanes 2–13, clones 1, 7, 10, 12, 17, 21, 22, 23, 24, 25, 26 and 27, respectively; Lane 14, pool composed of each of the 12 clones.
* Size (bp) for those that were not sequenced was estimated using Sigmagel Gel Analysis Software 1.0 (Jandel Scientific, SPSS Science, IL, USA). † Allele frequency determined by allele counts. ‡ Corrected values for unequal allelic amplification.

combinations of 6 of the 12 clones were also prepared randomly and analysed for SMMS2. Individual clones and pools were always amplified at the same time, under the same conditions and electrophoresed in the same gel to allow for valid comparison.

Data analysis

The observed heterozygosity (Ho) based on allele counts was compared to the expected heterozygosity (He) under Hardy-Weinberg equilibrium. Both the expected heterozygosity and the Chi-square were calculated using the Genetic Calculation Applets (Christensen, 2003). All other statistical testing was performed with the program SPSS for Windows V. 10.0.5 (SPSS, Inc., Chicago, IL, USA).

Allele intensity and molecular weight were determined using Sigmagel Gel Analysis Software v. 1.0 (Jandel Scientific, SPSS Science, Chicago, IL, USA). The intensity of the local integrated area (pixel units) was measured in triplicate at the centre of each band, and the mean of the areas was recorded. Gaussian curves were also used to facilitate definition of the area for local integration. The intensity of alleles in individual samples may vary slightly due to size or differential initiation of the polymerase (Daniels *et al*. 1998), so a correction factor was calculated for unequal allelic amplification (k) based on

https://doi.org/10.1017/S0031182005009066 Published online by Cambridge University Press

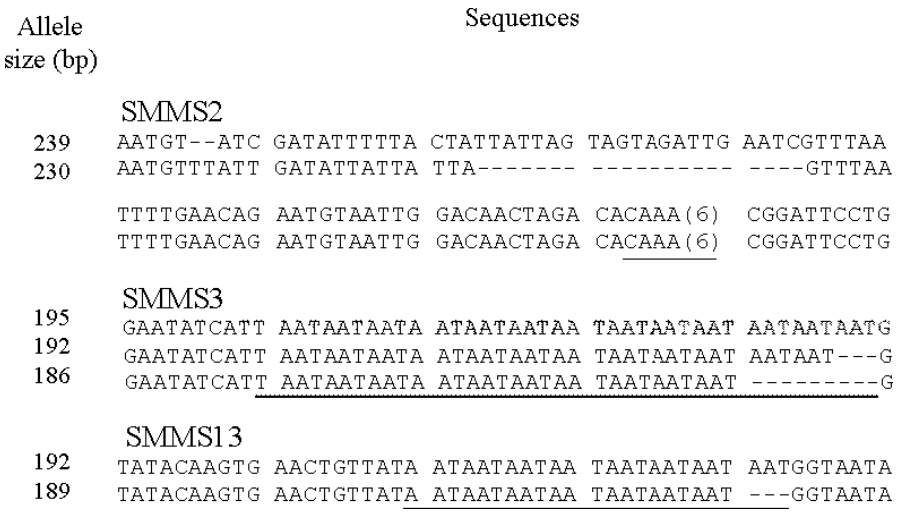

Fig. 2. Allele sequences for 3 microsatellite loci. Allele DNA was isolated from polyacrylamide gels and sequenced to determine the source and nature of variation. Underlined are the repeat regions corresponding to the microsatellite. Dashed regions represent the polymorphic insertions or deletions.

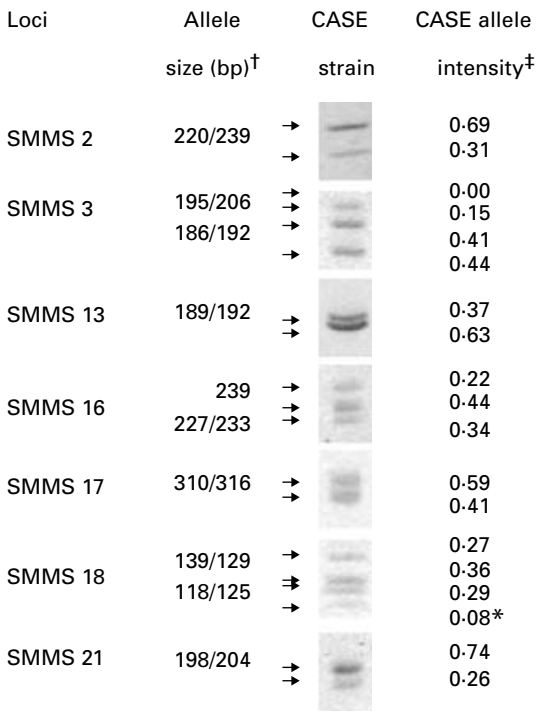

Fig. 3. Allele frequency estimation in the CASE strain pool. † Size (bp) for those that were not sequenced was estimated using Sigmagel Gel Analysis Software 1.0 (Jandel Scientific, SPSS Science, IL, USA). ‡ Corrected values for unequal allelic amplification. * This allele was not detected among the 12 clones we worked with.

the mean ratio of allele intensities for heterozygotes at each locus (i). The smallest allele was taken as the reference for each locus ($k = 1{\cdot}00$). The estimator $\overline{k}_i$ is assumed to be unbiased (Hoogendoorn *et al.* 1999; Le Hellard *et al.* 2002). The final allele intensities from the clones and pooled samples were corrected for by differential amplification using the corresponding correction factor.

True allele frequencies were determined by allele counting. The relative intensity of each band was also calculated by summing the intensity measured by Sigmagel for each allele and dividing this by the total intensity for all alleles. Statistical differences between values obtained for the clones and for the pooled samples were assessed by Chi-square test with Yates' correction. However, ANOVA F-test was used to test the overall concordance of the relative intensities obtained from 9 pools with allele frequencies different from those of the summed individual parasites clones.

RESULTS

Microsatellite characteristics

Out of 21 primer pairs for potential microsatellites, 9 amplified, were associated with polymorphic amplicons and behaved as single-copy loci. Seven of these were chosen for this study, since they produced the clearest patterns in pooled samples (Table 1).

For each cloned parasite, only 1 or 2 alleles were present, indicating that these microsatellites behave like single-copy loci. No stuttering was observed for these tri- and tetra-nucleotide repeats. The loci contained from 2–4 alleles for the 12 clones examined (Fig. 1).

The observed heterozygosity of these 7 microsatellites ranged from 0·25 to 0·83 (mean = 0·52 $\pm$0·23 S.D.) in this population. Hardy-Weinberg equilibrium was observed for each of the loci studied except for SMMS16 (Table 2). This may have been due to the small sample size represented by the 12 clones.

Locus sequence variation

Some alleles did not appear to differ by factors of 3 or 4 indicating that the heterogeneity in some cases

https://doi.org/10.1017/S0031182005009066 Published online by Cambridge University Press

Table 2. Observed (*Ho*) compared to expected heterozygosity (*He*) at each locus

Microsatellite locus	*Ho**	*He***	Chi-square; df†	*P* value‡
SMMS 2	0·50	0·50	<0·0001; 1	N.S.
SMMS 3	0·83	0·68	8·4444; 6	N.S.
SMMS 13	0·25	0·41	1·8719; 1	N.S.
SMMS 16	0·67	0·56	8·1088; 3	<0·05
SMMS 17	0·25	0·50	2·9581; 1	N.S.
SMMS 18	0·75	0·61	4·5000; 3	N.S.
SMMS 21	0·42	0·47	0·1481; 1	N.S.

* Based on allele counts from 12 clones from the CASE strain (Fig. 1).
** The expected heterozygosity (*He*) under Hardy-Weinberg equilibrium and Chi-square were calculated by using the Genetic Calculation Applets (Christensen, 2003).
† The degrees of freedom equaled 1 for 2 allele systems; 3 for 3 allele systems; and 6 for 4 allele systems. The results were rounded to 4 decimal places. Critical values of Chi-square for $P=0{\cdot}05$ and degree of freedom equal to 1, 3 and 6 were 3·84, 7·82 and 12·59, respectively.
‡ N.S., Not significant.

was due to mechanisms other than simple slippage of the template. The DNA sequence was determined for several alleles isolated from the amplified microsatellite loci SMMS 2, 3 and 13. The sequence confirmed that the correct region had been amplified and that for locus 3 and 13, the polymorphism represented the insertion or deletion of 1–3 repetitive units (Fig. 2). The 2 alleles of locus SMMS 2 in the CASE strain, however, did not differ at the tetranucleotide repeat region, but rather at upstream insertions or deletions for a total of 19 nucleotides difference between these alleles. This indel is a common polymorphism in several strains from the Americas (not shown).

Comparison between individual clones and pooled data

Intensities from each band from the amplified individual clones and pooled samples were measured and corrected for unequal allelic amplification. The mean k value for unequal amplification was $1{\cdot}16 \pm 0{\cdot}22$ S.D. (range 1·00–1·60). The clone allele intensity and pool allele intensity were similar at the level of significance of 5% for all loci either before or after correction, except for SMMS18 ($P<0{\cdot}05$) (Fig. 1). The difference between the clones and the pooled sample in relative intensities of the major allele from each locus ranged from 0 to 7% (mean $=2{\cdot}5\% \pm 2{\cdot}3\%$) and from 0 to 6% (mean $=2{\cdot}1\% \pm 2{\cdot}1\%$) for uncorrected and corrected values, respectively. The highest discrepancies were observed with SMMS18.

The effect of cycle number

The effect of the number of amplification cycles on the estimate of allele frequency was tested by varying the number of cycles (25, 30, 35 and 40) for locus SMMS 2. Comparing the pool allele intensities to the clone allele intensities, there was only a 0–2% difference. The difference between the pool allele intensity and the true allele frequency at different cycles ranged from 0 to 4%. The cycle number, therefore, did not affect the relative efficiency of amplification. Within this range, the plateau effect did not interfere with our measurements. At 35 and 40 cycles, there was an increase in the intensity and appearance of artifacts (extra bands) outside the range of microsatellite amplification. The increased number of cycles did produce a stronger signal.

Estimates of CASE strain allele frequencies

In order to assess how representative the 12 clones were of the CASE strain, DNA was extracted from one infection cohort and amplified for each of the 7 microsatellite loci. Since mice are infected in groups periodically, and eggs from these infections are used to continue the laboratory's life-cycle, each infection cohort (~250 worms) should be highly representative of the CASE strain. Differences in the relative intensities of the major allele in each locus for the two pools ranged from 6 to 36% (mean $=16{\cdot}3\% \pm 11{\cdot}6\%$ S.D.) (Fig. 3). In some cases the 'major' allele in the pool reversed its relative frequency to become the minor allele in the CASE strain, e.g. loci SMMS13 and SMMS16. Allele 206 from locus SMMS3 was evident in the pool of 12, but was not apparent in the cohort DNA. This allele has been observed faintly in the CASE strain DNA on other occasions where more cycles were used for amplification (not shown). On the other hand, allele 118 was observed only in locus SMMS18 only in DNA from the CASE strain (Fig. 3).

Allele frequency estimation

The true allele frequencies were computed by counting the number of each allele and dividing by the total number of alleles present in the individuals that were genotyped (Fig. 1). Relative intensity of each allele was calculated by summing the intensity at each allele and dividing by the total intensity for all alleles in all individuals. When the clone allele intensities and the pool allele intensities were compared, to the true allele frequency of each locus, all were similar, including SMMS18. Thus, the pooled values proved to be good estimators of the actual allele frequency. For the pooled samples, the relative intensities of the major alleles from each locus differed from the true allele frequency by 1–8% (mean $=4{\cdot}2\% \pm 2{\cdot}2\%$ S.D.) for uncorrected data

https://doi.org/10.1017/S0031182005009066 Published online by Cambridge University Press

Table 3. Clone and pooled allele intensity in 9 pools of 6 randomized samples using *Schistosoma mansoni* microsatellite locus 2 (SMMS2)

Pool number	Allele frequency*	Clone allele intensity	Pool allele intensity¶	Absolute difference	*P* value†,‡
1	0·25	0·32	0·29	0·03	N.S.
2	0·42	0·43	0·45	0·02	N.S.
3	0·42	0·47	0·47	0·00	N.S.
4	0·42	0·45	0·42	0·03	N.S.
5	0·42	0·43	0·40	0·03	N.S.
6	0·33	0·41	0·37	0·04	N.S.
7	0·58	0·64	0·64	0·00	N.S.
8	0·58	0·61	0·62	0·01	N.S.
9	0·58	0·54	0·63	0·09	$<0·01$

* Allele frequency determined by allele counts.
¶ Corrected values for unequal allelic amplification.
† Chi-square test with Yates' correction was used to compare the clone allele intensity to the pool allele intensity. The relative intensity for the smallest allele (220) is reported in the table. N.S., Non-significant.
‡ ANOVA F-test was used to compare the overall concordance among the relative intensities and showed Person's correlation between clone and pooled proportions (adjusted $r^2=0·944$, $P<0·01$) and between polled data and true allele frequency (adjusted $r^2=0·955$, $P<0·01$).

and by 0 to 7% (mean = 2·2% ± 2·2% S.D.) for the corrected data. Performing the correction for unequal amplification reduces the mean difference between the true allele frequencies and relative band intensities by 2%.

The reproducibility of allele frequency estimates was tested by comparing multiple pools to the results of genotyping their component individual clones. In the 9 pools of 6 clones, the true allele frequency of the smallest allele ranged from 0·25 to 0·58 for locus SMMS2, while the relative intensity ranged from 0·32 to 0·64 (Table 3). The difference between the clone and pooled relative intensities ranged from an absolute value of 0·00 to 0·09 (mean = 2·8% ± 2·7%). By Chi-squared test there were no significant differences between the expected intensities and those observed in the pools except in 1 case. Comparison over all 9 pools showed that the differences were not significant. The Pearson's correlation was very high between the sum of individual clones and pools for relative allele intensity (adjusted $r^2=0·944$; $P<0·01$). Most importantly, there was a good correlation between the pooled values and the true allele frequencies (adjusted $r^2=0·955$; $P<0·01$).

DISCUSSION

We report here on the validation of 7 new microsatellite markers of S. mansoni and their application to pooled DNA samples. Our data show that in pooled samples of DNA, these microsatellite loci are represented in proportion to their input DNA after PCR amplification. The accuracy of these estimates is not affected by the number of PCR cycles in the range from 25–40, nor is it affected by relative allele frequencies of the pooled samples. We provide correction factors for the microsatellites used here, and for other loci, a correction factor can easily be generated when DNA from heterozygote clones is available. Estimates without this correction can still be made with acceptable degrees of error. Comparing the relative intensities of the alleles on polyacrylamide gels, the difference between the proportions of amplifications of the 12 clones and their pool averaged 1·9% ± 1·9% (range 1–6%). Even though only spectrophotometric measurements were used in quantifying DNA, the relative allele intensities for the pooled samples differed from allele frequencies measured by allele counting by 2·2% ± 1·9%. This difference is similar to that found in other studies that genotyped pooled DNA (Sham *et al*. 2002).

Two factors in the selection of microsatellite loci contribute to the accuracy of the estimates. First, only tri- and tetranucleotide repeats were selected since they amplify with less stuttering than dinucleotide repeats (Hughes and Queller, 1993). In a pooled sample the faint extra bands associated with the target amplicons could not be discounted. Using single-copy locus markers also produces a cleaner, simpler profile on a gel than when there are multiple copies, and avoids the possibility of individual variation in locus copy number.

The genotyping of pooled DNA offers a way to increase the number of individuals tested and thereby increase the power of a study. While discrete genotyping data can certainly be obtained from individual worms, pooling allows hypothesis testing on a much larger sample. Currently, no chromosomal localizations have been published for *Schistosoma mansoni* microsatellites, so that they cannot yet be used for gene finding in association studies, but certain questions might be addressed within a genetic association framework. For example, whether

https://doi.org/10.1017/S0031182005009066 Published online by Cambridge University Press

different hosts (different mouse strains, primates, hamsters) have selection bias for particular genotypes of worms might be studied efficiently by genotyping the pool of worms from each. Another example where this approach could prove useful might be to genotype pools of worms from male and female mice to test for gross sex-based differences in infection or survival. In either case, a Chi-squared 'goodness-of-fit' test or logistic regression could be applied. Multiplying the allele frequencies by the number of worms would provide allele counts needed for these statistical tests. Such an approach would require analysis of large numbers of organisms for validity.

In addition to the number of subjects, the method of sampling is an essential component for any epidemiological study, and the sampling scheme should be guided by how the population is distributed in space. It is relatively easy to obtain a representative sample from an evenly distributed population with unrestricted movement. Many human parasites, however, are aggregated within individual hosts [infrapopulations (Bush *et al*. 1997)], and many have more than one kind of host each with its own distinct biology and distribution. Therefore, not only is there a need for sampling many parasites from one host, but many hosts from one geographical region should also be sampled in order to be representative. For population genetics, the typical indices, such as the between population fixation index (F_{ST}) and the effective population size (N_e) can be generated from allele frequencies in programs such as ARLEQUIN (Schneider, Roessli and Excoffier, 2000), or MLNe (Wang, 2001). Mixed-stock analysis employed by the program SPAM (Utter and Ryman, 1993) can use allele frequencies to compare the origin of different populations. Thus, large samples could be used to study the relationship between strains or the degree of interchange between natural populations. Comparing these indices could likewise serve as a way to monitor the effect of a control programme or ecological changes on schistosome populations. Finally, genetic association studies comparing populations for biologic characteristics (morbidity, praziquantel resistance, age or sex of host) can be performed with allele frequencies and standard nonparametric tests such as Fisher's exact test.

Although analysis of pooled samples is likely to be a powerful approach for population genetic epidemiology of *S. mansoni* and some other parasites, it is important to recognize its potential limitations. In most studies of pooled samples, the formation of the pool itself is the main source of error. If all parasites are used in forming the pool this is no longer a source of error. Finally, there is no way to assess Hardy-Weinberg proportions on pooled data. Approaches that use the expectation of Hardy-Weinberg proportions such as structure (Pritchard *et al*. 2000), probably cannot be applied to pooled data. As discussed earlier, however, important indices of population structure can be used with just allele frequency data. It is hoped that this work sets the stage for larger, more efficient studies of *S. mansoni* population genetics. We are extending this work to the analysis of schistosome egg DNA so that the technique can be applied to studies in human infection and may also decrease the need for animal studies.

The authors would like to thank Edilson Machado de Assis, M.S. at the Catholic University of Salvador, School of Engineering, Salvador-BA, Brazil for useful discussions on the study design and data analysis. This work was supported in part by NIH grant AI41680.

REFERENCES

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology* **215**, 403–410.

Bush, A. O., Lafferty, K. D., Lotz, J. M. and Shostak, A. W. (1997). Parasitology meets ecology on its own terms: Margolis *et al*. revisited. *The Journal of Parasitology* **83**, 575–583.

Christensen, K. (2003). Genetic calculation applets 3/31/01. [Online.]

Collins, H. E., Li, H., Inda, S. E., Anderson, J., Laiho, K., Tuomilehto, J. and Seldin, M. F. (2000). A simple and accurate method for determination of microsatellite total allele content differences between DNA pools. *Human Genetics* **106**, 218–226.

Daniels, J., Holmans, P., Williams, N., Turic, D., McGuffin, P., Plomin, R. and Owen, M. J. (1998). A simple method for analyzing microsatellite allele image patterns generated from DNA pools and its application to allelic association studies. *American Journal of Human Genetics* **62**, 1189–1197.

Hoogendoorn, B., Owen, M. J., Oefner, P. J., Williams, N., Austin, J. and O'Donovan, M. C. (1999). Genotyping single nucleotide polymorphisms by primer extension and high performance liquid chromatography. *Human Genetics* **104**, 89–93.

Hughes, C. R. and Queller, D. C. (1993). Detection of highly polymorphic microsatellite loci in a species with little allozyme polymorphism. *Molecular Ecology* **2**, 131–137.

Jarne, P. and Theron, A. (2001). Genetic structure in natural populations of flukes and snails: a practical approach and review. *Parasitology* **123**, S27–S40.

Le Hellard, S., Ballereau, S. J., Visscher, P. M., Torrance, H. S., Pinson, J., Morris, S. W., Thomson, M. L., Semple, C. A., Muir, W. J., Blackwood, D. H., Porteous, D. J. and Evans, K. L. (2002). SNP genotyping on pooled DNAs: comparison of genotyping technologies and a semi automated method for data storage and analysis. *Nucleic Acids Research* **30**, e74.

Pritchard, J. K., Stephens, M. and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics* **155**, 945–959.

Sambrook, J. and Russell, D. W. (2001). *Molecular Cloning: A Laboratory Manual*, 3rd Edn. Cold Spring Harbor Press, New York.

Schneider, S., Roessli, D. and Excoffier, L. (2000). ARLEQUIN: a software for population genetics data analysis. University of Geneva 2.000. [Online.]

Sham, P., Bader, J. S., Craig, I., O'Donovan, M. and Owen, M. (2002). DNA pooling: a tool for large-scale association studies. *Nature Reviews Genetics* **3**, 862–871.

Shaw, S. H., Carrasquillo, M. M., Kashuk, C., Puffenberger, E. G. and Chakravarti, A. (1998). Allele frequency distributions in pooled DNA samples: applications to mapping complex disease genes. *Genome Research* **8**, 111–123.

Smit, A. F. (2003). RepeatMasker open-3.0. [Online.]

Utter, F. and Ryman, N. (1993). Genetic markers and mixed stock fisheries. *Fisheries* **18**, 11–21.

Wang, J. (2001). A pseudo-likelihood method for estimating effective population size from temporally spaced samples. *Genetical Research* **78**, 243–257.

https://doi.org/10.1017/S0031182005009066 Published online by Cambridge University Press