Molecular Ecology (2004) **13**, 1423–1431 doi: 10.1111/j.1365-294X.2004.02159.x

Single nucleotide polymorphism (SNP) discovery in mammals: a targeted-gene approach

NICOLA AITKEN,* STEVEN SMITH,† CARSTEN SCHWARZ‡ and PHILLIP A. MORIN§
Laboratory for Conservation Genetics, Max Planck Institute for Evolutionary Anthropology, Inselstrasse 22, D-04103, Leipzig, Germany

Abstract

Single nucleotide polymorphisms (SNPs) have rarely been exploited in nonhuman and nonmodel organism genetic studies. This is due partly to difficulties in finding SNPs in species where little DNA sequence data exist, as well as to a lack of robust and inexpensive genotyping methods. We have explored one SNP discovery method for molecular ecology, evolution, and conservation studies to evaluate the method and its limitations for population genetics in mammals. We made use of 'CATS' (or 'EPIC') primers to screen for novel SNPs in mammals. Most of these primer sets were designed from primates and/or rodents, for amplifying intron regions from conserved genes. We have screened 202 loci in 16 representatives of the major mammalian clades. Polymerase chain reaction (PCR) success correlated with phylogenetic distance from the human and mouse sequences used to design most primers; for example, specific PCR products from primates and the mouse amplified the most consistently and the marsupial and armadillo amplifications were least successful. Approximately 24% (opossum) to 65% (chimpanzee) of primers produced usable PCR product(s) in the mammals tested. Products produced generally high but variable levels of readable sequence and similarity to the expected genes. In a preliminary screen of chimpanzee DNA, 12 SNPs were identified from six (of 11) sequenced regions, yielding a SNP on average every 400 base pairs (bp). Given the progress in genome sequencing, and the large numbers of CATS-like primers published to date, this approach may yield sufficient SNPs per species for population and conservation genetic studies in nonmodel mammals and other organisms.

Keywords: ascertainment, conservation genetics, CATS, population genetics, SNP

Received 16 October 2003; revision received 29 January 2004; accepted 29 January 2004

Introduction

Technologies and genetic markers for molecular ecology and evolutionary and conservation biology in the last decade have included DNA fingerprinting, mitochondrial DNA sequencing or restriction fragment length polymorphism (RFLP) analysis, nuclear gene sequencing and genotyping of various types of nuclear loci, such as microsatellites and amplified fragment length polymorphisms (AFLPs). Although each of these approaches is adequate and appropriate for certain types of questions, all suffer from some technical (cost, efficiency, accuracy, transferability, repeatability, etc.) and analytical (accuracy of analytical models, variability in mutation rates and patterns, difference in modes of inheritance) issues (e.g. Rosenbaum & Deinard 1998; Schlötterer & Pemberton 1998; Luikart & England 1999). In particular, the two most widely used marker types, mitochondrial DNA and microsatellites, suffer from technical and analytical issues that limit their application. Analytical issues include the limitations of a single locus and its inheritance patterns (mtDNA, Hare 2001; Harpending *et al*. 1998), high and variable mutation rates that are difficult to model for appropriate analysis (microsatellites, Excoffier & Yang 1999; Balloux & Lugon-Moulin 2002), limitations due to sample size (Rao 2001) and technical issues such as nuclear inserts of mitochondrial DNA (Numts, Bensasson *et al*. 2001) and microsatellite stutter bands, null alleles and

Correspondence: Phillip Morin. Fax: 858 546 7003; E-mail: Phillip.Morin@noaa.gov
Current addresses: *Applied Ecology Research Group, University of Canberra, ACT 2601, Australia, †Griffith University, QLD 4111, Australia, ‡Department of Anthropology, McMaster University, Hamilton, Ontario, Canada and §Southwest Fisheries Science Center, 8604 La Jolla Shores Dr, La Jolla, CA 92037, USA

allelic dropout (Navidi *et al*. 1992; Callen *et al*. 1993; Taberlet *et al*. 1996; Gagneux *et al*. 1997; Morin *et al*. 2001).

An ideal genetic marker for population and evolutionary studies would possess at least three properties. First, the markers should be abundant and distributed widely across the genome to avoid biases associated with single locus analysis (or use of few loci). Second, well-understood and well-characterized models of evolution should apply to facilitate analysis and interpretation. Finally, technical applications must allow data acquisition from many loci scored in large population samples, and the data must be comparable across laboratories using different genotype scoring methods or technologies (Sunnucks 2000).

One promising new type of marker, single nucleotide polymorphisms (SNPs), has many of the characteristics of an ideal marker (Vignal *et al*. 2002; Brumfield *et al*. 2003; Morin *et al*. 2004). A SNP is a change in the nucleotide composition of a DNA sequence at a single site, and these changes are found typically every 300–1000 base pairs (bp) in most genomes (Brouillette *et al*. 2000; Sachidanandam *et al*. 2001; Shubitowski *et al*. 2001). Recently, the rate and patterns of mutation in several genomes have been characterized extensively (Nachman & Crowell 2000; Ebersberger *et al*. 2002; Silva & Kondrashov 2002); the single nucleotide mutation rates seem to be relatively low ($\sim 10^{-8}$), similar across sites and in agreement with an infinite sites model.

Although SNPs are common they are typically biallelic, so individual locus information content is low. In population analyses, this must be compensated for by the use of many more SNPs (relative to microsatellites) to obtain statistical power. For parentage analysis and individual identification, the number of SNP loci needed to match the power of 10–15 microsatellites has been estimated at about 30–50, depending on the frequencies of the alleles (Chakraborty *et al*. 1999; Krawczak 1999; Fries & Durstewitz 2001; Glaubitz *et al*. 2003). It is likely that this number also will be adequate for population analysis, although larger numbers of SNPs will become feasible and probably desirable (Edwards & Beerli 2000; Pluzhnikov & Donnelly 1996; Luikart & England 1999; Kuhner *et al*. 2000; Nielsen 2000; Wakeley *et al*. 2001).

Despite some obvious advantages of SNPs and their increasing use in human and model organism studies, they have not been employed frequently to date in studies of nonmodel organisms. This is primarily because of technical limitations to finding SNPs in relatively unknown genomes, and producing genotypes efficiently and cost-effectively.

In the last few years, these technological hurdles have been largely overcome. SNP discovery in the absence of databases of comparative sequences for an organism of interest can still take place by sequencing random DNA fragments (Karl & Avise 1993; McLenachan *et al*. 2000; Bensch *et al*. 2002; Primmer *et al*. 2002; Nicod & Largiader 2003) or by using a targeted gene approach, with primers designed from conserved regions of aligned genes of at least two species (e.g. mouse and human) to amplify a less conserved region (e.g. an intron or 3′ UTR). This latter type of primers has been termed 'comparative anchor tagged sequences' ('CATS', Lyons *et al*. 1997) or 'exon priming intron crossing' ('EPIC', Palumbi & Baker 1994). The advantages of this approach include wide, current availability of primers (Palumbi & Baker 1994; Venta *et al*. 1996; Friesen *et al*. 1997; Lyons *et al*. 1997; Strand *et al*. 1997; Bagley & Gall 1998; Friesen *et al*. 1999; Brouillette *et al*. 2000; Shubitowski *et al*. 2001; Primmer *et al*. 2002), knowledge of the gene ortholog in which the SNPs are found, which could be useful for detecting selection on ecologically important genes (Crandall *et al*. 2000; Reed & Frankham 2001; McKay & Latta 2002; van Tienderen *et al*. 2002) and potentially broad application over a group of species with less per-sequence initial effort to find SNP loci than might be required using a random sequence approach (for review, see Morin *et al*. 2004). The use of intron sequences as a source of variation has been exploited extensively to date. Levels of variation appear adequate for use in both phylogenetic (e.g. DeBry & Seshadri 2001) and population genetic (e.g. Lessa 1992; Bierne *et al*. 2000) studies. Some limitations of this approach include the potential for amplifying paralogous genes from gene families or repetitive loci, which can result in incorrect inference of genotypes, and biased inference of historical events as a result of the effects of selection acting on the associated genes. Because of the relatively large number of SNP loci used for population history inference, this is unlikely to bias the overall analysis significantly, and has the advantage of allowing identification of genes involved in diversifying selection (Purugganan & Gibson 2003; van Tienderen *et al*. 2002). It is also worth noting that markers obtained randomly from the genome are not necessarily free from the effects of selection, but their anonymity precludes a priori inference of selection based on knowledge of linked gene functions.

Whichever sequence generation method is employed, SNP detection requires multiple sample sequences to be generated from each locus. Heterozygote sequencing has been demonstrated to be an effective method for detecting SNPs (Brouillette *et al*. 2000; Shubitowski *et al*. 2001), especially with the use of specialized software (e.g. POLYPHRED, Nickerson *et al*. 1997) to assist in SNP detection, either directly from all amplified samples, or on a subset of samples subsequent to high throughput (but potentially less accurate) methods of mutation screening (Oleykowski *et al*. 1998; Wolford *et al*. 2000; Brumfield *et al*. 2003).

This study was undertaken to determine the feasibility of using the targeted gene approach to find SNPs in any mammal species. We have used 202 previously published CATS loci to screen 16 species from the major mammalian lineages. Subsets of amplified loci were chosen for each species for DNA sequencing, and sequences were scored for quality and sequence similarity to the gene from which the primers were designed. Sequences were generated from multiple individuals of one species, the chimpanzee, as well as from a set of pooled samples from that species. Sequences were analysed in both directions to assess the relative ability to detect SNPs from one- vs. two-directional sequencing. A subset of high-confidence SNPs was chosen for assay design, and the assays were validated by genotyping the previously sequenced and additional individuals.

There are a number of established methods, varying in efficiency and cost, for SNP genotyping. Some methods appropriate for molecular ecology studies are compared in Morin *et al*. (2004); see Syvänen (2001) for a thorough review of SNP genotyping methods.

Materials and methods

Sixteen mammalian species were selected to represent most major mammalian clades (Murphy *et al*. 2001): sheep (*Ovis aries*), cow (*Bos taurus*), pig (*Sus scrofa*), dhole (*Cuon alpinus*), mole (*Asioscalops altaica*), mouse (*Mus musculus*), hamster (*Mesocricetus grise*), rabbit (O*rycytolagus cuni*), chimpanzee (*Pan troglodytes verus*), cat (*Felis catus*), baboon (*Papio hamadryas*), bat (*Plecotus auritus*) marmoset (*Saguinus oedipus*), armadillo (*Chaetophractus villosus*), elephant (*Loxodonta africana*) and opossum (*Didelphis virginiana*). A single sample of each species was used for CATS locus screening. Samples sources are given in Smith *et al*. (2002).

The 202 CATS primers used in this study were selected from previously published research (see supplementary material), and designed typically from alignments of two mammalian species (most often human and mouse) to amplify from exons across an intervening intron.

Initial polymerase chain reaction (PCR) conditions were the same for all primers. Reactions were carried out in the presence of 1 × Roche PCR buffer [10 mM Tris-HCl; 50 mM KCl (pH 8.3, 20 °C)], 200 µM each dNTP, 0.2 µM of each primer, 0.5 U Roche *Taq* polymerase, in a final volume of 15 µL. Both 1.5 and 2.5 mM $MgCl_2$ concentrations were used for all loci. The temperature profile was an initial denaturation at 94 °C for 2 min, followed by 94 °C for 60 s, 64 °C to 48 °C for 30 s, decreasing at 1 °C per cycle in a touchdown-style profile, 72 °C for 30 s, then 20 additional cycles with the annealing temperature fixed at 48 °C, followed by a final extension of 72 °C for 4 min. PCR products were visualized on a 1.5%:1% (w/v) NuSeive (TekNova):Agarose (Serva) gel stained with ethidium bromide. Loci were classified (0–3) according to whether they produced (0) no product; (1) a single band; (2) two bands; or (3) three or more bands. Loci producing a single band or a double band that could be optimized potentially to one band were considered 'usable' for direct sequencing.

For each species, at least 10 loci with product scores of one were randomly chosen and re-amplified for sequencing. Our goal was to obtain approximately 10 high-quality sequences, so for most species we continued to amplify loci until we obtained products suitable for sequencing approximately 10. Re-amplifications followed the above conditions and used the $MgCl_2$ concentration deemed optimal from the initial PCRs. PCRs were carried out in triplicate and products pooled prior to purification for sequencing. This was employed to minimize the chance of complete reaction failure and also to exclude nonspecific amplification occurring in a single reaction. Amplicons were purified using the High Pure PCR Purification Kit (Roche, Mannheim).

Sequencing was carried out using the ABI big-dye terminator cycle sequencing kit 2.0, and electrophoresed on an ABI 3700 automated sequencer (Applied Biosystems). Sequences were visualized using Bioedit (V. 5.0.6; Hall 1999). In order to assess homology with the target gene, sequences were compared to public database sequences using BLAST (Altschul *et al*. 1990) (http://www.ncbi.nlm.nih.gov/BLAST/).

For SNP discovery in one species (chimpanzee, *Pan troglodytes*), samples were obtained from 19 captive individuals from the Primate Foundation of Arizona and DNA from all 19 individuals was pooled at equimolar concentrations. PCR amplification of the same 11 loci sequenced initially for the chimpanzee was performed on the pooled DNA and also on DNA from six unrelated individuals (which also contributed to the pooled sample). These were carried out in triplicate under the same reaction conditions as above. For SNP detection, sequences were aligned and analysed using PHRAP and PHRED, and visualized using CONSED (Gordon *et al*. 1998). Forward and reverse sequences were analysed separately so that independent identification of potential SNPs in both directions could be assessed. Potential SNPs were identified as either nucleotide differences between aligned individuals at the same site or the occurrence of double peaks at a site in at least one individual or in the pool sample. Heterozygous SNPs were confirmed by identification of the polymorphic site in both sequence directions, or by conservative analysis of the sequence in only one direction if the other direction was not available. If the latter, we required that the available sequence displayed the SNP clearly and convincingly in order to minimize the chance of false SNP identification from a sequencing artifact.

The ABI PRISM® SNaPshot™ Multiplex System (Applied Biosystems) was used to genotype each SNP. PCRs and product purification were carried out according to the kit instructions, using new primers flanking the potential SNP and producing a fragment of approximately 100 bp. A third primer was designed abutting the 5′ end of the SNP. Using the kit, a primer extension reaction was carried out extending this third primer by the one (variable) nucleotide of the SNP using fluorophore-labelled ddNTPs (fam, tet, rox, tamra). The fragments were electrophoresed on an

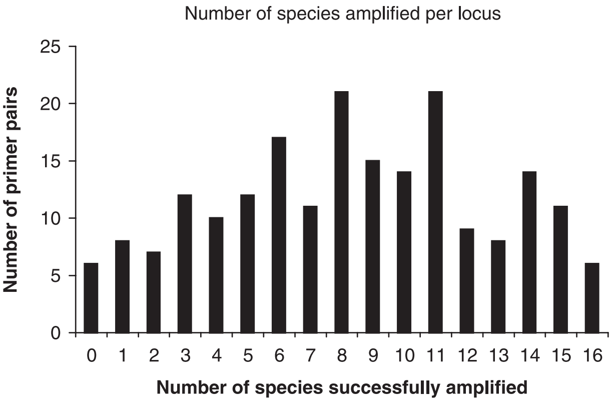

Fig. 1 Distribution of the total number of species amplified for each locus, summarized by number of loci amplifying any given number of species (0–16).

Applied BioSystems 310 Sequencer, producing one or two peaks varying in fluorescence depending on the incorporated fluorescent colour of the ddNTP at the SNP site.

Results

Amplifications

CATS primer pairs varied greatly in their ability to amplify products from multiple species. More than 50% of the primer pairs amplified putative homologues in at least half of the species screened. Six primer pairs produced 'usable' products in all mammals (defined as products receiving a score of 1 or 2 by the above scheme) and six pairs failed to produce amplification products in any species. Over all primer pairs, the median number of species yielding 'usable' amplifications per locus was eight. The distribution of amplification success (Fig. 1) can be seen to approximate normal with the bulk of the primers effective on an intermediate number of species.

Amplification success varied greatly across the species tested. In any one species, an average of 52% of the primer pairs can be expected to produce 'usable' product. Primers were most successful in the primates (65% in the chimpanzee and marmoset) and the mouse (62%) and success decreased with increasing phylogenetic distance from these species (25% in the opossum) (Fig. 2A).

To determine the 'best' 96 loci (convenient for easy handling in the lab), we ranked the loci by number of species each primer pair successfully amplified, where the number of species with score = 1 ≥ 5; score = 1 + 2 ≥ 6; score = 3 ≤ 5 (see supplemental material). Repeating the analysis using only these 'best' 96 loci resulted in a marked increase in the success rate of primers for each species (Fig. 2B); the average success of primers was 71% [range 30% (opossum)–89% (baboon)], with success decreasing with phylogenetic distance from human and mouse and consistent success with approximately half the species followed by a rapid drop-off.

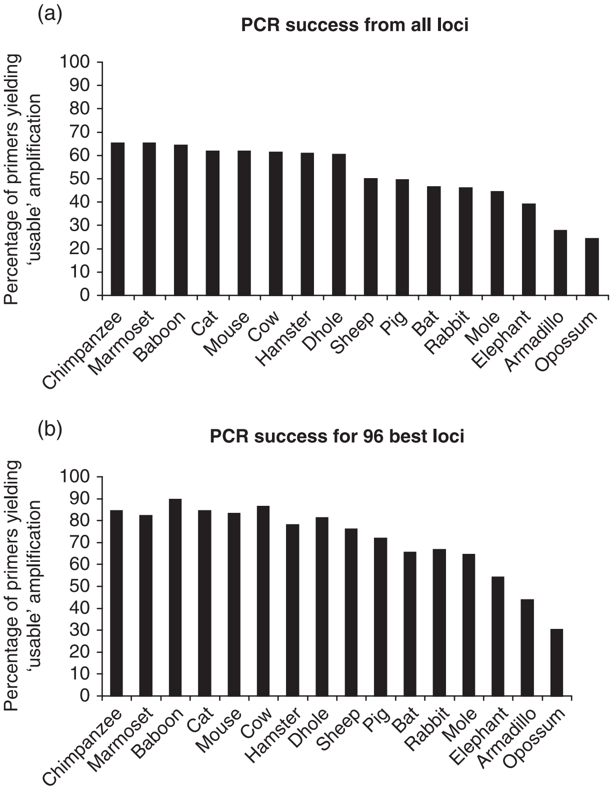

Fig. 2 Number of loci successfully amplified for each species, using (A) the complete set of 202 primer pairs and (B) the subset of the 96 'best' loci (those that produced PCR products from the most species). 'Usable' loci are defined as those yielding one or two PCR products that could potentially be used directly for sequencing, or optimized to yield products that could be sequenced.

Sequencing

All opossum amplicons were faint and deemed unsequencable because of the presence of secondary (or more) product(s) that would result in poor sequence. Therefore, subsequent analysis involves only the remaining 15 species.

From the randomly selected subset of primers that produced one band, an average of 17.3 (range 11–23) loci was required to produce adequate product for sequencing 8–13 loci per species. The more distantly related species required amplification of a larger number of loci to obtain roughly 10 bands (amplification products) that were adequate to proceed with sequencing. The average success was 61% with a range of 35% (bat) to 100% (chimpanzee) (Fig. 3). We did not sequence products, or obtained poor or no sequence, when the products were faint or missing from the agarose gels.

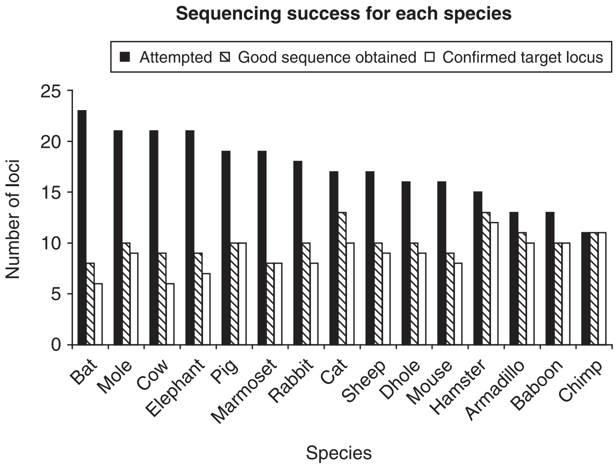

Fig. 3 Sequencing success for attempted sequencing of randomly selected PCR products for each species. Sequences were compared to the public databases (GenBank) using a BLAST search to determine if they matched closely the gene sequences originally used to design the CATS primers.

The confirmation of target gene amplification was high, ranging from 67% (cow) to 100% (chimp, baboon, marmoset, pig) (mean = 88%) of sequenced products determined to be the target gene based on the closest BLAST match (Fig. 3). This is likely to be an underestimation of actual matches, because there were a number of sequences that aligned with unlabelled GenBank submissions.

SNP detection

A larger number of individual sequences is required for SNP identification, as most SNPs occur at less than 50% frequency (although > 50% of SNPs have minor allele frequencies of at least 0.2; Marth *et al*. 2001). In order to evaluate this approach to SNP detection we limited our study to a single species, the chimpanzee, for which a large number of samples were available to us.

Twenty-six potential SNPs were identified from the sequence alignments of six of the 11 loci (no SNPs were detected in five of the loci). Potential SNPs could be classified into one of three categories: (i) The SNP was visible in both directions: eight (only two of which were independently identified in both directions; the remaining were identified in one direction, and confirmed upon close inspection of the opposite direction sequence); (ii) the SNP appeared in only one direction, in high quality sequence, and the reverse sequence was not available (e.g. when the forward and reverse sequences from long fragments did not overlap): four; and (iii) the SNP was visible in only one direction but either (a) the reverse direction contradicted it, (b) the reverse direction did not produce scorable or clean sequence or (c) the reverse direction was not available and sequence quality was not high: 14. In summary, 12 SNPs across six loci fell into categories 1 or 2, and 14 other possible SNPS were not verified by further sequencing or locus optimization. Of the 12 SNPs identified positively, seven were transitions and five were transversions.

Assay development and SNP validation

We developed SNaPshot assays for 19 SNP loci in order to confirm the presence of SNPs and assess the genotyping method. Five were developed from the set of 12 SNPs identified here, selected to represent independent loci (one SNP per locus) from which there was sufficient high-quality sequence for primer design (loci: TF1, MPO, FES, PFKM, IFNB1); a further four were designed from SNPs identified from chimpanzee sequences identified previously from CATS loci in our laboratory (FOS, IGH, CAT, MYH6) and 10 were designed from published sequence polymorphisms in chimpanzees (ApoB140, ApoB476, PSUY397, PSUY, Y85, sY19a, sY19b, SMCY, sY67, sY123, Deinard 1997; Stone *et al*. 2002). All assays were optimized and tested on a subset of chimpanzees, including several previously genotyped (sequenced) individuals (Smith *et al*. 2004). Two assays failed to amplify consistently or produce interpretable genotypes, and were dropped from further analysis (FES, PFKM). The remaining 17 assays were genotyped on a population of chimpanzees, and genotypes of the previously sequenced individuals were used to confirm the SNP genotype-calling method when possible. One assay (IFNB1) produced a spurious genotype product that precluded genotyping, and one assay (MPO) appeared to be monomorphic; this was determined to be an assay artefact rather than a sequencing artefact. Two loci (Y85, PSUY) failed for some individuals, presumably because of poor PCR. These are technical problems for assay development that can probably be overcome by redesign of the PCR primers, use of the opposite strand for the single-base extension primer and/or further reaction optimization. In the end, 14 of 19 assays appeared to genotype true polymorphic SNPs and provided a clean genotype for unknown and, when available, known (previously sequenced) individuals (Smith *et al*. 2004.). One locus (sY67) was monomorphic in our sample set, but has been shown to be polymorphic previously in a different sample set (Stone *et al*. 2002). GenBank Accession nos for the sequences generated in this study are: AF245195, AY528405, AY528406, AY528407, AF245196, AY528408, AY528409, AF244809, AF440146, AF440120, AF440119, AF440162, AF244810, AF440132 and AF440150.

Discussion

The approach we have described makes use of currently available conserved primers that amplify mammalian genes. Aside from the practical reasons of providing readily available primers for SNP discovery in a variety of species, the use of CATS loci allows identification of SNPs in genes of known function (and known genomic location in some species), so that some genomic information is associated with the loci even without prior genomic characterization of the target species.

Because ascertainment bias of SNPs can significantly affect the inference of historical demographic parameters (Wakeley *et al*. 2001), SNP discovery requires that SNPs be selected independent of frequency in a sample population representative of the study population. Low information content of some of these SNPs may require that additional SNPs be included in the study to obtain sufficient statistical power. Given that some ascertainment bias may still exist, a consistent set of rules for ascertainment and detailed records of the methods will facilitate analytical bias correction (Wakeley *et al*. 2001; Brumfield *et al*. 2003; Morin *et al*. 2004). In this study we have demonstrated that one may expect to amplify approximately 52% of the CATS loci employed here in mammalian species [range 24.26 (opossum) to 65.35 (chimpanzee and marmoset)] by our 'usable' criteria, or 61% (range 30.69 (opossum) to 73.76 (mouse)] for any amplification at all (i.e. scores of 1, 2 or 3). Sequencing and SNP discovery can be expected to yield on average one SNP per 400 bp if the variability is similar to chimpanzee or other previously screened mammalian species (Brouillette *et al*. 2000; Sachidanandam *et al*. 2001; Shubitowski *et al*. 2001). A subset of 'best' loci is expected to produce an average of 71% amplification success in placental mammals. Information on primer sequences and sources and amplification scores in the 16 species is provided in supplementary appendices.

If the target number of independently segregating SNPs is, for example, 50, then we would expect to be able to reach this number by screening between 140 (for chimpanzee; Fig. 4) and 755 loci (for bat) randomly from the whole set (assuming a set with the same characteristics and enough loci was available), or 109 (for chimpanzee) and 535 loci (for bat) from a set of loci with the characteristics of our 'best' loci set. For an 'average' mammal, with an overall ascertainment rate of 16.4%, we would expect to need to screen 306 loci (Fig. 4). We have selected only 202 loci from a more extensive and constantly growing sets of published loci, but further identification of CATS-like primers is clearly needed to increase the number of available loci. Additionally, design of primers to conserved regions of additional mammalian or other vertebrate species, and also to other types of genome regions (such as 3′ UTRs), might improve relative success rates for amplification and subsequent SNP discovery. Focused development of loci using the same principles from publicly available sequences of organisms related more closely to the target organisms might also increase the per locus success rate.

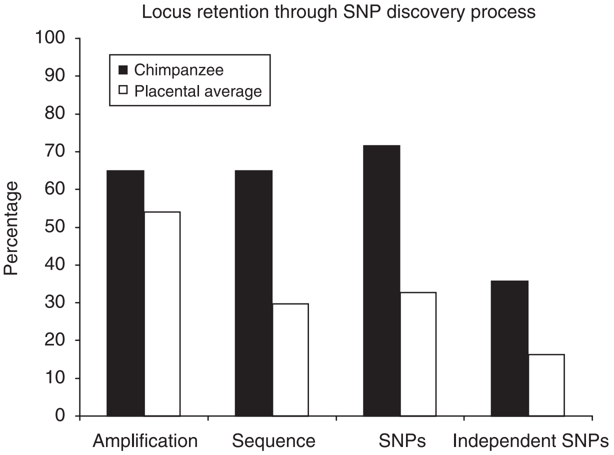

Fig. 4 Percentage of loci that produced usable PCR products and high quality sequences identified as the target gene are shown for chimpanzees and the average for the 15 placental mammals. The percentage of SNPs per locus and independent SNPs per locus sequenced (counting only one SNP in a locus) for chimpanzees were used to infer the SNP ascertainment rates for other placental animals, for which no data were available.

One of the hurdles to developing SNPs *de novo* for nonmodel organisms has been the perception that SNP discovery, assay development and genotyping would be considerably more expensive than for microsatellites. Table 1

Table 1 Estimates of time and current materials costs for development of (A) SNP loci and (B) microsatellite loci for population studies. SNP development is based on work performed in our laboratory, using the methods described in this study, assuming that about 150 CATS locus primer sets will yield approximately 80 independent SNPs (e.g. from chimpanzees). The labour and materials cost estimates are based on commercial development of an enriched genomic library (Genetic Identification Services; http://www.genetic-id-services.com), and conservative estimates of subsequent screening of clones, sequencing, assay design and assay optimization in the researcher's laboratory

(A) SNP detection, assay design, optimization			(B) Microsatellite detection, assay design, optimization		
Process description	Time (h)	Material exp.	Process description	Time (h)	Material exp.
Screen 150 nuclear loci	86	$7200	Sequence 50 clones	8	$10 300
SNP detection, assay design, optimization (for 80 loci)	473	$8800	Assay design and optimization (30 loci)	148	$2780
Total (months)	559 (3.5)	$16 000	Total (months)	156 (1)	$13 080

summarizes our estimated costs for these processes. There are trade-offs, such as the probable need to develop new SNPs for each species, whereas many microsatellites can be transferred among closely related species (Morin *et al*. 1998). With the targeted locus approach described here, however, the initial investment in primers can be spread over many species, so the relative cost of developing markers for several species, with equivalent statistical power (i.e. 50–80 SNPs vs. 10–15 microsatellites) may be substantially lower. The development of high-throughput genotyping methods with highly multiplexed PCR and/or genotyping assays will also result in lower costs per multilocus genotype, with potential savings in both disposables and labour (Chen *et al*. 2000; Syvänen 2001; Taylor *et al*. 2001).

In conclusion, we believe that there are many good reasons to employ SNPs for evolutionary, population and conservation studies (Vignal *et al*. 2002; Brumfield *et al*. 2003; Morin *et al*. 2004) and have demonstrated that, for most mammals, the targeted locus approach might provide an efficient and cost-effective method of discovering SNPs. For other species, this approach may become feasible as more genomic information becomes available, or random DNA fragment ascertainment approaches can be employed.

Acknowledgements

We are grateful to people who donated DNA samples and unpublished primers: Arati Iyengar; William Murphy, Laboratory of Genomic Diversity, National Cancer Institute; Jeff Rogers, Southwest Foundation for Biomedical Research; Jo Fritz, Primate Foundation of Arizona; Jean Wickings, Centre International de Recherches Medicales Franceville; Tamsyn Burland, School of Biological Sciences, Queen Mary, University of London; Patrick O'Grady and Martine Zilversmit, American Museum of Natural History. We are also thankful to the Max Planck Institute for Evolutionary Anthropology for support, Amanda Bowmann and Nicole Hedrick for technical support and Andy Dizon and Tasha Belfiore for comments on the manuscript.

Supplementary material

The following material is available from
http://www.blackwellpublishing.com/products/journals/suppmat/MEC/MEC2159/MEC2159sm.htm

Table S1. CATS Loci Amplification Summary

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *Journal of Molecular Biology*, **215**, 403–410.

Bagley MJ, Gall GA (1998) Mitochondrial and nuclear DNA sequence variability among populations of rainbow trout (*Oncorhynchus mykiss*). *Molecular Ecology*, **7**, 945–961.

Balloux F, Lugon-Moulin N (2002) The estimation of population differentiation with microsatellite markers. *Molecular Ecology*, **11**, 155–165.

Bensasson D, Zhang DX, Hartl DL, Hewitt GM (2001) Mitochondrial pseudogenes: evolution's misplaced witnesses. *Trends in Ecology and Evolution*, **16**, 314–321.

Bensch S, Akesson S, Irwin DE (2002) The use of AFLP to find an informative SNP: genetic differences across a migratory divide in willow warblers. *Molecular Ecology*, **11**, 2359–2366.

Bierne N, Lehnert SA, Bedier E, Bonhomme F, Moore SS (2000) Screening for intron length polymorphisms in penaeid shrimps using exon-primered intron-crossing (EPIC)-PCR. *Molecular Ecology*, **9**, 233–235.

Brouillette JA, Andrew JR, Venta PJ (2000) Estimate of nucleotide diversity in dogs with a pool-and-sequence method. *Mammalian Genome*, **11**, 1079–1086.

Brumfield RT, Beerli P, Nickerson DA, Edwards SV (2003) Single nucleotide polymorphisms (SNPs) as markers in phylogeography. *Trends in Ecology and Evolution*, **18**, 249–256.

Callen DF, Thompson AD, Shen Y *et al*. (1993) Incidence and origin of 'null' alleles in the $(AC)_n$ microsatellite markers. *American Journal of Human Genetics*, **52**, 922–927.

Chakraborty R, Stivers DN, Su B, Zhong Y, Budowle B (1999) The utility of short tandem repeat loci beyond human identification: implications for development of new DNA typing systems. *Electrophoresis*, **20**, 1682–1696.

Chen J, Iannone MA, Li MS *et al*. (2000) A microsphere-based assay for multiplexed single nucleotide polymorphism analysis using single base chain extension. *Genome Research*, **10**, 549–557.

Crandall KA, Bininda-Emonds ORP, Mace GM, Wayne RK (2000) Considering evolutionary processes in conservation biology. *Trends in Ecology and Evolution*, **15**, 290–295.

DeBry RW, Seshadri S (2001) Nuclear intron sequences for phylogenetics of closely related mammals: an example using phylogeny of *Mus*. *Journal of Mammalogy*, **82**, 280–288.

Deinard AS (1997) *The evolutionary genetics of the chimpanzees*. PhD Thesis, Yale University.

Ebersberger I, Metzler D, Schwarz C, Paabo S (2002) Genomewide comparison of DNA sequences between humans and chimpanzees. *American Journal of Human Genetics*, **70**, 1490–1497.

Edwards SV, Beerli P (2000) Perspective: gene divergence, population divergence, and the variance in coalescence time in phylogeographic studies. *Evolution*, **54**, 1839–1854.

Excoffier L, Yang Z (1999) Substitution rate variation among sites in mitochondrial hypervariable region I of humans and chimpanzees. *Molecular Biology and Evolution*, **16**, 1357–1368.

Fries R, Durstewitz G (2001) Digital DNA signatures for animal tagging. *Nature Biotechnology*, **19**, 508.

Friesen VL, Congdon BC, Kidd MG, Birt TP (1999) Polymerase chain reaction (PCR) primers for the amplification of five nuclear introns in vertebrates. *Molecular Ecology*, **8**, 2147–2149.

Friesen VL, Congdon BC, Walsh HE, Birt TP (1997) Intron variation in marbled murrelets detected using analyses of single-stranded conformational polymorphisms. *Molecular Ecology*, **6**, 1047–1058.

Gagneux P, Boesch C, Woodruff DS (1997) Microsatellite scoring errors associated with noninvasive genotyping based on nuclear DNA amplified from shed hair. *Molecular Ecology*, **6**, 861–868.

Glaubitz JC, Rhodes OE, Dewoody JA (2003) Prospects for inferring pairwise relationships with single nucleotide polymorphisms. *Molecular Ecology*, **12**, 1039–1047.

Gordon D, Abajian C, Green P (1998) Consed: a graphical tool for sequence finishing. *Genome Research*, **8**, 195–202.

Hau TA (1999) BioEdit: a user-friendly biological sequence alignment editor and analysis program for windows 95/98/NT. *Nucleic Acids Symposium Series*, **41**, 95–98.

Hare MP (2001) Prospects for a nuclear gene phylogeography. *Trends in Ecology and Evolution*, **16**, 700–706.

Harpending HC, Batzer MA, Gurven M *et al.* (1998) Genetic traces of ancient demography. *Proceedings of the National Academy of Sciences USA*, **95**, 1961–1967.

Karl SA, Avise JC (1993) PCR-based assays of Mendelian polymorphisms from anonymous single-copy nuclear DNA: techniques and applications for population genetics. *Molecular Biology and Evolution*, **10**, 342–361.

Krawczak M (1999) Informativity assessment for biallelic single nucleotide polymorphisms. *Electrophoresis*, **20**, 1676–1681.

Kuhner MK, Beerli P, Yamato J, Felsenstein J (2000) Usefulness of single nucleotide polymorphism data for estimating population parameters. *Genetics*, **156**, 439–447.

Lessa E (1992) Rapid surveying of DNA sequence variation in natural populations. *Molecular Biology and Evolution*, **9**, 323–330.

Luikart G, England PR (1999) Statistical analysis of microsatellite DNA data. *Trends in Ecology and Evolution*, **14**, 253–256.

Lyons LA, Laughlin TF, Copeland NG *et al.* (1997) Comparative anchor tagged sequences (CATS) for integrative mapping of mammalian genomes. *Nature Genetics*, **15**, 47–56.

Marth G, Yeh R, Minton M *et al.* (2001) Single-nucleotide polymorphisms in the public domain: how useful are they? *Nature Genetics*, **27**, 371–372.

McKay JK, Latta RG (2002) Adaptive population divergence: markers, QTL and traits. *Trends in Ecology and Evolution*, **17**, 285–291.

McLenachan PA, Stockler K, Winkworth RC *et al.* (2000) Markers derived from amplified fragment length polymorphism gels for plant ecology and evolution studies. *Molecular Ecology*, **9**, 1899–1903.

Morin PA, Chambers KE, Boesch C, Vigilant L (2001) Quantitative polymerase chain reaction analysis of DNA from noninvasive samples for accurate microsatellite genotyping of wild chimpanzees (*Pan troglodytes verus*). *Molecular Ecology*, **10**, 1835–1844.

Morin PA, Luikart G, Wayne RK, SNP Workshop Group (2004) Applications of SNPs in ecology, evolution, and conservation. *Trends in Ecology and Evolution*, in press.

Morin PA, Mahboubi P, Wedel S, Rogers J (1998) Rapid screening and comparison of human microsatellite markers in baboons: allele size is conserved, but allele number is not. *Genomics*, **53**, 12–20.

Murphy WJ, Eizirik E, O'Brien SJ *et al.* (2001) Resolution of the early placental mammal radiation using Bayesian phylogenetics. *Science*, **294**, 2348–2351.

Nachman MW, Crowell SL (2000) Estimate of the mutation rate per nucleotide in humans. *Genetics*, **156**, 297–304.

Navidi W, Arnheim N, Waterman MS (1992) A multiple-tubes approach for accurate genotyping of very small DNA samples by using PCR: statistical considerations. *American Journal of Human Genetics*, **50**, 347–359.

Nickerson DA, Tobe VO, Taylor SL (1997) PolyPhred: automating the detection and genotyping of single nucleotide substitutions using fluorescence-based resequencing. *Nucleic Acids Research*, **25**, 2745–2751.

Nicod JC, Largiader CR (2003) SNPs by AFLP (SBA): a rapid SNP isolation strategy for non-model organisms. *Nucleic Acids Research*, **31**, e19.

Nielsen R (2000) Estimation of population parameters and recombination rates from single nucleotide polymorphisms. *Genetics*, **154**, 931–942.

Oleykowski CA, Bronson Mullins CR, Godwin AK, Yeung AT (1998) Mutation detection using a novel plant endonuclease. *Nucleic Acids Research*, **26**, 4597–4602.

Palumbi SR, Baker CS (1994) Contrasting population structure from nuclear intron sequences and mtDNA of humpback whales. *Molecular Biology and Evolution*, **11**, 426–435.

Pluzhnikov A, Donnelly P (1996) Optimal sequencing strategies for surveying molecular genetic diversity. *Genetics*, **144**, 1247–1262.

Primmer CR, Borge T, Lindell J, Saetre GP (2002) Single-nucleotide polymorphism characterization in species with limited available sequence information: high nucleotide diversity revealed in the avian genome. *Molecular Ecology*, **11**, 603–612.

Purugganan M, Gibson G (2003) Merging ecology, molecular evolution, and functional genetics. *Molecular Ecology*, **12**, 1109–1112.

B-Rao, C (2001) Sample size considerations in genetic polymorphism studies. *Human Heredity*, **52**, 191–200.

Reed DH, Frankham R (2001) How closely correlated are molecular and quantitative measures of genetic variation? A meta-analysis. *Evolution*, **55**, 1095–1103.

Rosenbaum HC, Deinard AS (1998) Caution before claim: an overview of microsatellite analysis in ecology and evolutionary biology. In: *Molecular Approaches to Ecology and Evolution* (eds DeSalle R, Schierwater B), pp. 87–106. Birkhäuser, Boston.

Sachidanandam R, Weissman D, Schmidt SC *et al.* (2001) A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature*, **409**, 928–933.

Schlötterer C, Pemberton J (1998) The use of microsatellites for genetic analysis of natural populations — a critical review. In: *Molecular Approaches to Ecology and Evolution* (eds DeSalle R, Schierwater B), pp. 71–86. Birkhäuser-Verlag, Berlin.

Shubitowski DM, Venta PJ, Douglass CL, Zhou RX, Ewart SL (2001) Polymorphism identification within 50 equine gene-specific sequence tagged sites. *Animal Genetics*, **32**, 78–88.

Silva JC, Kondrashov AS (2002) Patterns in spontaneous mutation revealed by human–baboon sequence comparison. *Trends in Genetics*, **18**, 544–547.

Smith S, Aitken N, Morin PA (2004) Characterization of 15 SNP markers for chimpanzees (*Pan troglodytes*). *Molecular Ecology Notes*, in press.

Smith S, Vigilant L, Morin PA (2002) The effects of sequence length and oligonucleotide mismatches on 5′ exonuclease assay efficiency. *Nucleic Acids Research*, **30**, e111.

Stone AC, Griffiths RC, Zegura SL, Hammer MF (2002) High levels of Y-chromosome nucleotide diversity in the genus Pan. *Proceedings of the National Academy of Sciences USA*, **99**, 43–48.

Strand AE, Leebens-Mack J, Milligan BG (1997) Nuclear DNA-based markers for plant evolutionary biology. *Molecular Ecology*, **6**, 113–118.

Sunnucks P (2000) Efficient genetic markers for population biology. *Trends in Ecology and Evolution*, **15**, 199–206.

Syvänen AC (2001) Accessing genetic variation: genotyping single nucleotide polymorphisms. *Nature Reviews Genetics*, **2**, 930–942.

Taberlet P, Griffin S, Goossens B *et al.* (1996) Reliable genotyping of samples with very low DNA quantities using PCR. *Nucleic Acids Research*, **24**, 3189–3194.

Taylor JD, Briley D, Nguyen Q *et al.* (2001) Flow cytometric platform for high-throughput single nucleotide polymorphism analysis. *Biotechniques*, **30**, 661–666, 668–669.

van Tienderen PH, de Haan AA, van der Linden CG, Vosman B (2002) Biodiversity assessment using markers for ecologically important traits. *Trends in Ecology and Evolution*, **17**, 577–582.

Venta PJ, Brouillette JA, Yuzbasiyan-Gurkan V, Brewer GJ (1996) Gene-specific universal mammalian sequence-tagged sites: application to the canine genome. *Biochemistry and Genetics*, **34**, 321–341.

Vignal A, Milan D, SanCristobal M, Eggen A (2002) A review on SNP and other types of molecular markers and their use in animal genetics. *Genetics Selection Evolution*, **34**, 275–305.

Wakeley J, Nielsen R, Liu-Cordero SN, Ardlie K (2001) The discovery of single-nucleotide polymorphisms — and inferences about human demographic history. *American Journal of Human Genetics*, **69**, 1332–1347.

Wolford JK, Blunt D, Ballecer C, Prochazka M (2000) High-throughput SNP detection by using DNA pooling and denaturing high performance liquid chromatography (DHPLC). *Human Genetics*, **107**, 483–487.

Nicola Aitken, Steven Smith and Phillip Morin were members of the Laboratory for Conservation Genetics, which developed and applied genetics technologies for conservation until its closure in 2002. Carsten Schwarz screened CATS loci as part of a project with Phillip Morin in the Department of Evolutionary Genetics at the Max Planck Institute for Evolutionary Anthropology.