Review TRENDS in Ecology and Evolution Vol.19 No.4 April 2004

SNPs in ecology, evolution and conservation

Phillip A. Morin[1,4], Gordon Luikart[2], Robert K. Wayne[3] and the SNP workshop group*

[1]Laboratory for Conservation Genetics, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany
[2]Laboratoire d'Ecologie Alpine, Génomique des Populations et Biodiversité, Université J.-Fourier, Grenoble, France
[3]Department of Organismic Biology, Ecology & Evolution, University of California, Los Angeles, CA 90095, USA
[4]Southwest Fisheries Science Center, 8604 La Jolla Shores Drive, La Jolla, CA 92037, USA

Over the past two decades, new molecular genetic techniques have had substantial impacts on the fields of ecology, evolution and conservation. However, our current toolbox of genetic methodologies remains inadequate for answering many questions and there are significant technological and analytical limitations. We review the possible uses of single nucleotide polymorphisms (SNPs) as novel genetic markers for common questions in population genetics. Furthermore, we evaluate the potential of SNPs relative to frequently used genetic markers, such as microsatellite loci and mitochondrial DNA (mtDNA) sequences, and we discuss statistical power, analytical approaches, and technological improvements and limitations. Although ascertainment bias is a problem for some applications, SNPs can often generate equivalent statistical power whilst providing broader genome coverage and higher quality data than can either microsatellites or mtDNA, suggesting that SNPs could become an efficient and cost-effective genetic tool.

Since the early 1990s, nuclear MICROSATELLITE (see Glossary) loci and mitochondrial DNA (mtDNA) sequences have been the tools of choice in molecular studies in ecology and evolution (Figure 1). Both kinds of genetic marker represent rapidly evolving DNA sequences that are informative for answering population-level questions. However, the high information content, a result of high mutation rates, comes at a price. HOMOPLASY poses severe limitations on subsequent data analysis and, thus, the biological meaning and usefulness of the results [1]. Inferences drawn from mtDNA sequences are further limited by the fact that the mtDNA genome comprises a single maternally inherited locus. Microsatellite loci suffer from null alleles and mutation patterns that are variable, introducing ambiguity to data analyses. The loci can also be sparse in the genome and, thus, difficult to find in some species (e.g. mites [2]). By contrast, mutations observed as SINGLE NUCLEOTIDE POLYMORPHISMS (SNPs) are abundant and widespread in many species' genomes (coding and non-coding regions), and they evolve in a manner well described by simple mutation models, such as the infinite sites model [3] (Figure 1).

Here, we discuss how recently developed technology greatly facilitates SNP discovery, genotyping and analysis, with respect to reduced cost and effort, thus enabling the extraction of comparatively more or better quality information than is possible from currently employed marker systems. We then review the potential uses of SNP genotypes for common applications in population genetics. We also briefly touch on the promising uses of SNPs for detecting selection and molecular adaptation. The present work complements a recent review addressing the application of SNPs to phylogeography [4], and a detailed review of the use of SNPs in animal genetics [3].

Many technological and analytical problems still remain. In particular, ASCERTAINMENT BIAS [5,6] is a crucial issue in many SNP applications and needs to be addressed by technological improvements and the development of new analytical methods. In addition, to assess ecologically important traits using SNPs, a better understanding of which genes or gene families are important to individual fitness is necessary [7].

SNP discovery, genotyping, and information technology

There are two principle steps to the use of SNP markers: locus discovery (ascertainment) and genotyping. Although

Glossary

Ascertainment bias: the selection of loci from an unrepresentative sample of individuals, or using a particular method, which yields loci that are not representative of the spectrum of allele frequencies in a population.
CATS loci: comparative anchor tagged sites loci; PCR primers designed from conserved gene regions, based on alignment of the same gene sequences from different species. The PCR product usually contains a less conserved region that can be screened for variation.
Haplotype: single chromosomal (haplotypic) DNA sequence component in a diploid (having two chromosomal sets) individual.
Homoplasy: the parallel evolution of identical character states.
Microsatellite: short tandem repeat sequence, usually comprising variable numbers of repeats of 2–5 nucleotides (e.g. CA). Different numbers of repeats result in different lengths of alleles.
Single nucleotide polymorphism (SNP): nucleotide site in a DNA sequence where more than one nucleotide (G, A, T or C) is present in the population.
Single strand conformation polymorphism (SSCP): a common method of detecting differences in DNA sequences based on the electrophoretic migration behavior of single stranded DNA.

Corresponding author: Phillip A. Morin (Phillip.Morin@noaa.gov).
* Invited participants in the workshop 'Technical and analytical methods for wildlife genetics: developments for use of Single Nucleotide Polymorphisms (SNPs)', held in Leipzig, Germany, 11–13 September, 2002: (Fred W. Allendorf, Charles F. Aquadro, Tomas Axelsson, Mark Beaumont, Karen Chambers, Gregor Durstewitz, Thomas Mitchell-Olds, Per J. Palsbøll, Hendrik Poinar, Molly Przeworski, Barbara Taylor and John Wakeley).

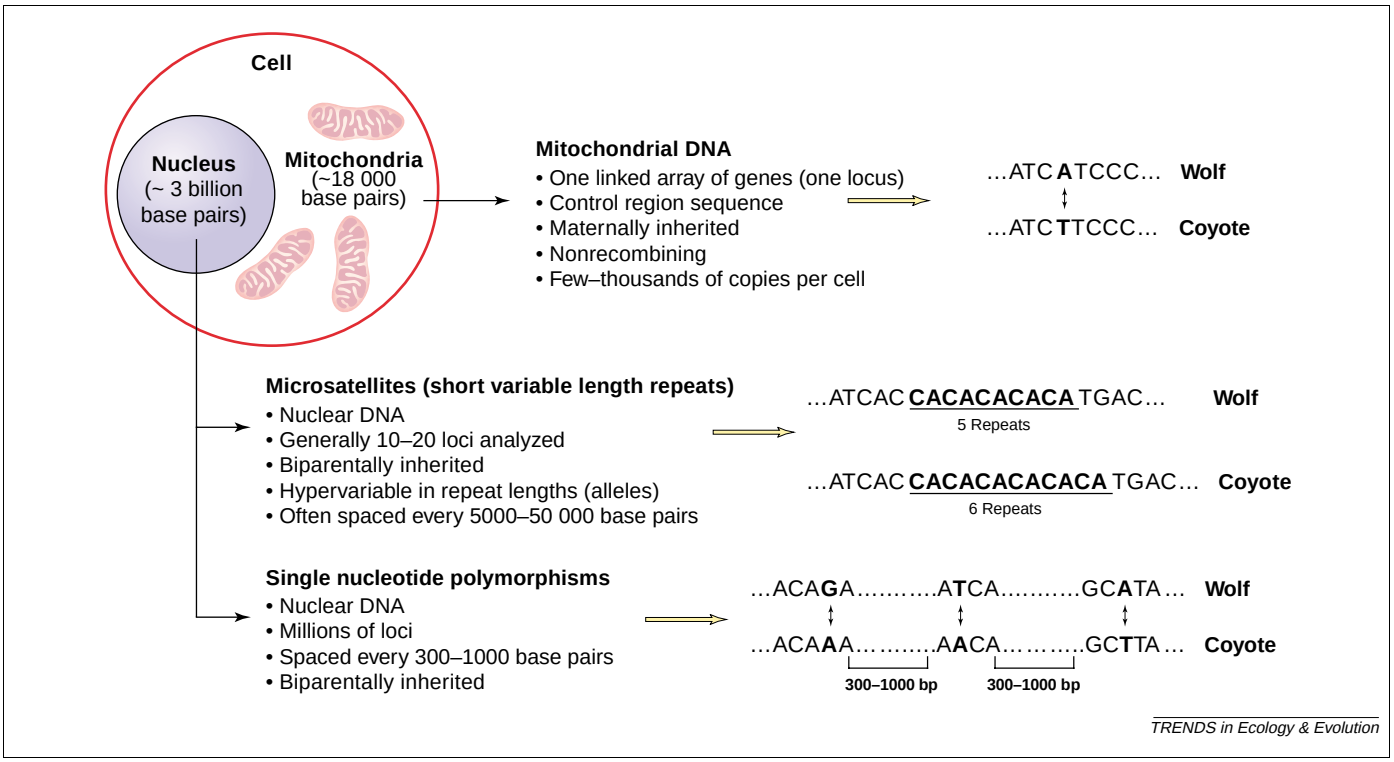

Figure 1. Comparison of the characteristics of mitochondrial DNA (mtDNA), microsatellites [63] and single nucleotide polymorphism (SNPs) [4] as genetic markers (with examples of possible DNA sequence differences between wolf-like canids; these types of differences can be found within or between taxa).

multiple technologies are available for both of these steps, the methods used must be selected based on criteria for the particular study to be performed. For population genetics of non-model organisms (i.e. those not subject to large-scale genome projects), studies often will include 50–100 loci or more and several hundred samples, and might use noninvasively obtained and degraded specimens.

Some of the available methods are more practical for non-model organisms than others (Box 1). Here, we review methods that are most feasible in academic laboratories; reviews that cover all technologies more fully are available (e.g. [8]).

Box 1. Generally applicable criteria for molecular markers and genotyping methods for population genetics studies of non-model organisms

- Must produce very high quality genotypes (e.g. for parentage analysis)
- Low ascertainment and per genotype costs
- The ability to use small amounts and/or low-quality DNA (e.g. noninvasive samples)
- Rapid ascertainment and assay development methods
- Relatively low per-assay development costs (e.g. unlabeled versus labeled oligonucleotides)
- Flexible platform for continuous or changing assay development and implementation
- Low sample handling requirements for producing genotypes
- The use of standard laboratory equipment already available, or access to specialized equipment through core facilities or private service organizations.
- An interface between a variety of genotyping platforms and a common database for markers and genotypes that enables both automated data transfer and manual input.

SNP discovery

SNP discovery is the process of finding the polymorphic sites in the genome of the species and populations of interest. In humans, much of the SNP discovery has been done *in silico*, meaning that genomic information from multiple individuals in the public databases is screened for the identification of putative polymorphisms (e.g. [9]). For most non-model organisms, SNPs have to be found through laboratory screening (e.g. sequencing) of segments of the genome from multiple individuals. The targeted gene or genomic region approach exploits the fact that there are often conserved sequences of genes or regions from multiple species (e.g. human and mouse) from which PCR primers can be designed to amplify the ORTHOLOGOUS gene regions in related species (called CATS LOCI [10,11]) (Box 2).

The number of genome segments that must be screened to discover ~50 SNPs depends on the density of SNPs across the genome. In many species, SNPs occur every 200–500 bp, suggesting that the screening of 75–100 independent genome segments of ~500–800 bp each should yield >50 independent SNPs [4,12]. If gene sequences for the taxonomic group of interest are not easily amplified, or if a more random genomic sample is desired, either a random sequence approach (Box 2), or methods based on sequencing of amplified fragment length polymorphisms (AFLPs) [13] can be used. These approaches are applicable to any organism, and produce large numbers of DNA fragments that can be sequenced readily for SNP screening with relatively little initial investment. Both methods require sequencing of multiple individuals and/or pools of samples to identify polymorphic loci (reviewed in [4]).

Box 2. SNP discovery and genotyping

SNP discovery

Targeted locus primers in the target gene have been termed 'CATS' [10] loci (Figure I). The same sets of CATS primers can be used to screen for single nucleotide polymorphisms (SNPs) in multiple related species ([11] and references therein), so that comparable genomic regions or genes can be surveyed in multiple taxa. Not all loci will work in all species, so that many loci need to be screened to find those that contain variable SNPs. The increasing availability of software and databases for comparative analysis of expressed sequence tags (ESTs) will facilitate development of primers in expressed genes [9]. Alternatively, random DNA fragments are incorporated into a genomic DNA library [67], and sequenced. Each fragment must be sequenced initially from a clone, and then primers designed for amplifying and sequencing the same fragment in multiple individuals to find polymorphic sites. SNPs will generally be in unknown genome regions, so targeted gene analysis and *a priori* knowledge of linkage between loci is not possible.

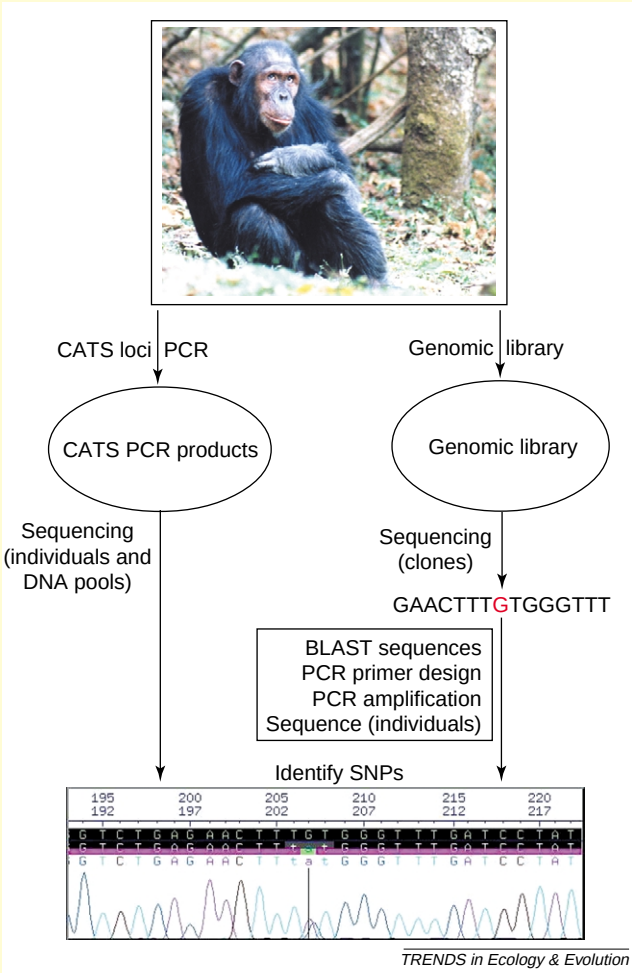

Figure I.

SNP genotyping with single base extension or allele-specific primer extension

SBE (single base extension) or ASPE (allele-specific primer extension) (Figure IIa): primer extension with only dideoxy nucleotides causes only the complementary nucleotide [shown as blue (C) or yellow (T) circles] to be added to the primer, situated just 3′ of the SNP site. Primer extension with deoxy nucleotides and highly specific polymerase enables allele-specific amplification when the 3′ terminal nucleotide of the two primers contains the SNP nucleotides. Alleles can be sorted and detected using various methods:

- Gel electrophoresis (Figure IIb): alleles sorted by size if they have different 5′ polynucleotide tails on them, and/or by color when allele-specific primers are dye labeled (green peak, G allele; black peak, A allele), or when different dye-labeled dideoxinucleotides are used (e.g. [68]). The alleles are labeled according to the template sequence rather than the complementary nucleotides added to the primers (e.g. G and A alleles, detected by incorporation of C and T dideoxynucleotides, respectively)
- Microarrays (Figure IIc): 5′ oligonucleotide 'tags' on SBE or ASPE primers are used to hybridize the extended products to complementary tags bound to spots on a microarray. Visualization can be multi-color for SBE (blue spot, G allele; yellow spot, A allele; green spot, G/A heterozygote), or single color with ASPE [69].
- Microsphere 'arrays' (Figure IId): as with arrays, SBE or ASPE primers can be tagged and hybridized to different colored microspheres with reverse complement oligonucleotide tags on them. Extension products are fluorescently labeled and the spheres (in red) are sorted by flow cytometry to identify the locus and genotypes [19]. The measured fluorescence for each allele is measured to determine the genotype; low fluorescenc for both alleles indicates a failed PCR reaction or the no template controls (NTC).
- Fluorescence polarization (FP) (Figure IIe): FP instruments measure differences in rotation rate of extended products because of different molecular composition (e.g. different incorporated nucleotides or primer tags) [70]. The plot indicates rotational differences for each allele plotted together to indicate the three possible genotypes and NTC.

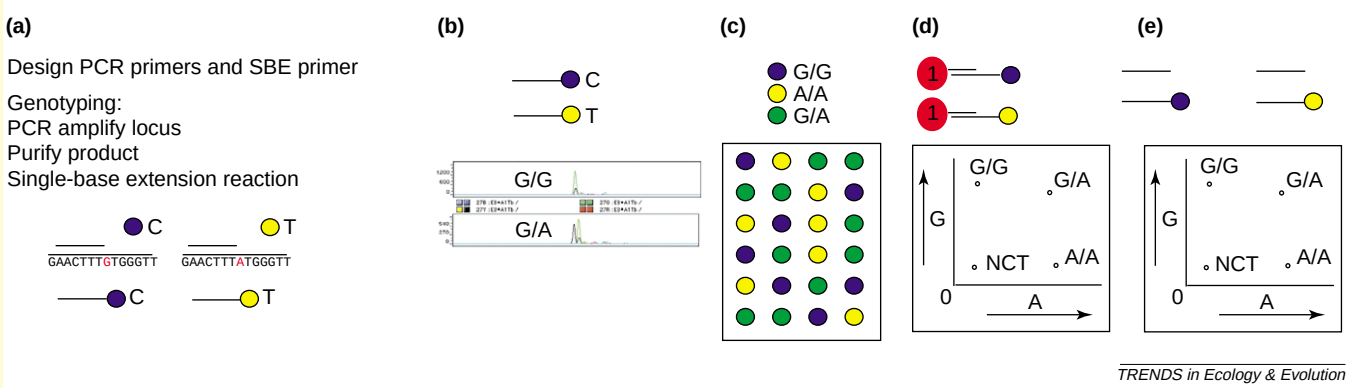

Figure II.

If a group of markers with intermediate frequency alleles is desired, selection of loci can involve screening of fewer individuals, and can use pooled DNA samples for the identification of SNPs [12,14]; in humans, >50% of discovered SNPs have minor allele frequencies of at least 0.2 [15]. Alternatively, to avoid ascertainment bias in allele frequencies, the best approach is to obtain high-quality sequence from a relatively large sample of individuals representing all of the populations in the study. Although more labor intensive, it can be facilitated by the use of alternative screening methods for SNP detection, such as SINGLE STRAND CONFORMATION POLYMORPHISM (SSCP), which enables rapid and inexpensive screening of more samples, but with less sensitivity, such that some SNPs might be missed.

SNP genotyping

The choice of method for genotyping SNP loci is dependent on many criteria. We present one set of relatively new options for generating SNP genotypes in non-model organisms that meets many of the criteria outlined in Box 1. Two other methods, PCR-RFLP and PCR-SSCP are 'old standards', enabling the detection of most SNPs using well established methods that use current standard laboratory equipment [16,17], but they are less suited to high-throughput genotyping. The newer method, primer extension [either single base extension (SBE) or allele-specific primer extension (ASPE)], meets many of the criteria, and can be performed with standard laboratory equipment, or made more efficient through the use of specialized equipment [8,18]. Both SBE and ASPE can be scaled up to enable hundreds to thousands of genotypes to be generated per day. However, all of the methods for SNP genotyping produce high-quality genotype data when proper controls are included.

Data acquisition, analysis, publication and public access

Crucial to the success of the use of population genetics as a means of generating genotype data is the use of laboratory information systems to assist in the accurate tracking of samples and data, and the development of public databases that enable presentation of genotype data in a fashion analogous to that of public sequence databases, such as GenBank. With microsatellite loci, standardization of allele sizes among laboratories is difficult (e.g. [19]), and public access and presentation of data has not yet become the standard.

With SNP data, the genotypes represent specific changes in the DNA sequence, and can be represented easily using the standard DNA code (G, A, T, C), or other coding systems if the orientation of the polymorphism is also identified [20]. The genotype codes are independent of the genotyping system, and can be standardized in public databases such that data can be compared directly among studies. SNP loci for humans and other species are currently presented in the 'dbSNP' database (http://www.ncbi.nlm.nih.gov/SNP), which enables information about the locus to be stored and searched, along with some information about allele frequencies in a sample population [21]. Unfortunately, it does not include actual SNP diploid genotype data.

Costs

Generating novel SNP loci requires a significant effort, particularly if the targeted locus approach (Box 2) is to be used for a single species, because a large investment in CATS primers is necessary (≥200 primer pairs, ~US$25 a pair). This is cost effective if several laboratories collaborate or if several or many species are screened with the same primers [11]. A potentially more cost-effective approach for individual species and random SNPs is the genomic library approach (Box 2). Both the target locus and the genomic library approach require sequencing of individuals and/or pooled samples to detect SNPs, but data indicate that SNPs are found in >50% of loci [4]. So 100 loci (each of 500–800 bp in length) can be sequenced for ≤US$6000, and 50 SBE assays can be assembled for ~US$3750. By comparison, discovery of 15 new microsatellites, using a commercially generated enriched library, would cost ~US$12 000 (http://www.genetic-id-services.com), and a new library would be required for each species unless they were closely related. SNP genotyping costs vary significantly by method, ranging from a few cents to >US$1.00 per genotype [8,18,22].

Population genetics

Here, we briefly compare the relative usefulness and limitations of SNPs relative to microsatellite loci for key applications in population genetics.

Estimating genetic variation

Precise estimation and comparison of genetic variation among populations requires a large number of SNPs relative to microsatellites because microsatellite loci typically have many alleles (~5–20), whereas two is the norm for SNP loci. This is especially true when testing for differences in variation using indices of allelic richness (e.g. [23]). A recent simulation-based study by Mariette *et al.* [24] found that four to ten times more biallelic markers were necessary compared with multi-allelic markers for reliably estimating genome-wide levels of variation. However, the study considered only dominant biallelic (AFLP) markers, which are less informative than are co-dominant biallelic markers (e.g. SNPs) and, thus, fewer SNPs than AFLPs are needed. In general, the required number of loci is difficult to assess *a priori* because each study has a different evolutionary context (e.g. [24,25]), and simulation studies are needed to further elucidate SNP numbers and characteristics for population genetics studies.

Ascertainment bias has the potential to introduce a systematic bias in estimates of variation within and among populations [26–28]. The protocol used to identify SNPs for a study must be recorded in detail, including the number and origin of individuals screened, to enable ascertainment bias to be assessed and potentially corrected [4]. However, this might add substantially to the effort and expertise required for data analysis. Using a test panel of individuals of wide geographical origin, and reporting monomorphic loci, in part, can often reduce ascertainment bias. Currently, monomorphic loci are rarely reported, especially for microsatellite-based

studies, which implies that comparisons between divergent populations or species are biased [4]. Ascertainment bias is most problematic for applications that use allele frequencies to estimate population size and demographic changes [4,6], and least important for individual identification, paternity analysis [29,30], and assignment tests (e.g. [31]) where intentional selection of high-heterozygosity markers provides greater statistical power.

Biases can also arise when transferring SNP markers across populations or between studies. For example, researchers might use only SNPs known to be the most polymorphic in one population or that are polymorphic in a small initial sample. This can upwardly bias estimates of diversity in the new study, or downwardly bias estimates if transferring SNPs from a low-diversity population to a high-diversity ancestral population [32] (Figure 2).

Individual identification, parentage and relatedness

Individual identification is an important step in the noninvasive monitoring of animal movements and abundance, in forensic applications and in behavioral studies. The power to identify individuals depends mainly on the number of independent markers and their heterozygosity, rather than on the number of alleles per locus [33]. The power of individual SNPs for individual identification is ~two to four times less than that of multi-allelic markers (Figure 3a), but the use of single-tube multiplex assays with small PCR products (<60–80 bp) could potentially produce better quality data more efficiently than would genotyping multiple microsatellites and would also provide equivalent or greater multilocus power.

The need for using more SNPs than microsatellites extends to estimation of relatedness, which requires even higher levels of precision than does individual identification. For example, paternity exclusion requires ~7–14 multi-allelic loci (expected heterozygosity (H_e)(0.60–0.80), but would require ~40–100 SNPs (H_e(0.20–0.40) to achieve a similar probability of paternity exclusion (Figure 3b). Blouin *et al.* [34] found that ~40–50 multi-allelic loci (H_e(0.60–0.75) are required to achieve a high likelihood (>0.90) of discriminating between unrelated and related individuals (e.g. half sibs). Simulation analysis by Glaubitz *et al.* [35] indicates that even 100 SNPs with a minor allele frequency of 0.20 is insufficient for distinguishing among all relationships except parent–offspring pairs.

Selection of high heterozygosity loci (approaching 0.50) provides the greatest power [29,30] for parentage analysis. Parentage and related analysis assume that loci are independent (unlinked) and that population allele frequencies are accurately estimated.

Population structure

Population structure is usually estimated from genetic distance measures such as F_{st}, R_{st} and Nei's D, and the statistical significance of the estimate assessed by permutation tests. The precision of such estimates is related to the degrees of freedom, which is one less than the number of alleles per locus (Figure 4) [36]. Theoretical work by Kalinowski [36] suggests similar levels of precision, with regard to estimating F_{st}, for either one locus with 11 alleles or ten loci each with only two alleles for a completely isolated population at equilibrium (but see [37,38]).

The advantage of using many SNP loci lies in a more representative sample of the entire genome and reduced interlocus sampling variance. Increasing the number of

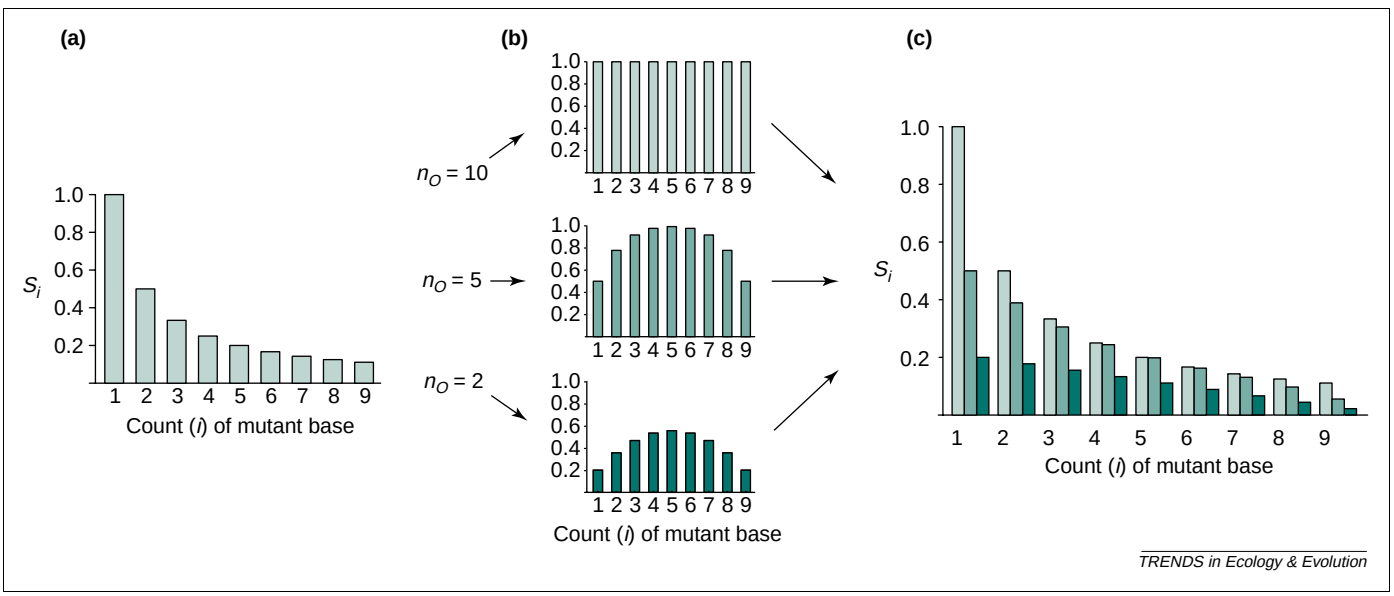

Figure 2. The ascertainment bias consequences of single nucleotide polymorphism (SNP) discovery in an initial population sample for numbers of SNPs and their estimated allele frequencies in the full population survey. It is assumed that the SNP discovery was performed with an initial sample of size n_O, using the notation of [6], and that a second survey included these samples plus enough additional samples to make the total sample size $n = 10$. **(a)** shows the expectations without an ascertainment screening step (i.e. if there was no pre-screening step, and all individuals in the population sample were sequenced). These are the expected numbers of SNPs in a sample of size ten from a population with theta = 1 (Theta = $4N_e\mu$; N_e = effective population size; μ = mutation rate per base pair per generation), divided into nine different allele frequency classes. **(b)** plots the chance that a SNP in each frequency class would be found in the initial screening for three different values of n_O. Clearly, when n_O is smaller, fewer SNPs are discovered and the ascertainment process preferentially finds high heterozygosity SNPs (i close to $n/2$). **(c)** compares the results for these three possible ascertainment schemes. When $n_O = n$ (light-green bars), no SNPs are missed and the pattern is identical to that in the leftmost panel, whereas when n_O is smaller (mid-green and dark green bars) fewer SNPs are found and there is an enrichment of SNPs with mutant base counts near the middle, relative to the unbiased case. Si is the number of SNP loci in each frequency class in the sample.

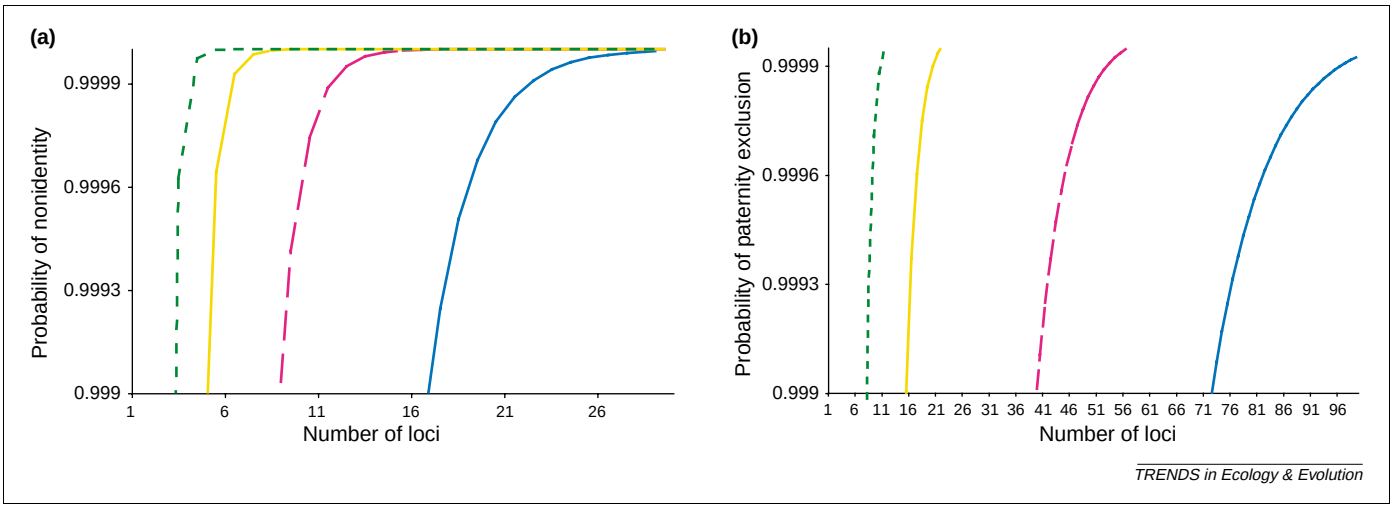

Figure 3. Statistical power versus heterozygosity for biallelic SNPs. **(a)** Relationship between the number of loci and the probability of nonidentity of genotypes (from different individuals), for each of four heterozygosity levels (H). The probability of nonidentity P(NID) is defined as the probability of randomly sampling two individuals and observing different genotypes. 10–20 SNPs (H = 0.2–0.4) are required to equal the power of four to six microsatellite loci (H = 0.6–0.8), for individual identification. For most wildlife forensics purposes, P(NID) must be ~0.999–0.9999 [64]. Lower values are appropriate for other purposes (e.g. for estimating population size). The equation used here is from [65] [Eqn 2: same allele frequencies, but here we compute P(NID) as [1 − *P*(ID)] instead of P(ID)]. **(b)** Relationship between the number of loci and the probability of paternity exclusion for each of four heterozygosity levels. There is a rapid increase in the number of loci needed as heterozygosity declines (from 0.60 to 0.40 and 0.20). For a SNP to have H > 0.32, the frequency of the rarer allele must be >0.20. The equation used here is for the general case with samples from two parents and one offspring when excluding one parent [66], with allele frequencies as in [65]. For (a) and (b): H(0.20 (blue line); 0.40 (pink line); 0.60 (yellow line); and 0.80 (dashed green line).

SNP loci also provides an opportunity to identify 'outliers' (e.g. loci under selection) [39]. By contrast, microsatellite loci are often subject to high mutation rates and, thus, homoplasy, and they can suffer from a n incomplete understanding of the underlying mutation model, yielding unreliable estimates of divergence times and gene flow among populations [36].

Ascertainment bias can be a serious issue for studies of population structure [6]. Preferentially discovering SNPs with high heterozygosity leads to an underestimation of the magnitude of structure. A possible explanation is that the mutations that created these high heterozygosity SNPs tend to be older than the mutations that are responsible for low heterozygosity SNPs. These older mutations are more likely to have had time to be distributed across the population by migration. This problem can be lessened by identifying SNPs from a large panel of individuals collected across all target populations [26].

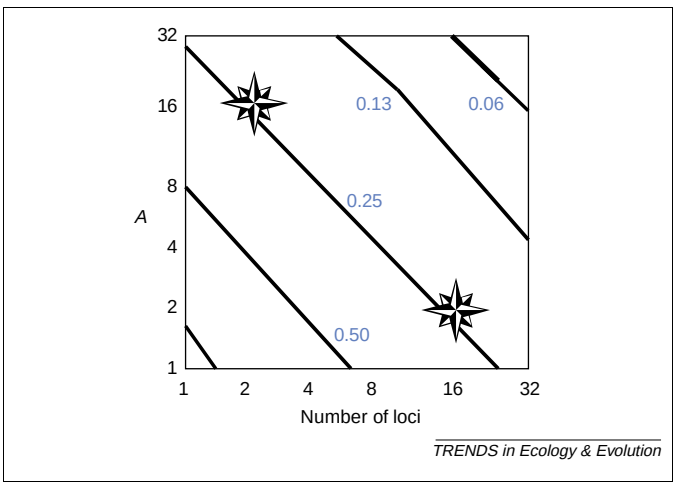

Figure 4. Theoretical relationship between the number of loci, the effective number of alleles (A = A_{total} − 1) and the coefficient of variation (CV, numbers above diagonal lines) for F_{st} estimates. This relationship is based on a demographic simulation where t = 50 generations, the effective population size (N_e) = 500. In theory, 16 alleles from two multi-allelic loci (A = 16 on the Y-axis, 2 on the X-axis) give approximately the same variance as does two alleles at each of 16 bi-allelic loci (A = 16 on the X-axis) (see star symbols; CV = 0.25). Diagonal lines represent a given CV (e.g. 0.25). The CV decreases as the number of loci or independent alleles increases. (Figure modified from [36])

Another kind of population-level analysis, population assignment tests, facilitates the identification of migrants and might enable estimation of current rates of dispersal [40]. In assignment tests, precision is positively correlated with heterozygosity [31] (and numbers of loci used), so it is again probable that many more SNPs than microsatellite loci will be required for a comparable degree of precision.

Population size

The abundance of animal populations can be estimated by mark–recapture methods using multi-locus genotypes as 'genetic tags' [41]. Because the target DNA sequence in SNP-based genotyping is appreciably shorter (e.g. 50–70 bp) than that in microsatellite-based genotyping (80–300 bp), degraded DNA samples should amplify more readily and, thus, facilitate use of noninvasive sampling methods.

The effective population size (N_e) can be estimated using several molecular-based statistical methods [37,42]. The short-term (current) effective size is most often estimated using the 'temporal variance' method. This method quantifies the standardized variance in allele frequencies (e.g. $F_{st\text{-}temporal}$) between two temporally spaced samples from one population. The relative precision provided by biallelic versus multi-allelic markers is the same as when estimating population structure from F_{st}. However, ascertainment bias should not affect temporal estimators of N_e, because these estimators simply monitor change in allele frequencies through time. Another short-term N_e-estimator based on genotypic disequilibrium between unlinked loci [37] should be relatively precise when using SNPs, because its precision

depends mainly on the number of loci (as well as sample size) and not on the number of alleles per locus. However, ascertainment bias affecting estimates of genotypic disequilibrium could occur if markers with high heterozygosity are favored.

Estimators of the long-term N_e often infer the effective size from the level of heterozygosity in a population, assuming mutation-drift equilibrium (e.g. [43]). Because the rate at which mutation drift equilibrium is reached is negatively correlated with mutation rate, SNPs are more likely to be affected by mutation–drift deviations than are high mutation rate markers, such as microsatellite loci. Thus, SNPs might provide more biased estimates of long-term N_e.

Changes in population size

A useful approach for detecting significant reductions in population size (population bottlenecks) is to monitor for loss of genetic variation. However, SNPs have only two alleles per locus and this substantially reduces the power to detect the loss of allelic richness (N_A). N_A has proven more sensitive than H_e (expected heterozygosity) for detecting population declines [44].

Another approach that detects recent population fluctuations requires only a single contemporary population sample and tests for an excess or deficit of alleles at even frequencies, relative to expectations under mutation–drift equilibrium (e.g. [45]). Ascertainment bias might be extremely problematic for this kind of approach. For example, choosing SNPs with high heterozygosity and, thus, even allele frequencies might generate false bottleneck signatures [46].

SNPs will often prove less useful than will microsatellite loci for detecting recent bottlenecks, because the bottleneck-induced removal of alleles will lead to monomorphism with SNPs more often than with microsatellites. However, monomorphic loci might still be informative provided that *a priori* estimates of effective population size or mutation rates are available [45]. To overcome the problem of low polymorphism, tightly linked SNPs that define HAPLOTYPES might be used. For example, multiple-haplotype (highly polymorphic) loci might be identified at a given locus if several linked SNPs can be assayed. Haplotype-based approaches are promising and are becoming popular in human population genetic studies (e.g. [47]).

For detecting population expansions, SNPs will be less useful over the short term than will microsatellites, because the accumulation of new mutations (and an excess of rare alleles) requires longer time periods for slowly evolving loci. However this deficiency of SNPs should be overcome by using many loosely linked SNPs and information from recombination and linkage disequilibrium.

In summary, the relative usefulness of SNPs in comparison to other molecular markers is context specific and more work is required to identify general rules, except in the cases of individual identification and parentage testing [29,30]. The uncertainty is due to the dependence of statistical power on the study objective, the test statistics and evolutionary history of the target populations, as well as sample sizes. Nonetheless, the above examples suggest that at least two to six times more SNPs will often be necessary to achieve the same resolution as achieved by microsatellite loci. Fortunately, large numbers of SNPs potentially can be studied owing to their high genome-wide abundance and the relative ease of SNP genotyping in automated formats, potentially improving cost effectiveness and data quality relative to microsatellites. For genetic distance based studies, F_{st} with SNPs is likely to be more accurate than F_{st} or R_{st} from microsatellites [36]. The use of many SNPs improves genome representation and detection of aberrant outlier loci, but potentially increases the risk of non-independence between loci.

Natural selection and conservation

Over the past two decades, many genetic surveys of natural populations have focused on neutral loci. Although this has provided new insights into the historical demography and evolution of populations [48], the missing element is an understanding of the dynamics of genes that affect fitness. In model species that have been the object of genomic sequencing efforts, the search for genes under selection is advancing [49,50]. For non-model species, SNPs might be useful for finding genes under selection and studying the dynamics of these genes in natural populations [51]. SNPs within exons and introns under divergent directional selection are predicted to have divergent allele frequencies exceeding those expected from neutral genes (e.g. [52]). Consequently, SNP surveys across populations experiencing divergent natural selection might yield a subset of SNPs having statistically higher variance in allele frequency as measured by F_{st} [53,54].

Two primary approaches have been used to identify and study genes or gene pathways (e.g. detoxifying pathways) influencing fitness, which might also be useful for non-model organisms. First, candidate genes of known function that might be predicted to influence fitness in a particular environment can be identified and sequenced, for example, genes encoding transferrin in salmon [7,55]. SNPs in candidate genes can potentially be used to detect selection. However, the degree of association between a SNP and the locus under selection depends on the intensity of the selective sweep, the distance between the SNPs and the locus under selection, the recombination rate and time since the sweep (e.g. [56]).

Whole-genome scans provide a second approach for identifying genomic regions under selection. An unresolved question is the number of SNPs that need to be discovered and genotyped to provide an effective genome scan for selection [57]. A focus on candidate gene regions (e.g. [54]) makes SNPs more practical, although, at present, this approach can be difficult in non-model species.

A possible third approach involves a large-scale SNP study based on a panel of a few dozen genes that are polymorphic and that might influence fitness. The effort required is, therefore, less than a full genomic scan but greater than that of a targeted candidate gene survey. Such comprehensive scans are motivated by a pressing need in conservation genetics for alternatives to neutral

markers as surrogates for levels of adaptive variation and divergence [7,58]. Although neutral markers enable inferences about historical demography and divergence owing to isolation and drift to be made, they are not highly correlated with levels of variation in fitness traits [59]. The association between quantitative traits (Q_{st}) and F_{st} at neutral loci is uncertain [60,61]. Consequently, genetic markers that are closely linked with genes influencing fitness might provide a better indicator of levels of adaptive variation within populations and their potential to respond to changing environmental conditions [7,58].

Currently, there are insufficient empirical data to evaluate the usefulness of the above SNP approaches (but see [62]). An essential question here is whether it might be more useful to just sequence entirely 30–40 genes of known function rather than go through the expense of SNP development to assay variation at these loci. Ascertainment bias and the problems of analysis of incomplete sequence information might argue for the former, but current technology and costs still prohibits the routine sequencing of so many genes in population samples. Furthermore, field-collected samples can contain degraded DNA at low concentration and, hence, the larger DNA fragments needed for complete sequencing might not be amplified readily by PCR.

Concluding statement

SNPs might rapidly become the marker of choice for many applications in population ecology, evolution and conservation genetics, because of the potential for higher genotyping efficiency, data quality, genome-wide coverage and analytical simplicity (e.g. in modeling mutational dynamics). They are not without their limitations, however, and might provide marginal additional, or even less, utility in some applications (e.g. relatedness). The widespread use of mtDNA in the 1980s and microsatellites in the 1990s provide examples of situations in which useful genetic markers appeared on the scene, outpacing the technology, theory and bioinformatic systems relevant to their use. SNP technology is before us now, and we have a unique opportunity to implement and standardize the technology, theory and bioinformatic systems that will facilitate the most efficient, economical and informative use of SNP markers within the scientific community. We hope that our review will encourage further investigation of the theoretical, analytical and technical advantages and limitations to SNP genotyping for molecular ecology and conservation genetics studies, and provide a useful framework for investigators in choosing genetic markers most appropriate for their study design.

Acknowledgements

We thank the Internet Science Education Project for funding the SNP workshop, Svante Pääbo and Mark Stoneking for useful presentations, and John Pollinger for designing Figure 1. We are grateful to Pascal Gagneux and three anonymous reviewers for helpful comments. Members of the SNP workshop group contributed extensively and nearly equally to the ideas presented here.

References

1 Hedrick, P.W. (1999) Perspective: highly variable loci and their interpretation in evolution and conservation. *Evolution* 53, 313–318
2 Navajas, M.J. *et al.* (1998) Microsatellite sequences are under-represented in two mite genomes. *Insect Mol. Biol.* 7, 249–256
3 Vignal, A. *et al.* (2002) A review on SNP and other types of molecular markers and their use in animal genetics. *Genet. Sel. Evol.* 34, 275–305
4 Brumfield, R.T. *et al.* (2003) Single nucleotide polymorphisms (SNPs) as markers in phylogeography. *Trends Ecol. Evol.* 18, 249–256
5 Carlson, C.S. *et al.* (2003) Additional SNPs and linkage-disequilibrium analyses are necessary for whole-genome association studies in humans. *Nat. Genet.* 33, 518–521
6 Wakeley, J. *et al.* (2001) The discovery of single-nucleotide polymorphisms – and inferences about human demographic history. *Am. J. Hum. Genet.* 69, 1332–1347
7 van Tienderen, P.H. *et al.* (2002) Biodiversity assessment using markers for ecologically important traits. *Trends Ecol. Evol.* 17, 577–582
8 Syvanen, A.C. (2001) Accessing genetic variation: genotyping single nucleotide polymorphisms. *Nat. Rev. Genet.* 2, 930–942
9 Ahren, D. *et al.* PHOREST: a web-based tool for comparative analyses of EST data manuscript. *Mol. Ecol. Notes* (in press)
10 Lyons, L.A. *et al.* (1997) Comparative anchor tagged sequences (CATS) for integrative mapping of mammalian genomes. *Nat. Genet.* 15, 47–56
11 Aitken, N. *et al.* Screening of CATS loci for SNP discovery in mammals. *Mol. Ecol.* (in press)
12 Sham, P. *et al.* (2002) DNA Pooling: a tool for large-scale association studies. *Nat. Rev. Genet.* 3, 862–871
13 Bensch, S. *et al.* (2002) The use of AFLP to find an informative SNP: genetic differences across a migratory divide in willow warblers. *Mol. Ecol.* 11, 2359–2366
14 Brouillette, J.A. *et al.* (2000) Estimate of nucleotide diversity in dogs with a pool-and-sequence method. *Mamm. Genome* 11, 1079–1086
15 Marth, G. *et al.* (2001) Single-nucleotide polymorphisms in the public domain: how useful are they? *Nat. Genet.* 27, 371–372
16 Neff, M.M. *et al.* (2002) Web-based primer design for single nucleotide polymorphism analysis. *Trends Genet.* 18, 613–615
17 Ren, J. (2000) High-throughput single-strand conformation polymorphism analysis by capillary electrophoresis. *J. Chromatogr. B Biomed. Sci. Appl.* 741, 115–128
18 Taylor, J.D. *et al.* (2001) Flow cytometric platform for high-throughput single nucleotide polymorphism analysis. *Biotechniques* 30, 661–666.668–669
19 LaHood, E.S. *et al.* (2002) Microsatellite allele ladders in two species of Pacific salmon: preparation and field-test results. *Mol. Ecol. Notes* 2, 187–190
20 Fries, R. and Durstewitz, G. (2001) Digital DNA signatures for animal tagging. *Nat. Biotechnol.* 19, 508
21 Sherry, S.T. *et al.* (2000) Use of molecular variation in the NCBI dbSNP database. *Hum. Mutat.* 15, 68–75
22 Morin, P.A. *et al.* (1999) High throughput single nucleotide polymorphism genotyping by the 5′ exonuclease assay. *Biotechniques* 27, 538–552
23 Leberg, P.L. (1992) Effects of a population bottleneck on genetic variation as measured by allozyme electrophoresis. *Evolution* 46, 477–494
24 Mariette, S. *et al.* (2002) Sampling within the genome for measuring within-population diversity: trade-offs between markers. *Mol. Ecol.* 11, 1145–1156
25 Pritchard, J.K. *et al.* (2000) Inference of population structure using multilocus genotype data. *Genetics* 155, 945–959
26 Akey, J.M. *et al.* (2003) The effect of single nucleotide polymorphism identification strategies on estimates of linkage disequilibrium. *Mol. Biol. Evol.* 20, 232–242
27 Kuhner, M.K. *et al.* (2000) The usefulness of single nucleotide polymorphism data for estimating population parameters. *Genetics* 156, 439–447
28 Nielsen, R. (2000) Estimation of population parameters and recombination rates from single nucleotide polymorphisms. *Genetics* 154, 931–942
29 Chakraborty, R. *et al.* (1999) The utility of short tandem repeat loci beyond human identification: implications for development of new DNA typing systems. *Electrophoresis* 20, 1682–1696

30 Krawczak, M. (1999) Informativity assessment for biallelic single nucleotide polymorphisms. *Electrophoresis* 20, 1676–1681
31 Manel, S. *et al.* (2000) Detecting wildlife poaching: identifying the origin of individuals using Bayesian assignment tests and multi-locus genotypes. *Conserv. Biol.* 16, 650–657
32 Schlötterer, C. and Harr, B. (2002) Single nucleotide polymorphisms derived from ancestral populations show no evidence for biased diversity estimates in *Drosophila melanogaster*. *Mol. Ecol.* 11, 947–950
33 Miller, C.R. *et al.* (2002) Assessing allelic dropout and genotype reliability using maximum likelihood. *Genetics* 160, 357–366
34 Blouin, M.S. *et al.* (1996) Use of microsatellite loci to classify individuals by relatedness. *Mol. Ecol.* 5, 393–401
35 Glaubitz, J.C. *et al.* (2003) Prospects for inferring pairwise relationships with single nucleotide polymorphisms. *Mol. Ecol.* 12, 1039–1047
36 Kalinowski, S.T. (2002) How many alleles per locus should be used to estimate genetic distances? *Heredity* 88, 62–65
37 Waples, R.S. (1991) Genetic methods for estimating the effective size of Cetacean populations. In *Genetic Methods for Estimating the Effective Size of Cetacean Populations* (Hoelzel, R., ed.), pp. 279–300, International Whaling Commission
38 Beaumont, M.A. and Nichols, R.A. (1996) Evaluating loci for use in the genetic analysis of population structure. *Proc. R. Soc. Lond. Ser. B* 263, 1619–1626
39 Allendorf, F.W. and Seeb, L.W. (2000) Concordance of genetic divergence among sockeye salmon populations at allozyme, nuclear DNA, and mitochondrial DNA markers. *Evolution* 54, 640–651
40 Paetkau, D. *et al.* (1995) Microsatellite analysis of population structure in Canadian polar bears. *Mol. Ecol.* 4, 347–354
41 Palsbøll, P.J. (1999) Genetic tagging: contemporary molecular ecology. *Biol. J. Linn. Soc.* 68, 3–22
42 Schwartz, M.K. *et al.* (1998) Using DNA to estimate population size: many methods, much potential, unknown utility. *Anim. Conserv.* 2, 321–323
43 Paetkau, D. *et al.* (1998) Gene flow between insular, coastal and interior populations of brown bears in Alaska. *Mol. Ecol.* 7, 1283–1292
44 Spencer, C.C. *et al.* (2000) Experimental evaluation of the usefulness of microsatellite DNA for detecting demographic bottlenecks. *Mol. Ecol.* 9, 1517–1528
45 Beaumont, M.A. (1999) Detecting population expansion and decline using microsatellites. *Genetics* 153, 2013–2029
46 Fay, J.C. and Wu, C.I. (1999) A human population bottleneck can account for the discordance between patterns of mitochondrial versus nuclear DNA variation. *Mol. Biol. Evol.* 16, 1003–1005
47 Stephens, M. *et al.* (2001) A new statistical method for haplotype reconstruction from population data. *Am. J. Hum. Genet.* 68, 978–989
48 Avise, J.C. (2000) Phylogeography. *The History and Formation of Species*, Harvard University Press
49 Sabeti, P.C. *et al.* (2002) Detecting recent positive selection in the human genome from haplotype structure. *Nature* 419, 832–837
50 Harr, B. *et al.* (2002) Hitchhiking mapping: a population-based fine-mapping strategy for adaptive mutations in *Drosophila melanogaster*. *Proc. Natl. Acad. Sci. U. S. A.* 99, 12949–12954
51 Luikart, G. *et al.* The power and promise of population genomics: from genotyping to genome typing. *Nat. Rev. Genet.* (in press)
52 Schlötterer, C. (2002) Towards a molecular characterization of adaptation in local populations. *Curr. Opin. Genet. Dev.* 12, 683–687
53 Lenormand, T. *et al.* (1998) Evaluating gene flow using selected markers: a case study. *Genetics* 149, 1383–1392
54 Kohn, M.H. *et al.* (2000) Natural selection mapping of the warfarin-resistance gene. *Proc. Natl. Acad. Sci. U. S. A.* 97, 7911–7915
55 Ford, M.J. (2000) Effects of natural selection on patterns of DNA sequence variation at the transferrin, somatolactin, and p53 genes within and among chinook salmon (*Oncorhynchus tshawytscha*) populations. *Mol. Ecol.* 9, 843–855
56 Barton, N.H. (2000) Genetic hitchhiking. *Philos. Trans. R. Soc. Lond. Ser. B* 355, 1553–1562
57 Beaumont, M.A. and Balding, D.J. (2000) Identifying adaptive genetic divergence among populations from genome scans. *Mol. Ecol.* (in press)
58 Crandall, K.A. *et al.* (2000) Considering evolutionary processes in conservation biology. *Trends Ecol. Evol.* 15, 290–295
59 Reed, D.H. and Frankham, R. (2001) How closely correlated are molecular and quantitative measures of genetic variation? A meta-analysis. *Evolution* 55, 1095–1103
60 Crnokrak, P. and Merilä, J. (2002) Genetic population divergence: markers and traits. *Trends Ecol. Evol.* 17, 501–501
61 Latta, R.G. and McKay, J.K. (2002) Genetic population divergence: markers and traits – response. *Trends Ecol. Evol.* 17, 501–502
62 Hamblin, M.T. *et al.* (2002) Complex signatures of natural selection at the Duffy blood group locus. *Am. J. Hum. Genet.* 70, 369–383
63 Valdes, A.M. *et al.* (1993) Allele frequencies at microsatellite loci: the stepwise mutation model revisited. *Genetics* 133, 737–749
64 Taberlet, P. and Luikart, G. (1999) Non-invasive genetic sampling and individual identification. *Biol. J. Linn. Soc.* 68, 41–55
65 Waits, L.P. *et al.* (2001) Estimating the probability of identity among genotypes in natural populations: cautions and guidelines. *Mol. Ecol.* 10, 249–256
66 Jamieson, A. and Taylor, S.C. (1997) Comparisons of three probability formulae for parentage exclusion. *Anim. Genet.* 28, 397–400
67 Primmer, C.R. *et al.* (2002) Single-nucleotide polymorphism characterization in species with limited available sequence information: high nucleotide diversity revealed in the avian genome. *Mol. Ecol.* 11, 603–612
68 Makridakis, N.M. and Reichardt, J.K. (2001) Multiplex automated primer extension analysis: simultaneous genotyping of several polymorphisms. *Biotechniques* 31, 1374–1380
69 Hirschhorn, J.N. *et al.* (2000) SBE-TAGS: an array-based method for efficient single-nucleotide polymorphism genotyping. *Proc. Natl. Acad. Sci. U. S. A.* 97, 12164–12169
70 Hsu, T.M. and Kwok, P.Y. (2003) Homogeneous primer extension assay with fluorescence polarization detection. *Methods Mol. Biol.* 212, 177–187