The utility of single nucleotide polymorphisms in inferences of population history

Robb T. Brumfield[1], Peter Beerli[2], Deborah A. Nickerson[2] and Scott V. Edwards[3]

[1]Museum of Natural Science, 119 Foster Hall, Louisiana State University, Baton Rouge, LA 70803, USA
[2]Department of Genome Sciences, Box 357730, University of Washington, Seattle, WA 98195, USA
[3]Department of Biology, Box 351800, University of Washington, Seattle, WA 98195, USA

Single nucleotide polymorphisms (SNPs) represent the most widespread type of sequence variation in genomes, yet they have only emerged recently as valuable genetic markers for revealing the evolutionary history of populations. Their occurrence throughout the genome also makes them ideal for analyses of speciation and historical demography, especially in light of recent theory suggesting that many unlinked nuclear loci are needed to estimate population genetic parameters with statistical confidence. In spite of having lower variation compared with microsatellites, SNPs should make the comparison of genomic diversities and histories of different species (the core goal of comparative biogeography) more straightforward than has been possible with microsatellites. The most pervasive, but correctable, complication to SNP analysis is a bias towards analyzing only the most variable loci, an artifact that is usually introduced by the limited number of individuals used to screen initially for polymorphisms. Although the use of SNPs as markers in population studies is still new, innovative methods for SNP identification, automated screening, haplotype inference and statistical analysis might quickly make SNPs the marker of choice.

Traditionally, phylogeography has used gene trees of non-recombining, uniparentally inherited LOCI (see Glossary), such as mitochondrial DNA or the vertebrate Y chromosome, to study the geographical distribution of genetic variation within species [1]. As evolutionary biologists have started to examine variation in recombining, biparentally inherited loci, a natural outgrowth of phylogeography is a shift from gene trees to analyses, based on COALESCENT THEORY, of multi-locus, recombining histories. This new discipline, dubbed historical demography [2,3] or statistical phylogeography [4], is concerned less with gene trees than with estimating population parameters such as genetic diversities, divergence times, growth rates and gene flow between populations. The shift in focus is, in part, a result of recent advances in population genetics, which suggest that, from a statistical standpoint, the ability of single-locus phylogeography to determine the timing of speciation events and the historical demography of populations has been overestimated [3–7]. The errors surrounding estimates of divergence times, rates of gene flow and population-size changes during speciation are all reduced substantially when information from multiple unlinked loci is combined [8,9]. With the move to analyses of multiple loci, phylogeographers must re-learn an old lesson: that the number of loci required to estimate the preceding parameters with statistical confidence can be soberingly large because of the high stochasticity of the gene tree of any single locus [10]. What is required is a suite of unlinked nuclear genetic markers that can capture a genome-wide picture of the population history [3,11–14]. The polymerase chain reaction (PCR) as well as fluorescent sequencing and fragment analysis technologies have catalyzed a revolution in the development of genetic markers for the analysis of natural populations. Emphasizing discoveries in non-model species, we discuss one emerging marker of great relevance to historical demography: single nucleotide polymorphisms (SNPs).

Glossary

Ascertainment bias: bias introduced into an analysis because of arbitrary decisions made during data sampling. In SNP or microsatellite studies, ascertainment bias can arise if only the most variable loci are analysed or if only a small panel of individuals is used to discover variation.
Coalescent theory: population genetic framework that allows one to calculate the probability of obtaining a given genealogical structure for many contemporary samples under many different population genetic models.
cSNP: SNP identified from a coding DNA region.
Diplotype: genetic data that cannot be assigned unambiguously to one of the two chromosomes in a diploid organism.
Haplotype: genetic data from a single chromosome (e.g. data from a single sperm or egg, or from the haploid mitochondrion); a set of sites linked on the same allele or chromosome.
Indel: a sequence gap caused by an insertion or deletion mutation.
Intron: within a protein-coding gene, intervening sequence that does not perform a coding function. These regions typically have a higher substitution rate than do the flanking coding sequence.
Loci: genetic regions of interest; here, we assume that a locus is part of a larger region that will freely recombine with other regions. Loci can be unlinked or tightly linked.
ncSNP: SNP identified from a noncoding DNA region.
Panel: initial group of individuals used to screen for variation.
Substitution: mutational change in the character state of a nucleotide.
Theta (Θ): a measure of genetic diversity equivalent to four times the product of the effective population size and the mutation rate per site per generation ($4N_e\mu$).

Corresponding author: Robb T. Brumfield (brumfld@lsu.edu).

http://tree.trends.com 0169-5347/03/$ - see front matter doi:10.1016/S0169-5347(03)00018-1

Approximately 90% of genetic variation in the human genome is in the form of SNPs [15], the result of point mutations that produce single base-pair differences (SUBSTITUTIONS or INDELS) among chromosome sequences. But, it has only been with the availability of large databases of overlapping sequences and the development of novel, large-scale SNP identification and screening technologies that their characterization and use as markers has accelerated [16–18]. The prevalence and distribution of SNPs in genomes has resulted in several large-scale initiatives to identify and characterize hundreds of thousands of SNPs from the genomes of model organisms such as *Mus musculus*, *Drosophila melanogaster*, *Caenorhabditis elegans* and *Arabidopsis thaliana* [19–24]. Currently, SNPs are used primarily in whole-genome linkage and association studies, but their ubiquity, tractable levels of variation and ease in screening suggest that they will dominate increasingly as markers for elucidating the evolutionary history of populations.

How one goes about identifying SNPs will depend largely on the organism of interest. Overlapping homologous sequences of non-model organisms are not prevalent in the public data bases [25], and so most researchers will have to find their own SNPs. Although the number of universal primers for the amplification of segments of nuclear protein-encoding genes in non-model species is increasing [26–28], even these are unlikely to provide enough loci as studies of population history begin to scale up. Researchers can utilize other strategies for finding SNPs in their focal species such as: (1) designing primers from conserved sequences of related species that are available online; or (2) sequencing anonymous nuclear loci [29], either through cloning [30] or PCR methods, such as amplified fragment length polymorphisms [31,32] and microsatellite discovery [27]. A comparison of SNP frequencies in a diverse collection of taxa illustrates that researchers can expect, on average, to sequence at least 200–500 base pairs of noncoding DNA to find a single NCSNP, and ~500–1000 base pairs to locate a CSNP (Table 1). Several new computer programs facilitate greatly the identification of heterozygous sites directly from sequence chromatograms (Box 1, Figs I,II), decreasing the need for allele-specific primers to detect heterozygotes. If a SNP gene tree is required, methods are available for inferring HAPLOTYPES from DIPLOTYPES statistically (Box 2), thereby bypassing laborious laboratory procedures for haplotype resolution.

Application of SNPs in studies of population history

It is only recently that a conceptual framework for the population genetic analysis of SNPs, founded appropriately

Table 1. Sampling statistics of nuclear SNPs from selected multi-locus studies

Taxon	Individuals orstrains	Nt[b] diversity ($\times 100$)	SNP freq. (bp^{-1})	Variable loci (%)	SNPs/locus (range)	Refs
Dolphins (four introns; 7118 bp)						
Lagenorhynchus obscurus	5	16.00	254	100	7.0 (5–9)	[52,53]
L. obliquidens	6	11.00	297	100	6.0 (2–9)	
L. acutus	3	7.00	2373	50	0.7 (0–2)	
Human						
All STSs (1139 STSs)	13	0.04	1001	–	0.24	[43,54]
EST sequences (705 STSs)		–	1159	–	0.23	
Random genomic sequence (434 STSs)		–	785	–	0.27	
Flycatchers (28 loci)						
Ficedula hypoleuca						
Intron (5753 bp)	2–7	1.80	205	79	2.0 (0–6)	
Msat flanking sequence (692 bp)	2–6	4.50	87	50	1.3 (0–5)	
Anonymous (2637 bp)	2–7	2.90	165	63	2.0 (0–6)	
Ficedula albicollis						
Intron (5780 bp)	2–8	2.50	156	86	2.6 (0–6)	
Microsatellite flanking sequence (718 bp)	2–5	5.70	72	50	1.7 (0–8)	
Anonymous (2666 bp)	2–6	2.30	190	63	1.7 (0–5)	
Murrelet (nine introns)						
Brachyramphus marmoratus	120	NA	80	100	6.8 (4–14)	[28,44,55]
Parrot (four ESTs)						
Psittacus erithacus	34	NA	314	75	1.7 (0–5)	[56]
Chicken (23 427 ESTs)[a]						
Gallus gallus	NA	NA	2119	NA	NA	
Fruit fly						
Ceratitis capitata (four loci; coding; 2037 bp)	11	1.1	52	100	9.8 (8–15)	[26]
Sugar beet (37 loci)						
Beta vulgaris	2 inbred strains					[57,58]
Coding and noncoding (18 002 bp)		7.60	130	59	1.4 (0–19)	
Coding (13 604 bp)		3.50	283	51	0.5 (0–7)	
Fungus (five loci, primarily coding)						
Coccidioides immitis (2384 bp)	8	NA	159	80	3.0 (0–6)	[59]

[a]H. Kim *et al.*, unpublished.
[b]Abbreviations: EST, expressed sequence tag; NA, not available; Nt, nucleotide; SNP, single nucleotide polymorphism; STS, sequence tagged site.

Box 1. Methods for inferring SNPs from sequence chromatograms

There are many laboratory and computational approaches to finding single nucleotide polymorphisms (SNPs) within a genome, but all involve some form of comparative analysis of the same DNA segment from different individuals or from different haplotypes. For example, DNA segments amplified by polymerase chain reaction (PCR) can be compared with one another based on conformational analysis, melting temperature analysis, the ability to be cut differentially with enzymes or chemicals, or by direct sequence analysis. Fluorescence-based sequencing has an increasingly important role in identifying SNPs, not only because of its high throughput and decreasing costs, but also because of the development of programs that automate the base-calling (Phred [60]), assembly (Phrap; http://www.phrap.org/) and finishing (Consed [61]) of large-scale projects. One of the significant advantages of the Phred base-caller is its ability to provide a quality score (i.e. an error rate probability) for each base-call in a sequence. Sequence quality scores that improve the speed and accuracy of identifying DNA variations among assembled sequences have also been incorporated into additional programs, such as Polybayes [62] and PolyPhred (http://droog.gs.Washington.edu/PolyPhred.html) [63,64].

Many projects have identified SNPs as high quality nucleotide mismatches between overlapping clone sequences from large-scale genome sequencing or from expressed sequence tags (ESTs; see Box Glossary) [65]. Polybayes is a program that applies a Bayesian inference to calculate the probability of a given site being polymorphic and has been applied on a large-scale to identify SNPs across the human genome [19]. PolyPhred is another program that operates together with Phred, Phrap and Consed to identify SNPs as high quality base mismatches in assembled cloned sequences (Fig. I). Importantly, PolyPhred can also detect SNPs as heterozygotes (two bases at a single position in the sequence) in diploid sequences amplified by PCR (Fig. II). Laboratory analyses suggest that dye primer sequencing, in which the sequencing primer is fluorescently labeled, produces more even chromatogram peaks than in dideoxy sequencing, resulting in more accurate detection of heterozygous sites [63,64]. Automated base-calling programs, such as Polyphred, have been used in only a handful of non-model species (e.g. red-winged blackbird [66]).

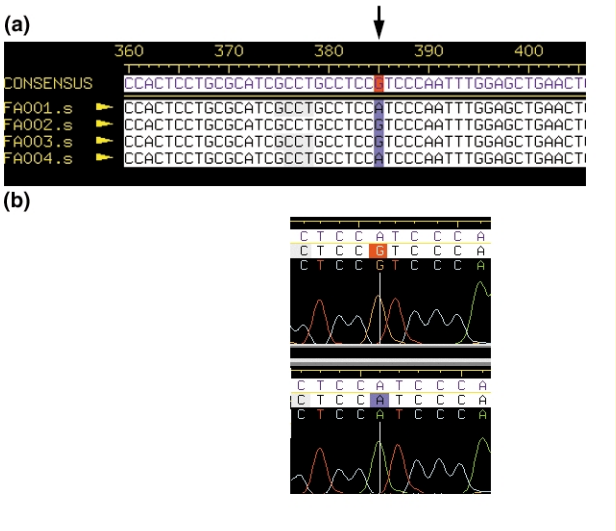

Fig. I. Consed view of sequences scanned with PolyPhred 4.0. (a) The consensus sequence has been tagged red at position 385 (below arrow), which indicates a high scoring polymorphic site within the compared sequences. (b) Consed also enables the user to view the underlying sequence trace data (below the consensus window).

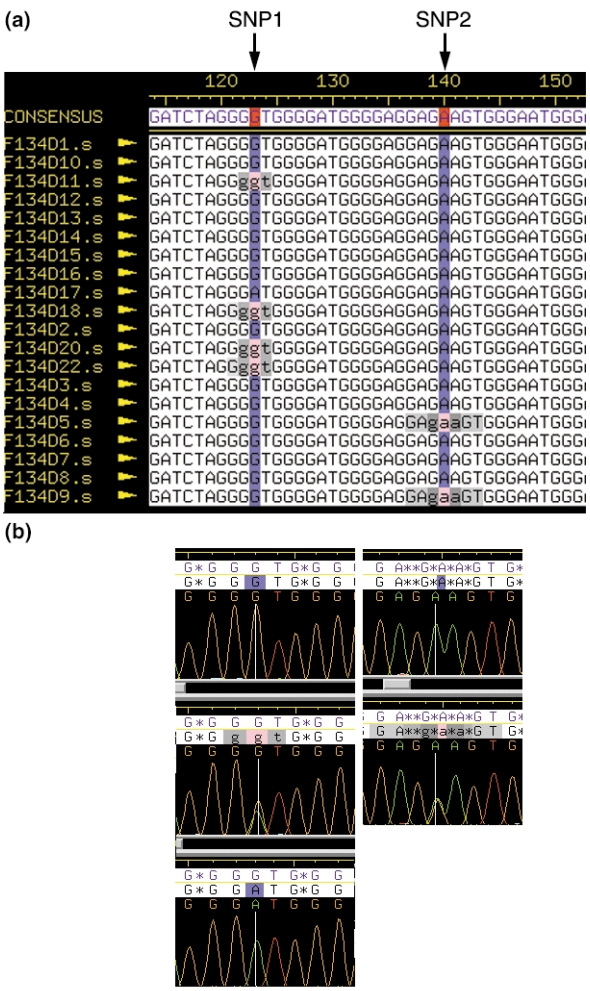

Fig. II. Heterozygote detection using chromatograms. (a) The detection of substitution polymorphisms by PolyPhred is based on the simple observation that the pattern of fluorescence-based dye incorporation at each position in a diploid sample is reproduced faithfully every time the same sequence is generated. Heterozygous positions (two different bases at the same location in a sequence – marked by pink on the individual traces) can be identified accurately based on both the predictable reduction or drop (~50%) in the height of a peak that occurs at a heterozygous position relative to a homozygous one and the presence of a second fluorescent peak at the heterozygous site. (b) A consed view of two high-quality SNPs (SNP1 and SNP 2) detected by PolyPhred and highlighted in red on the consensus of a group of assembled sequences. Notice that, for SNP1, alternative homozygotes are found among the sequences, whereas for SNP2, heterozygotes are detected in the presence of only one of the homozygous forms.

Glossary

Expressed sequence tag (EST): short single-copy DNA sequence derived from reverse transcribed messenger RNA (coding or cDNA). By definition, ESTs occur in coding regions.

on coalescent theory, has been developed [9,11]. Perhaps the most important nuance to SNP analysis is the need to correct for the ASCERTAINMENT BIAS [33] that arises as a by-product of how the SNPs are identified and/or screened [11]. Before starting a SNP study of population history, the researcher must make two decisions that will determine the necessary ascertainment correction. First, how is a SNP defined? The researcher could decide that every variable site is a SNP, or could set a cut-off frequency below which variable sites are not considered to be SNPs. The latter would be justifiable if the researcher wanted to eliminate sites that are variable because of infrequent

Box 2. Inferring haplotypes from diplotypes

In most cases, the chromosome assignment for sequence data generated from diploid organisms is unknown (Fig. I). Aside from direct laboratory analysis [67,68], the most common method of assignment is the inference of haplotypes by using computer programs [69–71]. These programs try to infer a minimal number of haplotypes using the frequencies of the polymorphic sites or heuristic approaches. Recently, Stephens *et al.* [71] developed a Bayesian method that assigns probabilities to possible haplotypes and to individual sites within a haplotype. An alternative method [72] does not infer the haplotype, but instead considers the haplotype uncertainty when estimating population size and recombination rates. This is accomplished by integrating over all possible haplotype assignments using MARKOV CHAIN MONTE CARLO (see Box Glossary) [73]: (1) the program starts with a specific configuration (A–G and T–A) and calculates the likelihood of this data configuration given the genealogy; (2) one of the heterozygous sites is then flipped at random (e.g. A–A and T–A) and the likelihood is recalculated; and (3) an acceptance/rejection scheme is used that takes the different probabilities into account, accepting the change when the new configuration has a higher likelihood, or, if it has a lower likelihood, when it surpasses a probability in proportion to how much worse the new configuration is. Such methods point to a new phase of historical demography in which a single inferred gene tree for a locus is of less interest than the population parameters averaged over all trees and haplotypes.

Glossary

Markov chain Monte Carlo: complicated problems can often not be solved analytically. Markov chain Monte Carlo methods make a random change to a start state and then compare the two states using a ratio of probabilities to accept the new state if the new state is better than the old or, alternatively, accept the new state with some probability that depends on that ratio. The next cycle starts from the new 'old' state.

Fig. I. The two single nucleotide polymorphisms (A/T; G/A) specify two heterozygous sites for the diplotype. The A at the first polymorphic site could sit on the same chromosome as either the G or the A from the second polymorphic site. It is necessary to resolve haplotypes to estimate recombination rates and linkage disequilibrium accurately, or to reconstruct a genealogical tree of haplotype relationships (gene tree).

sequencing errors. However, under this protocol, valuable data could be lost. Second, are all individuals in the study going to be sequenced for all loci (even the invariant ones)? Another widely used strategy is to identify SNPs from a PANEL of individuals that is limited in size and composition compared with the target samples [34]. This approach can save considerable time and money because the SNPs discovered can be screened in the target samples using non-sequence methods (e.g. probes or primers specific to the SNP). The drawback is that, by focusing on only those sites that are variable in the panel, the researcher might miss some SNPs, an uncertainty that must be corrected (Box 3). This bias was illustrated in a recent *Drosophila* study in which genetic diversity estimates differed depending on the geographical location of the panel [35].

The effects of not correcting for the method of ascertainment can be profound, especially if a panel is used. Because rare variants are less likely to be encountered in the panel, the pool of SNPs examined in target populations will be distorted towards sites with intermediate base frequency. From coalescent theory, these SNPs tend to occur on intermediate branches of the gene tree (assuming that the polymorphic sites are bi-allelic). Accordingly, the genealogical tree from a biased set of SNPs often lacks resolution towards the tips, making population samples appear more similar than they truly are. Migration rates inferred from a biased set of SNPs or loci will be overestimated because low-frequency, recently derived haplotypes that are unique to the populations are missed [34].

Why choose SNPs for studies of population history?

Mutation pattern

Unlike microsatellites, which have mutation rates per generation of the order of 10^{-4}, SNPs have relatively low mutation rates (10^{-8}–10^{-9}). Multiple mutations at a single site are thus unlikely, and so most SNPs are bi-allelic, a quality that facilitates high-throughput genotyping and minimizes recurrent substitutions at a single site that would confound the population history. The restriction to four character states might make SNPs less informative than microsatellites for parentage analyses [36] or for detecting fine-scale geographic structure, but this limitation can be offset by scoring more loci [37]. For linkage studies, approximately three times as many SNPs are needed in comparison to microsatellites [38]. The relative number of SNPs needed to estimate population genetic parameters with statistical confidence is likely to be the same, although we caution that the necessary simulation studies have not yet been performed.

We suggest that, when all population genetic and analytical considerations are weighed, SNPs are in fact superior to microsatellites for elucidating historical demography (Table 2). Mutation rates at microsatellite loci are difficult to estimate, and vary across loci and across alleles within the same locus [39]. Perhaps most importantly for comparative studies of historical demography, the difference in evolution and variation of the same microsatellite locus in even closely related species makes them ill suited for interspecific comparisons of genomic variability. Nuclear SNPs are measured on the same mutational scale as mitochondrial (mt) SNPs (substitutions per site), making intergenomic comparisons easier. Many more tests for deviations from neutrality, for population size changes and for recombination exist for SNP data than for microsatellites, and the fit of models to data is probably better for SNPs. The recent focus on microsatellites has caused an unconscious ascertainment bias in the estimation of genomic variability for many species. This undue focus creates an idiosyncratic description of genomic variation, insofar as microsatellites will not record the 'background' levels of variability that one would like to use to compare variation across species. In addition, measures of population differentiation, such as F_{ST}, can be very sensitive to the level of within-population variation, resulting in suspiciously low values in many microsatellite studies [40–42].

Unbiased sampling of genomic variation

SNPs offer historical demography a great opportunity for unbiased sampling of loci. As an extension of recent studies

Box 3. Ascertainment bias

Single nucleotide polymorphisms (SNPs) are ascertained by: (1) sequencing all individuals for an entire DNA region of interest (Fig. Ia) or (2) sequencing a panel (subset) of individuals, identifying variable sites, defining which of those sites will be considered SNPs, and then screening only those sites in a sample of interest by using SNP-specific primers or probes (Fig. Ib).

By using maximum likelihood, the ascertainment bias can be corrected for by conditioning on the specific details of the ascertainment [9,11,33,34,73]. The specific interpretation of how a SNP is defined influences the genealogy, and thus the estimation of population genetic parameters in a coalescent framework. In the example in Fig. IIa, 182 variable sites were identified after sequencing ten individuals. The effect of not recognizing all variable sites as SNPs on the tree topology can be striking. If all singleton mutations are discarded as sequencing errors, resolution is lost at the tips (Fig. IIb); an even more stringent interpretation of SNPs (Fig. IIc) would alter the tree more strongly. The ascertainment bias tends toward a pool of SNPs having intermediate frequencies, resulting in decreased resolution towards the tips of the tree. The effect on population genetic inference of not making an ascertainment correction when one is needed has been illuminated through simulation studies [9]. Currently, only a few computer programs, such as LAMARC (http://evolution.gs.washington.edu/lamarc.html), allow for ascertainment bias correction [74].

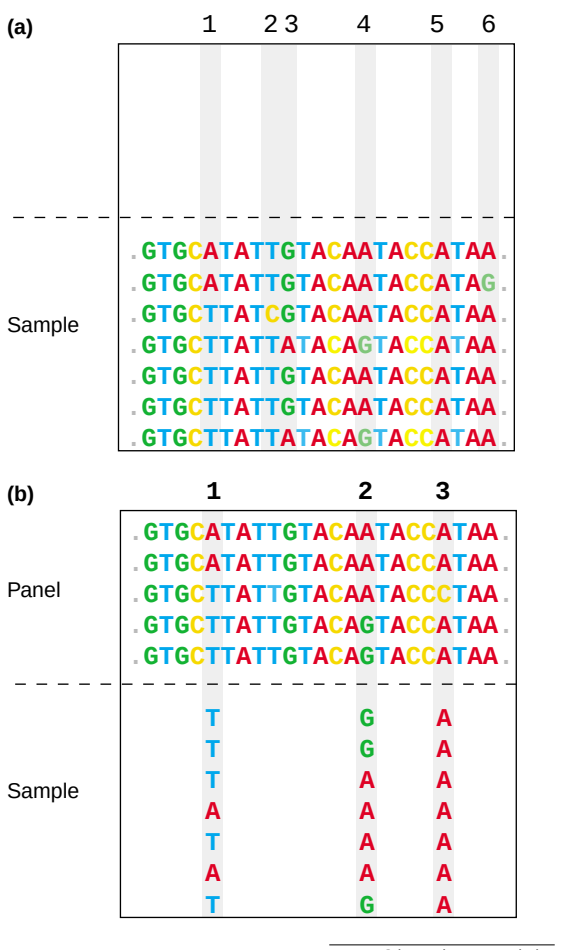

Fig. I. SNP screening in which (a) all individuals are sequenced or (b) a panel is used to identify SNPs.

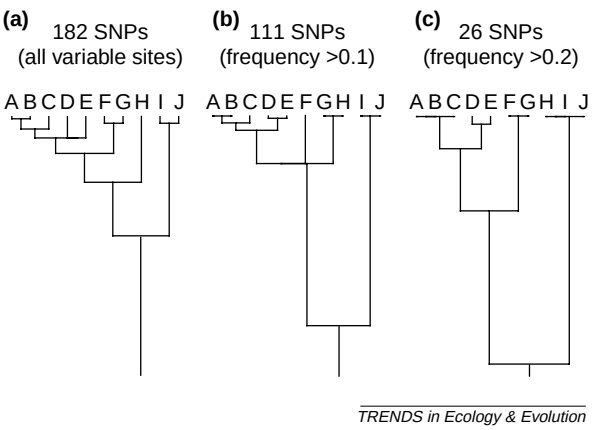

Fig. II. Effects on the genealogy of sampling (a) all SNPs in a DNA region or (b and c) a subset of SNPs.

that focus, for example, exclusively on INTRONS [3], we can envisage future studies in which SNP loci are sampled at random without reference to, or previous knowledge of, rates of mutation (Box 4). The least biased descriptions of genomic variation will come from studies that sample nucleotide sites blindly. Such studies are now being published based on humans, and indicate that, for example, ~1 in 1000 nucleotides are variable [43]. The low figure of nucleotide diversity might seem discouraging, yet it is precisely this sort of statistic that is invaluable for facilitating comparisons between species and for providing a measure of genome variability that is appropriate in

Table 2. Characteristics of four types of DNA marker for broad-scale historical demography and detection of population differentiation[a]

Characteristic	SNPs[b]	Msats	AFLPs	RFLPs
Within-population variation	**Low**	High	Unknown	**Low**
Among-locus variation	**Low**	High	Unknown	**Low**
Conducive to unbiased sampling among loci?	**Yes**	No	**Yes**	**Yes**
Susceptibility to molecular convergence	**Low**	High	Unknown	**Low**
Variation easy to interpret?	**Yes**	No	No	Moderate
Are the variations detected further reducible?	**No**	Yes	Yes	Yes
Easy comparison with mitochondrial DNA?	**Yes**	No	No	**Yes**

[a]Traits in bold indicate suggested advantages. Differences among markers in ease of assaying variation or of laboratory isolation are not considered.
[b]Abbreviations: AFLPs, amplified fragment length polymorphisms; Msats, microsatellites; RFLPs, restriction fragment length polymorphisms; SNPs, single nucleotide polymorphisms.

context. For example, a study of seabirds [44] detected SNPs at a frequency of ~1 in 80 nucleotides based on the nine loci studied, although this study focused exclusively on introns, and not on the totality of noncoding DNA.

Does this mean that phylogeographers need to devote more effort to finding variation that is useful for analysis in a study that uses SNPs than in one that uses microsatellites? A better question is whether microsatellites, and the concomitant focus on hypervariability, provide more complete or less complete measures of population variation than do SNPs (Table 2). But what about the fact that SNPs often show very low variability in the few species that have been examined? We suggest that this generalization (if it is supported in further studies) is telling us something about the level of naturally occurring genome-wide variation that a focus on microsatellite variation alone cannot. More side-by-side comparisons of population inferences from SNPs and microsatellites on the same samples are needed.

How SNPs should be sampled across the genome is another pressing issue. It is known that the number of independently segregating loci is the crucial parameter for estimating THETA (Θ) accurately [45], and presumably other population parameters as well [6]. The following question thus arises. If one had resources to sequence 10 kilobases (kb) per individual, is it better to sample two loci of 5 kb each or 20 loci of 500 base pairs – or, if feasible, 1 base pair at 10 000 loci? The answer depends on the focus of the study. It seems reasonable to assume that the encounter rate of SNPs per base pair should be the same under all schemes, unless of course variation in mutation rate across the genome [46] results in a focus on high-variation regions. If estimating the population history is the goal of a study, it seems that the last strategy should be superior. In fact, when estimating Θ from a randomly mating population, theory suggests that interrogating a single nucleotide site per locus across six to seven individuals, and maximizing the number of loci in this way, would be the most efficient approach [45]. In the face of among-locus rate variation, the last strategy will still be superior because only in this way will variation across the entirety of the genome be assayed in an unbiased way. Several recent studies have focused on SNP variation in the human X chromosome [47,48] or in single large (~10 kb) regions on autosomes [47–49], because it is relatively easy to score SNPs on the X chromosome by scoring only males (which are haploid for this chromosome), or because there is a specific interest in the gene region studied. However, studies focusing on SNP variation within a single genomic region cannot generalize the results easily to the variation across the entire genome – the clearest window on the total history of the species under study.

Recombination

A complication of an approach that assays variation at many loci across the genome is that recombination cannot be ignored, as it is in mitochondrial and human Y chromosome studies. Recombination influences both the interpretation and the sampling strategy of SNP

Box 4. A program for comparative studies of historical demography based on noncoding SNPs

As researchers move away from single-species phylogeographical questions to analyses of multiple, co-distributed species [75], it is useful to consider which types of marker will provide both the resolution and the comparative utility that such an approach requires. The scope of comparative phylogeography has expanded to include multilocus descriptions of genealogical patterns within and between species, as well as comparisons of historical demographies, patterns of gene flow and divergence times [3]. We suggest that single nucleotide polymorphisms (SNPs), particularly when standardized to noncoding DNA, provide the most appropriate tools for such a research program, and have several advantages over microsatellites, another prevalent marker. Implicit in our vision is a need to describe genome-wide variability of multiple species, without biases for or against variable loci, and in ways that maximize our ability to compare histories between species [76].

The research program

- Loci for analysis from each species to be analyzed should be chosen without regard to variability, except possibly that the more variable noncoding portion of the genome would be a focus.
- Heterozygous sites (SNPs) can be scored directly or indirectly and with appropriate software. A large number of relatively short loci (~500 base pairs typed for SNPs) are likely to provide more information than a few very long loci (e.g. 10 kilobases [49]) and, in principle, one should encounter SNPs at the same rate using either approach.
- For any given species, some loci will show no variation within a panel of individuals (ideally geographically widespread) on which loci are tested; such loci can be scored as 'invariant' for the panel and not analyzed for the entire sample. However, such loci are not discarded. Rather, using appropriate statistical methods [34], the unknown (but presumably low) variation at such loci can be incorporated without wasting effort and resources on discovering rare SNPs.
- The PCR primers for some fraction of the loci isolated in one species will work on homologous loci in other species in the study; the number of such 'common loci' between species should be maximized. However, many PCR primers will not work for homologous loci in other species in the study. Thus, the final data set in a comparative analysis will comprise a set of loci that are common to all of the species, and a set that is common only to a subset of species.

Provided that enough loci are analyzed, estimates of genomic history obtained by the above methods will not be comparable across species only when global differences in rates of mutation exist between the species compared. Such differences can be detected empirically by employing relative rate tests on the set of common loci. When rate differences cannot be shown to exist for common loci, it is reasonable to place the entire, nonoverlapping set of loci of all species on a common molecular yardstick.

Advantages to the research program

- Variability is assayed in an unbiased fashion across the entire genome.
- Variability and history of different species can be placed on a common mutational timescale.
- The units of measurement of nuclear SNPs are the same as those for mitochondrial SNPs, making intergenomic comparisons of demographic history more straightforward.
- By comparing rates of SNP evolution between species using relative rate tests on common loci, genome-wide differences in mutation rates between species can be detected and used to analyze datasets that employ only partially overlapping sets of loci between species.

variation. Nachman [46] has reviewed the dramatic effects of recombination on the level of SNP variation in humans, in which SNP variation is low in regions of low recombination, and high in regions of high recombination. The same pattern is likely to hold in many other species, both model [50] and non-model organisms. Although the precise interpretation of this pattern is still debated (it could be a result of positively selected sweeps, to negative 'background' selection reducing variation at linked sites, or to a mutagenic effect of recombination [51]), the fact that the nucleotide diversity of a given anonymous locus will depend crucially on the local recombination rate should figure prominently in the minds of phylogeographers, even if recombination rates are unknown, which is the case for most non-model species.

Conclusion

SNPs have the potential to place historical demography and speciation studies on a common molecular framework, one that is easily comparable to the decades of mtDNA work already undertaken [77]. Their simplicity, ease of modeling and sheer abundance will make them powerful contributors to the new era of using multiple biparentally inherited, potentially recombining loci to infer population histories. The challenge for evolutionary biologists will be to harness and assay variation at large numbers of unlinked SNPs, so as to overcome their lower mutation rates and what initial studies suggest might be a lower level of variation in natural populations than researchers in the field are accustomed to from using microsatellites (Box 5). Fortunately, new automated methods of SNP detection and a wealth of novel theoretical tools will allow researchers to probe the evolutionary history of populations in unprecedented detail.

Box 5. An example multilocus study of population history employing SNPs

The design of SNP studies is probably most developed for work on humans. With the human genome sequence in hand, one is able to choose loci that are far from potential genes and other targets of selection, and that are embedded in genomic regions with known physical distances between loci, similar recombination rates and base compositions. Choosing loci in this way maximizes the power of coalescent simulations to detect deviations from a specified demographic model. Pluzhnikov *et al.* [77] examined SNP variation in three human populations for ten unlinked genomic regions. Their locus sampling strategy minimized the among-locus variation in nucleotide diversity per base pair (see their Table I). For two of the populations, a combination of coalescent simulations involving distributions of recombination and mutation rates and the resultant multilocus means and variances of summary statistics allowed the authors to reject a model of constant population size, or even simple growth models, in favor of complex models involving growth, population structure, and/or bottlenecks. The authors noted that the data from uniparentally inherited, non-recombining loci in humans suggest a simpler growth scenario that is called into question by the multilocus SNP data, and that the picture from microsatellite analyses is even more incongruent, probably as a result of the difficulty of estimating mutation rates and patterns from these loci.

The use of standardized multilocus SNP data for species other than humans and flies [13] is still in its infancy, and the design of recent human studies could prove a useful guide to this work [78]. Although it will usually be impossible to know the regional recombination rate or genomic location of any given locus, judicious use of multilocus tests of neutrality and of recombination should advance the field considerably.

Acknowledgements

For providing helpful comments, we thank A. Di Rienzo, M. Hare, H.L. Gibbs, B. Jennings, C. Moritz, M. Nachman, R. Nielsen, P. Palsbøll, M. Slatkin, J. Wakeley and two anonymous reviewers. We thank L. Knowles for providing us with a prepublication copy of her manuscript. Work on this article was made possible in part by support from National Science Foundation grants DBI-9974235 (to R.T.B.), DEB 0108249 (to S.V.E. and P.B.), DEB 0129487 (to S.V.E.) and DEB9815650 (to J. Felsenstein); and National Institutes of Health grants HG-01436 and HL-66682 (to D.A.N.) and GM-51929 and HG-01989 (to J. Felsenstein).

References

1 Avise, J.C. (1994) *Molecular Markers, Natural History and Evolution*, Chapman & Hall
2 Emerson, B.C. *et al.* (2001) Revealing the demographic histories of species using DNA sequences. *Trends Ecol. Evol.* 16, 707–716
3 Hare, M.P. (2001) Prospects for nuclear gene phylogeography. *Trends Ecol. Evol.* 16, 700–706
4 Knowles, L.L. and Maddison, W.P. (2002) Statistical phylogeography. *Mol. Ecol.* 11, 2623–2635
5 Takahata, N. and Satta, Y. (2002) Pre-speciation coalescence and the effective size of ancestral populations. In *Modern Developments in Theoretical Population Genetics* (Slatkin, M. and Veuille, M., eds) pp. 52–71, Oxford University Press
6 Edwards, S.V. and Beerli, P. (2000) Perspective: gene divergence, population divergence, and the variance in coalescence time in phylogeographic studies. *Evolution* 54, 1839–1854
7 Rosenberg, N. and Nordborg, M. (2002) Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. *Nat. Rev. Genet.* 3, 380–390
8 Wakeley, J. and Hey, J. (1997) Estimating ancestral population parameters. *Genetics* 145, 847–855
9 Kuhner, M.K. *et al.* (2000) Usefulness of single nucleotide polymorphism data for estimating population parameters. *Genetics* 156, 439–447
10 Harpending, H.C. *et al.* (1998) Genetic traces of ancient demography. *Proc. Natl. Acad. Sci. U. S. A.* 95, 1961–1967
11 Nielsen, R. (2000) Estimation of population parameters and recombination rates using single nucleotide polymorphisms. *Genetics* 154, 931–942
12 Kliman, R.M. *et al.* (2000) The population genetics of the origin and divergence of the *Drosophila simulans* complex species. *Genetics* 156, 1913–1931
13 Machado, C.A. *et al.* (2002) Inferring the history of speciation from multilocus DNA sequence data: the case of *Drosophila pseudoobscura* and close relatives. *Mol. Biol. Evol.* 19, 472–488
14 Sunnucks, P. (2000) Efficient genetic markers for population biology. *Trends Ecol. Evol.* 15, 199–203
15 Collins, F.S. *et al.* (1998) A DNA polymorphism discovery resource for research on human genetic variation. *Genome Res.* 8, 1229–1231
16 International SNP Map Working Group, (2001) A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* 409, 928–933
17 Picoult-Newberg, L. *et al.* (1999) Mining SNPs from EST databases. *Genome Res.* 9, 167–174
18 Brookes, A.J. (1999) The essence of SNPs. *Gene* 234, 177–186
19 Sachidanandam, R. *et al.* (2001) A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* 409, 928–933
20 Lindblad-Toh, K. *et al.* (2000) Large-scale discovery and genotyping of single-nucleotide polymorphisms in the mouse. *Nat. Genet.* 24, 381–386
21 Hoskins, R.A. *et al.* (2001) Single nucleotide polymorphism markers for genetic mapping in *Drosophila melanogaster*. *Genome Res.* 11, 1100–1113
22 Cho, R.J. *et al.* (1999) Genome-wide mapping with biallelic markers in *Arabidopsis thaliana*. *Nat. Genet.* 23, 203–207
23 Wicks, S.R. *et al.* (2001) Rapid gene mapping in *Caenorhabditis*

elegans using a high density polymorphism map. *Nat. Genet.* 28, 160–164
24 Winzeler, E.A. *et al.* (1998) Direct allelic variation scanning of the yeast genome. *Science* 281, 1194–1197
25 Barnes, M.R. (2002) SNP and mutation data on the Web: hidden treasures for uncovering. *Comp. Funct. Genomics* 3, 67–74
26 Villablanca, F.X. *et al.* (1998) Invasion genetics of the Mediterranean fruit fly: variation in multiple nuclear introns. *Mol. Ecol.* 7, 547–560
27 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
28 Friesen, V.L. *et al.* (1999) PCR primers for the amplification of five nuclear introns in vertebrates. *Mol. Ecol.* 8, 2147–2149
29 Dietrich, W.F. *et al.* (1999) Identification and analysis of DNA polymorphisms. In *Genome Analysis: A Laboratory Manual* (Birren, B. *et al.*, eds), pp. 135–186, Cold Spring Harbor Laboratory Press
30 Karl, S.A. and Avise, J.C. (1993) PCR-based assays of Mendelian polymorphisms from anonymous single-copy nuclear DNA: techniques and applications for population genetics. *Mol. Biol. Evol.* 10, 342–361
31 Bradeen, J.M. and Simon, P.W. (1998) Conversion of an AFLP fragment linked to the carrot Y_2 locus to a simple, codominant, PCR-based marker form. *Theor. Appl. Genet.* 97, 960–967
32 McLenachan, P.A. *et al.* (2000) Markers derived from amplified fragment length polymorphism gels for plant ecology and evolution studies. *Mol. Ecol.* 9, 1899–1903
33 Fisher, R.A. (1934) The effects of methods of ascertainment upon the estimation of frequencies. *Ann. Hum. Genet.* 6, 13–25
34 Wakeley, J. *et al.* (2001) The discovery of single-nucleotide polymorphisms- and inferences about human demographic history. *Am. J. Hum. Genet.* 69, 1332–1347
35 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
36 Prodohl, P.A. *et al.* (1998) Genetic maternity and paternity in a local population of armadillos assessed by microsatellite DNA markers and field data. *Am. Nat.* 151, 7–19
37 Giordano, M. *et al.* (1999) Identification by denaturing high-performance liquid chromatography of numerous polymorphisms in a candidate region for multiple sclerosis susceptibility. *Genomics* 56, 247–253
38 Kruglyak, L. (1997) The use of a genetic map of biallelic markers in linkage studies. *Nat. Genet.* 17, 21–24
39 Macaubas, C. *et al.* (1997) The complex mutation pattern of a microsatellite. *Genome Res.* 7, 635–641
40 Hedrick, P.W. (1999) Perspective: highly variable loci and their interpretation in evolution and conservation. *Evolution* 53, 313–318
41 Charlesworth, B. (1998) Measures of divergence between populations and the effect of forces that reduce variability. *Mol. Biol. Evol.* 15, 538–543
42 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
43 Wang, D.G. *et al.* (1998) Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science* 280, 1077–1082
44 Congdon, B.C. *et al.* (2000) Mechanisms of population differentiation in marbled murrelets: historical versus contemporary processes. *Evolution* 54, 974–986
45 Pluzhnikov, A. and Donnelly, P. (1996) Optimal sequencing strategies for surveying molecular genetic diversity. *Genetics* 144, 1247–1262
46 Nachman, M.W. (2001) Single nucleotide polymorphisms and recombination rate in humans. *Trends Genet.* 17, 481–485
47 Harris, E.E. and Hey, J. (1999) X chromosome evidence for ancient human histories. *Proc. Natl. Acad. Sci. U. S. A.* 96, 3320–3324
48 Kaessmann, H. *et al.* (1999) Extensive nuclear DNA sequence diversity among chimpanzees. *Science* 286, 1159–1162
49 Yu, N. *et al.* (2001) Global patterns of human DNA sequence variation in a 10-kb region on chromosome 1. *Mol. Biol. Evol.* 18, 214–222
50 Aquadro, C.F. *et al.* (1994) Selection, recombination and DNA polymorphism in *Drosophila*. In *Non-Neutral Evolution: Theories and Molecular Data* (Golding, B. *et al.*, eds), pp. 46–56, Chapman & Hall
51 Lercher, M.J. and Hurst, L.D. (2002) Human SNP variability and mutation rate are higher in regions of high recombination. *Trends Genet.* 18, 337–340
52 Nei, M. (1987) *Molecular Evolutionary Genetics*, Columbia University Press
53 Hare, M.P. *et al.* (2002) Genetic evidence on the demography of speciation in allopatric dolphin species. *Evolution* 56, 804–816
54 Watterson, G.A. (1975) On the number of segregating sites in genetical models without recombination. *Theor. Popul. Biol.* 7, 256–276
55 Friesen, V.L. *et al.* (1997) Intron variation in marbled murrelets detected using analysis of single-stranded conformational polymorphism. *Mol. Ecol.* 6, 1047–1058
56 Shi, L. *et al.* (2001) Comparative DNA sequence analysis of genetic variation in the African grey parrot, *Psittacus erythacus*. *Genetica* 110, 227–230
57 Nei, M. and Li, W. (1979) Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proc. Natl. Acad. Sci. U. S. A.* 76, 5269–5273
58 Schneider, K. *et al.* (2001) SNP frequency and allelic haplotype structure of *Beta vulgaris* expressed genes. *Mol. Breed.* 8, 63–74
59 Koufopanou, V. *et al.* (1997) Concordance of gene genealogies reveals reproductive isolation in the pathogenic fungus *Coccidioides immitis*. *Proc. Natl. Acad. Sci. U. S. A.* 94, 5478–5482
60 Ewing, B. *et al.* (1998) Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res.* 8, 175–185
61 Gordon, D. *et al.* (1998) Consed: a graphical tool for sequence finishing. *Genome Res.* 8, 195–202
62 Marth, G.T. *et al.* (1999) A general approach to single-nucleotide polymorphism discovery. *Nat. Genet.* 23, 452–456
63 Nickerson, D.A. *et al.* (1997) PolyPhred: automating the detection and genotyping of single nucleotide substitutions using fluorescence-based resequencing. *Nucleic Acids Res.* 25, 2745–2751
64 Rieder, M.J. *et al.* (1998) Automating the identification of DNA variations using quality-based fluorescence re-sequencing: analysis of the human mitochondrial genome. *Nucleic Acids Res.* 26, 967–973
65 Buetow, K.H. *et al.* (1999) Reliable identification of large numbers of candidate SNPs from public EST data. *Nat. Genet.* 21, 323–325
66 Gasper, J.S. *et al.* (2001) Songbird genomics: analysis of 45 kb upstream of a polymorphic MHC class II gene in red-winged blackbirds (*Agelaius phoeniceus*). *Genomics* 75, 26–34
67 Ortí, G. *et al.* (1997) Detection and isolation of nuclear haplotypes by PCR-SSCP. *Mol. Ecol.* 6, 575–580
68 Hare, M.P. and Palumbi, S.R. (1999) The accuracy of heterozygous base calling from diploid sequence and resolution of haplotypes using allele-specific sequencing. *Mol. Ecol.* 8, 1750–1752
69 Clark, A.G. (1990) Inference of haplotypes from PCR-amplified samples of diploid populations. *Mol. Biol. Evol.* 7, 111–122
70 Excoffier, L. and Slatkin, M. (1995) Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Mol. Biol. Evol.* 12, 921–927
71 Stephens, M. *et al.* (2001) A new statistical method for haplotype reconstruction from population data. *Am. J. Hum. Genet.* 68, 978–989
72 Kuhner, M.K. and Felsenstein, J. (2000) Sampling among haplotype resolutions in a coalescent-based genealogy sampler. *Genet. Epidemiol.* 19, S15–S21
73 Nielsen, R. and Signorovitch, J. Correcting for ascertainment biases when analyzing SNP data: applications to the estimation of linkage disequilibrium. *Theor. Popul. Biol.* (in press)
74 Kuhner, M.K. *et al.* (2000) Maximum likelihood estimation of recombination rates from population data. *Genetics* 156, 1393–1401
75 Klicka, J. and Zink, R.M. (1997) The importance of recent ice ages in speciation: a failed paradigm. *Science* 277, 1666–1669
76 Brumfield, R.T. *et al.* (2001) Evolutionary implications of divergent clines in an avian (*Manacus*: Aves) hybrid zone. *Evolution* 55, 2070–2087
77 Pluzhnikov, A. *et al.* (2002) Inferences about human demography based on multilocus analyses of noncoding sequences. *Genetics* 161, 1209–1218
78 Przeworski, M. *et al.* (2000) Adjusting the focus on human variation. *Trends Genet.* 16, 296–302