The utility of single nucleotide polymorphisms in inferences of population history

Robb T. Brumfield  
*Louisiana State University*

Peter Beerli  
*University of Washington, Seattle*

Deborah A. Nickerson  
*University of Washington, Seattle*

Scott V. Edwards  
*University of Washington, Seattle*

Follow this and additional works at: https://repository.lsu.edu/biosci_pubs

**Recommended Citation**


This Article is brought to you for free and open access by the Department of Biological Sciences at LSU Scholarly Repository. It has been accepted for inclusion in Faculty Publications by an authorized administrator of LSU Scholarly Repository. For more information, please contact ir@lsu.edu.
The utility of single nucleotide polymorphisms in inferences of population history

Robb T. Brumfield¹, Peter Beerli², Deborah A. Nickerson² and Scott V. Edwards³

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

Ascertertainment 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.

Diploidy: 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 (4Nₑμ).
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.

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

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

*H. Kim et al., unpublished.

*aAbbreviations: EST, expressed sequence tag; NA, not available; Nt, nucleotide; SNP, single nucleotide polymorphism; STS, sequence tagged site.

http://tree.trends.com
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 Polyclad (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 Polyclad, have been used in only a handful of non-model species (e.g. red-winged blackbird [66]).

http://tree.trends.com
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 [89–71]. 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.

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 SNP's are measured on the same mutational scale as mitochondrial (mt) SNPs (substitutions per site), making intergenic 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

SNP's offer historical demography a great opportunity for unbiased sampling of loci. As an extension of recent studies...
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 differentiationa

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>SNPsb</th>
<th>Msats</th>
<th>AFLPs</th>
<th>RFLPs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Within-population variation</td>
<td>Low</td>
<td>High</td>
<td>Unknown</td>
<td>Low</td>
</tr>
<tr>
<td>Among-locus variation</td>
<td>Low</td>
<td>High</td>
<td>Unknown</td>
<td>Low</td>
</tr>
<tr>
<td>Conducive to unbiased sampling among loci?</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Susceptibility to molecular convergence</td>
<td>Low</td>
<td>High</td>
<td>Unknown</td>
<td>Low</td>
</tr>
<tr>
<td>Variation easy to interpret?</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Moderate</td>
</tr>
<tr>
<td>Are the variations detected further reducible?</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Easy comparison with mitochondrial DNA?</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
</tbody>
</table>

aTraits in bold indicate suggested advantages. Differences among markers in ease of assaying variation or of laboratory isolation are not considered.

bAbbreviations: 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 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.

**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 demographics, 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 for 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.

http://tree.trends.com
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 mutagenetic 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.

Acknowledgements
For providing helpful comments, we thank A. Di Rienzo, M. Hare, H.L. Gibbs, B. Jennings, C. Moritz, M. Nachman, R. Nielsen, P. Parsch, 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-9874235 (to R.T.B.), DEB 0108249 (to S.V.E. and P.B.), DEB 0129487 (to S.V.E.) and DEB 98115650 (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
17 Picout-Newberg, L. et al. (1999) Mining SNPs from EST databases. Genome Res. 9, 167–174

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. Piuzechnikov 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 1). 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 unparentally 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.
41 Wang, D.G. et al. (1998) Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. Science 280, 1077–1082
56 Schneider, K. et al. (2001) SNP frequency and allelic haplotype structure of Beta vulgaris expressed genes. Mol. Breed. 8, 63–74
63 Bu et al., K.H. et al. (1999) Reliable identification of large numbers of candidate SNPs from public EST data. Nat. Genet. 21, 323–325

http://tree.trends.com