Genetica (2009) 135:439–455
DOI 10.1007/s10709-008-9293-3

Multilocus phylogeography and phylogenetics using sequence-based markers

Patrícia H. Brito · Scott V. Edwards

Received: 9 April 2008 / Accepted: 28 June 2008 / Published online: 24 July 2008

Abstract We review recent trends in phylogeography and phylogenetics and argue that these two fields stand to be reunited by the common yardstick provided by sequence and SNP data and by new multilocus methods for phylogenetic analysis. Whereas the modern incarnation of both fields was spawned by PCR approaches applied to mitochondrial DNA in the late 1980s, the two fields diverged during the 1990s largely due to the adoption by phylogeographers of microsatellites, in contrast to the adoption of nuclear sequence data by phylogeneticists. Sequence-based markers possess a number of advantages over microsatellites, even on the recent time scales that are the purview of phylogeography. Using examples primarily from vertebrates, we trace the maturation of nuclear gene phylogeography and phylogenetics and suggest that the abundant instances of gene tree heterogeneity beckon a new generation of phylogenetic methods that focus on estimating species trees as distinct from gene trees. Whole genomes provide a powerful common yardstick on which both phylogeography and phylogenetics can assume their proper place as ends of a continuum.

Keywords Multilocus analysis · Indel · Intron · Genomic phylogeography · SNP · Species tree

Invited review for Genetica (ESF Workshop on Ecological Genomics, Tällberg, Sweden 11/07).

P. H. Brito · S. V. Edwards (✉)
Department of Organismic and Evolutionary Biology, Harvard University, 26 Oxford Street, Cambridge, MA 02138, USA
e-mail: sedwards@fas.harvard.edu

P. H. Brito
Instituto Gulbenkian de Ciência, Rua Quinta Grande, 6, Apartado 14, Oeiras 2781-901, Portugal
e-mail: pbrito@oeb.harvard.edu

Introduction

These are exciting times to be working in the fields of phylogeography and phylogenetics. We are witnessing an important revolution in the way data are collected as genomic tools are being transferred from studies on model organisms to studies focused on evolutionary or ecological models or on species of special conservation priority. As genomic approaches become cheaper and sequencing technologies allow more efficient surveys, it will soon be feasible to collect whole genome information for population samples of single species as well as genomic data for phylogenetic studies with dense taxonomic sampling. This advance comes as a long-awaited goal of the fields of phylogeography and phylogenetics as researchers strive for larger molecular data sets in order to address complicated problems in population history, demography and speciation. However, the availability of these large amounts of data also raises analytical challenges and encourages the development of new methods of analyses. Here we review these challenges in the context of their evolving fields. Due to our expertise, we emphasize vertebrate examples, highlighting results from invertebrates and plants, particularly model species, where applicable.

Transition from single to multiple loci: phylogeography

Phylogeography and phylogenetics may be seen as part of a continuum that crosses the species boundary (Avise et al. 1987). These two fields, however, have traditionally asked different questions and used different methodologies, in particular with reference to sampling methods: phylogeography is concerned with the analyses of evolutionary processes that occur at the population level for which extensive population sampling has always been advocated, whereas phylogenetics is interested in determining species or sister-clade relationships for which multiple sampling within species is, in general, considered less relevant. For this reason, it is not surprising that major methodological shifts, being those technological or analytical, would be experienced differently and at different times in each of these fields (Fig. 1).

With the widespread use of direct sequencing, mitochondrial DNA became the marker of choice for phylogeographic studies (Wilson et al. 1985). mtDNA typically has high evolutionary rates attributed to an inefficient mutation repair mechanism (Brown et al. 1979) that leads to on average high information content per base pair sequenced. Also, mtDNA is generally transmitted exclusively through the maternal line and hence genes in this molecule are single copy and do not generally undergo recombination (although see Ladoukakis and Zouros 2001a, b for recent, well-documented exceptions in animals). These features were considered ideal for the purpose of building gene genealogies that were used to infer the recent history of populations and species and to estimate population parameters associated with speciation, such as bottlenecks. The conceptual and technical simplicity of the phylogeographic approach facilitated the spread of the field and, early on, population subdivision and population structure were recognized by strong genealogical structure on intraspecific mtDNA phylogenies for a vast array of taxa (Fig. 2, Avise 2000).

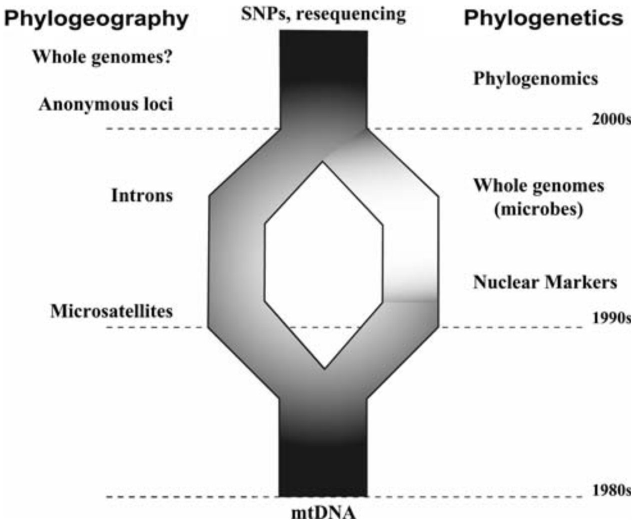

Fig. 1 Methodological shifts and evolution of the fields of phylogeography and phylogenetics. The figure depicts the original unity of these two fields conferred by a focus on mtDNA in the 1980 s. During the 1990 s, however, although mtDNA was still used in both phylogeography and phylogenetics, these fields somewhat separate, in part because of the disparate nuclear markers used by each field. With the advent of whole genomes and increasing access to the nuclear genomes of non-model species after the late 1990 s, phylogeography and phylogenetics stand to be reunited by the common yardstick provided by DNA sequences and SNP data

Simulations and empirical observations of the behavior of mitochondrial gene trees within populations (Avise 2000), as well as the rise of coalescent theory (Kingman 1982a, b), drove home the fact that gene trees are expected to differ, sometimes substantially, for each sampled locus. With the new awareness brought on by large scale phylogeographic analyses and gene tree heterogeneity, as well as early investigations of gene trees in the nuclear genome (e.g. Palumbi and Baker 1994; Hare and Avise 1998), criticisms of both the empirical and analytical aspects of single locus phylogeography became more common. These criticisms have drawn attention to the tendency of researchers to over-interpret single gene trees and to ignore coalescent stochasticity (Knowles and Maddison 2002). They have also highlighted the dependency of confidence intervals associated with estimates of demographic parameters, such as effective population sizes, divergence times, rates of gene flow, on the number of loci (Kuhner et al. 1998; Beerli and Felsenstein 1999). In the early 1990s, in anticipation of larger multilocus data sets, genealogical concordance across independent loci was advocated as a means of distinguishing between spurious phylogeographic breaks and true vicariant events, under the assumption that similar gene genealogies estimated from multiple loci must arise from a common historical event rather than from arbitrary divisions within populations (Avise and Ball 1990). Such an approach was also most appropriate when ancestral polymorphism was low or absent. Although the multilocus perspective envisioned by Avise and Ball was articulated nearly 20 years ago, assessing genealogical concordance in multi-locus studies is still a challenge in phylogeographic studies of non-model organisms.

Also common was the criticism that the exclusively maternal inheritance of mtDNA renders population inferences based solely on this marker unrepresentative of the whole genome. This criticism becomes most apparent when patterns of mtDNA variation can be contrasted directly with other haploid chromosomes, such as the Y chromosome in mammals, or with autosomes. While studies employing mtDNA and autosomal markers such as microsatellites are routine today, assessing male and female dispersal patterns through comparisons with sex chromosome variation is less common outside of studies on humans, mostly because the relevant sex-linked markers are not widely available for non-model species. By contrast, in studies on humans, Y-chromosome variation has yielded detailed comparisons with mtDNA, revealing systematic differences between the sexes in dispersal and population structure (e.g. Wilder et al. 2004). Nonetheless, together these criticisms accelerated the transition from

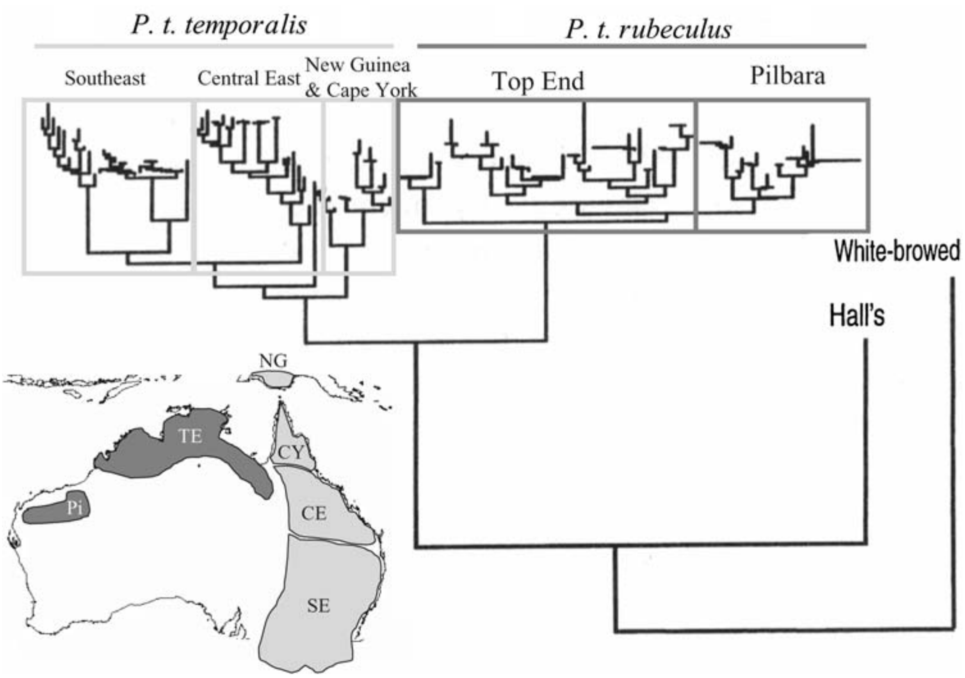

Fig. 2 Example of a mitochondrial phylogeographic study. Here population structure is inferred from strong genealogical structure on intraspecific mtDNA gene tree. Figure adapted from Edwards (1993). SE, Southeast; CE, Central East; CY, Cape York; NG, New Guinea; TE, Top End; Pi, Pilbara

single-locus to multilocus phylogeography in all taxonomic groups, and helped intensify the search for versatile markers in the nuclear genome.

Transition to multilocus phylogenetics

In phylogenetics the transition from single locus to multilocus analyses had different causes than for phylogeography, although many of the early debates also emerged from studies on mtDNA. The overall trend toward amassing larger data sets—particularly once entire mitochondrial genomes were accessible—was driven by the recognition that larger datasets would include more informative sites that would increase nodal support. One impetus to move beyond single locus phylogenetics emerged from debates on ‘total evidence’ and on whether it was better to combine different sources of data resultant either from different genes or from molecular and non-molecular datasets (morphology, behavior, etc.) (e.g. Kluge 1989; Bull et al. 1993). At the center of discussions about combining different molecular data sets was the conclusion that different genes could evolve in such radically different ways, and that their phylogenetic signals and substitution processes could be so divergent, that analyzing such partitions together could be analytically challenging (Bull et al. 1993). Such fears have been partly allayed in recent years by sophisticated methods for partitioning data and allowing models of nucleotide substitution to vary between partitions to enhance phylogenetic signal (Nylander et al. 2004). There was also the acknowledgement that gene trees could differ from one another, particularly in cases of rapid speciation, such situations were raised as a challenge to phylogenetics and in such cases combining data was discouraged (Bull et al. 1993). For example, mitochondrial paraphyly appears to be quite common (Funk and Omland 2003) and these situations will guarantee heterogeneity among nuclear gene trees.

The heterogeneity of the substitution process across genes was seen by some as a boon for phylogenetic analysis. Initial calls for using multiple loci for a particular phylogenetic problem came largely from a demonstrated superiority of data sets that encompassed different mutation rates and substitution patterns, each containing information for resolution of different levels of the tree (Cummings et al. 1995; Otto et al. 1996). By analyzing such markers in concatenated datasets, it was hoped that the potential conflict across loci would cancel out and that loci with different rates would resolve different portions of the clade of interest. These discussions did not specifically address the issue of genetic independence of the loci being studied, since heterogeneity of gene trees was generally considered minor. Instead these early discussions focused on variation in sequence substitution processes, and in many cases focused the discussion on mtDNA and assumed complete linkage of the loci being studied (Cummings et al. 1995).

The link between more genes, longer sequences, and increased nodal support was inherent in the practice of concatenation, which tacitly places a single gene tree on all genes, whether emanating from the nuclear or organelle genomes. Indeed, as phylogeneticists reached into the nuclear genome, the problem that gene trees may fundamentally differ from one another or from the species tree was tacitly resolved by assuming that common signals would emerge from concatenated datasets, and that the extent of gene tree heterogeneity was minor enough so as not to compromise phylogenetic analysis. Although we now know that concatenation can positively mislead phylogeny in some situations (Kubatko and Degnan 2007), concatenation is probably reasonable over a wide range of shapes and depths of phylogenetic trees, particularly when dealing with phylogenies characterized by long branches as measured in coalescent units (Fig. 3). In the phylogenetic tradition, those instances in which heterogeneity of gene trees for a particular data set was investigated, the discordance found was often ascribed not to independent realizations of the lineage sorting process but rather to sampling error at the level of nucleotides and nucleotide substitution models, an assumption inherent in most tests of congruence among data sets, such as likelihood ratio tests or the ILD test (Huelsenbeck 1996; Leigh et al. 2008). In most analyses, even today (except perhaps for microbial studies), the tacit assumption is that all genes do in fact have the same gene tree, that these gene trees are congruent with and converge on the species tree, or that concatenation will iron-out any inconsistencies in phylogenetic signal among genes (e.g. Rokas et al. 2003). Many researchers have found that the number of informative sites in any individual nuclear gene is small and its phylogenetic resolution poor, and have resolved to increase this resolution at the level of gene trees by concatenating multiple genes together.

Truly genomic approaches to phylogenetics (‘phylogenomics’) were driven by whole genome sequences of model organisms and have been progressively extended to other taxonomic groups due in part to the significant investment on tree-of-life initiatives (http://www.nsf.gov/pubs/2002/nsf02074/nsf02074.html). This new wave of molecular data raised new analytical challenges and has highlighted the need for new methods of analyses. One consequence of the using large numbers of independent loci in phylogenomics has been an increasing appreciation of the heterogeneity of gene trees, whether due to coalescent variation or to taxonomically less widespread causes such as horizontal gene transfer (Delsuc et al. 2005). Although the power of phylogenomics datasets has been impressive, they have lead to an increasing avoidance of the assumptions embodied in concatenation and of the sufficiency of single trees to describe all components of a genome (Rokas et al. 2003; Pollard et al. 2006). On a smaller and lower taxonomic scale, multilocus phylogenetic datasets have begun to emerge for non-model species and here, too, it has become obvious in many cases that assuming a single tree for all genes investigated is mistaken (Tosi et al. 2003; Bensch et al. 2006). However, even for non-model species, the impact of new sequencing technologies will be profound. Both Solexa and 454 sequencing approaches are beginning to be applied to non-model species (Binladen et al. 2007; Meyer et al. 2007; Vera et al. 2008), and these approaches will not only increase by several orders of magnitude the number of loci and

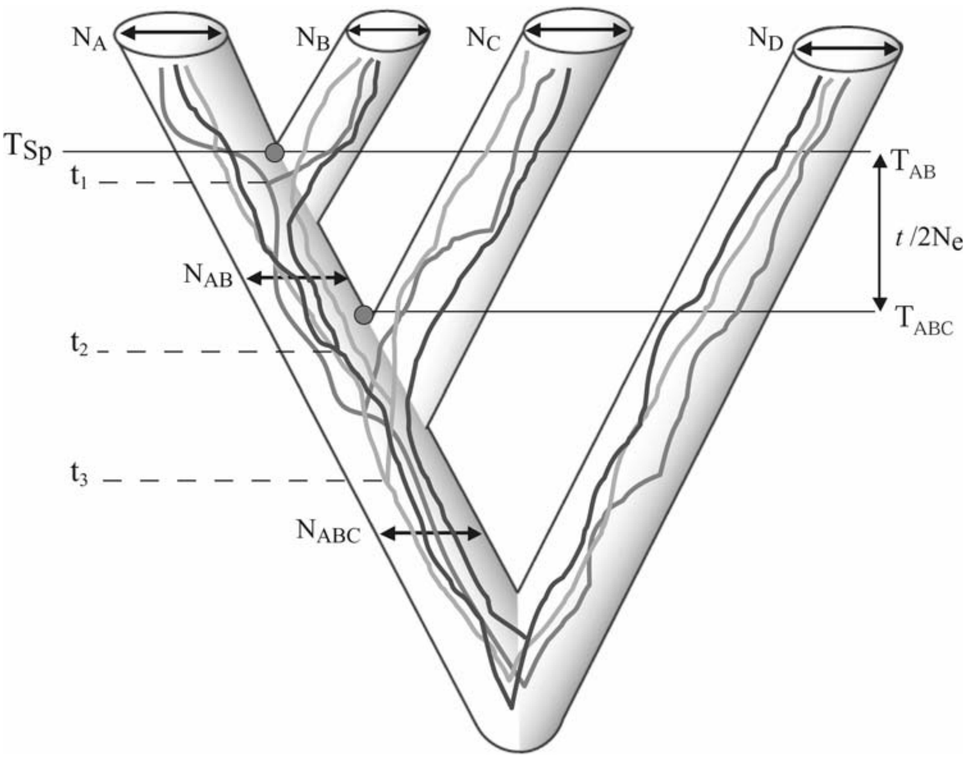

Fig. 3 Contrast between species tree and gene trees. This figure depicts important lessons and parameters derived from population-level studies and coalescent theory that are being used in field of phylogenetic inference of species trees. The figure shows that the stochastic sorting of alleles due to drift is dependent on both the effective population sizes of current (N_A – N_D) and ancestral populations (N_{AB} and N_{ABC}) and the species tree branch lengths, i.e., the times between the internodes (T_{ABC}–T_{AB}). Also, gene trees often differ from the species tree in topology and in branch lengths. T_{Sp} and dots represents speciation time between species A and B, and t_1, t_2, t_3 represent divergence times for genes sampled from species A–C

individuals that can be studied, but, because of their focus on single molecule sequencing, will avoid some of the problems with haplotype estimation that currently plague data sets produced by traditional dideoxy sequencing of diploid PCR products.

Multilocus phylogeography: state of the art

The rise of multilocus analyses is deeply embedded in the development of in vitro amplification methods (PCR) and, later, genomic approaches, as first applied in model organisms. As experimental approaches to the development of new markers become easier, researchers increasingly developed new markers for specific projects, rather than using primers designed from other species, such as the 'universal' cytochrome *b* primers of Kocher et al. (1989); (see also Zhang and Hewitt 2003). Such a 'bottom up' approach was especially useful when one preferred to emphasize a particular class of markers, such as microsatellites, but it is gradually being eroded as whole genomes become available for a greater diversity of representatives from the Tree of Life**.** Whole genomes make primer design for the focal species trivial, hence it is not surprising that many of the first multilocus phylogeographic studies were carried out with humans (e.g. Harpending et al. 1998; Pluzhnikov et al. 2002), *Drosophila* (e.g. Wang et al. 1997; Machado et al. 2002), *Arabidopsis* (e.g. Innan and Stephan 2000) and other model species (Whitfield et al. 2006; Hernandez et al. 2007). The availability of whole genomes and large-scale SNP surveys performed on model organisms promote the use of these resources among related species and clades. In primates, for example, the availability of the human and chimp genomes promoted multilocus phylogeographic analyses of other great apes (e.g. Osada and Wu 2005; Thalmann et al. 2006). Additionally whole genomes allow precise positioning of target loci relative to one another, in order to incorporate linkage into phylogeographic studies. Thus in birds, the availability of the chicken genome and the trace archive of the zebra finch genome has led to the design of more than 200 PCR primer pairs targeted toward conserved exons flanking introns evenly spaced throughout the avian genome, a major resource for avian population genetics (Backström et al. 2008).

Sequence-based markers

In an early review of the field of nuclear gene phylogeography, Zhang and Hewitt (2003) suggested that the success of the genetic studies of natural populations was in large part due to the widespread use of mtDNA and microsatellites. The former enabled the genealogical approach and the application of coalescent and phylogenetic tools for population-level questions, and the later made possible multilocus inferences of highly polymorphic markers. However, these authors recognized that important limitations of these two marker classes slowed the development of the field and that further developments waited for the development of new markers. In particular, they and others (Zink and Barrowclough 2008) have pointed out that microsatellites are challenging to study due to their complex mutation process; microsatellite mutation rates may vary across loci and across alleles within the same locus, and comparative analyses across taxa are often made impractical due to mutation bias and different rates of evolution among lineages (Rubinsztein et al. 1999; Ellegren 2004). Extremely high evolutionary rates, size homoplasy and other genotyping artifacts (e.g. null alleles and allele dropouts) also pose severe limitations to the analysis of microsatellites (Hedrick 1999; Rubinsztein et al. 1999). It is also the case that microsatellites often use statistics that cannot be directly compared with those used to describe mitochondrial DNA; even when parameters such as divergence time (τ) and population size (θ) are estimated using microsatellites, these parameters will have different units than those for mtDNA, based as they are on allele arrays rather than nucleotides; by contrast, sequence-based loci and mtDNA can be analyzed with exactly the same statistics with the same units, keeping in mind of course the difference in effective population size and mutation rate (see below). Finally, although the hypervariability of microsatellites was for a long time considered a boon, it was also recognized that such high variation only compromised the estimation of some basic demographic statistics, such as F_{ST}, which becomes less meaningful when the denominator (which usually includes intrapopulation variability) is high (Hedrick 1999, 2005). Furthermore, although selection can be studied with microsatellites, particularly when linked relationships among loci are considered, in general microsatellites do not lend themselves well to testing for the effects of natural selection, and there are now many more tests of selection for coding or noncoding DNA sequence than for microsatellites. In part due to the onslaught of genomic information, and the increasingly realistic goal of comparing phylogeographic histories on a common genome-wide scale, it has become evident that comparisons between different genomic regions, genes, and species are much easier with DNA sequence data that can be aligned across multiple hierarchical levels.

Nuclear coding sequences (or exons) typically show low levels of intraspecific variation, and due to the existence of many other appropriate classes of nuclear markers, they are rarely used for population inference. Among the classes of non-coding nuclear DNA, introns are now popular in multilocus phylogeographic analyses (e.g. Palumbi and Baker 1994; Friesen et al. 1997; Bensch et al. 2006). Taking advantage of the fact that introns sit between two highly conserved regions, many researchers have adopted the exon- primed – intron crossing PCR method (Palumbi 1996) to design primers based on exon data available on Genbank. This approach has been responsible for the collection of nuclear DNA across many taxonomic groups that span invertebrates, vertebrates, as well plants (see Hare 2001 and references therein). One advantage of using highly conserved primers is the assurance of amplifying orthologous copies, an issue that may be critical when multiple gene copies are a concern. Also, screening the same loci across many studies will enable comparative analysis of sequence variation that is useful for rate calibrations and comparative phylogeography. However, conserved primers may have the drawback of leading to loci that are adjacent to and tightly linked with sites under directional or background selection, or which may evolve slowly or violate neutrality assumptions inherent in many types of statistical analyses.

SNPs and resequencing approaches

Many authors have suggested that single nucleotide polymorphisms (SNP) will become the marker of choice for multilocus population analyses, whether as unlinked sites interrogated individually or as sets of linked sites whose variation is studied via resequencing (e.g. Brumfield et al. 2003; Zhang and Hewitt 2003). SNPs have simple patterns of variation and the potential for automated detection, but more importantly, SNPs have low mutation rates (10^{-8} to 10^{-9}, Brumfield et al. 2003) and thus low levels of homoplasy. In general large-scale SNP surveys have shown considerable promise for revealing fine-scale population history, for example in humans (Shriver et al. 2004, 2005). Recently the Illumina platform (Fan et al. 2003) was used to characterize over 1,000 SNP genotypes in honeybees that had been detected in trace archives of the genome and EST databases. This variation was used to unravel their phylogeographic history in remarkable detail (Whitfield et al. 2006). However, because SNPs usually are biallelic, each locus only defines one bipartition in the dataset and thus many SNPs are necessary to produce robust inferences of population history. Individual SNPs do not easily yield a result that is visualized by the locus-by-locus contrast of gene trees that was primed by the mitochondrial era of phylogeography. As a result, recent large-scale SNP analyses of phylogeography have utilized analytical approaches not based on gene trees. Rather, they have built population trees directly from SNP frequencies (as in studies on honeybees, Whitfield et al. 2006), or used, for example, the site frequency spectrum of the SNPs analyzed, particularly when estimating demographic parameters (Hernandez et al. 2007). SNPs also lend themselves well to classification approaches such as principle components or assignment clustering approaches such as STRUCTURE (Pritchard et al. 2000). The taxonomic applicability of large-scale SNP approaches is still restricted to close relatives of model organisms for which large-scale genome-sequencing projects are available, but new sequencing technologies will certainly make these markers a more viable option for studies of natural populations.

As an alternative to genotyping unlinked SNPs, as in the honeybee example, researchers have characterized stretches of DNA sequence, often anonymous in nature, with multiple and linked SNPs. Remarkably, the idea of using nuclear sequence variation, particularly of anonymous regions, was originally put forward as early as 1993 by Karl and Avise (1993), giving anonymous loci almost as much of an opportunity as microsatellites to become a workhorse for phylogeography. However, even by that early date, microsatellites had become available for a limited number of species (e.g. Tautz 1989; Jorde et al. 1997) and the seductiveness of their hypervariability caused Karl and Avise's approach to be ignored. Recently a number of studies have utilized anonymous regions of the genome to infer phylogeographic history with some success, and their results illustrate the advantages of linked SNPs over individual SNPs (e.g. Ross and Harrison 2002; Dettman et al. 2003; Jennings and Edwards 2005; Carstens and Knowles 2007b). Such an approach has also been used on a genome-wide scale (Hernandez et al. 2007). For example, gene tree heterogeneity can be readily visualized with anonymous loci or introns, as can variation in rates among loci (Fig. 4). The use of anonymous loci allows markers to be selected without reference to their polymorphism, a feature that some workers consider essential for providing an unbiased description of genomic variation (Brumfield et al. 2003). By contrast, most studies involving microsatellites deliberately throw out loci with low polymorphism so as to amplify the signal found within each locus. Such a practice certainly results in more information per locus, but also results in a biased view of variation whose relevance to the rest of the genome is unclear. Discarding low polymorphism loci amounts to ascertainment bias, a bias that can be addressed (e.g. Rosenblum and Novembre 2007), but which is generally recognized as a caveat in many studies that actively select the loci they study based on information from a subset of study organisms or loci (Nielsen et al. 2004).

In large-scale studies, whether researchers genotype SNPs individually or resequence homologous regions among individuals, the most recent wave of genome-scale SNP analyses reveal that the gene tree approach favored by phylogeographers for decades is now giving way to

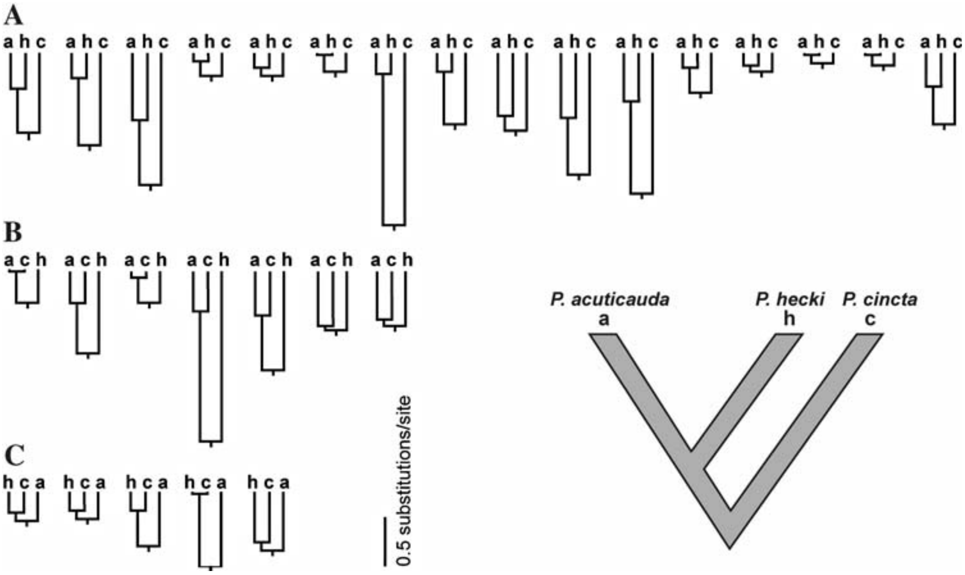

Fig. 4 Example of the heterogeneity in topology and coalescent times that can be recovered in multilocus phylogeography when several independent loci are sampled. Letters a, h, and c on gene trees (A, B, and C) correspond to *Poephila acuticauda*, *hecki*, and *cincta* as illustrated in the species tree at lower right. This figure was adapted from Jennings and Edwards (2005). Tree depicted on the lower right corner is the species tree as inferred from these data by Liu and Pearl (2007)

approaches informed by coalescent theory, but nonetheless abandoning the gene tree paradigm, primarily because recombination is viewed as a major roadblock to building gene trees. For example, Hernandez et al. (2007) used the rhesus macaque genome to sequence 166 regions in three populations of macaques totaling over 150 kb per individual. Yet they resisted making trees of these regions, favoring approaches that modeled the site frequency spectra of SNPs within these resequenced regions. It is unclear what the relative power of non-gene tree vs. gene tree approaches are, yet the pervasive issue of recombination cannot be ignored and may ultimately result in simulation approaches aimed at summary statistics such as the site frequency spectrum. Additionally, there have been few discussions of the relative merits of linked vs. unlinked SNPs in phylogeographic inference. Regardless, it is clear that patterns of linkage disequilibrium can provide important information on the phylogeographic history of populations (e.g. Voight et al. 2005), and that such history can be gleaned without making gene trees. Although many studies working on non-model species have utilized nuclear gene trees to inform their analyses, those working on model systems are often bypassing such approaches in favor of site-by-site analyses.

Indels

Insertion-deletion polymorphisms were recently proposed as useful genetic markers for studying natural populations (Väli et al. 2008). Genome-wide surveys across several taxonomic groups have revealed that indels are common in humans (Mills et al. 2006), *Drosophila* (Ometto et al. 2005), chicken (Brandström and Ellegren 2007), and canids (Väli et al. 2008). Väli and collaborators (Väli et al. 2008) using data from several breeds of dogs and wolf populations compared levels of heterozygosity between indels and microsatellites and found a strong correlation between the two, although, not surprisingly, levels of diversity at indels were substantially lower than for microsatellites. The advantage of using indels as molecular markers to study natural populations relies on them being relatively common and prone to easy methods of survey. However, the molecular evolution of length polymorphisms remains to be fully understood although the current progress in the development of explicit models of indel evolution (e.g. Keightley and Johnson 2004; Miklos et al. 2004) is already making a clear impact in the field of multilocus phylogenetics (Benavides et al. 2007). The recent interest in characterizing indel polymorphism in natural populations, as well as the clear utility of retroposons as intraspecific markers of ancestry (Boissinot et al. 2000; Shedlock et al. 2004; Konkel et al. 2007), will provide invaluable insight into population history and the evolutionary mechanisms of insertion and deletions.

Mutation rates and polymorphism

Initial concerns with the use of nuclear loci to study natural populations were fueled by the perception that low substitution rates would not yield enough polymorphism to infer robust gene trees or to estimate demographic parameters at the population level. Low substitution rate minimizes back mutations and thus homoplasy in the datasets, but at the cost of important phylogeographic signals being represented by only a few mutations. Whenever clades are defined by only a few diagnostic sites, traditional methods of tree inference may not unambiguously resolve haplotype relationships, and measures of nodal support in gene trees would not produce elevated statistics.

It is well known that, in many lineages, substitution rates in mtDNA are substantially higher than for nuclear DNA. Based on the expected correlation between substitution rates (D), as measured by $D = \mu t$ ($\mu =$ mutation rate, t is divergence time), it is widely believed that levels of polymorphism ($\theta = 2N_f\mu$, where N_f is the female effective population size) in mtDNA will be higher than those of nuclear DNA ($\theta = 4\ N\mu$, where N is the effective size of both sexes) by the same proportion as are substitution rates. But there are important reasons why this latter hypothesis might not be true, either due to natural selection on mtDNA, or the much smaller effective population size of organelle genomes compared to autosomal loci, and there has been no meta-analysis documenting this trend systematically for nucleotide diversity in nuclear and organelle genomes. Recently Bazin and collaborators (Bazin et al. 2006) showed across several thousand species that as population size increased, heterozygosity in allozymes continued to increase whereas mtDNA diversity plateaued, suggesting that natural selection routinely places a ceiling on mitochondrial diversity. In plants, rates of nucleotide substitution are often faster in nuclear loci than in the two organelle genomes (Wolfe et al. 1987), and in fruit flies the rates seem not to differ substantially (Caccone et al. 1988; see Table 1 in Zhang and Hewitt 2003). A genomic survey of intraspecific variation from samples of natural population of a urochordate show extremely high polymorphism that was attributed to a large effective population size (Small et al. 2007). Polymorphism across different regions of the nuclear genome may vary greatly due to factors such as recombination, functional constraints, or levels of selection (Nachman 2001), and the precise ratio of nuclear to mitochondrial diversity in nature will be an important topic for future study. Some studies, for example in lizards, have shown that interspecific divergence in nuclear genes is often less than that for mtDNA, even when intraspecific diversity for nuclear genes is high, a result that might implicate selection on mtDNA within species (Dolman and Moritz 2006; Rosenblum et al. 2007).

Statistical analysis of multilocus data

Phylogeographic analyses have come a long way since the first inferences of population history based on single gene trees. The concept of “statistical phylogeography” (Knowles and Maddison 2002) formalized the study of geographic patterns of genetic variation via gene trees by stressing the importance of defining phylogeographic hypotheses to be contrasted at the population level, rather than at the level of gene trees. Here, error due to gene tree inference and to the stochastic behavior of population genetic processes is assessed and taken into consideration. These ideas were already relevant for single locus phylogeography but became more pertinent for multilocus phylogeography mainly because multilocus datasets provide the power to accommodate the stochasticity attributable to the coalescent process (Edwards and Beerli 2000; Wakeley 2002).

Another important consequence for the analysis of multilocus datasets is that software packages have become more sophisticated over time, not only in terms of the number of parameters that can be simultaneously estimated but also the different phylogeographic scenarios, or null models, that can be tested in the statistical framework provided by coalescent theory (Nielsen and Wakeley 2001; Kuhner 2006; Hey and Nielsen 2007). The use of likelihood and Bayesian statistics are now common in phylogeographic inference (Hey and Machado 2003; Beaumont and Rannala 2004). Another trend is the increased use of summary statistics that combine information from all loci instead of relying on inferences based on individual tree topologies (Beaumont et al. 2002; Hickerson et al. 2007). This analytical shift is well described in Hey and Machado (2003), who noted a conflict between the older, gene tree based approaches to phylogeography, and the newer approaches that integrate out gene tree heterogeneity in their focus on parameters at the level of populations. For example, Slatkin’s s statistic (Slatkin and Maddison 1989), which was commonly used to estimate gene flow, and Nested Clade Analysis (Templeton 1998) are examples of methodologies that, due to their reliance on the inference of specific gene tree topologies, have become less popular. The coalescent models available now allow researchers to infer historical demographic fluctuations (Bensch et al. 2006), test hypotheses of lineage sorting versus introgression (Peters et al. 2007), examine colonization histories (e.g. Rosenblum et al. 2007), confirm ancient differentiation with introgression (Townsend et al. 2007), and estimate a model of population divergence that incorporates both contemporary distributions and historical associations (Knowles and Carstens 2007b). Many such approaches can even integrate information obtained from loci with different effective population sizes, such as mtDNA, X and Y chromosomes, and autosomes, leading to greater utility.

In addition, the advent of multilocus approaches has, if anything, caused researchers to move away from a strict reliance on gene trees, and it is now recognized that the signal in any one gene tree is less important than the sum of signals across gene trees and loci (Dolman and Moritz 2006). This shift in perspective has been driven in part by a diversity of statistical packages that infer demographic history not by measuring signals in fixed gene trees, but by integrating over gene trees and thereby incorporating the uncertainty in gene tree inference into the estimation process. Examples of packages that take this approach, often employing Bayesian statistics, include BEAST (Drummond and Rambaut 2007), msBayes (Hickerson et al. 2007), IM (Nielsen and Wakeley 2001), MIMAR (Becquet and Przeworski 2007), MIGRATE (Beerli and Felsenstein 1999), and BEST (Liu and Pearl 2007).

How many loci are necessary for phylogeographic inference?

The optimal sampling strategy among individuals, genes, and sites within genes for parameter estimation has been analyzed for estimates of genetic diversity (Pluzhnikov and Donnelly 1996; Felsenstein 2006; Carling and Brumfield 2007), gene flow (Beerli and Felsenstein 1999), and population growth rates (Kuhner et al. 1998). In general, all these studies corroborate the conclusion that increasing the number of loci has a critical effect on the accuracy of the parameter estimates, although in some cases increasing the number of individuals or the number of sites per sequence read is also important (Maddison and Knowles 2006). Felsenstein (2006) pointed out that when estimating θ for a single population, if the cost of a project depended solely on the number of sites sequenced, and if there were no technical impediments to sequencing a large number of linked or unlinked sites, it would be most efficient to focus on only 7 or 8 individuals, sampling a single nucleotide per locus until the study costs are used up! Such a sampling scheme would seem surprising to most phylogeographers, who usually prefer to increase the number of alleles (individuals) at the expense of additional loci. However, the few studies that have provided guidelines for optimal sampling (Felsenstein 2006; Carling and Brumfield 2007) have analyzed very simple situations that few populations will approximate. Empirical datasets, by contrast, have many more sources of variation that may elevate the number of individuals and loci needed to accurately estimate the population parameters. The optimal number of loci will likely depend on the complexity of the demography being inferred and how much it departs from a standard neutral model of a single population**.** This is an area in need of much more research and the precise allocation among individuals, loci and sites may be complex and scenario-dependent (Edwards and Beerli 2000; Hickerson et al. 2007; Peters et al. 2007).

Population genomics—Genomic phylogeography

The concept of “population genomics” was introduced to describe the process of sampling numerous loci within a genome to identify locus-specific effects from genome-wide effects (Black et al. 2001). Likewise, “genomic phylogeography” describes the simultaneous sampling of numerous loci across the genome to infer population history and estimate demographic parameters. Genomic phylogeography is distinguished from multilocus phylogeography by scale and degree. Multilocus studies usually focus on a few tens of markers; although a considerable improvement over single-locus analyses, such studies still only sparsely sample the full heterogeneity of the drift process, and inferences may be driven by a few outlier loci. In genomic phylogeography, by contrast, enough loci are screened to accurately estimate sampling distributions across loci, and locus-specific effects will be represented on the extreme values while genome-wide effects will fall into the centers of the distribution (Luikart et al. 2003). Such locus-specific effects may be due to selection, mutation, or recombination, whereas genome-wide effects are due to demographic processes such as gene flow, inbreeding, population growth, or bottlenecks, and that informs population history. Two main steps are involved in this process: (1) estimating genome-wide effects and (2) detecting outlier loci. Luikart et al. (2003) and Storz (2005) review ways for identifying outlier loci that uses simulated or empirical null distribution of summary statistics such as F_{ST} or homozygosity. These genomic approaches have been successfully applied to model species such as humans (Storz et al. 2004; Teshima et al. 2006), *Drosophila* (Harr et al. 2002), and maize (Teshima et al. 2006). The empirical distribution requires that enough loci be sampled (in the order of hundreds) to build robust null distributions and avoid erroneous identifications of perfectly good neutral loci. Examples or methods that use theoretical distributions are Ewens-Watterson test for neutrality (Ewens 1972; Watterson 1978), and the F_{ST}-outlier test developed by Beaumont and Nichols (1996).

Do we still need mtDNA?

With the increasing ease of obtaining multiple nuclear loci it is relevant to ask whether there is still a need for organelle phylogeography in the era of genomics. Zhang and Hewitt (2003) argue that nuclear data would not completely replace the use of organelle (mtDNA) data but they will be complementary in the sense that they reveal different aspects of a complex story at different depths of perception. Due to their different effective population size, all else being equal, neutral genetic drift will cause divergence between populations to be four-times slower at nuclear loci, and thus nuclear phylogeographic structure is expected to be smaller than organelle structure (Moore 1995). Also, mtDNA can yield exceptionally clear views of phylogeograhic history (e.g., Fig. 1); where mtDNA exhibits monophyly, nuclear genes invariably exhibit poly- or paraphyly (Zink and Barrowclough 2008). For these reasons, we do not expect phylogeographers to abandon mtDNA anytime soon. On the other hand, mtDNA is increasingly found to be the target of non-neutral processes, a claim that has been made for humans and many other species (Pluzhnikov et al. 2002; Bensch et al. 2006). In addition, mtDNA can be in complete linkage disequilibrium with chromosomal regions under strong selection (such as the W chromosome of birds) and thus prone to selective sweeps and Hill-Robertson effects. This was recently studied by Berlin et al. (2007), who showed that sequence diversity in a mitochondrial coding gene is in general reduced in birds when compared with mammals. Although some genome-informed phylogeographic studies have already abandoned mtDNA, invariably these studies build on previous work for which mtDNA was essential and provided the basic groundwork. Given the increasing ease of collecting large-scale DNA sequence data (Binladen et al. 2007), we suspect that the 'choice' of using nuclear or mitochondrial DNA will rapidly become moot.

Lessons and challenges from the nuclear genome

Recombination and estimating allelic phase

Recombination results in multiple histories within a single contiguously sampled genomic region, raising the caveat that gene tree approaches cannot be applied blindly. There are several strategies to deal with recombination. One may choose markers that typically show low levels of recombination, but usually such markers correspond to genomic regions with low polymorphism (Nachman 2001) and therefore may not be very informative for intraspecific studies. Alternatively, one may use one of the available methods, such as the four-gamete test to estimate the minimum number of recombination events for each sequence read and then choose the largest stretch of DNA sequence that complies with the non-recombining assumptions (Hudson and Kaplan 1985). This method, however, is sensitive to sample size and strong geographic structure, and assumes an infinite site model which might not apply in many cases where mutation rates are fairly high or effective population sizes are large.

Another strategy is to analyze the data with methods that incorporate recombination into the model of population history, such as those available in the computer package LAMARC (Kuhner 2006) or the newly proposed MIMAR approach for estimating speciation parameters in the two population case (Becquet and Przeworski 2007). Few such methods are available and few attempts have been made to incorporate recombination into the phylogeographic estimation process, except through coalescent simulations (Voight et al. 2005).

An additional technical hurdle in the use of nuclear loci for phylogeography is determining allelic phase; in fact, the issues of recombination and phase estimation are largely ignored in higher level phylogenetic studies, with some justification, since recombination is less likely to have effects on estimation when divergence times between species are long and limited sampling within species does not yield a set of sequences that show evidence of recombination. We will not review methods for haplotype inference, except to reiterate Zhang and Hewitt (2003), who pointed out that such approaches can be divided into probabilistic or computational methods, and empirical methods. The most commonly used empirical method is cloning; several clones are sequenced for each individual and true phases are determined after ruling out sequencing artifacts and in vitro recombination events. Probabilistic methods can be based on maximum-likelihood, parsimony or Bayesian algorithms, with the most widely used package being PHASE (Stephens et al. 2001).

Recent phylogeographic studies insist in estimating allelic phases because they frequently seek to build gene trees, which usually require resolved haplotypes for inference. However, as we have seen earlier, many large-scale resequencing studies have opted to analyze their resequencing data using site frequency spectra rather than gene trees, even when the opportunity for building gene trees is available. Thus the question of whether to phase or not to phase will be tightly linked with downstream approaches used to infer demographic history.

Multilocus phylogenetics

One of the most difficult challenges in phylogenetic reconstruction is the widespread occurrence of incongruence between alternative datasets. Empirical studies have shown that incongruence can occur at different taxonomic levels in the tree of life, and can be generated by both analytical and biological causes (Maddison 1997; Rokas et al. 2003). The reigning paradigm is one in which applying the correct optimality criteria, deep taxon sampling, and an increasing size of sequence data will together provide robust estimations of phylogenetic history even for the most difficult problems in phylogenetic reconstruction. This view is probably justified for a large diversity of phylogenetic situations. However, it has recently become evident that resolving many of the tree-of-life nodes is not just a matter of increased data collection (Nishihara et al. 2005; Rokas et al. 2005; Patterson et al. 2006; Rokas and Carroll 2006). Recently, efforts have been made to search for new types of data different from primary sequences that hopefully have the desired features of high signal-to-noise ratio that is required for historical inference in phylogenetic analyses. Some of these new types of data have been collectively called Rare Genomic Changes (RGC, Rokas and Holland 2000), and comprise chromosome rearrangements (Müller et al. 2003), gene order (O'Brien et al. 1998; Bourque and Pevzner 2002), retroelements (Shedlock et al. 2000), indels (Murphy et al. 2007), gene duplications (Bowers et al. 2003), etc. Methods involving RGC have been reviewed elsewhere (e.g. Rokas and Holland 2000; Delsuc et al. 2005; Edwards et al. 2005), and will not be reviewed further here. Rather, we emphasize how recent methodological advances for the analysis of sequence-based markers, in particular from coalescent theory, are transforming phylogenetic analysis and bridging the gap with phylogeography.

Although most phylogenetic analysis is focused on resolving relationships among species and clades, in fact most phylogenetic methods essentially estimate gene trees and simply assume a complete correspondence between the gene tree(s) and the species tree. This assumption will be correct without a large amount of horizontal gene transfer or interspecific gene flow, gene duplication, or incomplete lineage sorting of ancestral alleles (Maddison 1997). Traditional phylogenetic methods that were originally designed for single gene analyses are still the workhorse of phylogenomics, despite increasing evidence for heterogeneity in trees for different loci. There is a need for methods that account for individual and population-level processes that influence the multilocus behavior of the genome. Such processes depend on the effective population size, scaled by the mutation rate ($\theta = 4N_e\mu$), and the branch lengths (time between the internodes also scaled by the mutation rate μt) as measured in coalescent units (Fig. 3). Because different regions of an individual's genome may have different coalescent histories, different effective population sizes, and may be evolving at different rates, the genome is really a collection of gene trees that may be explained by many different topologies, all correct but all potentially different from the species tree and from one another (Fig. 4). In addition there is a need to extend two-lineage phylogeographic models, such as those incorporated in IM and MIMAR, to multi-lineage models, or to adopt mathematical approximations that make it easier to do so (Wakeley 2004).

Empirical studies of multilocus phylogenetics invariably reveal complex patterns among loci. Machado and Hey (2003), analyzed 16 loci in several closely related species of *Drosophila* and provided one of the first available empirical examples of the multiplicity of gene trees within and among closely related species. They demonstrated highly dissimilar gene genealogies leading to composite genomes that were explained by partial introgression between species or differential lineage sorting from polymorphic ancestors. Jennings and Edwards (2005) analyzed 30 randomly selected loci to study the speciation history of the Australian grass finches (*Poephila*), and from the reconstructed gene genealogies recovered all possible topologies with considerable variation in coalescent times (Fig. 4). Pollard et al. (2006) investigated the phylogenetic relationship among three species subgroups with the genus *Drosophila* using whole genomes, and found widespread incongruence in nucleotide and amino acid substitutions as well as in indels and gene tree topologies. These results were most easily explained by incomplete lineage sorting occurring along the short internodes spanning the two key speciation events, a conclusion that was supported by the observation that substitutions supporting the same tree were spatially clustered or in regions with low recombination.

As more empirical examples of multilocus approaches to the study of phylogenetics become available it is becoming clear that multiple genealogical histories are the norm rather than the exception, especially when cladogenetic events are separated by short internodes (Rokas and Carroll 2006) or when speciation is accompanied by differential introgression of the genome (Machado et al. 2002; Machado and Hey 2003; Hey and Nielsen 2004). All of this was predicted by empiricists and theoreticians decades ago (Kingman 1982a, b; Avise and Ball 1990), yet it is only recently that data sets illustrating these phenomena have become common, or that the consequences for phylogenetic analysis are becoming better understood (Kubatko and Degnan 2007).

Statistical phylogenetics: estimating species trees

Although phylogenetic methods for estimating gene trees from sequence data have achieved an enviable complexity, and are now able to deal with a myriad of nucleotide substitution models and rates of evolution, these models all focus on improving model complexity at the nucleotide level so as to produce a more accurate gene tree. Only recently have researchers devoted attention to the second critical process of inference in phylogenetics, inferring the species tree from a collection of gene trees. Over the past 20 years, the increased appreciation of coalescent heterogeneity, particularly among closely related species, has lead to a series of methods that directly address this heterogeneity to inform phylogenetics (Table 1). In this relatively young tradition, lessons from coalescent theory are being used to solve complex phylogenetic problems, such as hard polytomies in the tree of life (Rokas and Carroll 2006), studies of speciation in rapid radiations (Shaffer and Thomson 2007), and delimitation of species in the absence of complete lineage sorting (Knowles and Carstens 2007a). Most importantly, they have gradually led to a clear distinction between gene trees and species trees, and to the recognition that systematists are fundamentally interested in the latter (Fig. 1; Table 1).

Methods for inferring species trees from multiple gene trees have a reasonably deep history (Table 1), going back really to the dawn of phylogenetics, such as Cavalli-Sforza's inferences of phylogeny for human populations based on allozyme and blood group data (Cavalli-Sforza and Edwards 1967). Methods that use multilocus data to esimtat genetic distances among populations (summarized in Nei 1987), which are then used to construct a phylogenetic tree of species or populations, can be considered species tree approaches. Felsenstein's PHYLIP package (Felsenstein 1981)has for a long time contained a procedure, contml, which inferred phylogeny based on a genetic drift Brownian motion model of allele frequency change proposed by Cavalli-Sforza and Edwards (1967); this is one of the first phylogenetic models to incorporate an explicit population genetic model of drift. With the advent of mtDNA studies and a genealogical view of genetic variation, methods were developed for inferring species trees from the pattern of ancestry of allelic lineages shared between species (Takahata 1989), and such approaches have been extended to multilocus data (Liu 2006). Recent approaches to species tree inference have utilized MCMC simulation of genetic variation, either using allele frequencies and summary statistics such as F_{st} (Nielsen 1998; Nielsen et al. 1998) or by modeling gene trees explicitly (Liu and Pearl 2007). Whereas some of these approaches are applicable to only small data sets, others have wider applicability (see references in Table 1).

The most recent phase of species tree estimation has introduced approaches that incorporate models of stochastic sorting of alleles due to drift into the process of phylogenetic inference (Maddison and Knowles 2006; Liu and Pearl 2007). These methods promise to produce robust estimates of the species tree even in the face of extensive and continuing persistence of ancestral polymorphism in cases of no introgression. Maddison and Knowles (2006) showed that the signal of a species phylogeny may persists despite the lack of reciprocal monophyly and discordance among loci, and this signal can be used to infer the species tree using a parsimony criterion in which the species tree that is favored is the one that minimizes the total number of discordances summed over all gene trees. These authors investigate the utility of using minimizing-deep coalescences method for phylogenetic reconstruction using

Table 1 Summary of methods that estimate species trees from gene trees or allele frequencies

Method (Reference)	Methodological basis	Statistical model?	Accounts for error in gene tree?	Yields species tree branch lengths?	Yields effective population sizes?	Applicable to large data sets?
Continuous characters (contml; Felsenstein 1981)	Likelihood (allele frequencies)	Yes	N/A	Yes	No	Yes
Probability of incongruence (Pamilo and Nei 1988; Wu 1991; Chen and Li 2001)	Likelihood	Yes	No	Yes	Yes	No
Minimum divergence (Takahata 1989)	Parametric	Yes	No	No	No	No
Infinite sites model (Nielsen 1998)	Likelihood	Yes	N/A	Yes	Yes	No
F_{st} method (Nielsen et al. 1998)	Likelihood	Yes	N/A	Yes	Yes	No
Genetree parsimony (Page and Charleston 1997)	Parsimony	No	No	No	No	Yes
Deep coalescence (Maddison 1997; Maddison and Knowles 2006)	Parsimony	No	No	No	No	Yes
SINE method (Waddell et al. 2001)	Likelihood	Yes	No	No	Yes	No
Ancestral polymorphism (Hudson 1992; Waddell 2002)	Likelihood	Yes	N/A	Yes	Yes	Yes
Gene tree probabilities (Carstens and Knowles 2007a)	Likelihood	Yes	No	N/A	N/A	Yes
Bayesian Estimation of Species Trees (BEST) (Liu and Pearl 2007)	Bayesian	Yes	Yes	Yes	Yes	Yes
Maximum tree (Liu 2006)	Likelihood	Yes	No	Yes	No	Yes
Rooted triple consensus (Ewing et al. 2008)	Consensus	No	No	No	No	Yes
Sum and average criteria (Seo et al. 2005)	Likelihood	Yes	Yes	No	No	Yes

simulated datasets. However, a disadvantage of the deep coalescent method is that it does not account for the error in gene tree estimation. Other methods for estimating species trees from gene trees do not use an explicit model connecting gene to species trees but rather combine likelihoods of gene trees (Seo et al. 2005) or combine groups of gene trees via consensus approaches (Ewing et al. 2008; Table 1).

A hierarchical Bayesian method to estimate posterior distribution of the species tree given multiple and independent gene trees was devised to address some of these shortcomings (Edwards et al. 2007; Liu and Pearl 2007). Like the deep coalescence approach, the Bayesian method does not rely on concatenated datasets but instead estimates gene tree distributions in a model in which gene trees are partially correlated due to their common species tree history (Liu and Pearl 2007). This method is implemented in the program Bayesian estimation of Species Trees (BEST) and has been shown to perform well in simulations (Edwards et al. 2007) and can now handle multiple alleles per species (Belfiore et al. 2008; Liu et al. 2008).

The deep coalescence method of Maddison and Knowles (2006) has been applied to empirical datasets to estimate the species tree when both incongruence among gene trees and lack of reciprocal monophyly are present (Carstens and Knowles 2007a). Carling and Brumfield (2008) sampled a suit of 10 loci to investigate the phylogenetic relationship of four species of passerine buntings. These authors successfully applied both phylogenetic and population genetic (phylogeographic) methods to determine the species relationships and recovered the traditional phylogeny that contradicts the mtDNA inference previously made with cytochrome *b* data.

Species delimitation

Another topic in the field of molecular phylogenetics that has been recently influenced by population genetics, in particular coalescent theory, is the recognition of new species and the identification of species limits by using molecular sequence data. Sites and Marshall (2003, 2004) classified methods of species delimitation into tree-based methods and non-tree-based methods, with the former usually implying the need to recover reciprocally monophyletic clades between sister species. Hudson and Coyne (2002) investigated the length of time it takes for a pair of sister species to achieved reciprocal monophyly and suggested that not only species delimitation should not be carried out exclusively with mitochondrial genes (due to its smaller N_e) but also, they suggested that genealogical species should allow for less than 100% monophyly for a sample of loci it the entire genome. Knowles and Carstens (2007a) proposed a new method that uses coalescent simulations to test hypotheses about species limits that incorporates information from multiple loci without requiring reciprocal monophyly from each locus. These authors argue that species can be accurately identified with nuclear genes even when very recently derived but species delimitation cannot be done by simple visual inspection of the gene trees. More importantly, sampling many loci as well as multiple individuals has a substantial impact on whether species can be delimited with these probabilistic methods. When analyzed with only single loci such as mtDNA (as in DNA barcoding programs), species and their relationships may be overconfidently delimited, especially in recent radiations when reciprocal monophyly is not expected with nuclear genes, although further work is needed here.

Conclusions

The modern incarnations of phylogeography and phylogenetics were both spawned in the mitochondrial era that was greatly accelerated by the advent of PCR. But, as we have discussed, these two fields diverged somewhat in the 1990s, we believe because of the adoption by phylogeographers of markers that were ill-suited to phylogenetics, and—as we now know—possess a number of difficulties for phylogeography as well. An increased ease of developing loci from focal species, as well as the large amounts of sequence data made available in the era of whole genomes have begun to turn phylogeographers' attention away from microsatellites and toward sequence-based markers, such as anonymous loci, introns or individual SNPs. Sequence-based markers (including both linked and unlinked SNPs) have enormous promise for placing mitochondrial and nuclear gene histories on common mutational scales, and for mending the diverging technical paths that phylogeography and phylogenetics took during the 1990s. The use of sequence-based markers among closely related populations of the same species presents challenges; however, since recombination can render gene trees uninformative and spurious, the diversity of analytical approaches currently applied is partly a consequence of the conflict between the gene tree legacy left by studies on mtDNA and the unsuitability for many nuclear loci for strict phylogenetic analysis.

The heterogeneity in genealogical histories that become apparent upon most examinations of nuclear sequence data among closely related species raises additional challenges that theory is gradually beginning to address. Chief among these are the advent of powerful approaches for estimating species trees as distinct from gene trees. Coalescent theory is making inroads to approaches far above the species level, and it continues to provide a conceptual bridge for fruitful exchange between phylogeography and phylogenetics. Rapid advances in phylogeography and phylogenetics will require researchers to confront the rampant gene tree heterogeneity that is now the norm and to continue the exchange of perspectives that has recently been facilitated by sequence-based markers.

Acknowledgments PHB was supported by a postdoctoral fellowship awarded by the Fundação para a Ciência e Tecnologia, Portugal. We thank our lab colleagues and Carlos Bustamante for helpful discussion. We are grateful for the criticisms of two anonymous reviewers and to Craig Moritz and Rauri Bowie for their perspective and for providing detailed comments on the manuscript.

References

Avise JC (2000) Phylogeography. The history and formation of species. Harvard University Press, Cambridge, Massachusetts

Avise JC, Ball RM Jr (1990) Principles of genealogical concordance in species concepts and biological taxonomy. Oxf Surv Evol Biol 7:45–67

Avise JC, Arnold J, Ball RM et al (1987) Intraspecific phylogeography: the mitochondrial DNA bridge between population genetics and systematics. Annu Rev Ecol Syst 18:489–522

Backström N, Fagerberg S, Ellegren H (2008) Genomics of natural bird populations: a gene-based set of reference markers evenly spread across the avian genome. Mol Ecol 17:964–980. doi:10.1111/j.1365-294X.2007.03551.x

Bazin E, Glémin S, Galtier N (2006) Population size does not influence mitochondrial genetic diversity in animals. Science 312:570–572. doi:10.1126/science.1122033

Beaumont MA, Nichols RA (1996) Evaluating loci for use in the genetic analysis of population structure. Proc R Soc Lond B Biol Sci 263:1619–1626. doi:10.1098/rspb.1996.0237

Beaumont MA, Rannala B (2004) The Bayesian revolution in genetics. Nat Rev Genet 5:251–261. doi:10.1038/nrg1318

Beaumont MA, Zhang W, Balding DJ (2002) Approximate bayesian computation in population genetics. Genetics 162:2025–2035

Becquet C, Przeworski M (2007) A new approach to estimate parameters of speciation models with application to apes. Genome Res 17:1505–1519. doi:10.1101/gr.6409707

Beerli P, Felsenstein J (1999) Maximum-likelihood estimation of migration rates and effective population numbers in two populations using a coalescent approach. Genetics 152:763–773

Belfiore NM, Liu L, Moritz C (2008) Multilocus phylogenetics of a rapid radiation in the genus thomomys (Rodentia: *Geomyidae*). Syst Biol 57:294–310. doi:10.1080/10635150802044011

Benavides E, Baum R, McClellan D, Sites JW Jr (2007) Molecular phylogenetics of the lizard genus *Microlophus* (Squamata: Tropiduridae): aligning and retrieving indel signal from nuclear introns. Syst Biol 56:776–797. doi:10.1080/10635150701618527

Bensch S, Irwin DE, Irwin JH, Kvist L, Åkesson S (2006) Conflicting patterns of mitochondrial and nuclear DNA diversity in *Phylloscopus* warblers. Mol Ecol 15:161–171. doi:10.1111/j.1365-294X.2005.02766.x

Berlin S, Tomaras D, Charlesworth B (2007) Low mitochondrial variability in birds may indicate Hill-Robertson effects on the W chromosome. Heredity 99:389–396. doi:10.1038/sj.hdy.6801014

Binladen J, Thomas M, Gilbert P et al (2007) The use of coded PCR primers enables high-throughput sequencing of multiple homolog amplification products by 454 parallel sequencing. PLoS One 2:e197. doi:10.1371/journal.pone.0000197

Black WC, Baer CF, Antolin MF, DuTeau NM (2001) Population genomics: genome-wide sampling of insect populations. Annu Rev Entomol 46:441–469. doi:10.1146/annurev.ento.46.1.441

Boissinot S, Chevret P, Furano AV (2000) L1 (LINE–1) retrotransposon evolution and amplification in recent human history. Mol Biol Evol 17:915–928

Bourque G, Pevzner PA (2002) Genome-scale evolution: reconstructing gene orders in the ancestral species. Genome Res 12:26–36

Bowers JE, Chapman BA, Rong J, Paterson AH (2003) Unraveling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. Nature 422:433–438. doi:10.1038/nature01521

Brandström M, Ellegren H (2007) The genomic landscape of short insertion and deletion polymorphisms in the chicken (*Gallus gallus*) genome: a high frequency of deletions in tandem duplicates. Genetics 176:1691–1701. doi:10.1534/genetics.107.070805

Brown MW, George M Jr, Wilson AC (1979) Rapid evolution of animal mitochondrial DNA. Proc Natl Acad Sci USA 76:1967–1971. doi:10.1073/pnas.76.4.1967

Brumfield RT, Beerli P, Nickerson DA, Edwards SV (2003) The utility of single nucleotide polymorphisms in inferences of population history. Trends Ecol Evol 18:249–256. doi:10.1016/S0169-5347(03)00018-1

Bull JJ, Huelsenbeck JP, Cunningham CW, Swofford DL, Waddell PJ (1993) Partitioning and combining data in phylogenetic analysis. Syst Biol 42:384–397. doi:10.2307/2992473

Caccone A, Amato GD, Powell JR (1988) Rates and patterns of scnDNA and mtDNA divergence within the *Drosophila melanogaster* subgroup. Genetics 118:671–683

Carling MD, Brumfield RT (2007) Gene sampling strategies for multi-locus population estimates of genetic diversity (θ). PLoS One 2:e160. doi:10.1371/journal.pone.0000160

Carling MD, Brumfield RT (2008) Integrating phylogenetic and population genetic analyses of multiple loci to test species divergence hypotheses in Passerina buntings. Genetics 178:363–377. doi:10.1534/genetics.107.076422

Carstens BC, Knowles LL (2007a) Estimating species phylogeny from gene-tree probabilities despite incomplete lineage sorting: an example from *Melanoplus* grasshoppers. Syst Biol 56:400–411. doi:10.1080/10635150701405560

Carstens BC, Knowles LL (2007b) Shifting distributions and speciation: species divergence during rapid climate change. Mol Ecol 16:619–627. doi:10.1111/j.1365-294X.2006.03167.x

Cavalli-Sforza LL, Edwards AWF (1967) Phylogenetic analysis: models and estimation procedures. Evol Int J Org Evol 32:550–570. doi:10.2307/2406616

Chen FC, Li WH (2001) Genomic divergences between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees. Am J Hum Genet 68:444–456. doi:10.1086/318206

Cummings MP, Otto SP, Wakeley J (1995) Sampling properties of DNA sequence data in phylogenetic analysis. Mol Biol Evol 12:814–822

Delsuc F, Brinkmann H, Philippe H (2005) Phylogenomics and the reconstruction of the tree of life. Nat Rev Genet 6:361–375. doi:10.1038/nrg1603

Dettman JR, Jacobson DJ, Taylor JW (2003) A multilocus genealogical approach to phylogenetic species recognition in the model eukaryote *Neurospora*. Evol Int J Org Evol 57:2703–2720

Dolman G, Moritz C (2006) A multilocus perspective on refugial isolation and divergence in rainforest skinks (*Carlia*). Evol Int J Org Evol 60:573–582

Drummond AJ, Rambaut A (2007) BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol Biol 7:214. doi:10.1186/1471–2148-7-214

Edwards SV (1993) Long-distance gene flow in a cooperative breeder detected in genealogies of mitochondrial DNA sequences. Proc R Soc Lond B Biol Sci 252:177–185. doi:10.1098/rspb.1993.0063

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

Edwards SV, Jennings WB, Shedlock AM (2005) Phylogenetics of modern birds in the era of genomics. Proc R Soc Lond B Biol Sci 272:979–992. doi:10.1098/rspb.2004.3035

Edwards SV, Liu L, Pearl DK (2007) High-resolution species trees without concatenation. Proc Natl Acad Sci USA 104:5936–5941. doi:10.1073/pnas.0607004104

Ellegren H (2004) Microsatellites: simple sequences with complex evolution. Nat Rev Genet 5:435–445. doi:10.1038/nrg1348

Ewens WJ (1972) The sampling theory of selectively neutral alleles. Theor Popul Gen 3:87–112. doi:10.1016/0040-5809(72)90035-4

Ewing GB, Ebersberger I, Schmidt HA, von Haeseler A (2008) Rooted triple consensus and anomalous gene trees. BMC Evol Biol 8:118

Fan JB, Oliphant A, Shen R et al (2003) Highly parallel SNP genotyping. Cold Spring Harb Symp Quant Biol 68:69–78. doi:10.1101/sqb.2003.68.69

Felsenstein J (1981) Evolutionary trees from gene frequencies and quantitative characters: finding maximum likelihood estimates. Evol Int J Org Evol 35:1229–1242. doi:10.2307/2408134

Felsenstein J (2006) Accuracy of coalescent likelihood estimates: do we need more sites, more sequences, or more loci? Mol Biol Evol 23:691–700. doi:10.1093/molbev/msj079

Friesen V, Congdon B, Walsh H, Birt T (1997) Intron variation in marbled murrelets detected using analyses of single-stranded conformational polymorphisms. Mol Ecol 6:1047–1058. doi:10.1046/j.1365-294X.1997.00277.x

Funk DJ, Omland KE (2003) Species-level paraphyly and polyphyly: frequency, causes, and consequences, with insights from animal mitochondrial DNA. Annu Rev Ecol Evol Syst 34:397–423. doi:10.1146/annurev.ecolsys.34.011802.132421

Hare MP (2001) Prospects for nuclear gene phylogeography. Trends Ecol Evol 16:700–706. doi:10.1016/S0169-5347(01)02326-6

Hare M, Avise J (1998) Population structure in the American oyster as inferred by nuclear gene genealogies. Mol Biol Evol 15:119–128

Harpending HC, Batzer MA, Gurven M et al (1998) Genetic traces of ancient demography. Proc Natl Acad Sci USA 95:1961–1967. doi:10.1073/pnas.95.4.1961

Harr B, Kauer M, Schlötterer C (2002) Hitchhiking mapping—a population-based fine mapping strategy for adaptive mutations in *Drosophila melanogaster*. Proc Natl Acad Sci USA 99:12949–12954. doi:10.1073/pnas.202336899

Hedrick PW (1999) Perspective: highly variable loci and their interpretation in evolution and conservation. Evol Int J Org Evol 53:313–318. doi:10.2307/2640768

Hedrick PW (2005) A standardized genetic differentiation measure. Evol Int J Org Evol 59:1633–1638

Hernandez RD, Hubisz MJ, Wheeler DA et al (2007) Demographic histories and patterns of linkage disequilibrium in Chinese and Indian rhesus macaques. Science 316:240–243. doi:10.1126/science.1140462

Hey J, Machado CA (2003) The study of structured populations—new hope for a difficult and divided science. Nat Rev Genet 4:535–543. doi:10.1038/nrg1112

Hey J, Nielsen R (2004) Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of *Drosophila pseudoobscura* and *D. persimilis*. Genetics 167:747–760. doi:10.1534/genetics.103.024182

Hey J, Nielsen R (2007) Integration within the Felsenstein equation for improved Markov chain Monte Carlo methods in population genetics. Proc Natl Acad Sci USA 104:2785–2790. doi:10.1073/pnas.0611164104

Hickerson MJ, Stahl E, Takebayashi N (2007) msBayes: pipeline for testing comparative phylogeographic histories using hierarchical approximate Bayesian computation. BMC Bioinformatics 26(8):268

Hudson RR (1992) Gene trees, species trees and the segregation of ancestral alleles. Genetics 131:509–513

Hudson RR, Coyne JA (2002) Mathematical consequences of the genealogical species concept. Evol Int J Org Evol 56:1557–1565

Hudson RR, Kaplan NL (1985) Statistical properties of the number of recombination events in the history of a sample of DNA sequences. Genetics 111:147–164

Huelsenbeck JP (1996) A likelihood ratio test to detect conflicting phylogenetic signal. Syst Biol 45:92–98. doi:10.2307/2413514

Innan H, Stephan W (2000) The coalescent in an exponentially growing metapopulation and its application to *Arabidopsis thaliana*. Genetics 155:2015–2019

Jennings WB, Edwards SV (2005) Speciational history of Australian grass finches (*Poephila*) inferred from thirty gene trees. Evol Int J Org Evol 59:2033–2047

Jorde LB, Rogers AR, Watkins WS et al (1997) Microsatellite diversity and the demographic history of modern humans. Proc Natl Acad Sci USA 94:3100–3103. doi:10.1073/pnas.94.7.3100

Karl SA, Avise JC (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

Keightley PD, Johnson T (2004) MCALIGN: stochastic alignment of non-coding DNA sequences based on an evolutionary model of sequence evolution. Genome Res 14:442–450. doi:10.1101/gr.1571904

Kingman JFC (1982a) On the genealogy of large populations. J Appl Probab A 19:27–43. doi:10.2307/3213548

Kingman JFC (1982b) The coalescent. Stochast Process Appl 13:235–248. doi:10.1016/0304-4149(82)90011-4

Kluge AG (1989) A concern for evidence and a phylogenetic hypothesis of relationships among Epicrates (Boidae, Serpentes). Syst Zool 38:7–25. doi:10.2307/2992432

Knowles LL, Maddison WP (2002) Statistical phylogeography. Mol Ecol 11:2623–2635. doi:10.1046/j.1365-294X.2002.01637.x

Knowles LL, Carstens BC (2007a) Delimiting species without monophyletic gene trees. Syst Biol 56:887–895. doi:10.1080/10635150701701091

Knowles LL, Carstens BC (2007b) Estimating a geographically explicit model of population divergence. Evol Int J Org Evol 61:477–493. doi:10.1111/j.1558-5646.2007.00043.x

Kocher TD, Thomas WK, Meyer A et al (1989) Dynamics of mitochondrial DNA evolution in animals: amplification and sequencing with conserved primers. Proc Natl Acad Sci USA 86:6196–6200. doi:10.1073/pnas.86.16.6196

Konkel MK, Wang JX, Liang P, Batzer MA (2007) Identification and characterization of novel polymorphic LINE-1 insertions through comparison of two human genome sequence assemblies. Gene 390:28–38. doi:10.1016/j.gene.2006.07.040

Kubatko LS, Degnan JH (2007) Inconsistency of phylogenetic estimates from concatenated data under coalescence. Syst Biol 56:17–24. doi:10.1080/10635150601146041

Kuhner MK (2006) LAMARC 2.0: maximum likelihood and Bayesian estimation of population parameters. Bioinformatics 22:768–770. doi:10.1093/bioinformatics/btk051

Kuhner MK, Yamato J, Felsenstein J (1998) Maximum likelihood estimation of population growth rates based on the coalescent. Genetics 149:429–434

Ladoukakis ED, Zouros E (2001a) Recombination in animal mitochondrial DNA: evidence from published sequences. Mol Biol Evol 18:2127–2131

Ladoukakis ED, Zouros E (2001b) Direct evidence for homologous recombination in mussel (*Mytilus galloprovincialis*) mitochondrial DNA. Mol Biol Evol 18:1168–1175

Leigh JW, Susko E, Baumgartner M, Roger AJ (2008) Testing congruence in phylogenomic analysis. Syst Biol 57:104–115. doi:10.1080/10635150801910436

Liu L (2006) Reconstructing posterior distributions of a species phylogeny using estimated gene tree distributions, Doctoral dissertation, The Ohio State University

Liu L, Pearl DK (2007) Species trees from gene trees: reconstructing Bayesian posterior distributions of a species phylogeny using estimated gene tree distributions. Syst Biol 56:504–514. doi:10.1080/10635150701429982

Liu L, Pearl DK, Brumfield RT, Edwards SV (2008) Estimating species trees using multiple-allele DNA sequence data. Evol Int J Org Evol (in press)

Luikart G, England PR, Tallmon D, Jordan S, Taberlet P (2003) The power and promise of population genomics: from genotyping to genome typing. Nat Rev Genet 4:981–994. doi:10.1038/nrg1226

Machado CA, Hey J (2003) The causes of phylogenetic conflict in a classic *Drosophila* species group. Proc R Soc Lond B Biol Sci 270:1193–1202. doi:10.1098/rspb.2003.2333

Machado CA, Kliman RM, Markert JA, Hey J (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

Maddison W (1997) Gene trees in species trees. Syst Biol 46:523–536. doi:10.2307/2413694

Maddison W, Knowles LL (2006) Inferring phylogeny despite incomplete lineage sorting. Syst Biol 55:21–30. doi:10.1080/10635150500354928

Meyer M, Stenzel U, Myles S, Prüfer K, Hofreiter M (2007) Target high-throughput sequencing of tagged nucleic acid samples. Nucleic Acids Res 35:e97. doi:10.1093/nar/gkm566

Miklos I, Lunter GA, Holmes I (2004) A "long-indel" model for evolutionary sequence alignment. Mol Biol Evol 21:529–540. doi:10.1093/molbev/msh043

Mills RE, Luttig CT, Larkins CE et al (2006) An initial map of insertion and deletion (INDEL) variation in the human genome. Genome Res 16:1182–1190. doi:10.1101/gr.4565806

Moore W (1995) Inferring phylogenies from mtDNA variation: mitochondrial-gene trees versus nuclear-gene trees. Evol Int J Org Evol 49:718–726. doi:10.2307/2410325

Müller S, Hollatz M, Wienberg J (2003) Chromosomal phylogeny and evolution of gibbons (Hylobatidae). Hum Genet 113:1432–1203. doi:10.1007/s00439-003-0997-2

Murphy WJ, Pringle TH, Crider TA, Springer MS, Miller W (2007) Using genomic data to unravel the root of the placental mammal phylogeny. Genome Res 17:413–421. doi:10.1101/gr.5918807

Nachman MW (2001) Single nucleotide polymorphisms and recombination rate in humans. Trends Genet 17:481–485. doi:10.1016/S0168-9525(01)02409-X

Nei M (1987) Molecular evolutionary genetics. Columbia University Press, New York

Nielsen R (1998) Maximum likelihood estimation of population divergence times and population phylogenies under the infinite sites model. Theor Popul Biol 53:143–151. doi:10.1006/tpbi.1997.1348

Nielsen R, Wakeley J (2001) Distinguishing migration from isolation: a Markov Chain Monte Carlo approach. Genetics 158:885–896

Nielsen R, Mountain JL, Huelsenbeck JP, Slatkin M (1998) Maximum likelihood estimation of population divergence times and population phylogeny in models without mutation. Evol Int J Org Evol 52:669–677. doi:10.2307/2411262

Nielsen R, Hubisz MJ, Clark AG (2004) Reconstituting the frequency spectrum of ascertained single-nucleotide polymorphism data. Genetics 168:2373–2382. doi:10.1534/genetics.104.031039

Nishihara H, Satta Y, Nikaido M et al (2005) A retroposon analysis of Afrotherian phylogeny. Mol Biol Evol 22:1823–1833. doi:10.1093/molbev/msi179

Nylander JAA, Ronquist F, Huelsenbeck JP, Nieves-Aldrey JL (2004) Bayesian phylogenetic analysis of combined data. Syst Biol 53:47–67. doi:10.1080/10635150490264699

O'Brien TG, Kinnaird MF, Dierenfeld ES et al (1998) Gene translocation links insects and crustaceans. Nature 392:667–668. doi:10.1038/33580

Ometto L, Stephan W, De Lorenzo D (2005) Insertion/deletion and nucleotide polymorphism data reveal constraints in *Drosophila melanogaster* introns and intergenic regions. Genetics 169:1521–1527. doi:10.1534/genetics.104.037689

Osada N, Wu C-I (2005) Inferring the mode of speciation from genomic data: a study of the great apes. Genetics 169:259–264. doi:10.1534/genetics.104.029231

Otto SP, Cummings MP, Wakeley J (1996) Inferring phylogenies from DNA sequence data: the effects of sampling. In: Harvey PH, Leigh Brown AJ, Maynard Smith J, Nee S (eds) New uses for new phylogenies. Oxford University Press, New York, pp 103–115

Page R, Charleston M (1997) From gene to organismal phylogeny: reconciled trees and the gene tree/species tree problem. Mol Phyl Evol 7:231–240. doi:10.1006/mpev.1996.0390

Palumbi SR (1996) The polymerase chain reaction. In: DM Hillis, Moritz C, Mable BK (eds) Molecular systematics. Sinauer Associates, Inc, pp 205–247

Palumbi S, Baker C (1994) Contrasting population structure from nuclear intron sequences and mtDNA of Humpback whales. Mol Biol Evol 11:426–435

Pamilo P, Nei M (1988) Relationships between gene trees and species trees. Mol Biol Evol 5:568–583

Patterson N, Richter DJ, Gnerre S, Lander ES, Reich D (2006) Genetic evidence for complex speciation of humans and chimpanzees. Nature 441:1103–1108. doi:10.1038/nature04789

Peters JL, Zhuravlev Y, Fefelov I, Logie A, Omland KE (2007) Nuclear loci and coalescent methods support ancient hybridization as cause of mitochondrial paraphyly between gadwall and falcated duck (*Anas* spp.). Evol Int J Org Evol 61:1992–2006. doi:10.1111/j.1558-5646.2007.00149.x

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

Pluzhnikov A, Rienzo AD, Hudson RR (2002) Inferences about human demography based on multilocus analyses of noncoding sequences. Genetics 161:1209–1218

Pollard DA, Iyer VN, Moses AM, Eisen MB (2006) Widespread discordance of gene trees with species tree in *Drosophila*: evidence for incomplete lineage sorting. PLoS Genet 2:e173. doi:10.1371/journal.pgen.0020173

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155: 945–959

Rokas A, Carroll SB (2006) Bushes in the tree of life. PLoS Biol 4:1899–1904. doi:10.1371/journal.pbio.0040352

Rokas A, Holland PWH (2000) Rare genomic changes as a tool for phylogenetics. Trends Ecol Evol 15:454–459. doi:10.1016/S0169-5347(00)01967-4

Rokas A, Williams BL, King N, Carroll SB (2003) Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature 425:798–804. doi:10.1038/nature02053

Rokas A, Krüger D, Carroll SB (2005) Animal evolution and the molecular signature of radiations compressed in time. Science 310:1933–1938. doi:10.1126/science.1116759

Rosenblum EB, Novembre J (2007) Ascertainment bias in spatially structured populations: a case study in the eastern fence lizard. J Hered 98:331–336. doi:10.1093/jhered/esm031

Rosenblum EB, Hickerson MJ, Moritz C (2007) A multilocus perspective on colonization accompanied by selection and gene flow. Evol Int J Org Evol 61:2971–2985. doi:10.1111/j.1558-5646.2007.00251.x

Ross CL, Harrison RG (2002) A fine-scale spatial analysis of the mosaic hybrid zone between *Gryllus firmus* and *Grillus pennsylvanicus*. Evol Int J Org Evol 56:2296–2312

Rubinsztein DC, Amos B, Cooper G (1999) Microsatellite and trinucleotide-repeat evolution: evidence for mutational bias and different rates of evolution in different lineages. Philos Trans R Soc Lond B Biol Sci 354:1095–1099. doi:10.1098/rstb.1999.0465

Seo T-K, Kishino H, Thorne JT (2005) Incorporating gene-specific variation when inferring and evaluating optimal evolutionary tree topologies from multilocus sequence data. Proc Natl Acad Sci (USA) 102:4436–4441

Shaffer HB, Thomson RC (2007) Delimiting species in recent radiations. Syst Biol 56:896–906. doi:10.1080/10635150701772563

Shedlock AM, Milinkovitch MC, Okada N (2000) SINE evolution, missing data, and the origin of whales. Syst Biol 49:808–817. doi:10.1080/106351500750049851

Shedlock AM, Takahashi K, Okada N (2004) SINEs of speciation: tracking lineages with retroposons. Trends Ecol Evol 19:545–553. doi:10.1016/j.tree.2004.08.002

Shriver MD, Kennedy GC, Parra EJ et al (2004) The genomic distribution of population substructure in four populations using 8, 525 autosomal SNPs. Hum Genomics 1:274–286

Shriver MD, Mei R, Parra EJ et al (2005) Large-scale SNP analysis reveals clustered and continuous patterns of human genetic variation. Hum Genomics 2:81–89

Sites JW Jr, Marshall JC (2003) Delimiting species: a renaissance issue in systematic biology. Trends Ecol Evol 18:462–470. doi:10.1016/S0169-5347(03)00184-8

Sites JW Jr, Marshall JC (2004) Operational criteria for delimiting species. Annu Rev Ecol Syst 35:199–227. doi:10.1146/annurev.ecolsys.35.112202.130128

Slatkin M, Maddison WP (1989) A cladistic measure of gene flow inferred from the phylogenies of alleles. Genetics 123:603–613

Small KS, Brudno M, Hill MM, Sidow A (2007) Extreme genomic variation in a natural population. Proc Natl Acad Sci USA 104:5698–5703. doi:10.1073/pnas.0700890104

Stephens M, Smith N, Donnelly P (2001) A new statistical method for haplotype reconstruction from population data. Am J Hum Genet 68:978–989. doi:10.1086/319501

Storz JF (2005) Using genome scans of DNA polymorphism to infer adaptive population divergence. Mol Ecol 14:671–688. doi:10.1111/j.1365-294X.2005.02437.x

Storz JF, Payseur BA, Nachman MW (2004) Genome scans of DNA variability in humans reveal evidence for selective sweeps outside of Africa. Mol Biol Evol 21:1800–1811. doi:10.1093/molbev/msh192

Takahata N (1989) Gene genealogy in three related populations: consistency probability between gene and population trees. Genetics 122:957–966

Tautz D (1989) Hypervariability of simple sequences as a general source for polymorphic DNA markers. Nucleic Acids Res 17:6463–6471. doi:10.1093/nar/17.16.6463

Templeton AR (1998) Nested clades analyses of phylogeographic data: testing hypotheses about gene flow and population history. Genetics 7:381–397

Teshima KM, Coop G, Przeworski M (2006) How reliable are empirical genomic scans for selective sweeps? Genome Res 16:702–712. doi:10.1101/gr.5105206

Thalmann O, Fischer A, Lankester F, Pääbo S, Vigilant L (2006) The complex evolutionary history of gorillas: insights from genomic data. Mol Biol Evol 24:146–158. doi:10.1093/molbev/msl160

Tosi AJ, Morales JC, Melnick DJ (2003) Paternal, maternal, and biparental molecular markers provide unique windows onto the evolutionary history of macaque monkeys. Evol Int J Org Evol 57:1419–1435

Townsend AK, Rimmer CC, Latta SC, Lovette IJ (2007) Ancient differentiation in the single-island avian radiation of endemic Hispaniolan chat-tanagers (Aves: *Calyptophilus*). Mol Ecol 16:3634–3642. doi:10.1111/j.1365-294X.2007.03422.x

Väli Ü, Brandström M, Johansson M, Ellegren H (2008) Insertion-deletion polymorphisms (indels) as genetic markers in natural populations. BMC Genet 9:8

Vera JC, Wheat CW, Fescemyer HW et al (2008) Rapid transcriptome characterization for a nonmodel organism using 454 pyrosequencing. Mol Ecol 17:1636–1647. doi:10.1111/j.1365-294X.2008.03666.x

Voight BF, Adams AM, Frisse LA et al (2005) Interrogating multiple aspects of variation in a full resequencing data set to infer human population size changes. Proc Natl Acad Sci USA 102:18508–18513. doi:10.1073/pnas.0507325102

Waddell PJ (2002) Very fast algorithms for evaluating the stability of ML and Bayesian phylogenetic trees from sequence data. Genome Inform 13:82–92

Waddell PJ, Kishino H, Ota R (2001) A phylogenetic foundation for comparative mammalian genomics. Genome Inform 12:141–154

Wakeley J (2002) Inferences about the structure and history of populations: coalescents and intraspecific phylogeography. In: Singh R, Uyenoyama M, Jain S (eds) The evolution of population biology—Modern synthesis. Cambridge University Press, Cambridge

Wakeley J (2004) Recent trends in population genetics: more data! more math! simple models? J Hered 95:397–405. doi:10.1093/jhered/esh062

Wang R, Wakeley J, Hey J (1997) Gene flow and natural selection in the origin of *Drosophila pseudoobscura* and close relatives. Genetics 147:1091–1106

Watterson GA (1978) The homozygosity test of neutrality. Genetics 88:405–417

Whitfield CW, Behura SK, Berlocher SH et al (2006) Thrice out of Africa: ancient and recent expansions of the honey bee, *Apis mellifera*. Science 314:642–645. doi:10.1126/science.1132772

Wilder JA, Kingan SB, Mobasher Z, Pilkington MM, Hammer MF (2004) Global patterns of human mitochondrial DNA and Y-chromosome structure are not influenced by higher migration rates of females versus males. Nat Genet 36:1122–1125. doi:10.1038/ng1428

Wilson AC, Cann RL, Carr SM et al (1985) Mitochondrial DNA and two perspectives on evolutionary genetics. Biol J Linn Soc 26:375–400. doi:10.1111/j.1095-8312.1985.tb02048.x

Wolfe KH, Li W-H, Sharp PM (1987) Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. Proc Natl Acad Sci USA 84:9054–9058. doi:10.1073/pnas.84.24.9054

Wu CI (1991) Inferences of species phylogeny in relation to segregation of ancient polymorphisms. Genetics 127:429–435

Zhang D-X, Hewitt GM (2003) Nuclear DNA analyses in genetic studies of populations: practice, problems and prospects. Mol Ecol 12:563–584. doi:10.1046/j.1365-294X.2003.01773.x

Zink RM, Barrowclough GF (2008) Mitochondrial DNA under siege in avian phylogeography. Mol Ecol 17:2107–2121. doi:10.1111/j.1365-294X.2008.03737.x