Provided for non-commercial research and education use.
Not for reproduction, distribution or commercial use.

This article appeared in a journal published by Elsevier. The attached copy is furnished to the author for internal non-commercial research and education use, including for instruction at the authors institution and sharing with colleagues.

Other uses, including reproduction and distribution, or selling or licensing copies, or posting to personal, institutional or third party websites are prohibited.

In most cases authors are permitted to post their version of the article (e.g. in Word or Tex form) to their personal website or institutional repository. Authors requiring further information regarding Elsevier's archiving and manuscript policies are encouraged to visit:

http://www.elsevier.com/authorsrights

Author's personal copy

Molecular Phylogenetics and Evolution 69 (2013) 1172–1185

Contents lists available at SciVerse ScienceDirect

Molecular Phylogenetics and Evolution

journal homepage: www.elsevier.com/locate/ympev

Review

Integrating phylogenetics, phylogeography and population genetics through genomes and evolutionary theory

Asher D. Cutter *

Department of Ecology & Evolutionary Biology, University of Toronto, 25 Willcocks St., Toronto, ON M5S 3B2, Canada

ARTICLE INFO

Article history:
Received 6 April 2013
Revised 6 June 2013
Accepted 12 June 2013
Available online 22 June 2013

Keywords:
Phylogeography
Species trees
Speciation
Genome evolution
Coalescent theory

ABSTRACT

Evolutionary theory is primed to synthesize microevolutionary processes with macroevolutionary divergence by taking advantage of multilocus multispecies genomic data in the molecular evolutionary analysis of biodiversity. While coalescent theory bridges across timescales to facilitate this integration, it is important to appreciate the assumptions, caveats, and recent theoretical advances so as to most effectively exploit genomic analysis. Here I outline the connections between population processes and phylogeny, with special attention to how genomic features play into underlying predictions. I discuss empirical and theoretical complications, and solutions, relating to recombination and multifurcating genealogical processes, predictions about how genome structure affects gene tree heterogeneity, and practical choices in genome sequencing and analysis. I illustrate the conceptual implications and practical benefits of how genomic features generate predictable patterns of discordance of gene trees and species trees along genomes, for example, as a consequence of how regions of low recombination and sex linkage interact with natural selection and with the accumulation of reproductive incompatibilities in speciation. Moreover, treating population genetic parameters as characters to be mapped onto phylogenies offers a new way to understand the evolutionary drivers of diversity within and differentiation between populations. Despite a number of challenges conferred by genomic information, the melding of phylogenetics, phylogeography and population genetics into integrative molecular evolution is poised to improve our understanding of biodiversity at all levels.

© 2013 Elsevier Inc. All rights reserved.

Contents

1. Introduction . . . 1173
2. Gene genealogies and gene summaries . . . 1173
   2.1. Gene tree heterogeneity . . . 1173
   2.2. Population genetics parameters and their interpretation . . . 1174
3. Intra-genomic patterns of gene tree heterogeneity . . . 1174
   3.1. Implications of within-genome variation in effective population size . . . 1175
   3.2. Within-genome variation in gene flow . . . 1176
   3.3. Effects of modes of chromosomal inheritance on gene tree heterogeneity . . . 1176
   3.4. Taking advantage of predictability of intra-genomic patterns of molecular evolution . . . 1176
4. Coalescent theory and biological reality . . . 1177
   4.1. Population subdivision in the present and past . . . 1177
   4.2. Population fission and fusion and the 'braided river' of genealogical history . . . 1177
   4.3. Pedigrees violate coalescent assumptions . . . 1177
   4.4. Recombination, reticulation and coalescence . . . 1177
   4.5. Exploiting recombination for biological interpretation . . . 1178
   4.6. Violating the Kingman coalescent: multifurcations in gene trees (and species trees) . . . 1179
5. The essence of time . . . 1179
   5.1. Units of time: years, mutational divergence, coalescence . . . 1179
   5.2. Natural timescales of molecular evolution . . . 1180

* Fax: +1 416 978 5878.
E-mail address: asher.cutter@utoronto.ca

1055-7903/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.ympev.2013.06.006

Author's personal copy

6. Is marker choice for inference rendered moot? . . . 1180
6.1. Molecular markers in phylogenetics, phylogeography, and population genetics . . . 1180
6.2. Markers from high-throughput sequencing . . . 1181
6.3. Benefits of single-copy nuclear protein coding genes . . . 1182
7. Comparative phylogenetics of phylogeographic and population genetic properties . . . 1182
8. Cautions and future directions . . . 1182
Acknowledgments . . . 1183
References . . . 1183

1. Introduction

Students of molecular phylogenetics, phylogeography and population genetics are broadly interested in understanding the same thing: evolutionary change of genomes and the organisms that host them. Where they differ is in emphasis, whether trying to draw inference about biodiversity from the interspecies genealogical histories of organisms, explicitly seeking to explain the demographic history of populations, or to characterize how natural selection gets recorded in DNA sequences – despite the evolutionary intercalation of phylogeny, demographic history, and natural selection. The generality of their common dependence on molecular sequence data provides a clear means of unifying phylogenetics, phylogeography and population genetics in a view of integrative molecular evolution. Recent and ongoing attempts have partially harmonized these subdisciplines (Edwards, 2009; Knowles, 2009). With renewed urgency, technical advances in DNA sequencing bring this integration to the fore (Carstens et al., 2012; McCormack et al., 2013): it is clear that 'phylogenomics' and 'population genomics' are the new standard for analysis. This is well exemplified by the impressive integrative work on our own species into how selection, demography and speciation history all interact throughout human and primate genomes (McVicker et al., 2009; Campbell and Tishkoff, 2010; The 1000 Genomes Project Consortium, 2012). The infiltration of genomics into these disciplines means that we can accelerate our understanding of biodiversity from micro- to macro-evolutionary scales. The genomic data itself provides the raw material to do so, exposing the predictable features of genome structure that must be accounted for, ripe for exploitation by researchers, provided that we more fully appreciate and incorporate the empirical and theoretical complexities that genomes help reveal.

My aim here is to raise attention to the key conceptual molecular evolutionary issues that bridge across phylogenetics, phylogeography and population genetics, with particular emphasis on recent progress in molecular population genetics and genomics that will prove valuable in the analysis and interpretation of phylogeography and phylogeny. In this review, I begin by introducing the special importance of gene tree heterogeneity and non-genealogical summaries to genome-scale studies (Section 2). I then describe how genome structure varies in predictable ways to affect gene tree heterogeneity, and its implications and potential to be exploited for understanding evolutionary history (Section 3). This is followed by empirical and theoretical caveats about genealogical coalescence, with emphasis on recent advances (Section 4), and about how to integrate evolutionary views of time across studies of differing depths of divergence (Section 5). I then outline practical genomic approaches to addressing problems in phylogenetics, phylogeography and population genetics (Section 6) before offering a view to ways in which integrative genomic analysis can span these subdisciplines (Section 7), with some cautions and areas of promise for future discoveries (Section 8).

2. Gene genealogies and gene summaries

The importance of considering many loci in phylogenetics, phylogeography and population genetics is now well-accepted. But there are different perspectives on how to integrate multi-locus data for inference. Gene trees provide a superb visual and quantitative way to consider the evolutionary process, but genome-scale data shines a bright light on some inherent, well-known challenges to this way of summarizing the evolution of populations. Non-genealogical summaries of molecular evolution provide complementary and alternative methods for some applications. In this section, to set the stage for more detailed issues, I outline some fundamental attributes of using gene trees and non-genealogical population genetic metrics for evaluating the many trajectories of evolution that get recorded across the genome.

2.1. Gene tree heterogeneity

The independent genealogical realizations of distinct loci provides a powerful way of determining the phylogeny of species relationships even when the genealogies of many or most loci do not reflect the true branching of speciation events in history (Edwards et al., 2007; Degnan and Rosenberg, 2009; Kubatko et al., 2009; Salichos and Rokas, 2013). In the face of speciation events clustered close in time or in the very recent past, different genes can differ in branch lengths or topology owing to heterogeneity in how or if allelic variation of a given locus in the species' common ancestor passes to the descendant lineages (Pamilo and Nei, 1988; Maddison, 1997), creating so-called 'anomalous gene trees' (Degnan and Rosenberg, 2006; Degnan and Rosenberg, 2009). The prevalence of gene tree heterogeneity depends on the time between speciation events relative to ancestral population sizes, with large ancestral population sizes and gene flow between incipient species near the time of ancestral nodes exacerbating such effects (Kubatko and Degnan, 2007; Eckert and Carstens, 2008).

Incorporating this heterogeneity in coalescence among loci can remove misleading inferences that could result from concatenation of loci and helps reveal the true 'species tree' in the face of both recent and ancient radiations (Knowles and Maddison, 2002; Edwards et al., 2007; Kubatko and Degnan, 2007; Edwards, 2009; Salichos and Rokas, 2013). This requires collecting data for many loci from multiple individuals in each species (or population), ideally, weighted by information content (Heled and Drummond, 2010; Lanier and Knowles, 2012; Salichos and Rokas, 2013). Just as coalescent theory has proved powerful in understanding evolution within populations (Hein et al., 2004; Wakeley, 2009), the 'multispecies coalescent' process has emerged as an important paradigm in understanding divergence between species (Degnan and Rosenberg, 2009; Edwards, 2009). This view simply makes explicit the idea that genealogies are a property of evolving populations that, owing to the stochasticity of genetic drift and recombination among their constituent individuals, can differ from locus to locus despite sharing an identical population history of speciation. This

Author's personal copy

speciation history is the 'species tree' phylogeny that phylogeneticists and phylogeographers want to figure out, and that population geneticists want to be able to take for granted. Many methods are now available to implement species tree inference or simulation from independent gene trees (Ewing et al., 2008; Liu, 2008; Kubatko et al., 2009; Kuhner, 2009; Liu et al., 2009; Heled and Drummond, 2010; Hey, 2010; Liu et al., 2010; Excoffier and Foll, 2011; Crisci et al., 2012; Heled et al., 2013; Boussau et al., 2013). This approach is especially important for taxa that radiated recently or over a short interval of time, even if in the very distant past, because ancestral polymorphism in common ancestors will more conspicuously create heterogeneous 'anomaly zone' gene tree topologies and branch lengths that generate discordance with the true species tree (Edwards and Beerli, 2000; Degnan and Rosenberg, 2006; Kubatko and Degnan, 2007).

2.2. Population genetics parameters and their interpretation

A multilocus view of evolution is well-entrenched in population genetics. And yet, in contrast to the centrality of gene trees in phylogeography and phylogenetics, molecular population genetics largely employs 'summary statistics' as the basis for evolutionary inference, the importance of coalescent genealogies notwithstanding (Hey and Machado, 2003; Crisci et al., 2012). Key summary statistics include the population mutation rate ($\theta = 4N_e\mu$), the population recombination rate ($\rho = 4N_ec$), substitution rates (d_N and d_S), and various metrics of the site frequency spectrum (e.g. Tajima's D (Tajima, 1989; Achaz, 2009)). All of these metrics have been explored in terms of how demographic and selection histories affect them on average across loci (Crisci et al., 2012); it is taken for granted that these metrics must be computed for many loci, because the genealogy and summary metrics for any one locus may differ wildly from the central tendency driven by the particular history of the species.

Because this multilocus viewpoint has now spread, leading to statistical phylogeography and the species tree paradigm (Knowles and Maddison, 2002; Edwards, 2009; Knowles, 2009), it is valuable to consider population genetic summary statistics further. Among them, θ is most central and most relevant to phylogeography and phylogenetics. It is important to remember that each ancestral node in a phylogeny represents a population, so we must keep in our minds' eye the population processes that impinge on that ancestral species. The parameter θ describes the amount of genetic diversity in a population at the balance between mutational input (μ) and loss through genetic drift, with drift being inversely related to population size (N_e) (Wright, 1931; Charlesworth, 2009). Because coalescent and diffusion theory tell us that the average time to the most recent common ancestor for a sample of alleles at a locus equals $4N_e$ generations (Fig. 1), empirical measures of θ also provide information about how much gene tree heterogeneity will be observed owing to ancestral polymorphism shared with related species (Knowles and Maddison, 2002; Edwards, 2009). This connection between polymorphism (θ) and coalescent time presumes selective neutrality, so population geneticists typically quantify θ separately for amino acid replacement sites (non-neutral) and synonymous sites (neutral) for coding genes. The analog for divergence between species is d_S, which is a codon-informed measure of sequence divergence (Yang, 1997). From a phylogenetic perspective, discriminating divergence at synonymous (d_S) and non-synonymous sites (d_N) helps account explicitly for among-site rate heterogeneity using a biological rationale, rather than applying, say, a GTR + I + γ mutation model across all sites (Posada and Crandall, 1998; Arbogast et al., 2002). For species with very large populations, however, there can be weak selection even between synonymous codons in highly expressed genes (Duret, 2002; Hershberg and Petrov, 2008). This drives home the point that selective neutrality depends on population size (i.e. $N_e s < 1$), so large populations will have a smaller fraction of their genome left unconstrained by selection, especially in non-coding parts of the genome. The lack of a defined genetic code to help specify constraints in non-coding regions makes it challenging to model their molecular evolution in relation to theory. The effect of population size influences protein evolution, as well. On average, genes evolve slower (i.e. lower d_N/d_S) in species with larger N_e and higher θ (Ellegren, 2009), which could contribute to among-species rate heterogeneity under a GTR + I + γ mutation model in phylogenies of taxa with widely disparate population sizes.

The site frequency spectrum (SFS) of populations provides a non-genealogical way to relate molecular differences to demographic and selective effects on genomes. The SFS can be visualized as the histogram of allele frequencies at nucleotide sites across the genome, or by simple per-locus summaries thereof (e.g. Tajima's D) (Tajima, 1989; Achaz, 2009). Population size changes and selection affect the shape of the SFS in well-defined ways relative to a standard neutral model of molecular evolution (Myers et al., 2008; Crisci et al., 2012). Recent work extends this idea to multiple populations or closely-related species (the 'joint SFS') (Gutenkunst et al., 2009; Nielsen et al., 2009), making this approach more attractive to mainstream phylogeographers. These approaches can use diffusion theory or maximum likelihood frameworks, coupled with efficient coalescent simulations, to estimate population demographic parameters from sequence polymorphism data (Gattepaille et al., 2013). Especially as researchers begin to collect genome-scale polymorphism data for phylogeographic analysis, non-genealogical SFS-based methods can prove to be a valuable complement to explicitly genealogical methods.

3. Intra-genomic patterns of gene tree heterogeneity

Genomes are not homogeneous. With genomic analysis in phylogenetics, phylogeography and population genetics, it is becoming increasingly important to appreciate and embrace structural

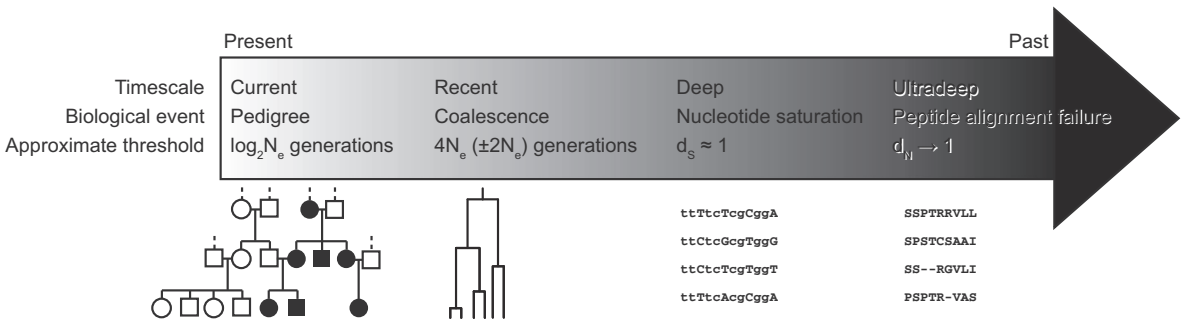

Fig. 1. The timescales of evolution. Evolution can be broken down into four broad timescales corresponding to key biological features associated with population genetic and molecular evolutionary processes. N_e = effective population size; d_S = number of synonymous-site substitutions per synonymous site; d_N = number of non-synonymous substitutions per non-synonymous site.

Author's personal copy

variation so as to arrive at the most appropriate biological interpretations. In this section, I describe three key features of genomic heterogeneity that will produce predictable effects on patterns of molecular evolution like gene tree heterogeneity. In turn, I discuss how we can think of effective population size and migration rates varying along chromosomes, how the mode of chromosomal inheritance affects population polymorphism and divergence between species, and how we can exploit these features.

3.1. Implications of within-genome variation in effective population size

The influx of phylogenomics and population genomics into the mainstream makes it important to understand and incorporate the biology of genome structure into our thinking about evolutionary analysis with molecular data. Some well-known population genetic phenomena can help make this connection. The genetic effective population size (N_e) of a real-world population reflects the census size (N_0) of an idealized reference population in which genetic drift operates at the same rate, so that N_e will not equal the census size in nature (N) to the extent that the real-world population has features that differ from that idealized 'Wright-Fisher' model population (e.g. inbreeding, biased sex ratios, population subdivision, natural selection) (Charlesworth, 2009). What may not be broadly recognized is that natural selection creates the appearance of heterogeneous N_e across the genome, such that those portions of the genome linked to a target of directional selection (both positive and negative) will effectively experience a smaller N_e (Charlesworth, 2009; Cutter and Payseur, 2013). This is used in some methods to scan the genome for signatures of recent adaptive evolution (Nielsen, 2005). Such perturbations from genetic hitchhiking effects are more extensive in regions of low recombination (e.g. near centromeres) and with higher gene density because more selected DNA is linked to a given locus (Cutter and Payseur, 2013). Consequently, gene trees of loci that occur in parts of the genome with little recombination will tend to have shorter coalescent depths (i.e. faster lineage sorting) (Fig. 2), with less likelihood of ancestral polymorphism and 'anomalous gene trees.' Loci that are more likely to be linked, as in regions of low recombination, also will tend to share the same gene tree. Thus, selection and recombination are expected to create predictable patterns along the genome in terms of gene tree heterogeneity and gene tree autocorrelation.

For a phylogeographic perspective, this effect of selection's effects on different parts of the genome also will result in population differentiation (e.g. F_{st}) tending to be greater in low recombination regions, even when there is no gene flow among the subpopulations (Charlesworth, 1998). Indeed, this has been observed empirically in a number of species (Begun and Aquadro, 1993; Carneiro et al., 2009; Keinan and Reich, 2010; Geraldes et al., 2011). The selection-linkage interaction also could yield a signature of molecular evolution that superficially looks like low recombination regions of the genome having a different population demography than genomic regions with high recombination. For example, this effect is observed in *C. elegans* as low recombination parts of chromosomes tend to have an excess of low frequency variants for the site frequency spectrum (Andersen et al., 2012), which a molecular demographer might otherwise attribute to a population expansion. These examples emphasize how integrating a deep familiarity with a genome's features helps to interpret evolutionary processes in molecular patterns.

Finally, I will mention the emerging population genetics literature devoted to the consequences of widespread adaptation across the genome for inferring recent demographic history and genetic

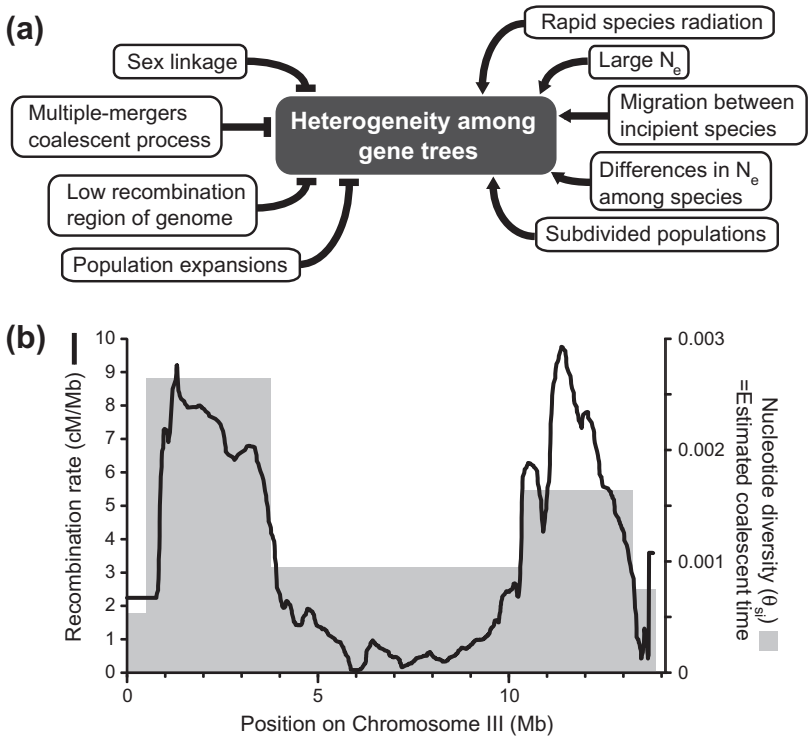

Fig. 2. Genome features affecting gene tree heterogeneity. (a) Heterogeneity among gene trees can be exacerbated by some factors (arrows) and mitigated by others (bars). (b) An example from *Caenorhabditis elegans* of recombination rate variation along a chromosome (black line, left axis), with corresponding differences in nucleotide diversity (gray bars, right axis) in different domains of high and low recombination rate. Nucleotide diversity is expected to be directly proportional to the time to the most recent common ancestor of the locus (coalescent time), which is inversely proportional to the likelihood of ancestral polymorphism generating discordance between gene trees and the species tree. Recombination rates calculated from physical and genetic map positions from Wormbase.org release WS220 as in Cutter et al. (2009). Nucleotide diversity values from Andersen et al. (2012).

Author's personal copy

targets of recent adaptation (Hahn, 2008; Olson-Manning et al., 2012). Specifically, there is growing appreciation for the possibility that some genome-wide patterns of polymorphism within populations that have been traditionally ascribed to demographic changes could be caused by pervasive selection, irrespective of chromosomal heterogeneity in recombination rate (Messer and Petrov, 2013). Moreover, selection can impinge on the genome in such a way that genealogical coalescence cannot be described with a single measure of effective population size (Neher, 2013). These issues present added challenges to disentangling causal recent evolutionary forces with molecular data.

3.2. Within-genome variation in gene flow

Hybridization and gene flow between incipient species, or between partially-isolated populations, also can generate predictable genomic heterogeneity in gene tree structure. Population genetic models of speciation predict that genes involved directly in reproductive isolation will occur disproportionately in genomic regions with low recombination (Butlin, 2005; Feder et al., 2012; Nachman and Payseur, 2012). Consequently, gene flow between incipient species will be reduced in those parts of the genome with restricted recombination, as observed empirically from plants to insects to vertebrates (Butlin, 2005; Nachman and Payseur, 2012), with correspondingly less propensity for discordance between gene trees and the population history of speciation. This provides another mechanism, in addition to the linked selection phenomenon described above, that contributes to parts of the genome experiencing higher recombination rates being more likely to have loci with 'anomalous gene trees.' Several approaches have been suggested to distinguish these alternative causes, with current evidence suggesting a prominent role of restricted gene flow for low recombination rate regions of the genome (Nachman and Payseur, 2012).

These issues are particularly compelling for phylogeographers interested in the process of speciation discussed as 'isolation with migration' (Becquet and Przeworski, 2009; Pinho and Hey, 2010; Strasburg and Rieseberg, 2010) or 'speciation with gene flow,' (Smith et al., 1997; Nosil et al., 2009; Sousa and Hey, 2013). Powerful methods are being developed to exploit population genomic data to identify and trace tracts of chromosomal haplotype blocks that are inherited in a common way, permitting a sensitive means of identifying gene flow, quantifying the amount and timing of gene flow between populations or closely-related species, and to visualize this process genealogically (Pool and Nielsen, 2009; Lawson et al., 2012). Such blocks of chromosomal haplotype structure also have been used in inference of very recent selection across the genome (Voight et al., 2006). As reference genomes become readily available, chromosome-scale patterns of molecular evolution also can help phylogeneticists to select loci for analysis that occur in genomic regions that should give less gene tree heterogeneity. Perhaps more importantly, understanding these issues will help interpretation of potential causes to heterogeneity among gene trees from features about where the loci occur in the genome, which will be increasingly important with phylogenomic analysis. For example, genomic regions with different amounts of gene flow should exhibit different degrees of among-locus heterogeneity in coalescent depths (Wakeley, 2003). With increasing divergence time in a dataset, however, structural features of the genome will themselves become less conserved across taxa, making it challenging to incorporate them into analyses in a simple way.

3.3. Effects of modes of chromosomal inheritance on gene tree heterogeneity

The mode of chromosomal inheritance forms another conspicuous axis of predictable genomic heterogeneity (Fig. 2). For example, the 4-fold difference in N_e between the mitochondrion and autosomal nuclear loci in diploid organisms provides one motivation encouraging animal phylogeographers to use mitochondrial markers to explore recent population and speciation histories, because coalescence will be 4-fold faster (Avise et al., 1987). In species with chromosomal sex determination, loci linked to X (and Z) chromosomes have ¾ the N_e of autosomal loci, but still benefit from recombination (unlike mitochondrial, chloroplast, Y and W chromosomes) (Charlesworth et al., 1987; Laporte and Charlesworth, 2002). As a result, we should expect gene tree heterogeneity to be less pervasive, and population differentiation to be stronger, for loci on the X chromosome compared to autosomes owing to the differential lineage sorting effects conferred by differences in N_e (Nachman and Payseur, 2012). The effective population recombination rate (ρ) of X-linked loci also is less than that of autosomal loci ($\rho_X = 3N_e c_X$ vs. $\rho_A = 4N_e c_A$), unless sex-averaged meiotic recombination rates compensate by being substantially higher on the X chromosome (i.e. $c_X > c_A$). Therefore, less gene flow conferred on low recombination regions of the genome through the mechanisms described above could also induce less shared polymorphism between populations and incipient species (Nachman and Payseur, 2012), with a correspondingly lower incidence of anomalous gene trees for X-linked loci. Z-linked genes in shorebirds show this advantage for phylogenetic reconstruction with species tree methods (Corl and Ellegren, 2013). An important caveat to the expected differences in N_e for distinct patterns of chromosomal inheritance is that it depends on the relative variance in reproductive success between males and females, which may differ across taxa (e.g. owing to differences in polygamy versus polygyny) (Charlesworth, 2001; Laporte and Charlesworth, 2002). Demographic size changes can influence the X and autosomes differently, which might be exploited for improved phylogeographic inferences about population growth (Pool and Nielsen, 2007; Pool and Nielsen, 2008).

Moreover, a 'large X effect' of a greater than expected density of reproductive incompatibility factors linked to the X chromosome is known in incipient species pairs and hybrid zones of a variety of organisms (Charlesworth et al., 1987; Coyne, 1992; Masly and Presgraves, 2007; Presgraves, 2008). Genes associated with reproductive isolation will more likely have genealogies that reflect the species tree (Ting et al., 2000), and this logic fuels genome-wide scans for 'speciation genes' that show particularly strong differentiation between close relatives (Rosenberg, 2002; Feder et al., 2012). As a result, this should lead to less gene flow of X-linked loci between closely-related species and populations, again yielding shorter coalescence times and fewer anomalous gene trees for loci located on the X chromosome. Haldane's rule, the phenomenon whereby the heterogametic sex of hybrids between incipient species will most commonly have reduced fitness, also is generally considered to involve sex chromosomes (Haldane, 1922; Orr, 1997). This could accentuate differences between autosomes and sex chromosomes. These X-autosome differences have empirical support in a variety of taxa, including mammals, birds and insects (Presgraves, 2008; Nachman and Payseur, 2012).

3.4. Taking advantage of predictability of intra-genomic patterns of molecular evolution

Concerted genomic patterns of gene tree heterogeneity, supported by underlying evolutionary bases, offer a predictive means of connecting across the levels of population genetic properties to phylogenetic phenomena. Genome scale data used in phylogeographic and phylogenetic studies requires this integration with evolutionary theory to fully understand the causes and consequences of gene tree heterogeneity. For groups of species with well-assembled genomes, chromosomes can be scanned for

Author's personal copy

patterns of incongruent gene trees in association with genomic features (Prasad et al., 2013). Reciprocally, the molecular population genetic objective of delineating selection along the lengths of chromosomes can benefit from further integration of phylogenetic divergence and within-population variation (Wilson et al., 2011).

4. Coalescent theory and biological reality

The embracing of a coalescent framework in phylogeography and phylogenetics represents an important achievement (Edwards, 2009; Knowles, 2009), but it also is important to acknowledge some of the assumptions underlying standard coalescent theory and recent advances. I will touch on just a few examples here, related to factors important for population subdivision within species, very recent-time pedigrees, intralocus recombination, and non-bifurcating gene trees. For brevity, I will largely omit discussion of several other intriguing and relevant factors, like horizontal gene transfer, hybridization, gene duplication and migration.

4.1. Population subdivision in the present and past

When a species is structured into many subpopulations, then it matters how individuals are sampled to recreate their coalescent history. Specifically, we expect allele copies to coalesce more quickly when they are sampled from within a subpopulation than between them (Wakeley, 1999); this has implications for erroneously estimating effective population size and expected coalescent times if individuals from different subpopulations are unintentionally pooled in analysis. Fortunately, this effect can be modeled (Wakeley, 1999), and different 'sampling schemes' of individuals can be exploited for phylogeographic and population genetic inference. One can extract the partially-independent information of different types of population samples by considering explicitly only those individuals from a single local subpopulation, pooling individuals from all subpopulations, or subsampling a single individual from each subpopulation (Wakeley and Aliacar, 2001; Stadler et al., 2009; Cutter et al., 2012). Generally, population subdivision with ongoing migration will create more variance among loci in coalescent depths than expected for a single population, whereas there will be less variation among loci for populations lacking gene flow, as for distinct species (Wakeley, 2003). Moreover, the presence of population structure within common ancestors will have a profound influence on expected coalescence times (Wakeley and Hey, 1997; Edwards and Beerli, 2000; Eriksson and Manica, 2012). The potential influence of ancestral population structure on species tree inference is one area requiring further investigation.

In coalescent modeling of speciation and species trees, some care must be taken in conceiving of population splitting events (speciation). In one sense, it is simplest to consider a scenario in which extant and ancestral populations all have the same population size (e.g. Knowles and Carstens, 2007). But looking backward in time in this context, at that moment of fusion of two species into one, this would yield a corresponding population size contraction (or, looking forward in time, a single population of size N_e would split into two populations each of size N_e, thus doubling the gene copies involved in the coalescent process at a polymorphic locus). Such an assumption about population size change at the time of speciation would lead to more rapid coalescence than one might otherwise expect. Alternatively, one could envision ancestral populations being the sum of the sizes of their daughters (Rosenberg, 2002). This circumvents the coupling of speciation and altered rate of coalescence, but would generate species with unrealistically large population sizes as we trace the species tree back in time through many common ancestors.

4.2. Population fission and fusion and the 'braided river' of genealogical history

Just as a single population can become fragmented into subpopulations, those subpopulations could fuse into a single population. Looking backward in time, the restricted gene flow among historical subpopulations relative to the single present-day population prolongs the time to coalescence, as does recombination within a population (Fig. 3). Graphically, the evolutionary history of populations could look much like a braided river, with a combination of reticulation owing to recombination and migration between subpopulations, the fission and fusion of subpopulations over time, and different tributary lineages ultimately coalescing in their common ancestors (Fig. 3). Incorporating ancestral population structure into species tree inference presents a continuing challenge that genome-scale data and new phylogeographic techniques may help to address (Wakeley and Hey, 1997; Edwards and Beerli, 2000; Lawson et al., 2012). Even with relatively simple demographic histories, however, it is difficult to estimate with confidence ancestral population sizes and migration rates (Jennings and Edwards, 2005; Becquet and Przeworski, 2009; Hey, 2010; Strasburg and Rieseberg, 2010; Heled et al., 2013), reinforcing the need for further development in this area.

4.3. Pedigrees violate coalescent assumptions

A conceptual limitation of coalescent theory is that it neglects the fact that alleles pass through a single pedigree of individuals (Wakeley et al., 2012). The fact of the pedigree violates the independence assumption that gives a $1/N_e$ probability of descent from a common ancestor each generation for a pair of random alleles, although this assumption proves to be a very good approximation for timescales $>\log_2 N_e$ generations ago (e.g. >17 generations for $N_e = 10^5$). As a consequence, demographic effects on very recent timescales will yield a higher variance on genealogies than expected under standard coalescent theory (Wakeley et al., 2012). This timescale is pertinent to some tests for contemporary adaptive evolution in genomes and for some questions in phylogeography, conservation, and landscape genetics (Manel et al., 2003; Allendorf et al., 2010; Storfer et al., 2010).

4.4. Recombination, reticulation and coalescence

Within the populations that comprise any given point along a species tree, recombination occurs among individuals. Recombination creates genealogical reticulation, as it will cause the focal piece of DNA to have more ancestors rather than less (Hudson, 1983; Nordborg, 2000). Consequently, a single unidirectional bifurcating gene tree will no longer accurately describe the relationships among ancestors and descendants; this is a good conceptual reason, among others, to avoid phylogenetic analysis of a concatenation of different genes and motivates species tree approaches to phylogenetic reconstruction (Brito and Edwards, 2009; Edwards, 2009; Salichos and Rokas, 2013). Recombination can also occur within loci, however, so a bifurcating gene tree diagram will give a misleading view of the history even within a locus for phylogenies (Schierup and Hein, 2000) and population variation (Huson and Bryant, 2006; Huson and Scornavacca, 2011).

Reticulating gene networks provide one graphical solution to this problem by making recombination visually explicit (Fig. 3), and tests for recombination make it explicit statistically (Hudson and Kaplan, 1985; Huson and Bryant, 2006; Huson and Scornavacca, 2011). Recombination also has been integrated into coalescent theory (the 'ancestral recombination graph') and into coalescent simulation programs, making it possible to test for effects of recombination (Hudson, 1983). Fortunately, within-locus

Author's personal copy

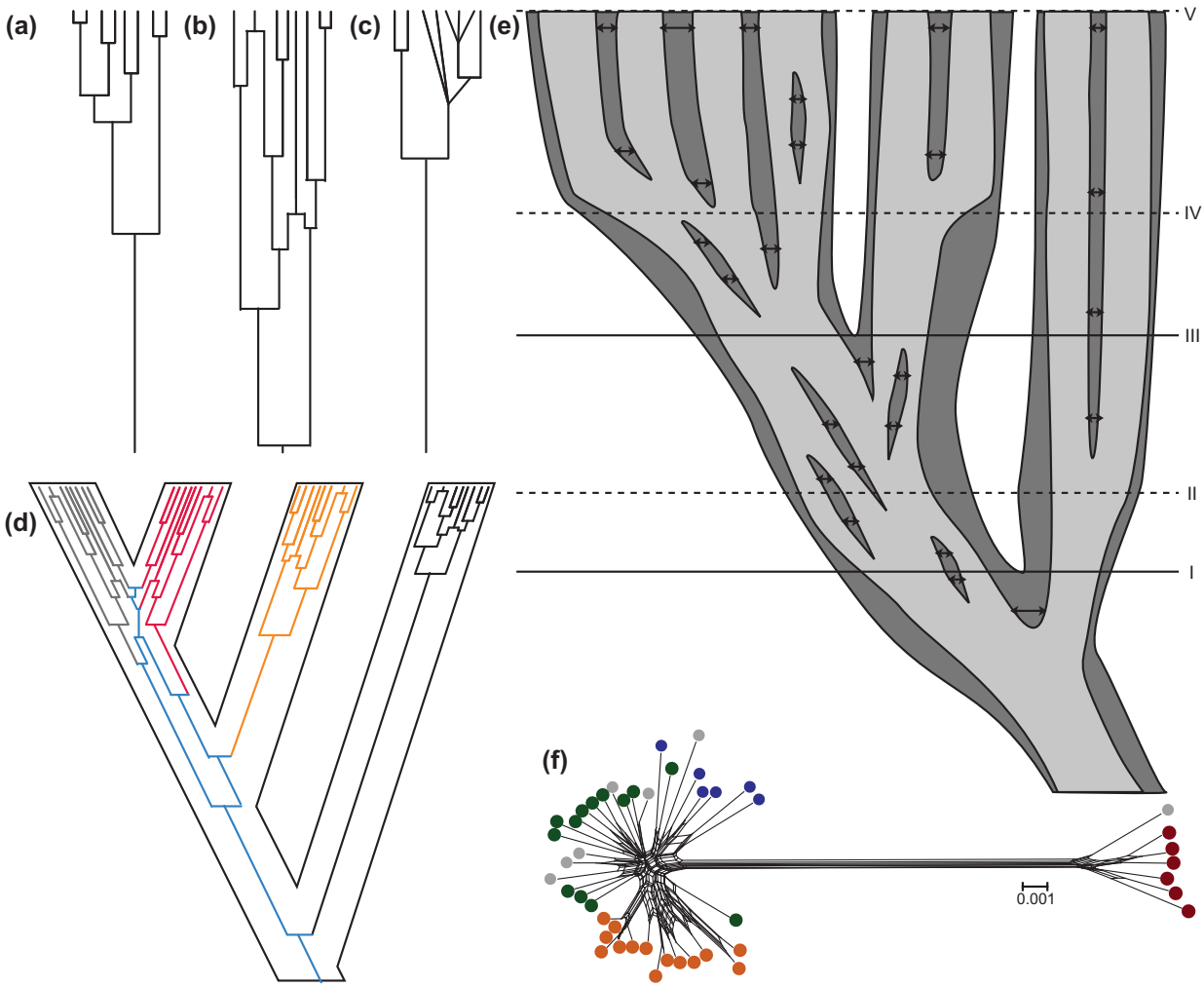

Fig. 3. Gene trees and species trees. Diagram of a standard bifurcating (a), reticulating (b) and multifurcating genealogy (c). Recombination tends to produce deeper coalescence (b), whereas multiple-mergers accelerate coalescence (c). (d) Ancestral polymorphism in a 4-taxon gene tree with recombination. (e) Current and historical population subdivision creates a braided evolutionary history when viewing the population tree (light gray) superimposed on the overall species tree (dark gray). Gene flow between subpopulations within a species at a given point in time is represented by horizontal arrows. The most ancestral speciation event occurred at time T_I, which was preceded by a period of gene flow between the incipient species' populations. The left-hand species was comprised of two populations at the time of speciation (T_I), but by time T_{II} populations had fused and split to yield three subpopulations. Another speciation event occurred at time T_{III}, with each of the new descendant species comprising a single population each. Times T_{IV} and T_V indicate the persistent population subdivision in the right-most species, whereas the left-most species history is marked by complex episodes of population fission and fusion. (f) Network diagram illustrating reticulation within and among 17 nuclear loci for *Caenorhabditis remanei* (left cluster of sequences with blue, green and orange points) and its divergence with cryptic species *C.* sp. 23 (right cluster of red points) (Dey et al., 2012). The different colors identify genetically differentiated subpopulations connected by gene flow. The strongly restricted gene flow between North America and Europe (blue, green, orange) with China (red) helped discriminate the cryptic species, which was subsequently confirmed with genetic crosses.

recombination has only a weak influence on species tree inference and is weaker in its effects for more ancient speciation events (Lanier and Knowles, 2012), although clearly this must depend on the length of the locus. At the extreme, we know concatenation of loci can mislead phylogenetic inference (Degnan and Rosenberg, 2006; Edwards et al., 2007; Kubatko and Degnan, 2007; Salichos and Rokas, 2013), so the potential for a problematic effect depends on the incidence of recombination making independent genetic histories relative to the information density captured in those genetic segments (i.e. in population genetic terms, it depends on ρ/θ). For some types of high throughput sequencing datasets, this will prove a serious concern. Some phylogeographic approaches analyze substrings of sequence of non-recombinant haplotype blocks, but this is not always feasible, especially for large populations having rapid decay of linkage disequilibrium along chromosomes. At the interspecies level, multi-locus methods for inferring the species tree phylogeny essentially integrate across different genealogies made independent by recombination between loci, pulling the signal of speciation history from the noise of stochastic coalescence.

4.5. Exploiting recombination for biological interpretation

Recombination in a dataset is sometimes viewed as a problem to be circumvented, but it can provide powerful information about population biology. We can take advantage of it to help inform a key aim of molecular phylogenetics and phylogeography: to use DNA sequence information to help discriminate cryptic species that defy obvious morphological differentiation. I will sidestep issues of species definition per se, except to recognize that it is infeasible in many cases to strictly (i.e. experimentally) apply the biological species concept for species delimitation and that it is indefensible to use a single non-recombining locus to impute species membership (Hudson and Turelli, 2003; Coyne and Orr, 2004). Consequently, genealogical evidence of interbreeding is a powerful tool for identifying partly-isolated populations and fully-isolated species (Coyne and Orr, 2004; Yang and Rannala, 2010). In sexual species, interbreeding is coupled with recombination, so it makes most sense biologically and conceptually to evaluate multiple unlinked loci for recombination and concordance of genealogical

Author's personal copy

history. Recombination affects many features that we can quantify in molecular data, including genealogical reticulation and linkage disequilibrium. *Caenorhabditis* nematodes provide a useful example of how recombination information in multilocus molecular data in a morphologically constrained group has helped both to reveal cryptic species (Dey et al., 2012) and to retain animals as a single species in spite of enormous molecular differences (Cutter et al., 2012; Dey et al., 2013) (Fig. 3). In the cryptic species example of *C. remanei* and *C.* sp. 23, patterns of linkage disequilibrium, genetic differentiation, and genetic networks for nearly 20 nuclear loci pointed to the presence of two distinct species despite the lack of morphological divergence of taxonomic characters or discrimination in simple laboratory mating tests (Dey et al., 2012). Subsequent genetic crosses showing strong F2 hybrid breakdown then validated the occurrence of cryptic species (Dey et al., 2012). On the other hand, analysis of more than 20 nuclear loci in *C. brenneri* showed extreme sequence polymorphism between alleles of magnitude nearly as great as the divergence between *C. remanei* and *C.* sp. 23 (Dey et al., 2013). Intriguingly, this observation cannot be explained by cryptic species, based on patterns of linkage disequilibrium, reticulating genetic networks, and genetic crosses, and so points to a single 'hyperdiverse' biological species with extensive gene flow and an enormous population size (Dey et al., 2013). This also shows how the notion of a standard yardstick of sequence divergence can fail to discriminate species boundaries, even within a focal taxonomic group. Biologically relevant demarcation of species boundaries with molecular sequence data has important implications for downstream analyses of diversification rates, particularly in groups with low resolution from morphology.

4.6. Violating the Kingman coalescent: multifurcations in gene trees (and species trees)

Standard coalescent theory (the 'Kingman coalescent' (Kingman, 1982)) presumes the idealized Wright-Fisher model of reproduction to yield the expectation of the probability of coalescence each generation of $1/N_e$ noted above (Wakeley, 2009). However, this assumption is not appropriate for species with highly skewed distributions of offspring among individuals, as is the case for some marine broadcast spawning organisms (Eldon and Wakeley, 2006; Der et al., 2012). Instead of the standard Kingman coalescent, this scenario of high stochastic variance in reproductive output can be modeled with a multiple-mergers coalescent process (Eldon and Wakeley, 2006; Der et al., 2012) (Fig. 3). That is, the standard coalescent assumption of a bifurcating gene tree is invalidated, even for a non-recombinant segment of DNA, so a true representation of the gene genealogy itself will be one with multifurcations (i.e. genealogical polytomies) (Eldon and Wakeley, 2006; Der et al., 2012).This alternative depiction of evolution leads to faster coalescence in genealogies, less population genetic diversity, and different site frequency spectra within populations than is found under standard coalescent theory (Eldon and Wakeley, 2006). This is most pertinent to species with very high stochastic variance in reproductive output (e.g. oysters with enormous numbers of larvae, of which a tiny fraction survive) (Hedgecock, 1994; Eldon and Wakeley, 2006). Although this population demography can be very important for understanding various features of phylogeographic history, it is not yet clear how important this effect might be on species tree inference.

Bifurcating genealogies also will differ from biological reality if there is a 'hard polytomy' in the species tree (i.e. multiple descendant species split simultaneously from a single common ancestor), because any given bifurcating gene tree will not recapitulate the species tree (Slowinski, 2001). The hypothesis of a hard polytomy may be assessed explicitly with many independent loci by testing for equal representation of all possible gene tree topologies at the given putatively polytomous node (Slowinski, 2001). However, if the species tree contains a hard polytomy and the descendant species differ in population size (N_e), then genealogies might strongly, yet spuriously, support a bifurcating species tree because genealogies of species with a small population will coalesce sooner (Fig. 4). Some species tree reconstruction methods can incorporate information on different extant population sizes into coalescent heterogeneity (Liu and Pearl, 2007; Heled and Drummond, 2010), although an area for future work will be to determine the sensitivity of species tree inference to population size differences among species. Changes in population size along a lineage also induce distinct departures from standard coalescent expectations for a population at equilibrium. For example, expanding populations will tend to show a more uniform coalescence time and contracting or structured populations will have more heterogeneous coalescence times among loci (Wakeley, 2003). Explicit simulations of the coalescent process among recently-separated species of differing size with software that can accommodate arbitrary histories (Hudson, 2002; Hey, 2010; Excoffier and Foll, 2011; Heled et al., 2013) could help test whether heterogeneous population sizes among species will affect species tree reconstruction.

5. The essence of time

The focal time depth can differ drastically among, or even within, studies of phylogeny, phylogeography and population genetics. To integrate the interpretation of patterns of molecular evolution from these perspectives, an explicit depiction of timescale is crucial. Because time can be represented in a variety of ways, here I outline how different units of time are linked to each other, identify evolutionarily-motivated temporal reference-points, and summarize how these issues impinge on analyses of molecular evolution from population variation through to deep phylogenetic history.

5.1. Units of time: years, mutational divergence, coalescence

When translating between the evolutionary levels of population dynamics and phylogenies, it is useful to clarify several points related to time; specifically, different units of time and different natural time scales. We are most personally familiar with time measured in years. Two evolutionarily motivated ways of measuring time, however, are in coalescent units (i.e. units of N_e

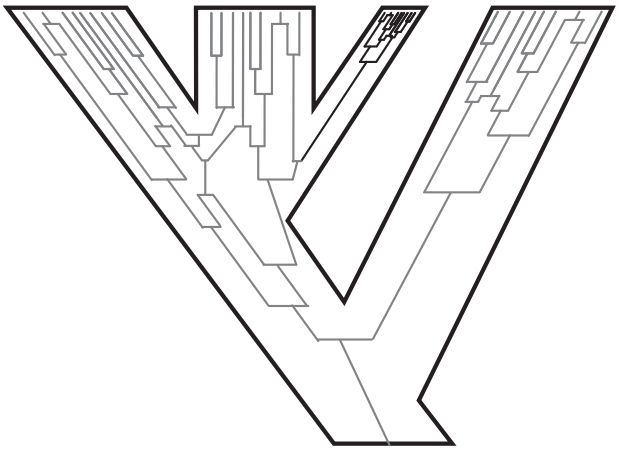

Fig. 4. Population size differences among species can affect gene tree topology and patterns of gene tree heterogeneity. For example, given a hard polytomy with one descendant population being much smaller, the faster coalescence in the small population could tend to produce gene trees giving strong support for a bifurcating species tree.

Author's personal copy

generations) and in mutational units (i.e. k substitutions). Mutational units are familiar to researchers working with DNA sequence data and connect to the neutral theory of molecular evolution through the concept of the molecular clock: neutral mutations will accumulate as fixed differences between species at the neutral mutation rate, so that given a per site mutation rate μ each generation, $k = \mu T$ substitutions will accrue on a lineage after T generations (Kimura, 1983). There are many well-known caveats and corrections necessary when applying this simple logic to actual empirical data that I will not delve into here, except to say that for recent timescales this kind of measure must account for polymorphisms (Pulquerio and Nichols, 2007). And, finally, it is coalescent units that are least familiar to most people, and yet crucial to integrating population genetics, phylogeography and phylogenetics. Standard coalescent theory in population genetics shows that the expected time to the most recent common ancestor for a sample of alleles at a diploid nuclear locus is $4N_e$ generations, with a standard deviation of $\sim 2N_e$ generations (Tavare, 1984; Wakeley, 2009). Consequently, using the effective population size as the base unit of time in generations provides a natural scale to connect to the genealogical process that generates divergence between independently evolving lineages. So, just as we can convert to calendar time from rates of radiometric decay for geologic dating of fossil beds, we can convert from biologically relevant timing of coalescent units, provided we know something about the generation times and population sizes of the species under consideration.

5.2. Natural timescales of molecular evolution

We can also think of several natural, if somewhat loosely defined, timescales of evolution (Fig. 1). First we have the 'current' timescale: what is happening now to individuals in the population and in their immediate ancestors that make up their pedigree over the last $\sim \log_2 N_e$ generations (i.e. within the last 5–20 generations) (Wakeley et al., 2012). Next we have the 'recent' timescale, extending to $\sim 4N_e$ generations into the past to the expected coalescent time of population variation. Most loci are expected to coalesce within $5N_e$ generations, making lineage sorting issues less important for species separated longer ago than this (Degnan and Rosenberg, 2009). However, population structure within a species could extend this recent coalescent timescale further into the past. Third are 'shallow' phylogenetic timescales up to the point at which silent sites become saturated with substitutions (i.e. divergence at synonymous sites, $d_S \sim 1$), with a 'deep' timescale comprised of species comparisons longer ago than this. Finally, we can consider 'ultradeep' phylogenetic depths at which even amino acid substitutions provide little resolution to phylogenetic reconstruction.

Timescale is important when testing for selection history and for weighing the role of ancestral polymorphism in species tree inference. For example, d_N/d_S ratios are most appropriate at timescales for which it may safely be assumed that fixed substitutions greatly outweigh polymorphisms within the species under consideration, but have not saturated; this approach is inappropriate for analyzing intraspecies polymorphism (Rocha et al., 2006; Kryazhimskiy and Plotkin, 2008). Explicit tests of more recent selection make use of a plethora of population genetic information, from contrasts of polymorphisms to fixed differences (e.g. HKA and MK tests), site frequency spectra (e.g. Fay & Wu's H), linkage disequilibrium and homozygosity, and combinations of these features (Nielsen, 2005; Crisci et al., 2012). Incomplete lineage sorting because of the coalescent process will be most pronounced at shorter timescales, leading to extensive heterogeneity among gene trees, although bursts of speciation over short timespans that occurred at deeper timescales also can similarly create gene tree heterogeneity (Edwards and Beerli, 2000; Edwards, 2009). Analysis of single-copy nuclear protein-coding genes provides the powerful advantage of a common yardstick of well-defined molecular change across loci and taxa, in the form of population polymorphism (θ_S) and divergence (d_S) as estimated at synonymous-sites. These measures connect most easily to assumptions of coalescent theory and a molecular clock.

6. Is marker choice for inference rendered moot?

High-throughput sequencing (HTS) technologies are rapidly settling debates over marker choice in phylogeography and phylogenetics, given their cost advantage per unit data (McCormack et al., 2013). However, genomic approaches shift the burden from data acquisition to data analysis, and this is a challenging transition for many researchers – especially given that bioinformatics tools for many evolutionary applications have not yet matured to the easy accessibility of time-tested software that has been the workhorse of molecular phylogenetics, phylogeography and population genetics (Excoffier and Heckel, 2006). Even with HTS approaches, we still face the competing demands of the fraction of the genome to include versus sample size of individuals per population or species (Lemmon et al., 2012; McCormack et al., 2013). And, many flavors of genome-scale sequencing have emerged (Ekblom and Galindo, 2011; McCormack et al., 2013). With all this in mind, we should consider what features will maximize the scope of biological insight and can ease the transition from traditional (PCR and Sanger sequencing) to modern (e.g. Illumina) molecular analysis for phylogenetics, phylogeography and population genetics.

6.1. Molecular markers in phylogenetics, phylogeography, and population genetics

Despite the population genomics revolution being more than 5 years old (Begun et al., 2007; Charlesworth, 2010), most published studies on molecular phylogeography and phylogeny continue to make use of mitochondrial markers (Fig. 5). Only about half of such studies appear to use nuclear loci, although their relative incidence is growing steadily (Fig. 5). Because this pattern may continue, it is worth reiterating several key issues associated with different molecular markers to help encourage the adoption of HTS approaches that will permit direct comparisons across diverse taxonomic groups. The limitations of focusing on the mitochondrial (or chloroplast) genome for phylogenetic and phylogeographic analysis are well-known (Hey and Machado, 2003; Ballard and Whitlock, 2004; Hurst and Jiggins, 2005; Balloux, 2010) and also apply to DNA barcoding (DeSalle et al., 2005), so I will simply underscore the importance of incorporating many genealogical realizations in species ancestry by using many unlinked nuclear loci. With respect to inferring population and phylogenetic history, the number of independent loci is the key unit of replication. With this aim in mind, many researchers opt to use microsatellite loci (a.k.a. simple sequence repeats), which tend to be highly polymorphic, or to use ribosomal subunit genes, for which PCR primers often work across many taxa. While valuable for some applications, these marker types suffer some unfortunate properties. Specifically, the mutational dynamics at these markers makes it difficult or impossible to compare many biologically relevant features across taxa, owing to the homoplasy and locus-specific mutational parameters of microsatellites and lack of a genetic code and propensity for concerted evolution in multi-copy ribosomal RNA genes. Single-copy protein-coding genes linked to nuclear chromosomes circumvent these drawbacks to mitochondrial, microsatellite and ribosomal loci. Consequently, HTS methods targeting coding genes will be valuable for a broad range of questions with practical utility across a broad range of timescales, and, will help ease transition to genome-scale analyses.

Author's personal copy

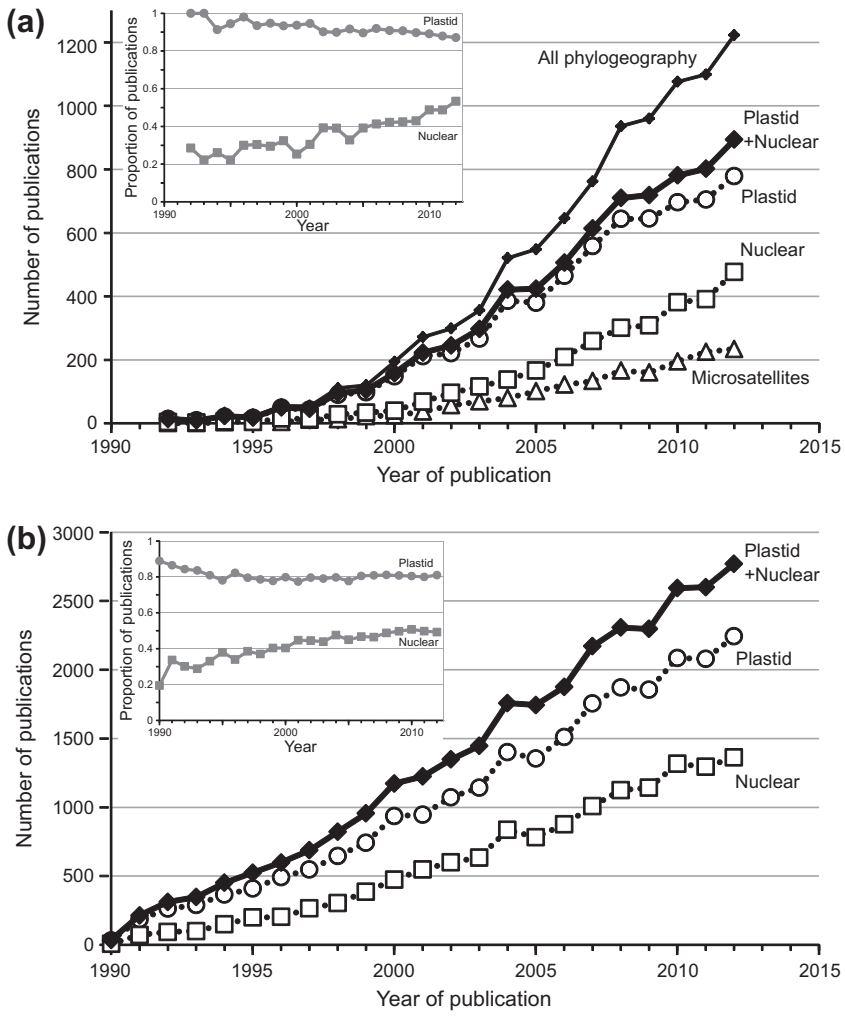

Fig. 5. Phylogeographic and phylogenetic studies using mitochondrial and chloroplast versus nuclear loci. (a) Use of plastid loci continues to outpace nuclear loci in phylogeographic publications, although the proportion of studies using plastid is declining (inset). Data based on ISI Web of Science queries (June 6, 2013) for "phylogeograph* and (mitochondri* or chloroplast)" (Plastid), "phylogeograph* and (nuclear or microsat* or repeat)" (Nuclear), "phylogeograph* and (mitochondri* or nuclear or microsat* or repeat)" (Plastid + Nuclear), "phylogeograph* and (microsat* or repeat)" (Microsatellites), "phylogeograph*" (All phylogeography). (b) Phylogenetic studies also implicate greater use of plastid loci, but with nuclear loci comprising a larger fraction of studies (inset). Data based on Web of Science queries (June 6, 2013) for "phylogen* and (mitochondri* or chloroplast)" (Plastid), "phylogen* and nuclear" (Nuclear), "phylogen* and (mitochondri* or chloroplast or nuclear)" (Plastid + Nuclear).

6.2. Markers from high-throughput sequencing

Technical aspects of HTS approaches for evolution have been reviewed thoroughly elsewhere (Glenn, 2011; McCormack et al., 2013), so here I just give the briefest synopsis. Conceptually, there are five key approaches relevant to our discussion: (i) whole genome sequencing, (ii) transcriptome sequencing (i.e. RNAseq) (De Wit et al., 2012), (iii) restriction digest sequencing (e.g. RADseq) (Hohenlohe et al., 2010), (iv) PCR amplicon sequencing, and (v) target enrichment sequencing (Mamanova et al., 2010; Faircloth et al., 2012; Lemmon et al., 2012). (i) Whole-genome sequencing is best suited to species with small, well-characterized genomes for which it is desired to delve into detailed molecular evolutionary analysis of many genomic features (Langley et al., 2012), and so it is unlikely to be the choice for most applications in phylogenetics and phylogeography. Low-coverage genome sequencing is a cost-effective strategy with utility for some questions in population genetics and phylogeography (Kulathinal et al., 2009; Lynch, 2009; Cutler and Jensen, 2010; The 1000 Genomes Project Consortium, 2010; Kofler et al., 2011), but is not suitable for deeper time depths.

The other four approaches provide ways of minimizing genomic overkill, each with benefits and drawbacks. (ii) Sequencing of transcriptomes is a powerful way to obtain information from many discrete loci (coding genes), although it is burdened with the costs and attention required of working with RNA as well as several complications of the huge span of gene expression levels, of potentially erroneous de novo assembly of genes for non-model organisms, and difficulties in distinguishing orthology, paralogy and allelism (De Wit et al., 2012; Ilut et al., 2012). Nevertheless, transcriptomes already have proved valuable for phylogenetics by improved branching resolution from the inclusion of hundreds or thousands of loci rather than just a handful (Dunn et al., 2008; Hittinger et al., 2010), albeit not incorporating within-species coalescence (Edwards, 2009). Nucleotide polymorphisms pertinent to phylogeographic and population genetic questions also can be extracted from transcriptome resequencing of many individuals from populations or species, but this bears similar complications (De Wit et al., 2012; Ilut et al., 2012; Gayral et al., 2013). (iii) RADseq and related approaches are a modern incarnation of AFLP (Hohenlohe et al., 2010; Lemmon and Lemmon, 2012). It benefits from randomly sampling many loci in the genome, but the selective context of those loci (i.e. non-coding vs. genic) is often unknown and orthologous loci do not extend well across species. (iv) Sequencing of PCR amplicons via HTS has the virtue of high control over what loci are to be used, but generally does not fully exploit

Author's personal copy

the high capacity and potential for low input overhead afforded by HTS (McCormack et al., 2013). (v) Recent developments in sequence capture with custom oligo-microarrays, followed by HTS, offer a powerful direction forward for phylogenetics and phylogeography (Faircloth et al., 2012; Lemmon et al., 2012; McCormack et al., 2012). This sequence target enrichment approach can share a key virtue of transcriptome sequencing in targeting coding genes, which make for very convenient units for downstream analysis (although any useful sequences may be targeted (McCormack et al., 2012)). The main drawback to sequence capture methods is the up-front effort and cost in developing the tiling arrays.

6.3. Benefits of single-copy nuclear protein coding genes

Population geneticists have typically quantified variation in a random collection of single-copy nuclear coding genes (Nordborg et al., 2005), or groups of such genes that share some key biological feature (e.g. chemosensory genes, genes involved in small RNA processing, or genes with reproductive functions) (Swanson and Vacquier, 2002; Jovelin, 2009; Obbard et al., 2009). Phylogeographers generally avoid such loci, in favor of loci with a greater density of polymorphisms (e.g. introns, microsatellites, mitochondrial or chloroplast genes), and phylogeneticists avoid them in favor of loci more likely to have universal PCR primers (e.g. ribosomal subunits). Why do population geneticists love coding genes, given some clear practical compromises? Perhaps the key benefit is that the genetic code imposes a clear model of molecular evolution that distinguishes amino-acid replacement sites from synonymous sites, with a straightforward biological interpretation and connection to theory. This allows explicit tests for selection rather than presumption of neutrality; the selective forces on non-coding or RNA gene sequences are much more challenging to specify. Given how pervasive selection can be, even in non-coding regions, this is a serious issue (Andolfatto, 2005; Hahn, 2008). Polymorphism and divergence at synonymous sites provides a natural yardstick for comparison across all species, although care must be taken at deep time depths when nucleotide site saturation sets in and for species with very large population sizes in which even synonymous sites can be subject to constraint (Cutter et al., 2013). With neutral evolution of synonymous sites, the neutral theory of molecular evolution can be applied in sophisticated ways. And by taking for granted the inclusion of many loci, there is explicit recognition of the heterogeneity among loci in coalescent histories. Tests for within and between locus recombination and linkage disequilibrium provide further windows into demographic and selective history beyond what is afforded by nucleotide differences, and the recognition of reticulating genealogies motivates site frequency spectrum analysis rather than imposing bifurcating genealogies on the loci (Crisci et al., 2012).

Here I have extolled some virtues of using many single-copy protein-coding genes, and HTS approaches that target them, in analyses of phylogeny, phylogeography and population genetics. With the genomic scope now afforded by HTS, I think these benefits will generally outweigh their lower density of polymorphisms for the purposes desired for phylogeographic and phylogenetic reconstruction. Coding genes also are amenable to genomic analysis of duplication and loss in a framework that integrates gene trees and the species tree (Boussau et al., 2013). However, we must also distinguish the roles of first-pass and final-word studies. HTS approaches certainly are a cost effective way of gaining an enormous dataset with high statistical power. But they require a daunting initial outlay of investment for many researchers. Consequently, there is still an important place for smaller scale analysis to glimpse the basic trends in a species (or group of species) to identify potentially interesting patterns worthy of, or requiring, further hypothesis testing on a genomic scale. For this purpose, mitochondrial or any convenient marker makes sense for a preliminary analysis.

7. Comparative phylogenetics of phylogeographic and population genetic properties

As genome-scale phylogeographic analysis takes root, an exciting direction will be to integrate intra-species population processes into a phylogenetic framework. In other words, we can begin to map population genetic parameters about demography and selection onto phylogenies to characterize their evolution and test for correlated trait evolution (Lynch, 2007; Lartillot and Poujol, 2011; Drummond et al., 2012; Leffler et al., 2012). Phylogenetic tests for correlated trait evolution attempt to identify the common factors that drive the evolution of traits across many groups of organisms (Harvey and Pagel, 1991). This 'comparative population genetics' contrasts with current applications of 'comparative phylogeography' that test for consistent genealogical splits among different species with overlapping ranges that could be explained by common biogeographic factors (Arbogast and Kenagy, 2001). To effectively treat population genetic parameters as phylogenetic characters requires data that is directly comparable across species. Hence, it will be invaluable to use molecular features with uniform interpretability, such as for single-copy protein-coding genes. Researchers have already started this enterprise, for example, by mapping effective population size (i.e. θ from synonymous-site nucleotide polymorphism) onto phylogenies to test for ecological triggers of genetic diversity (Leffler et al., 2012; Hazzouri et al., 2013) and by testing for a correlation between population size and rates of adaptive protein substitution (Gossmann et al., 2012) and mutation rate (Sung et al., 2012). However, such studies need to incorporate appropriate phylogenetically independent contrasts as standard operating procedure, as in modern analysis of phenotypic character evolution (Garland et al., 1992). Further work could test for how other metrics summarizing demographic and selection features are distributed along phylogenies, such as for intra-specific population differentiation (e.g. F_{st}), N_e/N ratios (Frankham, 1995; Palstra and Ruzzante, 2008), bottleneck propensity, or intra-genomic patterns of selection (Cutter and Payseur, 2013) to test explicit hypotheses about ecological or other drivers. Scaling up to phylogeographic studies on whole-genomes would provide additional intriguing characters to map onto phylogenies, building off of such tests as those relating effective population size (estimated from nucleotide polymorphism) with genome size and other attributes of genome architecture (Lynch and Conery, 2003; Lynch, 2007; Whitney et al., 2010; Whitney and Garland, 2010; Ai et al., 2012; Boussau et al., 2013).

8. Cautions and future directions

Sequencing technology is constantly improving, with the next phase anticipated to include long-read HTS (each read being many kilobases long) in contrast to current short-read HTS methods (each read a few hundred bases long) (Niedringhaus et al., 2011). This will surely be a boon for evolutionists of all stripes to ramp up to full genome analysis, but will simultaneously present challenges to practitioners of phylogenetics and phylogeography. The locus-centric view will begin to crumble as sequence lengths render untenable standard assumptions about recombination in datasets: analytical methods typically presume either perfect linkage of markers or complete independence of markers. In the face of chromosome-length sequence data, the notion of a focal locus becomes especially nebulous and arbitrary. Moreover, we must grapple with how to accommodate autocorrelation of molecular evolution along lengths of chromosomes owing to partial linkage. If clever, we can use the fact of partial linkage to improve our understanding of the history of demography, selection and divergence (Pool and Nielsen, 2009; Lawson et al., 2012). Fortunately, genomic data provide the

Author's personal copy

raw material to incorporate the true complexities of genome structure and the evolutionary process.

In the meantime, researchers must balance the financial and logistical factors with their grand notions of an ideal genomic dataset. For many, it is a challenge to match the new standard of phylogenomics and population genomics with adequate computational infrastructure, bioinformatics skillsets, funding, and biological samples. Pooling of DNA from many individuals (without individual barcode tags) used in HTS is valuable for a number of population genetic questions (Cutler and Jensen, 2010; Kofler et al., 2011); it remains to be seen whether this approach could be exploited for application to species tree phylogenetic reconstruction. With the new ease by which we can 'grind and find' genomes, as with allozymes in the 1970s and mtDNA in the 1990s, what is the right balance between discovery and hypothesis testing? This depends on how much researchers value the solutions to figuring out how *does* evolution work (process) vs. how *did* evolution work (pattern). In methodological choices to address population genetic and phylogeographic problems, it will be especially valuable to keep an eye toward those that are most suitable for future meta-analysis to map directly-comparable population features onto phylogenies.

While species tree methods for phylogenetic and phylogeographic inference represent a powerful way forward (Edwards, 2009; Knowles, 2009; Crisci et al., 2012; Drummond et al., 2012), they remain to be tested for sensitivity to a number of biological and theoretical realities (Fig. 2). These include issues of ancestral population structure, differences among species in population size and demographic change, and biological features that violate the Kingman coalescent. With whole-genome data, autocorrelation of gene trees along chromosomes provides a feature that could be implemented to improve species tree inference.

All of the world's biodiversity originated as genomic variability within populations that has since partitioned into the divergence between populations and species that we see today. This is what all phylogeneticists, phylogeographers and population geneticists seek to understand. It may seem to phylogeneticists interested in deep- and ultradeep-time diversification that population processes are of little relevance. I argue that thinking from the perspective of the populations living at even those deeply ancestral nodes, in terms of the microevolutionary processes associated with speciation, is invaluable to understand the often complex patterns in data so as to better interpret biological diversification. Characterizing the joint effects of selection and demography and phylogeny are increasingly important for making sense of genome-scale data, as is clear from the beachhead of work on our own species (McVicker et al., 2009; Campbell and Tishkoff, 2010; The 1000 Genomes Project Consortium, 2012). It is therefore essential that these subdisciplines weave into a unified view of integrative molecular evolution as our empirical ability to interrogate genomes has caught up with, or surpassed, existing evolutionary theory.

Acknowledgments

I am grateful to the Center for Systems Biology and the Department of Organismic and Evolutionary Biology at Harvard University for their accommodations. A.D.C. is supported by the Natural Sciences and Engineering Research Council of Canada and by a Canada Research Chair.

References

Achaz, G., 2009. Frequency spectrum neutrality tests: one for all and all for one. Genetics 183, 249–258.
Ai, B., Wang, Z.S., Ge, S., 2012. Genome size is not correlated with effective population size in the Oryza species. Evolution 66, 3302–3310.
Allendorf, F.W., Hohenlohe, P.A., Luikart, G., 2010. Genomics and the future of conservation genetics. Nat. Rev. Genet. 11, 697–709.
Andersen, E.C., Gerke, J.P., Shapiro, J.A., Crissman, J.R., Ghosh, R., Bloom, J.S., Felix, M.A., Kruglyak, L., 2012. Chromosome-scale selective sweeps shape *Caenorhabditis elegans* genomic diversity. Nat. Genet. 44, 285–290.
Andolfatto, P., 2005. Adaptive evolution of non-coding DNA in Drosophila. Nature 437, 1149–1152.
Arbogast, B.S., Kenagy, G.J., 2001. Comparative phylogeography as an integrative approach to historical biogeography. J. Biogeogr. 28, 819–825.
Arbogast, B.S., Edwards, S.V., Wakeley, J., Beerli, P., Slowinski, J.B., 2002. Estimating divergence times from molecular data on phylogenetic and population genetic timescales. Annu. Rev. Ecol. Syst. 33, 707–740.
Avise, J.C., Arnold, J., Ball, R.M., Bermingham, E., Lamb, T., Neigel, J.E., Reeb, C.A., Saunders, N.C., 1987. Intraspecific phylogeography: the mitochondrial-DNA bridge between population-genetics and systematics. Annu. Rev. Ecol. Syst. 18, 489–522.
Ballard, J.W.O., Whitlock, M.C., 2004. The incomplete natural history of mitochondria. Mol. Ecol. 13, 729–744.
Balloux, F., 2010. The worm in the fruit of the mitochondrial DNA tree. Heredity 104, 419–420.
Becquet, C., Przeworski, M., 2009. Learning about modes of speciation by computational approaches. Evolution 63, 2547–2562.
Begun, D.J., Aquadro, C.F., 1993. African and North American populatoins of *Drosophila melanogaster* are very different at the DNA level. Nature 365, 548–550.
Begun, D.J., Holloway, A.K., Stevens, K., Hillier, L.W., Poh, Y.P., Hahn, M.W., Nista, P.M., Jones, C.D., Kern, A.D., Dewey, C.N., Pachter, L., Myers, E., Langley, C.H., 2007. Population genomics: whole-genome analysis of polymorphism and divergence in *Drosophila simulans*. PLoS Biol. 5, e310.
Boussau, B., Szollosi, G.J., Duret, L., Gouy, M., Tannier, E., Daubin, V., 2013. Genome-scale coestimation of species and gene trees. Genome Res. 23, 323–330.
Brito, P.H., Edwards, S.V., 2009. Multilocus phylogeography and phylogenetics using sequence-based markers. Genetica 135, 439–455.
Butlin, R.K., 2005. Recombination and speciation. Mol. Ecol. 14, 2621–2635.
Campbell, M.C., Tishkoff, S.A., 2010. The evolution of human genetic and phenotypic variation in Africa. Curr. Biol. 20, R166–R173.
Carneiro, M., Ferrand, N., Nachman, M.W., 2009. Recombination and speciation: loci near centromeres are more differentiated than loci near telomeres between subspecies of the European rabbit (*Oryctolagus cuniculus*). Genetics 181, 593–606.
Carstens, B., Lemmon, A.R., Lemmon, E.M., 2012. The promises and pitfalls of next-generation sequencing data in phylogeography. Syst. Biol. 61, 713–715.
Charlesworth, B., 1998. Measures of divergence between populations and the effect of forces that reduce variability. Mol. Biol. Evol. 15, 538–543.
Charlesworth, B., 2001. The effect of life-history and mode of inheritance on neutral genetic variability. Genet. Res. 77, 153–166.
Charlesworth, B., 2009. Fundamental concepts in genetics: effective population size and patterns of molecular evolution and variation. Nat. Rev. Genet. 10, 195–205.
Charlesworth, B., 2010. Molecular population genomics: a short history. Genet. Res. 92, 397–411.
Charlesworth, B., Coyne, J.A., Barton, N.H., 1987. The relative rates of evolution of sex-chromosomes and autosomes. Am. Nat. 130, 113–146.
Corl, A., Ellegren, H., 2013. Sampling strategies for species trees: the effects on phylogenetic inference of the number of genes, number of individuals, and whether loci are mitochondrial, sex-linked, or autosomal. Mol. Phylogenet. Evol. 67, 358–366.
Coyne, J.A., 1992. Genetics and speciation. Nature 355, 511–515.
Coyne, J.A., Orr, H.A., 2004. Speciation. Sinauer, Sunderland, MA.
Crisci, J.L., Poh, Y.P., Bean, A., Simkin, A., Jensen, J.D., 2012. Recent progress in polymorphism-based population genetic inference. J. Hered. 103, 287–296.
Cutler, D.J., Jensen, J.D., 2010. To pool, or not to pool? Genetics 186, 41–43.
Cutter, A.D., Payseur, B.A., 2013. Genomic signatures of selection at linked sites: unifying the disparity among species. Nat. Rev. Genet. 14, 262–274.
Cutter, A.D., Dey, A., Murray, R.L., 2009. Evolution of the *Caenorhabditis elegans* genome. Mol. Biol. Evol. 26, 1199–1234.
Cutter, A.D., Wang, G.-X., Ai, H., Peng, Y., 2012. Influence of finite-sites mutation, population subdivision and sampling schemes on patterns of nucleotide polymorphism for species with molecular hyperdiversity. Mol. Ecol. 21, 1345–1359.
Cutter, A.D., Jovelin, R., Dey, A., 2013. Molecular hyperdiversity and evolution in very large populations. Mol. Ecol. 22, 2074–2095.
De Wit, P., Pespeni, M.H., Ladner, J.T., Barshis, D.J., Seneca, F., Jaris, H., Therkildsen, N.O., Morikawa, M., Palumbi, S.R., 2012. The simple fool's guide to population genomics via RNA-Seq: an introduction to high-throughput sequencing data analysis. Mol. Ecol. Res. 12, 1058–1067.
Degnan, J.H., Rosenberg, N.A., 2006. Discordance of species trees with their most likely gene trees. PLoS Genet. 2, e68.
Degnan, J.H., Rosenberg, N.A., 2009. Gene tree discordance, phylogenetic inference and the multispecies coalescent. Trends Ecol. Evol. 24, 332–340.
Der, R., Epstein, C., Plotkin, J.B., 2012. The dynamics of neutral and selected alleles when the offspring distribution is skewed. Genetics 191, 1331–1344.
DeSalle, R., Egan, M.G., Siddall, M., 2005. The unholy trinity: taxonomy, species delimitation and DNA barcoding. Philos. Trans. R. Soc. B 360, 1905–1916.
Dey, A., Jeon, Y., Wang, G.-X., Cutter, A.D., 2012. Global population genetic structure of *Caenorhabditis remanei* reveals incipient speciation. Genetics 191, 1257–1269.

Author's personal copy

Dey, A., Chan, C.K.-W., Thomas, C.G., Cutter, A.D., in press. Nucleotide hyperdiversity defines populations of *Caenorhabditis brenneri*. Proc. Natl. Acad. Sci. USA.

Drummond, A.J., Suchard, M.A., Xie, D., Rambaut, A., 2012. Bayesian phylogenetics with BEAUti and the BEAST 1.7. Mol. Biol. Evol. 29, 1969–1973.

Dunn, C.W., Hejnol, A., Matus, D.Q., Pang, K., Browne, W.E., Smith, S.A., Seaver, E., Rouse, G.W., Obst, M., Edgecombe, G.D., Sorensen, M.V., Haddock, S.H.D., Schmidt-Rhaesa, A., Okusu, A., Kristensen, R.M., Wheeler, W.C., Martindale, M.Q., Giribet, G., 2008. Broad phylogenomic sampling improves resolution of the animal tree of life. Nature 452, 745–749.

Duret, L., 2002. Evolution of synonymous codon usage in metazoans. Curr. Opin. Genet. Dev. 12, 640–649.

Eckert, A.J., Carstens, B.C., 2008. Does gene flow destroy phylogenetic signal? The performance of three methods for estimating species phylogenies in the presence of gene flow. Mol. Phylogenet. Evol. 49, 832–842.

Edwards, S.V., 2009. Is a new and general theory of molecular systematics emerging? Evolution 63, 1–19.

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

Edwards, S.V., Liu, L., Pearl, D.K., 2007. High-resolution species trees without concatenation. Proc. Natl. Acad. Sci. USA 104, 5936–5941.

Ekblom, R., Galindo, J., 2011. Applications of next generation sequencing in molecular ecology of non-model organisms. Heredity 107, 1–15.

Eldon, B., Wakeley, J., 2006. Coalescent processes when the distribution of offspring number among individuals is highly skewed. Genetics 172, 2621–2633.

Ellegren, H., 2009. A selection model of molecular evolution incorporating the effective population size. Evolution 63, 301–305.

Eriksson, A., Manica, A., 2012. Effect of ancient population structure on the degree of polymorphism shared between modern human populations and ancient hominins. Proc. Natl. Acad. Sci. U.S.A. 109, 13956–13960.

Ewing, G., Ebersberger, I., Schmidt, H., von Haeseler, A., 2008. Rooted triple consensus and anomalous gene trees. BMC Evol. Biol. 8, 118.

Excoffier, L., Foll, M., 2011. Fastsimcoal: a continuous-time coalescent simulator of genomic diversity under arbitrarily complex evolutionary scenarios. Bioinformatics 27, 1332–1334.

Excoffier, L., Heckel, G., 2006. Computer programs for population genetics data analysis: a survival guide. Nat. Rev. Genet. 7, 745–758.

Faircloth, B.C., McCormack, J.E., Crawford, N.G., Harvey, M.G., Brumfield, R.T., Glenn, T.C., 2012. Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales. Syst. Biol. 61, 717–726.

Feder, J.L., Gejji, R., Yeaman, S., Nosil, P., 2012. Establishment of new mutations under divergence and genome hitchhiking. Philos. Trans. R. Soc. B 367, 461–474.

Frankham, R., 1995. Effective population-size/adult-population size ratios in wildlife: a review. Genet. Res. 66, 95–107.

Garland, T., Harvey, P.H., Ives, A.R., 1992. Procedures for the analysis of comparative data using phylogenetically independent contrasts. Syst. Biol. 41, 18–32.

Gattepaille, L.M., Jakobsson, M., Blum, M.G.B., 2013. Inferring population size changes with sequence and SNP data: lessons from human bottlenecks. Heredity 110, 409–419.

Gayral, P., Melo-Ferreira, J., Glemin, S., Bierne, N., Carneiro, M., Nabholz, B., Lourenco, J.M., Alves, P.C., Ballenghien, M., Faivre, N., Belkhir, K., Cahais, V., Loire, E., Bernard, A., Galtier, N., 2013. Reference-free population genomics from next-generation transcriptome data and the vertebrate-invertebrate gap. PLoS Genet. 9.

Geraldes, A., Basset, P., Smith, K.L., Nachman, M.W., 2011. Higher differentiation among subspecies of the house mouse (*Mus musculus*) in genomic regions with low recombination. Mol. Ecol. 20, 4722–4736.

Glenn, T.C., 2011. Field guide to next-generation DNA sequencers. Mol. Ecol. Res. 11, 759–769.

Gossmann, T.I., Keightley, P.D., Eyre-Walker, A., 2012. The effect of variation in the effective population size on the rate of adaptive molecular evolution in eukaryotes. Genome Biol. Evol. 4, 658–667.

Gutenkunst, R.N., Hernandez, R.D., Williamson, S.H., Bustamante, C.D., 2009. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. PLoS Genet. 5, e1000695.

Hahn, M.W., 2008. Toward a selection theory of molecular evolution. Evolution 62, 255–265.

Haldane, J.B.S., 1922. Sex ratio and unisexual sterility in hybrid animals. J. Genet. 12, 101–109.

Harvey, P.H., Pagel, M.D., 1991. The Comparative Method in Evolutionary Biology. Oxford University Press, New York.

Hazzouri, K.M., Escobar, J.S., Ness, R.W., Killian Newman, L., Randle, A.M., Kalisz, S., Wright, S.I., 2013. Comparative population genomics in *Collinsia* sister species reveals evidence for reduced effective population size, relaxed selection, and evolution of biased gene conversion with an ongoing mating system shift. Evolution 67, 1263–1278.

Hedgecock, D., 1994. Does variance in reproductive success limit effective population sizes of marine organisms? In: Beaumont, A.R. (Ed.), Genetics and Evolution of Aquatic Organisms. Chapman & Hall, London, pp. 122–134.

Hein, J., Schierup, M.H., Wiuf, C., 2004. Gene Genealogies, Variation and Evolution: A Primer in Coalescent Theory. Oxford University Press, Oxford.

Heled, J., Drummond, A.J., 2010. Bayesian inference of species trees from multilocus data. Mol. Biol. Evol. 27, 570–580.

Heled, J., Bryant, D., Drummond, A., 2013. Simulating gene trees under the multispecies coalescent and time-dependent migration. BMC Evol. Biol. 13, 44.

Hershberg, R., Petrov, D.A., 2008. Selection on codon bias. Annu. Rev. Genet. 42, 287–299.

Hey, J., 2010. Isolation with migration models for more than two populations. Mol. Biol. Evol. 27, 905–920.

Hey, J., Machado, C.A., 2003. The study of structured populations – new hope for a difficult and divided science. Nat. Rev. Genet. 4, 535–543.

Hittinger, C.T., Johnston, M., Tossberg, J.T., Rokas, A., 2010. Leveraging skewed transcript abundance by RNA-Seq to increase the genomic depth of the tree of life. Proc. Natl. Acad. Sci. USA 107, 1476–1481.

Hohenlohe, P.A., Bassham, S., Etter, P.D., Stiffler, N., Johnson, E.A., Cresko, W.A., 2010. Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. PLoS Genet. 6, e1000862.

Hudson, R.R., 1983. Properties of a neutral allele model with intragenic recombination. Theor. Popul. Biol. 23, 183–201.

Hudson, R.R., 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics 18, 337–338.

Hudson, R.R., Kaplan, N.L., 1985. Statistical properties of the number of recombination events in the history of a sample of DNA-sequences. Genetics 111, 147–164.

Hudson, R.R., Turelli, M., 2003. Stochasticity overrules the "three-times rule": genetic drift, genetic draft, and coalescence times for nuclear loci versus mitochondrial DNA. Evolution 57, 182–190.

Hurst, G.D.D., Jiggins, F.M., 2005. Problems with mitochondrial DNA as a marker in population, phylogeographic and phylogenetic studies: the effects of inherited symbionts. Proc. R. Soc. B 272, 1525–1534.

Huson, D.H., Bryant, D., 2006. Application of phylogenetic networks in evolutionary studies. Mol. Biol. Evol. 23, 254–267.

Huson, D.H., Scornavacca, C., 2011. A survey of combinatorial methods for phylogenetic networks. Genome Biol. Evol. 3, 23–35.

Ilut, D.C., Coate, J.E., Luciano, A.K., Owens, T.G., May, G.D., Farmer, A., Doyle, J.J., 2012. A comparative transcriptomic study of an allotetraploid and its diploid progenitors illustrates the unique advantages and challenges of RNA-seq in plant species. Am. J. Bot. 99, 383–396.

Jennings, W.B., Edwards, S.V., 2005. Speciational history of Australian grass finches (Poephila) inferred from thirty gene trees. Evolution 59, 2033–2047.

Jovelin, R., 2009. Rapid sequence evolution of transcription factors controlling neuron differentiation in Caenorhabditis. Mol. Biol. Evol. 26, 2373–2386.

Keinan, A., Reich, D., 2010. Human population differentiation is strongly correlated with local recombination rate. PLoS Genet. 6, e1000886.

Kimura, M., 1983. The neutral theory of molecular evolution. Cambridge University Press, New York.

Kingman, J.F.C., 1982. On the genealogy of large populations. J. Appl. Probab. 19A, 27–43.

Knowles, L.L., 2009. Statistical phylogeography. Annu. Rev. Ecol. Evol. Syst. 40, 593–612.

Knowles, L.L., Carstens, B.C., 2007. Delimiting species without monophyletic gene trees. Syst. Biol. 56, 887–895.

Knowles, L.L., Maddison, W.P., 2002. Statistical phylogeography. Mol. Ecol. 11, 2623–2635.

Kofler, R., Pandey, R.V., Schlotterer, C., 2011. PoPoolation2: identifying differentiation between populations using sequencing of pooled DNA samples (Pool-Seq). Bioinformatics 27, 3435–3436.

Kryazhimskiy, S., Plotkin, J.B., 2008. The population genetics of dN/dS. PLoS Genet. 4, e1000304.

Kubatko, L.S., Degnan, J.H., 2007. Inconsistency of phylogenetic estimates from concatenated data under coalescence. Syst. Biol. 56, 17–24.

Kubatko, L.S., Carstens, B.C., Knowles, L.L., 2009. STEM: species tree estimation using maximum likelihood for gene trees under coalescence. Bioinformatics 25, 971–973.

Kuhner, M.K., 2009. Coalescent genealogy samplers: windows into population history. Trends Ecol. Evol. 24, 86–93.

Kulathinal, R.J., Stevison, L.S., Noor, M.A.F., 2009. The genomics of speciation in Drosophila: diversity, divergence, and introgression estimated using low-coverage genome sequencing. PLoS Genet. 5.

Langley, C.H., Stevens, K., Cardeno, C., Lee, Y.C.G., Schrider, D.R., Pool, J.E., Langley, S.A., Suarez, C., Corbett-Detig, R.B., Kolaczkowski, B., Fang, S., Nista, P.M., Holloway, A.K., Kern, A.D., Dewey, C.N., Song, Y.S., Hahn, M.W., Begun, D.J., 2012. Genomic variation in natural populations of *Drosophila melanogaster*. Genetics 192, 533–598.

Lanier, H.C., Knowles, L.L., 2012. Is recombination a problem for species-tree analyses? Syst. Biol. 61, 691–701.

Laporte, V., Charlesworth, B., 2002. Effective population size and population subdivision in demographically structured populations. Genetics 162, 501–519.

Lartillot, N., Poujol, R., 2011. A phylogenetic model for investigating correlated evolution of substitution rates and continuous phenotypic characters. Mol. Biol. Evol. 28, 729–744.

Lawson, D.J., Hellenthal, G., Myers, S., Falush, D., 2012. Inference of population structure using dense haplotype data. PLoS Genet. 8, e1002453.

Leffler, E.M., Bullaughey, K., Matute, D.R., Meyer, W.K., Segurel, L., Venkat, A., Andolfatto, P., Przeworski, M., 2012. Revisiting an old riddle: what determines genetic diversity levels within species? PLoS Biol. 10, e1001388.

Lemmon, A.R., Lemmon, E.M., 2012. High-throughput identification of informative nuclear loci for shallow-scale phylogenetics and phylogeography. Syst. Biol. 61, 745–761.

Lemmon, A.R., Emme, S.A., Lemmon, E.M., 2012. Anchored hybrid enrichment for massively high-throughput phylogenomics. Syst. Biol. 61, 727–744.

Liu, L., 2008. BEST: Bayesian estimation of species trees under the coalescent model. Bioinformatics 24, 2542–2543.

Author's personal copy

Liu, L., Pearl, D.K., 2007. Species trees from gene trees: reconstructing Bayesian posterior distributions of a species phylogeny using estimated gene tree distributions. Syst. Biol. 56, 504–514.

Liu, L., Yu, L.L., Pearl, D.K., Edwards, S.V., 2009. Estimating species phylogenies using coalescence times among sequences. Syst. Biol. 58, 468–477.

Liu, L.A., Yu, L.L., Edwards, S.V., 2010. A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. BMC Evol. Biol. 10, 302.

Lynch, M., 2007. The Origins of Genome Architecture. Sinauer, Sunderland, MA.

Lynch, M., 2009. Estimation of allele frequencies from high-coverage genome-sequencing projects. Genetics 182, 295–301.

Lynch, M., Conery, J.S., 2003. The origins of genome complexity. Science 302, 1401–1404.

Maddison, W.P., 1997. Gene trees in species trees. Syst. Biol. 46, 523–536.

Mamanova, L., Coffey, A.J., Scott, C.E., Kozarewa, I., Turner, E.H., Kumar, A., Howard, E., Shendure, J., Turner, D.J., 2010. Target-enrichment strategies for next-generation sequencing. Nat. Methods 7, 111–118.

Manel, S.P., Schwartz, M.K., Luikart, G., Taberlet, P., 2003. Landscape genetics: combining landscape ecology and population genetics. Trends Ecol. Evol. 18, 189–197.

Masly, J.P., Presgraves, D.C., 2007. High-resolution genome-wide dissection of the two rules of speciation in Drosophila. PLoS Biol. 5, e243.

McCormack, J.E., Faircloth, B.C., Crawford, N.G., Gowaty, P.A., Brumfield, R.T., Glenn, T.C., 2012. Ultraconserved elements are novel phylogenomic markers that resolve placental mammal phylogeny when combined with species-tree analysis. Genome Res. 22, 746–754.

McCormack, J.E., Hird, S.M., Zellmer, A.J., Carstens, B.C., Brumfield, R.T., 2013. Applications of next-generation sequencing to phylogeography and phylogenetics. Mol. Phylogenet. Evol. 66, 526–538.

McVicker, G., Gordon, D., Davis, C., Green, P., 2009. Widespread genomic signatures of natural selection in Hominid evolution. PLoS Genet. 5, e1000471.

Messer, P.W., Petrov, D.A., 2013. Frequent adaptation and the McDonald-Kreitman test. Proc. Natl. Acad. Sci. USA 110, 8615–8620.

Myers, S., Fefferman, C., Patterson, N., 2008. Can one learn history from the allelic spectrum? Theor. Popul. Biol. 73, 342–348.

Nachman, M.W., Payseur, B.A., 2012. Recombination rate variation and speciation: theoretical predictions and empirical results from rabbits and mice. Philos. Trans. R. Soc. B 367, 409–421.

Neher, R.A., 2013. Genetic draft, selective interference, and population genetics of rapid adaptation. arXiv:1302.1148 [q-bio.PE].

Niedringhaus, T.P., Milanova, D., Kerby, M.B., Snyder, M.P., Barron, A.E., 2011. Landscape of next-generation sequencing technologies. Anal. Chem. 83, 4327–4341.

Nielsen, R., 2005. Molecular signatures of natural selection. Annu. Rev. Genet. 39, 197–218.

Nielsen, R., Hubisz, M.J., Hellmann, I., Torgerson, D., Andres, A.M., Albrechtsen, A., Gutenkunst, R., Adams, M.D., Cargill, M., Boyko, A., Indap, A., Bustamante, C.D., Clark, A.G., 2009. Darwinian and demographic forces affecting human protein coding genes. Genome Res. 19, 838–849.

Nordborg, M., 2000. Linkage disequilibrium, gene trees and selfing: an ancestral recombination graph with partial self-fertilization. Genetics 154, 923–929.

Nordborg, M., Hu, T.T., Ishino, Y., Jhaveri, J., Toomajian, C., Zheng, H., Bakker, E., Calabrese, P., Gladstone, J., Goyal, R., Jakobsson, M., Kim, S., Morozov, Y., Padhukasahasram, B., Plagnol, V., Rosenberg, N.A., Shah, C., Wall, J.D., Wang, J., Zhao, K., Kalbfleisch, T., Schulz, V., Kreitman, M., Bergelson, J., 2005. The pattern of polymorphism in *Arabidopsis thaliana*. PLoS Biol. 3, e196.

Nosil, P., Funk, D.J., Ortiz-Barrientos, D., 2009. Divergent selection and heterogeneous genomic divergence. Mol. Ecol. 18, 375–402.

Obbard, D.J., Welch, J.J., Kim, K.W., Jiggins, F.M., 2009. Quantifying adaptive evolution in the Drosophila immune system. PLoS Genet. 5.

Olson-Manning, C.F., Wagner, M.R., Mitchell-Olds, T., 2012. Adaptive evolution: evaluating empirical support for theoretical predictions. Nat. Rev. Genet. 13, 867–877.

Orr, H.A., 1997. Haldane's rule. Annu. Rev. Ecol. Syst. 28, 195–218.

Palstra, F.P., Ruzzante, D.E., 2008. Genetic estimates of contemporary effective population size: what can they tell us about the importance of genetic stochasticity for wild population persistence? Mol. Ecol. 17, 3428–3447.

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

Pinho, C., Hey, J., 2010. Divergence with gene flow: models and data. Annu. Rev. Ecol. Evol. Syst. 41, 215–230.

Pool, J.E., Nielsen, R., 2007. Population size changes reshape genomic patterns of diversity. Evolution 61, 3001–3006.

Pool, J.E., Nielsen, R., 2008. The impact of founder events on chromosomal variability in multiply mating species. Mol. Biol. Evol. 25, 1728–1736.

Pool, J.E., Nielsen, R., 2009. Inference of historical changes in migration rate from the lengths of migrant tracts. Genetics 181, 711–719.

Posada, D., Crandall, K.A., 1998. MODELTEST: testing the model of DNA substitution. Bioinformatics 14, 817–818.

Prasad, A.B., Mullikin, J.C., Green, E.D., Progra, N.C.S., 2013. A scalable and flexible approach for investigating the genomic landscapes of phylogenetic incongruence. Mol. Phylogenet. Evol. 66, 1067–1074.

Presgraves, D.C., 2008. Sex chromosomes and speciation in Drosophila. Trends Genet. 24, 336–343.

Pulquerio, M.J.F., Nichols, R.A., 2007. Dates from the molecular clock: how wrong can we be? Trends Ecol. Evol. 22, 180–184.

Rocha, E.P.C., Smith, J.M., Hurst, L.D., Holden, M.T.G., Cooper, J.E., Smith, N.H., Feil, E.J., 2006. Comparisons of dN/dS are time dependent for closely related bacterial genomes. J. Theor. Biol. 239, 226–235.

Rosenberg, N.A., 2002. The probability of topological concordance of gene trees and species trees. Theor. Popul. Biol. 61, 225–247.

Salichos, L., Rokas, A., 2013. Inferring ancient divergences requires genes with strong phylogenetic signals. Nature 497, 327–331.

Schierup, M.H., Hein, J., 2000. Consequences of recombination on traditional phylogenetic analysis. Genetics 156, 879–891.

Slowinski, J.B., 2001. Molecular polytomies. Mol. Phylogenet. Evol. 19, 114–120.

Smith, T.B., Wayne, R.K., Girman, D.J., Bruford, M.W., 1997. A role for ecotones in generating rainforest biodiversity. Science 276, 1855–1857.

Sousa, V., Hey, J., 2013. Understanding the origin of species with genome-scale data: modelling gene flow. Nat. Rev. Genet. 14, 404–414.

Stadler, T., Haubold, B., Merino, C., Stephan, W., Pfaffelhuber, P., 2009. The impact of sampling schemes on the site frequency spectrum in nonequilibrium subdivided populations. Genetics 182, 205–216.

Storfer, A., Murphy, M.A., Spear, S.F., Holderegger, R., Waits, L.P., 2010. Landscape genetics: where are we now? Mol. Ecol. 19, 3496–3514.

Strasburg, J.L., Rieseberg, L.H., 2010. How robust are "Isolation with Migration" analyses to violations of the IM model? A simulation study. Mol. Biol. Evol. 27, 297–310.

Sung, W., Ackerman, M.S., Miller, S.F., Doak, T.G., Lynch, M., 2012. Drift-barrier hypothesis and mutation-rate evolution. Proc. Natl. Acad. Sci. USA 109, 18488–18492.

Swanson, W.J., Vacquier, V.D., 2002. The rapid evolution of reproductive proteins. Nat. Rev. Genet. 3, 137–144.

Tajima, F., 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123, 585–595.

Tavare, S., 1984. Line-of-descent and genealogical processes, and their applications in population-genetics models. Theor. Popul. Biol. 26, 119–164.

The 1000 Genomes Project Consortium, 2010. A map of human genome variation from population-scale sequencing. Nature vol. 467, pp. 1061–1073.

The 1000 Genomes Project Consortium, 2012. An integrated map of genetic variation from 1,092 human genomes. Nature vol. 491, pp. 56–65.

Ting, C.T., Tsaur, S.C., Wu, C.I., 2000. The phylogeny of closely related species as revealed by the genealogy of a speciation gene, Odysseus. Proc. Natl. Acad. Sci. USA 97, 5313–5316.

Voight, B.F., Kudaravalli, S., Wen, X.Q., Pritchard, J.K., 2006. A map of recent positive selection in the human genome. PLoS Biol. 4, 446–458.

Wakeley, J., 1999. Nonequilibrium migration in human history. Genetics 153, 1863–1871.

Wakeley, J., 2003. Inferences about the structure and history of populations: coalescents and intraspecific phylogeography. In: Singh, R.S., Uyenoyama, M.K. (Eds.), Evolution of Population Biology. Cambridge University Press, Cambridge.

Wakeley, J., 2009. Coalescent theory: an introduction. Roberts and Company Publishers, Greenwood Village, CO.

Wakeley, J., Aliacar, N., 2001. Gene genealogies in a metapopulation. Genetics 159, 893–905.

Wakeley, J., Hey, J., 1997. Estimating ancestral population parameters. Genetics 145, 847–855.

Wakeley, J., King, L., Low, B.S., Ramachandran, S., 2012. Gene genealogies within a fixed pedigree, and the robustness of Kingman's coalescent. Genetics 190, 1433–1445.

Whitney, K.D., Garland, T., 2010. Did genetic drift drive increases in genome complexity? PLoS Genet. 6, e1001080.

Whitney, K.D., Baack, E.J., Hamrick, J.L., Godt, M.J.W., Barringer, B.C., Bennett, M.D., Eckert, C.G., Goodwillie, C., Kalisz, S., Leitch, I.J., Ross-Ibarra, J., 2010. A role for nonadaptive processes in plant genome size evolution? Evolution 64, 2097–2109.

Wilson, D.J., Hernandez, R.D., Andolfatto, P., Przeworski, M., 2011. A population genetics-phylogenetics approach to inferring natural selection in coding sequences. PLoS Genet. 7, e1002395.

Wright, S., 1931. Evolution in Mendelian populations. Genetics 16, 97–159.

Yang, Z.H., 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput. Appl. Biosci. 13, 555–556.

Yang, Z.H., Rannala, B., 2010. Bayesian species delimitation using multilocus sequence data. Proc. Natl. Acad. Sci. USA 107, 9264–9269.