doi:10.1111/evo.12118

MORE ACCURATE PHYLOGENIES INFERRED FROM LOW-RECOMBINATION REGIONS IN THE PRESENCE OF INCOMPLETE LINEAGE SORTING

James B. Pease[1,2] and Matthew W. Hahn[1,3]

[1] *Department of Biology, Indiana University, Bloomington, Indiana 47405*

[2] *E-Mail: jbpease@indiana.edu*

[3] *School of Informatics and Computing, Indiana University, Bloomington, Indiana 47405*

Received December 5, 2012
Accepted March 20, 2013

When speciation events occur in rapid succession, incomplete lineage sorting (ILS) can cause disagreement among individual gene trees. The probability that ILS affects a given locus is directly related to its effective population size (N_e), which in turn is proportional to the recombination rate if there is strong selection across the genome. Based on these expectations, we hypothesized that low-recombination regions of the genome, as well as sex chromosomes and nonrecombining chromosomes, should exhibit lower levels of ILS. We tested this hypothesis in phylogenomic datasets from primates, the *Drosophila melanogaster* clade, and the *Drosophila simulans* clade. In all three cases, regions of the genome with low or no recombination showed significantly stronger support for the putative species tree, although results from the X chromosome differed among clades. Our results suggest that recurrent selection is acting in these low-recombination regions, such that current levels of diversity also reflect past decreases in the effective population size at these same loci. The results also demonstrate how considering the genomic context of a gene tree can assist in more accurate determination of the true species phylogeny, especially in cases where a whole-genome phylogeny appears to be an unresolvable polytomy.

KEY WORDS: ***Drosophila*, mitochondria, phylogenomics, primates, sex chromosomes.**

In molecular phylogenetics, the ultimate goal is the reconstruction of the evolutionary history of a group of species from molecular sequence data (Felsenstein 2004; Edwards 2009). Such species trees are often inferred from one or more gene trees, each of which describes the evolutionary relationships among homologous loci in the sampled species. However, individual gene trees may differ from the true species tree for a number of reasons, including long-branch attraction, hybridization, horizontal gene transfer, or duplication and loss of undetected paralogs (reviewed in Maddison 1997; Degnan and Rosenberg 2009). In addition, when three or more species have a relatively short time between speciation events, incomplete lineage sorting (ILS) is a common cause of discordant gene trees (Degnan and Salter 2005; Degnan and Rosenberg 2009; Hobolth et al. 2011).

In the most general sense, ILS is the failure of two lineages to coalesce within a population, instead having their most recent common ancestor (MRCA) in an ancestral population (Hudson 1983b; Tajima 1983). In order for ILS to affect gene trees among three or more taxa, multiple lineages must be maintained between speciation events without coalescing, instead coalescing in a pre-speciation ancestral population (Fig. 1A, dashed line). Because multiple lineages are maintained in the internode population, the possibility exists that lineages will coalesce first with an outgroup lineage instead of with a lineage leading to its sister taxon. Because ILS depends on the maintenance of polymorphisms between speciation events, it is inversely proportional to $\tau/2N_e$: the relative time between speciation events (τ; i.e., the length of the internode branch) and the effective population size (N_e) of the

EVOLUTION
INTERNATIONAL JOURNAL OF ORGANIC EVOLUTION

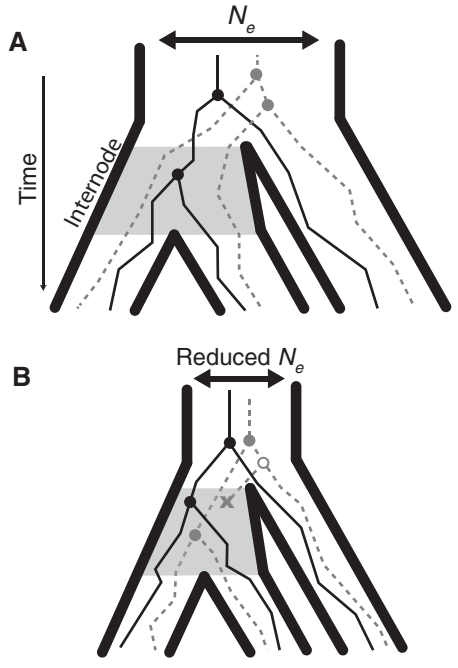

Figure 1. Species phylogeny (solid outline) in relation to gene trees (thin internal lines). Concordant gene trees have the same topology as the species tree (solid black line). When the effective population size (N_e) is large (A) and the internode time (shaded area) is short, lineages may not coalesce in the internode ancestral population and could lead to a discordant gene tree (dotted gray line). When N_e is reduced (B) lower levels of polymorphism are maintained in the internode population, and lineages are more likely to go extinct (x), decreasing the likelihood of sampling discordant gene trees.

internode population. This implies that larger population sizes or shorter relative times between speciation events both increase the probability of having ILS at a locus (Pamilo and Nei 1988). Because τ is a function of the species phylogeny, it is constant across all regions of the genome. The probability of ILS occurring in a given genomic region, however, is directly proportional to N_e in that region, as N_e determines the time to the MRCA for lineages in a population (Fig. 1A vs. B).

There are multiple reasons why N_e may vary from locus to locus. Under neutrality, the time to the MRCA varies stochastically among loci, depending solely on the coalescent process (Hudson 1990). In the presence of strong selection—either positive selection fixing new mutations or negative selection removing new mutations—N_e is consistently lowered at linked sites (Maynard Smith and Haigh 1974; Charlesworth et al. 1993). Although strong natural selection can occur at any position in the genome, the effects of this selection are magnified in regions with low recombination (i.e., low rates of crossing-over), as a much larger region is affected by any selected variant (Kaplan et al. 1989; Charlesworth 2012). Therefore, among a set of randomly chosen loci, the action of strong selection should cause N_e to be negatively correlated with the recombination rate. Indeed, studies in multiple organisms, including *Drosophila melanogaster* (Begun and Aquadro 1992; Haddrill et al. 2007; Campos et al. 2012; Langley et al. 2012) and *Homo sapiens* (Nachman et al. 1998; Stajich and Hahn 2005; Cai et al. 2009; McVicker et al. 2009) have shown that low-recombination regions (LRRs) experience a reduction in levels of nucleotide diversity relative to divergence, an indication that N_e is reduced due to linked selection. In the absence of strong selection, N_e is not expected to vary with recombination in a consistent manner (Hudson 1983a).

Taken together, the expected relationships between N_e and both ILS and recombination suggest that regions of lower recombination will show lower levels of ILS (expressed as a lower proportion of incorrect gene trees). This hypothesis is supported by previous results comparing the degree of ILS within regions expected to experience strong selection (coding sequences) to regions not experiencing strong selection (noncoding sequences; Hobolth et al. 2011; Scally et al. 2012), and among regions differing in the rate of recombination (Hobolth et al. 2011; Prüfer et al. 2012). In addition to LRRs, we also expect to sample less ILS (i.e., fewer gene trees supporting incorrect species relationships) on sex chromosomes due to lower N_e. This implies that regions on sex chromosomes should have a lower level of ILS than autosomes, and furthermore that LRR regions on the X chromosome—as well as the entire Y chromosome—should have the lowest ILS among all nuclear loci. Finally, the mitochondrial genome has been shown to be generally nonrecombining and to have relatively low N_e in both humans and *Drosophila* (Meiklejohn et al. 2007; Galtier et al. 2009; Piganeau and Eyre-Walker 2009; Rand 2011). Even though the entire mitochondrial genome represents only a single locus, and therefore a single realization of the coalescent process, it should also be enriched for support of the true species phylogeny. However, mitochondrial introgression is common in many lineages and may affect its agreement with the species phylogeny.

Based on the above considerations, we hypothesized that LRRs should on average exhibit lower levels of ILS compared to other loci. We tested this hypothesis in three genome-wide phylogenetic datasets with varying degrees of ILS. First, we examined the human–chimpanzee–gorilla clade (HCG), which shows a strong majority (70%) of nucleotides in protein-coding regions supporting humans and chimpanzees as sister taxa with gorilla as the outgroup—denoted (HC)G*—and equal support for the two alternative trees (Scally et al. 2012). (Note: "*" denotes our "putative species tree" for each clade.) Second, we considered the subgroup composed of *D. melanogaster*, *Drosophila erecta*, and *Drosophila yakuba* (the "Dmel clade"), which shows a

plurality—but not a majority—of protein-coding nucleotides (44.7%) supporting *D. erecta* and *D. yakuba* as sister taxa, denoted (EY)M* (Pollard et al. 2006). Again, an approximately equal number of sites support the two alternative topologies (Pollard et al. 2006). Finally, we considered the subgroup composed of *Drosophila simulans*, *Drosophila sechellia*, and *Drosophila mauritiana* (the "Dsim clade"), which is characterized by frequent hybridization events and an extremely short interval between the speciation events (Nunes et al. 2010; Legrand et al. 2011). *Drosophila sechellia* and *D. mauritiana* are island species in the Seychelles and Mauritius archipelagos, respectively, and are inferred to have split from an ancestral population on Madagascar, where *D. simulans* is still prevalent. A recent analysis determined the species relationship as (*simulans*, *sechellia*)*mauritiana* [(SC)M*] from an alignment of all autosomes (Garrigan et al. 2012). However, after accounting for recurrent hybridizations among the species, this article proposed that the Dsim clade represented a true (or "hard") polytomy, with each possible topology represented approximately equally.

For each of the three clades, our analysis found increased support for a single phylogeny with decreasing recombination rate. We also found support for reduced ILS on the X chromosome in HCG, but in the Dmel and Dsim clades, the X chromosome appears to have a more complex history, possibly due to its role in species isolation. The mitochondrial genomes in all three datasets also supported the nuclear LRR-predicted species phylogenies. Therefore, in clades where ILS-discordant gene trees are a possibility, an analysis that considers the recombination environment at each locus can provide added context for gene tree discordance that may indicate the true species phylogeny.

Materials and Methods

HCG GENE TREES AND RECOMBINATION RATES

For the HCG clade, we used a dataset reporting the proportion of sites that support alternative topologies derived from a CoalHMM analysis of $n = 1{,}998$ regions of a five-way genome alignment, including orangutan and macaque as outgroups (Scally et al. 2012). Any regions overlapping another region (these all occurred on chromosome 1) were excluded, leaving $n = 1{,}961$ regions. Each region contained 10^6 sites, but these sites could be spread over highly variable lengths of the genome. Therefore, only the $n = 1885$ regions where (1) the coordinates of the sampled genomic section were <2 Mb apart, and (2) the region did not span the centromere, were retained for the final dataset. The overall recombination rate for each region was calculated as the average of all point estimates of recombination within the region (International HapMap Consortium et al. 2007). Each region contained between 598 and 3482 (mean = 1622) point estimates of recombination.

DMEL GENE TREES AND RECOMBINATION RATES

For the Dmel clade, we used a dataset composed of rooted ML gene trees from alignments of $n = 8838$ protein-coding regions in *D. melanogaster*, *D. yakuba*, *D. erecta*, and outgroup *Drosophila ananassae* (Pollard et al. 2006). We used recombination rates from genome-wide estimates for each protein-coding gene in *D. melanogaster* (Langley et al. 2012). Only genes with available recombination rates ($n = 6811$) were retained in the dataset. We further restricted our analysis to $n = 5949$ genes that appear on the same chromosome in both *D. melanogaster* and *D. yakuba* (McQuilton et al. 2012). We also required that no gene overlap another gene by more than 10% of its total length, leaving $n = 5532$ genes. For chromosome 4 (the nonrecombining "dot" chromosome), data from the 37 ML gene trees in the original dataset was used. All 37 loci were consistently on chromosome 4 in both *melanogaster* and *yakuba*.

DSIM GENE TREES AND RECOMBINATION RATES

For the Dsim clade, we used a dataset composed of polarized SNP counts for $n = 23{,}773$ nonoverlapping 5 kb aligned genomic regions of *D. simulans*, *D. sechellia*, *D. mauritiana*, and outgroup *D. melanogaster* (Garrigan et al. 2012). To avoid errors resulting from small numbers of SNPs, only regions with ≥20 SNPs ($n = 21{,}636$) were included, following Garrigan et al. (2012). Phylogenies for these 5 kb regions were determined using the 4sp software package (http://kimura.biology.rochester.edu/software/4sp/4sp.tar.gz, Garrigan et al. 2012).

Recombination rates were estimated from the *D. melanogaster* recombination rates for homologous genes (Langley et al. 2012). We calculated the recombination rate for each 5 kb region as $r = \frac{\sum_i p_i r_i}{\sum_i p_i}$. In this equation, p_i is the proportion of sites in a given 5 kb region overlapped by a given gene, and r_i is the recombination rate for that gene; the product of these is summed over all genes overlapping the given region and then divided by the total proportion of genic sites in the region. For 5 kb regions without any genes in them, if there was a single gene located within 1 kb we estimated the recombination rate as equal to that gene. Recombination rates for $n = 9{,}848$ windows were determined, and all other regions were excluded.

Using the methods outlined in Garrigan et al. (2012), we calculated the global and local log-likelihood of introgression between species using 4sp. A standard likelihood-ratio test was applied to each 5 kb window and any window with likelihood-ratio test P-value < 0.029 was considered putatively introgressed. After exclusion of introgressed regions, $n = 9130$ regions remained in the final dataset.

GENE DENSITY

Because genic regions have reduced ILS, we need to control for the gene density of each region when considering the effect

of recombination rate. For the HCG clade, we determined the number of genic positions in each genomic region by counting the number of nucleotides in each 1 Mb aligned region that fell within a coding region in the human hg19 genome assembly (Fujita et al. 2011). The proportion of genic sites in each region ranged from 0% ($n = 47$) to 100% ($n = 1$), with a mean of 47%. In the Dmel clade, only coding regions were used, so no correction was needed. For the Dsim clade, we determined the number of genic positions in each genomic region by comparison with all protein-coding gene coordinates in the *D. melanogaster* genome (v5.31) via FlyBase (McQuilton et al. 2012).

MITOCHONDRIAL GENE TREES

Mitochondrial genomes for human (NC_012920.1), chimpanzee (NC_001643.1), gorilla (NC_001645.1), and outgroup orangutan (NC_001646.1) were aligned using CLUSTALW 2.1 (Larkin et al. 2007). Mitochondrial genomes of *D. melanogaster* (NC_001709.1), *D. erecta* (BK006335.1), *D. yakuba* (NC_001322.1), and outgroup *D. ananassae* (BK006336.1) were aligned by the same method. Mitochondrial genomes of *D. simulans* (NC_005781.1), *D. mauritiana* (NC_005779.1), *D. sechellia* (NC_005780.1), and outgroup *D. melanogaster* were separately aligned by the same method. We inferred an ML whole-mitochondria phylogeny from these nucleotide alignments with RaXML using the GTRGAMMA substitution model (Stamatakis 2006).

STATISTICAL ANALYSES

Autocorrelation, linear regression, logistic regression, Spearman's rank correlation, and Student's *t*-test were performed using the *acf*, *lm*, *glm*, *cor*, and *t.test* functions of R (http://www.R-project.org). All other analyses were performed with custom python scripts through MySQL databases.

Results

LOW-RECOMBINATION REGIONS

HCG autosomal aligned regions ($n = 1838$) showed an average of 69.5% of sites with support for the (HC)G* phylogeny, similar to the proportion of sites reported for protein-coding regions (Fig. 1C; Scally et al. 2012). However, those autosomal regions in the lowest quintile of recombination rates (LRR_{20}) showed significantly increased support (72.8%) for (HC)G* (Fig. 2A; $P = 4.9 \times 10^{-14}$, *t*-test). Mean support for (HC)G* in the LRR_{20} regions is also significantly higher than the next quintile (LRR_{40}, $P = 0.016$, *t*-test). The mean recombination rate in LRR_{20} regions is 0.62 cM/Mb (±0.009 SE), whereas the recombination rate in LRR_{40} is 1.06 (±0.005 SE), reflecting a real difference in recombination rates for LRR^{20}. As can also be seen in Figure 2A, there is an almost linear decrease in support for the (HC)G* topology with increasing recombination rate ($\rho = -0.24$, $P < 2.2 \times 10^{-16}$, Spearman's rank correlation), with each of the two alternative trees showing equal increases in frequency (data not shown). We also found strong first-order autocorrelation coefficients for each chromosome: in effect, neighboring windows showed highly similar results. A conservative reanalysis of every fourth window still showed both a significant relationship between recombination rate and ILS ($\rho = -0.23$, $P = 2.3 \times 10^{-7}$) and a significant difference in support for LRR_{20} regions ($P = 2.2 \times 10^{-4}$, *t*-test).

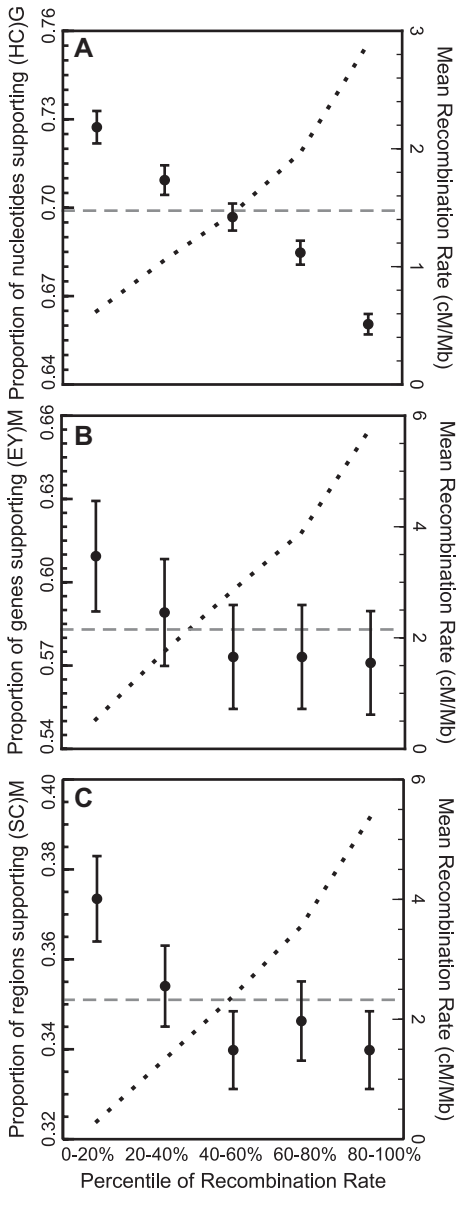

Figure 2. Regions with low-recombination rates show increased support for the putative species tree in each clade compared to the mean support (dashed line). As the recombination rate increases (dotted line), the support declines and incomplete lineage sorting (ILS) increases. This pattern is consistent in the human–chimpanzee–gorilla clade (HCG) clade (A), Dmel clade (B), and Dsim clade (C).

Similarly, in the Dmel clade, the (EY)M* topology is supported by 58.3% of all autosomal genes ($n = 4685$), increasing to 60.9% of autosomal genes in LRR^{20} supporting this tree (Fig. 2B; $\chi^2 = 3.19$, df $= 1$, $P = 0.07$). Lower recombination rates generally correlate with increased support for (EY)M*, although nonsignificantly ($\beta = -0.016$, $P = 0.29$, logistic regression). Neither of the alternative trees shows this decrease in support with increasing recombination ($\beta = 0.017$ and 0.005, respectively). Of autosomal genes with an estimated recombination rate of zero, 64.7% (101/156) support (EY)M*. Both the HCG and Dmel datasets match our prediction that LRRs will exhibit lower levels of ILS and therefore more strongly support the putative species tree.

Previous analysis in the Dsim clade was unable to identify a bifurcating species tree, with all three possible trees having essentially equal support after controlling for introgression (Garrigan et al. 2012). Among non-introgressed, autosomal 5 kb windows sampled ($n = 7725$), the support for the topology (SC)M* is 35.1%. However, among LRR_{20} regions the support increases significantly to 37.2% (Fig. 2C; $\chi^2 = 4.27$, df $= 1$, $P = 0.039$), and in LRR_{10} to 40.0% ($\chi^2 = 4.38$, df $= 1$, $P = 0.036$). Both alternative topologies had decreased support in LRR_{20} regions. More generally, support for (SC)M* inversely correlates with recombination ($\beta = -0.023$, $P = 0.07$, logistic regression). As in Dmel, support for neither alternative species tree shows a negative relationship with recombination ($\beta_{S(CM)} = 0.016$, $\beta_{(SM)C} = 0.007$). We found no significant autocorrelation among windows in the Dsim dataset, and therefore did not correct for any non-independence among windows. Assuming that much of the phylogenetic discordance in the Dsim clade is due to ILS, the regions of lowest recombination indicate that (SC)M* is the true species relationship, where *D. mauritiana* diverged first from a *D. simulans–sechellia* ancestral population. This result also agrees with the initial autosomal ML tree in Garrigan et al. (2012).

Because ILS is lower in genic regions (Hobolth et al. 2007; Prüfer et al. 2012; Scally et al. 2012), a confounding variable may be the genic content of each genomic region in the HCG and Dsim clades because the data were calculated in equal-length windows including both coding and noncoding sequences. However, the proportion of genic sites in a genomic region (p) does not correlate significantly with the recombination rate in either the HCG ($R^2 = 0.054$, linear regression) or Dsim clades ($R^2 = 0.0014$). Even when only values of p between 0.1 and 0.9 are considered, we find no significant correlation (HCG: $R^2 = 0.084$; Dsim: $R^2 = 0.0010$). Therefore, we find that the proportion of genic positions in each window (i.e., gene density) does not appear to be a confounding variable in our results concerning the relationship between recombination rate and ILS.

Choosing the size of genomic regions used to study ILS is a balance between having sufficient sequence data to produce confident gene trees, and making the regions small enough to capture heterogeneity within the genome. We considered the possibility that the 1 Mb regions in the HCG dataset are too large to capture the finer-scale dynamics of recombination and ILS. To test the potential effects of estimating ILS in larger aligned regions, we used the recombination and phylogenetic data from the Dsim 5 kb regions to calculate average recombination and average support for (SC)M* in 1 Mb regions across the complete genome. Among autosomal 1 Mb regions ($n = 83$), those regions with lowest recombination tended to show stronger support for (SC)M* ($\rho = -0.244$, $P = 0.029$, Spearman's rank correlation). This suggests that the relationship between recombination and ILS appears to be observable and consistent even when relatively large genomic regions are sampled. Although supportive of results using 1 Mb windows, we should also note that the HCG dataset measures the proportion of sites that support the putative species tree within each window, rather than estimating a single tree. Although using the proportion of sites within a window does not adequately control for associations among sites, it may be preferred to simply assigning each window to a single topology. Indeed, because all windows show a majority of sites supporting (HC)G*, the proportion of sites more informatively describes the uncertainty present in the dataset.

X CHROMOSOMES

In the HCG clade, 84.3% of the sites on the X chromosome predict the (HC)G* tree, compared to 69.6% of autosomal regions (Fig. 3A; $P = 1.2 \times 10^{-26}$, t-test; see also Scally et al. 2012). Furthermore, 86.2% of LRR_{20} regions on the X chromosome provide support for (HC)G* ($P = 0.022$, t-test). Support is therefore maximized in this clade when using LRRs on the X chromosome, which should (all other things being equal) have the lowest N_e in the nuclear genome, aside from the Y chromosome (Y-linked sequences are not available for many of the species considered here).

In the Dmel clade, only 49.0% of genes on the X chromosome ($n = 846$) predict (EY)M*, whereas this tree is supported by 58.3% of autosomal genes ($n = 4686$; Fig. 3B; $\chi^2 = 124.8$, df $= 1$, $P < 2.2 \times 10^{-16}$). Among LRR_{20} genes on the X chromosome, 50.8% support (EY)M* ($\chi^2 = 0.24$, df $= 1$, $P = 0.6$). Similarly, in the Dsim clade, only 31.3% of X-linked regions ($n = 1402$) provide support for (SC)M* versus 42.4% for (SM)C (Fig. 3C). Autosomal regions ($n = 7728$) support (SC)M* at 35.1% and (SM)C at 36.2%. However, among LRR^{10} regions on the X chromosome, support increases to 44.2% for (SC)M* and drops to 29.2% for (SM)C. Although the X chromosome in Dsim overall indicates (SM)C, LRRs among X-linked loci shift toward the autosomal LRR prediction of (SC)M* ($\chi^2 = 11.7$, df $= 1$, $P = 6.2 \times 10^{-4}$).

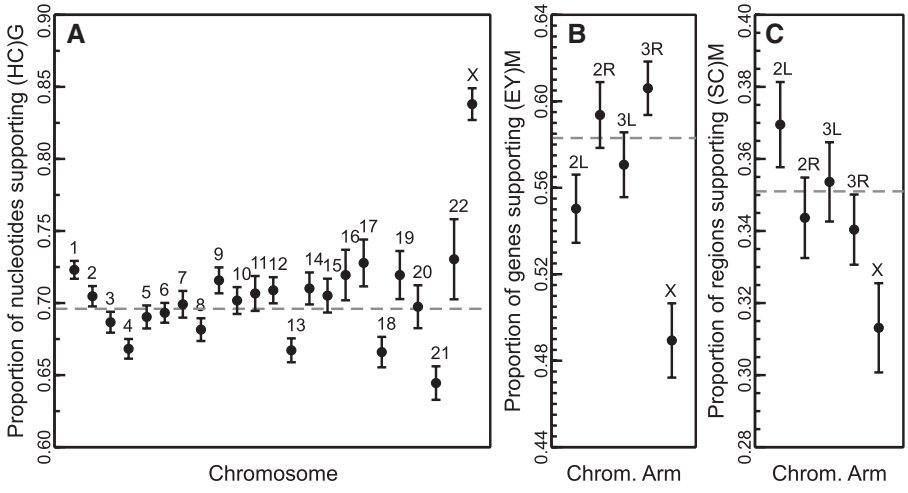

Figure 3. X-linked regions in the human–chimpanzee–gorilla clade (HCG) alignment (A) show markedly increased support for the putative species tree over the autosomal average (dashed line). However, X-linked genes in the Dmel clade (B) and X-linked regions in the Dsim clade (C) show decreased support relative to autosomes.

MITOCHONDRIA AND THE DOT CHROMOSOME

Maximum-likelihood trees for the HCG and Dmel mitochondrial alignments both returned the putative species tree. *Drosophila* chromosome 4 (the "dot" chromosome) does not show any evidence of crossing-over but does have gene conversion (Comeron et al. 2012). Of the 37 trees constructed for genes on chromosome 4, 25 support (EY)M* (67%), 7 support (EM)Y (19%), and 5 support (MY)E (14%). Thus, for the HCG and Dmel mitochondria and the *Drosophila* dot chromosome—as in the recombining portion of the nuclear genome examined here—there is strong support for a single tree ((HC)G* and (EY)M*, respectively).

For the Dsim clade, the mitochondrial genome strongly supports the (SC)M* tree. Therefore, even in the clade with the most ambiguous results from the nuclear genome, the mitochondrial genome shows clear support for a single species tree. Our results here differ from the mitochondrial analysis in Garrigan et al. (2012) because the *D. mauritiana* mitochondria used in that study was a haplotype introgressed recently from *D. simulans*. Our trees are derived from the reference genomes.

Discussion

One of the most studied sets of species relationships is the one between humans, chimpanzees, and gorillas (HCG). The first sequence-based HCG species trees were inferred from single genes (e.g., Ferris et al. 1981; Brown et al. 1982; Hixson and Brown 1986; Koop et al. 1986; Saitou and Nei 1986). With single-gene species tree inference, the assumption is that the sampled gene's evolution history matches speciation events, yielding a gene tree is that is concordant (topologically identical) with the species tree. These early studies were often equivocal about the relationships within the HCG clade, but all agreed that this uncertainty was likely due to a relatively short interval between speciation events. With the exponential rise in sequencing, HCG phylogenetic datasets expanded to 5 genes (Koop et al. 1989), 25 genes (Takahata et al. 1995), 45 genes (Satta et al. 2000), 53 homologous regions of ~25 kb each (Chen and Li 2001; Chen et al. 2001), 129 homologous regions (Osada and Wu 2005), and an 18.3 Mb concatenated alignment (Patterson et al. 2006). These datasets continued to confirm that the distance between speciation events was relatively short in the HCG clade, but generally favored the divergence of gorilla from a human–chimpanzee ancestral population [(HC)G*]. Most recently, with the completion of the gorilla genome, 70.1% of nucleotides in protein-coding regions were found to support (HC)G*, with 15% support each for (HG)C and (HC)G (Scally et al. 2012).

As with the HCG clade, many phylogenetic datasets have grown in size over the past decade, often leading to a concomitant rise in the evidence for discordance among individual gene trees. Although much of this discordance can be due to artifacts arising from tree inference methods or simply to a lack of resolution, incomplete lineage sorting in internode populations can lead to gene trees that contradict the true species tree for biological reasons. Therefore, even in cases where whole-genome data are obtained, multiple topologies may be supported by hundreds of different loci each (e.g., White et al. 2009). In these cases, no further gene trees can be collected, and we must therefore turn to other pieces of evidence to help in inferring the species tree from the individual gene trees.

In this study, we tested the hypothesis that specific genomic regions have reduced ILS, and as a result are more likely to give a topology matching the species tree. Specifically, we asked

whether LRRs, the X chromosome, and/or the mitochondrial genome showed evidence for reduced ILS due to reductions in N_e. To address these questions we used three clades: one with a strong consensus species tree (HCG), one with a moderate consensus tree (Dmel), and one previously declared a hard polytomy (Dsim). In the HCG and Dmel datasets, the gene trees inferred from LRRs showed increased support for the putative consensus tree. In the Dsim clade, the LRRs showed relatively stronger support for a single species tree, one that the mitochondrial genome also supports.

The use of LRRs to predict the species phylogeny in part assumes that recombination rates are conserved across all species in the clade. Without some conservation of recombination rates across the clade, genes in the ancestral populations experiencing ILS would experience recombination rates that may not be predicted by current recombination rates; we would therefore expect no association between recombination measured in extant taxa and reduced N_e in ancestral populations. Although still a relatively open question, it appears that recombination rates among closely related species are generally conserved across large genomic loci (Dumont and Payseur 2008, 2011; Smukowski and Noor 2011). Between humans and chimpanzee, a very high correlation in inferred recombination rates is seen between orthologous 1 Mb intervals, even though finer-scale variation in the recombination rate exists within these intervals (Auton et al. 2012). Because there is undoubtedly some variation in recombination rates among lineages—especially in the Dsim clade (Cattani and Presgraves 2012)—the relationships we see between recombination and reduced ILS may reflect decreased power due to noise in this predictor variable. It is also highly likely that the method used here would not work for ILS that has occurred much further back in time, such as in the base of the metazoans (Rokas et al. 2005), as we cannot accurately estimate the recombination rates in these ancient genomes.

We also expected that the X chromosome would have reduced ILS, because it has a lower N_e compared to autosomes. Although data from the primate X chromosome supports this hypothesis, the *Drosophila* X appears to have a more complex history. One factor may be the outsized role of the X chromosome in reproductive isolation—the so-called large X-effect (Coyne and Orr 2004). It has been demonstrated that hybrid incompatibilities accumulate on the X chromosome faster than on the autosomes in comparisons between *D. mauritiana* and *D. sechellia* (Masly and Presgraves 2007; McNabney 2012), between *D. mauritiana* and *D. simulans* (Davis and Wu 1996; Nunes et al. 2010), and between *D. melanogaster* and *D. simulans* (Presgraves 2003; Cattani and Presgraves 2012). In addition, fewer loci on the X chromosome show evidence for introgression among the species in the Dsim clade (Garrigan et al. 2012), further supporting its role in reproductive isolation. However, why this involvement in hybrid incompatibility manifests itself as a lower proportion of trees reflecting the species phylogeny is unclear. In fact, it has previously been proposed that loci involved in incompatibility should more faithfully reflect the species tree (Ting et al. 2000), although such genes would have to be involved only in the more recent speciation event to do so. One possibility is that the most common gene tree in both *Drosophila* clades is the outcome of massive hybridization after the initial speciation events, with only the X chromosome retaining the original, bifurcating species tree. Unfortunately, we do not currently have the ability to test this hypothesis.

Our results are consistent with the hypothesis that the levels of reduced heterozygosity seen in regions of low recombination really do indicate lower N_e, and not just decreased mutation rates (e.g., Hellmann et al. 2003). Many different pieces of evidence support the role of strong selection in reducing N_e, especially in LRRs, including a decreased efficacy of natural selection on weakly selected variants (Betancourt et al. 2002; Hey and Kliman 2002; Betancourt et al. 2009; Gossmann et al. 2011). To our knowledge, however, this is some of the first evidence from topological concordance that is consistent with reduced N_e in LRRs. Our results imply not only that strong selection is reducing N_e in these regions currently, but also that recurrent selection has been reducing N_e since at least the time when the internode populations existed. In fact, our expectation about the relationship between N_e and ILS was predicated on the assumption that N_e was also reduced in LRRs in the ancestral internode populations. This recurrent selection may be either negative selection against deleterious mutation or positive selection on advantageous mutations (or a mixture of the two), and may differ qualitatively between the clades considered here (cf. Hahn 2008). Previous studies examining the Dmel and HCG datasets found mixed effects of measures of selection on protein-coding genes in predicting ILS. Pollard et al. (2006) found no effect of d_N/d_S on the probability of discordance, whereas Scally et al. (2012) found that genes with lower d_N/d_S showed less discordance. This latter relationship suggests that variation in d_N/d_S when it is less than 1 is more indicative of the strength of negative selection than the incidence of positive selection.

We can also use these data to make rough calculations about the variation in N_e among genomic regions: if the probability of sampling a gene tree that matches the species tree is equivalent to $1 - (2/3)e^{-\tau/2N_e}$ (Hudson 1983b), then variation in the observed proportion of concordant gene trees among regions with different recombination rates should be due in part to variation in $2N_e$. For the HCG dataset, setting $\tau = 1$ for simplicity, our results imply an approximately 20% reduction in N_e going from LRR_{100} to LRR_{20}. This value is strikingly similar to the estimated average reduction in nucleotide diversity caused by linked selection in the human–chimpanzee ancestral population (19–26% on the

autosomes; McVicker et al. 2009) and may therefore accurately reflect variation in N_e across the genome.

The finding that LRRs more often indicate the true species tree could be used to boost the inference of species relationships from whole genome data in two ways. Quantitatively, gene trees could be weighted by recombination rates to estimate the species tree more accurately. Qualitatively, if regions of low recombination (LRRs, sex chromosomes) and no recombination (mitochondria, microchromosomes) all support a particular phylogeny more strongly than the genomic average, this would provide additional support for that species phylogeny. These pieces of data could provide crucial contextual information for resolving species trees when ILS is prevalent in a clade. The future of phylogenetics inevitably lies in the inference of phylogenies from genome-wide sequence data. Therefore, an important challenge is to develop phylogenetic models that can handle not only the heterogeneous process of molecular evolution across the genome, but that can also use these differences to enhance inferences of the species phylogeny. This implies that some of the most crucial developments in phylogenetic models will not necessarily be improvements in sequence alignment or gene tree inference, but rather in models that contextually interpret genomic heterogeneity in sequence evolution.

ACKNOWLEDGMENTS

The authors thank D. Pollard and G. Yuh Chwen Lee for providing data for the Dmel clade, and D. Garrigan for providing Dsim data. L. Kubatko and two anonymous reviewers also made very helpful suggestions that improved the article. This research was supported by the Indiana University Genetics, Molecular and Cellular Sciences Training Grant (T32-GM007757) and a fellowship from the Alfred P. Sloan Foundation to MWH.

LITERATURE CITED

Auton, A., A. Fledel-Alon, S. Pfeifer, O. Venn, L. Ségurel, T. Street, E. M. Leffler, R. Bowden, I. Aneas, J. Broxholme, et al. 2012. A fine-scale chimpanzee genetic map from population sequencing. Science 336:193–198.

Begun, D. J., and C. F. Aquadro. 1992. Levels of naturally occurring DNA polymorphism correlate with recombination rates in *D. melanogaster*. Nature 356:519–520.

Betancourt, A. J., D. C. Presgraves, and W. J. Swanson. 2002. A test for faster X evolution in *Drosophila*. Mol. Biol. Evol. 19:1816–1819.

Betancourt, A. J., J. J. Welch, and B. Charlesworth. 2009. Reduced effectiveness of selection caused by a lack of recombination. Curr. Biol. 19:655–660.

Brown, W. M., E. M. Prager, A. Wang, and A. C. Wilson. 1982. Mitochondrial DNA sequences of primates: tempo and mode of evolution. J. Mol. Evol. 18:225–239.

Cai, J. J., J. M. Macpherson, G. Sella, and D. A. Petrov. 2009. Pervasive hitchhiking at coding and regulatory sites in humans. PLoS Genet. 5:e1000336.

Campos, J. L., B. Charlesworth, and P. R. Haddrill. 2012. Molecular evolution in nonrecombining regions of the *Drosophila melanogaster* genome. Genome Biol. Evol. 4:278–288.

Cattani, M. V., and D. C. Presgraves. 2012. Incompatibility between X chromosome factor and pericentric heterochromatic region causes lethality in hybrids between *Drosophila melanogaster* and its sibling species. Genetics 191:549–559.

Charlesworth, B. 2012. The effects of deleterious mutations on evolution at linked sites. Genetics 190:5–22.

Charlesworth, D., M. T. Morgan, and B. Charlesworth. 1993. Mutation accumulation in finite outbreeding and inbreeding populations. Genet. Res. 61:39–56.

Chen, F. C., and W. H. Li. 2001. Genomic divergences between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees. Am. J. Hum. Genet. 68:444–456.

Chen, F. C., E. J. Vallender, H. Wang, C. S. Tzeng, and W. H. Li. 2001. Genomic divergence between human and chimpanzee estimated from large-scale alignments of genomic sequences. J. Hered. 92:481–489.

Comeron, J. M., R. Ratnappan, and S. Bailin. 2012. The many landscapes of recombination in *Drosophila melanogaster*. PLoS Genet. 8:e1002905.

Coyne, J. A., and H. A. Orr. 2004. Speciation. Sinauer Associates, Sunderland, MA.

Davis, A. W., and C. I. Wu. 1996. The broom of the sorcerer's apprentice: the fine structure of a chromosomal region causing reproductive isolation between two sibling species of *Drosophila*. Genetics 143:1287–1298.

Degnan, J. H., and N. A. Rosenberg. 2009. Gene tree discordance, phylogenetic inference and the multispecies coalescent. Trends Ecol. Evol. 24:332–340.

Degnan, J. H., and L. A. Salter. 2005. Gene tree distributions under the coalescent process. Evolution 59:24–37.

Dumont, B. L., and B. A. Payseur. 2008. Evolution of the genomic rate of recombination in mammals. Evolution 62:276–294.

———. 2011. Genetic analysis of genome-scale recombination rate evolution in house mice. PLoS Genet. 7:e1002116.

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

Felsenstein, J. 2004. Inferring phylogenies. Sinauer Associates, Sunderland, MA.

Ferris, S. D., A. C. Wilson, and W. M. Brown. 1981. Evolutionary tree for apes and humans based on cleavage maps of mitochondrial DNA. Proc. Natl. Acad. Sci. USA 78:2432–2436.

Fujita, P. A., B. Rhead, A. S. Zweig, A. S. Hinrichs, D. Karolchik, M. S. Cline, M. Goldman, G. P. Barber, H. Clawson, A. Coelho, et al. 2011. The UCSC Genome Browser database: update 2011. Nucleic Acids Res. 39:D876–D882.

Galtier, N., B. Nabholz, S. Glémin, and G. D. Hurst. 2009. Mitochondrial DNA as a marker of molecular diversity: a reappraisal. Mol. Ecol. 18:4541–4550.

Garrigan, D., S. B. Kingan, A. J. Geneva, P. Andolfatto, A. G. Clark, K. R. Thornton, and D. C. Presgraves. 2012. Genome sequencing reveals complex speciation in the *Drosophila simulans* clade. Genome Res. 22:1499–1511.

Gossmann, T. I., M. Woolfit, and A. Eyre-Walker. 2011. Quantifying the variation in the effective population size within a genome. Genetics 189:1389–1402.

Haddrill, P. R., D. L. Halligan, D. Tomaras, and B. Charlesworth. 2007. Reduced efficacy of selection in regions of the *Drosophila* genome that lack crossing over. Genome Biol. 8:R18.

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

Hellmann, I., I. Ebersberger, S. E. Ptak, S. Pääbo, and M. Przeworski. 2003. A neutral explanation for the correlation of diversity with recombination rates in humans. Am. J. Hum. Genet. 72:1527–1535.

Hey, J., and R. M. Kliman. 2002. Interactions between natural selection, recombination and gene density in the genes of *Drosophila*. Genetics 160:595–608.

Hixson, J. E., and W. M. Brown. 1986. A comparison of the small ribosomal RNA genes from the mitochondrial DNA of the great apes and humans: sequence, structure, evolution, and phylogenetic implications. Mol. Biol. Evol. 3:1–18.

Hobolth, A., O. F. Christensen, T. Mailund, and M. H. Schierup. 2007. Genomic relationships and speciation times of human, chimpanzee, and gorilla inferred from a coalescent hidden Markov model. PLoS Genet. 3:e7.

Hobolth, A., J. Y. Dutheil, J. Hawks, M. H. Schierup, and T. Mailund. 2011. Incomplete lineage sorting patterns among human, chimpanzee, and orangutan suggest recent orangutan speciation and widespread selection. Genome Res. 21:349–356.

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

———. 1983b. Testing the constant-rate neutral allele model with protein sequence data. Evolution 37:203–217.

———. 1990. Gene genealogies and the coalescent process. Pp. 1–44 *in* D. Futuyma and J. Antonovics, eds. Oxford surveys in evolutionary biology. Oxford University Press, Oxford, U.K.

International HapMap Consortium, K. A. Frazer, D. G. Ballinger, D. R. Cox, D. A. Hinds, L. L. Stuve, R. A. Gibbs, J. W. Belmont, A. Boudreau, P. Hardenbol, S. M. Leal, et al. 2007. A second generation human haplotype map of over 3.1 million SNPs. Nature 449:851–861.

Kaplan, N. L., R. R. Hudson, and C. H. Langley. 1989. The "hitchhiking effect" revisited. Genetics 123:887–899.

Koop, B. F., M. Goodman, P. Xu, K. Chan, and J. L. Slightom. 1986. Primate η-globin DNA sequences and man's place among the great apes. Nature 319:234–238.

Koop, B. F., D. A. Tagle, M. Goodman, and J. L. Slightom. 1989. A molecular view of primate phylogeny and important systematic and evolutionary questions. Mol. Biol. Evol. 6:580–612.

Langley, C. H., K. Stevens, C. Cardeno, Y. C. Lee, D. R. Schrider, J. E. Pool, S. A. Langley, C. Suarez, R. B. Corbett-Detig, B. Kolaczkowski, et al. 2012. Genomic variation in natural populations of *Drosophila melanogaster*. Genetics 192:533–598.

Larkin, M. A., G. Blackshields, N. P. Brown, R. Chenna, P. A. McGettigan, H. McWilliam, F. Valentin, I. M. Wallace, A. Wilm, R. Lopez, et al. 2007. Clustal W and Clustal X version 2.0. Bioinformatics 23:2947–2948.

Legrand, D., T. Chenel, C. Campagne, D. Lachaise, and M. L. Cariou. 2011. Inter-island divergence within *Drosophila mauritiana*, a species of the *D. simulans* complex: past history and/or speciation in progress? Mol. Ecol. 20:2787–2804.

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

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

Maynard Smith, J., and J. Haigh. 1974. Hitch-hiking effect of a favorable gene. Genet. Res. 23:23–35.

McNabney, D. R. 2012. The genetic basis of behavioral isolation between *Drosophila mauritiana* and *D. sechellia*. Evolution 66:2182–2190.

McQuilton, P., S. E. St Pierre, and J. Thurmond. 2012. FlyBase 101—the basics of navigating FlyBase. Nucleic Acids Res. 40:D706–714.

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

Meiklejohn, C. D., K. L. Montooth, and D. M. Rand. 2007. Positive and negative selection on the mitochondrial genome. Trends Genet. 23:259–263.

Nachman, M. W., V. L. Bauer, S. L. Crowell, and C. F. Aquadro. 1998. DNA variability and recombination rates at X-linked loci in humans. Genetics 150:1133–1141.

Nunes, M. D. S., P. O. Wengel, M. Kreissl, and C. Schlotterer. 2010. Multiple hybridization events between *Drosophila simulans* and *Drosophila mauritiana* are supported by mtDNA introgression. Mol. Ecol. 19:4695–4707.

Osada, N., and C. I. Wu. 2005. Inferring the mode of speciation from genomic data: a study of the great apes. Genetics 169:259–264.

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

Patterson, N., D. J. Richter, S. Gnerre, E. S. Lander, and D. Reich. 2006. Genetic evidence for complex speciation of humans and chimpanzees. Nature 441:1103–1108.

Piganeau, G., and A. Eyre-Walker. 2009. Evidence for variation in the effective population size of animal mitochondrial DNA. PLoS One 4:e4396.

Pollard, D. A., V. N. Iyer, A. M. Moses, and M. B. Eisen. 2006. Widespread discordance of gene trees with species tree in *Drosophila*: evidence for incomplete lineage sorting. PLoS Genet. 2:e173.

Presgraves, D. C. 2003. A fine-scale genetic analysis of hybrid incompatibilities in *Drosophila*. Genetics 163:955–972.

Prüfer, K., K. Munch, I. Hellmann, K. Akagi, J. R. Miller, B. Walenz, S. Koren, G. Sutton, C. Kodira, R. Winer, et al. 2012. The bonobo genome compared with the chimpanzee and human genomes. Nature 486:527–531.

Rand, D. M. 2011. Population genetics of the cytoplasm and the units of selection on mitochondrial DNA in *Drosophila melanogaster*. Genetica 139:685–697.

Rokas, A., D. Krueger, and S. B. Carroll. 2005. Animal evolution and the molecular signature of radiations compressed in time. Science 310:1933–1938.

Saitou, N., and M. Nei. 1986. The number of nucleotides required to determine the branching order of three species, with special reference to the human-chimpanzee-gorilla divergence. J. Mol. Evol. 24:189–204.

Satta, Y., J. Klein, and N. Takahata. 2000. DNA archives and our nearest relative: the trichotomy problem revisited. Mol. Phylogenet. Evol. 14:259–275.

Scally, A., J. Y. Dutheil, L. W. Hillier, G. E. Jordan, I. Goodhead, J. Herrero, A. Hobolth, T. Lappalainen, T. Mailund, T. Marques-Bonet, et al. 2012. Insights into hominid evolution from the gorilla genome sequence. Nature 483:169–175.

Smukowski, C. S., and M. A. F. Noor. 2011. Recombination rate variation in closely related species. Heredity 107:496–508.

Stajich, J. E., and M. W. Hahn. 2005. Disentangling the effects of demography and selection in human history. Mol. Biol. Evol. 22:63–73.

Stamatakis, A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22:2688–2690.

Tajima, F. 1983. Evolutionary relationship of DNA sequences in finite populations. Genetics 105:437–460.

Takahata, N., Y. Satta, and J. Klein. 1995. Divergence time and population size in the lineage leading to modern humans. Theor. Popul. Biol. 48:198–221.

Ting, C. T., S. C. Tsaur, and C. I. Wu. 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.

White, M. A., C. Ané, C. N. Dewey, B. R. Larget, and B. A. Payseur. 2009. Fine-scale phylogenetic discordance across the house mouse genome. PLoS Genet. 5:e1000729.

Associate Editor: L. Kubatko