MOLECULAR ECOLOGY

Molecular Ecology (2016) **25**, 2361–2372 doi: 10.1111/mec.13544

GENOMICS OF HYBRIDIZATION

Reticulate evolutionary history and extensive introgression in mosquito species revealed by phylogenetic network analysis

DINGQIAO WEN,* YUN YU,* MATTHEW W. HAHN†‡ and LUAY NAKHLEH*§

*Department of Computer Science, Rice University, Houston, TX 77005, USA, †Department of Biology, Indiana University, Bloomington, IN 47405, USA, ‡School of Informatics and Computing, Indiana University, Bloomington, IN 47405, USA, §Department of BioSciences, Rice University, Houston, TX 77005, USA

Abstract

The role of hybridization and subsequent introgression has been demonstrated in an increasing number of species. Recently, Fontaine *et al.* (*Science*, 347, 2015, 1258524) conducted a phylogenomic analysis of six members of the *Anopheles gambiae* species complex. Their analysis revealed a reticulate evolutionary history and pointed to extensive introgression on all four autosomal arms. The study further highlighted the complex evolutionary signals that the co-occurrence of incomplete lineage sorting (ILS) and introgression can give rise to in phylogenomic analyses. While tree-based methodologies were used in the study, phylogenetic networks provide a more natural model to capture reticulate evolutionary histories. In this work, we reanalyse the *Anopheles* data using a recently devised framework that combines the multispecies coalescent with phylogenetic networks. This framework allows us to capture ILS and introgression simultaneously, and forms the basis for statistical methods for inferring reticulate evolutionary histories. The new analysis reveals a phylogenetic network with multiple hybridization events, some of which differ from those reported in the original study. To elucidate the extent and patterns of introgression across the genome, we devise a new method that quantifies the use of reticulation branches in the phylogenetic network by each genomic region. Applying the method to the mosquito data set reveals the evolutionary history of all the chromosomes. This study highlights the utility of 'network thinking' and the new insights it can uncover, in particular in phylogenomic analyses of large data sets with extensive gene tree incongruence.

Keywords: *Anopheles gambiae*, hybridization, incomplete lineage sorting, introgression, phylogenetic networks

Received 3 October 2015; revision received 15 December 2015; accepted 6 January 2016

Introduction

In a recent study, Fontaine *et al.* (2015) conducted phylogenomic analyses of the species complex including the malaria vector, *Anopheles gambiae*. The authors reported a reticulate evolutionary history of this group, including extensive introgression patterns across all four autosomal chromosome arms. They inferred a species tree based on the X chromosome and used information on sequence divergence from the autosomes to hypothesize three hybridization events. This study of recently diverged species highlighted two processes that can be at play during evolution and must be accounted for in phylogenomic analyses. On the one hand, the low levels of divergence mean that species can hybridize and that their genomes may carry introgressed genetic material. On the other hand, the short times between speciation events mean that incomplete lineage sorting (ILS) is likely to occur. Phylogenomic

Correspondence: Luay Nakhleh, Fax: 713 348 3959; E-mail: nakhleh@rice.edu

analyses must account for the possibility of both of these processes acting to understand the evolutionary history of rapid radiations.

The multispecies coalescent (MSC) (Degnan & Rosenberg 2009) has recently emerged as a powerful model for gene genealogies inside the branches of a species tree, including when ILS is involved. A wide array of methods have been devised for inferring species trees from multilocus data sets based on the MSC, including maximum-likelihood (Kubatko *et al.* 2009; Wu 2012) and Bayesian approaches (Liu 2008; Heled & Drummond 2010). Nakhleh (2013) reviews some of the recent computational developments in this area. However, these models only consider ILS as a cause of incongruence. If reticulation events are a cause of incongruence, not only would these methods fail to detect them, but they would also propose very inaccurate branch lengths and population sizes in order to explain all the incongruence.

When reticulation alone is considered, all incongruence among estimated gene trees is taken to be due to reticulation, and a phylogenetic network is inferred that can combine, or reconcile, all these trees (Nakhleh 2010). A phylogenetic network extends the species tree topological model by allowing for reticulation events via nodes in the network with two parents. Recently developed parsimony methods can infer phylogenetic networks with the smallest number of reticulations required to reconcile a set of conflicting gene trees (van Iersel *et al.* 2010; Wu 2010). If ILS is the sole cause of gene tree incongruence, these methods will overestimate the number of reticulations and will incorrectly infer the timing of reticulations. In particular, if there is extensive incongruence in the data set—which is becoming a common theme of almost all phylogenomic analyses—these methods will result in overly complex phylogenetic networks.

In addition to the aforementioned phylogenomic analysis of mosquitoes, several recent studies have highlighted the co-occurrence of ILS and reticulation in many clades (Eriksson & Manica 2012; Moody & Rieseberg 2012; Staubach *et al.* 2012; The Heliconious Genome Consortium 2012; Liu *et al.* 2014; Marcussen *et al.* 2014). Therefore, developing methods that account for the two processes, rather than assume one or the other, has become essential.

We recently extended the MSC so that the process operates within the branches of a phylogenetic network that includes reticulation events (Yu *et al.* 2012, 2014). Under this extended model—the multispecies network coalescent, or MSNC—it becomes possible to infer a phylogenetic network while accounting simultaneously for both ILS and reticulation (Yu *et al.* 2012, 2013b, 2014). This work neither assumes nor requires knowledge of an underlying species tree, unlike several methods that can differentiate the two processes only if the correct species branching order is known (Green *et al.* 2010; Durand *et al.* 2011; Pease & Hahn 2015). Instead, it infers a phylogenetic network from the gene trees. While there are likelihood-based computations that account for gene flow and work directly from sequences, they exist for very limited cases (Hearn *et al.* 2014). Phylogenetic networks subsume trees and, consequently, using them allows for new evolutionary analyses (Bapteste *et al.* 2013).

Here we use these methods, as implemented in PHYLONET (Than *et al.* 2008), to infer the reticulate evolutionary history of the six genomes in the *A. gambiae* species complex. Our analyses reveal a reticulate evolutionary history of the species that encompasses the species tree used in Fontaine *et al.* (2015), but that also posits some different reticulation events. These results demonstrate the power that phylogenetic networks provide not only for understanding how species and genomes evolve, but also for better understanding how genes evolve. The PHYLONET software package (http://bioinfo.cs.rice.edu/phylonet) implements all the utilities that facilitated the analyses reported here.

Materials and methods

The Anopheles gambiae data

We downloaded the MAF genome alignment from high-depth field samples of *Anopheles* species from Dryad (doi: 10.5061/dryad.f4114). The data consist of one genome from each of the species *Anopheles gambiae* (G), *Anopheles coluzzii* (C), *Anopheles arabiensis* (A), *Anopheles quadriannulatus* (Q), *Anopheles merus* (R) and *Anopheles melas* (L). *Anopheles christyi* serves as the outgroup for rooting the gene trees. We used the data set for two different tasks: phylogenetic network inference and introgression detection for each genomic window.

As the phylogenetic network inference method of Yu *et al.* (2014) assumes independent loci, we sampled loci (genomic windows) so that every two loci were at least 64 kb apart. On average, loci were about 3.4 kb in length, with about 1000 loci having length smaller than 2000 bases (Fig. S1, Supporting information shows the histogram of locus length frequencies). While the 64 kb sampling window was held constant during the sampling, the lengths of loci were determined by the data, as the chromosomes were partitioned and not contiguous in the original data. The number of loci we sampled from chromosomes 2L, 2R, 3L, 3R and X are 669, 849, 564, 709 and 228, respectively. To separate the 2La (20 521 765–42 163 507) and 3La (14 452 080–35 641 019) inversions from other regions of the 2L and 3L chromosomes, we

used the reference (PEST) genome coordinates provided by Fontaine *et al.* (2015). For chromosome 2L, 308 of the 669 sampled loci are from the 2La region. For chromosome 3L, 299 of the 564 sampled loci are from the 3La region. For each locus, we estimated 100 bootstrap gene trees using RAXML8 (Stamatakis 2014) under the GTRGAMMA model. We used the bootstrap trees directly in the network inference process as discussed in Yu *et al.* (2014) and described below. To obtain an estimate of how much signal there is in the data, for each locus we computed the majority-rule consensus and counted the number of internal branches in the resulting tree. As each gene tree is rooted and has six leaves, a fully resolved tree would have four internal branches. Of the majority-rule consensus trees of the loci, 82 had zero internal branches (a star phylogeny), 382 had one internal branch, 624 had two internal branches, 840 had three internal branches, and 863 had four internal branches (i.e. were fully resolved).

The phylogenetic network model

A phylogenetic network extends the rooted phylogenetic tree model by incorporating nodes with two parents, also called reticulation nodes, to allow for hybridization; see Fig. 1 for an illustration. More formally, the topology of a phylogenetic network is a rooted, directed, acyclic graph. The node with no parents corresponds to the root, nodes with single parents correspond to speciation or divergence events, and nodes with two parents are reticulation nodes. The leaves (the nodes with no children) of the phylogenetic network are labelled uniquely by a set of taxa of interest. In our case, each leaf will be labelled uniquely by a species name. If a collection of subpopulations, instead of different species, is being modelled, then each leaf would be labelled uniquely by a subpopulation name or label. This topological model is similar to the admixture graph model proposed in Reich *et al.* (2009). The branch lengths of the phylogenetic network are given in coalescent units, where one unit equals $2N_e$ generations, where N_e is the effective population size. In the admixture graph model of Reich *et al.* (2009), branch lengths correspond to genetic drift values that measure variation in allele frequency corresponding to random sampling of alleles from generation to generation in a finite-size population. Given the way branch lengths are modelled in the phylogenetic network, population sizes and generation times are not assumed to be constant across the branches of the phylogenetic network. As a consequence, the phylogenetic network need not be ultrametric; that is, the sum of branch lengths in coalescent units from the root to a leaf can vary depending on the choice of the leaf. However, it is important to point out that in using gene tree topologies alone in our likelihood framework (i.e. disregarding their branch lengths), the individual population sizes and number of generations associated with a branch cannot be identified. A branch length of k coalescent units corresponds to any of an infinite number of combinations of branch lengths and numbers of generations.

Tracing the evolution of a set of lineages from the leaves of a species tree backward in time, there is always a unique path towards the root. This is not so in the case of phylogenetic networks, as reticulation nodes give rise to multiple paths. Therefore, while the topology and branch lengths of a species tree are sufficient to capture the probability distribution of gene tree topologies, that is not the case for phylogenetic networks. To complete the model, we associate with the branches incoming into a reticulation node an inheritance probability that turns the phylogenetic network into a model that describes a full probability distribution over the gene trees (Yu *et al.* 2012, 2014). While these inheritance probabilities are similar to the admixture proportions (f) in the model of Reich *et al.* (2009), they are not identical. As defined in Yu *et al.* (2012, 2014), inheritance probabilities could be locus specific (we use here one value for all loci for computational feasibility), as the proportion of admixture might vary from one locus to another. It is unclear locus-specific admixture can be modelled in the admixture graph model. Furthermore, in the model of Yu *et al.* (2012, 2014), and looking backward in time, any number of lineages could follow one of the two parents at a reticulation node and all remaining lineages would follow the other parent. This is more general than the model of Reich *et al.* (2009). Admixture graphs and admixture proportions are also used in the model of Pickrell & Pritchard (2012).

Phylogenetic network inference

Let Ψ be a phylogenetic network (with its branch lengths) and Γ be the inheritance probabilities. Further, let g be the gene tree, or 'local genealogy', estimated

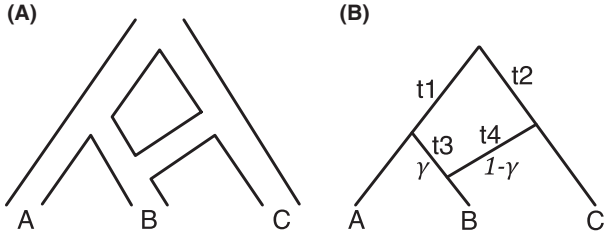

Fig. 1 The phylogenetic network model. (A) Phylogenetic network on three taxa A, B and C, showing hybridization between B and C. (B) The abstract model of the phylogenetic network, with the lengths of internal branches (t1, t2, t3 and t4) and inheritance probability (γ).

from the sequence alignment of some locus. We denote by $H_\Psi(g)$ the set of all coalescent histories of g given the phylogenetic network Ψ (Yu *et al.* 2012). Figure 2 shows a gene tree and two possible coalescent histories that explain its evolution within the branches of the phylogenetic network in Fig. 1. Each coalescent history h in the set $H_\Psi(g)$ defines a mapping of the genealogy g onto the phylogenetic network Ψ; $P(h\,|\,\Psi,\Gamma)$, the probability of that coalescent history, can be computed as in Yu *et al.* (2012). Then, the probability of g given Ψ and the inheritance probabilities Γ is given by

$$P(g|\Psi,\Gamma) = \sum_{h\in H_\Psi(g)} P(h|\Psi,\Gamma)$$

Notice that this formulation accounts for the effects of genetic drift on the gene tree topologies by allowing for deep coalescence. Finally, given m gene tree $\mathcal{G} = \{g_1, g_2, \ldots, g_m\}$ estimated from m independent loci, the likelihood of the model (Ψ, Γ) is given by

$$\mathcal{L}(\Psi,\Gamma|\mathcal{G}) = P(\mathcal{G}|\Psi,\Gamma) = \prod_{i=1}^{m} P(g_i|\Psi,\Gamma)$$

If instead of a single gene tree we have a set of gene trees inferred per locus (e.g. the set of bootstrap replicates), then $P(g_i|\Psi,\Gamma)$ is summed over all trees for inferred for that locus and normalized by the number of trees. In our analyses here, we use 100 bootstrap trees per locus. Yu *et al.* (2014) devised a hill-climbing search heuristic for obtaining a maximum-likelihood estimate (Ψ , Γ), from the data, $\mathcal{G}$.

The models of Reich *et al.* (2009) and Pickrell & Pritchard (2012) assume sequence data. While the method of Reich *et al.* (2009) assumes a known phylogenetic tree, the method of Pickrell & Pritchard (2012) employs a hill-climbing heuristic to search for the graph structure as well as the branch lengths and admixture proportions. In this sense, our inference method is similar to the latter. However, our method assumes gene tree estimates for the input data and poses no constraints on the topologies it searches or how it searches the space [e.g. the hill-climbing heuristic of Pickrell & Pritchard (2012) first searches for an optimal tree, then for the single optimal migration event to add, and then for the second one].

Peter (2015) recently reported on connections between admixture graph models and coalescent-based statistics using gene trees. While our computational framework makes use of gene tree estimates, more work needs to be performed to establish the similarities and differences between the two lines of work. Furthermore, although the search allows for evaluating hypotheses with and without reticulation, the model does not allow for distinguishing between reticulation and ancestral structure. The recent work of Lohse & Frantz (2014) uses a likelihood-based framework to distinguish between gene flow and ancestral structure, and the method is applicable to three genomes. The states of the discrete-time Markov chain in the model of Lohse & Frantz (2014) are similar to the ancestral configurations in the work of Yu *et al.* (2013b).

Quantifying introgression

For a branch, e, in phylogenetic network, Ψ, and a coalescent history, h, of a gene tree, g, we define the indicator function $\mathbb{I}(e \in h) = 1$ if e is 'used' by h (i.e. at least one lineage enters e under coalescent history h) and $\mathbb{I}(e \in h) = 0$ otherwise. For example, let e be the branch that indicates hybridization between B and C in the phylogenetic network shown in Fig. 2. For the coalescent history h in Fig. 2B, we have $\mathbb{I}(e \in h) = 0$ since that coalescent history does not use the branch e. However, for the coalescent history h in Fig. 2C, we have $\mathbb{I}(e \in h) = 1$ since that coalescent history uses the branch e.

Finally, for every coalescent history, h, of a gene tree, g, we define $\omega(h) = P(h)/P(g)$. As $P(g) = \sum_{h\in H_\Psi(g)} P(h)$, it follows that $0 \le \omega(h) \le 1$ and $\sum_{h\in H_\Psi(g)} \omega(h) = 1$. In other words, the probabilities of all coalescent histories of a gene tree sum to 1.

We are interested in quantifying for each branch, e, that is incident with a reticulation node in Ψ whether

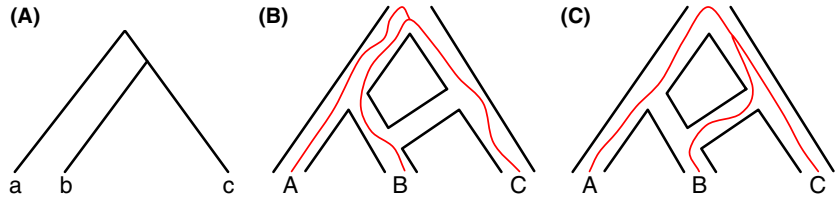

Fig. 2 Gene trees within the branches of a phylogenetic network. (A) A gene tree at one locus sampled from the three taxa, A, B and C. (B) A coalescent history of the gene tree within the branches of the phylogenetic network of Fig. 1. This coalescent history reflects a scenario of incomplete lineage sorting, but no introgression. (C) A coalescent history of the gene tree that reflects a scenario involving introgression. Note that there are other possible coalescent scenarios for this gene tree given the network.

locus L was transferred across e (i.e. whether its gene tree g 'used' branch e). We first set the inheritance probability to 1 for every branch in the phylogenetic network. We then compute $\alpha_g(e)$ according to

$$\alpha_g(e) = \sum_{h \in H_\Psi(g)} (\mathbb{I}(e \in h) \times \omega(h)) \quad (1)$$

When multiple trees are inferred per locus (e.g. bootstrap trees), the formula is modified so that $\alpha_g(e)$ is summed over all such trees, and then the value is normalized by the sum of probabilities of those trees. Notice that $0 \quad \alpha_g(e) \leq 1$ for every e. In other words, $\alpha_g(e)$ can be interpreted as the probability that locus L followed branch e (based on the gene tree that was estimated for that locus).

Ultimately, we are interested in inferring whether locus L was transferred across edge e or not; that is, we are interested in a binary outcome. To achieve this, we use a threshold τ and obtain $\beta_g(e) \in \{0, 1\}$ as follows:

$$\beta_g(e) = \begin{array}{ll} 1 & \text{if } \alpha_g(e) \geq \tau \\ 0 & \text{if } \alpha_g(e) < \tau \end{array} \quad (2)$$

In the Results section, we plot the values of β and discuss the choice of τ.

Results

Evaluating the phylogeny of Fontaine et al. (2015)

To compute the likelihood of the phylogenetic network proposed in fig. 1C by Fontaine *et al.* (2015), we used PHYLONET to optimize the branch lengths and inheritance probabilities of the phylogenetic network using the bootstrap gene tree estimates from the sampled loci on the autosomes (see Materials and methods). The optimized inheritance probabilities are shown in Fig. 3A and the log likelihood of the network is 12443.636. The main observation from this result is that the inheritance probabilities on the reticulation edges between *Anopheles arabiensis* and the common ancestor of *Anopheles coluzzii* and *Anopheles gambiae* are very high. This indicates that the autosomes give a strong signal that these three species are grouped together. Indeed, gene tree analyses conducted in fig. 2 by Fontaine *et al.* (2015) clearly show that an overwhelming majority of gene trees built from the autosomes support an [*A. arabiensis* (*A. coluzzii*, *A. gambiae*)] grouping.

As the three reticulation events reported in fig. 1C of Fontaine *et al.* (2015) were hypothesized based on analyses of gene trees and coalescence times, we set out to test what reticulation events PHYLONET would detect if we fix the species tree topology of the original study. To achieve this, we searched the space of all phylogenetic networks that could be formed by adding up to three reticulation edges to the fixed species tree topology. Branch lengths and inheritance probabilities for each resulting network were optimized during the search to maximize the phylogenetic networks' likelihoods. The optimal network identified is shown in Fig. 3B.

This analysis reveals similar patterns to those of the previous analysis with respect to *A. arabiensis*, *A. coluzzii* and *A. gambiae*. However, the optimal phylogenetic network now posits a reticulation edge from *Anopheles quadriannulatus* to *Anopheles merus*, which is in the opposite direction of that in the network of Fig. 3A. It is important to note here that the likelihood improves significantly with only this difference, pointing to strong support in the data for this direction of the reticulation. Given these results, we set out to infer phylogenetic networks under maximum likelihood without restricting the search to either the phylogenetic network or underlying species tree of Fontaine *et al.* (2015).

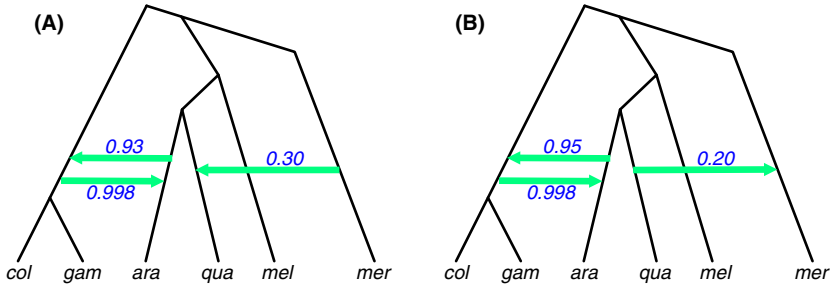

Fig. 3 Phylogenetic inference results based on the reticulate evolutionary history in fig. 1C of Fontaine *et al.* (2015). (A) The phylogenetic network topology reported in Fontaine *et al.* (2015) was used and only its branch lengths and inheritance probabilities were optimized to maximize the likelihood. The log likelihood of this optimized network is 12443.36. (B) Only the underlying species tree of Fontaine *et al.* (2015) was used, and search under maximum likelihood was conducted to identify the three top reticulation events. The log likelihood of this optimized network is 12382.18. The numbers on the horizontal edges indicate the estimated inheritance probabilities.

Phylogenetic network inference from the gene tree data

We inferred phylogenetic networks from the (bootstrap) gene tree data on the autosomes by searching for the optimal phylogenetic networks with 0, 1, 2 and 3 reticulations. The results are shown in Fig. 4. The phylogenetic tree shown in Fig. 4A amounts to treating all incongruence as a result of only ILS and no introgression. In this case, the inference is based on maximum likelihood under the MSC. While this tree seems, at first glance, very different from the species tree in Fontaine *et al.* (2015), it can be obtained from the latter tree by one simple move: grouping the (*A. coluzzii*, *A. gambiae*) clade with *A. arabiensis* as a sibling of *A. quadriannulatus*. Once again, the data under the maximum-likelihood criterion support such a grouping in the species tree, which differs from the grouping supported by data from the X chromosome. It is worth mentioning that when we inferred a species tree under the MSC using only the X chromosome data, the tree agreed with that in Fontaine *et al.* (2015).

The optimal phylogenetic single-reticulation network (Fig. 4B) consisted of the optimal phylogenetic tree with the addition of a hybridization between *A. quadriannulatus* and *A. merus* (direction from the former to the latter). The inheritance probability of this additional reticulation edge is 0.21. The optimal phylogenetic network with two reticulations posits an additional reticulation edge from *Anopheles melas* to *A. merus*. The estimated inheritance probability of this horizontal edge in Fig. 4C is 0.42.

The optimal phylogenetic network with three reticulations is the optimal two-reticulation network with an additional reticulation from *A. quadriannulatus* to *A. gambiae*. The likelihood of this network (Fig. 4D) is significantly higher than that of any of the other networks. The optimized inheritance probability of this additional reticulation is 0.03. This reticulation edge is mainly supported by the 2La inversion region.

The orange dotted reticulation edge in Fig. 4D is inferred from the X chromosome data alone, with inheritance probability of about 0.73 (this reticulation edge is not supported by the autosome data). Indeed, the gene tree analysis (fig. 2 in Fontaine *et al.* 2015) demonstrates that about 64% of the X chromosome support gene genealogies that group (*A. coluzzii*, *A. gambiae*) as a separate clade from the clade (*A. arabiensis*, *A. quadriannulatus*).

Finally, it is important to highlight the effect of simultaneously accounting for ILS and introgression. When only ILS is accounted for (Fig. 4A), the branch lengths (in coalescent units) of the phylogeny are estimated to be very short (see Fig. S2, Supporting information). This is easy to explain as all incongruence among the gene trees in this case is assumed to be due to ILS whose extent depends on the branch lengths: the shorter they are, the more likely the phylogenetic network given that the data have extensive amounts of incongruence. As hybridization events are added to the phylogenetic networks, not only does the likelihood improve, but the

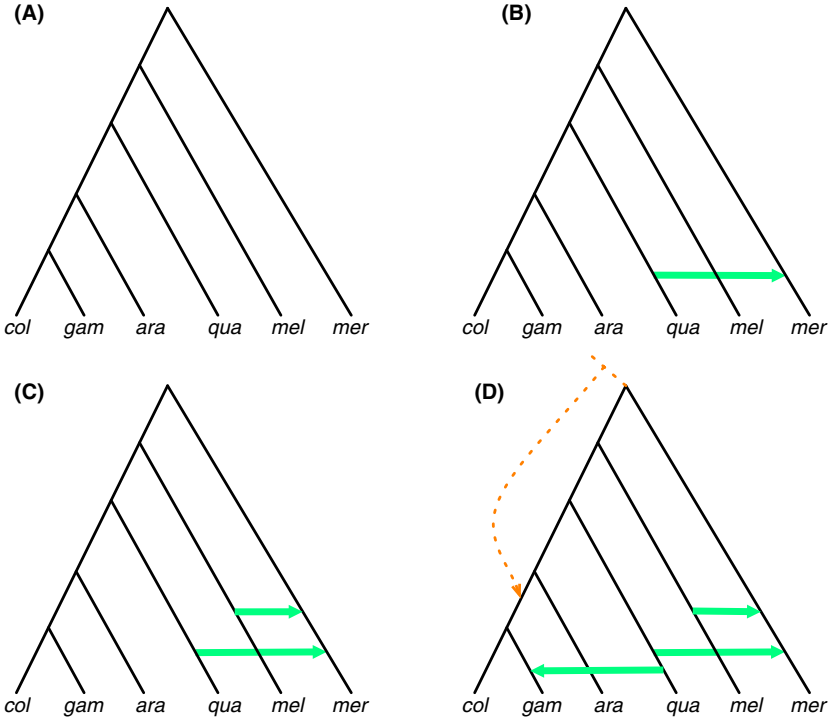

Fig. 4 Phylogenetic inference results when the reticulate evolutionary history in fig. 1C of Fontaine *et al.* (2015) is not used to guide the search. Here, the bootstrap gene trees from the sampled loci were used to infer optimal networks under maximum likelihood with 0, 1, 2 and 3 reticulation events. (A) The maximum-likelihood tree (network with 0 reticulations) estimate; the log likelihood of this tree is 12650.07. (B) The maximum-likelihood network with 1 reticulation that the search identified; the log likelihood of this network is 12401.37. (C) The maximum-likelihood network with two reticulations that the search identified; the log likelihood of this network is 12363.07. (D) The maximum-likelihood network with three reticulations that the search identified; the log likelihood of this network is 12295.53. The orange, dotted reticulation edge was identified using the data from the X chromosome.

branches also become longer (see Fig. S2, Supporting information). In this case, branch lengths need not be exceedingly short to fit the data, as the hybridization events help to explain a large portion of the incongruence.

Individual evolutionary histories along the chromosomes

To detect and quantify which branches are associated with each individual locus across the chromosomes, we used the data described in Materials and methods, but processed it differently. We first optimized the branch lengths and inheritance probabilities of the phylogenetic networks based on the autosome data and X chromosome data separately. We covered each chromosome with nonoverlapping 10 kb windows, and in each window inferred 100 bootstrap trees using RAXML8. Using the bootstrap trees for each window, we calculated the inheritance probabilities along each branch, α, on the respective optimized phylogenetic network as described above. Plots of the α values for each chromosome and each reticulation edge are provided in Figs S3–S7 (Supporting information).

Note that, from the phylogenetic network alone, we cannot determine which branch comes from the species tree, and which from introgression (see Discussion). In fact, we can speak of multiple 'parental species trees' within the network (Yu *et al.* 2012). Therefore, whether a branch is horizontal or vertical in a phylogenetic network is arbitrary without additional information. Without this assignment, the α values simply represent the inheritance probabilities along each reticulation edge, with no judgement about whether this represents 'introgression'. Here we have visualized the phylogenetic network to conform to the species tree proposed in Fontaine *et al.* (2015). Given this species tree, all four horizontal edges considered below do correspond to introgression.

To discretize the α values and to determine which branch a locus followed, we used $\tau = 0.7$ in eqn (2). We view the estimated phylogenetic network and α's in terms of fuzzy, or soft, clustering (Bezdek 2013). In standard clustering, each point in a data set is placed in a single cluster; this is what is called hard clustering. In fuzzy clustering, each point could be associated with more than one cluster, and a membership weight for each point-cluster pair denotes the strength of its association with the cluster. In the case of phylogenetic networks, and under the likelihood setting defined in Yu *et al.* (2012, 2014) and used here, each parental species tree inside the network can be viewed as the centroid of a cluster, and $\alpha_g(e)$ for locus L can be viewed as the probability that locus L is associated with the cluster of all parental species trees that 'use' reticulation edge e. Under this interpretation, we convert the fuzzy association of loci to parental species trees into hard assignments by choosing for each locus the cluster of parental species trees with the highest α value. As there are two possible clusters (the cluster of all parental species trees that use e and the cluster of all other parental species trees), the hard assignment is achieved by assigning the locus to the parental species tree with $\alpha > 0.5$. To avoid many false assignments, we used the more stringent value of $\tau = 0.7$; introgression plots based on this threshold are shown in Figs 5 and 6. Results based on $\tau = 0.5$ are given in Figs S8 and S9 (Supporting information).

As shown in Fig. 5, most of the histories that follow the edge from *A. quadriannulatus* to *A. gambiae* come from the 2La inversion region (see Discussion). For the introgression from *A. quadriannulatus* to *A. merus*, the 2L, 2R and 3R chromosomes have approximately the same percentage of genetic material that is inherited along this reticulation edge. The 3La inversion region has a high inheritance probability across this reticulation edge, contributing to a high total percentage of introgression on chromosome 3L. The 'chromoplot' in fig. 4 of Fontaine *et al.* (2015) also clearly shows that the 3La inversion has introgressed between *A. quadriannulatus* and *A. merus*. For the introgression from *A. melas* to *A. merus*, the reticulate signal is approximately uniform along each chromosome, and all chromosomes indicate a similar fraction of histories following the minor edge. Notice, however, that the two reticulation edges to *A. merus* are dependent: an *A. merus* lineage can follow at most one of them, but not both. Indeed, Fig. 5 shows a complementary pattern of evolution in the 3La inversion—very dense along one of the edges and very sparse along the other. For the fourth reticulation edge, almost all windows on each autosome followed this edge, indicating that most of the genetic material from *A. coluzzii* and *A. gambiae* follow this history. These panels were omitted from Fig. 5 because there was no spatial pattern of introgression.

For the X chromosome, Fig. 6 indicates very few histories across reticulation edges. Approximately 26% of loci follow the edge that corresponds to panel 4, the introgression from *A. arabiensis* into the ancestor of the (*A. coluzzii*, *A. gambiae*) clade. This is the same edge that 99.8% of the autosomes followed. A large proportion of introgressed histories on the X are in the 15–19 MB region, outside of the inversions distinguishing these species. These results are in clear agreement with Fontaine *et al.* (2015), with most of the X chromosome evolving down the 'species tree', and very little introgression.

Finally, when the less stringent threshold of 0.5 is used to discretize the inheritance probabilities (Figs S8

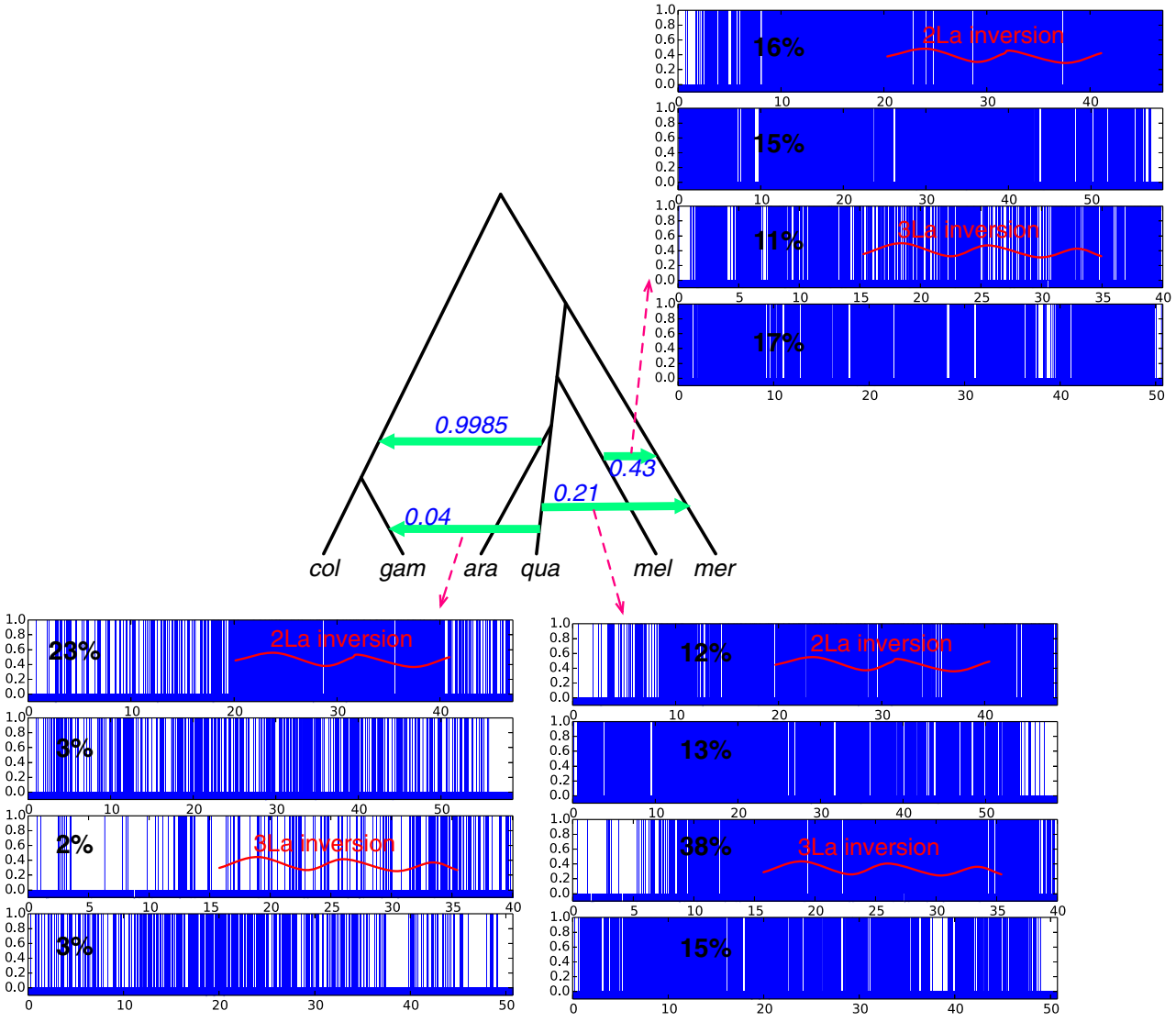

Fig. 5 The histories of each locus along the autosomes for each of three reticulation edges. The inheritance probability of each horizontal edge is shown. In each panel, the *x*-axis corresponds to the position along the chromosome, and the *y*-axis corresponds to whether the branch was followed by the locus (value 1) or not followed (value 0). For each reticulation edge, the four panels from top to bottom correspond to chromosomes 2L, 2R, 3L and 3R, respectively. The percentage within each plot is the fraction of loci along the chromosome arm that followed the respective edge. For the fourth edge with inheritance probability 99%, the panel is omitted, as it is very dense due to high rates of introgression, which are 86%, 85%, 89% and 78% for chromosomes 2L, 2R, 3L and 3R, respectively.

and S9, Supporting information), an additional 9–16% of autosomal windows and about 6–11% of X chromosome windows follow the minor branches. However, the qualitative patterns across each chromosome do not change very much.

Discussion

Fontaine *et al.* (2015) analysed the genomes of six members of the *Anopheles gambiae* species complex and found that extensive species/gene tree incongruence was due to both ILS and multiple introgression events. The authors presented an evolutionary history of the species with three major hybridization events posited across the branches of a species tree supported almost solely by the X chromosome. These analyses used estimated gene trees for different genomic regions as the basis for all inferences, but did not have a unified probabilistic approach to inferring reticulation events. In this study, we employ phylogenetic network methods to infer the evolutionary history of the species as well as the introgression patterns across the chromosomes.

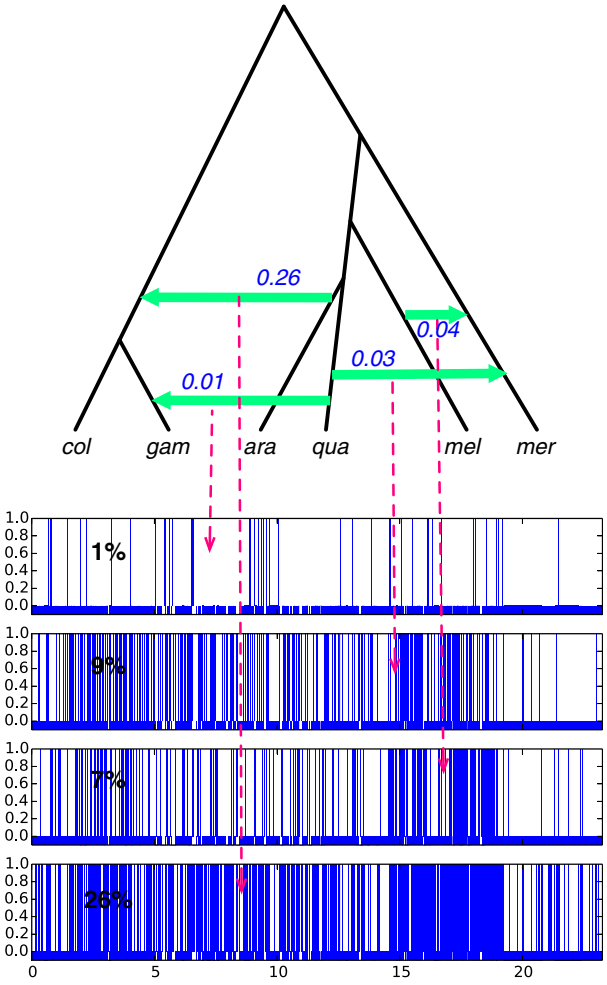

Fig. 6 The introgressed regions in the X chromosome across each of four reticulation edges. The inheritance probability calculated from the X chromosome for each horizontal edge is shown. In each panel, the *x*-axis corresponds to the positions along the chromosome, and the *y*-axis corresponds to introgressed (value 1) or nonintrogressed (value 0). The percentage within each plot is that of the introgressed parts of the X chromosome along the respective edge.

These methods are based on a model that extends the MSC (Degnan & Rosenberg 2009) to phylogenetic networks (which we call the multispecies network coalescent, or MSNC), thus accounting for ILS simultaneously with reticulation.

The methods revealed a reticulate evolutionary history that resembles the phylogeny reported by Fontaine *et al.* (2015), but disagrees in some of the inferred relationships. In particular, there were three main differences found here. First, the direction of introgression of one of the events identified in the previous study (from *Anopheles merus* into *Anopheles quadriannulatus*) has been reversed. Second, a major introgression event (from *Anopheles melas* into *A. merus*) was identified using PHYLONET. Re-examination of the data from Fontaine *et al.* (2015) supports both of these updated inferences. However, the third major difference—an edge connecting *A. quadriannulatus* and *A. gambiae* in the 2La region—is likely caused by balancing selection, and not introgression. Trans-specific balancing selection can mimic introgression in topology-based analyses (Liu *et al.* 2014), and that is almost certainly what is happening here. The 2La inversion is a balanced polymorphism that pre-dates the origin of the species complex, and differential loss of inversion arrangements places *A. quadriannulatus* and *A. gambiae* together in trees made from this region (Fontaine *et al.* 2015). In general, though, the analyses carried out here serve to highlight the utility of using a phylogenetic network. When there are so many hybridization events going on in a single clade, it is simply too hard (or impossible) to identify all such events by hand. In addition, using the machinery of phylogenetic networks allowed us to infer the extent and distribution of introgression for each genomic window. Such a task is not possible given only gene tree topologies for each window.

In addition to inferring introgression events, Fontaine *et al.* (2015) used information external to the gene tree topologies themselves to choose one topology as 'the' species tree. However, in most such cases it may not be possible to make this designation, nor does the phylogenetic network make this choice on its own. Clark & Messer (2015) noted that 'given that the bulk of the genome has a network of relationships that is different from this true species tree, perhaps we should dispense with the tree and acknowledge that these genomes are best described by a network, and that they undergo rampant reticulate evolution'. Indeed, this is what phylogenetic network-based analyses provide: they reconstruct networks and use them for subsequent analyses without designating any particular tree or path inside these networks as the species tree. In other words, phylogenetic networks naturally capture the ambiguity and challenge associated with delineating the exact speciation events in the presence of extensive introgression. However, this also means that denoting any single branch as introgressed or not is arbitrary. It may be more helpful simply to acknowledge the different routes any particular gene tree may have taken through the network.

We can further illustrate this issue with the phylogenetic network we inferred and discussed above. Figure 7 shows two different interpretations (out of many more possible interpretations) of the phylogenetic network inferred in Fig. 4D. In each panel, we have highlighted with thicker lines branches that correspond to a 'species tree', and with arrows the resulting hybridization events. Figure 7A shows the species tree proposed by Fontaine *et al.* (2015), along with four

reticulation events. Under this interpretation, the (*Anopheles coluzzii*, *A. gambiae*) clade splits directly from the root of the phylogeny and (*A. quadriannulatus*, *Anopheles arabiensis*) form a monophyletic group. Subsequently, hybridizations occurred between these two clades. Under the second interpretation, illustrated in Fig. 7B [*A. arabiensis* (*A. coluzzii*, *A. gambiae*)] form a clade whose sister taxon is *A. quadriannulatus*. Based on this interpretation, hybridization occurred between the (*A. coluzzii*, *A. gambiae*) clade and a species outside this group, and another hybridization occurred between *A. quadriannulatus* and *A. gambiae*.

The use of gene tree topologies without branch lengths to infer phylogenetic networks does not provide the power to distinguish between the two scenarios in Fig. 7. While coalescence time estimates on the gene trees could be informative about differentiating between these two scenarios, the likelihood framework we use here is very sensitive to the coalescence time estimates. It has been shown that coalescence times are poorly estimable in practice for individual gene trees, which results in poor estimates of the species phylogeny based on criteria that make use of these individual time estimates (DeGiorgio & Degnan 2014). If appropriate whole-genome or whole-chromosome data are available, however, clear hypotheses about coalescence times can distinguish between the species tree and introgression events (Fontaine *et al.* 2015).

The approach we used here relies on estimated gene trees for inferring phylogenetic networks and introgression patterns. As gene tree estimates are likely to have errors, it is important to account for this factor as it gives rise to signals that masquerade as ILS, introgression or both. In all analyses conducted and reported here, the set of all bootstrap trees for each locus was used to account for uncertainty in the gene tree estimates. While phylogenetic network inference could be robust to low levels of error in gene tree estimates—especially when a large number of loci are used—our introgression pattern analyses (like those in Fig. 5) are not robust to gene tree errors, as individual loci are analysed independently of all others. This further emphasizes the need to carefully account for gene tree uncertainty in such analyses.

As we illustrated above, when there is extensive gene tree incongruence in a data set, phylogenomic analyses that account only for ILS (i.e. analyses under the MSC) will estimate very short branches in the species tree, very large effective population sizes or both. Therefore, extra caution must be taken when interpreting the branch lengths on estimated species trees from MSC methods, particularly when they are estimated to be very small. Fortunately, when reticulation is also accounted for as a potential cause of incongruence, the branch lengths can be estimated much more accurately.

All the analyses reported here were conducted using the software package PHYLONET (Than *et al.* 2008), which implements all the methods described above for inferring phylogenetic networks and analysing data in their context. Currently, the computational requirements involved in calculating the likelihood of a phylogenetic network present the major hurdle in analysing larger data sets, and for inferring larger numbers of reticulation events. Yu *et al.* (2013a) introduced a parsimony criterion for inferring phylogenetic networks in the presence of ILS. While inference and network evaluation based on this criterion are much faster than under likelihood, they are also less accurate. More recently, inference based on pseudo-likelihood has been introduced (Solıs-Lemus & Ane 2015; Yu & Nakhleh 2015), yet performance analyses are still required to establish the full merit of this approach. In the future, we hope that even very large clades will be amenable to analysis by phylogenetic networks, opening up new possible inferences for a wide range of taxa.

Acknowledgements

This work was supported in part by NSF grant CCF-1302179 to LN and NIH grant R01-AI076584 to MWH and Nora Besansky. The data analyses were conducted on Data Analysis and Visualization Cyberinfrastructure (DAVinCI), which is funded by

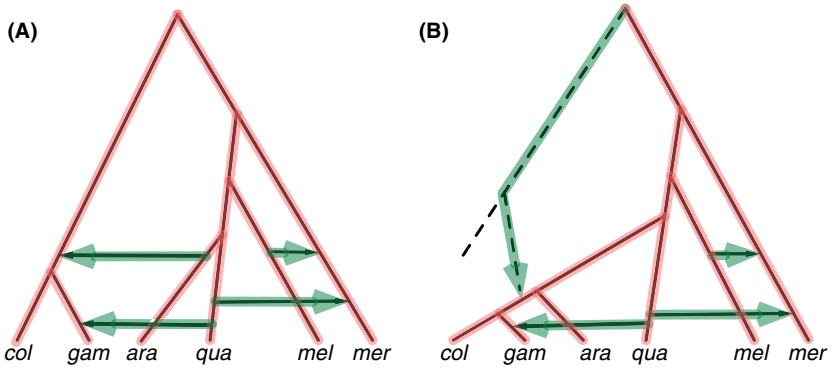

Fig. 7 (A) The species tree of Fontaine *et al.* (2015) is shown, as highlighted by thick lines inside the estimated phylogenetic network in Fig. 4D. The reticulation edges are highlighted by thick arrows. (B) A different species tree is highlighted by thick lines, along with the resulting hybridization events highlighted by thick arrows.

NSF under grant OCI-0959097 and Rice University, and BlueBioU, which is funded by NIH award NCRR S10RR02950, an IBM Shared University Research (SUR) Award in partnership with CISCO, Qlogic and Adaptive Computing, and Rice University. We thank James Pease for assistance with the genomic data.

References

Bapteste E, van Iersel L, Janke A *et al.* (2013) Networks: expanding evolutionary thinking. *Trends in Genetics*, **29**, 439–441.

Bezdek JC (2013) *Pattern Recognition with Fuzzy Objective Function Algorithms*. Springer Science & Business Media, New York.

Clark AG, Messer PW (2015) Conundrum of jumbled mosquito genomes. *Science*, **347**, 27–28.

DeGiorgio M, Degnan JH (2014) Robustness to divergence time underestimation when inferring species trees from estimated gene trees. *Systematic Biology*, **63**, 66–82.

Degnan J, Rosenberg N (2009) Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends in Ecology & Evolution*, **24**, 332–340.

Durand EY, Patterson N, Reich D, Slatkin M (2011) Testing for ancient admixture between closely related populations. *Molecular Biology and Evolution*, **28**, 2239–2252.

Eriksson A, Manica A (2012) Effect of ancient population structure on the degree of polymorphism shared between modern human populations and ancient hominins. *Proceedings of the National Academy of Sciences of the USA*, **109**, 13956–13960.

Fontaine MC, Pease JB, Steele A *et al.* (2015) Extensive introgression in a malaria vector species complex revealed by phylogenomics. *Science*, **347**, 1258524.

Green RE, Krause J, Briggs AW *et al.* (2010) A draft sequence of the neandertal genome. *Science*, **328**, 710–722.

Hearn J, Stone GN, Bunnefeld L, Nicholls JA, Barton NH, Lohse K (2014) Likelihood-based inference of population history from low-coverage de novo genome assemblies. *Molecular Ecology*, **23**, 198–211.

Heled J, Drummond AJ (2010) Bayesian inference of species trees from multilocus data. *Molecular Biology and Evolution*, **27**, 570–580.

van Iersel L, Kelk S, Rupp R, Huson D (2010) Phylogenetic networks do not need to be complex: using fewer reticulations to represent conflicting clusters. *Bioinformatics*, **26**, i124–i131.

Kubatko LS, Carstens BC, Knowles LL (2009) STEM: species tree estimation using maximum likelihood for gene trees under coalescence. *Bioinformatics*, **25**, 971–973.

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

Liu K, Steinberg E, Yozzo A, Song Y, Kohn M, Nakhleh L (2014) Interspecific introgressive origin of genomic diversity in the house mouse. *Proceedings of the National Academy of Sciences of the USA*, **112**, 196–201.

Lohse K, Frantz LA (2014) Neandertal admixture in Eurasia confirmed by maximum-likelihood analysis of three genomes. *Genetics*, **196**, 1241–1251.

Marcussen T, Sandve SR, Heier L *et al.* (2014) Ancient hybridizations among the ancestral genomes of bread wheat. *Science*, **345**, 1250092.

Moody M, Rieseberg L (2012) Sorting through the chaff, nDNA gene trees for phylogenetic inference and hybrid identification of annual sunflowers (*Helianthus* sect. *Helianthus*). *Molecular Phylogenetics and Evolution*, **64**, 145–155.

Nakhleh L (2010) Evolutionary phylogenetic networks: models and issues. In: *The Problem Solving Handbook for Computational Biology and Bioinformatics* (eds Heath L, Ramakrishnan N), pp. 125–158. Springer, New York, New York.

Nakhleh L (2013) Computational approaches to species phylogeny inference and gene tree reconciliation. *Trends in Ecology & Evolution*, **28**, 719–728.

Pease J, Hahn M (2015) Detection and polarization of introgression in a five-taxon phylogeny. *Systematic Biology*, **64**, 651–662.

Peter B (2015) Admixture, population structure and F-statistics. *bioRxiv* doi: 10.1101/028753.

Pickrell JK, Pritchard JK (2012) Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genetics*, **8**, e1002967.

Reich D, Thangaraj K, Patterson N, Price AL, Singh L (2009) Reconstructing indian population history. *Nature*, **461**, 489–494.

Solıs-Lemus C, Ane C (2015) *Inferring phylogenetic networks with maximum pseudolikelihood under incomplete lineage sorting*. arXiv preprint arXiv:1509.06075.

Stamatakis A (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, **30**, 1312–1313.

Staubach F, Lorenc A, Messer PW, Tang K, Petrov DA, Tautz D (2012) Genome patterns of selection and introgression of haplotypes in natural populations of the house mouse (*Mus musculus*). *PLoS Genetics*, **8**, e1002891.

Than C, Ruths D, Nakhleh L (2008) PhyloNet: a software package for analyzing and reconstructing reticulate evolutionary relationships. *BMC Bioinformatics*, **9**, 322.

The Heliconious Genome Consortium (2012) Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. *Nature*, **487**, 94–98.

Wu Y (2010) Close lower and upper bounds for the minimum reticulate network of multiple phylogenetic trees. *Bioinformatics*, **26**, 140–148.

Wu Y (2012) Coalescent-based species tree inference from gene tree topologies under incomplete lineage sorting by maximum likelihood. *Evolution*, **66**, 763–775.

Yu Y, Nakhleh L (2015) A maximum pseudo-likelihood approach for phylogenetic networks. *BMC Genomics*, **16**, S10.

Yu Y, Degnan JH, Nakhleh L (2012) The probability of a gene tree topology within a phylogenetic network with applications to hybridization detection. *PLoS Genetics*, **8**, e1002660.

Yu Y, Barnett R, Nakhleh L (2013a) Parsimonious inference of hybridization in the presence of incomplete lineage sorting. *Systematic Biology*, **62**, 738–751.

Yu Y, Ristic N, Nakhleh L (2013b) Fast algorithms and heuristics for phylogenomics under ILS and hybridization. *BMC Bioinformatics*, **14**, S6.

Yu Y, Dong J, Liu KJ, Nakhleh L (2014) Maximum likelihood inference of reticulate evolutionary histories. *Proceedings of the National Academy of Sciences of the USA*, **111**, 16448–16453.

All authors designed the research, conducted the analyses, and wrote the manuscript. DW ran the experiments.

Data accessibility

Input NEXUS files for the PHYLONET analyses conducted for this study: DRYAD entry doi: 10.5061/dryad.tn47c.

Data from Fontaine *et al.* (2015): DRYAD entry doi: 10.5061/dryad.f4114.

Supporting information

Additional supporting information may be found in the online version of this article.

Fig. S1 Histogram of the frequencies of lengths (in kb) of the loci used to infer the phylogenetic networks.

Fig. S2 Estimated branch lengths for the optimal networks in Fig. 4 in the main text, without the reticulation edge that is supported by the X chromosome (optimized to maximize the likelihood of the networks based on the data from the autosomes).

Fig. S3 The introgression probabilities [α values based on eqn (1) in the main text] for chromosome 2L across each of the three reticulation edges.

Fig. S4 The introgression probabilities [α values based on eqn (1) in the main text] for chromosome 2R across each of the three reticulation edges.

Fig. S5 The introgression probabilities [α values based on eqn (1) in the main text] for chromosome 3L across each of the three reticulation edges.

Fig. S6 The introgression probabilities [α values based on eqn (1) in the main text] for chromosome 3R across each of the three reticulation edges.

Fig. S7 The introgression probabilities [α values based on eqn (1) in the main text] for chromosome X across each of the four reticulation edges.

Fig. S8 The introgressed regions in each of the chromosomes across each of three reticulation edges when threshold value $\tau = 0.5$ is used to eqn (2) in the main text.

Fig. S9 The introgressed regions in the X chromosome across each of four reticulation edges when threshold value $\tau = 0.5$ is used to eqn (2) in the main text.