MOLECULAR ECOLOGY
RESOURCES

Molecular Ecology Resources (2012) **12**, 179–184 doi: 10.1111/j.1755-0998.2011.03065.x

JML: testing hybridization from species trees

SIMON JOLY
Institut de recherche en biologie végétale, Université de Montréal and Jardin botanique de Montréal, 4101 Sherbrooke Est, Montréal, Quebec, Canada H1X 2B2

Abstract

I introduce the software JML that tests for the presence of hybridization in multispecies sequence data sets by posterior predictive checking following Joly, McLenachan and Lockhart (2009, American Naturalist 174, e54). Although their method could potentially be applied on any data set, the lack of appropriate software made its application difficult. The software JML thus fills a need for an easy application of the method but also includes improvements such as the possibility to incorporate uncertainty in the species tree topology. The JML software uses a posterior distribution of species trees, population sizes and branch lengths to simulate replicate sequence data sets using the coalescent with no migration. A test quantity, defined as the minimum pairwise sequence distance between sequences of two species, is then evaluated on the simulated data sets and compared to the one estimated from the original data. Because the test quantity is a good predictor of hybridization events, departure from the bifurcating species tree model could be interpreted as evidence of hybridization. Software performance in terms of computing time is evaluated for several parameters. I also show an application example of the software for detecting hybridization among native diploid North American roses.

Keywords: Gene trees, hybridization, introgression, Rosa, species trees, the multispecies coalescent

Received 3 June 2011; revision received 9 August 2011; accepted 15 August 2011

Introduction

Hybridization is an important evolutionary process (Arnold 1997; Barton 2001). Its role in speciation (Rieseberg 1997; Rieseberg *et al.* 2003; Seehausen 2004; Mallet 2007) and adaptation (Arnold 2004; Joly & Schoen 2011) is understood theoretically and has also been confirmed experimentally. Yet, the role of hybridization is hard to confirm in many instances because it is often difficult to find statistical evidence for hybridization. Here, the term hybridization is used in the broad sense. That is, it refers both to the event, the successful mating between individuals from two distinct species, and its outcomes: hybrid speciation and introgression, where introgression is the transfer of genetic material between species via sexual reproduction. Typically, hybridization is detected using measures of gene tree incongruence (Funk & Omland 2003), either among gene trees or between the gene tree and the species tree, although other processes can be in cause. Thus, distinguishing between hybridization and other processes resulting in gene tree incongruence is a critical issue in evolutionary biology. A specific question that has received much attention is that of distinguishing incongruence caused by introgression from that caused by incomplete lineage sorting. Incomplete lineage sorting arises when ancestral polymorphisms present in the ancestral species have not been completely sorted out by genetic drift in the daughter species, resulting in nonmonophyletic species. Even though several methods have been described to address this problem, none provide a clear and general test for the presence of hybridization (reviewed in Joly *et al.* 2009).

Joly *et al.* (2009) proposed a method based on the idea that incomplete lineage sorting imposes a limit to the minimum expected distance between sequences of two species because the sequences compared have been diverging since the speciation event. Such limit does not exist for introgressed sequences. Consequently, it should be possible to statistically identify introgressed sequences when the pairwise distance between sequences found in two distinct species is smaller than that expected under a lineage sorting scenario. Simulations have confirmed that this statistic is able to detect introgression, although the success rate depends on several parameters: the relative timing of the hybridization and of speciation events, the population sizes and the sequence length (Joly *et al.* 2009). The method of Joly *et al.* (2009) has the potential to be applied on any data set, but the lack of software implementing the method has limited its use. Here, I introduce the software JML that implements the posterior predictive approach of Joly *et al.* (2009). I also improve the original approach by accounting for the uncertainty in the species tree topology.

Correspondence: Simon Joly, Fax: +1 514-343-2288;
E-mail: simon.joly@umontreal.ca

Formal description of the test

In JML, posterior predictive checking is used to test for the presence of hybridization. The software uses as input a posterior distribution of species trees (S) with branch lengths (l) and population sizes (θ). This posterior distribution is generally defined as

$$P(S, l, \theta | D) \propto \int_G \left(\prod_{i=1}^{n} P(d_i | g_i) P(g_i | S) \right) P(S) dG.$$

D is the data that consist of n multiple sequence alignments (d_i). The equation integrates over all possible gene trees (G) for all alignments, and g_i represents one specific gene tree. $P(d_i \mid g_i)$ is the likelihood of the data given the gene tree (Felsenstein 1981), $P(g_i \mid S)$ is the multispecies coalescent (Rannala & Yang 2003; Degnan & Rosenberg 2009) and $P(S)$ is the prior on species trees.

Replicated data sets are simulated from the posterior distribution $P(S, l, \theta \mid D)$. A test quantity is then estimated on the observed data and on the simulated data sets to see how well the model is consistent with the data. This approach of posterior predictive checking is commonly used in Bayesian analyses to check the adequacy of a model (Gelman *et al.* 2004); if the test quantity estimated on the observed data departs strongly from the quantities estimated from the simulated data, then we can conclude that the model is inadequate. Here, the test quantity used is the minimum pairwise distance between sequences of two species (*minDist*), which has been shown to be a useful quantity for detecting hybridization (Joly *et al.* 2009). In the presence of hybridization, *minDist* can sometimes be much smaller than that expected in a scenario without hybridization. Suppose that *minDist(AB)* represents *minDist* between species A and B on the observed data and that $minDist(AB)^{sim}$ represents *minDist* between species A and B on simulated data. The P-value for hybridization between species A and B is

$$p = \Pr(minDist(AB) < minDist(AB)^{sim}).$$

The probability is taken over the posterior distribution of parameters S, l and θ (i.e. $P(S, l, \theta \mid D)$) and the posterior predictive distribution of $minDist(AB)^{sim}$. This probability can be approximated by simulation. If we simulate M data sets from the posterior distribution $P(S, l, \theta \mid D)$, we can calculate $minDist(AB)^{sim(m)}$ on each simulated data set m and the P-value is the proportion of these m simulations for which $minDist(AB) < minDist(AB)^{sim(m)}$. If the model is good, then $\Pr(minDist(AB) < minDist(AB)^{sim}) \approx 0.5$. On the contrary, a small P-value will indicate that the model does not fit the data well. Because a small value is characteristic of hybrid sequences in a data set, one can tentatively conclude that the inaccuracy of the model is because of the presence of hybrid sequences.

Implementation

Incorporating species tree topology uncertainty in posterior predictive checking represents an improvement compared to the original description of the method where the species tree topology was fixed (Joly *et al.* 2009). This is performed using as input the posterior distribution obtained from *BEAST analyses (Drummond & Rambaut 2007; Heled & Drummond 2010). *BEAST is a Bayesian method that estimates the posterior distribution of species trees, branch lengths and population sizes using sequence information from multiple genes. Note that posterior distributions from other programmes could also be used in JML as long as the tree file is in the same format as the *BEAST nexus format. For the simulations, species trees (with branch lengths and populations sizes) are sampled from the stationary phase of the Markov Chain Monte Carlo.

For each species tree, a gene tree is then simulated using the coalescent. The code for the gene tree simulation routine was adapted from MCMCcoal (Yang 2007). The number of gene copies simulated per species is the same as in the original data set. The user can scale the species tree population sizes using a heredity scalar to reflect the effective population size of the marker being simulated. Similarly, the mutation rate of the species tree can also be scaled for the simulations to allow the possibility that the mutation rate of the marker being simulated is not the same as the mutation rate implied in the species tree.

Sequences are then simulated on the gene tree. This was implemented by adapting the code of the software seq-gen 1.3.2 (Rambaut & Grassly 1997), which allows any nucleotide substitution model to be used. This procedure is repeated for all species trees of the posterior distribution (or a subset of them). Finally, JML outputs the posterior predictive distribution of the smallest distances between sequences of any two species of the data set, from which P-values could be estimated. JML can also output the exact P-value for each pairwise species comparison if the empirical sequence data set is given.

Interpretation and multiple comparisons

Different approaches can be used for interpreting results from posterior predictive checking. An intuitive one is to interpret the P-value(s) directly. The P-values estimated by JML are posterior probabilities (Gelman *et al.* 2004) and can be interpreted as the probability that the model will generate a minimum distance between sequences of two species smaller than that observed from the data, given the data. However, appealing this interpretation, it could lead to statistical issues when multiple tests are performed. Indeed, the need to correct for multiple statistical testing (Rice 1989) diminishes the likelihood of finding statistically significant results. This is especially problematic for the present application because the large variance in mutation rate for short sequences (Edwards & Beerli 2000), combined with the difficulty to get long nucleotide sequence stretches that lack the evidence of recombination in practice, results in power issues (Joly *et al.* 2009). The problem is even more acute when the approach is used in an explorative way, that is, if there are no a priori hypotheses of hybridization to test and if JML is only used to investigate the presence of hybridization in the data set. In such cases, all pairwise species comparisons can be tested simultaneously and the statistical power will be greatly affected. To minimize power issues, it could thus be important to specify hybridization hypotheses a priori without reference to the sequence data.

There is an alternative interpretation of posterior predictive checking, which is to see 'how particular aspects of the data would be expected to appear in replications' (Gelman *et al.* 2004). For instance, we could evaluate the overall adequacy of a model by assessing whether there are some aspects of the data that are not well predicted by the model. To do this, it would be of interest to report all observed distances that have a low probability of being observed, e.g. distances with $P < 0.1$ (this value is arbitrary and can be fixed by the user). This could indicate species comparisons where the model cannot adequately predict the observed minimum distances. If there were several of those instances, one could thus conclude that a strictly bifurcating species tree model is not adequate, probably because of the presence of hybridization. Note, however, that this is not the same as concluding that there has been hybridization between two given species. With such interpretations of posterior predictive distributions, the type I error is less of a concern because we use posterior predictive checking to evaluate the fit of the model rather than to test a specific hypothesis (Gelman *et al.* 2004).

Regardless of the multiple comparison issues associated with posterior predictive checking, there are two points that should always be kept in mind when interpreting results from JML. First, posterior predictive checking is a test of the model and not of hybridization. If one rejects the model (bifurcating species tree without gene flow), this may well be because of the presence of hybridization, although it could also be due to other properties of the data such as undetected gene duplication (Maddison 1997), population substructure along the branches of the phylogeny (Machado *et al.* 2002) and parallel evolution (Joly *et al.* 2010). The second point to take into account is that a lack of evidence for hybridization with JML should not be interpreted as an absolute absence of hybridization in the data set because (i) a lack of statistical significance can also be caused by a lack of data and (ii) not all hybridization events leave a detectable molecular signature (Joly *et al.* 2006, 2009).

Performance

Thorough simulations regarding the performance of the test statistic have already been conducted for several parameters such as sequence length, population size, speciation time and time of the hybridization event (Joly *et al.* 2009). Here, I report results on the impact of different parameter values on computing time. The parameters investigated were the number of species (5, 10, 15), the number of sequences per species (5, 10, 15), the number of simulations (1000, 2000, 4000) and the sequence length (500, 1000, 1500). Random species trees were simulated under a birth and death model with the R package 'geiger' (Harmon *et al.* 2008); the birth and death parameters were set to 0.00025 and 0.000125, respectively, and the phylogeny was evolved for 0.01 units of time. These settings resulted in phylogenies with a tree depth (time × mutation rate) similar to that of empirical data sets (Joly *et al.* 2009). The first phylogenies obtained with five, ten and fifteen extant species were retained for the simulations (extinct species were pruned from the tree). Mutational population sizes ($\theta = 4N_e\mu$) for the branches of the tree were generated randomly by sampling from a truncated normal distribution with mean and standard deviation of 0.005, with a lower cut-off of 0.0001. Again, this is similar to empirical observations. These phylogenies were treated as 'fixed' and JML generated simulated data sets (using the GTR + I + Γ substitution model) using combinations of the parameters mentioned earlier. Because repeated runs had very small coefficients of variation (0.5%), only one full run was performed for each combination of parameters. Simulations were performed on a HP desktop computer with an Intel core2 duo CPU at 2.33 GHz with 2 Gb of RAM.

The results show that the computing time for a complete run grows linearly with the number of data sets simulated (data not shown) and with the sequence length (Fig. 1a). In contrast, the computing time increases according to a power function relative to the number of species and relative to the number of sequences per species (Fig. 1b).

An application example—North American roses

To give an application example of the software, I reanalyse here sequence data from three nuclear genes for the native diploid roses of North America. Three nuclear

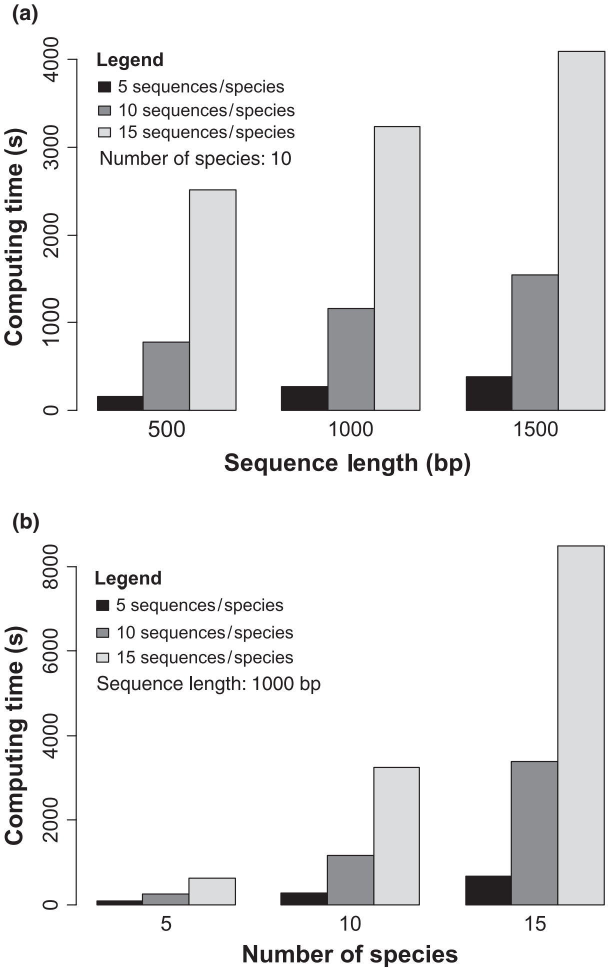

Fig. 1 Performance of the JML software in terms of computing time for (a) different sequence lengths and number of sequences per species, keeping the number of species to 10, and for (b) different number of species and sequences per species, keeping the sequence length to 1000 bp.

genes (*GAPDH*, *TPI* and *MS*) have been sequenced for 46 individuals from eight species and have been analysed with distances and gene tree parsimony approaches (Joly & Bruneau 2006, 2009). Alleles within individuals were obtained through direct sequencing or via cloning when an individual was heterozygous for a gene (Joly & Bruneau 2006). Previous studies showed that there might be introgressed sequences in the data set, i.e. some sequences in one species are often either identical or one mutation away from a sequence of another species (Joly & Bruneau 2006). Yet, no formal tests of hybridization have been conducted to date.

Previous studies could not find evidence of recombination in these data sets (Joly & Bruneau 2006), and thus, the three genes could be analysed integrally. Species tree analyses were performed in *BEAST. The nucleotide substitution model used was the one that received the highest Akaike Information Criteria (AIC) score in MODELTEST 3.7 (Posada & Crandall 1998) when fitted on a maximum likelihood tree obtained from five independent searches in Garli 1.0 (Zwickl 2006) with a GTR + $I + \Gamma$ substitution model. A strict clock was used for all genes; the rate of the *GAPDH* gene was set to 1, and the rate of the other genes was estimated relative to *GAPDH*. Population sizes were modelled as constant along branches. More details on parameters and priors can be found in Data S1 (Supporting information). The analysis was run for 10^7 generations, recording the trees and parameters every 10^4 generations, and the first million generations were discarded as burnin. Independent runs converged on the same parameter values and species tree topologies.

The species tree obtained with *BEAST (Fig. 2) was identical to one of the two most parsimonious species trees obtained by gene tree parsimony (Joly & Bruneau 2009). The branch support was relatively high for most nodes, but there is nevertheless clearly some uncertainty in the tree topology which was clearly worth accounting for in the hybridization tests. The wide branches along the backbone of the tree are likely the results of gene tree incongruence, which could be caused by either incomplete lineage sorting or hybridization.

The species trees (with branch length and population sizes) estimated by *BEAST were then input into JML and posterior predictive distributions generated for *minDist* between all species for all genes. For each gene,

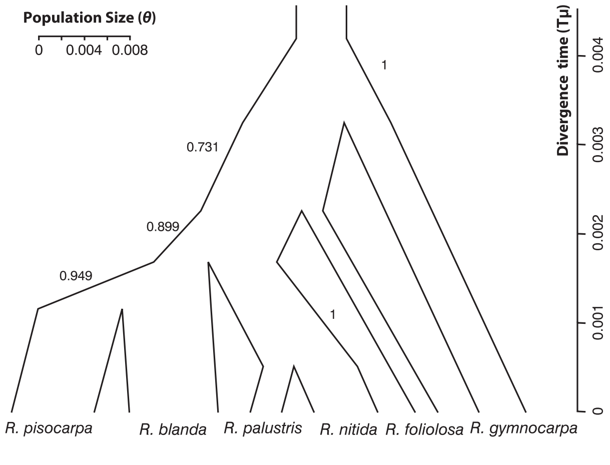

Fig. 2 Species tree of diploid North American roses obtained with *BEAST. The branch widths are proportional to the estimated population sizes, and the branch lengths are proportional to their divergence times (both median estimates). The variations in population sizes along the branches are a consequence of the graphical representation; population sizes were constant along branches, and the correct population sizes are those at the beginning of the branches. Numbers besides branches represent the posterior probabilities of the groups. The outgroup (*Rosa setigera* and *Rosa multiflora*) is not shown.

Table 1 List of distances with P-values < 0.1 according to the posterior predictive distributions

Gene	Individual 1	Individual 2	Obs. distance	*P*-value
TPI	*Rosa pisocarpa* 847	*Rosa gymnocarpa* 543	0	0.0529
TPI	*R. pisocarpa* 847	*R. gymnocarpa* 751	0	0.0529
TPI	*R. pisocarpa* 847	*R. gymnocarpa* 767	0	0.0529
TPI	*Rosa blanda* 741	*R. gymnocarpa* 543	0	0.0812
TPI	*R. blanda* 741	*R. gymnocarpa* 751	0	0.0812
TPI	*R. blanda* 741	*R. gymnocarpa* 767	0	0.0812

The number designing the individual is the accession number. See Joly *et al.* (2006) for more details on accessions.

sequences of the same length as the original ones were simulated according to the best substitution model and parameter values as determined by the AIC in ModelTest. The relative mutation rate used in the simulations for each gene was set to the median posterior value obtained from the *BEAST analyses. The species trees from the first million generations were discarded as burnin in JML, and the remaining 9000 trees were used for the simulations. Because I did not have a specific hypothesis of hybridization to test, I decided to investigate the overall fit of the model and report all observed distances that had a probability < 0.1 of being generated by the posterior distribution.

Six distances between alleles were smaller than the 10th quantile in the posterior predictive distributions (Table 1). These involved one individual of *Rosa blanda* (incl. *Rosa woodsii*) and one of *Rosa pisocarpa*, each with three individuals of *Rosa gymnocarpa* for the *TPI* gene. Although the observed distances are not statistically significant at the 5% level, they are small enough to suggest that the model does not explain these observations very well. In other words, although there is not statistical evidence for a hybridization event between *R. gymnocarpa* and *R. blanda/R. pisocarpa*, the data suggest this could be the case. Hybridization could have occurred in different ways, but most likely towards *R. gymnocarpa* given that *R. gymnocarpa* sequences are nested with a *R. blanda/ R. pisocarpa* clade (see Fig. S1, Supporting information), whereas the species trees suggest *R. gymnocarpa* is basal to the other species (Fig. 2). Because both *R. blanda* and *R. pisocarpa* share the introgressed allele, the hybridization event could have occurred between either of these species and *R. gymnocarpa* or between the ancestor of *R. blanda* and *R. pisocarpa* and *R. gymnocarpa*. More data are needed to confirm these hypotheses. For instance, the addition of genes might help to narrow down the confidence intervals of the species tree and perhaps provide stronger statistical results in the future.

One interesting observation from this example is that although there were several cases of shared alleles between species (*Rosa nitida* and *Rosa palustris* (*TPI*, *MS*, *GAPDH*), *R. pisocarpa* and *R. blanda* (*TPI*, *MS*, *GAPDH*), *R. blanda* and *Rosa foliolosa* (*MS*), *R. blanda* and *R. nitida* (*TPI*); see Fig. S1, Supporting information), none of these were found to be significant. In other words, even relatively good evidence for the presence of hybridization such as identical sequences between nonsister species does not mean that it is necessarily caused by hybridization. Owing to stochasticity in the coalescent process and in the mutation rates for short sequences, it is relatively difficult to statistically infer hybridization events from empirical data. In the present example, only one possible instance of hybridization was confirmed. In this case, identical sequences were found in a putative hybrid formed between two of the most diverged species in the group.

This application example shows why it is important to test hybridization hypotheses. Lack of significance could mean that hybridization is not responsible for the observed pattern, but it could also stimulate the gathering of additional data to eventually obtain statistical support for hybridization hypotheses. The statistical approach implemented in JML should thus help researchers to attain a better knowledge regarding the presence of hybridization in their study groups and hopefully contribute to better understand the contribution of hybridization to evolution.

Availability

JML is written in C++ and is released under the GNU General Public License 3+. Source code and precompiled binaries can be downloaded from http://www.plantevolution.org/jml.html. The manual of JML version 1.0 is available as Appendix S1 (Supporting information).

Acknowledgements

I want to thank colleagues that showed interest in the original method and that motivated the development of JML. I am also thankful for the useful comments provided by Peter Lockhart and three anonymous reviewers. This work was supported by an NSERC discovery grant.

References

Arnold ML (1997) *Natural Hybridization and Evolution*. Oxford University Press, New York.

Arnold ML (2004) Transfer and origin of adaptations through natural hybridization: were Anderson and Stebbins right? *The Plant Cell*, **16**, 562–570.

Barton NH (2001) The role of hybridization in evolution. *Molecular Ecology*, **10**, 551–568.

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

Drummond AJ, Rambaut A (2007) BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evolutionary Biology*, **7**, 214.

Edwards SV, Beerli P (2000) Gene divergence, population divergence, and the variance in coalescence time in phylogeographic studies. *Evolution*, **54**, 1839–1854.

Felsenstein J (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution*, **17**, 368–376.

Funk DJ, Omland KE (2003) Species-level paraphyly and polyphyly: frequency, causes, and consequences, with insights from animal mitochondrial DNA. *Annual Reviews in Ecology, Evolution, and Systematics*, **34**, 397–423.

Gelman A, Carlin JB, Stern HS, Rubin DB (2004) *Bayesian Data Analysis*. Chapman & Hall, Boca Raton, FL.

Harmon LJ, Weir JT, Brock CD, Glor RE, Challenger W (2008) GEIGER: investigating evolutionary radiations. *Bioinformatics*, **24**, 129–131.

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

Joly S, Bruneau A (2006) Incorporating allelic variation for reconstructing the evolutionary history of organisms from multiple genes: an example from *Rosa* in North America. *Systematic Biology*, **55**, 623–636.

Joly S, Bruneau A (2009) Measuring branch support in species trees obtained by gene tree parsimony. *Systematic Biology*, **58**, 100–113.

Joly S, Schoen DJ (2011) Migration rates, frequency-dependent selection and the self-incompatibility locus in *Leavenworthia* (Brassicaceae). *Evolution*, **65**, 2357–2369.

Joly S, Starr JR, Lewis WH, Bruneau A (2006) Polyploid and hybrid evolution in roses east of the Rocky Mountains. *American Journal of Botany*, **93**, 412–425.

Joly S, McLenachan PA, Lockhart PJ (2009) A statistical approach for distinguishing hybridization and incomplete lineage sorting. *The American Naturalist*, **174**, e54–e70.

Joly S, Pfeil BE, Oxelman B, McLenachan PA, Lockhart PJ (2010) Correction. *The American Naturalist*, **175**, 621–622.

Machado CA, Kliman RM, Markert JA, Hey J (2002) Inferring the history of speciation from multilocus DNA sequence data: the case of *Drosophila pseudoobscura* and close relatives. *Molecular Biology and Evolution*, **19**, 472–488.

Maddison WP (1997) Gene trees in species trees. *Systematic Biology*, **46**, 523–536.

Mallet J (2007) Hybrid speciation. *Nature*, **446**, 279–283.

Posada D, Crandall KA (1998) Modeltest: testing the model of DNA substitution. *Bioinformatics*, **14**, 817–818.

Rambaut A, Grassly NC (1997) Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Computer Applications in the Biosciences*, **13**, 235–238.

Rannala B, Yang Z (2003) Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. *Genetics*, **164**, 1645–1656.

Rice WR (1989) Analysing tables of statistical tests. *Evolution*, **43**, 223–225.

Rieseberg LH (1997) Hybrid origins of plant species. *Annual Reviews in Ecology and Systematics*, **28**, 359–389.

Rieseberg LH, Raymond O, Rosenthal DM *et al.* (2003) Major ecological transitions in wild sunflowers facilitated by hybridization. *Science*, **301**, 1211–1216.

Seehausen O (2004) Hybridization and adaptive radiation. *Trends in Ecology and Evolution*, **19**, 198–207.

Yang Z (2007) *MCMCcoal: Markov Chain Monte Carlo Coalescent Program*. University College London, London.

Zwickl DJ (2006) Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion. Ph.D. dissertation, The University of Texas at Austin.

Supporting Information

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

Appendix S1 JML, version 1.0.

Data S1 xml file used for the *BEAST analyses of the North American roses.

Fig. S1 Individual gene trees of the North American roses.

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting information supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.