

Accepted Manuscript

Does gene flow destroy phylogenetic signal? The performance of three methods for estimating species phylogenies in the presence of gene flow

Andrew J. Eckert, Bryan C. Carstens

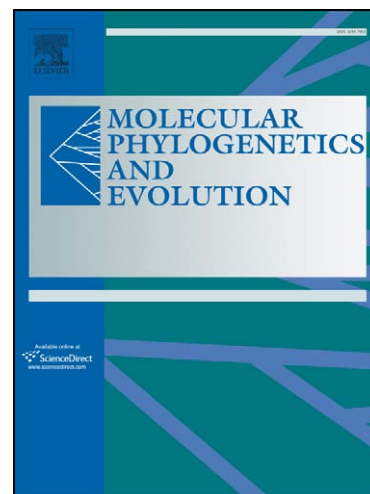
PII: S1055-7903(08)00456-9
DOI: [10.1016/j.ympev.2008.09.008](https://doi.org/10.1016/j.ympev.2008.09.008)
Reference: YMPEV 3016

To appear in: *Molecular Phylogenetics and Evolution*

Received Date: 21 March 2008
Revised Date: 8 September 2008
Accepted Date: 12 September 2008

Please cite this article as: Eckert, A.J., Carstens, B.C., Does gene flow destroy phylogenetic signal? The performance of three methods for estimating species phylogenies in the presence of gene flow, *Molecular Phylogenetics and Evolution* (2008), doi: [10.1016/j.ympev.2008.09.008](https://doi.org/10.1016/j.ympev.2008.09.008)

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.



1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29

Running Header: Eckert and Carstens: Gene flow and estimating species phylogenies

**Does gene flow destroy phylogenetic signal? The performance of three methods
for estimating species phylogenies in the presence of gene flow**

Andrew J. Eckert¹ and Bryan C. Carstens²

¹Section of Evolution and Ecology, University of California at Davis, One Shields
Avenue, Davis, CA 95616

²Department of Biological Sciences, 202 Life Sciences Building, Louisiana State
University, Baton Rouge, LA 70803

Correspondence:

Bryan C. Carstens

Department of Biological Sciences
202 Life Sciences Building
Louisiana State University
Baton Rouge, LA 70803
e-mail: carstens@lsu.edu
phone: (225) 578-0960

1 Abstract

2 Incomplete lineage sorting has been documented across a diverse set of taxa ranging
3 from song birds to conifers. Such patterns are expected theoretically for species
4 characterized by certain life history characteristics (e.g. long generation times) and
5 those influenced by certain historical demographic events (e.g. recent divergences). A
6 number of methods to estimate the underlying species phylogeny from a set of gene
7 trees have been proposed and shown to be effective when incomplete lineage sorting
8 has occurred. The further effects of gene flow on those methods, however, remain to be
9 investigated. Here, we focus on the performance of three methods of species tree
10 inference, ESP-COAL, minimizing deep coalescence (MDC), and concatenation, when
11 incomplete lineage sorting and gene flow jointly confound the relationship between gene
12 and species trees. Performance was investigated using Monte Carlo coalescent
13 simulations under four models (*n*-island, stepping stone, parapatric, and allopatric) and
14 three magnitudes of gene flow ($N_e m = 0.01, 0.10, 1.00$). Although results varied by the
15 model and magnitude of gene flow, methods incorporating aspects of the coalescent
16 process (ESP-COAL and MDC) performed well, with probabilities of identifying the
17 correct species tree topology typically increasing to greater than 0.75 when five more
18 loci are sampled. The only exceptions to that pattern included gene flow at moderate to
19 high magnitudes under the *n*-island and stepping stone models. Concatenation
20 performs poorly relative to the other methods. We extend these results to a discussion
21 of the importance of species and population phylogenies to the fields of molecular
22 systematics and phylogeography using an empirical example from *Rhododendron*.

1 **Key words:** COAL; gene flow; estimating species phylogenies; incomplete lineage
2 sorting; minimizing deep coalescence; *Rhododendron*

3

4 **1. Introduction**

5 The fundamental goal of systematics is to understand the process of lineage
6 divergence that leads to the formation of new species. Since Maddison (1997) there has
7 been growing acceptance among systematists that gene genealogies are not always
8 congruent with species phylogenies (e.g. the actual pattern of lineage splitting and
9 descent from common ancestors). It is now widely recognized that processes such as
10 gene duplication (Fitch, 1970), lateral transfer (Cummings, 1994) and incomplete
11 lineage sorting (Tajima, 1983; Takahata and Nei, 1985; Hudson, 1992) can lead to
12 incongruence between gene trees and species trees, and empirical examples of each
13 process exist (cf. Syring et al., 2007 for an example of incomplete lineage sorting). This
14 realization has prompted the development of approaches designed to estimate species
15 phylogenies despite the process that presumably caused the incongruence. For
16 example, gene tree parsimony (Slowinski and Page, 1999) was developed to account
17 for gene duplication, while the minimization of deep coalescence (MDC; Maddison,
18 1997), COAL (Degnan and Salter, 2005), and BEST (Edwards et al., 2007; Liu and
19 Pearl, 2007) were designed in part to estimate species phylogeny when the discord
20 between the gene trees and species tree is a result of the incomplete sorting of
21 ancestral polymorphisms.

22 At the initial stages of divergence, incomplete lineage sorting is ubiquitous and
23 likely produces the majority of gene-species tree discord among closely related

1 lineages. This is a direct outcome of population-level processes; consequently, the
2 developers of methods have incorporated statistical models derived from the coalescent
3 (Kingman, 1982; Hudson, 1990) into species-level phylogenetic analyses to account for
4 these processes. However, for many empirical systems it is also these lineages that
5 exchange migrants, particularly when they occur in sympatry. Since genetic
6 polymorphism shared among lineages can result from either retained ancestral
7 polymorphism or a gene copy introduced into the population via gene flow (Slatkin and
8 Maddison, 1987), it is often difficult to determine which process produced the shared
9 polymorphism. Fully statistical treatments of coalescence, gene flow, and divergence
10 are currently available only for pairwise comparisons between two lineages (Nielsen &
11 Wakeley, 2001; Hey and Nielsen, 2004, 2007; Hey, 2006).

12 It is an understatement to suggest that the biologist who wishes to estimate
13 species phylogeny in a system where details such as (a) the number of lineages, (b) the
14 relationship among lineages, and (c) the amount of gene flow are unclear is currently
15 faced with a difficult task. Methods that estimate a species phylogeny using some
16 approach derived from the coalescent must be robust to at least moderate levels of
17 gene flow (e.g. levels that not be easily recognizable) to be of any use to the majority of
18 empirical biologists, or the use of such methods may result in spurious conclusions
19 about the actual pattern of lineage divergence. The data we present in this manuscript
20 were collected out of a desire to explore how the phylogenetic signal contained in DNA
21 sequence data is affected by gene flow in recently diverged lineages. Does gene flow
22 destroy phylogenetic signal entirely, or are some methods able to accurately estimate
23 species phylogeny when some of the shared polymorphisms result from gene flow? In

1 order to explore this issue, we explore approaches based on the coalescent that use
2 estimated gene trees as input in an attempt to isolate gene flow as the sole factor
3 affecting phylogenetic accuracy.

4

5 **2. Materials and Methods**

6

7 *2.1 Statistical inference of species trees from gene trees*

8

9 A renewed interest exists in the development and interpretation of statistical
10 methods for the inference of species trees from gene trees (Maddison and Knowles,
11 2006). A myriad of innovative approaches have been developed (Slatkin and Maddison,
12 1989; Maddison, 1997; Page and Charleston, 1997; Slowinski and Page, 1999; Liu and
13 Pearl, 2006; Edwards et al., 2007; Carstens and Knowles, 2007), as well as applied to
14 empirical questions in phylogeography and systematics (Knowles and Carstens 2007;
15 Brumfield et al., 2008; Carling and Brumfield, 2008). Here, we focus on two methods for
16 estimating species phylogenies at relatively low levels of lineage divergence. The first
17 seeks to identify the species tree that maximizes the probability of a set of genealogies
18 given the species tree (Maddison, 1997), as implemented in COAL (Degnan and Salter,
19 2005) and as applied by Carstens and Knowles (2007). The second method, described
20 by Maddison (1997) and implemented in the MESQUITE software package (Maddison
21 and Maddison, 2004) minimizes the amount of deep coalescence to estimate the
22 species phylogeny. Hereafter we refer to these approaches as ESP-COAL and MDC,
23 respectively.

1 ESP-COAL is a maximum-likelihood approach to the inference of species trees
2 from a set of gene trees. Maddison (1997) noted that the likelihood (L) of a species tree
3 inferred from n independent gene loci could be written as:

$$4 \quad L(D | ST) = \prod_{n=1}^n \left(\sum_{GT} [\Pr(D | GT) \Pr(GT | ST)] \right) \quad (\text{eqn. 1})$$

5 where D are the sequence data, ST is the species tree, and GT is the gene tree. Note
6 that the summation is over all possible GT for each of the n loci. The first expression of
7 the inner product is the likelihood of the data given a gene tree, which can be computed
8 by standard phylogenetic software. The second expression of the inner product
9 represents the probability of a gene tree given a species tree. This quantity can only be
10 calculated for some sample configurations using the mathematical theory for the neutral
11 coalescent (Tajima, 1983, 1989; Hudson, 1983; Takahata, 1989; Rosenberg, 2002;
12 Yang, 2002; Wall, 2003). Degnan and Salter (2005), however, devised a combinatoric
13 approach for the calculation of this probability, with the limitation that it results in the
14 probability of the gene tree topology, not considering branch lengths, conditional on a
15 species tree topology with known branch lengths. Rigorous maximization of the
16 likelihood function would require joint searches through the state space of all possible
17 gene and species tree topologies and their branch lengths using some form of
18 importance sampling (Felsenstein, 2004). In order to approximate the maximization of
19 the likelihood function as defined above, we followed the method of Maddison (1997)
20 and Carstens and Knowles (2007), which searches for the ST topology conferring the
21 highest probability for the observed gene trees.

22 The second approach to estimating species phylogeny (MDC), also described by
23 Maddison (1997), uses a heuristic search to identify the species phylogeny that

1 minimizes the amount of deep coalescence (e.g. incomplete lineage sorting). This
2 approach can be accurate in the absence of gene flow under certain assumptions
3 concerning the species tree topology (Degnan and Rosenberg, 2006; Maddison and
4 Knowles, 2006), but has not been explicitly explored given varying levels of gene flow.
5 Like the ESP-COAL approach, it evaluates the pattern of coalescence without
6 considering the branch lengths of the genealogies.

8 *2.2 Parameters and models of gene flow*

9
10 ESP-COAL is accurate when ancestral polymorphisms are segregating within
11 species that otherwise conform to a bifurcating phylogenetic tree (Carstens and
12 Knowles, 2007), particularly when the depth of the species tree is $3N_e$ or greater.
13 However, the signature of ancestral polymorphisms segregating within descendant
14 lineages due strictly to genetic drift is complicated when gene flow, either recent or
15 historical, has occurred among lineages (Slatkin and Maddison, 1987).

16 We devised four basic models of gene flow in order to elucidate the effects of this
17 process on phylogenetic analyses: n -island, stepping stone, and two models of
18 historical gene flow (Fig. 1). The models of historical gene flow were formulated to
19 reflect scenarios of either allopatric or parapatric speciation. Historical gene flow
20 occurred strictly between sister lineages and was modeled as a burst of gene flow
21 directly after (parapatric) or $0.5xN_e$ generations after speciation (allopatric), where x is
22 the length of time between successive speciation events (Fig. 1). The duration of these
23 bursts was controlled by the parameter d , and we incorporated a relatively short period

1 of divergence with gene flow ($0.1N_e$) as well as a longer period ($0.5N_e$). For each model,
2 we assumed three different magnitudes of gene flow as measured by the effective
3 number of migrants per generation ($N_e m = 0.01, 0.10, \text{ or } 1.00$).

4 As shown previously, the power of the ESP-COAL and MDC depend upon the
5 number of unlinked loci used in the analysis and the depth of the species tree
6 (Maddison and Knowles, 2006; Carstens and Knowles, 2007). Therefore, we varied the
7 number of sampled loci from two to ten and the depth of the species tree ($2N_e$ or $6N_e$),
8 as well as the effective population sizes ($N_e = 10,000$ or $100,000$). These parameter
9 treatments were considered in a fully factorial design, yielding 72 different treatment
10 combinations, for each of which we analyzed the accuracy of the ESP-COAL and MDC
11 across samples of two to ten loci (Table S1; online supplemental data).

13 *2.3 Canonical species tree*

14
15 We assumed a single, fully resolved, pectinate species tree with four taxa for our
16 simulations [$((c:0.75,(a:0.375,b:0.375):0.375):0.25,d:1.00)$]. The fourth taxon (taxon d)
17 was designated as the outgroup. The ingroup taxa can then be characterized by three
18 possible rooted tree topologies. In all cases, the relative branch lengths conform to a
19 molecular clock and were defined as pictured in Fig. 1.

21 *2.4 Monte Carlo coalescent simulations*

22

1 Monte Carlo coalescent simulations were used to generate simulated
2 genealogies under each treatment. For each of the parameter combinations, we
3 simulated 500 genealogies consisting of five gene sequences sampled from each of
4 four species in the canonical species tree. Using those simulated genealogies, we
5 generated data sets consisting of two to ten loci selected at random without
6 replacement. The genealogies within these data sets were assumed to be estimated
7 without error, thus negating the need to calculate the quantity $\Pr(D|GT)$ in ESP-COAL.
8 We also assumed that all lineages were consistent and equal in their effective
9 population size. Furthermore, at each speciation event, daughter lineages were
10 assumed to instantaneously grow to the size of the ancestral population. For example,
11 an ancestral population with an N_e of 10,000 was assumed to speciate into two
12 daughter lineages each with an N_e of 10,000. While this is clearly a simplified model of
13 the process of lineage divergence, our aim here is to explore the effects of gene flow,
14 and only gene flow, on these methods. All coalescent simulations were carried out using
15 SIMCOAL v. 2.0 (Excoffier et al., 2000).

16

17 *2.5 Performance of ESP-COAL, MDC, and concatenation in estimating species* 18 *phylogeny*

19

20 We followed the approach of Carstens and Knowles (2007) to explore the power
21 and accuracy of ESP-COAL. Briefly, this method involves five steps – (1) estimation of
22 the $\Pr(GT/ST)$ using COAL for each of the 500 simulated gene trees, (2) sampling
23 between two and ten loci, (3) calculating the sum of the probability of the gene trees

1 given the species tree for each possible species tree, (4) performing approximate
2 likelihood ratio tests (LRTs) to evaluate the significance of the correct species topology
3 versus the two incorrect topologies (Anisimova and Gascuel, 2006), and (5) counting
4 the number of times the LRTs identified the correct species topology out of the
5 replicated simulations. This process was repeated for each of the 72 parameter
6 combinations.

7 We used a similar approach to explore the performance of the MDC method
8 given varying types and amounts of gene flow. For each of the 72 treatments, we
9 randomly selected between two and ten genealogies and used MESQUITE ver. 2.5
10 (Maddison and Maddison, 2004) to estimate the species phylogeny. The genealogies
11 were treated as unrooted, and a heuristic search with nearest-neighbor interchange
12 branch swapping and max-trees set to 100 was used to explore species treespace. As
13 above, we used the actual simulated genealogies in the searches.

14 Lastly, we analyzed the 72 treatments by concatenating data from between two
15 and ten loci and estimating phylogeny using maximum-likelihood (ML). Sequence data
16 were simulated using SIMCOAL under an HKY model of sequence evolution with $ts/tv =$
17 3.0 and simulations were conditioned on an expectation of $\theta = 20$, which resulted in
18 between 40 and 60 segregating sites in each data simulated data set. For each
19 concatenated data set, PAUP* (Swofford, 2002) was used to estimate the phylogeny
20 using ML with a heuristic search, maxtrees = 10, and the HKY model of sequence
21 evolution used to simulate the data. A tree filter was then used to determine what
22 proportion of estimated phylogenies matched the species phylogeny that was used to
23 simulate the data. Each treatment was replicated 100 times.

1

2 2.6 An empirical example from *Rhododendron*

3

4 The genus *Rhododendron* contains approximately 1,000 species distributed
5 throughout the Northern Hemisphere (Cox and Cox, 1997). Recent phylogenetic work
6 has elucidated the placement of *Rhododendron* within the Ericaceae, defined generic
7 boundaries, and revised the taxonomy within some of the largest subgeneric clades
8 (Kron et al., 2002; Milne, 2004; Goetsch et al., 2005). At the species level, however,
9 many relationships remain ambiguous, even when phylogenies are inferred from
10 multilocus nuclear markers. This ambiguity is likely caused by a combination of
11 historical gene flow and incomplete lineage sorting when the species under
12 consideration are or were geographically proximal.

13 Here, we concentrate on inferring the species phylogeny of a group of four
14 *Hymenanthes* rhododendrons (*Rhododendron macrophyllum* D. Don ex G. Don, *R.*
15 *catawbiense* Michx., *R. caucasicum* Pall., and *R. brachycarpum* D. Don ex G. Don)
16 distributed across eastern Asia, Europe, and North America using data from two RNA
17 polymerase genes (*RPB2* and *RPC1*) and two chloroplast genes (*trnK* and *trnL-trnF*).
18 Data for the chloroplast genes were obtained from Milne (2004) and were concatenated
19 due to the uniparental inheritance and lack of recombination for the chloroplast genome.
20 The sample size for each species ranged from one to five gene copies per locus. All
21 gene trees were rooted using *R. aureum* Georgi as an outgroup.

22 Estimation of the species tree topology for the four *Rhododendron* species was
23 carried out using the three methods described above. For ESP-COAL, we selected the

1 HKY model of DNA sequence evolution using DT-MODSEL (Minin et al. 2003), and
2 estimated gene trees with ML using PAUP*. Parameters of the HKY model were
3 estimated concomitantly with the gene tree topology and branch lengths. The likelihood
4 score associated with the maximum-likelihood estimate (MLE) of the gene tree was
5 equated to the $\Pr(D|GT)$. The nonparametric bootstrap (Felsenstein, 1985) with 1,000
6 replicates was used to assess the reliability of nodes within inferred genealogies as well
7 as the concatenated tree topology. The topology of the MLE gene tree was evaluated
8 on each of the 15 possible rooted species tree topologies using COAL. The species tree
9 topology conferring the largest probability of the MLE gene tree was taken as the best
10 estimate. Analyses using MDC and concatenation were conducted as described above
11 for these data with gene and species trees being rooted with *R. aureum*.

12 While our analyses of simulated data used the actual genealogies, the analysis
13 of empirical data utilized estimates of the actual *Rhododendron* genealogies for these
14 loci, and includes an additional source of statistical error associated with the estimation
15 of the gene tree. To explore the potential magnitude of this effect, we chose the
16 simulation treatments likely to mirror the evolutionary history of the four *Rhododendron*
17 species described above (parapatric model, $N_e = 10,000$, $N_e m = 0.10$, $d = 0.10$, and ST
18 depth = $2N_e$), and conducted the ESP-COAL and MDC analyses using estimated rather
19 than the actual genealogies. Data sets were simulated under these conditions based on
20 the conclusions of Milne (2004) who showed that the divergence time among the
21 species considered here ranges from one to two million years ago, or approximately
22 $2N_e$ generations. Nucleotide data were simulated under the HKY model, maximum-
23 likelihood searches in PAUP* were used to estimate the genealogies for 500 simulated

1 data sets, and two to ten gene trees were picked at random and analyzed as described
2 in section 2.5. Thus, any decrease in the performance of ESP-COAL and MDC likely
3 results from the contribution of the addition source of error in the estimated genealogies.

4

5 **3. Results**

6

7 *3.1 Performance of ESP-COAL*

8

9 The type and magnitude of gene flow affected the ability to infer the correct ST
10 topology using ESP-COAL (Fig. 2a). In general, models of historical gene flow did not
11 greatly degrade the phylogenetic accuracy, regardless of the magnitude ($N_e m = 0.01$ to
12 1.00) or duration ($0.1 \times N_e$ or $0.5 \times N_e$ generations) of gene flow. The probability of
13 identifying the correct ST never dipped below 0.70 for any parameter combination for
14 either the parapatric or allopatric models. In contrast, phylogenetic accuracy was low
15 under the n -island and stepping stone models, with the magnitude of the performance
16 drop depending more on the magnitude of gene flow than on the number of loci
17 examined. For example, the two locus data sets for the n -island model at a magnitude
18 of $N_e m = 1.00$ only identified the correct ST with probabilities of 0.00 to 0.03 depending
19 upon the value of N_e , but the phylogenetic accuracy did not improve when 10 loci were
20 used (Table S1, online supplemental data). However, at lower magnitudes of gene flow
21 (e.g. $N_e m = 0.01$), ESP-COAL retained reasonable power ($\Pr[\text{ST}_{\text{correct}}|\text{GT}] > 0.72$) to
22 identify the correct species topology, so long as more than five loci were sampled (Fig.
23 2a, Tables S1-S3, online supplemental data).

1

2 *3.2 Performance of MDC*

3

4 With one exception, phylogenetic accuracy was high in the MDC analyses across
5 all models of gene flow and parameter combinations (Tables S4 - S6, online
6 supplemental data). The lowest probability of estimating the true species tree for the
7 stepping stone, parapatric, and allopatric models of gene flow was 0.69, and values
8 were typically much higher so long as four loci were used. The performance of MDC
9 was also correlated with the assumed value of N_e , the depth of the ST, and the number
10 of sampled loci, with higher probabilities associated with larger values of each of those
11 quantities. As in the ESP-COAL analysis, the n -island model presented the greatest
12 difficulty to the analysis, although this was correlated with the strength of migration
13 (Table S4, online supplemental data). With $N_e m = 0.01$, MDC was reasonably accurate
14 (e.g. accuracy greater than 0.70) regardless of the numbers of loci. Accuracy decreased
15 with $N_e m = 0.10$, but still trended upwards as loci were added. However, when $N_e m =$
16 1.00, the probability of identifying the correct ST was never greater than 0.34 (Fig. 2b).

17

18 *3.3 Performance of concatenation*

19

20 The identification of the correct species topology when data were concatenated
21 was severely affected by both the magnitude and type of gene flow. In general,
22 concatenation performed poorly under the n -island and stepping stone models even
23 when the number of sampled loci increased (Fig. 2c). As the magnitude of gene flow

1 increased under these models, probabilities of identifying the correct ST reached zero
2 for all data sets (Tables S7 - S9, online supplemental data). For the models of historical
3 gene flow, concatenation achieved reasonable power only when the number of loci
4 sampled was greater than five, when gene flow was limited in magnitude ($N_e m < 1.00$)
5 and duration ($d = 0.10$), or when the depth of the ST increased.

6 7 3.4 A comparison among methods

8
9 Noticeable differences existed among the three methods considered for the
10 inference of the species phylogeny (Table 1; Fig. 2). Both ESP-COAL and MDC
11 outperformed concatenation, especially for the n -island and stepping stone models (Fig.
12 2). Moreover, MDC performed better than ESP-COAL for those models, while ESP-
13 COAL performed better for the allopatric and parapatric models. This difference was
14 asymmetric, with MDC exhibiting much larger increases ($\Delta \Pr[\text{ST}_{\text{correct}}|\text{GT}] = 0.10 - 0.35$)
15 relative to ESP-COAL under the n -island model than for ESP-COAL relative to MDC
16 under the allopatric and parapatric models ($\Delta \Pr[\text{ST}_{\text{correct}}|\text{GT}] = 0.05 - 0.15$). In both
17 cases, the differences between ESP-COAL and MDC narrowed as the number of
18 sampled loci increased.

19 The greatest degradation of phylogenetic signal occurred when gene flow was
20 simulated using the n -island model, and as such this scenario provides the clearest
21 illustration of the effects of varying parameters such as $N_e m$, N_e , and ST depth. For
22 example, phylogenetic accuracy decreases at a much faster rate as the amount of gene
23 flow increases when ESP-COAL is used as opposed to MDC (Fig. 3a). While effective

1 population size does not appear to have a large effect on phylogenetic accuracy (Fig.
2 3b), both ESP-COAL and MDC perform better under the n -island model when the
3 species tree depth is shallow ($2N_e$) rather than deeper ($6N_e$; Fig. 3c). This finding is in
4 contrast to the general trend demonstrated for MDC (Maddison and Knowles, 2006) and
5 ESP-COAL (Carstens and Knowles, 2007), where accuracy increases with increasing
6 species tree depth.

8 3.5 *An empirical example from Rhododendron*

9
10 Non-monophyly among the four *Rhododendron* species was apparent for both
11 RNA polymerase data sets (Figs. 4a-c), despite differences in the number of
12 polymorphisms, length of the aligned sequences, and sample size among sampled loci
13 (Table 2). Estimated genealogies for each locus were fully resolved and differed in
14 topology, with samples from *R. macrophyllum* often paraphyletic (Fig. 4). Analyses
15 using ESP-COAL and MDC identified different species tree topologies as the most likely
16 (Fig. 5a-b). However, the number of deep coalescences between the topology identified
17 by ESP-COAL and that identified using MDC differed by only a single event (17 vs. 18).
18 For the ESP-COAL analyses, the optimal tree was significantly better than the second
19 best tree when compared using an approximate likelihood ratio test ($-2\Delta\ln L = 6.03$, $df =$
20 1 , $P = 0.014$). Concatenation of the three data sets still resulted in a tree where the four
21 species were not monophyletic (Fig. 5c).

22 Do the phylogenetic accuracy values reported here (Sup. Tables 1 - 6) change
23 appreciably when genealogies are estimated from sequence data, as opposed to using

1 the actual genealogies? We reanalyzed the case thought to most closely approximate
2 the *Rhododendron* data ($N_e=10,000$, $N_e m = 0.10$, $d = 0.10$, ST depth = $2N_e$) using gene
3 trees estimated from sequence data simulated on our actual genealogies. Using this
4 approach, accuracy decreased by slightly less than 10% in the ESP-COAL analyses
5 and by slightly less than 2% in the MDC analyses (Fig. 6). We expect this decrease to
6 be influenced by such factors as the number of variable sites in each locus as well as
7 the topology of the species tree. This example illustrates one possible application of
8 power analyses to phylogenetic investigations in empirical systems.

10 **4. Discussion**

12 *4.1 Explanation of results*

14 Incomplete lineage sorting has emerged as a common problem for phylogenetic
15 inference at the species level. Given the volume of mathematical theory predicting this
16 phenomenon (cf. Pamilo and Nei, 1988; Rosenberg, 2002, 2003), this may not be
17 surprising. Several methods of inferring species phylogenies from gene trees have
18 incorporated the stochastic process of incomplete lineage sorting (Maddison, 1997;
19 Degnan and Salter, 2005; Liu and Pearl, 2007). While these methods are clearly at
20 early stages of development, they appear to perform well when incomplete lineage
21 sorting is the only process contributing to the discord between gene trees and species
22 trees (Maddison and Knowles, 2006; Carstens and Knowles, 2007; Edwards et al.,
23 2007). Here, we explore the further confounding effects of gene flow on the ability to

1 identify the correct species tree topology using three methods – ESP-COAL, MDC, and
2 concatenation. The results presented here suggest that methods derived from the
3 coalescent are robust to the confounding effects of various models of gene flow and
4 particularly to gene flow of lower magnitudes.

5 The greatest degradation of phylogenetic signal occurred under the n -island
6 model with moderate to high levels of gene flow. In the n -island model all extant
7 lineages have the same probabilities of sharing migrants per generation, and at
8 sufficiently high levels of gene flow the notion of an underlying species phylogeny
9 probably loses validity. The net effect of this process is to scramble the gene copies
10 among lineages to such a magnitude that the gene tree topology in reference to the
11 species topology becomes highly reticulate. However, at low magnitudes of gene flow
12 ($N_e m = 0.01$), there remains some phylogenetic signal that can be recovered.
13 Surprisingly, MDC performed quite well even at moderate levels of gene flow under the
14 n -island model. This observation may partially be explained by the argument that the
15 process of gene copy exchange among lineages may mirror the sorting of ancestral
16 polymorphism when gene flow is equal among demes and large in magnitude (as in our
17 simulations). Since deep coalescences become uncommon (Nielsen, 2005) as gene
18 flow increases in magnitude under the n -island model, unequal migration among demes
19 may present significant difficulties to MDC that are not predicted by our simulation
20 results. In the case where gene flow is low, deep coalescences are more common, and
21 gene copies have to wait to coalesce until they are in the same species lineage.

22 The stepping stone model presented more difficulties for ESP-COAL than MDC
23 at high levels of gene flow ($N_e m = 1.00$). ESP-COAL can be misled when stepping stone

1 migration mimics a spurious pattern of gene coalescence, especially between non-sister
2 species, in multiple loci because the probability of these gene trees would be maximized
3 under an incorrect species tree. Because MDC counts the number of deep coalescent
4 events, and uses this as an optimality criterion across all gene trees, it may be less
5 severely impacted by biological processes (like stepping stone migration) that increase
6 the discord among loci. Of course, this explanation is valid only when divergence is not
7 extremely recent or characterized by rapid radiations (cf. Degnan and Rosenberg,
8 2006).

9 Both ESP-COAL and MDC performed exceptionally well under models of
10 historical gene flow (Tables S2 – S3, S5 – S6, online supplemental data). In these
11 models, factors such as the magnitude of the gene flow, the depth of divergence, and
12 the effective population size exert less influence on phylogenetic accuracy than they do
13 in the n -island or stepping stone models. In taxa where divergence occurs in parapatry,
14 or where secondary contact has occurred, this suggests that methods for estimating
15 species phylogeny from gene trees will be able to accurately estimate the species
16 phylogeny even at the initial stages of divergence and at relatively high levels of
17 migration. Concatenation, on the other hand, performed poorly for most models, even
18 as the number of sampled loci increased. This may partially be an artifact of the
19 simulation framework since concatenation analyses were confounded with the
20 estimation of the true genealogy, as well as the fact that gene flow is a genome-wide
21 phenomenon so that additional loci do not necessarily clarify phylogenetic relationships.

22 The optimism gleaned from these results, however, needs to be hedged with the
23 realization that recent theoretical and simulation studies have shown that certain gene

1 and species tree characteristics can lead to positively misleading results. For example,
2 Degnan and Rosenberg (2006) prove that for species trees with ≥ 5 tips there always
3 exists a region of the branch length parameter space which guarantees that the most
4 likely gene tree will not match the true underlying species tree. These optimal gene
5 trees that differ from the underlying species tree were coined as anomalous gene trees
6 (AGTs) and were shown to occur most often when deep internal branches within the
7 true species tree were extremely short. For species trees with four tips and asymmetric
8 topologies, as presented here (cf. Fig. 1), AGTs occur when the two internal branch
9 lengths are approximately less than 0.156 coalescent time units or in the case of
10 populations consistent with Wright-Fisher mating, $0.156N_e$ generations. The relative
11 branch lengths used in the true species tree for our simulations, however, did not fall
12 below this threshold (Fig. 1). Similarly, Kubatko and Degnan (2007) show that
13 concatenated data sets suffer from similar problems of convergence to the incorrect
14 species topology when single individuals are sampled and internal branches are short
15 relative to external branches on the underlying species topology. The direct
16 ramifications of the aforementioned studies is that methodologies relying on gene-
17 species tree discordance (e.g. MDC) can quickly become misleading when divergence
18 is recent or species radiations are rapid. It would be interesting in future work, therefore,
19 to examine the effect of gene flow within this anomaly zone, because it is precisely
20 those populations which diverged recently that are likely to share migrants.

21

22 *4.2 Implications for empirical studies*

23

1 The results presented here are of interest to two broad groups of molecular
2 systematists; those who construct large phylogenies that include clades of closely-
3 related species, and those who conduct phylogeographic investigations. For the former,
4 concatenation across loci may lead to statistical error in tip clades where the species
5 from which the exemplars are sampled have a history of gene flow. While the degree to
6 which error in the tip clades contributes to error at deeper nodes is unknown, it may be
7 prudent to estimate species phylogeny for the tip clades using an approach such as
8 MDC or ESP-COAL, and then to conduct the broader analysis using this estimate of
9 species phylogeny as a constraint. Alternately, systematists could first conduct the
10 broad analysis, and then double-check the results for poorly supported tip clades using
11 one of these approaches.

12 Phylogeography studies have long assumed that the population structure implied
13 by a single locus (commonly mitochondrial or chloroplast DNA in animals or plants,
14 respectively) can be used as a proxy for the actual population structure. Since Avise
15 (2000) and Knowles and Maddison (2001), an influx of methodological approaches
16 derived from the coalescent have been incorporated into phylogeographic investigations
17 (Kuhner et al., 1998; Beerli and Felsenstein, 1999, 2001; Pritchard et al., 2000; Nielsen
18 and Wakeley, 2001; Ranala and Yang, 2003; Hey and Nielsen, 2004, 2007; Kuhner,
19 2006; Hickerson et al., 2007). While the main objective of those methodologies is to
20 estimate the parameters of interest using gene trees as a nuisance parameter, many
21 phylogeographers would benefit directly from the ability to estimate robust species or
22 population phylogenies. Our results suggest that, even at low levels of divergence,
23 direct estimates the population phylogeny can be accurate given a modest number of

1 loci (but see Degnan and Rosenberg [2006]). A model of the population phylogeny can
2 be used as the basis for other analyses, particularly those that aim to test hypotheses
3 pertaining to population demography using parametric bootstrapping (Knowles and
4 Carstens, 2007).

5 Extending the utility of these methods will also aid in the inference of species
6 relationships for many systematic and evolutionary questions across a broad set of
7 taxa. This is true especially for organisms with large generation times, effective
8 population sizes, and recent divergences where incomplete lineage sorting has been
9 shown to be common (Syring et al., 2007). By using tree inference methods that
10 incorporate incomplete lineage sorting into their formulation, rather than trying to
11 diagnose the fact that incomplete lineage sorting has occurred using standard
12 phylogenetic methods, (e.g. maximum parsimony) which inherently assume an
13 unknown bi- or multi-furcating tree without reticulations, systematists can formulate and
14 test biogeographic and evolutionary hypotheses previously unable to be addressed due
15 to the lack of a reliable species topology.

16 We illustrate this change in focus by analyzing intraspecific DNA sequence data
17 from two RNA polymerase and two chloroplast gene regions obtained from four
18 *Rhododendron* species. The relationships among these species changes depending
19 upon the gene region analyzed and whether the data are concatenated or collapsed into
20 one consensus sequence per species (cf. Milne, 2004; Goetsch et al., 2005). In the best
21 species tree identified using ESP-COAL, *R. caucasicum* is sister to the remaining three
22 taxa, with *R. cawtabiense* being placed as sister to the *R. macrophyllum* and *R.*
23 *brachycarpum* clade (Fig. 5a). Alternatively, the topology requiring the fewest number of

1 deep coalescences places *R. macrophyllum* as sister to *R. brachycarpum* which
2 together are sister to a clade composed of *R. catawbiense* and *R. caucasicum*. As
3 pointed out previously, however, this topology required only one less deep coalescence
4 as opposed to the topology identified by ESP-COAL (Fig. 5). It is important to note,
5 moreover, that both of these topologies differ from those estimated by Milne (2004) and
6 Goetsch et al. (2005) suggesting that incomplete lineage sorting and gene flow may
7 complicate inference of shallow phylogenetic structure within this clade. These
8 topologies also provide scaffolds from which to formulate and test biogeographic (Ree
9 and Smith, 2008) and divergence time (Sanderson, 2002) hypotheses previously
10 hampered by a lack of monophyly.

11 The results presented here suggest that ESP-COAL and MDC perform well with
12 respect to inference of a species topology when gene flow and incomplete lineage
13 sorting are occurring among closely related species. Our simulations, however, did not
14 encompass all possible configurations and magnitudes of gene flow or sampling
15 designs. For example, unequal rates of gene flow among lineages could result in a
16 reduction of the statistical power. Historical demographic events may also change our
17 results and the effect of such events on the power of these methods remains to be
18 investigated. Furthermore, the effect of incomplete species sampling remains unknown
19 relative to the power of all three methods examined here. The magnitudes and models
20 of gene flow investigated here, however, mirror those commonly used in population
21 genetic and phylogeographic inference and our results provide an initial step forward in
22 understanding the effects of common population processes on the ability to infer
23 phylogenetic trees for closely related species and populations. Understanding the

1 effects of such processes will not only aid in the inference of robust species topologies
2 for taxonomic and conservation studies, but will also contribute to the emerging field of
3 statistical phylogeography.

4

5 **Acknowledgements**

6

7 We would like to thank Benjamin Hall and Wennie Chou for providing the
8 *Rhododendron* DNA sequences. Special thanks to Gabriel Rosa and John Liechty for
9 assistance with the Department of Plant Sciences computing cluster located at the
10 University of California, Davis and with PERL scripting. We thank Amy Litt, Jeffrey
11 Oliver, and one anonymous reviewer for providing insightful comments that significantly
12 improved this manuscript.

1 **References**

2

3 Avise, J. C. 2000. *Phylogeography: The history and formation of species*. Harvard
4 University Press, Cambridge, MA.

5

6 Beerli, P., Felsenstein, J. 1999. Maximum likelihood estimation of migration rates and
7 population numbers of two populations using a coalescent approach. *Genetics* 152:763-
8 773.

9

10 Beerli, P., Felsenstein, J. 2001. Maximum likelihood estimation of a migration matrix
11 and effective population sizes in n subpopulations by using a coalescent approach.
12 *Proc. Nat. Acad. Sci. USA* 98:4563-4568.

13

14 Brumfield, R. T., Liu, L., Lum, D. E., Edwards, S. V. 2008. Comparison of species tree
15 methods for reconstructing the phylogeny of bearded manakins (Aves: Pipridae:
16 *Manacus*) from multilocus sequence data. *Syst. Biol.*, *in press*.

17

18 Carling, M. D., Brumfield, R. T. 2008. Integrating phylogenetic and population genetic
19 analyses of multiple loci to test species divergence hypotheses in *Passerina* buntings.
20 *Genetics* 178:363-377.

21

22 Carstens, B.C., Knowles, L. L. 2007. Estimating phylogeny from gene tree probabilities
23 in *Melanoplus* grasshoppers despite incomplete lineage sorting. *Syst. Biol.* 56:400-411.

- 1
- 2 Cox, P.A., Cox, K.N.E. 1997. *The Encyclopedia of Rhododendron Species*. Glendoick
3 Publishing, Perth, Scotland.
- 4
- 5 Cummings, M. P. 1994. Transmission patterns of eukaryotic transposable elements:
6 Arguments for and against horizontal transfer. *Trends Ecol. Evol.* 9:141-145.
- 7
- 8 Degnan, J. H., Salter, L. A. 2005. Gene tree distributions under the coalescent process.
9 *Evolution* 59:24-37.
- 10
- 11 Degnan, J. H., Rosenberg, N. A. 2006. Discordance of species trees with their most
12 likely gene trees. *PLoS Genetics* 3:e68.
- 13
- 14 Edwards, S. V., Liu, L., Pearl, D. K. 2007. High-resolution species trees without
15 concatenation. *Proc. Natl. Acad. Sci. USA* 104:5936-5941.
- 16
- 17 Excoffier, L., Novembre, J., Schneider, S. 2000. SIMCOAL: a general coalescent
18 program for the simulation of molecular data in interconnected populations with arbitrary
19 demography. *J Hered.* 91:506-9.
- 20
- 21 Felsenstein, J. 1985. Confidence limits on phylogenies. An approach using the
22 bootstrap. *Evolution* 39:783-789.
- 23

- 1 Felsenstein, J. 2004. Inferring Phylogenies. Sinauer Assoc. Sunderland, MA.
2
- 3 Fitch, W. M. 1970. Distinguishing homologous from analogous proteins. Syst. Zool.
4 19:99-113.
5
- 6 Goetsch, L., Eckert, A.J., Hall, B.D. 2005. The molecular systematics of *Rhododendron*
7 (Ericaceae): A phylogeny based upon RPB2 gene sequences. Syst. Bot. 30:616-626.
8
- 9 Hickerson, M.J., Stahl, E., Takebayashi, N. 2007. msBayes: A flexible pipeline for
10 comparative phylogeographic inference using approximate Bayesian computation
11 (ABC). BMC Bioinformatics 8: e268.
12
- 13 Hey J., 2006. Recent advances in assessing gene flow between diverging populations
14 and species. Curr. Opin. Genet. Dev.16:592-6.
15
- 16 Hey, J., Nielsen, R. 2004. Multilocus methods for estimating population sizes, migration
17 rates and divergence time, with applications to the divergence of *Drosophila*
18 *pseudoobscura* and *D. persimilis*. Genetics 167:747-760.
19
- 20 Hey, J., Nielsen, R. 2007. Integration within the Felsenstein equation for improved
21 Markov chain Monte Carlo methods in population genetics. Proc. Natl. Acad. Sci. U S A
22 104:2785-90.
23

- 1 Hudson, R. R. 1983. Testing the Constant-Rate Neutral Allele Model with Protein
2 Sequence Data. *Evolution* 37:203-217.
- 3
- 4 Hudson, R. R. 1990. Gene genealogies and the coalescent process. Pp. 1-44 in D. J.
5 Futuyma & J. Antonovics, eds. *Oxford Survey Evolutionary Biology*. Oxford University
6 Press, New York.
- 7
- 8 Hudson, R. R. 1992. Gene trees, species trees and the segregation of ancestral alleles.
9 *Genetics* 131:509-512.
- 10
- 11 Kingman, J. F. C. 1982. The coalescent. *Stochastic Process. Appl.* 13:235-248.
- 12
- 13 Knowles, L. L., Carstens, B. C. 2007. Estimating a geographically explicit model of
14 population divergence for statistical phylogeography. *Evolution* 61:477-493.
- 15
- 16 Kron, K. A., Judd, W. S., Stevens, P. F., Crayn, D. M., Anderberg, A. A., Gadek, P. A.,
17 Quinn, C. J., Luteyn, J. L. 2002. Phylogenetic classification of Ericaceae: Molecular and
18 morphological evidence. *Bot. Rev.* 68:335-423.
- 19
- 20 Kubatko, L. S., Degnan, J. H. 2007. Inconsistency of phylogenetic estimates from
21 concatenated data under coalescence. *Syst. Biol.* 56:17-24.

22

- 1 Kuhner, M. K. 2006. LAMARC 2.0: Maximum likelihood and Bayesian estimation of
2 population parameters. *Bioinformatics* 22:768-770.
- 3
- 4 Kuhner, M. K., J. Yamato, and J. Felsenstein. 1998. Maximum likelihood estimation of
5 population growth rates based on the coalescent. *Genetics* 149: 429-434.
- 6
- 7 Liu, L., Pearl, D. K. 2006. Species trees from gene trees: Reconstructing Bayesian
8 posterior distributions of a species phylogeny using estimated gene tree distributions.
9 Technical report #53, Ohio State University.
- 10
- 11 Maddison, W. P. 1997. Gene trees in species trees. *Syst. Biol.* 46:523-536.
- 12
- 13 Maddison, W. P., Knowles, L.L. 2006. Inferring phylogeny despite incomplete lineage
14 sorting. *Syst. Biol.* 55:21-30.
- 15
- 16 Maddison, W. P., Maddison, D. R. 2004. MESQUITE: a modular system for evolutionary
17 analysis. Version 1.01. available at <http://mesquiteproject.org>
- 18
- 19 Minin, V., Abdo, Z., Joyce, P., Sullivan, J. 2003. Performance-based selection of
20 likelihood models for phylogeny estimation. *Syst. Biol.* 52:674-683.
- 21
- 22 Nielsen, R. 2005. Molecular signatures of natural selection. *Ann. Rev. Genet.* 39:197-
23 218.
- 24

- 1 Milne, R. I. 2004. Phylogeny and biogeography of *Rhododendron* subsection *Pontica*, a
2 group with a tertiary relict distribution. *Mol. Phylo. Evol.* 33:389-401.
3
- 4 Nielsen, R., Wakeley, J. W. 2001. Distinguishing Migration from Isolation: an MCMC
5 Approach. *Genetics* 158:885-896.
6
- 7 Page, R. D. M., Charleston, M. 1997. From gene to organismal phylogeny: Reconciled
8 trees and the gene tree/species tree problem *Mol. Phylo. Evol.* 7: 231-240
9
- 10 Pamilo, P. Nei, M. 1988. Relationships between gene trees and species trees. *Mol.*
11 *Biol. Evol.* 5:568-583.
12
- 13 Pritchard, J. K., Stephens, M., Donnelly, P. 2000. Inference of population structure
14 using multilocus genotype data. *Genetics* 155:945-959.
15
- 16 Rannala, B., Yang, Z. 2003. Bayes estimation of species divergence times and
17 ancestral population sizes using DNA sequences from multiple loci. *Genetics* 164:1645-
18 1656.
19
- 20 Ree, R. H. Smith, S. A. 2008. Maximum-likelihood inference of geographic range
21 evolution by dispersal, local extinction, and cladogenesis. *Syst. Biol.* 57:4-414.
22

- 1 Rosenberg, N. A. 2002. The probability of topological concordance of gene trees and
2 species trees. *Theor. Pop. Biol.* 61:225-247.
3
- 4 Rosenberg, N. A. 2003. The shapes of neutral gene genealogies in two species:
5 Probabilities of monophyly, paraphyly, and polyphyly in a coalescent model. *Evolution*
6 57:1465-1477.
7
- 8 Sanderson, M. J. 2002. Estimating absolute rates of molecular evolution and divergence
9 times: a penalized likelihood approach. *Mol. Biol. Evol.* 19:101-109.
10
- 11 Slatkin, M., Maddison, W.P. 1989. A cladistic measure of gene flow inferred from
12 phylogenies of alleles. *Genetics* 123:603-613.
13
- 14 Slowinski, J. B., Page, R. D. M. 1999. How should species phylogenies be inferred from
15 sequence data? *Syst. Biol.* 48:814-825.
16
- 17 Swofford, D.L. 2002. PAUP*. Phylogenetic Analysis Using Parsimony (and other
18 methods). Version 4. Sinauer Associates, Sunderland, Massachusetts.
19
- 20 Syring, J., Farrell, K., Businsky, R., Cronn, R., Liston, A. 2007. Widespread
21 genealogical nonmonophyly in species of the *Pinus* subgenus *Strobus*. *Syst. Biol.*
22 56:163-181.
23

- 1 Tajima, F. 1983. Evolutionary relationship of DNA sequences in finite populations.
2 Genetics 105:437-460.
3
- 4 Tajima, F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA
5 polymorphism. Genetics 123:585-595.
6
- 7 Takahata, N., Nei, M. 1985. Gene genealogy and variance of interpopulation nucleotide
8 differences. Genetics 110:325-344.
9
- 10 Takahata, N. 1989. Gene Genealogy in Three Related Populations: Consistency
11 Probability Between Gene and Population Trees. Genetics 122:957-966.
12
- 13 Wakeley, J., Hey, J. 1997. Estimating ancestral population parameters. Genetics 145:
14 847-855.
15
- 16 Wall, J. D. 2003. Estimating ancestral population size and divergence times. Genetics
17 163:395-404.
18
- 19 Yang, Z. 2002. Likelihood and Bayes estimation of ancestral population sizes in
20 hominoids using data from multiple loci. Genetics 162:1811-1823.
21

1 **Table 1.** Comparison of phylogenetic accuracy across four models of gene flow for $N_e m$
 2 = 0.10, ST depth = $2N_e$, $N_e = 100,000$, and $d = 0.10$. CONC = concatenated data
 3 approach.

4

	Number of Loci								
	2	3	4	5	6	7	8	9	10
<i>n</i>-island									
ESP-COAL	0.30	0.51	0.43	0.63	0.60	0.65	0.72	0.64	0.61
MDC	0.52	0.55	0.60	0.65	0.70	0.74	0.74	0.75	0.74
CONC	0.00	0.02	0.03	0.04	0.09	0.12	0.14	0.24	0.23
stepping stone									
ESP-COAL	0.74	0.78	0.86	0.95	0.94	0.98	0.94	0.98	0.98
MDC	0.76	0.86	0.90	0.97	0.98	0.99	0.98	0.98	0.99
CONC	0.03	0.03	0.11	0.15	0.17	0.21	0.17	0.23	0.27
allopatric									
ESP-COAL	0.97	0.94	0.99	1.00	1.00	1.00	1.00	1.00	1.00
MDC	0.79	0.88	0.86	0.91	0.95	0.96	0.97	0.98	1.00
CONC	0.11	0.20	0.28	0.40	0.44	0.51	0.63	0.68	0.70
parapatric									
ESP-COAL	0.95	0.97	1.00	0.98	1.00	0.98	0.99	1.00	1.00
MDC	0.72	0.82	0.86	0.90	0.94	0.94	0.95	0.99	0.99
CONC	0.08	0.18	0.24	0.32	0.33	0.40	0.42	0.41	0.48

5

6

7

8

1 **Table 2.** Description of sample sizes, sequence diversity, and GenBank accession
 2 numbers for the data used to infer relationships among *R. brachycarpum*, *R.*
 3 *catawbiense*, *R. caucasicum*, and *R. macrophyllum*. All inferred trees were rooted using
 4 *R. aureum* as an outgroup. Sequence diversity is reported as the total number of
 5 polymorphic sites followed by the alignment length in parentheses. Insertion-deletion
 6 polymorphisms were ignored for all analyses.

	<i>RPB2</i>	<i>RPC1</i>	Chloroplast (<i>trnK</i> , <i>trnL-trnF</i>)
Sample size			
<i>R. brachycarpum</i>	5	2	1
<i>R. catawbiense</i>	5	4	1
<i>R. caucasicum</i>	2	2	1
<i>R. macrophyllum</i>	5	5	1
<i>R. aureum</i>	1	1	1
Total	18	14	5
Sequence diversity	37 (1301)	71 (1596)	18 (2718)
GenBank accessions	EU822187-EU822204	EU822173-EU822186	AY494173-AY494176 AY496914-AY496917

8

1 **Figure 1.** Models of gene flow showing (A) n -island, (B) stepping stone, (C) historical
2 allopatric, and (D) historical parapatric gene flow. Shown for each model are the species
3 phylogeny (bold outlined), as well as an example genealogy contained within the
4 species phylogeny. Branch lengths are standardized to a species tree depth of 1.00.
5 The models are differentiated by when gene flow occurs; this is represented on the tree
6 through the use of shaded rectangles. The n -island and stepping stone models are
7 differentiated by which lineages exchange migrants. This is shown with the curved lines
8 connecting the terminal lineages.

9
10 **Figure 2.** Performance of (A) ESP-COAL, (B) MDC, and (C) concatenation for the four
11 models of gene flow as shown in Figure 1. All models had $N_e m = 0.10$, $N_e = 10,000$, $d =$
12 0.10 , and a ST depth of $2N_e$ generations.

13
14 **Figure 3.** Comparison of ESP-COAL and MDC assuming the n -island model of gene
15 flow when (A) N_e is 10,000, the depth of the species tree is equal to $2N_e$, and migration
16 rates vary from $N_e m = 0.01$ to 1.00, (B) $N_e m$ is 0.10, the depth of the species tree is
17 equal to $2N_e$, and N_e varies from 10,000 to 100,000, and (C) N_e is 10,000, $N_e m$ is 0.10,
18 and the depth of the species tree varies from $2N_e$ to $6N_e$.

19
20 **Figure 4.** An empirical example from *Rhododendron*. Maximum-likelihood genealogies
21 for (A) *RPB2*, (B) *RPC1*, and (C) the chloroplast *trnK* and *trnL-trnF* gene regions.
22 Numbers above or below branches are bootstrap support values (%). Only support
23 values $\geq 70\%$ are shown.

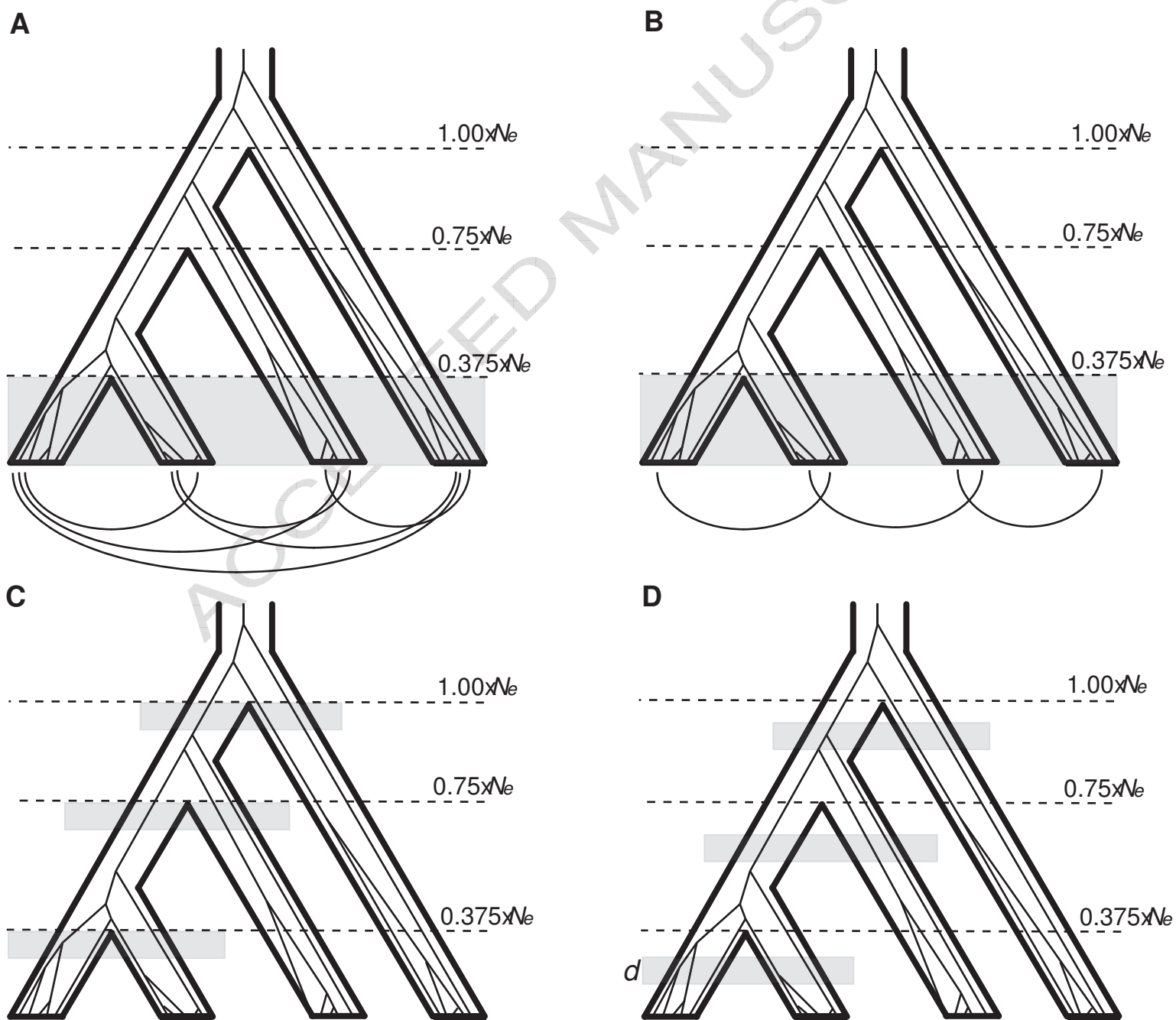
1

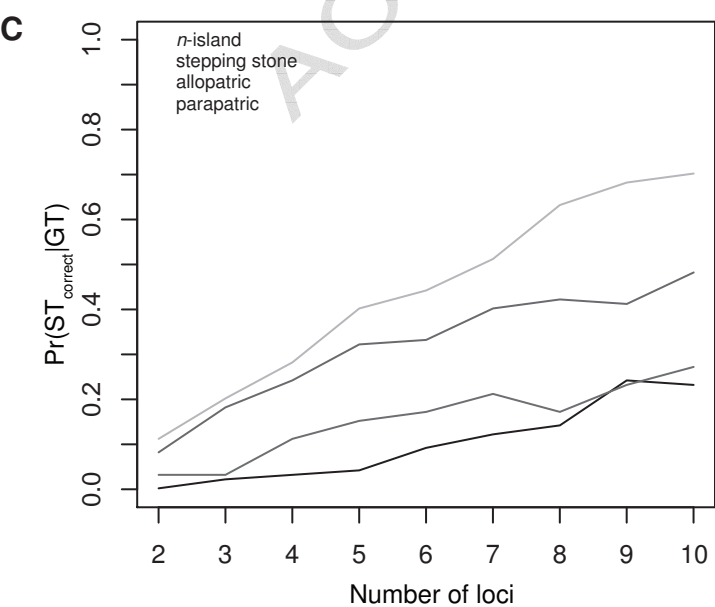
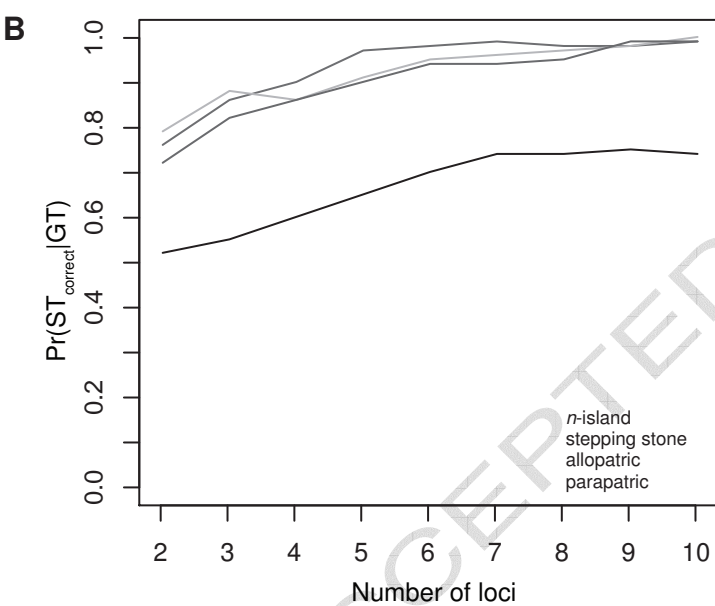
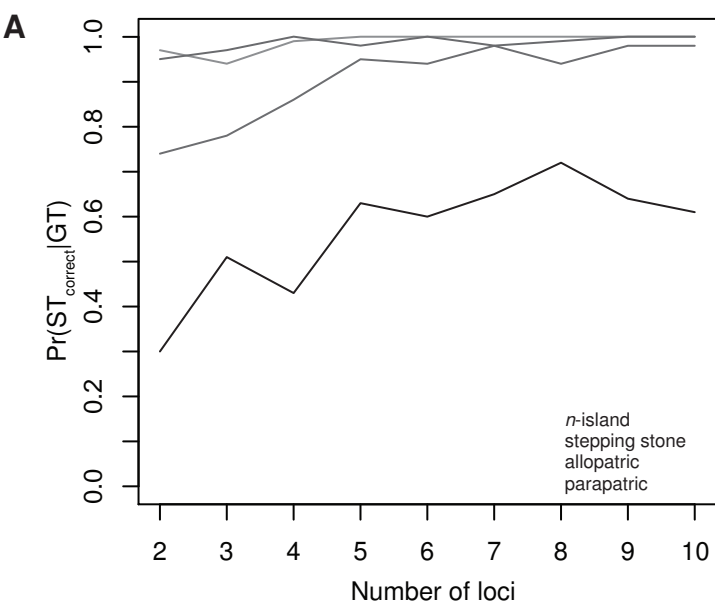
2 **Figure 5.** An illustration of the three phylogenetic hypotheses corresponding to the
3 evolutionary relationships among four *Rhododendron* species inferred using (A) ESP-
4 COAL, (B) MDC, and (C) concatenation. Numbers above or below branches are
5 bootstrap support values (%). Only support values $\geq 70\%$ are shown. Maximum-
6 likelihood gene trees (cf. Fig. 4) for each locus are contained within the species trees.

7

8 **Figure 6.** An illustration exploring the effect that error in the estimation of gene trees
9 has on the performance of ESP-COAL and MDC. Shown are results from conditions
10 that correspond to the empirical *Rhododendron* data ($\theta = 20$, $N_e = 10,000$, $N_e m = 0.10$, d
11 $= 0.10$), as well as a parapatric model of gene flow and a species tree depth equal to
12 $2N_e$

Figure 1





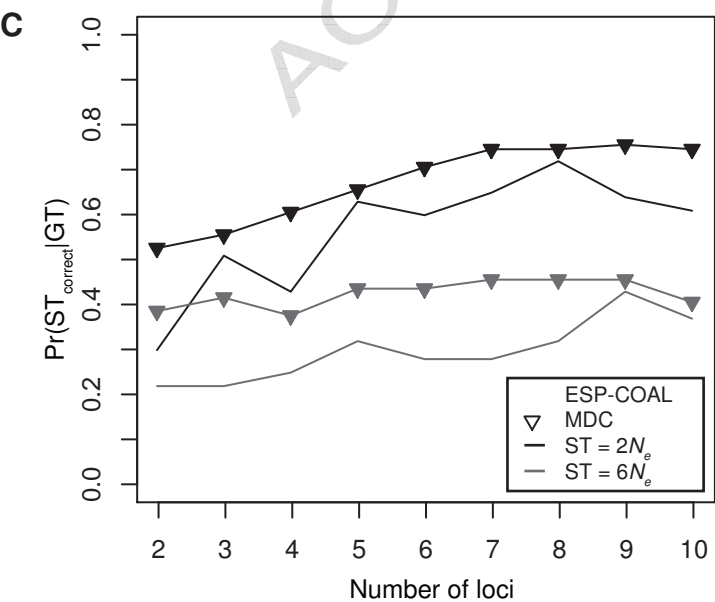
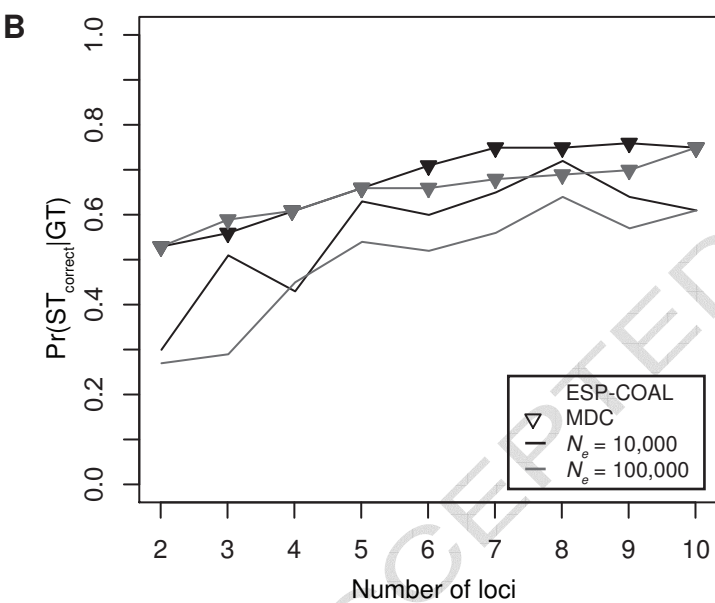
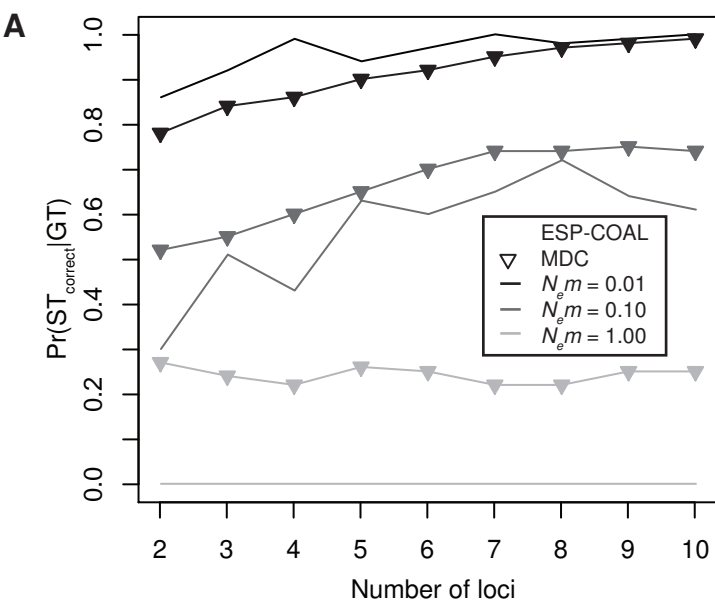
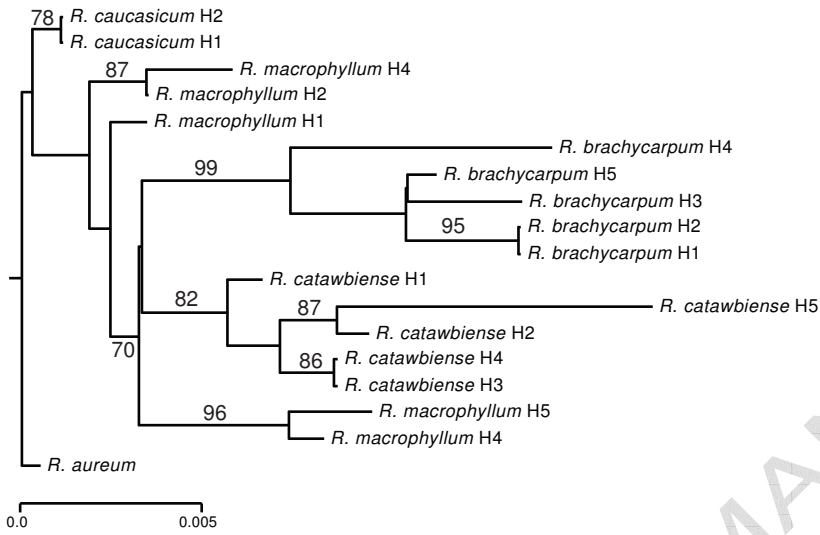
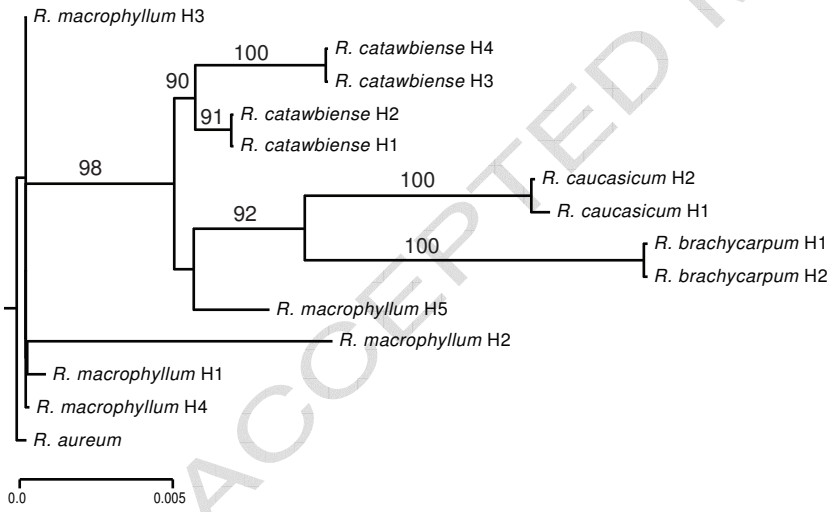


Figure 4

A



B



C

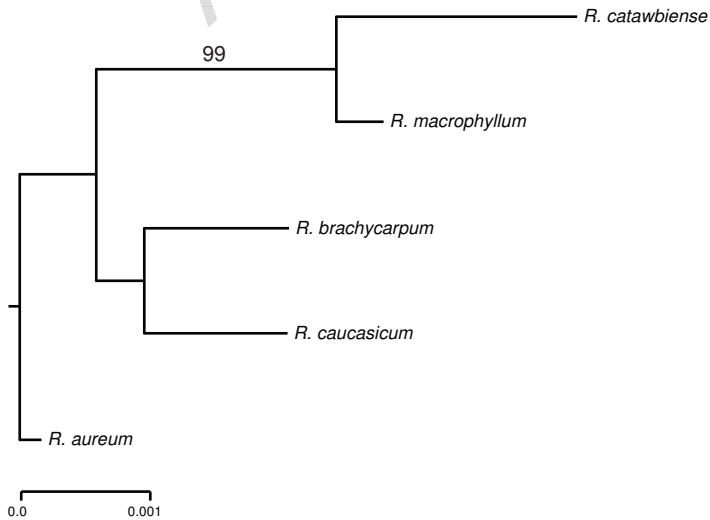
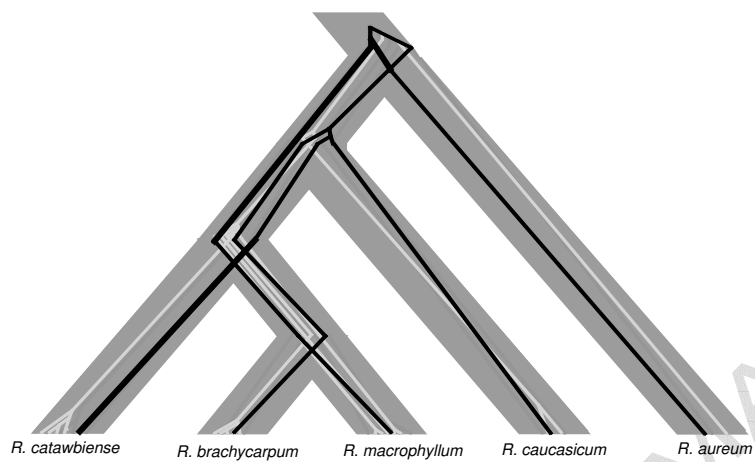
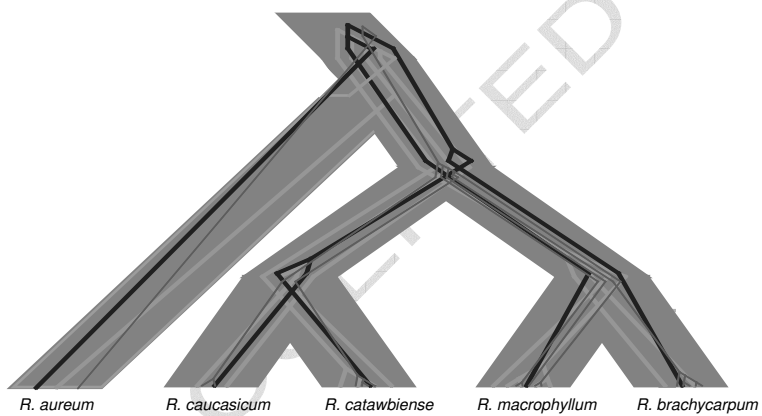


Figure 5

A



B



C

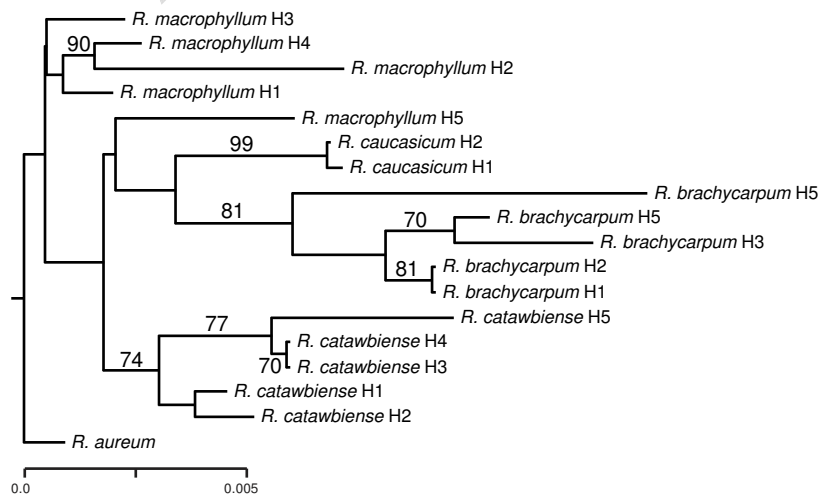


Figure 6

