Running head:

PRINCIPAL TREE INFERENCE FROM GENOMIC BLOCKS

Title:

Using genomic location and coalescent simulation to investigate gene tree discordance in *Medicago* L.

Sousa, F.[1,‡], Bertrand, Y.J.K.[1,‡], Doyle, J.J.[2], Oxelman, B.[1] and Pfeil, B.E.[1,*]

[1]Department of Biological and Environmental Sciences, University of Gothenburg, Box 461, 40530 Gothenburg, Sweden

[2]Department of Plant Biology, Cornell University, 404 Mann Library Building, Ithaca, NY 14853, USA

[*]Corresponding author: bernard.pfeil@gu.se

[‡]Equally shared contributions

ABSTRACT

Several well-documented evolutionary processes are known to cause conflict between species level phylogenies and gene level phylogenies. Three of the most challenging processes for species tree inference are incomplete lineage sorting, hybridisation and gene duplication, which may result in unwarranted comparisons of paralogous genes. Several existing methods have dealt with these processes but none has yet been able to untangle all three at once. Here, we propose a step-wise method by which these processes can be discerned using information on genomic location coupled with coalescent simulations. In the first step, highly discordant genes within genomic blocks (putative paralogues) are identified and excluded from the dataset and, in the second step, blocks of linked genes are grouped according to their hybrid history. Existing multispecies coalescent software can then be applied to recover the principal tree(s) that make up the species tree/network without violating the underlying model. The potential of the approach is evaluated on simulated data derived from a species network composed of nine species, of which one is of hybrid origin, and displaying a single gene duplication that leads to paralogous comparisons. We apply our method to an empirical set of 12 genes from seven species sampled in the plant genus *Medicago* that display phylogenetic discordance. We identify the causes of the discordance and demonstrate that the *Medicago orbicularis* lineage experienced an episode of ancient hybridisation. Our results

show promise as a new way to explore phylogenetic sequence data that can significantly improve species tree inference in presence of hybridisation and undetected paralogy or other causes leading to extremely discordant gene trees.

KEYWORDS: phylogenetic incongruence, species tree, gene tree, coalescent simulation, genomic location, paralogy, hybridisation, incomplete lineage sorting, principal tree

Species tree inference from molecular data has, in recent years, substantially shifted away from the concatenation of sequences from several genes to approaches based on the multispecies coalescent (Rannala and Yang, 2003; Degnan and Rosenberg, 2009). The multispecies coalescent assumes that each gene has its own phylogenetic history (Kingman, 1982) conditioned on the common underlying species phylogeny. However, several evolutionary processes may induce significant differences in phylogenies from unlinked genes, of which three of the best understood are: incomplete lineage sorting (ILS), homoploid hybridisation and paralogy (Maddison 1997).

Incomplete lineage sorting occurs when alleles coalesce more deeply than species divergence, as a consequence of stochastic differential assortment of ancestral polymorphism (e.g., Takahata and Nei 1985; Pamilo and Nei 1988; Maddison 1997). Although the effect of ILS on phylogenies can be profound (Degnan and Salter 2005), especially in sexually reproducing organisms with large populations and when the time between speciation events is short (Pamilo and Nei 1988; Nei and Kumar 2000), ILS itself is assumed to be ubiquitous, as an unavoidable consequence of neutral population processes. Gene tree incongruence has been explained by ILS in several studies that have applied coalescent-based methodologies (Pollard et al. 2006; Carstens and Knowles 2007; Syring et al. 2007; Blanco-Pastor et al. 2012).

Homoploid hybridisation, the formation of hybrids between different species at the same ploidy level, can result in the introgression of alleles from one species into another, or even in the formation of new species that contain alleles from both parents (e.g., Buerkle and Rieseberg 2008; Hermansen et al. 2011; Bosse et al. 2014). Hybridisation has been widely documented in animals and plants (Rieseberg 1997; Mallet 2007), with several studies demonstrating introgression of alleles or single nucleotide polymorphisms (SNPs) while taking into account ILS (Buckley et al. 2006; Peters et al. 2007; Blanco-Pastor et al. 2012; Phillips et al. 2013; Ramadugu et al. 2013; Good et al. 2015).

Genes are frequently subjected to duplication (the formation of paralogues) and loss. When paralogues from different species are mistaken for orthologues, their comparison may result in invalid inferences regarding species relationships (e.g., Doyle 1992; Oxelman et al. 2004).

Genealogical discordance due to paralogy can be further complicated by concerted evolution (Sanderson and Doyle 1992; Nei and Rooney 2005). Many examples drawn from eukaryotic species show that both tandem and whole genome duplications are very common (e.g., Lynch and Connery 2000; Cui et al. 2006; Chen et al. 2013; Van Zee et al. 2016) and can be followed by rapid gene loss (Blanc and Wolfe 2004). Tandem duplications can be extensive in some diploid organisms, even ones with relatively small genomes, such as *Arabidopsis thaliana* (Blanc and Wolfe 2004). The frequent absence of a full set of duplicated genes, either due to gene loss or sampling bias, fosters unintentional paralogous comparisons. We denote unrecognized instances of paralogy as "paralogy-affected" genes throughout the paper.

The phylogenetic effect of each of these three confounding processes has been widely studied (e.g., Maddison and Knowles 2006; Edwards et al. 2007; Holland et al. 2008; Maureira-Butler et al. 2008; Joly et al. 2009; Meng and Kubatko 2009; Joly et al. 2010; Heled and Drummond 2010; Jones et al. 2013; Yu et al. 2013). Several methods implement species tree inference that accommodates ILS, by using the multispecies coalescent (e.g., BEST, Liu and Pearl 2007; STEM, Kubatko et al. 2009; *BEAST, Heled and Drummond 2010; MP-EST, Liu et al. 2010; PhyloNet, Liu et al. 2014; ASTRAL, Mirarab et al. 2014), with two of these incorporating hybridisation in species tree inference (STEM-hy, Kubatko 2009, and PhyloNet). Another approach that infers species trees despite incongruence, but without modelling any specific causal process, is BUCKy (Ané et al. 2007). Other approaches determine whether ILS alone can explain the observed gene tree incongruence, and thus infer hybridisation on a case-by-case basis rather than by co-estimation (Buckley et al. 2006; Maureira-Butler et al. 2008; the JML method, Joly et al. 2009; the ABBA-BABA test, Green et al. 2010; Bertrand et al. 2015).

Models that accommodate gene duplications in a coalescent framework have been proposed (Rasmussen and Kellis 2012), but only for cases involving the presence of both duplicates at least in some gene trees. However, no method, step-wise or simultaneous, has yet accommodated all three sources of gene tree incongruence (ILS, hybridisation and paralogy). Instead, each method handles only one or at most two of these sources of incongruence, and requires the assumption that the other(s) are not operating (or do not affect the outcome). Given that each of these sources of incongruence, when not handled appropriately, can mislead species tree inference (e.g., Oxelman et al. 2004; Kubatko and Degnan 2007; Reid et al. 2012), it is critical that all three can be diagnosed when acting in concert.

The legume genus *Medicago* L. (Fabaceae) exhibits severe phylogenetic incongruence among nuclear and chloroplast markers (Bena 2001; Steele and Wojciechowski 2003; Maureira-Butler et al. 2008; Steele et al. 2010; Yoder et al. 2013; Sousa et al. 2014; Sousa et al. 2016) that

has been attributed, at least partially, to hybridisation (Maureira-Butler et al. 2008; Yoder et al. 2013; Sousa et al. 2016), although paralogy has not been formally excluded. A survey of published *Medicago* phylogenies indicated that *M. orbicularis* is one of the taxa putatively involved in ancient hybridisation, possibly involving the lineages that include *M. truncatula*, *M. ciliaris* and *M. arabica* (Sousa et al. 2016). The present study investigates gene tree incongruence involving the phylogenetic position of *M. orbicularis* using a method that untangles topological variation caused by ILS, paralogy and hybridisation. We explore the hypothesis that genomic location can be used to identify possible cases of paralogy by comparing groups of genes that are physically associated (tightly linked), and to recognize hybridisation by comparing groups of genes that are physically distant (unlinked). We use target-enrichment methods (Guschanski et al. 2013; Smith et al. 2014; Stull et al. 2013; Sousa et al. 2014) to obtain sequences of multiple loci from each sampled specimen.

To evaluate the performance of our method, we simulate data under a species network in the presence of paralogy, hybridisation and ILS. We then (1) identify genes likely to be paralogy-affected and remove them from further analyses, (2) group genes with a shared parental origin (due to hybrid parentage) and (3) reconstruct species phylogenies, while accounting for ILS, and confirm the different parental origins of taxa with hybrid histories through multi-species coalescent inference. Both the paralogy detection and hybridisation detection steps are based on the test developed by B.E.P. and described in Maureira-Butler et al. (2008) that assumes a null hypothesis of ILS and derives null distributions based on gene trees used as surrogates for the unknown species tree. We show that likely cases of paralogy in *Medicago* that could mislead species tree inference are successfully detected. We confirm that a hybridisation signal can be discerned in *Medicago orbicularis* only if genomic location is taken into account.

METHODS

SAMPLING, ALIGNMENT ASSEMBLY AND GENE TREE INFERENCE

We sampled seven species of *Medicago* (*M. arabica* (L.) Huds., *M.ciliaris* (L.) All., *M. granadensis* Willd., *M.intertexta* (L.) Mill., M. *littoralis* Rohde ex Loisel., *M.orbicularis* (L.) Bartal., *M. truncatula* Gaertn.) and two outgroups (based on e.g., Steele et al. 2010), *Melilotus neapolitanus* Ten. and *Mel. sulcatus* Desf. We used 12 markers (Supp. Table 1), of 2-3 Kb in length, located in four unlinked blocks of three genes (hereafter referred to as blocks A, B, C and D), with < 30 kb span between genes inside a block and each block located on a different chromosome, based on the *M. truncatula* genome (Young et al. 2011) and previously tested for phylogenetic utility (Sousa et al. 2014). We considered that four blocks of three genes each was an

adequate dataset to explore the use of genomic location for the identification of reticulation using our newly proposed method. The 12 genes selected showed, in a preliminary gene tree analysis, evidence for alternative positions of *Medicago orbicularis*. DNA extraction, target-enrichment and Illumina sequencing were performed as reported previously (Sousa et al. 2014). Raw sequence reads were submitted to the European Nucleotide Archive repository with the following accession numbers: *Medicago arabica*: ERS719973, *M. ciliaris*: ERS719974, *M. intertexta*: ERS719975, *M. granadensis*: ERS719976, *M. littoralis*: ERS719977, *M. orbicularis*: ERS719978, *Melilotus sulcatus*: ERS511667, *Melilotus neapolitanus*: ERS511668. *Medicago truncatula* sequences were obtained from the reference genome at www.medicagohapmap.org.

CLC Assembly Cell v.4.0.13 software (CLC Bio, Aarhus, Denmark) was used to remove adapter sequences and filter reads for quality, with a phred-score threshold of 20. CLC-mapper was employed to map reads from individual samples to the genomic reference sequence for each gene. Alleles were phased using the program Samtools phase (Li et al. 2009) and phased reads were re-assembled into contigs using CLC-assembler. Sequences were aligned in Geneious Pro (v.5.3.6), using the Geneious alignment tool with default parameters and five iterations. Alignments were tested for recombination with RDP4 (Martin et al. 2010). Recombination events were accepted when detected by at least two methods at P-values <0.01 (with Bonferroni correction) in presence of topological conflict; the corresponding recombinant fragments were removed.

In order to estimate appropriate substitution models, we analysed the alignments, previously filtered for recombination, in MrBayes v.3.1.2 (Huelsenbeck and Ronquist 2005) using model averaging over the GTR model family (Huelsenbeck et al. 2004) and gamma rate variation. Each analysis was run for 3M generations, mixing and ESS values were controlled in Tracer v1.6.0 (Rambaut and Drummond 2007). A run was deemed successful if examination of the parameter files in Tracer showed no indications of a lack of convergence and displayed high ESS values (above 100) for all parameters of interest. The "sump" function in MrBayes was then used on each run, with a specified burn-in (specific to each gene based on the Tracer results, not reported here), to verify which substitution model had the highest posterior probability. Ultrametric gene trees were then obtained in BEAST v. 1.8.0 (Drummond and Rambaut 2007) using the inferred substitution model and gamma rate variation, an uncorrelated lognormal relaxed clock with a ucld.mean rate normal prior with mean=3.6×10^{-9} (Sousa et al. 2014) and a normal root age prior with mean=15.9×10^{6} (Lavin et al. 2005; Maureira-Butler et al. 2008).

UNDERLYING ASSUMPTIONS FOR DETECTING ILS, HYBRIDISATION AND PARALOGY

i) Chromosomal blocks become fixed in a hybrid genome

We assume that lineages that have undergone hybridisation eventually fix large chromosomal blocks (= genomic blocks) derived from each parental lineage. Once blocks are fixed, i.e. become homozygous, recombination acts only on segments with the same parental origin. Coupled with the emergence of new mutations, this causes alleles at each locus within fixed blocks to accumulate differences and potentially undergo independent ILS with respect to speciation events subsequent to the hybridisation event. Despite the effect of ILS, for each gene tree within a block, the alleles will still display a closer relationship to the parental lineage of the block, rather than to the other parental lineage (i.e., of other genomic blocks). Each alternative parental relationship is henceforth referred to as a principal tree (PT) (Holland et al. 2008). The result of a single hybridisation episode and subsequent fixation of genomic blocks is a genome that is a mosaic of blocks, and within a block all the genes are evolving according to one of two principal trees (PT1 or PT2).

ii) Fixed genomic blocks are typically large

The two main mechanisms that produce mosaic genomes (homoploid hybrid speciation (HHS) and introgression) are expected to promote the rapid fixation of some genomic blocks: by population bottlenecking and drift (e.g., Buerkle and Rieseberg 2008) for HHS, and by selection in the case of introgression, where the majority of introgressed regions are lost due to drift, but a few that carry adaptive mutations may be rapidly fixed by selection (e.g., Liu et al. 2015). Genomes containing large hybrid blocks have been observed in cases of introgression as well as HHS. The introgression of anticoagulant-resistance alleles into some house mouse populations from the Algerian mouse have resulted in a few large introgressed blocks (some >10 Mb) (Song et al. 2011; Liu et al. 2015). The average genomic block size of recently introgressed (c. 200 years ago) Asian-descendant alleles following hybridisation of Asian domestic pig into a background of European breeds is over 100 000 bp (Bosse et al. 2014), although fixation has not occurred at every locus. Another case where the results of introgression have not completely stabilised is that of Neanderthal alleles into modern humans, as blocks with a median of 129 000 bp have been inferred to occur (Sankararaman et al. 2014). Three homoploid hybrid North American sunflower species display large blocks (Ungerer et al. 1998; Rieseberg et al. 2003), and the homoploid hybrid Italian sparrow has inherited most of its Z chromosome from the house sparrow (Trier et al. 2014).

iii) Genomic rearrangements do not disrupt small genomic blocks

Gene synteny within small genomic blocks in unlikely to be disrupted by genomic rearrangements, which are relatively rare. Large, chromosome-scale translocations or inversions will carry entire genomic blocks intact, and so can also be ignored for our purposes.

iv) Paralogue retention outlasts polymorphism retention

Duplicated gene copies can arise within genomic blocks, e.g., by tandem duplication (e.g., Innes et al. 2008). We expect that when paralogues are maintained for long periods, and loss is random in different lineages, the corresponding gene topologies will be more discordant than expected by ILS alone (e.g., see Supp. Fig. 1). Indeed, paralogous gene copies can persist for much longer than selectively neutral allelic polymorphisms. For example, many duplicated genes (20-50% of total gene content) originating by whole genome duplication (WGD) can be preserved for hundreds of millions of years (Lynch and Force 2000; Maere et al. 2005). In the legume *Glycine*, 77% of low copy genes in a 1 Mb duplicated region have been retained for at least 5 Ma and maybe as long as 10 Ma (Innes et al. 2008). In contrast, the typical duration for ancestral polymorphism retention (< 5 × *Ne* generations, Rosenberg 2003) is much less than 5 Ma under typical plant generation times and effective population sizes (*Ne* × generation time is usually < 1 million: Maureira-Butler et al. 2008; Strasburg and Rieseberg 2008; Foxe et al. 2009; Lundemo et al. 2009; Gossmann et al. 2010; Blanco-Pastor et al. 2012; Ramadugu et al. 2013; Pérez-Collazos et al. 2015). However, discordance caused by paralogy can also be minimal and approach the level of incongruence obtained through ILS. Such cases of paralogy are difficult to detect but are also unlikely to create problematic violations of the multispecies coalescent model (see Results). Thus, paralogy identification is most important when phylogenetic discordance is higher than that expected due to ILS, but is also most easily uncovered under these conditions.

It should be noted that any other cause of very discordant gene trees that operates on a single gene at a time (and not on entire blocks), whether biological (e.g. recombination, base composition bias) or methodological (e.g. model misspecification, alignment errors) could be mistaken for paralogy by our method. Therefore, an inference of a "paralogy-affected" gene should be treated with caution. However, the use of long informative alignments, recombination testing, quality control in laboratory procedures, dense taxon sampling, among other things, will reduce the chance that other causes will produce extremely discordant gene trees, leaving gene duplication and loss dynamics as (arguably) the most likely contender. For the purposes of describing our method, we assume that best practice procedures have been followed at every stage of analysis, and thus use the inference of "paralogy-affected" genes as a shorthand to indicate that this is the most likely cause. However, we acknowledge that other causes cannot be ruled out given our methodology.

Outline of the method to detect ILS, Hybridisation and Paralogy

The proposed method consists of four main steps: 1) Gene tree inference; 2) Paralogy test - identification and exclusion of highly-discordant trees within genomic blocks; 3) Hybridisation test - identification of hybridisation between genomic blocks and grouping of blocks sharing the same

parental history; 4) Inference of principal trees. Steps 1 and 4 rely on established methods for tree inference. Steps 2 (Paralogy test) and 3 (Hybridisation test) as well as the coalescent simulations tests used for gene pairwise comparisons, are explained below.

Paralogy Test

We considered a sampling scheme where each genomic block has three genes (triplet) (Fig. 1). Pairwise comparisons (Supp. Fig. 1) were performed between the observed gene trees within each genomic block, using the coalescent simulation test, for a total of three comparisons (Supp. Fig. 2). A gene tree was inferred as likely to be affected by paralogy if ILS was rejected in both pairwise comparisons with the other two gene trees of the triplet (Fig. 1, also Supp. Fig. 2). All genes identified as paralogy-affected were removed from further analysis. When ILS was rejected in all three pairwise comparisons, the whole block was excluded, because this case does not allow the identification of the paralogy-affected gene nor does it produce any remaining compatible pairs of genes. In contrast, if the block was not excluded, genes that differ by ILS alone where further subjected to the hybridization test.

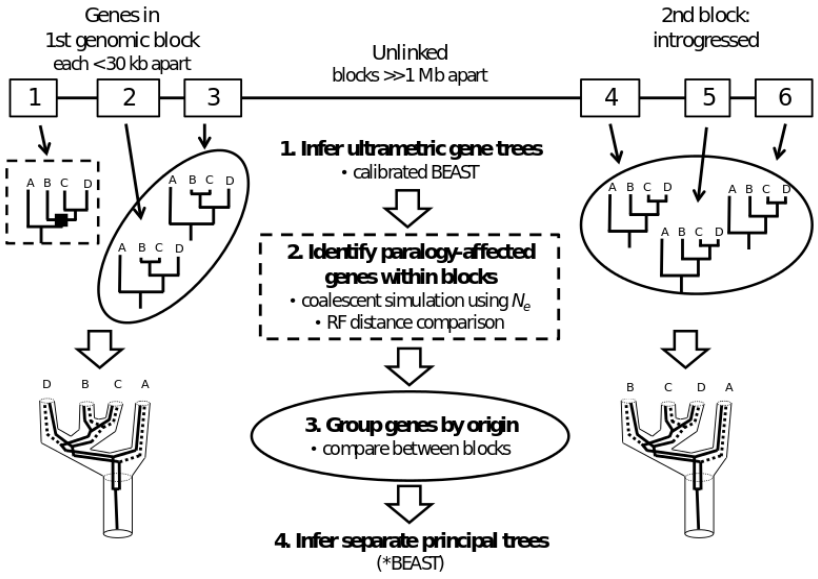


Figure 1.

Coalescent Simulation Tests

All pairwise comparisons used in our method are based on a parametric bootstrapping

procedure that compares the topological distance between two observed gene trees with the topological distances expected, under a coalescent process, between the simulated gene trees from two modelled species trees (Maureira-Butler et al. 2008; Blanco-Pastor et al. 2012; Ramadugu et al. 2013). In the original implementation (Maureira-Butler et al. 2008), the observed distance was calculated directly from a point estimate of each gene tree, and thus did not account for tree inference uncertainty. In order to take this source of uncertainty into account, we instead derived a distribution of distances from a posterior distribution of Bayesian trees. All tree distances in this test are calculated using the unweighed Robinson-Foulds (RF) metric (Robinson and Foulds 1981), a.k.a symmetric distance.

The test can be summarised as follows: Let G_1 and G_2 be two observed gene trees, and let D_1 and D_2 be the respective post burn-in gene tree posterior distributions. The topological distance between G_1 and G_2 (observed gene tree distances: Supp. Fig 2) is represented by a distribution, with a 95% credibility interval, of the distances between 100 trees drawn randomly from D_1 (D_1^1 to D_1^{100}) and 100 trees drawn randomly from D_2 (D_2^1 to D_2^{100}) – 100 x 100 RF distances. As the true species trees that contain G_1 and G_2 are unknown, they are modelled by treating G_1 and G_2 as surrogate species trees. From each tree drawn from D_1 (D_1^1 to D_1^{100}) and from D_2 (D_2^1 to D_2^{100}), 100 gene trees are simulated, under the multispecies coalescent, using Kingman's neutral coalescent process, resulting in simulated sets, S_1^1 to S_1^{100} and S_2^1 to S_2^{100}, each with 100 trees. The ILS null is obtained by calculating the RF distance between each tree drawn from D_1 and D_2 and the respective simulated set (D_1^1x S_1^1, D_1^2x S_1^2, … , D_1^{100}x S_1^{100}), generating two distributions (for G1 and G2) of 100 x 100 RF distances each (referred to as coalescent distance distributions, see Supp. Fig 2). Gene tree distances are then used as a test statistic for an upper-tailed test with a given critical value. We define the p% critical value to be the value of the simulated null distributions that (1-p)% of the simulated values exceed. The null hypothesis of ILS is rejected if the test statistic (gene tree distances) exceeds the p% critical value of both coalescent distance distributions.

The effect of population sizes in simulations for a similar coalescent-based test was assessed in Bertrand et al. (2015), where it was found that the performance of the test improved when the population sizes used for simulating the gene trees were significantly smaller than the real population size. One explanation that can account for this observation is that gene lineages have coalescent times extending farther back than species lineages (Maddison and Knowles 2006). Consequently, using a gene tree as a surrogate species tree for gene tree simulation will tend to further push back in time the divergence ages on the simulated trees. It ensues that the use of a surrogate tree introduces more sequence variation in the simulations than what would be obtained in an actual species tree. Instead of evaluating the population size needed for satisfactory ILS

293 acceptance/rejection rates in our coalescent simulation test, we opted for changing the critical value
294 of the one-tailed test, and thus performed it with p% critical values of 65%, 75%, 85% and 95%.
295
296
297 *Hybridisation Test*
298 After the exclusion of putative paralogy-affected genes, we tested for gene tree discordance
299 between genomic blocks (Fig. 1). The coalescent simulation test was used for pairwise comparisons
300 between genes from different blocks (maximum of 3 x 3 gene comparisons). We accept
301 hybridisation as the source of the observed discordance only when ILS was rejected in at least two
302 pairwise comparisons between genes from different blocks. Other decision rules were tested
303 (accepting hybridisation if ILS was rejected in at least one pairwise comparison, or only if all
304 pairwise comparisons rejected ILS) but these proved to have worse rates of type I or type II error
305 (not shown). Genes from genomic blocks that share the same PT were pooled for species tree
306 inference (Fig. 1).
307
308 VALIDATION OF THE METHOD
309 *Data Simulation*
310 The simulation protocol is outlined in Figure 2 and Supplementary Figure 3. We used a
311 simulated nine-taxon network, including one species of hybrid origin and one gene duplication (Fig.
312 2a). Principal tree 1 (PT1) was generated in Mesquite v.2.75 (Maddison and Maddison 2006),
313 assuming a total tree depth of 2 000 000 generations, though a pure birth (Yule) process (Fig. 2b).
314 Principal tree 2 (PT2) was generated from PT1 by grafting the hybrid taxon to the alternative
315 parental lineage, at the same tree depth (Fig. 2c).
316 Our approach consisted in analysing replicates (Supp. Fig. 3) that contain data simulated
317 under a range of conditions. Each replicate modelled four unlinked genomic blocks, each with three
318 non-homologous genes. For the first two blocks, the genes evolved according to PT2, according to
319 PT1 for the third block and in the last block two genes evolved according to PT1 and one gene was
320 paralogy-affected. Paralogous trees were obtained from PT1, by duplication of the clade containing
321 species 2, 3, 4, 6, 7 and 8, followed by the random pruning of one copy from each taxon descended
322 from the gene duplication (i.e., the locus was returned to single copy in all descendants, Fig. 2d).
323 We simulated gene trees in each block using Kingman's neutral coalescent process
324 implemented in the Dendropy python library (Sukumaran and Holder 2010). On paralogous trees,
325 the random pruning of one copy of each terminal taxon produced trees with RF distance from PT1
326 ranging from a minimum of 0 (identical) to a maximum of 8. We aimed at evaluating the influence

of the topological effect introduced by paralogy on our method. To this purpose, we simulated 100 trees for each of the possible distances and grouped them into five paralogy bins corresponding to the RF values of [0, 2, 4, 6, 8], in order to observe the general trends across the results.

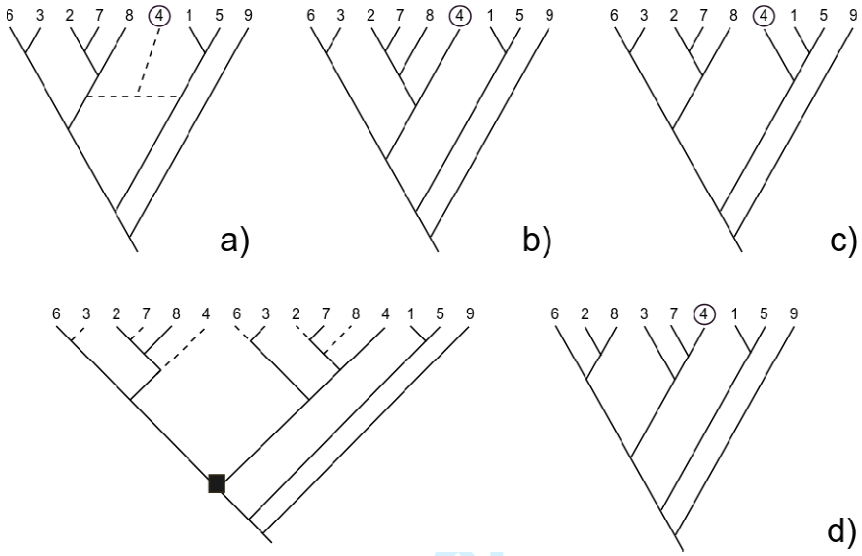


Figure 2

Tree depth was also included as a source of variation among replicates. The simulation of gene trees was performed under a fixed PT depth in number of generations, but varied in terms of coalescent units (CU; i.e., species tree branch length/N_e) by modifying the effective population size (N_e) in the simulations. To this effect, we selected a branch in the trees and varied the global N_e in order to obtain the desired length for this reference branch. This branch extended from the root to the node where the hybrid lineage (taxon 4) attached to the parental lineage in each PT (equivalent in length in PT1 and PT2). The length of this reference branch in CUs is the key parameter that determines the difficulty of detecting hybridisation and distinguishing it from ILS. ILS will, on average, be more prevalent when the branch length in CUs is short. Deep coalescence of alleles from taxon 4 in either PT (i.e., where polymorphisms were held throughout the reference branch) can result in gene trees where this taxon is placed sister to the remaining species, and thus ILS alone can produce identical topologies from gene trees derived from either PT. Given that the entire tree was varied in a systematic way (by changing N_e during the simulations), the reference branch length also determines the relative difficulty of detecting paralogy (along with the paralogy bin, i.e., detection should be easier with greater RF distances). Five different levels of CU, ranging from 2 to 6, with a unit increment, were tested. For a species with one generation/year and N_e of 50 000, this represents a realistic interval between speciation events, i.e., 100 000 to 300 000 years.

In order to include realistic parameters values in our simulation, we acquired nucleotide substitution models and rates from 50 empirical nucleotide alignments of low copy nuclear markers (Sousa et al. 2014), each with the same six species (four *Medicago* species + two species of its sister genus, *Melilotus*). Substitution models were inferred in jModelTest v. 2.1.4 (Darriba et al. 2012). The empirical distributions of mean rates were modelled with a lognormal distribution, from which the mu and sigma parameters were retrieved, using the distribution fitting software EasyFit 5.3 (MathWave Technology). For each simulated tree, we randomly sampled one of the 50 empirical genes to obtain the nucleotide substitution model and the 'meanrate' lognormal distribution parameters. Trees were converted from generations to substitutions by sampling a rate for each branch from the associated lognormal distribution. Random selection of substitution models and rates simulated the complexity expected from empirical data. Their effect was not evaluated systematically and we treated them as nuisance parameters downstream, focusing instead on the effect of ILS and duplication. We simulated sequence alignments of 1 500 bp using SeqGen (Rambaut and Grass 1997) for a total of 30 000 alignments: 100 replicates × 12 genes × 5 CUs × 5 paralogy bins. For each replicate of 12 genes, we therefore simulated alignments comparable to our empirical data with respect to sequence length per locus and number of loci, as well as model of sequence evolution and tree lengths.

Modelling of the Tree Root Prior

In our simulation we used a tree depth of 2 M generations and assumed a generation time of one year. The synthetic alignments were analysed in BEAST, which requires either time or rate prior information. In an application of our test to empirical data, a fossil (for example) could be used to calibrate the root of the species tree. However in our simulation, because we calibrated gene trees instead of species trees, there is a mismatch between species tree and gene tree node ages that needs to be taken into account. We therefore used an exponential prior distribution on the crown node to calibrate the gene trees estimated from the simulated data sets, offset at 2 Ma (as though we had a fossil date for the node) and with a mean equal to $2N_e$ (to include the variation expected due to the coalescent).

However, we could not infer N_e directly from our data, since we had simulated a single allele per taxon. To determine the error associated with N_e estimates from allelic data, we simulated 100 alleles per taxon from PT1 with SeqGen at population sizes of 10^3 and 10^4. Using 100 simulated alleles for one species and inferring N_e in DNAsp (Librado and Rozas 2009), we observed up to 50% of deviation in the recovered N_e from the values used for allele simulation. Thus, the mean of the exponential distribution for the tree root prior was obtained by first

calculating the mean N_e used for the gene tree simulations (across all CUs), and adding 50% to incorporate the expected degree of uncertainty. This value was then multiplied by two ($2N_e$ being the expected mean of exponentially distributed coalescent times) and added to the constraint, which equated to a mean root age of 2.87 Ma.

Gene Tree Inference from Simulated Alignments

Each simulated alignment was analysed in BEAST v1.8.0. under a substitution model obtained from jModelTest, a Yule prior on the branching process (a species branching process, because we simulated one sequence per species), and an uncorrelated lognormal relaxed clock with a ucld.mean rate prior of 3.6×10^{-9} (Sousa et al. 2014). We used a starting tree with a topology equivalent to PT1 (to speed up the search, given the large number of replicates to analyse) and a root age of 2 Ma. Each dataset was analysed with an MCMC of 5 M generations and the maximum clade credibility tree was summarized in TreeAnnotator (Drummond and Rambaut 2007) with 10% burn-in samples discarded. The resulting gene trees were subjected to paralogy and hybridisation tests, and to the subsequent grouping of genes according to their underlying PT. All tests were implemented in a Python 2.7 pipeline relying on methods from the Dendropy (Sukumaran and Holder 2010), Networkx (Hagberg et al. 2008) and Biopython (Cock et al. 2009) libraries and parallelized using MPI for Python (Dalcín et al. 2008).

Grouping of Genomic Blocks and Species Tree Inference

Genomic blocks differing by larger values than what can be obtained from ILS alone are considered to originate from different PTs. In each replicate, the expected result of the hybridisation test is the identification of two PTs (PT1 and PT2, as per our simulations). We evaluated whether the genes categorized into two principal tree groups would recover PT1 and PT2. To this effect we estimated species trees from each group of genes, in ten replicates drawn from the bin RF=8 at CU=2 (the hardest tree length for the test), using *BEAST. A lognormal distribution with mean=3.61×10^{-9} was set for substitution rates prior (Sousa et al. 2014). For population size we chose a normal distribution for prior with a mean=4.34×10^{5} (the mean population size value used for simulation with a 50% error increase, as used for data analysis) and sd=5×10^{4}, truncated to 2×10^{6}. A normal prior was applied to the species tree root height (mean=2.87×10^{6}, sd=0.5×10^{6}). Analyses were run for 50M MCMC generations with three separate chains for each analysis.

APPLYING THE METHOD TO THE *MEDICAGO* DATASET

Paralogy and hybridisation tests were run on the 12 inferred gene trees, in the same way as

for the simulated data, with acceptance at p% critical values of 95%, 85%, 75%, 65% of the null distributions. To generate the ILS null distribution we used the effective population size of N_e = 240 000 (Maureira-Butler et al. 2008). After running the tests and removing putative paralogy-affected genes, we grouped the genes that shared the same principal tree. Genes within each group were subjected to species tree analysis in *BEAST. Chains were run for 50 M generations using the inferred substitution model for each gene, an uncorrelated lognormal relaxed clock with a ucld.mean rate normal prior with mean=3.6×10^{-9}, a Yule model species tree prior and a normal prior on the species-tree root height with mean=15.9×10^{6} and sd=2.7×10^{6}.

In order to determine how important the new part of the testing procedure was – the paralogy test – we also ran the hybridisation test by itself on the *Medicago* data. Thus, inside a block, genes were presumed to differ by ILS alone, and we did not test for within-block incongruence. Each gene in a block was tested against all the genes in the other blocks. Blocks were handled separately if two pairwise rejections between blocks occurred and the resulting groups of blocks were analysed separately in *BEAST, as before. We also compared these analyses with two other approaches that do not use any paralogy or hybridisation test. Firstly, we ran a 12 gene *BEAST analysis (i.e., with gene trees assumed to be unlinked and thus independent due to recombination among loci). Secondly, we ran a concatenation of the 12 genes in BEAST. In both approaches each gene was endowed with its own substitution and clock model.

RESULTS

SIMULATED DATA

Paralogy Test

We explored the effect of two variables on the ability to detect paralogy with our method, namely: 1) the amount of topological variation introduced by a gene duplication with subsequent random losses of one paralogous locus, such that all taxa returned to single copy status, and 2) the length of a reference branch measured in CU units.

The power of the method increased with the depth of the tree (Fig. 3). For settings with large paralogy induced topological variation (RF=8) we found that lowering the detection critical value was effective to identify the paralogy-affected gene (Supp. Fig. 4). At lower cv, type II error (failure to detect real paralogy) decreased and the rate of successful detection of genuine paralogy increased (Supp. Fig. 4). This occurred with only a minor increase in type I error (either due to paralogue misidentification or to the incorrect exclusion of the whole genomic block). For the shortest tree (CU=2) at 95% cv, type I error ranged from zero to ~10% for block exclusion and ~5% for misidentification (Supp. Fig. 4). At 65% cv and for the shortest tree (CU=2), successful detection of

paralogy-affected genes in the RF=8 bin dropped to ~ 40% (Fig. 3), highlighting that shallow tree depth is the most challenging situation for coalescent-based methods (see also Maddison and Knowles 2006). Without paralogy-affected genes present in a block, the test performed well at all cv and CUs, showing less than 10% type I error in the most difficult case (CU=2) at 65% cv (Fig. 4).

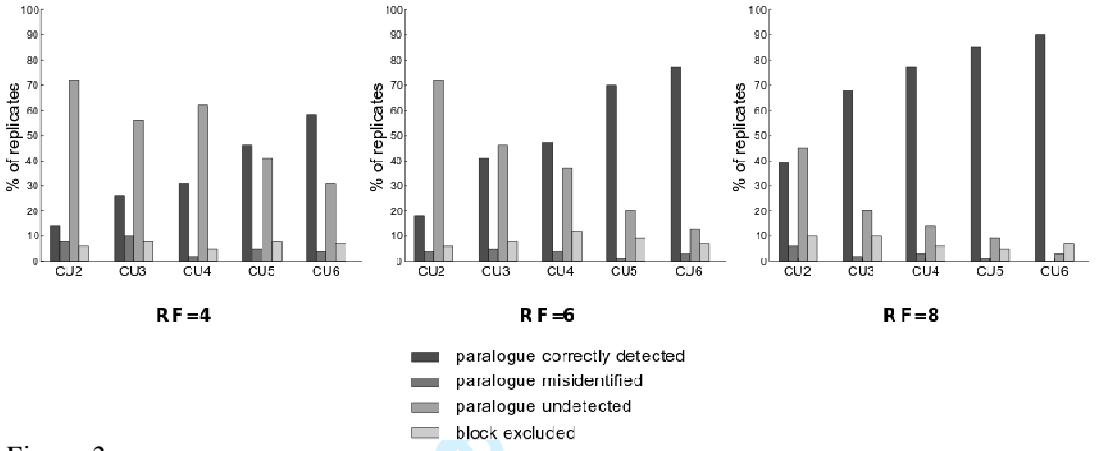


Figure 3.

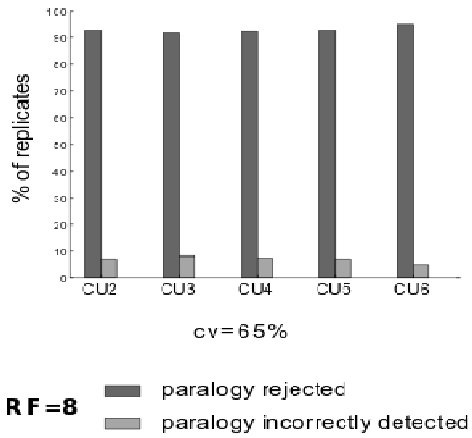


Figure 4.

Hybridisation Test

We focus on the results for the paralogy setting that introduces the most severe discordance (RF=8). The hybridisation test, like the paralogy test, performed better as the cv was lowered (Supp. Fig. 5). The rate of correct hybridisation detection at 65% cv approached 95% in the most favourable cases, CU=6, (Fig. 5) and was still of ~ 75% for the most difficult case, CU =2 (Fig. 5), with the failure rate to detect hybridisation not exceeding 10%. The false detection of hybridisation between blocks that share the same PT, occurred at less than 10% at any cv.

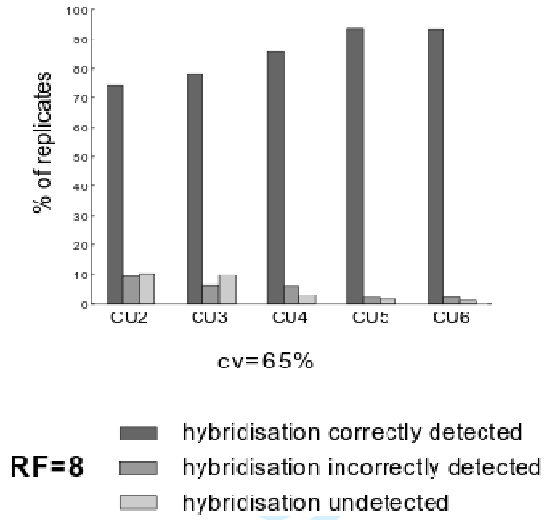


Figure 5.

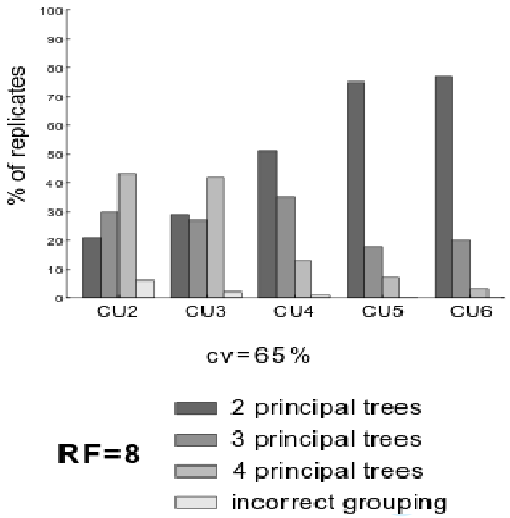
Grouping of Genomic Blocks

Blocks containing genes that differed due to ILS alone (in the hybridisation test) were assumed to be derived from the same origin. Genes in these blocks were then pooled for species tree inference. In our simulation settings, with four genomic blocks tracking two PTs, the hybridisation test can potentially output one to four PTs. Failure to recover the two initial grouping occurs when blocks are incorrectly clustered or wrongly segregated. Lower critical values produced more accurate results (Supp. Fig. 6). At 65% cv and CU=6, genomic blocks were correctly pooled into two groups corresponding to the two initial PTs, in over 75% of cases (Fig. 6). At this cv, the rate of wrong segregation increased at lower CUs (Fig. 6). However, the incorrect grouping of blocks that were simulated from different PTs – which represents the most severe kind of error – was only ~ 5% in the most challenging case at the lowest CU (Fig. 6).

In the more difficult cases, the method sometimes failed to group blocks originating from the same principal tree (i.e., it predicted either three or four PTs when only two PTs were present): this over-segregation rose from ~25% at CU=6 to ~75% at CU=2 across the 65% critical value cases in bin RF=8 (Fig. 6). Failure to group blocks from the same PT is not of major concern, however, as gene blocks that have been separated by our method can still cluster a posteriori when they are shown share the same tree in separate species tree analyses.

Species trees inferred from each group of blocks in replicates at CU=2 (RF=8 and cv=65%) recovered the hybrid taxon in the correct parental lineage in 100% of cases, with full branch support in 79% of cases and at the correct position in the parental lineage in 64% of the trees (not shown). Three out of 32 analyses did not converge for any of the three chains and were dismissed. The exact topology of the input PT for the gene simulation was recovered in 45% of cases. Both PTs were

493 recovered in eight out of 10 replicates, with full support in seven replicates. Failure to recover both
494 PTs in two replicates resulted from lack of convergence of the MCMC.



495 Figure 6.

496

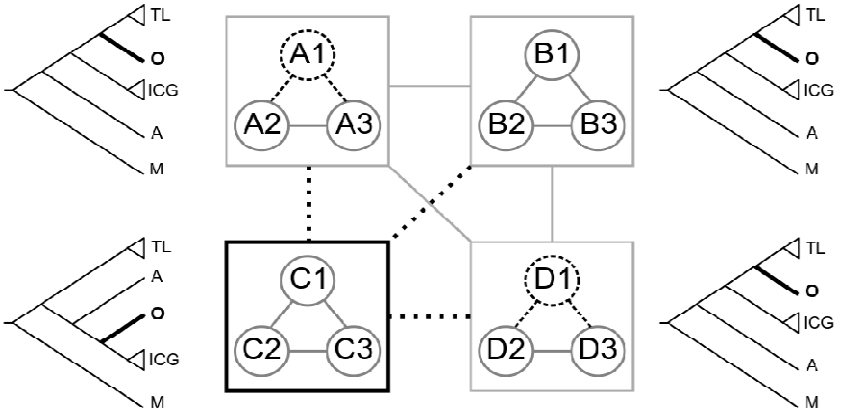
497 *MEDICAGO ORBICULARIS* DATASET

498 BEAST trees for the individual markers (not shown) revealed alternative supported
499 placements (posterior probability –PP– >95%) of *M. orbicularis*: either 1) within the (*M. truncatula*
500 - M. *littoralis* clade) (genes A1, A3, B1, B2 and D3) or 2) within the (*M. intertexta* - *M. ciliaris* - *M.*
501 *granadensis*) clade (genes C1, C2). A2 showed no supported placement for *M. orbicularis*. For B3,
502 the topology was the same as for the other genes of block B, but with marginal support for the
503 branching of *M. orbicularis*. In C3, *M. orbicularis* was placed with low support as sister to the *M.*
504 *truncatula* clade; this relationship was not recovered in C1 or C2. In D1, *M. orbicularis* appeared as
505 sister to the remaining *Medicago* species and in D2 it was placed as sister to the *M. intertexta* clade,
506 albeit with no support in either case. The (*M. truncatula* - *M. littoralis*) clade was always retrieved,
507 whereas *M. granadensis* was separated from *M. intertexta* and *M. ciliaris* in gene A2. *Medicago*
508 *arabica* was recovered, with good support, as sister to *M. orbicularis* (A1), to (*M. orbicularis* + *M.*
509 *truncatula*) (B2), to *M. granadensis* (B1, C3, D1) and to the remaining *Medicago* species (B3, C2,
510 D2). Thus, we observed a clear lack of concordance between several well supported topologies.

511 We applied our paralogy and hybridisation detection method to the *Medicago* data at critical
512 values of 95%, 85% and 75%. We obtained the exclusion of two genes (A1, D1) as probably
513 paralogy-affected, and the recovery of two groups of blocks (blocks A, B and D versus block C)
514 tracking two different principal trees. With a cv of 65%, the paralogy test excluded block B. When
515 species trees were inferred from each block, the topology alternated between one shared by block
516 A, B and D (*M. orbicularis* + *truncatula* clade) and a different one recovered with C (*M.orbicularis*

517 + *ciliaris* clade) (Fig. 7; Supp. Fig. 7). This result was concordant with the outcome of the
518 hybridisation test.

519 The grouping of blocks A, B and D was recovered in the hybridisation test for all cvs above
520 75% and was in agreement with the species tree topologies obtained from each block after the
521 exclusion of paralogy-affected genes (Fig 7). Since our simulation demonstrated a higher type I
522 error at 65% cv, we opted for keeping B in the (A, D) group. We also estimated the depth of this
523 hybridisation problem, based on individual gene trees, to be up to eight CUs (not shown), a tree
524 depth for which the paralogy-test performed well in the simulated data at cvs above 75%.



525 Figure 7.
526
527 The species tree inferred from group (A, B, D) (Supp. Fig. 7) displayed the same topology
528 as the one obtained from its individual blocks. Whereas the position of *M. orbicularis* was poorly
529 supported in the individual block analyses, its support increased in the group analysis, reaching a PP
530 of 0.94. For block C the support for the alternative position of *M. orbicularis* was high (PP=0.93).
531 *M. orbicularis* branched as sister to the *M. truncatula* clade in five genes across three blocks, and as
532 sister to the *M. intertexta* clade in two genes from one block. These alternative relationships were
533 also obtained in the species tree (after exclusion of paralogy-affected genes) when we analysed the
534 individual blocks and the grouped blocks. The position of *M. arabica* also differed between the two
535 recovered species trees: on the tree inferred from group (A, B, D), *M. arabica* appeared as sister to
536 all other *Medicago* species (PP=0.71), and on the tree inferred from block C, it branched as sister to
537 the (*M. orbicularis* + *M. intertexta*) clade (PP=0.93). On the gene trees, this species was involved in
538 four different well supported topologies and no block contained at least two genes that tracked the

same history for this taxon. When *M. arabica* was removed from the analyses, the position of *M. orbicularis* was unchanged in the two trees obtained from blocks (A, B, D) and C (not shown).

When the hybridisation test was performed without filtering for paralogy-affected genes, the block groupings and species trees inference steps were adversely affected. Block A was separated from B and D, which grouped together as before (Supp. Fig. 7). Block A alone returned a different topology than that obtained from (A, B, D), namely that *M. orbicularis* was weakly supported as sister to *M. arabica*, and together these were placed as sister to the *truncatula* clade with marginal support (Supp. Fig. 7); B and D returned the same position for *M. orbicularis* as before, but with lower support (PP=0.84) than in the (A, B, D) species tree with paralogy-affected genes removed. (Supp. Fig. 7). The result from C was unchanged.

Gene A1 appeared to drive the placement of *M. orbicularis* as sister to *M. arabica* in the block A, as it is the only gene in the block displaying this relationship, while the co-analysis of A2 and A3 inferred *M. orbicularis* as sister to the *truncatula* clade, but with lower support (0.71; Supp. Fig. 7). In fact, each of blocks A, B and D alone and together (with suspected paralogy-affected genes removed in each case) recovered this relationship, although with weak support from individual blocks (Supp. Fig. 7). If *M. orbicularis* sister grouping to the *truncatula* clade is correct (as it appears to be), then including suspected paralogues degraded support, negatively impacted the block grouping and returned a block with a different topology. What was a well-supported relationship became poorly supported or weakly contradicted to the extent that no clear interpretation could be obtained from this group of genes. With respect to hybrid origins, one parentage was recovered as before (block C), but the other parental lineage was not discovered. Thus, hybridisation could not be inferred from the *BEAST analyses *unless* the paralogy test had been performed and suspected paralogy-affected genes removed.

The analysis of 12 separated genes in *BEAST recovered the same topology as from blocks (B, D) (Supp. Fig. 7). However, the supports for including *M. orbicularis* in the *M. truncatula* clade and *M. arabica* in the *M. ciliaris* clade, were lower (PPs varied from 0.84 to 0.75, and 0.55 to 0.52, respectively). The concatenation of all 12 genes in BEAST returned a topology identical to blocks (A, B, D), but with 1.0 PP on each node, and again *Medicago orbicularis* was sister to the *M. truncatula* clade (Supp. Fig. 7). Both methods suggested a single possible relationship compatible with one of the parental origins, but failed to provide evidence for a hybridization event.

DISCUSSION

RECOVERY OF DIFFERENT PRINCIPAL TREES IN *MEDICAGO* AFTER PARALOGY AND HYBRIDISATION TESTING

573 Applying our method to the *Medicago* dataset resulted in the recovery of two principal trees.
574 Our results corroborate the initial hypothesis that hybridisation occurred in the *Medicago*
575 *orbicularis* lineage, as it is recovered in two clear alternative positions in the two principal trees.
576 The paralogy detection step indicated the presence of two putatively paralogy-affected genes: In
577 gene D1, *Medicago orbicularis* branched deeper than in any other gene, which is indeed consistent
578 with a hypothesis of paralogy. The other putative paralogy-affected gene, A1, recovered *M. arabica*
579 sister to *M. orbicularis* with good support at a shallow position. This would be more consistent with
580 recent hybridisation between these two species than with paralogy, but the gene was nevertheless
581 effectively identified within the respective genomic block as a highly-discordant gene tree.
582 Removal of suspected paralogy-affected genes was critical to correctly infer the parentage of the
583 hybrid lineage and recover the two principal trees. The hybridisation detection step indicated that
584 block C evolved under a different principal tree from the remaining blocks and species tree
585 reconstruction confirms the results of the hybridisation test: the tree recovered from block C
586 recovered *M. orbicularis* as sister to the *M. ciliaris* clade, whereas the remaining blocks
587 (individually or together) always placed *M. orbicularis* as sister to the *M. truncatula* clade. This
588 result shows that the use of genomic location can reveal hybridisation signal that would otherwise
589 be overlooked. None of the *Medicago orbicularis* alleles were identical to any of the alleles in the
590 two putative parental lineages, nor appeared at shallow positions within these lineages, which
591 suggests that the two principal trees obtained through our tests are compatible with a hypothesis of
592 ancient (rather than recent) hybridisation involving the *M. orbicularis* lineage. Six previously
593 published dated gene trees suggest that the hybridisation episode(s) probably took place more than
594 1.7 Ma years ago (Sousa et al. 2016).
595
596 *Taxon and Gene Sampling Effects on the Method*
597 One limitation of the metric used in the current method is that the Robinson-Foulds distance
598 could return only a minimum value (RF=2) for an unstable taxon placed at alternative positions on
599 two sister lineages, a situation that could be due to hybridisation. Such a small value is attained
600 when no other taxon branches between the unstable taxon and the divergence of the two sister
601 lineages. Our method would likely mistake this small distance for the effect of ILS in almost any
602 realistic example. To overcome this limitation, after we identified an unstable taxon as a putative
603 hybrid, it was necessary to sample a species sister to the putative hybrid, or one branching earlier
604 from the parental branch. The two putative parental lineages of *M. orbicularis* are represented by
605 the *M. intertexta* clade and the *M. truncatula* clade in our sample.
606 We included *M. arabica* in our sampling as this species appeared to be related to *M.*

orbicularis in previous phylogenies (e.g., Sousa et al. 2016) and was expected to increase the RF distance between alternative trees with respect to the *M. orbicularis* placement. However, the placement of *M. arabica* in the two recovered principal trees is also unstable, which suggests that *M. arabica* may also be involved in a hybridisation process. On the tree recovered from blocks (A, B, D), the low support for the placement of *M. arabica* (PP=0.71) may be caused by conflicting hybridisation signal in *M. arabica* among the three grouped blocks. This alternative placement of *M. arabica* did not affect the grouping of blocks as the corresponding topological difference, given the current sampling, is minimal (RF=2). Testing for hybridisation in *M. arabica* would, therefore, require a different sampling of taxa. This illustrates another limitation of our method, namely that only one independent hybridisation history can be effectively detected at a time. This limitation is due to the fact that genomic blocks can trace different hybridisation histories for different species, i.e., the genomic blocks tracing the same history in one lineage may trace more than one history in another lineage. If hybridisation occurs in more than one sampled lineage, our method may actually fail to group any genomic blocks. In such cases, a much higher number of genomic blocks need to be sampled in order to increase the chances of obtaining all principal trees that describe all the hybridisation events. Alternatively, lineages of suspected hybrid origin can be included one by one in analyses involving other lineages that do not appear to have hybrid histories (Sousa et al. 2016). Individual signals of hybridisation should be much more easily recovered and interpreted than overlapping signals.

Modelling Nuisance Parameters and Coalescent Stochasticity

We added complexity to our simulation with several sources of variation, such as substitution model variation and model selection uncertainty, substitution rate variation, gene branch rate variation, gene tree inference uncertainty, coalescent stochasticity and random gene copy loss after gene duplication. Thus, the simulations covered a realistic and challenging range of parameters for our method. The effect of coalescent stochasticity on tree-distance distributions, which may contain a tail of extreme values driven by very improbable topologies, needs to be overcome by choosing an adequate cv for the one-tailed tests. At high cv, the tests lose power, even at high CUs. This implies that previous applications of the test (for hybrid detection only) were probably conservative, i.e., favoured the ILS null (Maureira-Butler et al. 2008; Blanco-Pastor et al. 2012; Ramadugu et al. 2013). We found that decreasing the cv to 65% resulted in a decrease of type II error without a proportional increase of type I error.

Relevance of Paralogy Detection

If the proportion of paralogy-affected genes in a sample is low relative to non-affected genes, undetected paralogy may only marginally impact species-tree inference, as enough orthologous genes should swamp the paralogous signal. However, if this proportion is high, which may be the case when many taxa but few genes are sampled, or when WGD has recently occurred in an ancestor, identifying paralogues becomes more important. In such cases our method is able to detect paralogy-affected genes that would be the most detrimental for species tree inference.

Importantly, paralogy caused by WGD could likely also be detected by our method. If the reference organism used to determine genomic location has genes X, Y, Z in a genomic block, WGD in a related lineage will produce two blocks with X, Y, Z and X', Y', Z'. With random gene loss in each descendent, the gene trees may differ significantly. For example, a gene tree containing both Z/Z' copies among taxa (but only one copy sampled per taxon) can carry a markedly different signal than the tree produced from pure X and Y sequences (e.g., if the X' and Y' copies were lost before taxon divergence). If the sequences are recovered using primers or gene capture, the fact that some of these gene loci do not reside in the same physical block will be undetected, and would be handled as linked loci (in the absence of further information, i.e., for the majority of samples that would lack a physical map). The gene tree containing a mixture of Z/Z' sequences could be identified by our paralogy-test as being inconsistent with the other genes of the block, although the paralogous copies were not derived from tandem duplication but from WGD instead.

If a physical map for these genes were available, then the paralogy of Z' would be already known. So our test is applicable when the physical position of genes is well estimated in the ancestor of the whole lineage, but has been subsequently modified by WGD in some sub-lineages. This may be a fairly common situation, where a physical map is known for one model organism that is used to represent an entire clade that is not excessively old, which would imply small departure from the ancestral genomic map. Given how common WGD is in some lineages (e.g., flowering plants, Soltis et al. 2015 and references therein), our test may be especially useful for detecting WGD-derived paralogy.

Future Prospects

The use of genomic blocks to investigate hybridisation can be improved upon, and larger blocks containing a higher number of genes should be considered, albeit with caution. The genome-wide SNP study of several *Medicago* species by Yoder et al. (2013) is instructive with regard to this point. That study found that different genomic regions contained some strongly supported alternative trees compared to the entire genome-wide data set of over 82 000 SNPs. This is congruent with the idea that different genomic blocks contain different histories due to ancient

hybridisation. The genomic blocks used by Yoder et al. (2013) were fairly large (containing 500 SNPs spread over an average of 1.7 Mb of contiguous genomic sequence) and can clearly include many coalescent histories, given that linkage disequilibrium decays after only a few thousand nucleotides in *M. truncatula* (Branca et al. 2011). However, our results here (and with further taxon sampling, not shown) suggest that genomic blocks as small as 60 kb can contain more than one historical signal due to hybridisation. It is therefore possible that most of the large genomic regions used by Yoder et al. (2013) contain several hybridisation histories. This might explain why few large genomic blocks (defined as four consecutive windows, covering around 4.2 Mb) were found that displayed a single clear conflicting signal with the genome-wide tree. We contend that any genomic block as large or larger than 1.7 Mb (the smallest genomic block size use by Yoder et al. 2013) is highly likely to contain more than one parental origin. If so, then the sparse sampling of SNPs across these genomic blocks may not be the best strategy to unravel different histories due to ancient hybridisation. A much denser sampling of characters over the scale of 10s to 100s of kbp might be a better approach to infer resolved gene trees of much smaller genomic partitions that could be compared with other genomic partitions. This suggests that our test method can be improved by including, e.g., two, three or four times as many genes per genomic block, but extending the genomic blocks to maintain a similar density of sampled genes per length of DNA. The paralogy detection step in our methods could then serve to identify physical boundaries within a genomic block that separates two sets of genes with alternative hybrid histories.

Other future developments for our method should be centred on the use of alternative metrics that take into account other parameters besides tree topology, such as branch length or support values for each branch. A more sensitive metric would help overcome possible limitations caused by taxon sampling, which results in lack of power in gene tree pairwise comparisons. Other gene sampling approaches, which do not require such comprehensive a priori genomic information as the one available to us (a completely annotated genome), should also be pursued. Given the current developments in enrichment and sequencing techniques, it should not be difficult to implement sampling strategies that extend on ours, to include the sequencing of long contiguous regions that would enable the use of genomic location for a better understanding of complex phylogenetic problems.

CONCLUSIONS

Using linked and unlinked loci we found evidence of ancient hybridisation in *Medicago orbicularis*, confirming that the evolutionary history of this genus is affected by reticulation (Maureira-Butler et al. 2008). It is likely that introgressive hybridisation has occurred in many

lineages within the genus (Sousa et al. 2016). Thus, any attempt to represent the phylogeny of *Medicago* as a single dichotomous tree requires caution. The method used here to investigate incongruence in *Medicago* introduces a new angle to species tree inference in the presence of paralogy or other sources of highly discordant trees in addition to ILS and hybridisation. So far, phylogenetic inference methods have dealt with these two processes separately. Here, we tackle both processes sequentially and introduce the use of genomic location as an indicator of reticulate gene inheritance. Our approach does not rely on information about the (as yet) unknown species tree in order to test for paralogy and hybridisation. This is an advantage compared to other methods, because we do not need to infer, a priori, a species tree that might itself be misled by paralogy or hybridisation in order to assess the presence of misleading factors (e.g., Joly et al. 2009). We also do not need to specify in advance which lineages might be affected by hybridisation (unlike e.g., Kubatko et al. 2009; Jones et al. 2013). Finally, rather than attempting to reconcile paralogy- or hybridisation-affected tree topologies with the coalescent process (Rasmussen and Kellis 2012; Yu et al. 2013), we sort and deal with each process separately, as paralogy-affected genes are simply excluded and the remaining loci are analysed according to a shared history, allowing for a more complete perspective of speciation and diversification processes.

ACKNOWLEDGEMENTS

This work was supported by grants from the Swedish Research Council (grant 2009-5206), the Royal Swedish Academy of Sciences, Lars Hiertas Minne fund, The Royal Physiographic Society in Lund, Helge Ax:son Johnsons fund, and the Lundgrenska fund to B.E.P; from the P. A. Larssons fund and Lars Hiertas Minne fund to F.S.

REFERENCES

Ané C., Larget B., Baum D.A., Smith S.D., Rokas A. 2007. Bayesian estimation of concordance among gene trees. Mol. Biol. Evol. 24:412-426.

Blanc G., Wolfe K.H. 2004. Functional divergence of duplicated genes formed by polyploidy during *Arabidopsis* evolution. Pl. Cell 16:1679-1691.

Blanco-Pastor J.L., Vargas P., Pfeil B.E. 2012. Coalescent simulations reveal hybridization, incomplete lineage sorting in Mediterranean *Linaria*. PloS One 7:e39089.

Bena G. 2001. Molecular phylogeny supports the morphologically based taxonomic transfer of the "medicagoid" *Trigonella* species to the genus *Medicago* L. Pl. Syst. Evol. 229:217-236.

741 Bertrand Y.J.K., Scheen A.-C., Marcussen T., Pfeil B.E., Sousa F., Oxelman B. 2015. Assignment
742 of homoeologues to parental genomes in allopolyploids for species tree inference, with an example
743 from *Fumaria* (Papaveraceae). Syst. Biol. 64: 448-471.

744 Bosse M., Megens H.-J., Madsen O., Frantz L.A.F., Paudel Y., Croojimans R.P.M.A., Groenen
745 M.A.M. 2014. Untangling the hybrid nature of modern pig genomes: a mosaic derived from
746 biogeographically distinct, highly divergent *Sus scrofa* populations. Mol. Ecol. 23:4089-4102.

747 Branca A., Paape T.D., Zhou P., Briskine R., Farmer A.D., Mudge J., Bharti A.K., Woodward J.E.,
748 May G.D., Gentzbittel L., Ben C., Denny R., Sadowsky M.J., Ronfort J., Bataillon T., Young N.D.,
749 Tiffin P. 2011. Whole-genome nucleotide diversity, recombination,, linkage disequilibrium in the
750 model legume *Medicago truncatula*. Proc. Natl. Acad. Sci. U.S.A. 108:E864-E870.

751 Buckley T.R., Cordeiro M., Marshall D.C., Simon C. 2006. Differentiating between hypotheses of
752 lineage sorting, introgression in New Zealand alpine cicadas (*Maoricicada* Dugdale). Syst. Biol.
753 55:411-425.

754 Buerkle C.A., Rieseberg L.H. 2008. The rate of genome stabilization in homoploid hybrid species.
755 Evolution 62:266-275.

756 Carstens B.C., Knowles L.L. 2007. Estimating species phylogeny from gene-tree probabilities
757 despite incomplete lineage sorting: an example from *Melanoplus* grasshoppers. Syst. Biol. 56:400-
758 411.

759 Chen J., Huang Q., Gao D., Wang J., Lang Y., Liu T., Li B., Bai Z., Luis Goicoechea J., Liang C.,
760 Chen C., Zhang W., Sun S., Liao Y., Zhang X., Yang L., Song C., Wang M., Shi J., Liu G., Liu J.,
761 Zhou H., Zhou W., Yu Q., An N., Chen Y., Cai Q., Wang B., Liu B., Min J., Huang Y., Wu H., Li
762 Z., Zhang Y., Yin Y., Song W., Jiang J., Jackson S.A., Wing R.A., Wang J., Chen M. 2013. Whole-
763 genome sequencing of *Oryza brachyantha* reveals mechanisms underlying *Oryza* genome
764 evolution. Nat. Commun. 4:1595 doi: 10.1038/ncomms2596.

765 Cock P.J, Antao T., Chang J.T., Chapman B.A., Cox C.J., Dalke A., Friedberg I., Hamelryck T.,
766 Kauff F., Wilczynski B., de Hoon M.J. 2009. Biopython: freely available Python tools for
767 computational molecular biology, bioinformatics. Bioinformatics 25:1422-1423.

768 Cotton J.A., Page R.D. 2005. Rates, patterns of gene duplication, loss in the human genome. Proc.
769 Roy. Soc. London, Ser. B, Biol. Sci. 272:277-283.

770 Cui L., Wall P.K., Leebens-Mack J.H., Lindsay B.G., Soltis D.E., Doyle J.J., Soltis P.S., Carlson
771 J.E., Arumuganathan K., Barakat A., Albert V.A., Ma H., dePamphilis C.W. 2006. Widespread
772 genome duplications throughout the history of flowering plants. Genome Res. 16:738-749.

773 Dalcín L., Paz R., D'Elia M.S.J. 2008. MPI for Python: Performance improvements, MPI-2
774 extensions. J. Parallel Distrib. Comput. 68:655-662.

775 Darriba D., Taboada G.L., Doallo R., Posada D. 2012. jModelTest 2: more models, new heuristics,
776 parallel computing. Nature methods 9:772-772.

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

779 Degnan J.H., Salter L.A. 2005. Gene tree distributions under the coalescent process. Evolution
780 59:24-37.

781 Doyle J.J. 1992. Gene trees and species trees: molecular systematics as one-character taxonomy.
782 Syst. Bot. 17:144-163.

783 Drummond A.J., Rambaut A. 2007. BEAST: Bayesian evolutionary analysis by sampling trees.
784 BMC Evol. Biol. 7:214.

785 Edwards S.V., Liu L., Pearl D.K. 2007. High-resolution species trees without concatenation. Proc.
786 Natl. Acad. Sci. U.S.A. 104:5936-5941.

787 Foxe J.P., Slotte T., Stahl E.A., Neuffer B., Hurka H., Wright S.I. 2009. Recent speciation
788 associated with the evolution of selfing in *Capsella*. Proc. Natl. Acad. Sci. U.S.A. 106:5241-5245.

789 Good J.M., Vanderpool D., Keeble S., Bi K. 2015. Negligible nuclear introgression despite
790 complete mitochondrial capture between two species of chipmunks. Evolution 69: 1961-1972.

791 Gossmann T.I., Song B.-H., Windsor A.J., Mitchell-Olds T., Dixon C.J., Kapralov M.V., Filatov
792 D.A., Eyre-Walker A. 2010. Genome wide analyses reveal little evidence for adaptive evolution in
793 many plant species. Mol. Biol. Evol. 27:1822–1832.

794 Green R.E., Krause J., Briggs A.W., Maricic T., Stenzel U., Kircher M., Patterson N., Li H., Zhai
795 W., Fritz M.H., Hansen N.F., Durand E.Y., Malaspinas A.S., Jensen J.D., Marques-Bonet T., Alkan
796 C., Prüfer K., Meyer M., Burbano H.A., Good J.M., Schultz R., Aximu-Petri A., Butthof A., Höber
797 B., Höffner B., Siegemund M., Weihmann A., Nusbaum C., Lander E.S., Russ C., Novod N.,

798 Affourtit J., Egholm M., Verna C., Rudan P., Brajkovic D., Kucan Z., Gusic I., Doronichev V.B.,
799 Golovanova L.V., Lalueza-Fox C., de la Rasilla M., Fortea J., Rosas A., Schmitz R.W., Johnson
800 P.L., Eichler E.E., Falush D., Birney E., Mullikin J.C., Slatkin M., Nielsen R., Kelso J., Lachmann
801 M., Reich D., Pääbo S. 2010. A draft sequence of the Neandertal genome. Science 328:710-722.

802 Guschanski K., Krause J., Sawyer S., Valente L.M., Bailey S., Finstermeier K., Sabin R., Gilissen
803 E., Sonet G., Nagy Z.T., Lenglet G., Mayer F., Savolainen V. 2013. Next-generation museomics
804 disentangles one of the largest primate radiations. Syst. Biol. 62: 539–554.

805 Hagberg A., Swart P., Chult D. 2008. Exploring network structure, dynamics, function using
806 networkx. Techical report. Los Alamos, NM: Los Alamos National Laboratory (LANL).

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

809 Hermansen J.S., Saether S.A., Elgvin T.O., Borge T., Hjelle E., Sætre, G.P. 2011. Hybrid speciation
810 in sparrows I: phenotypic intermediacy, genetic admixture, barriers to gene flow. Molec. Ecol.
811 20:3812-3822.

812 Holland B.R., Benthin S., Lockhart P.J., Moulton V., Huber K.T. 2008. Using supernetworks to
813 distinguish hybridization from lineage-sorting. BMC Evol. Biol. 8:202.

814 Huelsenbeck J.P., Larget B., Alfaro M.E. 2004. Bayesian phylogenetic model selection using
815 reversible jump Markov chain Monte Carlo. Molec. Biol. Evol. 21:1123-1133.

816 Huelsenbeck J.P., Ronquist F. 2005. Bayesian analysis of molecular evolution using MrBayes. In
817 Statistical methods in molecular evolution. New York: Springer. p. 183-226

818 Innes R.W., Ameline-Torregrosa C., Ashfield T., Cannon E., Cannon S.B., Chacko B., Chen
819 N.W.G., Couloux A., Dalwani A., Denny R., Deshpande S., Doyle J.J., Egan A., Geffroy V.,
820 Glover N., Hans C.S., Howell S., Ilut D., Jackson S., Lai H., Mammadov J., Martin del Campo S.,
821 Metcalf M., Nguyen A., O'Bleness M., Pfeil B.E., Podicheti R., Ratnaparkhe M.B., Roe B.A.,
822 Saghai Maroof M.A., Samain S., Sanders I., Ségurens B., Sévignac M., Sherman-Broyles S.,
823 Thareau V., Tucker D.M., Walling J., Wawrzynski A., Yi J., Young N.D. 2008. Differential
824 accumulation of retroelements, diversification of NB-LRR disease resistance genes in duplicated
825 regions following polyploidy in the ancestor of soybean. Plant Physiol. 148:1740-1759.

826 Joly S., McLenachan P.A., Lockhart P.J. 2009. A statistical approach for distinguishing

827 hybridization, incomplete lineage sorting. Am. Nat. 174:E54-E70.

828 Joly S., Pfeil B.E., Oxelman B., McLenachan P.A., Lockhart P.J. 2010. A statistical approach for
829 distinguishing hybridization, incomplete lineage sorting: correction. Am. Nat. 175:621-622.

830 Jones G., Sagitov S., Oxelman B. 2013. Statistical inference of allopolyploid species networks in
831 the presence of incomplete lineage sorting. Syst. Biol. 62:467-478.

832 Kingman J.F.C. 1982. The coalescent. Stochastic processes, their applications 13:235-248.

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

835 Kubatko, L. S. 2009. Identifying hybridization events in the presence of coalescence via model
836 selection. Syst. Biol. 58:478-488.

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

839 Lavin M., Herendeen P.S., Wojciechowski M.F. 2005. Evolutionary rates analysis of Leguminosae
840 implicates a rapid diversification of lineages during the Tertiary. Syst. Biol. 54:575-594.

841 Librado P., Rozas J. 2009. DnaSP v5: a software for comprehensive analysis of DNA
842 polymorphism data. Bioinformatics 25:1451-1452.

843 Li H., Handsaker B., Wysoker A., Fennell T., Ruan J., Homer N., Marth G., Abecasis G., Durbin
844 R., 1000 Genome Project Data Processing Subgroup. 2009. The sequence alignment/map format,
845 SAMtools. Bioinformatics 25:2078-2079.

846 Liu K., Dai J., Truong K., Song Y., Kohn M.H., Nakhleh L. 2014. An HMM-based comparative
847 genomic framework for detecting introgression in eukaryotes. PLoS Comp. Biol. 10:e1003649

848 Liu K.J., Steinberg E., Yozzo A., Song Y., Kohn M.H., Nakhleh L. 2015. Interpecific introgressive
849 origin of genomic diversity on the house mouse. Proc. Natl. Acad. Sci. U.S.A. 112:196-201.

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

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

Lundemo S., Falahati-Anbaran M., Stenøien H.K. 2009. Seed banks cause elevated generation times, effective population sizes of *Arabidopsis thaliana* in northern Europe. Mol. Ecol. 18:1795-2811.

Lynch M., Connery J.S. 2000. The evolutionary fate, consequences of duplicate genes. Science 290:1151-1155.

Lynch M., Force A. 2000. The probability of duplicate gene preservation by subfunctionalisation. Genetics 154:459-473.

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

Maddison W.P., Knowles L.L. 2006. Inferring phylogeny despite incomplete lineage sorting. Syst. Biol. 55:21-30.

Maddison W.P., Maddison D.R. 2006. Mesquite: a modular system for evolutionary analysis, version 1.12.

Maere S., Bodt S.D., Raes J., Casneuf T., Montagu M.V., Kuiper M., Van de Peer, Y. 2005. Modelling gene, genome duplications in eukaryotes. Proc. Natl. Acad. Sci. U.S.A. 102:5454–5459.

Mallet J. 2007. Hybrid speciation. Nature 446:279-283.

Martin D.P., Lemey P., Lott M., Moulton V., Posada D., Lefeuvre P. 2010. RDP3: a flexible, fast computer program for analyzing recombination. Bioinformatics 26:2462-2463.

Maureira-Butler I.J., Pfeil B.E., Muangprom A., Osborn T.C., Doyle J.J. 2008. The reticulate history of *Medicago* (Fabaceae). Syst. Biol. 57:466-482.

Meng C., Kubatko L.S. 2009. Detecting hybrid speciation in the presence of incomplete lineage sorting using gene tree incongruence: A model. Theor. Pop. Biol. 75:35-45.

Mirarab S., Reaz R., Bayzid M.S., Zimmermann T., Swenson M.S., Warnow T. 2014. ASTRAL: genome-scale coalescent-based species tree estimation. Bioinformatics 30:i541-i548.

Nei M., Kumar S. 2000. Molecular evolution, phylogenetics. Oxford University Press, New York.

Nei M., Rooney A.P. 2005. Concerted, birth-and-death evolution of multigene families. Annual Rev. Genet. 39:121.

Oxelman B., Yoshikawa N., McConaughy B.L., Luo J., Denton A.L., Hall B.D. 2004. RPB2 gene

881 phylogeny in flowering plants, with particular emphasis on asterids. Mol. Phylogenet. Evol. 32:462-
882 479.

883 Pamilo P., Nei M. 1988. Relationships between gene trees, species trees. Mol. Biol. Evol. 5:568-
884 583.

885 Pérez-Collazos E., Segarra-Moragues J.G., Villar L., Catalán P. 2015. Ant pollination promotes
886 spatial genetic structure in the long-lived plant *Borderea pyrenaica* (Dioscoreaceae). Biol. J. Linn.
887 Soc. 116:144-155.

888 Peters J.L., Zhuravlev Y., Fefelov I., Logie A., Omland K.E. 2007. Nuclear loci, coalescent
889 methods support ancient hybridization as cause of mitochondrial paraphyly between gadwall,
890 falcated duck (*Anas* spp.). Evolution 61:1992-2006.

891 Phillips M.J., Haouchar D., Pratt R.C., Gibb G.C., Bunce M. 2013. Inferring kangaroo phylogeny
892 from incongruent nuclear, mitochondrial genes. PloS One 8:e57745.

893 Pollard D.A., Iyer V.N., Moses A.M., Eisen M.B. 2006. Widespread discordance of gene trees with
894 species tree in *Drosophila*: evidence for incomplete lineage sorting. PLoS Genet. 2:1634-1647.

895 Ramadugu C., Pfeil B.E., Manjunath K.L., Lee R.F., Maureira-Butler I.J., Roose M.L. 2013.
896 Coalescence simulation testing of hybridization versus lineage sorting in *Citrus* (Rutaceae) using
897 six nuclear genes. PloS One 8:e68410.

898 Rambaut A., Drummond A.J. 2007. Tracer v1.4.

899 Rambaut A., Grass N.C. 1997. Seq-Gen: an application for the Monte Carlo simulation of DNA
900 sequence evolution along phylogenetic trees. Computer applications in the biosciences: CABIOS
901 13:235-238.

902 Rannala B., Yang Z. 2003. Bayes estimation of species divergence times, ancestral population sizes
903 using DNA sequences from multiple loci. Genetics 164:1645-1656.

904 Rasmussen M.D., Kellis M. 2012. Unified modeling of gene duplication, loss,, coalescence using a
905 locus tree. Genome research, 22:755-765.

906 Reid N., Demboski J.R., Sullivan J. 2012. Phylogeny estimation of the radiation of western north
907 American chipmunks (*Tamias*) in the face of introgression using reproductive protein genes. Syst.
908 Biol. 61:44-62.

909 Rieseberg L.H. 1997. Hybrid origins of plant species. Annu. Rev. Ecol. Syst. 28:359-389.

910 Rieseberg L.H., Raymond O., Rosenthal D.M., Lai Z., Livingstone K., Nakazato T., Durphy J.L.,
911 Schwarzbach A.E., Donovan L.A., Lexer C. 2003. Major ecological transitions in wild suflowers
912 facilitated by hybridization. Science 301:1211-1216.

913 Robinson D., Foulds L. 1981. Comparison of phylogenetic trees. Math. Biosci. 53:131-147.

914 Rosenberg N.A. 2003. The shapes of neutral gene genealogies in two species: probabilities of
915 monophyly, paraphyly,, polyphyly in a coalescent model. Evolution 57:1465-1477.

916 Sanderson M.J., Doyle J.J. 1992. Reconstruction of organismal, gene phylogenies from data on
917 multigene families: concerted evolution, homoplasy, confidence. Syst. Biol., 41:4-17.

918 Sankararaman S., Mallick S., Dannemann M., Prufer K., Kelso J., Paabo S., Patterson N., Reich D.
919 2014. The genomic landscape of Neanderthal ancestry in present-day humans. Nature 507:354–357.

920 Sousa F., Bertrand Y.J.K., Nylinder S., Oxelman B., Eriksson J.S., Pfeil B.E. 2014. Phylogenetic
921 properties of 50 nuclear loci in *Medicago* (Leguminosae) generated using multiplexed sequence
922 capture, next-generation sequencing. PloS One 9:e109704.

923 Sousa F., Bertrand Y.J., Pfeil B.E. 2016. Patterns of phylogenetic incongruence in *Medicago* found
924 among six loci. Pl. Syst. Evol. 302:493-513.

925 Smith B.T., Harvey M.G., Faircloth B.C., Glenn T.C., Brumfield R.T. 2014. Target capture,
926 massively parallel sequencing of ultraconserved elements for comparative studies at shallow
927 evolutionary time scales. Syst. Biol. 63: 83–95.

928 Soltis P.S., Marchant D.B., Van de Peer Y., Soltis D.E. 2015. Polyploidy, genome evolution in
929 plants, 35:119-125.

930 Song Y., Endepols S., Klemann N., Richter D., Matuschka F.-R., Shih C.-H., Nachman M.W.,
931 Kohn M.H. 2011. Adaptive introgression of anticoagulant rodent poison resistance by hybridization
932 between Old World mice. Cur. Biol. 21:1296-1301.

933 Steele K.P., Wojciechowski M.F. 2003. Phylogenetic analyses of tribes Trifolieae, Vicieae, based
934 on sequences of the plastid gene matK (Papilionoideae: Leguminosae). Advances Legume Syst.
935 Part 10:355-370.

936 Steele K.P., Ickert-Bond S.M., Zarre S., Wojciechowski M.F. 2010. Phylogeny, character evolution

in *Medicago* (Leguminosae): Evidence from analyses of plastid trnK/matK, nuclear GA3ox1 sequences. Amer. J. Bot. 97:1142-1155.

Strasburg J.L., Rieseberg L.H. 2008. Molecular demographic history of the annual sunflowers *Helianthus annuus*, *H. petiolaris* - Large effective population sizes, rates of long-term gene flow. Evolution 62:1936-1950.

Stull G.W., Moore M.J., Mandala V.S., Douglas N.A., Kates H.R., Qi X., Brockington S.F., Soltis P.S., Soltis D.E., Gitzendanner M.A. 2013. A targeted enrichment strategy for massively parallel sequencing of angiosperm plastid genomes. Applications in Plant Sciences 1:1-7.

Sukumaran J., Holder M.T. 2010. DendroPy: a Python library for phylogenetic computing. Bioinformatics 26:1569-1571.

Syring J., Farrell K., Businsky R., Cronn R., Liston A. 2007. Widespread genealogical nonmonophyly in species of *Pinus* subgenus Strobus. Syst. Biol. 56:163-181.

Takahata N., Nei M. 1985. Gene genealogy, variance of interpopulational nucleotide differences. Genetics 110:325-344.

Trier C.N., Hermansen J.S., Sætre G.-P., Bailey R.I. 2014. Evidence for mito-nuclear and sex-linked reproductive barriers between the hybrid Italian sparrow and its parent species. PLoS Genetics 10: e1004075. doi:10.1371/journal.pgen.1004075

Ungerer M.C., Baird S.J.E., Pan J., Rieseberg L.H. 1998. Rapid hybrid speciation in wild sunflowers. Proc. Natl. Acad. Sci. U.S.A. 95:11757-11762.

Van Zee J.P., Schluter J.A, Schluter S., Dixon P., Brito Sierra C.A., Hill C.A. 2016. Paralog analyses reveal gene duplication events and genes under positive selection in *Ixodes scapularis* and other ixodid ticks. BMC Gen. 17:241 doi: 10.1186/s12864-015-2350-

Yoder J.B., Briskine R., Mudge J., Farmer A., Paape T., Steele K., Weiblen G.D., Bharti A.K., Zhou P., May G.D., Young N.D., Tiffin P. 2013. Phylogenetic signal variation in the genomes of *Medicago* (Fabaceae). Syst. Biol. 62:424-438.

Young N.D., Debelle F., Oldroyd G.E.D., Geurts R., Cannon S.B., Udvardi M.K., Benedito V.A., Mayer K F.X., Gouzy J., Schoof H., Van de Peer Y., Proost S., Cook D.R., Meyers B.C., Spannagl M., Cheung F., De Mita S., Krishnakumar V., Gundlach H., Zhou S., Mudge J., Bharti A.K., Murray J.D., Naoumkina M.A., Rosen B., Silverstein K.A.T., Tang H., Rombauts S., Zhao P.X.,

966 Zhou P., Barbe V., Bardou P., Bechner M., Bellec A., Berger A., Berges H., Bidwell S., Bisseling
967 T., Choisne N., Couloux A., Denny R., Deshpande S., Dai X., Doyle J.J., Dudez A., Farmer A.D.,
968 Fouteau S., Franken C., Gibelin C., Gish J., Goldstein S., Gonzalez A.J., Green P.J., Hallab A.,
969 Hartog M., Hua A., Humphray S.J., Jeong D., Jing Y., Jocker A., Kenton S.M., Kim D., Klee K.,
970 Lai H., Lang C., Lin S., Macmil S.L., Magdelenat G., Matthews L., McCorrison J., Monaghan E.L.,
971 Mun J., Najar F.Z., Nicholson C., Noirot C., O'Bleness M., Paule C.R., Poulain J., Prion F., Qin B.,
972 Qu C., Retzel E.F., Riddle C., Sallet E., Samain S., Samson N., Sanders I., Saurat O., Scarpelli C.,
973 Schiex T., Segurens B., Severin A.J., Sherrier D.J., Shi R., Sims S., Singer S.R., Sinharoy S., Sterck
974 L., Viollet A., Wang B., Wang K., Wang M., Wang X., Warfsmann J., Weissenbach J., White D.D.,
975 White J.D., Wiley G.B., Wincker P., Xing Y., Yang L., Yao Z., Ying F., Zhai J., Zhou L., Zuber A.,
976 Denarie J., Dixon R.A., May G.D., Schwartz D.C., Rogers J., Quetier F., Town C.D., Roe B.A.
977 2011. The *Medicago* genome provides insight into the evolution of rhizobial symbioses. Nature
978 480:520-524.

979 Yu Y., Barnett R.M., Nakhleh L. 2013. Parsimonious inference of hybridization in the presence of
980 incomplete lineage sorting. Syst. Biol. 62:738-751.

981

Figure Legends

Figure 1: Cartoon illustration of the overall test scheme. Top: the physical position of genes within genomic blocks and between genomic blocks is indicated on a chromosome. Middle: individual ultrametric gene trees (with sequences from species A–D at the gene tree tips) are inferred (step 1) and compared within blocks (step 2). Significant differences among them are explained by gene duplication (black rectangle in gene tree 1 on left) and losses (not shown). Paralogy-affected gene trees (dashed box on left around gene tree 1) are discarded. Bottom: compatible genes are grouped by a common origin from their parental source (if hybridisation has occurred) (step 3), both within blocks (shown) and among blocks (not shown). Incompatible blocks of genes (left-hand pair versus right-hand group of three genes in this example) are kept separate for principal tree estimation using the multi-species coalescent model (step 4).

Figure 2: Hypothetical species phylogeny with a hybrid taxon, taxon 4 (panel a), that can be decomposed into principal tree 1 (panel b) and principal tree 2 (panel c). Paralogous gene trees are obtained by removing taxa from the clade descending from a gene duplication event, such that each taxon retains only one copy of the duplicated gene (panel d). Thus, the mismatch between the gene tree (panel d) and the principal tree (panel b) is due to differential gene loss after the gene duplication event.

Figure 3: Paralogy test for bins RF=4, RF=6 and RF=8 at critical value 65%, for blocks that contain a single paralogy-affected gene. When the paralogy-affected gene is *misidentified*, another gene (not harbouring paralogues) is removed instead. When the paralogy-affected gene is *undetected*, all three genes in the block are retained for down-stream analyses. *Block excluded* refers to when all three gene trees show significant differences to one another and are all excluded, because no pair of compatible genes could be identified.

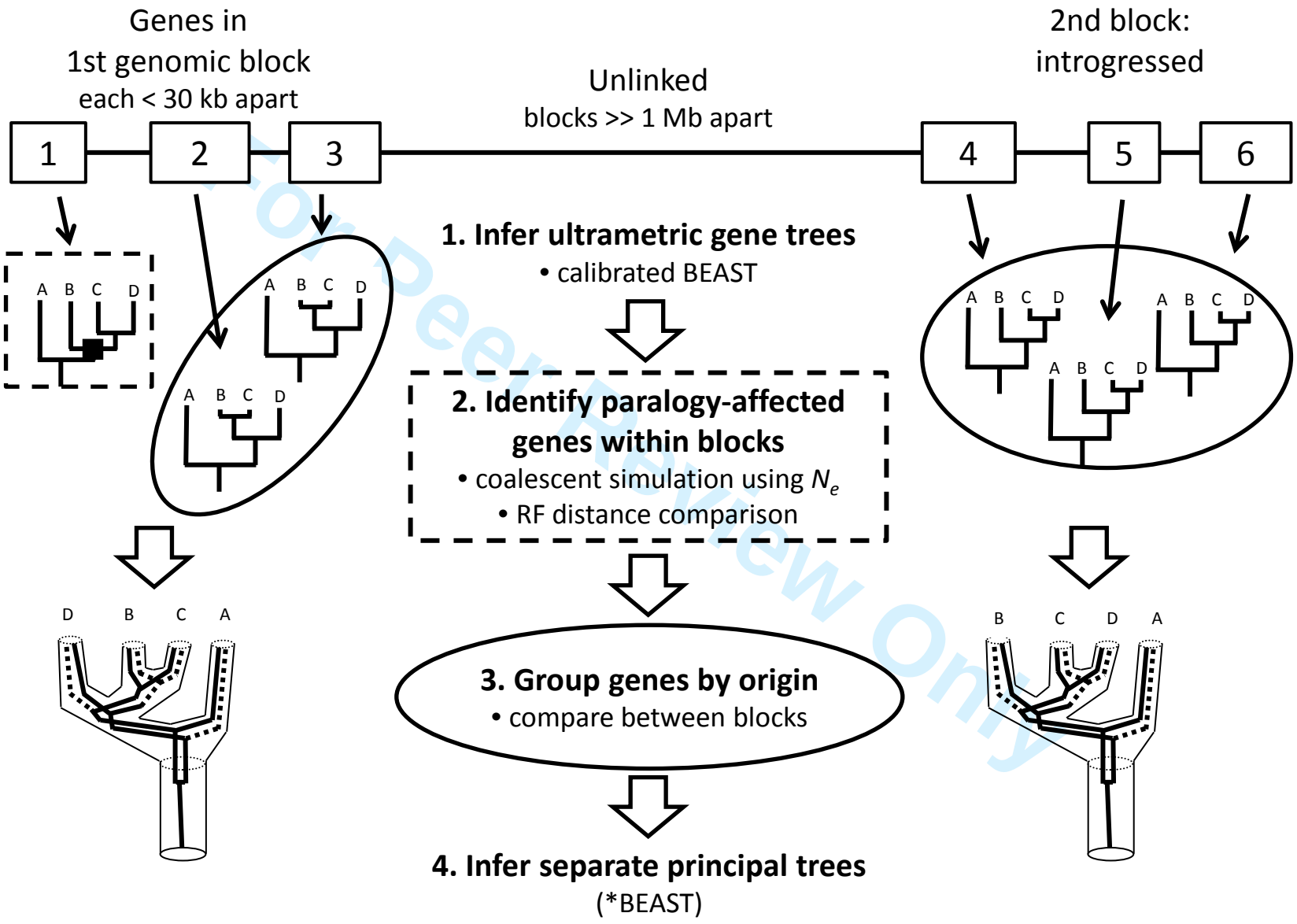
Figure 4: Paralogy test at critical value 65% for blocks without paralogy-affected genes. Paralogy *rejected* means that all three genes are correctly retained for down-stream analyses. Paralogy *incorrectly detected* means that one gene was falsely identified as paralogy-affected and incorrectly excluded from further analyses.

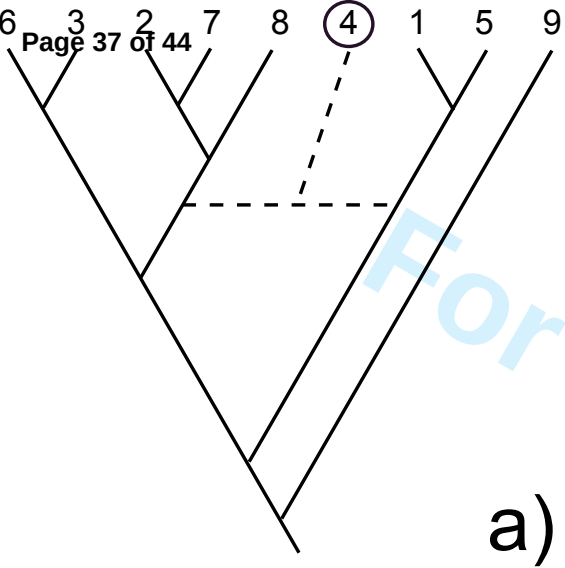
Figure 5: Hybridisation test using only those genes and blocks that were not identified as being paralogy-affected in the previous step (cv 65% paralogy bin RF=8). Both hybridisation and earlier

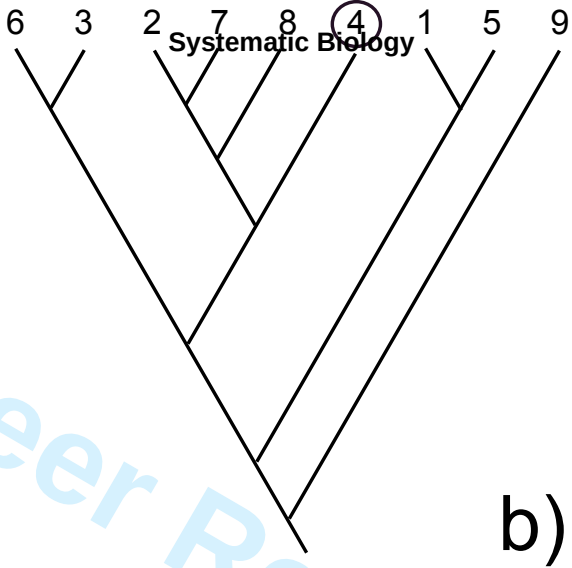
paralogy tests used the same critical value. Hybridisation *correctly detected* separates blocks of genes into the correct groups based on the principal tree of origin. When hybridisation is *incorrectly detected*, some separation of blocks of genes occurs (compatible with hybridisation), but the groupings are incorrect. *Undetected* hybridisation means that no separation of blocks of genes has occurred, even though two principal trees have generated the blocks of genes.

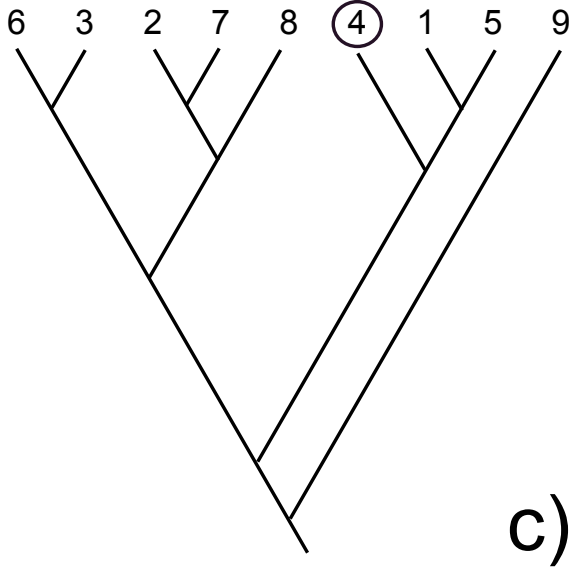
Figure 6: The number of principal trees found after the hybridisation test (incorporating the outcome of the paralogy test) for cv 65% and paralogy bin RF=8. Both tests use the same critical value. Correct detection (Fig. 5) sometimes results in the over-separation of genomic blocks into more than two implied principal trees, although each group contains blocks of genes derived from only one principal tree each. An *incorrect grouping* places blocks together than do not originate from the same single principal tree.

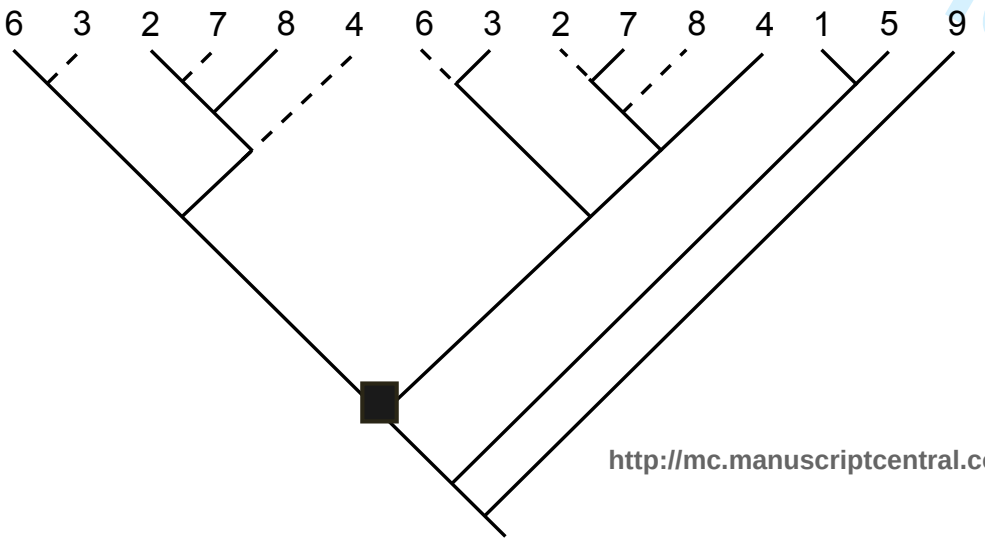
Figure 7: A schematic representation of the paralogy and hybridisation test results on genes from four genomic blocks sampled from *Medicago orbicularis* and relatives. Circles represent individual genes and dashed circles represent excluded genes (following the paralogy test). Squares represent genomic blocks. Dashed lines represent ILS rejection between genes within blocks (implying paralogy) and between blocks (implying hybridisation). Cladograms next to each block illustrate which of the two principal tree topologies were recovered with *BEAST for that block. Labels on cladograms correspond to the following taxa: TL – *M. truncatula*, *M. littoralis*; O – *M. orbicularis*; ICG – (*M. intertexta*, *M. ciliaris*, *M. granadensis*) clade; A – *M. arabica*; M – *Melilotus* (outgroup).

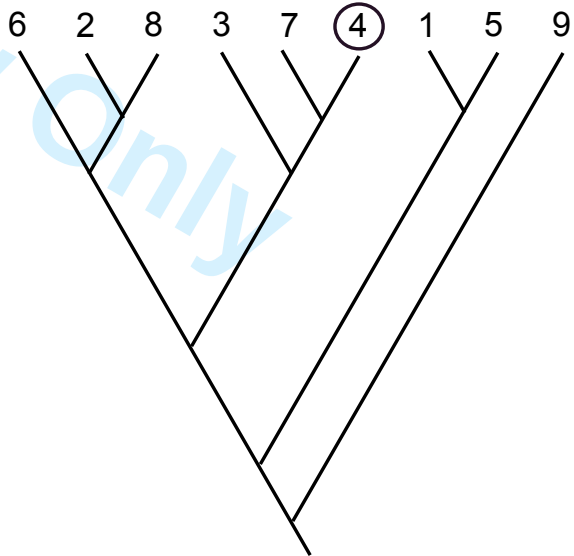


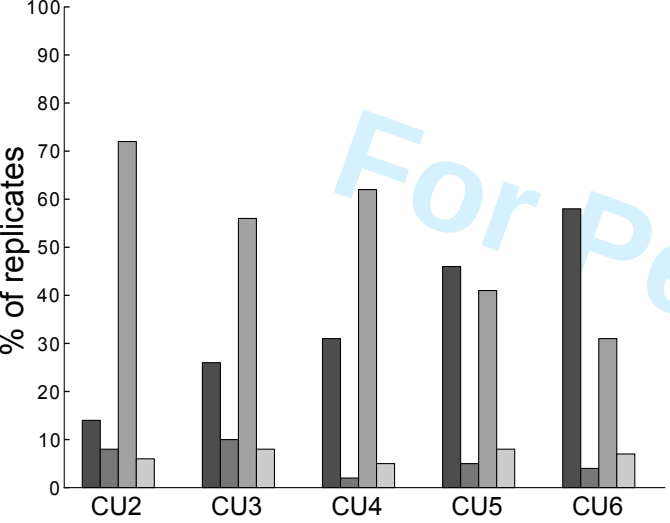




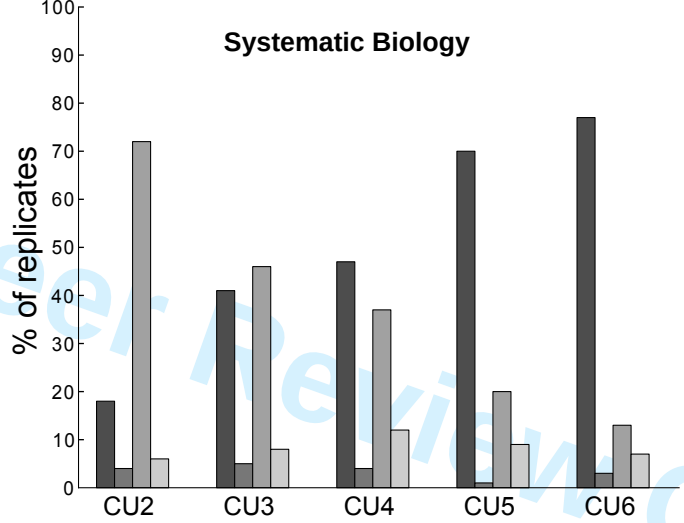




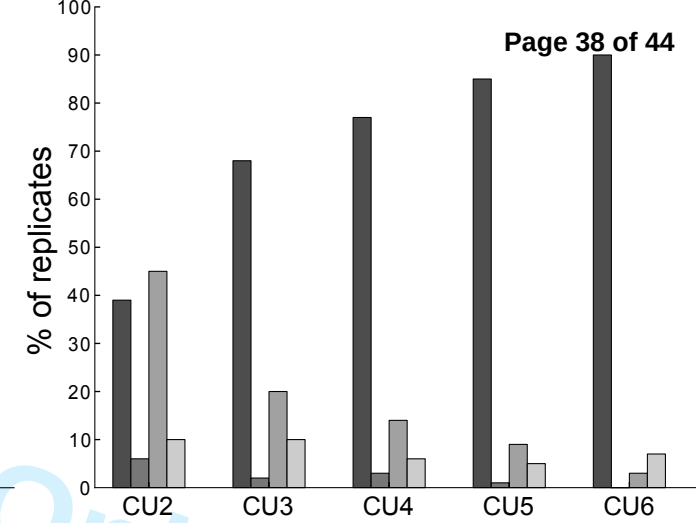




R F =4



R F =6



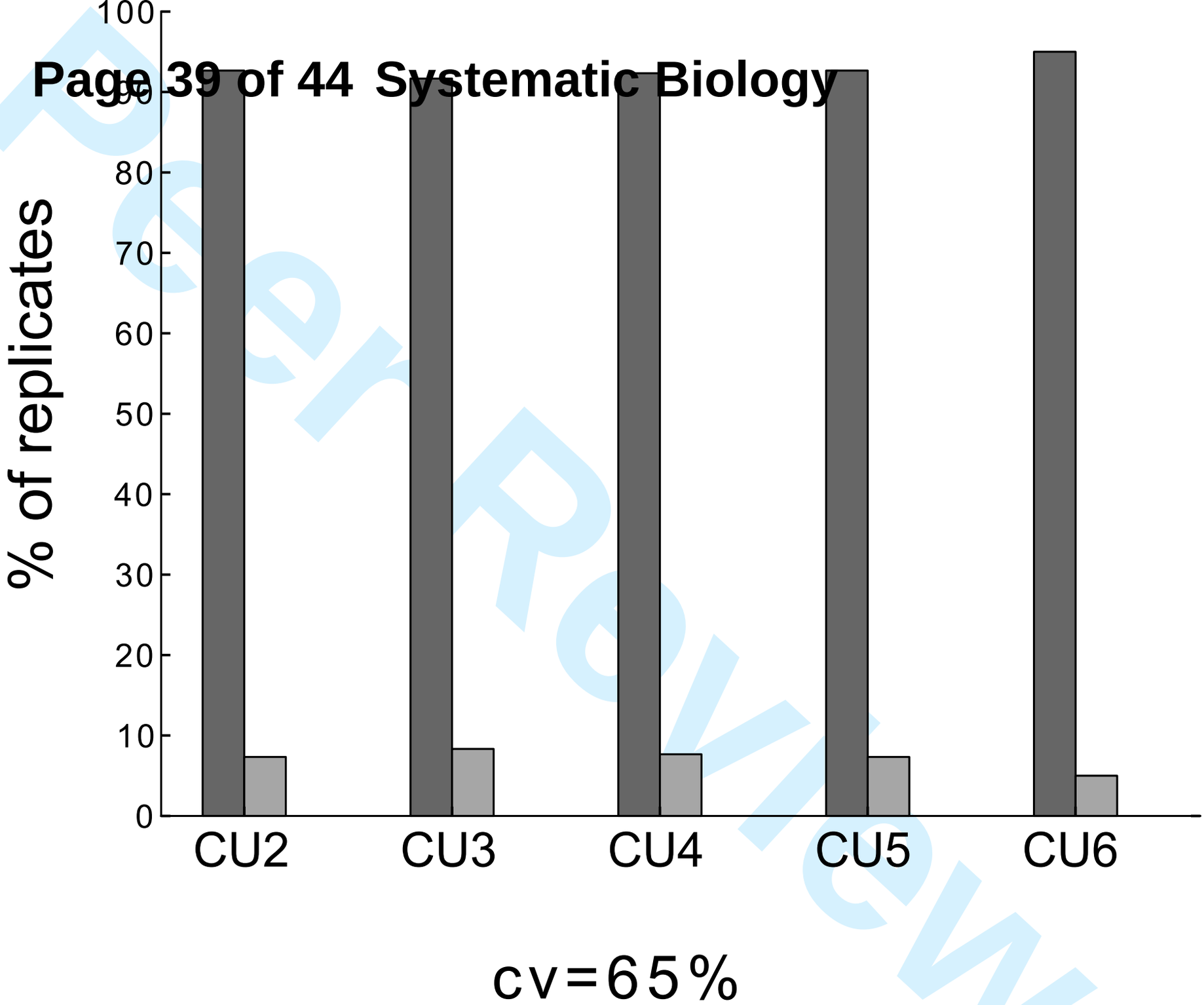
R F =8

paralogue correctly detected

paralogue misidentified

paralogue undetected

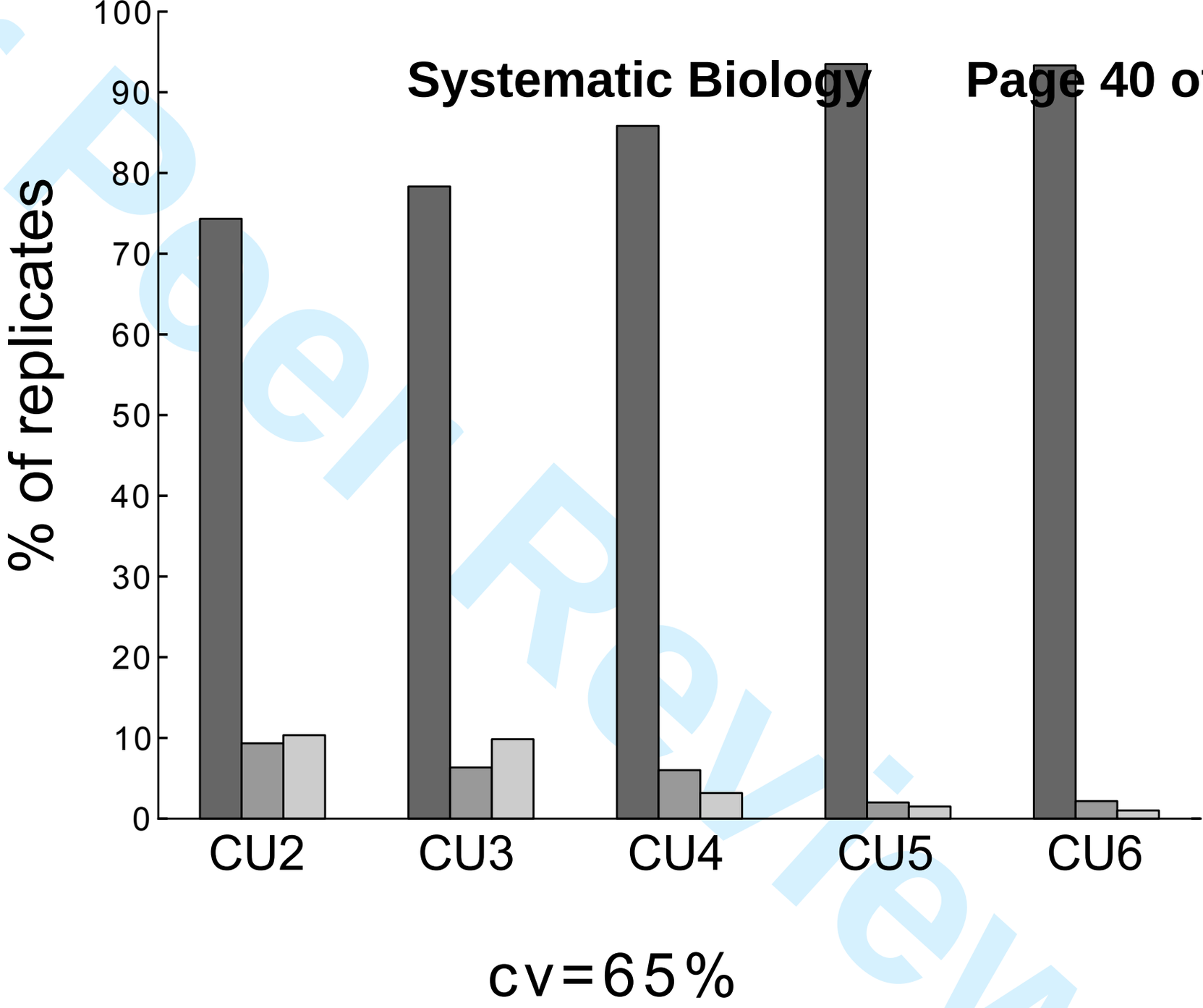
block excluded



paralogy rejected

paralogy incorrectly detected

RF=8



hybridisation correctly detected

hybridisation incorrectly detected

hybridisation undetected

RF=8