



**HAL**  
open science

# The Multilocus Multispecies Coalescent: A Flexible New Model of Gene Family Evolution

Qiuyi Li, Celine Scornavacca, Nicolas Galtier, Yao-Ban Chan

► **To cite this version:**

Qiuyi Li, Celine Scornavacca, Nicolas Galtier, Yao-Ban Chan. The Multilocus Multispecies Coalescent: A Flexible New Model of Gene Family Evolution. *Systematic Biology*, 2020, 10.1093/sysbio/syaa084 . hal-02972075

**HAL Id: hal-02972075**

**<https://hal.science/hal-02972075v1>**

Submitted on 2 Dec 2020

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

*Systematic Biology* (2020), 0, 0, pp. 1-38  
doi:10.1093/sysbio/output

# The Multilocus Multispecies Coalescent: A Flexible New Model of Gene Family Evolution

QIUYI LI<sup>1,†</sup>, CELINE SCORNAVACCA<sup>2,†</sup>, NICOLAS GALTIER<sup>2</sup>, AND YAO-BAN CHAN<sup>1,\*</sup>

<sup>1</sup> *School of Mathematics and Statistics / Melbourne Integrative Genomics, The University of Melbourne, Melbourne, 3010, Australia*

<sup>2</sup> *Institut des Sciences de l'Evolution, Université Montpellier, CNRS, IRD, EPHE, Montpellier, 34095, France*

\*[yaoban@unimelb.edu.au](mailto:yaoban@unimelb.edu.au)

† *These authors contributed equally to this work.*

## ABSTRACT

1 Incomplete lineage sorting (ILS), the interaction between coalescence and speciation, can  
2 generate incongruence between gene trees and species trees, as can gene duplication (D),  
3 transfer (T) and loss (L). These processes are usually modelled independently, but in  
4 reality, ILS can affect gene copy number polymorphism, i.e., interfere with DTL. This has  
5 been previously recognised, but not treated in a satisfactory way, mainly because DTL  
6 events are naturally modelled forward-in-time, while ILS is naturally modelled  
7 backwards-in-time with the coalescent. Here we consider the joint action of ILS and DTL  
8 on the gene tree/species tree problem in all its complexity. In particular, we show that the  
9 interaction between ILS and duplications/transfers (without losses) can result in patterns  
10 usually interpreted as resulting from gene loss, and that the realised rate of D, T and L  
11 becomes non-homogeneous in time when ILS is taken into account. We introduce  
12 algorithmic solutions to these problems. Our new model, the *multilocus multispecies*  
13 *coalescent* (MLMSC), which also accounts for any level of linkage between loci, generalises  
14 the multispecies coalescent model and offers a versatile, powerful framework for proper  
15 simulation and inference of gene family evolution.

16

17 *Key words:* Gene duplication, gene loss, horizontal gene transfer, incomplete lineage  
18 sorting, multispecies coalescent, hemiplasy, recombination

19 Species trees and gene trees are two important kinds of phylogenetic trees which are  
20 key to the study of gene and genome evolution. A species tree depicts the evolutionary  
21 history of a set of organisms, whereas a gene tree depicts the evolutionary history of a gene  
22 family within a set of organisms. A gene tree is thus constrained by the species tree, which  
23 describes the evolution of the organisms containing the gene family.

24 When considering speciations as the only possible events shaping species histories  
25 (e.g., no hybridisation or reassortment possible), internal nodes of species trees represent  
26 only speciation events, and branch lengths represent divergence times or substitutions per  
27 site. Nodes of gene trees, in contrast, can be the result of a combination of diverse  
28 processes, such as variation in gene copy number or gene transfers between species, in  
29 addition to speciations. For this reason, gene trees often differ from species trees, which  
30 can be seen as a problem — if the goal is to infer the species tree — or a source of  
31 information — if the goal is to study molecular evolution.

32 These evolutionary processes can include several ‘gene-range’ events such as gene  
33 duplications, gene losses, and horizontal gene transfers. A gene duplication (D) is an event  
34 in which a single gene copy gives rise to two copies at distinct loci: the parent locus and a  
35 new (child) locus. In contrast, a gene loss (L) removes a gene from the genome. Horizontal  
36 gene transfer (T) occurs when a gene from one species enters the genome of another  
37 contemporary species, which can occur (frequently in bacteria, for example) through a  
38 number of biological mechanisms such as transformation, transduction, and conjugation.  
39 Collectively, we refer to these processes as ‘DTL’. These events can occur multiple times,  
40 allowing the gene tree to possibly differ greatly from the species tree. Examples of these  
41 events are given in Figure 1.

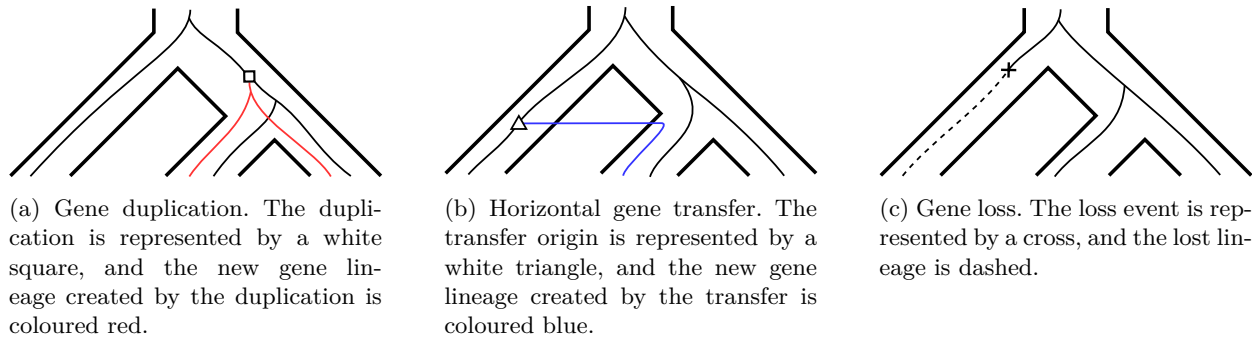


Figure 1: Tree representations of a gene duplication, a horizontal gene transfer, and a gene loss, respectively. The gene lineages (finer lines) evolve within a species tree (outer ‘tubes’).

42 In addition to these events, a gene tree can also be incongruent to the species tree  
43 due to a phenomenon known as incomplete lineage sorting (ILS, Maddison, 1997). When a  
44 population of individuals undergoes several speciations in a relatively short time,  
45 polymorphism (different alleles) maintained throughout this time may eventually fix in  
46 different descendant lineages. This can produce discrepancies between the gene tree and  
47 species tree. ILS is more likely to occur in branches of the species tree (i.e., ancestral  
48 species) that represent small time spans and/or large population sizes (Pamilo and Nei,  
49 1988). An example of ILS is given in Figure 2. Hemiplasy (Avice and Robinson, 2008) is a  
50 term used to refer to the species tree/gene tree conflicts that result from incomplete  
51 lineage sorting.

52 Coalescent theory (Kingman, 1982) provides a genealogical interpretation of ILS  
53 that helps to connect this phenomenon to gene tree-species tree discordance. A key point is  
54 that the age of the common ancestor of two gene copies sampled in two different species is  
55 older (in the absence of DTL) than the time of speciation between the two species. This is  
56 due to the existence of polymorphism in the ancestral species. When speciations occur far  
57 apart in time from each other, ancestral polymorphism is unlikely to result in differences in  
58 the topology of gene trees and the species tree, although it will create differences in branch  
59 lengths. If, however, two or more speciations occur in a time interval of the order of  
60 coalescence times, then coalescence and speciation may interact. This can cause not only

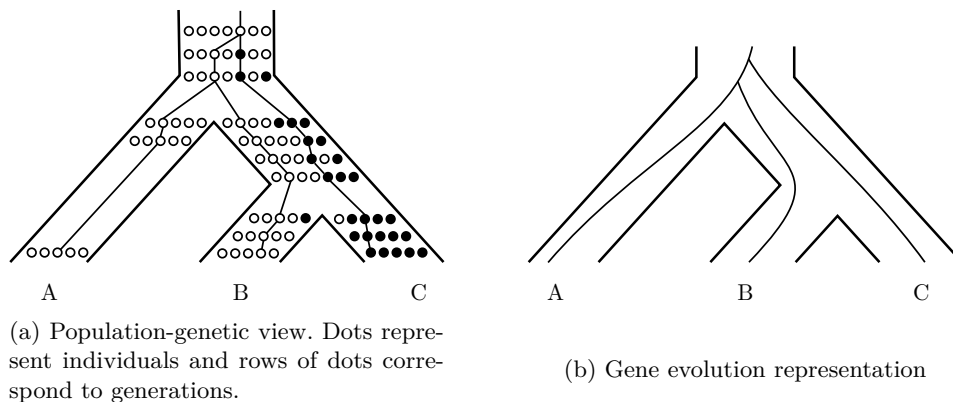


Figure 2: Example of incomplete lineage sorting (ILS). The original population contains a single white allele for the gene of interest. First, a mutation leads to a new black allele at the locus, then the first speciation takes place, rapidly followed by a second one. As the white and black alleles still coexist when the second speciation takes place, both alleles may be fixed in separate descendant species, resulting in a gene history which differs from the species history.

61 branch lengths, but also topology, to differ between gene trees and species trees, as in  
62 Figure 2.

63 The multispecies coalescent model (MSC, Rannala and Yang, 2003) predicts the  
64 effect of ILS on gene tree branch lengths and topology as a function of the effective  
65 population size and timing of speciations. ILS and the MSC have received much attention  
66 after the discovery that classical phylogenetic methods are inconsistent in a subset of the  
67 parameter space (Kubatko and Degnan, 2007; Roch and Steel, 2015), and in the context of  
68 the study of convergent evolution (Guerrero and Hahn, 2018).

69 So far, ILS has mainly been considered separately from other sources of conflict  
70 between gene and species trees. However, ILS can interact in complex and often  
71 unintuitive ways with the processes of gene duplication, loss, and transfer, as first noted by  
72 Rasmussen and Kellis (2012). This is because DTL events spend some time in a  
73 polymorphic stage in a population, when individuals differ from each other in terms of  
74 gene copy number, before they become fixed. If speciations occur during this transient  
75 period of polymorphism, then the issue of lineage sorting becomes even more complex than  
76 in the one-locus case. For example, an allele which is not yet fixed can be lost, as shown in

77 Figure 3. It is also possible that a newly created locus does not fix in all descendant  
78 species, as shown in Figure 4. Thus modelling ILS together with DTL requires greater  
79 flexibility than simply modelling each process individually. We refer to the discrepancies in  
80 gene copy number that result from the interaction between ILS and DTL as ‘copy number  
81 hemiplasy’ (or CNH for short). We note here that Rasmussen and Kellis’ model considers  
82 DTL and ILS independently, and does not model hemiplasy. Therefore, it cannot handle  
83 the scenarios shown in Figures 3 and 4.

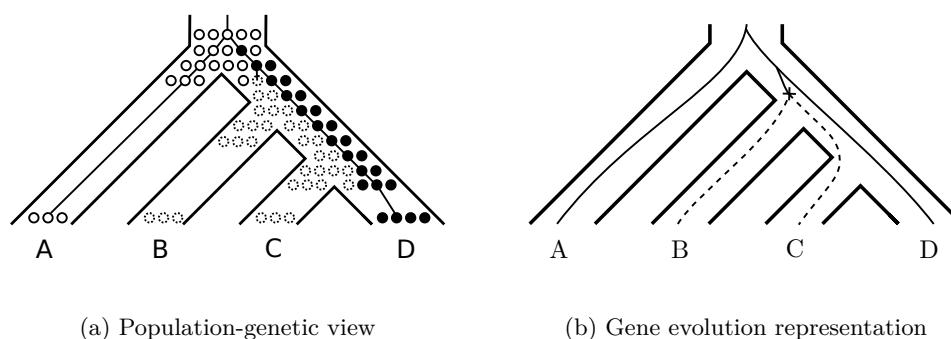


Figure 3: Example of ILS interacting with loss. With only one loss event, two descendant species may end up with an empty locus (represented by dotted circles) due to the presence of ILS.

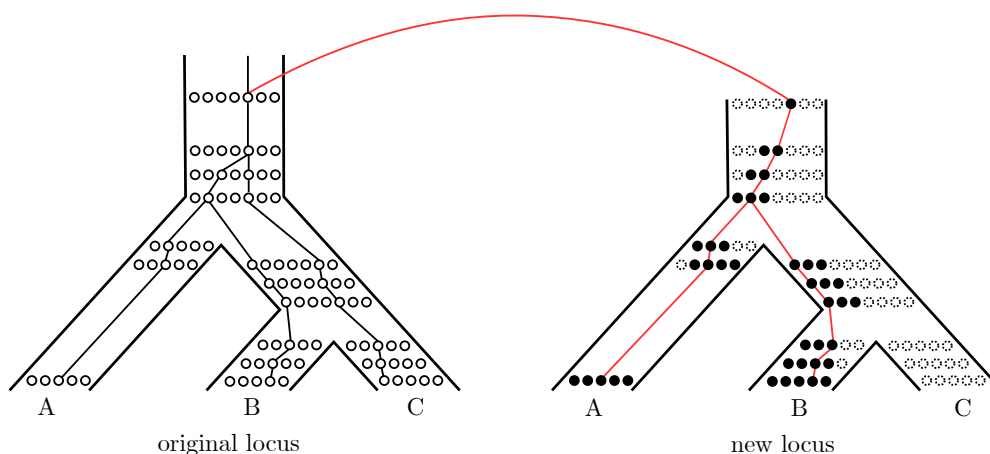


Figure 4: Example of copy number hemiplasy. A duplication arises in an ancestral species, and then two successive speciation events occur. This results in some species retaining two copies of the gene, while others have only one, without any gene loss event. Here, the original gene descends to all extant species (*A*, *B*, and *C*) in the original locus, but the duplicated gene only fixes in species *A* and *B*. Species *B* ends up with the same number of gene copies as *A*, while being more closely related to *C*.

84 Discrepancies between a gene tree and a species tree due to DTL and/or ILS are  
85 frequently analysed via mappings from the gene tree into the species tree, called  
86 *reconciliations* (Goodman et al., 1979). More formally, a reconciliation between a gene tree  
87  $G$  and species tree  $S$  is a mapping of the nodes of  $G$  into the nodes (or a sequence of  
88 nodes) of  $S$  (Doyon et al., 2011), respecting some constraints that depend on the  
89 evolutionary model chosen. This gives rise to the problem of reconciliation inference, where  
90 we seek to reconstruct the ‘true’ reconciliation from the gene and species trees.

91 There are two main paradigms for reconciliation inference: parsimonious and  
92 probabilistic (Doyon et al., 2011). In the parsimonious approach, we assign a cost to each  
93 evolutionary event, and search for a most parsimonious reconciliation, i.e., one which  
94 induces the lowest total cost. In the probabilistic approach, a stochastic model of evolution  
95 is assumed, and either the reconciliation with the maximum likelihood under this model is  
96 found, or a Bayesian approach is used to sample the posterior reconciliation space. In  
97 general, the probabilistic approach is more accurate but less time efficient, since it  
98 typically requires the estimation of many parameters, e.g., effective population sizes and  
99 branch lengths.

100 In our discussion, it is important to distinguish between three concepts:

- 101 • The *model* of evolution — a specification of what can (and cannot) happen in the  
102 evolution of a gene family;
- 103 • A specification of the statistical processes involved in producing the effects captured  
104 by the model, which we call here the *generating process* of the model, which is  
105 necessary to support a probabilistic method; and
- 106 • A *method* to choose an optimal reconciliation under some criterion (parsimonious or  
107 probabilistic).

108 Every reconciliation method assumes a particular model of gene family evolution, limiting  
109 the potential sources of gene tree-species tree discrepancy. However, the underlying model

110 is not always explicitly specified, which is necessary for a proper understanding and  
111 comparison of reconciliation methods.

112 The first reconciliation models considered only gene duplications and losses as the  
113 sources of discrepancy (Goodman et al., 1979; Zhang, 1997; Arvestad et al., 2004; Durand  
114 et al., 2006; Rasmussen and Kellis, 2011). More recently, horizontal gene transfer has been  
115 included in these models (Arvestad et al., 2009; Doyon et al., 2010; David and Alm, 2011;  
116 Tofigh et al., 2011; Sjöstrand et al., 2013), increasing both their complexity and realism.  
117 Statistically, these models are usually generated by birth-death processes running inside  
118 the species tree. On the other hand, the multispecies coalescent (Rannala and Yang, 2003)  
119 generates a model where discrepancies between gene and species trees are only due to ILS.

120 Only a few models (and corresponding methods to infer reconciliations) exist which  
121 attempt to unify these processes. The DLCoalRecon and DLCpar methods (Rasmussen and  
122 Kellis, 2012; Wu et al., 2014; Du et al., 2019; Mawhorter et al., 2019) consider ILS together  
123 with duplications and losses, overlaid on a model called DLCoal. SimPhy (Mallo et al.,  
124 2015) is a simulator based on the DLCoal model that additionally considers transfers.  
125 Schrempf and Szöllősi (2018) also consider these possible events, but use a different model  
126 that we call here the *haplotype tree model*. Lastly, the IDTL (Chan et al., 2017) and  
127 Notung (Stolzer et al., 2012) methods incorporate both ILS and transfers with duplications  
128 and losses, but again using different underlying models. Each of these models are slightly  
129 different from each other, and some of the papers (Chan et al., 2017; Du et al., 2019)  
130 compare the various models. However, they all have their limitations, which are discussed  
131 in Section “Existing Models of Evolution”. In particular, none of them can appropriately  
132 model copy number hemiplasy, although CNH is at the heart of the ILS/DTL interaction.

133 When modelling ILS together with DTL, one must also consider the issue of genetic  
134 linkage between gene copies. A new gene copy arising from a duplication often appears  
135 close to the parent gene on the chromosome, in which case the two loci are expected to  
136 follow correlated coalescent processes. The strength of the correlation is controlled by the



137 amount of recombination between the two loci; in the absence of recombination, the two  
138 genealogies will be the same. The joint coalescent process of partly linked loci in a single  
139 population is well characterised (Hudson, 1983), but the connection with speciation and  
140 ILS has only been rarely considered so far, the model of Slatkin and Pollack (2006) being  
141 one notable exception. (See the Supplementary Material for details about their model and  
142 how it relates to ours.) A realistic model of gene family evolution should account for the  
143 possible existence of linkage between duplicated gene copies.

144 In this paper, we propose a new gene family evolution model, called the *multilocus*  
145 *multispecies coalescent*, or MLMSC for short. (This should not be confused with the  
146 multilocus coalescent, or MLC for short, introduced by Rasmussen and Kellis (2012).) This  
147 model generalises the multispecies coalescent to gene families, and is designed to capture  
148 all possible scenarios that can arise through ILS, DTL, and interaction between these  
149 processes. The MLMSC combines forward- and backward-in-time modelling in order to  
150 properly account for copy number hemiplasy and linkage between loci. Importantly, we  
151 show that the realised rates of D, T and L become non-homogeneous as these processes  
152 interact with ILS, and we introduce a solution to this problem. The MLMSC model is  
153 more flexible and predicts a more diverse range of biological patterns than existing models  
154 of gene family evolution.

## 155 EXISTING MODELS OF EVOLUTION

156 We consider the problem of modelling gene family evolution in a phylogenetic  
157 context. Given a set of species, we assume that we have sampled exactly one haploid  
158 genome per species, and these genomes contain all the gene copies descending from one  
159 particular gene present at the root of the species tree, i.e., a complete gene family. We aim  
160 to model the genealogical relationships between these gene copies, assuming that it has  
161 been shaped by speciations, DTL, and ILS. In this section, we first review the existing  
162 models which have previously addressed this problem.

### *DLCOal Model*

163

164 The reconciliation methods DLCOalRecon and DLCpar are both based on the  
165 DLCOal model (Rasmussen and Kellis, 2012). In this model, when a gene duplication  
166 occurs, the child and parent gene copies evolve independently of each other, i.e., the two  
167 loci are unlinked. Biologically, this arises when there is a sufficient level of recombination  
168 between the two loci that they can be considered to evolve completely independently.

169 To generate this model statistically, given a species tree, we first generate a tree  
170 under a duplication-loss birth-death process on the species tree. Beginning with the  
171 original species tree, at each duplication the tree is copied from that point onwards and  
172 attached to itself (where it may be subject to further duplications), and at each loss the  
173 tree is truncated. This produces the so-called *locus tree*, which depicts the bifurcating  
174 evolution of all loci containing a copy of the gene.

175 The gene tree is then constructed by applying a multispecies coalescent process  
176 within the locus tree, with the caveat that only one gene copy can ‘travel’ along the branch  
177 connecting a duplicated locus back to its parent. In other words, all the lineages in one  
178 particular locus must coalesce into one lineage before coalescing with a copy in its parent  
179 locus — the so-called *bounded coalescent process*. This does not permit copy number  
180 hemiplasy — a duplicated gene is transmitted to all descendant species at the new locus,  
181 barring gene loss. Under this model, ILS does not interfere with gene duplication and loss  
182 in generating variation in gene copy number among species.

183 An example of this model is given in Figure 5, in which a new locus (with  
184 descendants in *B* and *C*) is created by duplication. The duplicated gene is then fixed in all  
185 species at the new locus.

186 It is easy to extend this formulation to include gene transfers, and indeed SimPhy  
187 (Mallo et al., 2015) is a simulator that generates the DLCOal model, extended to transfers  
188 (note that SimPhy does not provide a method to find an optimal reconciliation).

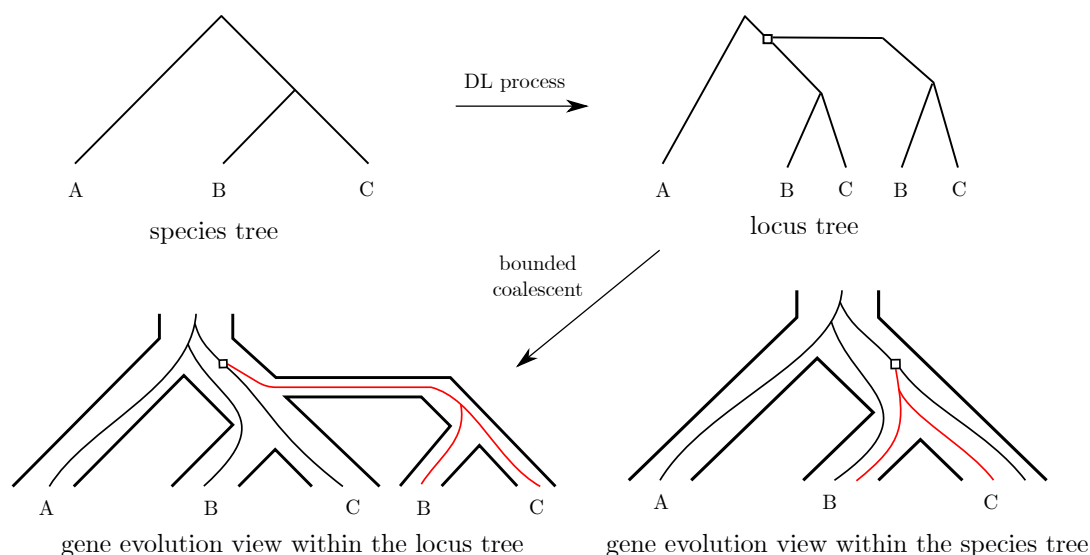


Figure 5: An example of the DLCoal model. Given a species tree, a locus tree is generated by applying a duplication-loss birth-death process on the species tree: the tree is copied from the point of duplication (white square) and attached to the original tree. The gene tree is then constructed under a bounded coalescent process within the locus tree, which requires all the lineages in the new locus (in red) to coalesce before entering the parent locus.

189

### *Haplotype Tree Model*

190 The haplotype tree model was introduced by Schrepf in a talk at the SBE 2018  
191 conference (Schrepf and Szöllösi, 2018). In contrast to the DLCoal model, the haplotype  
192 tree model assumes that there is no recombination. This implies that a duplicated gene  
193 must undergo the exact same genealogical history as its parent gene — a strong  
194 assumption.

195 To generate this model statistically, we first generate a tree under the multispecies  
196 coalescent model on the species tree, obtaining a so-called *haplotype tree*. The gene tree is  
197 then obtained by performing a duplication-loss birth-death process on the haplotype tree.  
198 An example is given in Figure 6. Note that, while some sort of copy number hemiplasy is  
199 allowed in this model, it is restricted: a gene that duplicates must be sorted into exactly  
200 the same descendant species as its parent gene. For instance, it is not allowed for a duplicated  
201 gene to be sorted into all descendant species if the parent gene is not.

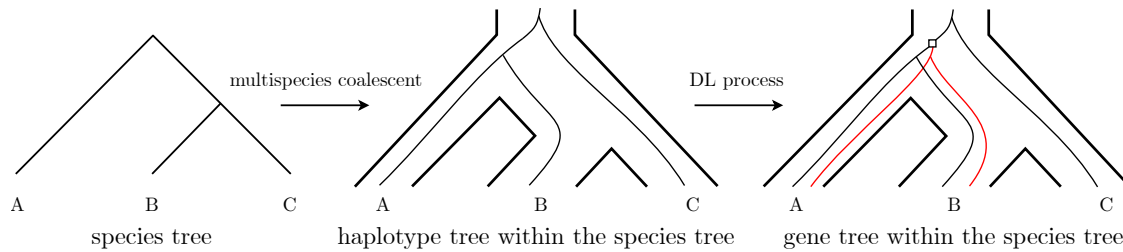


Figure 6: An example of the haplotype tree model. Given a species tree, a haplotype tree is constructed by applying a multispecies coalescent process within the species tree. The gene tree is then generated under a duplication-loss birth-death process on the haplotype tree: the haplotype tree is copied from the point of duplication (white square) and attached to the original tree. The new genes created by the duplication are coloured red.

202

### *IDTL Model*

203

204

205

206

207

208

209

210

211

212

213

214

215

216

The IDTL reconciliation method (Chan et al., 2017) takes a parsimony approach to the reconciliation problem. It defines events forward in time and assumes an underlying gene family evolution model which shares some elements with both the locus and haplotype tree models. Note that, while the model is specified, no generating process has been specified by Chan et al. (2017), since this is not necessary for a parsimony method.

One possible process based on the underlying model is to generate a tree under the multispecies coalescent model on the species tree, followed by a DTL birth-death process, as is done in the haplotype tree model. However, for each new locus, a new multispecies coalescent process is simulated as in the DLCoal model. As in the haplotype tree model, a limited form of copy number hemiplasy is allowed where a duplicate gene must be sorted into exactly the same descendant species as the parent gene. An example of this is given in Figure 7.

This model deals with recombination in an inconsistent way; this will be discussed further in Section “MLMSC vs IDTL Model”.

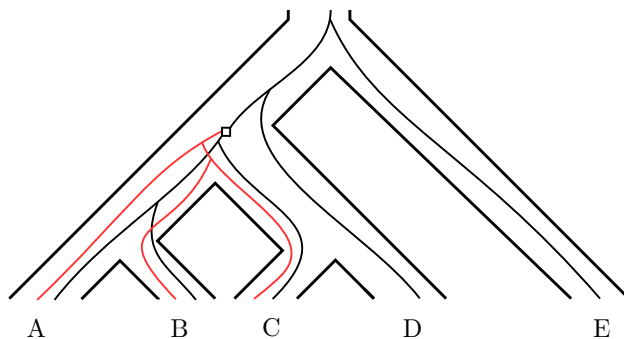


Figure 7: An example of the IDTL model. An allele which is sorted into species  $A$ ,  $B$ , and  $C$  is duplicated. The duplicated gene must then be sorted into the same species, but the coalescent process may be different from the parent gene, resulting in a subtree (coloured red) with differing topology.

217

### *Notung*

218

219

220

221

222

223

224

225

The Notung method (Vernot et al., 2008; Stolzer et al., 2012) was one of the earliest methods to unify the processes of DTL and ILS. It is primarily intended to be used with non-binary species trees, i.e., situations where the branching order is unclear. Because of this, ILS in Notung is only allowed at a polytomy (node with more than two descending lineages) in the species tree, and each possible sorting of genes at the polytomy is considered to be equally likely. Thus there is no explicit modelling of alleles co-existing inside a species branch. As with the IDTL method, Notung is a parsimony method and does not have a formal generating process.

226

### THE MLMSC MODEL

227

228

229

230

The models of gene family evolution in the literature do not model copy number hemiplasy or linked loci. Here we introduce a new gene family evolution model, the *multilocus multispecies coalescent* (MLMSC), for this purpose. We will specify a generating process for this model in the next section.

231

232

One of the primary conceptual difficulties in unifying ILS with a DTL model has been that ILS is traditionally generated by a backwards-in-time coalescent model, whereas

233 the birth-death process used to generate DTL events runs forwards in time. The MLMSC  
234 instead uses a Wright-Fisher model for the population-level genealogical process, which  
235 also runs forwards in time and is therefore easier to merge with a birth-death process.

236 We first define some terminology: in the MLMSC model, one or more *species* exist,  
237 each of which are represented by a *population* consisting of a number of haploid  
238 *individuals*, or *members*. Each of these individuals carry one or more *loci*, which are fixed  
239 positions in the genome in which a *gene* may reside. We trace the evolution of a single *gene*  
240 *family*; that is, all descendants of a single gene, located in a locus inside an individual  
241 belonging to a single ancestral species. (For the purposes of the model, it is unimportant  
242 how this gene originated.) In order to describe the model, we describe the evolution of  
243 species, loci, and finally genes within populations in turn.

#### 244 *The Evolution of Species*

245 New species are created through *speciation*. When a speciation occurs, two new  
246 species are created, each of which is represented by a population which contains the same  
247 loci as the parent species. The two new species then continue evolving independently (as  
248 described below) as if they were continuations of the parent species, as in the MSC  
249 (Rannala and Yang, 2003).

#### 250 *The Evolution of Loci*

251 New loci are created through *gene duplication* (D) or *gene transfer* (T):

- 252 • A duplication occurs in a locus of an individual in a species. When it occurs, a new  
253 locus is created in that species. The new locus may be *linked* or *unlinked* to its  
254 parent locus (resulting from *linked duplication* or *unlinked duplication*, respectively);  
255 in the former case, it is also (indirectly) linked to all loci that its parent is linked to.  
256 The new locus is unlinked to all other existing loci.

- 257 • When a transfer occurs, the same process is followed as for a duplication, except that  
258 a new locus is instead created in a different but contemporary species of the  
259 originating individual, and this locus is unlinked to any existing locus.

260 A locus is lost in a species when there is no longer any individual of the species  
261 carrying a gene copy at that locus.

262 We can thus partition the set of all loci into *sets of linked loci*, which contain loci  
263 linked to each other but not to any locus outside the set. These sets can be of size 1, i.e.,  
264 contain a single locus which is unlinked to all other loci. Note that loci within such a set  
265 are connected by a hierarchical relationship, with one locus (the one created first) being  
266 ancestral to the whole set (the *root locus* of the set), and each of the other loci descending  
267 from a parent locus within the set.

### 268 *The Evolution of Genes Within Populations*

269 Within a population, genes are transmitted from generation to generation according  
270 to a Wright-Fisher model with recombination and gene loss. More precisely, for each locus,  
271 each individual in a generation inherits that locus from a parent individual in the previous  
272 generation. The parent individual is chosen independently for each set of linked loci.

273 Within each set of linked loci, a parent individual is first randomly chosen for the root  
274 locus of the set. Then, recursively, each other locus may (with a certain probability) either  
275 descend from the same individual as its parent locus, or (representing a recombination  
276 between the two loci) another randomly chosen individual from the previous generation.  
277 An example of this process is given in Figure 8.

278 Loci within a set of linked loci are therefore considered such that parent loci are  
279 treated before descendant loci. This order reflects the history of locus creation, and is not  
280 always interpretable in terms of the relative positions of loci along a linear chromosome.  
281 Any pair of linked loci thus evolves according to the two-locus Wright-Fisher model with  
282 recombination (Griffiths, 1991).

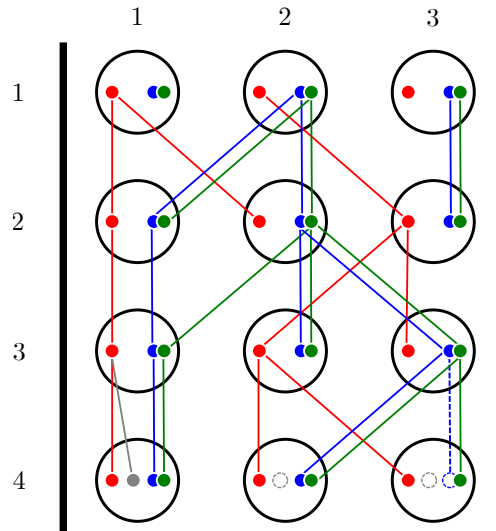


Figure 8: An example of the Wright-Fisher model with recombination and gene loss. In the first generation, each individual contains three loci: the blue and green loci are linked to each other and are inherited together, while the red locus is unlinked to the others and is inherited separately. At generation 3, a recombination event occurs between the blue and green loci, and individual 1 inherits these loci from different ancestors. At generation 4, an unlinked duplication occurs in individual 1, and a new locus (in grey) is created in all individuals in the population; however, only individual 1 contains a gene copy at that locus. A gene loss also occurs in the blue locus of individual 3.

283 When an individual inherits a locus from an individual in the previous generation,  
284 with fixed probability the gene is lost, i.e., is not transmitted to the child individual (see  
285 Figure 8). When there is no longer any individual that carries a gene at a given locus in a  
286 population, that locus is lost from the species.

287 New genes are created through gene duplication and gene transfer, and passed  
288 through speciation:

- 289 • When a duplication or transfer occurs, a new locus is created in the population. The  
290 newly created gene is initially present only in a single individual in this locus (the  
291 individual undergoing duplication, or receiving the transfer).
- 292 • When a speciation occurs, the first generation of each of the child species descends  
293 independently from the last generation of the parent species as described above, and  
294 the process continues independently in the two newly created species.



295 THE GENERATING PROCESS OF THE MLMSC MODEL

296 We now detail a statistical generating process of the MLMSC model under the  
297 following assumptions:

- 298 • A neutral Wright-Fisher model of evolution (no selection or segregation);
- 299 • Constant, large population size;
- 300 • Constant rates of duplication, transfer, loss, and recombination;
- 301 • Constant probability of a duplication being linked or unlinked.

302 Although the MLMSC model uses the Wright-Fisher model of evolution within a  
303 population, the coalescent (which approximates the Wright-Fisher model for  $N \gg n$ ) is  
304 used to generate the gene trees. Firstly, we introduce the concepts of *unilocus trees* and  
305 *haplotype trees and forests*.

306 *Unilocus trees model the history of duplications, transfers, and speciations.* For each locus,  
307 we define its *unilocus tree* as the subtree of the species tree rooted at the species where the  
308 locus first appears. The leaves of a unilocus tree thus correspond to the same locus in  
309 different species, and they are the only extant species which can inherit this locus. The  
310 history of gene duplications, gene transfers, and speciations in the MLMSC is thus stored  
311 as a collection of disconnected unilocus trees.

312 Note that this idea only slightly differs from the locus tree of the DLCoal model  
313 (see above): a collection of unilocus trees can equivalently be seen as a decorated locus  
314 tree, where the decorations are duplications/locus changes. For ease of understanding, in  
315 the following we consider them as disconnected trees.

316 *Haplotype trees and forests model the genealogies of lineages within a unilocus tree.* For a  
317 given locus, we model the genealogical relationships of gene copies across species via a  
318 collection of *haplotype trees*, each of which depicts the genealogy of the genes for a set of  
319 species. A *haplotype forest* is a set which contains either one haplotype tree or a number of

320 disconnected haplotype trees, i.e., haplotype trees whose sets of leaves are disjoint.

321 Generally, only one of the trees in the haplotype forest will actually describe the evolution  
322 of the gene, in which case we refer to it as *the* haplotype tree for the locus.

323 The haplotype forest for a locus is constrained by (evolves within) the unilocus tree  
324 for that locus. Haplotype trees and forests will be used to track the presence of different  
325 alleles in the populations (via ILS, CNH, and losses). Note that in our model, DTL events  
326 are modelled forwards in time and then haplotypes are modelled backwards in time.

327 To generate a gene tree from a species tree under the MLMSC model, we start with  
328 the species tree  $S$  as the unilocus tree for the original locus, and generate a haplotype tree  
329 for it. Then we recursively generate new events, new loci (with unilocus trees), and  
330 haplotype trees and forests for those loci, until all loci have been created. The haplotype  
331 trees are then concatenated together to form the full gene tree.

### 332 *Generating a Haplotype Tree for the Original Locus*

333 In the original locus, the unilocus tree is the original species tree  $S$ . For this locus  
334 only, the haplotype tree is generated according to the standard multispecies coalescent,  
335 starting from a single copy of the gene in each leaf of the tree. We also set the haplotype  
336 forest to be the set containing only the haplotype tree.

337 For loci created by duplication or transfer, a more complex process is required. We  
338 first describe how to generate new loci and unilocus trees, then return to generating  
339 haplotype trees within those unilocus trees.

### 340 *Simulating Surviving Events*

341 After we have generated a haplotype tree and forest for a locus, we simulate DTL  
342 events which occur in that locus. It is important to realise that we do not want to simulate  
343 all DTL events that occur, because the vast majority of these events will simply fail to fix  
344 in the population, and thus be unobserved. Instead, we only wish to simulate events which

345 survive to the present day and are observed in at least one sampled individual. (We use the  
346 terms ‘surviving’ and ‘observed’ interchangeably, as a lineage which survives in an  
347 unsampled individual is undetectable.) In order to do this, we must consider the  
348 probability of survival of each event.

349 In the MLMSC model, the survival probability is not constant. This can be seen by  
350 considering the following simple example: a duplication occurs ‘just before’ a speciation  
351 into two species leaves. In this case, the duplication can survive because it is observed in  
352 either of the descendant species (Figure 9a,b) and possibly in both (Figure 9c). Since the  
353 probabilities of being observed in either species are close to independent, the total survival  
354 probability is roughly twice that of a duplication which occurs in a terminal branch of the  
355 species tree (Figure 9d).

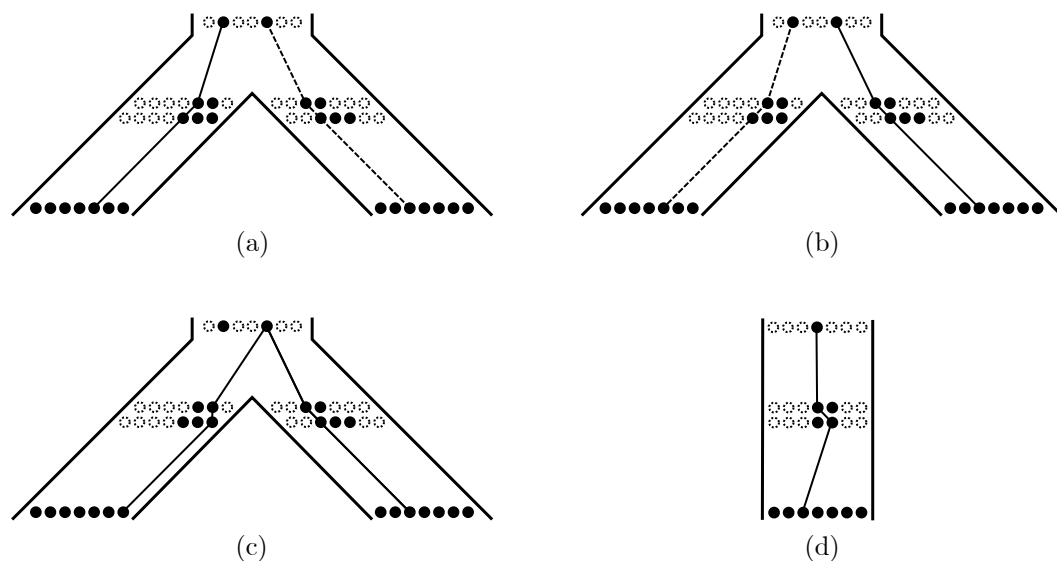


Figure 9: Although duplications occur at a constant rate, they do not survive with constant probability. (a-b) Duplication ‘just before’ a speciation. A duplication in either of the filled individuals will survive (we use the convention that the dashed line is not chosen as the haplotype tree). (c) Only a duplication in one of the individuals will survive, but this scenario requires a near-immediate coalescence and is unlikely. (d) Duplication in a species leaf. Only a duplication in one of the individuals can survive.

356 To simulate surviving duplication and transfer events at the correct rate, we  
357 introduce the *coalescent-rate process*. The simplest version of this process — applied to

358 unlinked duplications — runs a coalescent process in the unilocus tree, and then simulates  
359 events at constant rate on the branches of the trees obtained in this way. The events are  
360 then considered to occur in the corresponding branches of the unilocus tree (see Figure  
361 10). This allows us to simulate surviving duplications at the correct rate. Some  
362 modifications must be made for transfers and linked duplications, and we give further  
363 details in the Supplementary Material.

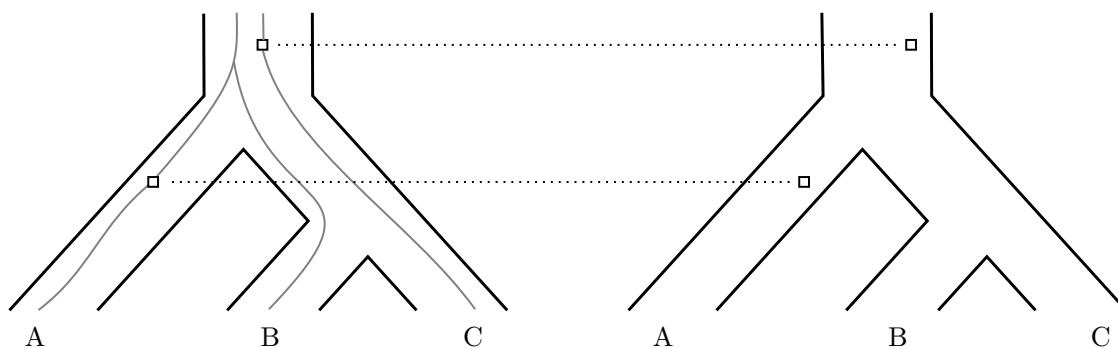


Figure 10: An example of the coalescent-rate process. Firstly, ‘temporary’ trees are sampled from the multispecies coalescent, and then events are sampled at constant rate on the branches of these trees. Finally, the ‘temporary’ trees are removed and the sampled events are considered to occur in the corresponding branches of the unilocus tree.

364 Unlike duplications and transfers, losses which are observed must occur on a  
365 surviving gene lineage, and thus they are sampled from the haplotype tree, instead of the  
366 unilocus tree, at a constant rate. Observe that this allows us to lose an allele which is not  
367 completely fixed in the population, resulting in CNH.

### 368 *Generating New Loci and Unilocus Trees*

369 Once DTL events have been generated, the effect of each event is applied in a  
370 forwards-in-time order. The haplotype tree is truncated at each loss event. At each  
371 duplication or transfer event, a new locus is created, with a corresponding unilocus tree.  
372 The unilocus tree shows the evolution of all species which could possibly contain the locus  
373 (i.e., all descendants of the species where the locus is created). This is the subtree of the

374 species tree  $S$ , starting from the time (and branch) of the creation of the locus (by  
375 duplication or transfer).

376 *Generating a Haplotype Forest — Unlinked Loci*

377 Once we have generated the unilocus tree for a new locus, we then simulate the  
378 haplotype tree and forest. If the new locus is unlinked to the parent locus (i.e., it is created  
379 by transfer or unlinked duplication), then it evolves completely independently from its  
380 parent. To generate the haplotype tree and forest for such a locus, we introduce a new  
381 process called the *incomplete multispecies coalescent*; here, we do not require that all  
382 extant genes have to coalesce to their most recent common ancestor by the time of  
383 origination, i.e., at the root of the unilocus tree. Instead, we simply stop the coalescent  
384 process at the time of origination of the locus, thus producing a haplotype forest. One of  
385 these trees is then randomly chosen to be the haplotype tree. In this way, a  
386 duplicated/transferred gene copy does not have to be transmitted to all descendant  
387 species, resulting in CNH. An example of this process is shown in Figure 11.

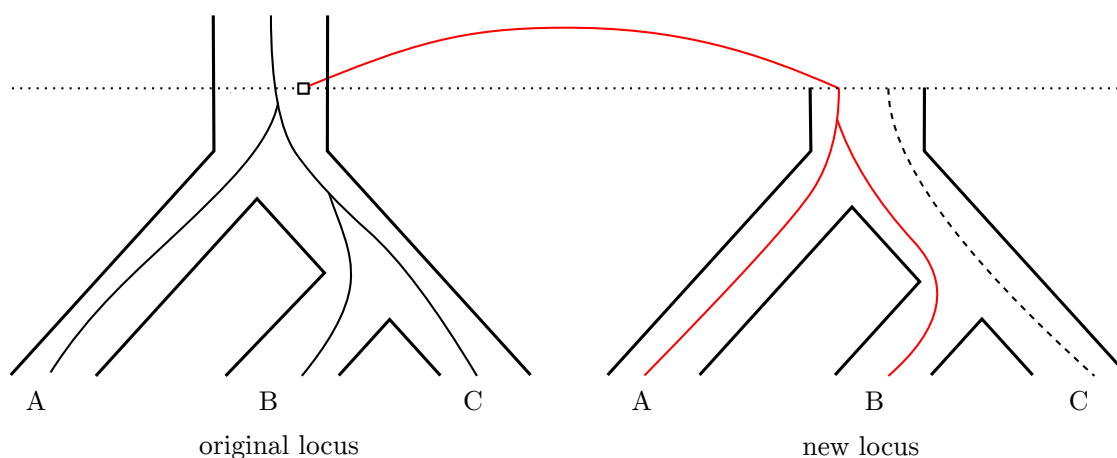


Figure 11: An example of the incomplete multispecies coalescent. Two trees (red and dashed) are generated in the new locus using the multispecies coalescent. The red tree is randomly chosen to be the haplotype tree for this locus, while the dashed tree is part of the haplotype forest but is not the haplotype tree. The root of the red tree represents the individual carrying the new copy.

388 *Generating a Haplotype Forest — Linked Loci*

389 To generate the haplotype tree and forest in a locus created by linked duplication,  
390 we introduce a new process called the *linked coalescent*. In this process, the haplotype  
391 forest in the parent locus is mirrored into the new unilocus tree and used as a pre-existing  
392 genealogy which can be followed by the lineages in the new locus. The incomplete  
393 coalescent is then run, starting from a single lineage in each extant species which is  
394 coalesced with the pre-existing genealogy. Gene lineages which are coalesced with the  
395 pre-existing genealogy must follow it. In this way, the genealogy of the new locus depends  
396 on the genealogy of the existing locus; in the absence of recombination, the two will be  
397 identical.

398 To model the effects of recombination, lineages in the new locus which are coalesced  
399 with the pre-existing genealogy can ‘uncoalesce’ from it at a fixed rate, representing a  
400 recombination event between the new locus and its parent. In contrast, lineages in the new  
401 locus which coalesce with each other cannot uncoalesce, because they represent actual  
402 lineages. Lineages which are not coalesced with the pre-existing genealogy can coalesce  
403 with each other, or with the pre-existing genealogy, at rates consistent with the ordinary  
404 coalescent. An example of this process is given in Figure 12.

405 This process produces the haplotype forest for the new locus, which may contain  
406 more than one tree. We choose one uniformly at random to become the new haplotype tree  
407 for that locus. This haplotype tree must then be (potentially) joined back to the haplotype  
408 tree in the parent locus, as detailed below.

409 *Assembling the Full Gene Tree*

410 Once all events, unilocus trees, and haplotype trees have been generated, we  
411 assemble the full gene tree by concatenating each haplotype tree in a locus to the  
412 haplotype tree in its parent locus, starting from loci which have no descendants.

413 Of particular note is that a duplication may either be *ancestral* (the duplicating

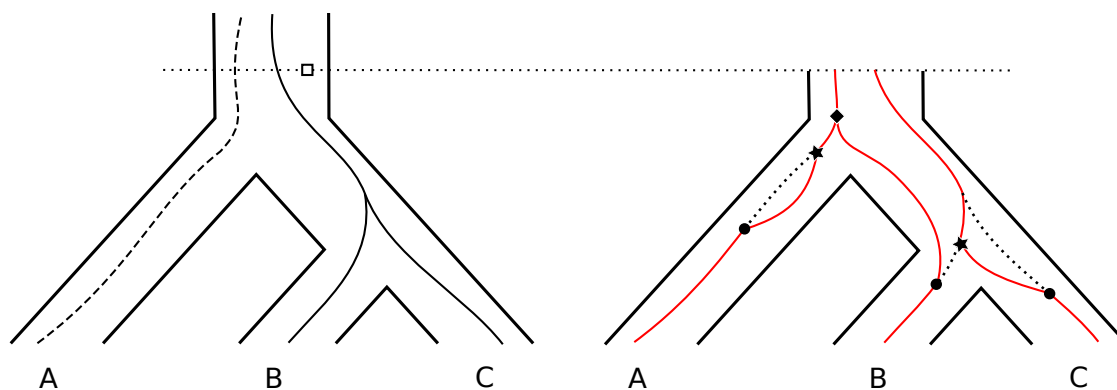


Figure 12: The linked coalescent. On the left, we have the original unilocus tree and its haplotype forest (the haplotype tree is solid and other trees in the haplotype forest are dashed); on the right, the new unilocus tree with a linked coalescent process run on it. The coalescent process starts linked with the genealogy copied from the parent locus in all species, then ‘uncoalesces’ due to recombination (black circles) in all species at different times, coalesces back to the mirrored genealogy (black stars) in  $A$  and the ancestor of  $B$  and  $C$ , and, finally, the lineages that descend to  $A$  and  $B$  coalesce with each other (black diamond).

414 individual is a direct ancestor of a sampled individual in the genealogy of the parent  
415 locus), or *non-ancestral* (the lineage of the duplicating individual does not survive, or is  
416 not sampled, in the parent locus); see Figure 13. For unlinked loci, it can be assumed that  
417 only non-ancestral duplications occur, as the probability of survival in the parent locus,  
418 which is independent of the existence of the duplicated locus, is extremely small and can  
419 be safely ignored. However, when modelling linked loci, it is important to distinguish  
420 between these two cases.

421 By associating lineages in the new locus with lineages in the parent locus, the  
422 linked coalescent provides a natural way to determine if a linked duplication is ancestral or  
423 non-ancestral. If the haplotype tree of a linked locus is coalesced (at the time of its  
424 creation) with a lineage copied from the parent locus, the duplication is ancestral. In this  
425 case, the haplotype tree of the child locus is joined directly to the lineage with which it is  
426 coalesced in the parent locus at the time of the event (Figure 13a).

427 On the other hand, if the locus is unlinked or if the haplotype tree is not coalesced  
428 with a copied lineage, the duplication is non-ancestral. In this case, we can treat the

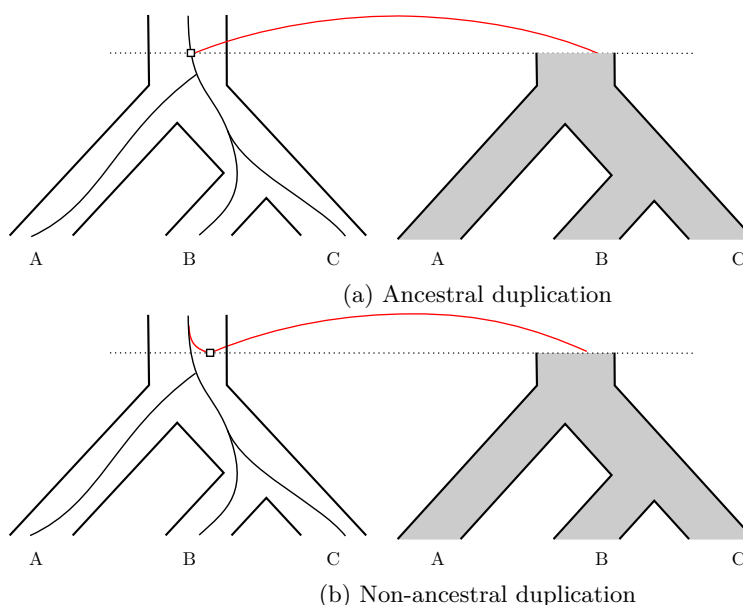


Figure 13: Ancestral vs non-ancestral duplications. (a) A duplication is ancestral if the duplicating individual is a direct ancestor of a sampled individual in the genealogy of the parent locus. (b) A duplication is non-ancestral if the lineage of the duplicating individual does not survive, or is not sampled, in the parent locus.

429 duplicating (or transferring) individual as another member of the population in the parent  
430 locus at the time of the event. We then follow this lineage backwards in time via the  
431 incomplete coalescent until it coalesces with the haplotype forest in the parent locus, or we  
432 reach the time of origination of the parent locus itself (Figure 13b).

433 Observe that this means that a locus can be entirely lost via sorting effects only.  
434 Consider the example in Figure 14: in a locus 1, originating at time  $t_1$ , a duplication occurs  
435 at time  $t_2$ , resulting in locus 2. Under the incomplete coalescent, it might be that the  
436 duplicating lineage fails to coalesce with any lineages in locus 1 by time  $t_1$ . This can be  
437 interpreted as a duplication arising in an individual who does not actually contain any  
438 gene copy at that locus — i.e., an impossibility. In this case, the duplication is discarded  
439 and no new locus is created.

440 When all haplotype trees have either been attached to the haplotype trees of their  
441 parents or discarded, we are left with one remaining tree starting in the original locus,  
442 which we take as the full simulated gene tree.



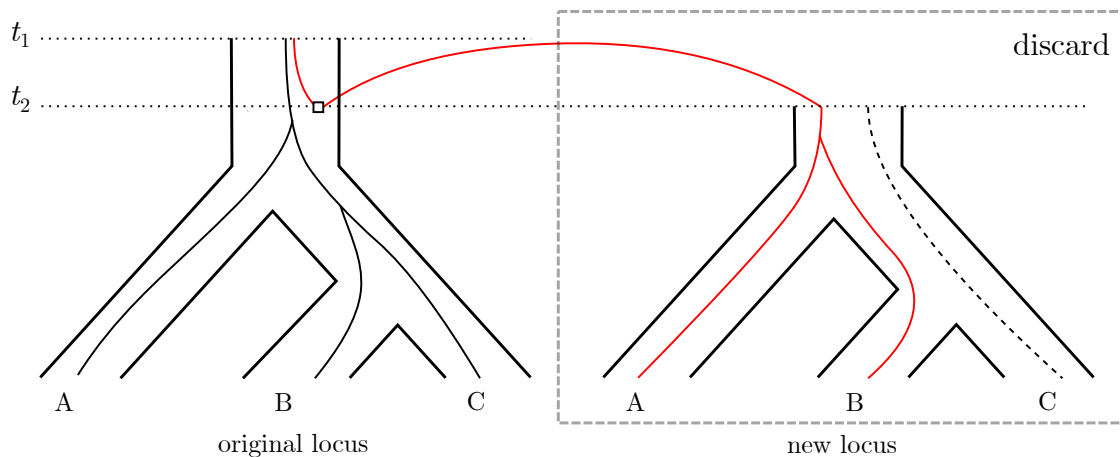


Figure 14: The new haplotype tree does not coalesce with the haplotype forest in the parent locus by the time of its creation, so the duplication is discarded. (Here we use the convention “dashed implies not chosen as haplotype tree”.)

	Transfers	CNH	Recombination	Generating process	Reconciliation method
MLMSC	✓	✓	All	✓	×
DLCoal	×	×	Infinite	✓	✓
Haplotype tree model	×	×	None	✓	×
IDTL	✓	×	Limited	×	✓
Notung	✓	×	Infinite	×	✓

Table 1: Comparison of the models in the literature.

443 More details on the generating process, including a full example, formal notation,  
 444 and pseudocode, are given in the Supplementary Material. We have implemented a  
 445 simulator for this process, available at <http://github.com/QiuyiLi/MLMSC>.

#### 446 MODEL COMPARISON

447 In this section, we discuss the limitations of the current models in the literature,  
 448 and show how the MLMSC model is subject to none of these limitations. Our discussion is  
 449 summarised by, but not limited to, Table 1.

*MLMSC vs DLCoal*

450

451 The DLCoal model simulates losses on the locus tree, instead of the haplotype tree.  
452 Assigning a loss to a locus, instead of a branch of the haplotype tree, means that the loss  
453 of an allele due to lineage sorting cannot be modelled. Consider Figure 15a: here, two  
454 alleles were present in the ancestral population of *B*, *C*, and *D*, one of which was  
455 subsequently lost, i.e., replaced by the null allele. The null allele is then sorted into species  
456 *B* and *C*. The DLCoal model can only produce this gene tree by invoking at least two  
457 losses, as in Figure 15b. In order to properly capture this scenario, losses must be placed  
458 on the haplotype tree rather than the locus (or unilocus) tree, which is exactly what is  
459 done in the MLMSC model.

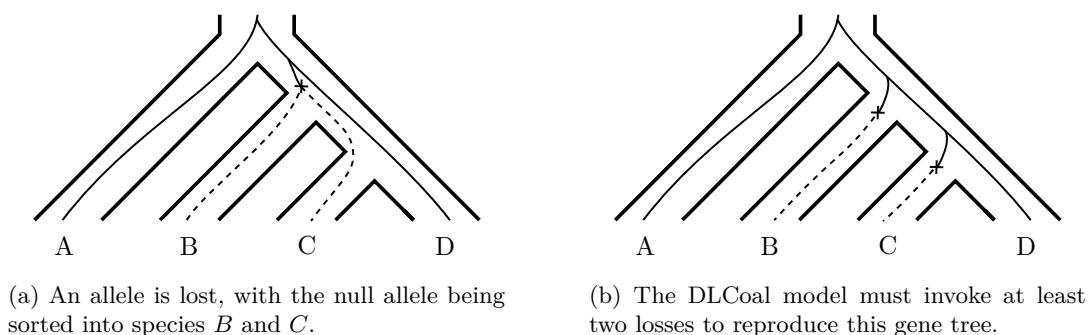


Figure 15: DLCoal cannot model lost alleles.

460 In a similar vein, the DLCoal model also assigns duplications to the locus tree.  
461 These duplications must then rejoin the haplotype tree via the multispecies coalescent; in  
462 other words, they are assumed to be non-ancestral. As discussed above, this is a reasonable  
463 consequence of the model assumption that all loci evolve independently (i.e., are unlinked),  
464 since the probability of an ancestral unlinked duplication is vanishingly small. However, in  
465 the more realistic case that some duplicated loci may be linked to their parent loci, the  
466 survival of a duplicated gene is correlated to the survival of its parent gene, and the  
467 probability of an ancestral linked duplication, knowing that the new duplicate exists, is  
468 non-negligible. Thus it is important to model both ancestral and non-ancestral  
469 duplications, as the MLMSC model does.

470 Finally, in the DLCoal model, a duplicated gene is either lost, or fixed in all possible  
471 descendant species. This means there can be no copy number hemiplasy. For example,  
472 DLCoal cannot model the scenario in Figure 16a: an additional loss is needed, as in Figure  
473 16b. By use of the incomplete coalescent, the MLMSC model can model both scenarios.

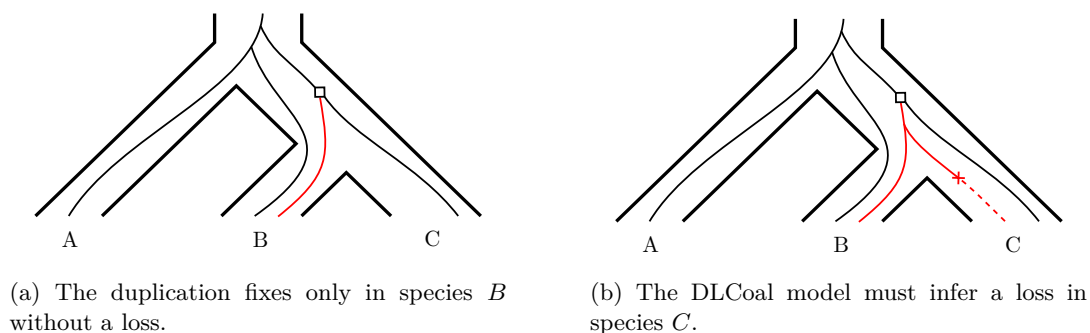


Figure 16: DLCoal cannot model copy number hemiplasy.

474

### *MLMSC vs Haplotype Tree Model*

475

476

477

478

479

480

In the haplotype tree model, a duplicated gene is assumed to have exactly the same genealogy as its parent gene; the model cannot model duplicated genes with different coalescent histories. For this to occur, we must assume that there is no recombination between loci, which is too restrictive. It is more realistic and flexible to allow the loci to be linked, where evolution in the loci are dependent but not necessarily identical, or unlinked, where evolution is completely independent. This is done in the MLMSC model.

481

482

483

484

485

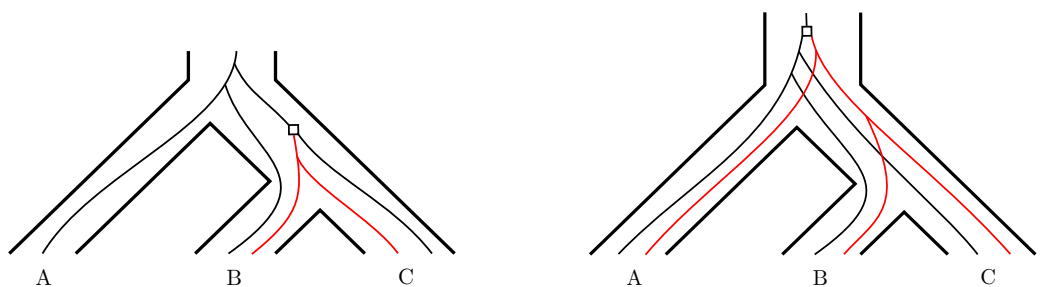
486

487

488

Because the haplotype tree model applies the coalescent first, duplications and losses are assigned to gene lineages, rather than loci as in the DLCoal model. It is reasonable to apply duplications to gene lineages when there is no recombination (and thus only ancestral duplications can be observed). But, as discussed above, a more realistic model is to allow recombination, and model both ancestral and non-ancestral duplications, as done by the MLMSC model. On the other hand, the disadvantages of the DLCoal model do not apply to the haplotype tree model; for example, it can model both scenarios in Figure 15.

489 The haplotype tree model also does not fully allow for copy number hemiplasy;  
490 instead, keeping with the assumption of fully dependent loci, it enforces a restricted  
491 version wherein a duplicate must undergo the same coalescent process, and therefore be  
492 sorted into the same species as the parent gene. For example, the haplotype tree model  
493 cannot model the scenario in Figure 17a, where the duplicated gene is sorted into different  
494 species than the parent gene. It also cannot model the scenario in Figure 17b, where the  
495 duplicated gene undergoes a different genealogy from the parent gene. The MLMSC model  
496 can accommodate both of these scenarios, assuming either unlinked loci or linked loci with  
497 recombination.



(a) The haplotype tree model cannot have a duplication which is sorted into different species from the parent lineage.

(b) The haplotype tree model cannot have a duplication which has a different genealogy than the parent lineage.

Figure 17: Limits of the haplotype tree model. Duplicated lineages are in red, while the parent lineages are in black.

### 498 *MLMSC vs IDTL Model*

499 The IDTL model offers a kind of ‘halfway house’ between the DLCoal and  
500 haplotype tree models, where duplications, transfers, and losses are applied to gene  
501 lineages rather than loci, and duplicates can be sorted in different ways from their parent  
502 genes. However, some of the assumptions made in the model are computationally  
503 convenient but biologically questionable — for example, a duplicated gene must be sorted  
504 into the same species as its parent gene (i.e., no recombination allowed), but it may be  
505 sorted in a different way (i.e., recombination is allowed). For example, Figure 16a is also

506 not allowed by the IDTL model. The MLMSC model is based explicitly on a model of the  
507 coalescent with recombination, and therefore incorporates recombination events in each  
508 species explicitly, allowing a greater range of scenarios.

509 *MLMSC vs Notung*

510 As discussed above, Notung only allows ILS at a polytomy, and each possible  
511 sorting of genes at the polytomy is equivalent and equally likely; there is no ‘correct’  
512 sorting which agrees with a specified species tree. In particular, this means that ILS is not  
513 penalised, and so any gene tree-species tree discrepancy which can be attributed to ILS is  
514 attributed to it, with the remaining discrepancies then explained by DTL. In contrast,  
515 MLMSC allows ILS everywhere in the tree, with probabilities based on the branch lengths  
516 of the species tree, and balances that with the DTL processes. It also specifies an  
517 underlying ‘true’ binary species tree which is always the most likely outcome for the gene  
518 sorting. Internal branch lengths in the MLMSC can be made arbitrarily short, which  
519 effectively covers what Notung represents by polytomies.

520 For computational convenience, the Notung model assumes that whenever a transfer  
521 occurs, the parent lineage must survive to the present day. This means that transfers must  
522 be ancestral. In reconciliation terminology, this means that there are no ‘transfer-loss’  
523 events. As discussed above, this is not a realistic assumption, as only the transferred  
524 lineage needs to be observed; in fact, assuming selective neutrality, the chance of the parent  
525 lineage also surviving in the parent locus is  $O(\frac{1}{2N})$ , where  $2N$  is the effective population  
526 size. The MLMSC model has no such restriction and can model transfer events accurately.

527 DISCUSSION

528 The MLMSC model generalises the multispecies coalescent to include duplications,  
529 transfers, and losses. By using the incomplete coalescent, haplotype forests, and disjoint  
530 unilocus trees, the MLMSC accounts for scenarios in which ancestral gene copy number

531 polymorphisms are incompletely sorted among the descendant species, in contrast with  
532 existing models. By allowing for both linked and unlinked loci, the MLMSC model can also  
533 model more complex evolutionary scenarios than the existing models, while taking  
534 recombination into account in a natural way. Finally, the MLMSC recognises the fact, so  
535 far ignored, that the realised rate of duplication, transfer and loss becomes  
536 non-homogeneous if ILS is at work.

537 MLMSC is a particularly versatile model, which captures many evolutionary  
538 processes: speciation, gene duplication, gene loss, gene transfer, and genetic linkage. In this  
539 respect, MLMSC is more powerful than any of the existing models of gene family evolution  
540 (see Table 1). Note that what one can and cannot model has a strong impact on the  
541 accuracy of the estimation of species and gene trees, as pointed out for instance by  
542 Boussau and Scornavacca (2020). We suggest that, in the presence of CNH, existing  
543 reconciliation methods might lead to biased estimates of gene trees, species trees, and/or  
544 evolutionary parameters, since every instance of a non-fixed duplication would incorrectly  
545 cost one loss if CNH is not explicitly accounted for.

546 Not all these processes, however, are necessarily relevant to every biological system.  
547 The prevalence of ILS, and consequently of CNH, depends on the ratio of effective  
548 population size to branch lengths (Pamilo and Nei, 1988), both of which vary by orders of  
549 magnitude among taxa and datasets. Scornavacca and Galtier (2017), for instance,  
550 suggested that ILS is only a minor determinant of conflicts between gene trees in  
551 phylogenetic analyses of the Mammalia clade, in which effective population sizes are  
552 relatively small (Romiguier et al., 2014) and branches typically represent millions of  
553 generations. DTL alone might be a sufficient model of gene family evolution at this scale.  
554 However, even in mammals, studies of closely related species have demonstrated the  
555 potential importance of ILS when short time frames are considered (Hobolth et al., 2011;  
556 Kutschera et al., 2014), suggesting that, as larger and larger data sets are generated and  
557 analysed, the pertinence of MLMSC should increase. Similarly, the modelling of genetic

558 linkage between duplicates might be superfluous when relatively small and ancient gene  
559 families are considered, but should be crucial for the analyses of large and dynamic gene  
560 families, such as venom toxins (Fry et al., 2009) or olfactory receptors (Niimura et al.,  
561 2014), in which paralogous gene copies are often organized in clusters across the genome  
562 (Olender et al., 2020).

563 To assess the impact of CNH on gene evolution, we simulated gene trees on the  
564 fungal species tree used by Rasmussen and Kellis (2012), using three different duplication  
565 and loss rates ( $10^{-10}$ ,  $5 \times 10^{-10}$ , and  $10 \times 10^{-10}$ , duplication and loss rates are assumed to  
566 be equal), three effective population sizes ( $10^7$ ,  $5 \times 10^7$ , and  $10 \times 10^7$ ), and 0.9 years per  
567 generation. For comparison, Rasmussen and Kellis (2012) used a duplication rate of  
568  $7.32 \times 10^{-10}$ , a loss rate of  $8.59 \times 10^{-10}$ , an effective population size of  $10^7$ , and 0.9 years  
569 per generation. For each set of parameters, we ran 500 simulations and calculated the  
570 proportion of unilocus trees for which the haplotype forest contained more than one  
571 haplotype tree (i.e., CNH occurred). From Figure 18, we see that, for this data set, CNH is  
572 far from negligible, and as expected, its impact is higher when the effective population size  
573 is large — ranging from a bit less than 15% for  $N_e = 10^7$ , to more than 50% for  $N_e = 10^8$ .  
574 On the other hand, the proportion of CNH stays constant when varying the duplication  
575 and loss rate. This indicates that CNH is an important phenomenon which must be taken  
576 into account when modelling biological systems.

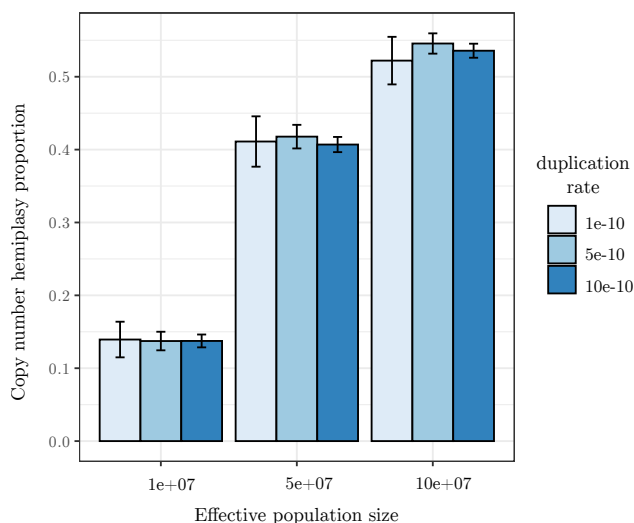


Figure 18: The proportion of CNH in our simulated data, varying the duplication and loss rate and the effective population size.

577 An implicit assumption in the MLMSC is that we can determine the linked or  
578 unlinked status of each locus solely with respect to its parent locus. In this respect, it is not  
579 fully cognisant of the linear structure of the chromosome. For example, if a locus (say locus  
580 1) has two linked duplications in rapid succession (say loci 2 and 3), then one of the three  
581 loci must lie between the other two. If the order is (say) 1–2–3, then a recombination event  
582 between loci 1 and 2 must imply a recombination event between loci 1 and 3; this kind of  
583 dependency is not modelled by the MLMSC, which effectively represents the relations  
584 between linked loci as a tree rather than a linear structure. Moreover, in a scenario such as  
585 this, it may not be reasonable to assume that the recombination rate between loci 1 and 2  
586 is equal to the recombination rate between 1 and 3. A full model which incorporates the  
587 linear structure of the chromosome would have to explicitly model the position of the gene  
588 copies, which we have elected not to do here for reasons of computational convenience.

589 We note that our goal here is not to correctly model recombination to its full  
590 extent, but to handle the case where the original locus and the duplicated one are next to  
591 each other and may thus have linked histories. In addition to the above, recombination is  
592 only allowed between (not within) duplicates. Traditional population genetic models, such



593 as the coalescent with recombination and its approximations such as the sequentially  
594 Markov coalescent, permit more flexibility on this matter: recombination is allowed  
595 anywhere within a linear genome, and no pre-knowledge about the genealogy in neighbour  
596 loci is required. These models also have their drawbacks: all genomes are assumed to be  
597 aligned, with insertions and creations of new loci not allowed, and they assume that  
598 recombination occurs at a constant rate throughout the genome, ignoring the existence of  
599 hotspots or variation in rates due to gene function. One could say that our model does  
600 include this somewhat by considering a locus as a non-recombinant segment and only  
601 allowing recombination outside it.

602 Another limitation of our model is that it only accounts for a portion of the events  
603 shaping species and gene histories. For example, reticulated events such as hybridisation or  
604 reassortment, whole genome duplication events, and gene conversions are not modelled.  
605 Neither is “transfer from the dead” (Szöllősi et al., 2013), which is an important  
606 consideration when modelling transfers. There are no theoretical barriers to modelling this  
607 process, but it is unclear how to do so in a realistic and definitive manner. Two possibilities  
608 are to (a) maintain a separate ‘dead’ species which can only receive and donate transfers,  
609 or (b) allow transfers to ‘jump’ forward in time (i.e., allow the transfer origin to be selected  
610 at a time previous to the transfer target). For (a), we would need to set transfer rates to  
611 (and from) the ‘dead’ species; since we lack information about the structure of these  
612 species, a coalescent-rate process could not be used. For (b), we would need to determine  
613 the appropriate distribution of time between the transfer origin and target. Both of these  
614 models would go ‘outside’ the Wright-Fisher process which we are attempting to capture  
615 with the MLMSC model; thus, in order to maintain the elegance of our model, we have  
616 elected not to model this process here, but intend to include this in future work.

617 The analysis of gene families is an increasingly important component of  
618 comparative genomics (see e.g., Scornavacca et al., 2020, part 3). With so many fully  
619 sequenced genomes/transcriptomes available, the proportion of one-to-one orthologs in

620 typical multi-gene, multi-species data sets will become smaller and smaller, and most of  
621 the information about the history of genomes and organisms must be extracted from  
622 multi-copy gene families. This is particularly true of organisms having large, complex  
623 genomes and undergoing frequent hybridisation and large-scale duplications, such as the  
624 economically important angiosperms (Glémin et al., 2019; Stull et al., 2020) and fish (Alda  
625 et al., 2019; Campbell et al., 2020), to name just a few.

626 In this context, the development of the general and arguably more realistic MLMSC  
627 model opens up promising research avenues. Simulations under the MLMSC can be used to  
628 assess the accuracy of current inference algorithms, and may help to confirm or contradict  
629 a number of published results. For instance, we suggest that neglecting ILS in phylogenetic  
630 analyses of gene families might bias the estimation of the timing of speciation/duplication.  
631 Patterns of gene loss subsequent to a whole-genome duplication, which is often interpreted  
632 in terms of evolution of gene function, could also be affected by ILS to an extent that  
633 remains to be quantified. Finally, in principle, the MLMSC model and its generating  
634 process could be used in an inference framework, i.e., serve as the basis for the  
635 development of new reconciliation algorithms, or parameter estimation methods (perhaps  
636 via Approximate Bayesian Computation). This would presumably be of great utility, while  
637 requiring substantial additional work.

#### 638 ACKNOWLEDGEMENTS

639 YBC and CS acknowledge the Partenariat Hubert Curien (Hubert Curien  
640 Partnership) for providing grants for collaborative travel. CS was funded by grant  
641 ANR-19-CE45-0012 from the Agence Nationale de la Recherche. QL would like to  
642 acknowledge Geoffrey Law for providing assistance with programming.

643

REFERENCES

- 644 Alda, F., Tagliacollo, V. A., Bernt, M. J., Waltz, B. T., Ludt, W. B., Faircloth, B. C.,  
645 Alfaro, M. E., Albert, J. S., and Chakrabarty, P. (2019). Resolving deep nodes in an  
646 ancient radiation of neotropical fishes in the presence of conflicting signals from  
647 incomplete lineage sorting. *Syst. Biol.*, 68(4):573–593.
- 648 Arvestad, L., Berglund, A.-C., Lagergren, J., and Sennblad, B. (2004). Gene tree  
649 reconstruction and orthology analysis based on an integrated model for duplications  
650 and sequence evolution. In *RECOMB 2004*, pages 326–335. ACM.
- 651 Arvestad, L., Lagergren, J., and Sennblad, B. (2009). The gene evolution model and  
652 computing its associated probabilities. *J. ACM*, 56(2):1–44.
- 653 Avise, J. C. and Robinson, T. J. (2008). Hemiplasy: a new term in the lexicon of  
654 phylogenetics. *Syst. Biol.*, 57(3):503–507.
- 655 Boussau, B. and Scornavacca, C. (2020). Reconciling Gene trees with Species Trees. In  
656 Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic  
657 Era*, pages 3.2:1–3.2:23. No commercial publisher | Authors open access book.
- 658 Campbell, M. A., Buser, T. J., Alfaro, M. E., and López, J. A. (2020). Addressing  
659 incomplete lineage sorting and paralogy in the inference of uncertain salmonid  
660 phylogenetic relationships. *PeerJ*, 8:e9389.
- 661 Chan, Y., Ranwez, V., and Scornavacca, C. (2017). Inferring incomplete lineage sorting,  
662 duplications, transfers and losses with reconciliations. *J. Theor. Biol.*, 432:1–13.
- 663 David, L. A. and Alm, E. J. (2011). Rapid evolutionary innovation during an archaean  
664 genetic expansion. *Nature*, 469(7328):93.
- 665 Doyon, J.-P., Ranwez, V., Daubin, V., and Berry, V. (2011). Models, algorithms and  
666 programs for phylogeny reconciliation. *Brief. Bioinform.*, 12(5):392–400.

- 667 Doyon, J.-P., Scornavacca, C., Gorbunov, K. Y., Szöllősi, G. J., Ranwez, V., and Berry, V.  
668 (2010). An efficient algorithm for gene/species trees parsimonious reconciliation with  
669 losses, duplications and transfers. In *RECOMB International Workshop on*  
670 *Comparative Genomics*, pages 93–108. Springer.
- 671 Du, H., Ong, Y. S., Knittel, M., Mawhorter, R., Liu, N., Gross, G., Tojo, R.,  
672 Libeskind-Hadas, R., and Wu, Y.-C. (2019). Multiple optimal reconciliations under  
673 the duplication-loss-coalescence model. *IEEE/ACM Trans. Comput. Biol. Bioinform.*
- 674 Durand, D., Halldórsson, B. V., and Vernet, B. (2006). A hybrid micro–macroevolutionary  
675 approach to gene tree reconstruction. *J. Comput. Biol.*, 13(2):320–335.
- 676 Fry, B. G., Roelants, K., Champagne, D. E., Scheib, H., Tyndall, J. D., King, G. F.,  
677 Nevalainen, T. J., Norman, J. A., Lewis, R. J., Norton, R. S., et al. (2009). The  
678 toxicogenomic multiverse: convergent recruitment of proteins into animal venoms.  
679 *Annu. Rev. Genomics Hum. Genet.*, 10:483–511.
- 680 Glémin, S., Scornavacca, C., Dainat, J., Burgarella, C., Viader, V., Ardisson, M., Sarah,  
681 G., Santoni, S., David, J., and Ranwez, V. (2019). Pervasive hybridizations in the  
682 history of wheat relatives. *Sci. Adv.*, 5(5):eaav9188.
- 683 Goodman, M., Czelusniak, J., Moore, G. W., Romero-Herrera, A. E., and Matsuda, G.  
684 (1979). Fitting the gene lineage into its species lineage, a parsimony strategy  
685 illustrated by cladograms constructed from globin sequences. *Syst. Biol.*,  
686 28(2):132–163.
- 687 Griffiths, R. C. (1991). The two-locus ancestral graph. *Lecture Notes-Monograph Series*,  
688 pages 100–117.
- 689 Guerrero, R. F. and Hahn, M. W. (2018). Quantifying the risk of hemiplasy in  
690 phylogenetic inference. *Proc. Natl. Acad. Sci. U.S.A.*, 115(50):12787–12792.

- 691 Hobolth, A., Dutheil, J. Y., Hawks, J., Schierup, M. H., and Mailund, T. (2011).  
692 Incomplete lineage sorting patterns among human, chimpanzee, and orangutan suggest  
693 recent orangutan speciation and widespread selection. *Genome Res.*, 21(3):349–356.
- 694 Hudson, R. R. (1983). Properties of a neutral allele model with intragenic recombination.  
695 *Theor. Popul. Biol.*, 23(2):183–201.
- 696 Kingman, J. F. (1982). On the genealogy of large populations. *J. Appl. Probab.*,  
697 19(A):27–43.
- 698 Kubatko, L. S. and Degnan, J. H. (2007). Inconsistency of phylogenetic estimates from  
699 concatenated data under coalescence. *Syst. Biol.*, 56(1):17–24.
- 700 Kutschera, V. E., Bidon, T., Hailer, F., Rodi, J. L., Fain, S. R., and Janke, A. (2014).  
701 Bears in a forest of gene trees: phylogenetic inference is complicated by incomplete  
702 lineage sorting and gene flow. *Mol. Biol. Evol.*, 31(8):2004–2017.
- 703 Maddison, W. P. (1997). Gene trees in species trees. *Syst. Biol.*, 46(3):523–536.
- 704 Mallo, D., de Oliveira Martins, L., and Posada, D. (2015). Simphy: phylogenomic  
705 simulation of gene, locus, and species trees. *Syst. Biol.*, 65(2):334–344.
- 706 Mawhorter, R., Liu, N., Libeskind-Hadas, R., and Wu, Y.-C. (2019). Inferring  
707 pareto-optimal reconciliations across multiple event costs under the  
708 duplication-loss-coalescence model. *BMC Bioinform.*, 20(20):639.
- 709 Niimura, Y., Matsui, A., and Touhara, K. (2014). Extreme expansion of the olfactory  
710 receptor gene repertoire in african elephants and evolutionary dynamics of orthologous  
711 gene groups in 13 placental mammals. *Genome Res.*, 24(9):1485–1496.
- 712 Olender, T., Jones, T. E., Bruford, E., and Lancet, D. (2020). A unified nomenclature for  
713 vertebrate olfactory receptors. *BMC Evol. Biol.*, 20:1–12.
- 714 Pamilo, P. and Nei, M. (1988). Relationships between gene trees and species trees. *Mol.*  
715 *Biol. Evol.*, 5(5):568–583.

- 716 Rannala, B. and Yang, Z. (2003). Bayes estimation of species divergence times and  
717 ancestral population sizes using DNA sequences from multiple loci. *Genetics*,  
718 164(4):1645–1656.
- 719 Rasmussen, M. D. and Kellis, M. (2011). A bayesian approach for fast and accurate gene  
720 tree reconstruction. *Mol. Biol. Evol.*, 28(1):273–290.
- 721 Rasmussen, M. D. and Kellis, M. (2012). Unified modeling of gene duplication, loss, and  
722 coalescence using a locus tree. *Genome Res.*, 22(4):755–765.
- 723 Roch, S. and Steel, M. (2015). Likelihood-based tree reconstruction on a concatenation of  
724 aligned sequence data sets can be statistically inconsistent. *Theor. Popul. Biol.*,  
725 100:56–62.
- 726 Romiguier, J., Gayral, P., Ballenghien, M., Bernard, A., Cahais, V., Chenuil, A., Chiari,  
727 Y., Dernas, R., Duret, L., Faivre, N., et al. (2014). Comparative population genomics  
728 in animals uncovers the determinants of genetic diversity. *Nature*, 515(7526):261–263.
- 729 Schrempf, D. and Szöllösi, G. J. (2018). Phylogenetic incongruences - opportunities to  
730 improve the reconstruction of a dated tree of life.
- 731 Scornavacca, C., Delsuc, F., and Galtier, N. (2020). Phylogenetics in the genomic era.
- 732 Scornavacca, C. and Galtier, N. (2017). Incomplete lineage sorting in mammalian  
733 phylogenomics. *Syst. Biol.*, 66(1):112–120.
- 734 Sjöstrand, J., Arvestad, L., Lagergren, J., and Sennblad, B. (2013). Genphylodata:  
735 realistic simulation of gene family evolution. *BMC Bioinform.*, 14(1):209.
- 736 Slatkin, M. and Pollack, J. L. (2006). The concordance of gene trees and species trees at  
737 two linked loci. *Genetics*, 172(3):1979–1984.
- 738 Stolzer, M., Lai, H., Xu, M., Sathaye, D., Vernot, B., and Durand, D. (2012). Inferring  
739 duplications, losses, transfers and incomplete lineage sorting with nonbinary species  
740 trees. *Bioinformatics*, 28(18):i409–i415.

- 741 Stull, G. W., Soltis, P. S., Soltis, D. E., Gitzendanner, M. A., and Smith, S. A. (2020).  
742 Nuclear phylogenomic analyses of asterids conflict with plastome trees and support  
743 novel relationships among major lineages. *Am. J. Bot.*
- 744 Szöllősi, G. J., Tannier, E., Lartillot, N., and Daubin, V. (2013). Lateral gene transfer  
745 from the dead. *Syst. Biol.*, 62(3):386–397.
- 746 Tofight, A., Hallett, M., and Lagergren, J. (2011). Simultaneous identification of  
747 duplications and lateral gene transfers. *IEEE/ACM Trans. Comput. Biol. Bioinform.*  
748 (*TCBB*), 8(2):517–535.
- 749 Vernot, B., Stolzer, M., Goldman, A., and Durand, D. (2008). Reconciliation with  
750 non-binary species trees. *J. Comput. Biol.*, 15(8):981–1006.
- 751 Wu, Y.-C., Rasmussen, M. D., Bansal, M. S., and Kellis, M. (2014). Most parsimonious  
752 reconciliation in the presence of gene duplication, loss, and deep coalescence using  
753 labeled coalescent trees. *Genome Res.*, 24(3):475–486.
- 754 Zhang, L. (1997). On a Mirkin-Muchnik-Smith conjecture for comparing molecular  
755 phylogenies. *J. Comput. Biol.*, 4(2):177–187.