1   To appear in *Digital Scholarship in the Humanities*

2

3

4   Oral Fairy Tale or Literary Fake? Investigating the Origins

5   of Little Red Riding Hood Using Phylogenetic Network

6   Analysis

7

8   Jamshid Tehrani[1][†], Quan Nguyen[2], Teemu Roos[2][†]

9   [1] Department of Anthropology, Durham University, South Road, Durham, DH1 3LE

10   [2] Department of Computer Science and Helsinki Institute for Information Technology,

11   FI-0014 University of Helsinki, PO Box 68, Helsinki.

12   [†] Authors for correspondence. J. Tehrani: jamie.tehrani@dur.ac.uk Teemu Roos:

13   teemu.roos@cs.helsinki.fi

14

15

16

18

19

20

21

22   *Abstract*

23

24   The evolution of fairy tales often involves complex interactions between oral and

25   literary traditions, which can be difficult to tease apart when investigating their

26   origins. Here, we show how computer-assisted stemmatology can be productively

27   applied to this problem, focusing on a long-standing controversy in fairy tale

28   scholarship: did Little Red Riding Hood originate as an oral tale that was adapted by

29   Perrault and the Brothers Grimm, or is the oral tradition in fact derived from literary

30   texts? We address this question by analysing a sample of 24 literal and oral versions

31   of the fairy tale Little Red Riding Hood using several methods of phylogenetic

32   analysis, including maximum parsimony and two network-based approaches

33   (NeighbourNet and TRex). While the results of these analyses are more compatible

34   with the oral origins hypothesis than the alternative literary origins hypothesis, their

35   interpretation is problematised by the fact that none of them explicitly model lineal

36   (i.e. ancestor-descendent) relationships among taxa. We therefore present a new

37   likelihood-based method, PhyloDAG, which was specifically developed to model

38   lineal as well as collateral and reticulate relationships. A comparison of different

39   structures derived from PhyloDAG provided a much clearer result than the

40   maximum parsimony, NeighbourNet or TRex analyses, and strongly favoured the

41   hypothesis that literary versions of Little Red Riding Hood were originally based on

42   oral folktales, rather than vice versa.

43

44   **1. Introduction**

45

46   Recent years have witnessed a boom in computational approaches to the reconstruction of

47   literary traditions, fuelled by the adoption of phylogenetic techniques from evolutionary

48   biology and the development of custom-made software for textual analysis (Howe et al.,

49   2001; Roos & Heikkilä, 2009). So far, research in this field has focused on the transmission

50   histories of hand-copied manuscripts, where the accumulation of errors and occasional

51   innovations can be modelled as a branching process analogous to the diversification of

52   biological lineages by descent with modification. Recently, it has been argued that a similar

53   approach can shed light on the evolution of oral traditions, such as folktales (Tehrani, 2013),

54   legends (Stubbersfield & Tehrani, 2013) and myths (d'Huy, 2013). Although these stories are

55   not literally copied in the way that manuscripts or DNA sequences are, their basic plot

56   elements, motifs, characters and symbols exhibit clear evidence of both fidelity of

57   transmission as well as cumulative change through time. Recent case studies (Tehrani, 2013)

58   demonstrate that careful analyses of these features make it possible to reconstruct deep and

59   robust stemmata, which can in turn yield potentially crucial insights into the origin and

60   development of oral tales.

61

62   One of the key issues in this area concerns the complex interactions between oral and

63   literary traditions, which are often difficult to disentangle. For example, it is well known that,

64   historically, many so-called fairy tales (i.e. traditional short stories containing fantastical or

65   magical elements) have been adapted by writers inspired by oral story-tellers and vice versa.

66   In such cases, it can be extremely problematic to establish in which medium a given tale

67   originated. While most folklorists have tended to assume that fairy tales are rooted in oral

68   tradition, some scholars have argued that they may in fact be derived from written texts. Most

69   notably, Ruth Bottigheimer (Bottigheimer, 2002, 2010) proposed that fairy tales are a

70   primarily literary genre that was invented by the sixteenth century writer Giovanni Francesco

71   Straparola and subsequently popularised by other authors such as Basile, Perrault and the

72   Brothers Grimm. While these authors presented their stories as though they were borrowed

73   from the tales told by common folk, Bottigheimer suggests this was simply a stylistic ruse,

74   and that the direction of transmission was much more likely to be the other way around. In

75 support of this point, she highlights that the earliest literary versions of fairy tales were

76 written centuries earlier than the supposedly more authentic oral versions collected by

77 folklorists. Bottigheimer's controversial thesis has been rejected by most experts (Ben-Amos,

78 Ziolkowski, Silva, & Bottigheimer, 2010), who point out that absence of evidence hardly

79 constitutes evidence for absence, especially given that oral traditions, by definition, lack a

80 written record. However, by the same token, nor can it be proved that oral fairy tales predate

81 the earliest written versions. In this paper, we show how techniques developed in computer-

82 assisted stemmatology can help break this impasse, and shed new light on the missing links

83 between oral and literary traditions in fairy tales.

84

85   Our case study focuses on a tale whose origin has long been the subject of intense

86 controversy: Little Red Riding Hood. The tale, which is classified as ATU 333 in the Aarne-

87 Thompson-Uther (ATU) Index of International Tale Types, famously tells the story of a

88 young girl who is attacked by a wolf disguised as her grandmother. There are numerous

89 theories about the source of the tale, from pre-Christian sun myths (Saintyves, 1989) or

90 medieval coming-of-age rites (Verdier, 1978) to Chinese folk tradition (Haar, 2006). While

91 these ideas remain difficult to substantiate, the modern tradition of Little Red Riding

92 Hood/ATU 333 can be traced back to 1697, when the first classic version of the story, *Le*

93 *Petit Chaperon Rouge*, was published by the French author Charles Perrault in his collection

94 of purportedly traditional stories, *Histoires ou Contes du Temps Passé* (Tales of Past Times)

95 (1697). A second classic version of Little Red Riding Hood (*Rotkäppchen*) was published in

96 1813 in the first volume of Jacob and Wilhelm Grimm's *Kinder und Hausmärchen*

97 (Children's and Household Tales) (1812). In this version, unlike Perrault's, Little Red and her

98 grandmother are rescued by a passing huntsman, who slices open the villain's stomach and

99 sews it up again with stones. Although, like the other tales in that volume, Rotkäppchen was

100 ostensibly collected from ordinary German peasant folk, Grimm scholars have established

101 that the brothers' source for the tale was actually an educated woman of French-Huguenot

102   descent named Marie Hassenpflug, who was almost certainly familiar with Perrault's

103   enormously popular *Contes* (Zipes, 1993).

104

105   While the Perrault and Grimm tales provided the model from which all subsequent

106   literary Little Red Riding Hoods are derived, the origins of the oral tradition of ATU 333, and

107   its relationship to these two "classic" versions, are much less well understood. Most

108   folklorists believe that Perrault based his tale on a traditional French werewolf tale, probably

109   from his mother's native region of Touraine, which was the site of a series of werewolf trials

110   in the sixteenth and seventeenth centuries (Zipes, 1993, p. 20). It is claimed that variants of

111   the tale survived into the nineteenth and twentieth centuries in the oral literatures of south-

112   east France, the Alps and northern Italy (Delarue, 1951; Rumpf, 1989). These tales,

113   commonly referred to as simply 'The Story of Grandmother' (following Delarue 1951) are

114   typically more gory than Perrault's censored version – for example, the girl is tricked into

115   eating some of her grandmother's remains. More importantly, rather than being a helpless

116   victim, the girl typically outwits the wolf/werewolf by tricking him into letting her go outside

117   to urinate. Although the provenance and antiquity of the tradition remains unknown, it has

118   been suggested that it may go back to medieval times. This is supported by an eleventh

119   century Latin poem by Egbert of Liége, which relates a local Walloon folktale in which a

120   young girl encounters a wolf in the woods, and is saved by the supernatural protection

121   afforded by her red tunic, a baptism gift from her godfather, (Ziolkowski, 1992). Although it

122   is debateable as to whether or not this tale represents a direct ancestor to Little Red Riding

123   Hood (Berlioz, 1991), the echo of common motifs like the young girl in the woods, the

124   villainous wolf, the red outfit given to her by a relative, etc. certainly point to some kind of

125   historical connection between them.

126

127   Nevertheless, other researchers are extremely sceptical that the oral variants held up

128   by folkorists can be regarded as "independent" descendents of the pre-Perraudian oral

129   tradition. Instead, they suggest that, like the Brothers Grimm version, these tales are more

130   likely to be vernacular interpretations of published texts. For example, in an essay that

131   strongly resonates with Bottigheimer's ideas, Hüsing (1989) writes that Little Red Riding

132   Hood "represents one of the loveliest French literary tales, perhaps being the most successful

133   fake that we have in the entire genre", which nonetheless lacks the characteristic stylistic

134   features of authentic oral fairy tales (such as incompleteness). Similarly, Berlioz (1991) and,

135   indeed, Bottigheimer herself (2010, p. 64), argue that there is no evidence to suggest that

136   Little Red Riding Hood existed in oral tradition prior to the publication of Perrault's *Contes* at

137   the end of the seventeenth century.

138

139   In this paper, we aim to shed more light on these issues by taking a quantitative

140   stemmatological approach to investigate the relationships between oral and literary traditions

141   of Little Red Riding Hood. Our study builds on Tehrani's (2013) recent phylogenetic analyses

142   of the ATU 333 type tales, which investigated the relationships between oral European

143   variants (plus Perrault and Grimm) to similar stories from other parts of the world, especially

144   Africa and East Asia. Tehrani's study did not, however, address the question of whether Little

145   Red Riding Hood originated in an oral or literary medium, nor did it examine interactions

146   between the two traditions of ATU 333. Below, we outline how these issues were tackled in

147   this study.

148   **2. Materials**

149

150   A total of 23 texts of Little Red Riding Hood were selected for analysis (see 'Sources' in

151   Appendix A). To be clear, the aim of the analyses was not to produce a comprehensive

152   stemma of the Little Red Riding Hood tradition – which would involve hundreds, if not

153   thousands of texts – but to investigate a specific problem concerning the relationship of oral

154   versions of the tale to literary versions. Specifically, we sought to test whether Perrault based

155 his tale on a pre-existing oral tradition, or if both the oral and literary traditions derive from

156 the classic versions of Perrault and the Grimms published in the seventeenth and nineteenth

157 centuries respectively.

158

159   Our dataset included 12 Franco-Italian oral tales collected in the nineteenth and

160 twentieth centuries that cover most of the major variations in the plot and character found in

161 the folk traditions of these regions. For example, in some cases Little Red Riding Hood lacks

162 her characteristic red hood and is simply described as a young girl. In many variants the

163 protagonist outwits the villain to escape, but in others she is eaten. The character of villain,

164 meanwhile, can take several forms, such as a wolf, witch or werewolf. In one group of Italian

165 tales (three of which are included here) known as 'Catterinetta' – formerly categorized as a

166 distinct subtype of ATU 333 (Aarne & Thompson, 1961) – the villain is actually the relative

167 that the girl went to visit (usually an aunt or uncle). She/he takes revenge on the girl for eating

168 the food that was in her basket and replacing them with cakes made from donkey dung. The

169 dataset also included Egbert's 11th century poem, the classic versions of Little Red Riding

170 Hood published by Perrault and the Brothers Grimm in the seventeenth and nineteenth

171 centuries respectively, five examples of literary versions of Little Red Riding Hood from the

172 late nineteenth and early twentieth centuries sampled from the deGrummond's Children's

173 Literature Research Collection curated by the University of Southern Mississippi

174 (http://www.usm.edu/media/english/fairytales/lrrh/lrrhhome.htm), and three oral variants

175 from beyond the hypothesised ATU 333 cradle (two from Portugal and one from Lusatia in

176 modern day Poland) that are thought to be based on literary texts, and which provide another

177 useful point of comparison with the Franco-Italian oral versions.

178

179   Next, we constructed a matrix that coded the presence or absence of 58 traits (or, in

180 phylogenetic parlance, "characters") identified in the 23 texts. The traits included features

181 such as the red hood worn by the girl, the character of the wolf, the girl being eaten and so on

182   (the full list of characters and the matrix are provided in Appendix A). The matrix only

183   included traits that occurred in at least two tales, which might give clues about common

184   ancestry. Traits that occurred in just a single text were excluded, since these would not be

185   informative about relationships.

186

187   The matrix was analysed using several methods of phylogenetic/stemmatic

188   reconstruction, each of which are described in the sections below. We predicted that, if the

189   oral origins hypothesis is correct, then the literary tradition instigated by Perrault and also

190   comprising the Grimms' *Rotkäppchen*, later published versions and oral copies from Portugal

191   and Lusatia, should constitute a distinct lineage nested within a larger family of Franco-Italian

192   folktales. Conversely, if the latter are derived from textual sources, they would be expected to

193   comprise a lineage (or lineages) that split off from the literary tradition instigated by Perrault

194   and continued by the Brothers Grimm. In the last analysis we introduce a method,

195   PhyloDAG, that directly tests for ancestor-descendent relationships, while also allowing us to

196   incorporate contamination between texts and/or oral traditions.

197

198   **3. Phylogenetic Tree Analysis**

199

200   Our first analysis employed the most-widely used method for reconstructing relationships

201   among texts in stemmatology, maximum-parsimony (Howe et al. 2001). Maximum

202   parsimony involves finding the tree(s) that minimises the number of evolutionary changes

203   required to explain shared traits among a group of taxa (in this case, versions of Little Red

204   Riding Hood) under a branching model of descent with modification. We carried out the

205   maximum parsimony analysis in the software program PAUP 4.0* (Swofford, 1998). The

206   results are shown in Figure 1.

207

208   **Fig. 1 "Parsimony tree" about here.**

209

210   The tree is rooted using the oldest text, Egbert's 11th century poem ("Latin"), as an outgroup.

211   Under the oral origins hypothesis, Egbert's text represents the earliest known witness of the

212   oral tradition of ATU 333 prior to Perrault, so it can be assumed that all the other texts (both

213   oral and literary) are descended from a common ancestor of more recent origin. Under the

214   literary origins hypothesis, Egbert's text would be excluded from the Little Red Riding Hood

215   tradition, which is assumed to have originated six centuries later. Thus, both hypotheses

216   would position Egbert's text as an outgroup with respect to the other texts.

217

218   The tree indicates that the literary versions of Little Red Riding Hood form a clade, or branch,

219   that also includes the three oral "copies" from Portugal and Lusatia, as well as an Italian tale

220   called *Three Girls*. Although the latter is technically a folktale, it is much closer to literary

221   versions of ATU 333 than traditional versions of 'The Story of Grandmother' (for example,

222   the girl is eaten and then subsequently cut out of the wolf's stomach), and is probably derived

223   from published texts. The literary clade forms part of a larger grouping that comprises

224   variants of the Franco-Italian tale 'The Story of Grandmother', but excludes variants of the

225   Italian 'Catterinetta' tale (represented by *Catterinetta, Serravalle* and *UncleWolf*), which form

226   a separate lineage splitting off at the root of the tree. Thus, as predicted by the oral origins

227   hypothesis, the results of the maximum parsimony analysis suggest that the literary texts

228   share a last common ancestor (LCA) of more recent origin than the LCA of the oral variants.

229

230   It is worth noting, however, that there are some inconsistencies between the tree and existing

231   knowledge and theories about the Little Red Riding Hood tradition. For example, one of the

232   literary variants (*Goldenhood*) and a Portuguese oral "copy" (*Consigliere*) form a clade that

233   appears to be descended from a common ancestor of more ancient origin than Perrault. Since

234   the literary tradition is known to have originated with Perrault, this anomaly can probably be

235   attributed to an error of the maximum parsimony estimation, possibly as a consequence of

236   contamination (or "reticulation" in phylogenetic jargon) between the literary and oral

237   traditions. Contamination is likely to be common in fairy tale traditions as multiple oral and

238   literary versions of a tale may circulate at the same time within and between geographical

239   areas, and sometimes get mixed together (e.g. Tehrani 2013). Since the underlying model

240   used in maximum parsimony analysis does not explicitly allow for horizontal transmission

241   across lineages, it can sometimes erroneously interpret similarities that result from this

242   process as primitive traits (i.e. the traits exhibited by the hybrid taxon are assumed to be

243   inherited from an ancestral taxon that existed before the lineages leading to the two donor

244   taxa split), thereby "dragging" highly contaminated variants deeper into the structure of the

245   tree. This effect might similarly explain the position of one of the oral variants, *Joisten*, which

246   is claimed to have borrowed traits from literary texts (Zipes, 1993, pp. 5-6), but appears in

247   this tree to have split off from the LCA of the oral and literary tradition prior to the

248   emergence of the latter. Another issue with maximum parsimony analysis is that it focuses

249   solely on reconstructing collateral phylogenetic relationships (i.e. relationships based on

250   common descent), rather than ancestor-descendent relationships. Consequently, it is not clear

251   from the tree whether the position of Perrault should be interpreted as ancestral or collateral

252   with respect to the other literary variants, while the position of the Grimm text is similarly

253   ambiguous. These examples highlight the need to be cautious in drawing strong conclusions

254   from the topology of the parsimony tree, or indeed other methods that assume a pure

255   branching model of evolution.

256

257   **4. Network Analysis**

258

259   Phylogenetic networks provide an alternative approach to reconstructing cultural and

260   biological evolution where relationships are not strictly tree-like. A number of methods for

261   detecting different kinds of reticulation events have been proposed (Morrison, 2011). Many of

262   the methods are specific to certain mechanisms, for instance, recombination and therefore not

263   necessarily appropriate for modeling fairy tale traditions where the blending process is rather

264   poorly understood and probably varies significantly from case to case.

265

266   Below, we present results from two popular network methods, NeighborNet and T-

267   Rex. In addition, we present a new method, PhyloDAG, which is based on maximum

268   likelihood analysis and allows generic directed networks or DAGs (directed acyclic graph).

269   We also apply a parametric bootstrap test to compare a number of network hypotheses

270   obtained by the PhyloDAG method.

271   *4.1 NeighborNet Analysis*

272

273   A popular method for studying data that may involve reticulation is NeighborNet (Bryant &

274   Moulton, 2003), (Huson & Bryant, 2006). In the terminology of Morrison (2011),

275   NeighborNet is a data-display method. In other words, it does not attempt to construct a

276   genealogical hypothesis that accurately represents the actual evolutionary history. Rather it

277   attempts to represent the possibly conflicting phylogenetic signals in the data, so that non-

278   tree-like structures may result either by actual reticulation or by other mechanisms such as

279   evolutionary reversal or convergent evolution. Neither does the NeighborNet attempt to

280   suppress statistically insignificant signals in the data which tends to result in very complex

281   networks with a large number of non-tree-like structures.

282

283   Figure 2 shows the NeighborNet obtained for the data in our study by using the

284   SplitsTree4 software[1]. The network shows similar clusters to the maximum parsimony

285   analysis, distinguishing the literary variants (including the Portuguese and Lusatian oral

286   copies) from Franco-Italian oral versions of 'The Story of Grandmother' and versions of the

287   Italian 'Catterinetta' tale, which form a separate group. The "boxiness" of the network

288   suggests probable lines of contamination within and between these sub-groups. However, the

289   network has the typical problem associated with this method, which is that the middle part of

290   the network is a very complex dense mesh of interconnected points that correspond to various

291   weak conflicting signals in the data. Furthermore, all the most of the extant versions (the

292   labelled points) are at the end of a long edge, suggesting that none of them (except perhaps

293   one root node) are ancestors of the others. This makes is very hard to interpret the result in a

294   way that would be informative for the questions we are presently considering. In particular,

295   we can tell almost nothing from the network about the influence of Perrault and the Brothers

296   Grimm on the oral tradition, or vice versa.

297

298   **Fig. 2 "NeighborNet" about here.**

299   *4.2 T-Rex Analysis*

300

301   Another technique from phylogenetics that can be used to model reticulation is T-Rex (Boc,

302   Diallo, & Makarenkov, 2012). It starts from a tree structure and by comparing the pairwise

303   distances computed from the data to the distances expected based on the tree, it identifies

304   parts of the tree that fail to accurately match the distances in the data. In case certain groups

305   of taxa are more similar to each other than the tree would lead us to expect, a reticulation

306   edge may be introduced. The underlying tree structure is obtained by Neighbor-Joining

307   (Saitou & Nei, 1987). The number of reticulation edges can be chosen by the user. We chose

308   to include five of them in an attempt to discover the most significant contamination events.

309

310   The result of the T-Rex analysis is shown in Figure 3. The backbone phylogeny is

311   largely similar to the parsimony tree, and indicates that the literary versions of Little Red

312   Riding Hood form a branch that split from the lineage leading to modern oral variants of the

313   traditional Franco-Italian tale 'The Story of Grandmother'. Versions of the Italian tale

314   'Catterinetta' form a sister group to these tales. One notable difference between the T-Rex

315 phylogeny and the parsimony tree is the position of *ThreeGirls*. As mentioned above,

316 *ThreeGirls* is an Italian oral tale that shares notable features in common with the

317 Grimms'*Rotkäppchen*. Whereas the parsimony analysis indicated that *ThreeGirls* was likely

318 to be derived from literary texts (as per the Portuguese and Lusatian oral versions of ATU

319 333), T-Rex suggests that *ThreeGirls* is descended from an oral ancestor that preceded the

320 literary tradition, but has been contaminated by the latter (N.B. although the reticulation edges

321 in T-Rex are undirected, the well-documented influence of literary fairy tales – particularly

322 the Grimms' *Kinder und Hausmärchen* – on European oral traditions (Zipes, 2013) support

323 this interpretation). This is consistent with the NeighbourNet graph, which grouped

324 *ThreeGirls* with oral variants, but indicated substantial conflict in the data surrounding its

325 relationships to other tales. The T-Rex analysis proposed several other reticulation edges that

326 suggest substantial mixing within regions between literary and oral traditions of ATU 333,

327 notably between Perrault's classic text and French oral tales, and between the Italian variants

328 of 'The Story of Grandmother' and 'Catterinetta'. More puzzlingly, the structure also

329 suggests contamination from the Egbert's medieval poem and a modern literary version of

330 Little Red Riding Hood (*CupplesLeon*). Since a careful reading of both texts revealed no

331 obvious link between them (e.g. characteristic features of the medieval version that occur in

332 *CupplesLeon* but not in the Perrault or Grimm tales from which it is certainly derived)) we

333 assume this to be an estimation error (the precise cause of which would require a more

334 detailed deconstruction of the search algorithm that is beyond the scope of the current paper).

335 A more general problem with the interpretation of the results of the T-Rex analysis is that,

336 like the parsimony and NeighbourNet structures, all the variants are represented as leaf nodes.

337 Consequently, it is not easy to evaluate direct lines of descent between historical and modern

338 variants, most particularly the relationships of Perrault and the Brothers Grimm to literary and

339 oral tales that were published/recorded more recently.

340

341 **Fig. 3 "T-Rex" about here.**

342 *4.3 PhyloDAG*

343

344 We will now propose an alternative approach to network analysis. Our approach is likelihood

345 based and, as we will show below, it solves many of the issues in existing network and tree-

346 based methods.

347

348   Likelihood based phylogenetic inference involves a probabilistic sequence evolution

349 model characterizing the evolutionary process. A popular example of such a model is the

350 Jukes-Cantor model (Jukes & Cantor, 1969) that gives the probability of the four DNA

351 symbols, A,T,G, and C, changing into other symbols or remaining unchanged in a certain

352 period of time, and also depending on the mutation rate. Given such a model, the likelihood

353 of a phylogenetic tree is obtained as the probability that the observed data sequences are

354 produced when the tree structure is fixed and the lineages evolve independently according to

355 the sequence evolution model and branching occurs according to the tree structure. The

356 maximum likelihood method for phylogenetic inference attempts to find the tree structure,

357 including the edge lengths that determine the expected amount of change along each edge, for

358 which the likelihood is the highest possible.

359

360   Strimmer and Moulton (2000) describe a simple extension of the likelihood defined

361 for phylogenetic trees that is also applicable to networks, hence allowing reticulation edges to

362 be added into a tree. We improve and extend the method by Moulton and Strimmer in two

363 ways. First, we introduce a more efficient technique for approximating the likelihood of

364 phylogenetic network. Second, we propose a simple search procedure that considers

365 additional reticulation edges in a given tree structure and also estimates the edge lengths by a

366 simple sampling technique. As a result, our method which we call PhyloDAG operates in a

367 similar fashion as T-Rex: it takes as input a matrix of character data such as DNA sequences

368 or a set of features, and an initial tree structure, and produces a network where a given

369 number of reticulation edges have been added to the tree, together with its likelihood value. In

370   contrast to T-Rex, however, PhyloDAG can be used to evaluate tree and network structures

371   where some of the extant taxa are placed at internal nodes so that they represent ancestors of

372   some of the other taxa. For a more detailed description of the PhyloDAG method, see

373   Appendix B. Different network or tree structures can be compared using a statistical test

374   known as the parametric bootstrap, which we will also outline below, see Appendix C.

375

376   We start the PhyloDAG method with a parsimony tree, Fig. 1, obtained from data

377   matrix in Table II. We then use PhyloDAG to evaluate its likelihood (setting the number of

378   reticulation edges to zero). The parsimony tree yields log-likelihood the value –863.4. [2]

379

380   Next, we manipulated the topology of the tree to explore different scenarios

381   concerning the origins of the literary and oral traditions of ATU 333. This involved moving

382   the Perrault and Grimm texts into different internal positions in the tree where they would be

383   either ancestral to both the oral and literary variants, or ancestral to the literary variants and

384   collateral to the oral variants (i.e. descended from a common oral ancestor). We did not

385   attempt manipulations which are incompatible with existing knowledge about the tales, such

386   as the chronology of the literary variants (for example, we did not experiment with making

387   Grimm's 1812 tale ancestral to Perrault's 1697 version). It is important to note that these

388   manipulations alone will not, as a rule, yield a higher likelihood score than a normal tree. This

389   is because any such manipulated tree is equivalent to a special case of a tree where the taxon

390   in the internal position is in fact a leaf node but the edge pointing it has length zero. Hence,

391   the likelihood value of the tree where the taxon is a leaf node will never be lower than the

392   likelihood of the tree where it is an internal node when the edge lengths in both models are

393   optimimized so as to maximize the likelihood. The interesting question is whether a

394   hypothesis involving observed ancestral taxa is better when we allow possible contamination,

395   i.e., reticulation edges in addition to the tree. The PhyloDAG method provides a tool for

396   answering this question.

397

398   We used PhyloDAG to search for reticulation edges that improve the likelihood

399   score. As a starting point for the search, we use different variations of the parsimony tree

400   (Fig. 1) where either Perrault or Grimms is moved into an ancestral position, considering a

401   number of different nearby positions just above or next to the position of the said taxa in the

402   parsimony tree. The search produced 11 alternative structures, which we label by *a, b, c, d, e,*

403   *f, g, h, i, j,* and *k.* Figures 5 and 6 show respectively networks *c* and *d,* which are of particular

404   importance for our discussion below. The other networks are given for completeness in

405   Appendix D.[3]

406

407   As an indication of how well the models "fit" the data, we report the log-likelihood

408   value of each of the models. For example, the log-likelihood of network *c* is –862.4, and the

409   log-likelihood of network *d* is –865.5. Networks *b, c* and g achieve a higher log-likelihood

410   value than the parsimony tree (–863.4). However, the likelihood values should not be taken to

411   be the final evaluation of the models because of two reasons. First, the likelihood evaluation

412   is approximate due to the random sampling procedure included in the method (see Appendix

413   B). Second, perhaps more importantly, the log-likelihood score tends to favor complex

414   models because they have more adjustable parameters that make it easier to achieve high log-

415   likelihood values for most data sets. To provide a statistically sound goodness-of-fit measure,

416   below we propose to use a parametric bootstrap technique.

417   *4.4 Parametric Bootstrap*

418

419   It is important to note that a network hypothesis is typically more complex than a tree

420   hypothesis (it has more parameters), which may lead to so called over-fitting: choosing a too

421   complex hypothesis considering its statistical support. To avoid over-fitting, we applied a

422   parametric bootstrap test to compare the tree hypotheses and the different network

423   hypotheses; for more details, see Appendix C.

424

Table I summarizes the results of the bootstrap test. The results are not unanimous
but there is a relatively strong (considering the small sample size) signal indicating that
models b, c, and g have the best statistical support. Among them, model c (fourth row in
Table I, and Fig. 4) fares especially well, and is only rejected with low statistical confidence
when compared to models b and g, while the latter two are both rejected in more
comparisons. All three models place *Perrault* in an internal position that makes it ancestral to
all the literary variants. However, there is some disagreement regarding the position of the
Grimms' tale: Model b (see Appendix D) has *Grimm* as a terminal node, whereas both c and g
place *Grimm* as an ancestral source for subsequent literary versions. Although the bootstrap
test was unable to discriminate between these possibilities, previous research into the history
of Little Red Riding Hood strongly support the latter scenario (Zipes, 1993).

TABLE I. STATISTICAL HYPOTHESIS TEST RESULTS (PARAMETRIC BOOTSTRAP). ROWS: NULL HYPOTHESIS.
COLUMNS: ALTERNATIVE HYPOTHESIS. 'tree': PARSIMONY TREE. '·': NOT REJECTED. '+': REJECTED AT
SIGNIFICANCE LEVEL 0.05. '*': REJECTED AT SIGNIFICANCE LEVEL 0.01.

NULL HYPOTHESIS	ALTERNATIVE HYPOTHESIS											
	tree	a	b	c	d	e	f	g	h	i	j	k
tree		*	*	*	*	+	*	*	*	·	*	·
a	·		*	*	*	*	*	*	*	·	*	*
b	·	·		+	+	+	+	+	+	·	*	+
c	·	·	+		·	·	·	+	·	·	·	·
d	·	+	*	*		+	·	*	·	·	*	+
e	+	*	*	+	*		*	*	+	·	*	*
f	+	*	*	*	·	*		*	+	·	+	·
g	+	·	+	·	*	*	*		·	·	+	·
h	*	*	*	*	*	*	*	*		*	*	*
i	*	*	*	*	*	*	*	*	*		*	*
j	*	*	*	*	*	*	*	*	·	·		*
k	*	*	*	*	*	*	*	*	·	·	*	

Fig. 4 "PhyloDAG network c" about here.

445   More significantly, all three models *b*, *c*, and *g* are consistent with the oral origins

446   hypothesis. The literary tradition instigated by Perrault (placed as an internal node in all three

447   models) is represented as an offshoot of a lineage that also gave rise to the French and Italian

448   tale 'The Story of Grandmother'. The models further suggest that the variants of the Italian

449   tale of *Catterinetta* comprise a separate group that split from the other oral and literary

450   variants prior to *Perrault*. However, the models show that these various subgroups of ATU

451   333 did not develop in isolation of one another. All three indicate contamination both within

452   and between the literary and oral traditions of the tale. For example, like the T-Rex structure,

453   models *b*, *c*, and *g*, all suggest reticulation played an important role in the tale *ThreeGirls*.

454   However, whereas the T-Rex analysis suggested that *ThreeGirls* was descended from an oral

455   ancestor that preceded the first written versions of Little Red Riding Hood, the PhyloDAG

456   models are more consistent with the parsimony results, which situated the tale within the

457   literary group. Specifically, models *b*, *c*, and *g*, indicate that *ThreeGirls* is descended from the

458   Grimm's text, which was mixed with elements from oral tradition (notably the Italian

459   *Catterinetta* tale, as shown in models *c* and *g*, with which it shares distinctive motifs like

460   angering the villain by replacing the contents of the basket). Contamination also appears to be

461   evident in the Portuguese tale *Consigliere* and French literary tale *Goldenhood*, which might

462   explain their anomalous positions in the parsimony tree, which made them a sister clade to the

463   Perrauldian literary tradition. As explained earlier, reticulation can be a major source of error

464   in inferring phylogenetic trees, for example by dragging affected taxa deeper into the

465   structure of the tree. By incorporating reticulation edges in PhyloDAG, we found that models

466   in which Perrault was ancestral to *Consigliere* and *Goldenhood* fitted the data much better

467   than models in which these tales formed a sister clade, i.e. *a* and *e*, which were rejected in all

468   the bootstrap comparisons with every other model except one (*i*, discussed below).

469

470   We analysed six structures that supported the alternative literary origins hypothesis.

471   Among them, the one that is best supported by the data – albeit not as well as the oral origins

472   models, *b, c,* and *g* – is model *d*, see Fig. 5. The other network structures are given in

473   Appendix D. Models *f, i* and *k* represent Perrault as the ancestor of all modern versions of

474   ATU 333, including the literary variants and the oral tales 'The Story of Grandmother' and

475   'Catterinetta'. Model *f* represents the Grimm tale as a leaf node, while in *i* and *k* the Grimm

476   tale is shifted into different internal positions within the PhyloDAG. In the bootstrap

477   comparisons, all three models are rejected against the tree and the oral origin scenarios

478   represented in *b, c* and *g*. Models *d, h* and *j* represent Perrault as the ancestor of the literary

479   variants of Little Red Riding Hood and the oral tale 'The Story of Grandmother', but not of

480   versions of 'Catterinetta', which consistently come out as a sister group to the other tales in the

481   analyses. The Grimm tale is positioned as a leaf node in model *d* and as an internal node in *h*

482   and *j*. Model *d* is supported against the parsimony tree, but rejected with high statistical

483   support against the oral origins models *b, c,* and *g*. Models *h* and *j* are rejected in all the

484   comparisons.

485

486   **Fig. 5 "PhyloDAG network d" about here.**

487

488   In sum, the inclusion of lineal and reticulate relationships using PhyloDAG produced

489   a number of structures that fit the data better than the parsimony tree. Structures consistent

490   with the oral origins hypothesis were less frequently rejected in the bootstrap comparisons

491   than those that are consistent with the literary origins hypothesis, with all three of the top

492   performing models (*b, c* and *g*) falling into the former category. However, it should be noted

493   that the evidence from the bootstrap test comparisons is not all in one direction, since models

494   *b* and *g* (oral) are rejected against *d* and *f* (literary). On the other hand, model *c* (oral) is

495   supported with high statistical confidence against both literary origins models. Thus, overall,

496   the results of the PhyloDAG analyses indicate that the literary tradition of Little Red Riding

497   Hood has its roots in oral folktales, rather than the other way around.

498

5. Conclusions

Our aim in this paper has been to shed light on a complex question in the historiography of fairy tales: is it possible to identify whether particular stories originated as traditional folktales or authored texts? We have proposed that a useful strategy for addressing this question is to adopt the kind of quantitative, computational approach that has been so successfully used to reconstruct manuscript stemmata. Our case study focused on testing two long-standing competing hypotheses about the origins of Little Red Riding Hood. The first suggests the tale originally evolved in French and Italian oral tradition, adapted by Charles Perrault in the late seventeenth century, and subsequently copied by The Brothers Grimm to establish the classic form of the tale found in present day popular culture. The second hypothesis proposes that the tale was a literary invention in the first place, and that "traditional" variants collected by folklorists are actually adaptations of Perrault's and Grimm's texts.

We initially tested these hypotheses by analysing 23 oral and literary variants of Little Red Riding Hood/ATU 333 using one the most popular methods in computer-assisted stemmatology – maximum parsimony analysis. While the general structure of the tree returned by this analysis seemed to be more compatible with the oral origins hypothesis than the literary origins hypothesis, this conclusion is mitigated by two problems with interpreting the results: firstly, maximum parsimony does not incorporate reticulation (contamination), which can lead to errors in estimating phylogenetic relationships; secondly, the method does not model lineal (ancestor-descendent) relationships among observed taxa, making it difficult to draw firm conclusions about the role of classic historic texts (i.e. Perrault and Grimm) on contemporary literary and oral variants. Alternative methods for modelling reticulate evolution, such as NeighbourNet and T-Rex, provide a means for addressing the first of these problems but not the second. As such, their usefulness for addressing the question in hand

526 turned out to be limited. We therefore introduced a new approach – PhyloDAG – which

527 handles both lineal and reticulate relationships in a statistically sound way. This enabled us to

528 compare different models for the evolution of Little Red Riding Hood and directly test the

529 oral hypothesis against the literary hypothesis. Our results pointed strongly toward the former,

530 with the best models indicating that Perrault adapted his tale from oral folktales, rather than

531 vice versa.

532

533   Of course, we cannot extrapolate any general conclusions about the origins of fairy

534 tales from a single case study. It is entirely possible – likely, even – that other tales originated

535 in a literary medium before passing into oral tradition, as suggested by Bottigheimer. What

536 we have shown here is that the problem of establishing these facts is far from intractable, and

537 can be solved using principled and powerful computational methods. We anticipate that the

538 application of these methods will generate new insights into the origins and development of

539 different types of fairy tale, as well as other kinds of cultural traditions (Lipo, O'Brien,

540 Collard, & Shennan, 2006; Mace, Holden, & Shennan, 2005).

541
542
543

544 **Endnotes**

1 The SplitsTree4 software is available at www.splitstree.org .
2 We follow the convention to give likelihood values in logarithmic scale, so that probabilities, which are always less than one, become negative numbers.
3 We chose to include all 11 networks in order to give an indication of the range of possible network hypotheses we considered and to quantify the statistical uncertainty by means of the bootstrap test.
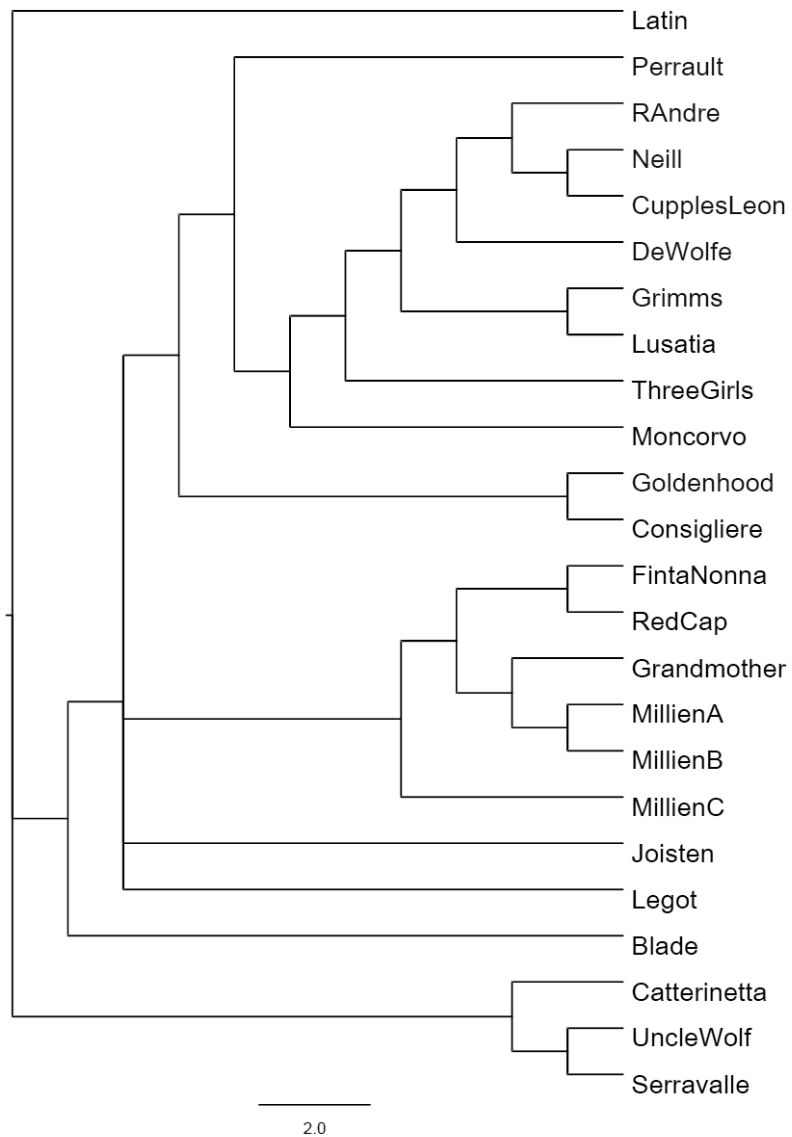
545 **References**
546
547 Aarne, A., & Thompson, S. (1961). *The Types of the Folktale. A Classification and*
548   *Bibliography* (Vol. 3). Helsinki: FF Communications.
549 Ben-Amos, D., Ziolkowski, J. M., Silva, F. Vaz da., & Bottigheimer, R. (2010). Special Issue:
550   The European Fairy-Tale Tradition between Orality and Literacy. *Journal of*
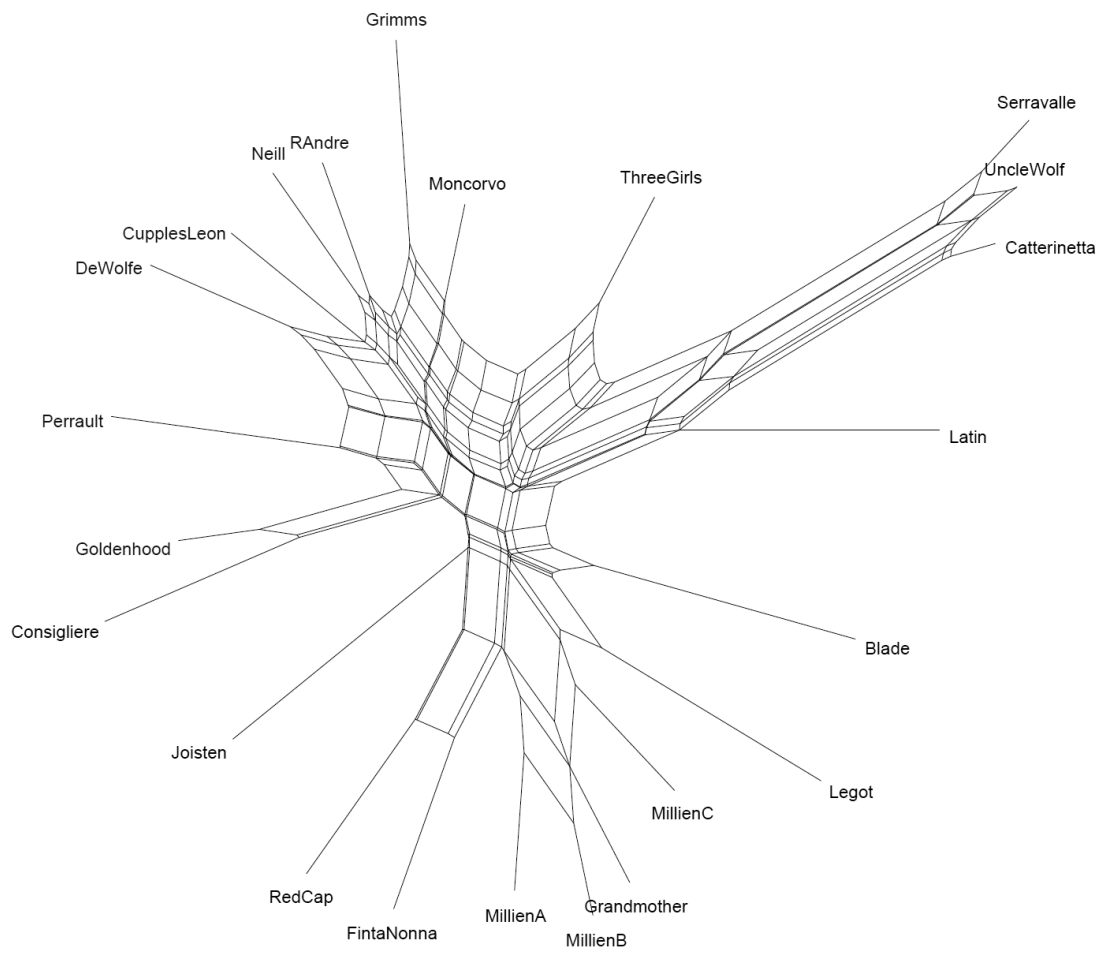551   *American Folklore, 123*(490).

552   Berlioz, Jaques. (1991). Un Petit chaperon rouge médiéval? 'La petite fille épargnée pa les
553   loups' dans la Fecunda ratis d'Egbert de Liège (début du XIe siècle). *Marvels and*
554   *Tales, 5*(2), 246–262.
555   Boc, Alix, Diallo, Alpha Boubacar, & Makarenkov, Vladimir. (2012). T-REX: a web server
556   for inferring, validating and visualizing phylogenetic trees and networks. *Nucleic*
557   *Acids Research, 40*(W1), W573-W579. doi: 10.1093/nar/gks485
558   Bottigheimer, R.B. (2002). *Fairy Godfather: Straparola, Venice, and the Fairy Tale*
559   *Tradition*: University of Pennsylvania Press, Incorporated.
560   Bottigheimer, R.B. (2010). *Fairy Tales: A New History*: State University of New York Press.
561   d'Huy, J. (2013). A phylogenetic approach to mythology and its archaeological consequences.
562   *Rock Art Research 30*(1), 115-118.
563   Delarue, P. (1951). Les contes marveilleux de Perrault et la tradition populaire: I. Le petit
564   chaperon rouge. *Bulletin folklorique d'Ile-de-France*, 221-228, 251-260, 283-291.
565   Grimm, J, & Grimm, W. (1812). *Children's and Household Tales*. Gottingen.
566   Haar, B.J. (2006). *Telling Stories: Witchcraft And Scapegoating in Chinese History*: Brill
567   Academic Pub.
568   Howe, C. J., Barbrook, A. C., Spencer, M., Robinson, P., Bordalejo, B., & Mooney, L. R.
569   (2001). Manuscript evolution. *Trends Genet, 17*(3), 147-152.
570   Husing, G. (1989). Is Little Red Riding Hood a Myth? In A. Dundes (Ed.), *Little Red Riding*
571   *Hood: A Casebook* (pp. 64-71). Madison: University of Wisconisn Press.
572   Huson, Daniel H., & Bryant, David. (2006). Application of Phylogenetic Networks in
573   Evolutionary Studies. *Mol Biol Evol, 23*(2), 254-267. doi: 10.1093/molbev/msj030
574   Lipo, C., O'Brien, M., Collard, M., & Shennan, S. J. (Eds.). (2006). *Mapping our ancestors:*
575   *phylogenetic approaches in anthropology and prehistory*. New Brunswick: Aldine
576   Transaction.
577   Mace, R., Holden, C., & Shennan, S. (Eds.). (2005). *The Evolution of Cultural Diversity – A*
578   *Phylogenetic Approach*. London: UCL Press.
579   Morrison, David. (2011). *Introduction to Phylogenetic Networks*. http://www.rjr-
580   productions.org/Networks/index.html: RJR Productions.
581   Perrault, C. (1697). *Histoires ou Contes du temps passé*.
582   Roos, Teemu, & Heikkilä, Tuomas. (2009). Evaluating methods for computer-assisted
583   stemmatology using artificial benchmark data sets. *Literary and Linguistic*
584   *Computing, 24*(4), 417-433. doi: 10.1093/llc/fqp002
585   Rumpf, M. (1989). *Little Red Riding Hood, A Comparative Study* (Vol. 17). Bern: Artes
586   Populares.
587   Saintyves, Paul. (1989). Little Red Riding Hood or The Little May Queen. In A. Dundes
588   (Ed.), *Little Red Riding Hood: A Casebook* (pp. 71-88). Madison: Wisconsin
589   University Press.
590   Stubbersfield, Joseph, & Tehrani, Jamshid. (2013). Expect the Unexpected? Testing for
591   Minimally Counterintuitive (MCI) Bias in the Transmission of Contemporary
592   Legends: A Computational Phylogenetic Approach. *Social Science Computer Review,*
593   *31*(1), 90-102. doi: 10.1177/0894439312453567
594   Swofford, D.L. (1998). PAUP* 4. Phylogenetic Analysis Using Parsimony (*and Other
595   Methods). Version 4. Sunderland: Sinauer.
596   Tehrani, Jamshid J. (2013). The Phylogeny of Little Red Riding Hood. *PLoS ONE, 8*(11),
597   e78871. doi: 10.1371/journal.pone.0078871
598   Verdier, Yvonne. (1978). Le Petit Chaperon Rouge dans las tradition orale. *Cahiers de*
599   *Litterature Orale, 4*, 17-55.
600   Ziolkowski, J. M. (1992). A fairy tale from before fairy tales: Egbert of Liege's "De puella a
601   lupellis seruata" and the medieval background of "Little Red Riding Hood".
602   *Speculum, 67*(3), 549-575.
603   Zipes, J. (1993). *The Trials and Tribulations of Little Red Riding Hood*. New York:
604   Routledge.

605   Zipes, J. (2013). *The Golden Age of Folk and Fairy Tales: From the Brothers Grimm to*
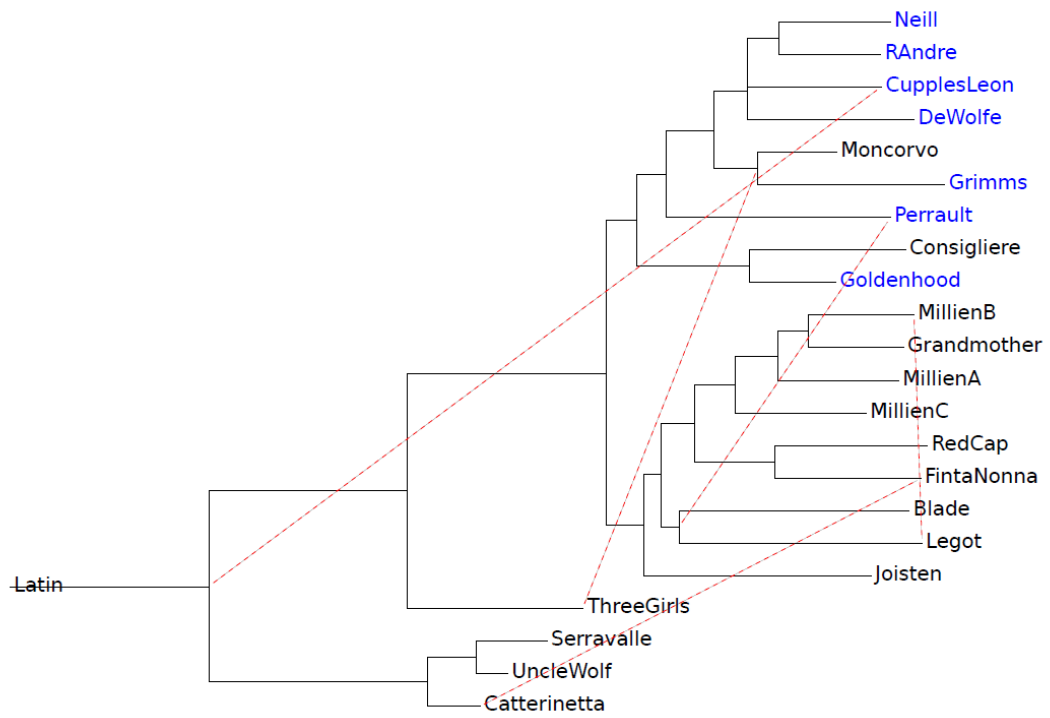606   *Andrew Lang*: Hackett Publishing.
607
608
609

610 **Figures**

Latin
Perrault
RAndre
Neill
CupplesLeon
DeWolfe
Grimms
Lusatia
ThreeGirls
Moncorvo
Goldenhood
Consigliere
FintaNonna
RedCap
Grandmother
MillienA
MillienB
MillienC
Joisten
Legot
Blade
Catterinetta
UncleWolf
Serravalle

2.0

611
612   Fig. 1 Parsimony tree. Log-likelihood −863.4.
613

Grimms

Serravalle

Neill RAndre

UncleWolf

Moncorvo

ThreeGirls

CupplesLeon

Catterinetta

DeWolfe

Perrault

Latin

Goldenhood

Consigliere

Blade

Joisten

Legot

MillienC

RedCap

Grandmother

FintaNonna

MillienA

MillienB

614
615   Fig. 2 NeighborNet. The network is obtained by Splitstree4 (Huson and Bryant, 2006) with
616   default settings.

617

Fig. 3 T-Rex. The underlying Neighbor-Joining tree is shown with solid black lines and five additional reticulation edges are shown with dotted red lines.

618
619
620
621
622
623
624
625

626
627 Fig. 4 PhyloDAG network c. Log-likelihood -862.4.
628
629

630

Fig. 5 PhyloDAG network *d*. Log-likelihood −865.5.

631
632
633
634

Appendix A. Data

635

636

Sources

637

638

Taxon name	Reference
Perrault	Perrault, C. (1697). "Le Petit Chaperon Rouge" *Histoire ou contes du temps passe.*
Grimm	Grimm J. & Grimm W. (1812). "Rotkäppchen". *Kinder- und Hausmärchen*. Gottingen, no. 26
Lusatia	A. H. Wratislaw (1889) "Little Red Hood". Sixty Folk-Tales from Exclusively Slavonic Sources London: Elliot Stock, pp. 97-100
Neill	Neill, J. (1908). *Little Red Riding Hood.* Chicago: Reilly & Lee Co. Downloaded from The University of Southern Mississippi Little Red Riding Hood Project: http://www.usm.edu/media/english/fairytales/lrrh/lrrhhome.htm
Randre	Andre, R. (1888). *Red Riding Hood.* New York: McLoughlin Bros. Downloaded from The University of Southern Mississippi Little Red Riding Hood Project: http://www.usm.edu/media/english/fairytales/lrrh/lrrhhome.htm
CupplesLeon	Gruelle J. B. (1916). *All About Little Red Riding Hood.* New York: Cupples & Leon. Downloaded from The University of Southern Mississippi Little Red Riding Hood Project: http://www.usm.edu/media/english/fairytales/lrrh/lrrhhome.htm
DeWolf	DeWolfe (1890). *Red Riding Hood and Cinderella*. DeWolfe, Fiske, and Co. Downloaded from The University of Southern Mississippi Little Red Riding Hood Project: http://www.usm.edu/media/english/fairytales/lrrh/lrrhhome.htm
Goldenhood	Marelles, C. 1895. "The True Story of Little Goldenhood". *Andrew Lang, The Red Fairy Book, 5th edition.* London and New York: Longmans, Green, & Co. pp. 215-19
Consigliere	Vaz da Silva, F. (1995). Capuchinho vermelho em Portugal. Estudos de Literatura Oral 1, p. 38-58
Moncorvo	Vasconcellos, L. (n.d.) "O Chapelinho Encarnado". Translated by Sara Silva. Courtesy of Isabel Cardigos and the Centro de Estudos Ataíde Oliveira
ThreeGirls	Calvino, I. (1956, trans. 1980 by G. Martin) "The Wolf and the Three Girls". *Italian Folktales.* Harmondsworth: Penguin, pp.26-27
MillenA	Millen, A. (1887). 'Little Red Riding Hood: Version 1'. Zipes, J. 2013. *The Golden Age of the Folk and Fairy Tales*. Indianapolis: Hackett. P 170-1

MillenB	Millen, A. (1887). 'Little Red Riding Hood: Version 2' zipes, J. 2013. *The Golden Age of the Folk and Fairy Tales*. Indianapolis: Hackett. P 172
MillenC	Millen, A. (1887). 'The Little Girl and the Wolf' zipes, J. 2013. *The Golden Age of the Folk and Fairy Tales*. Indianapolis: Hackett. P 173
Grandmother	Delarue, P. (1956). "The Story of Grandmother". *The Borzoi Book of French Folktales*. New York: Alfred Knopf, pp. 230-233.
FintaNonna	Calvino, I. (1956, trans. 1980 by G. Martin) "The False Grandmother". *Italian Folktales.* Harmondsworth: Penguin, pp.116-117
RedCap	Schneller, C. (1867, trans. 2007 by D. Ashliman). "Cappelin Rosso". *Märchen und Sagen aus Wälschtirol: Ein Beitrag zur deutschen Sagenkunde.*Innsbruck: Verlag der Wagner'schen Universitäts-Buchhandlung, pp. 9-10
Blade	Blade, Jean-Francois. (1886). 'The Wolf and the Child' zipes, J. 2013. *The Golden Age of the Folk and Fairy Tales.* Indianapolis: Hackett. P 169
Legot	Legot M. (1885). 'Little Red Riding Hood: The Version of Tourangelle'. Zipes, J. 2013. *The Golden Age of the Folk and Fairy Tales.* Indianapolis: Hackett. p167
Joisten	Joisten, C. Untitled. Recounted in Zipes, J. (1993) *The Trials and Tribulations of Little Red Riding Hood*. New York: Routledge, pp. 5-6.
Serravalle	Rumpf, M. (1958) "Caterinella: Ein italienisches Warnmärchen," Serravalle variant. ***Fabula*** 1: 76-84
UncleWolf	Calvino, I. (1956, trans. 1980 by G. Martin) "Uncle Wolf". *Italian Folktales.* Harmondsworth: Penguin, pp.49-50.
Catterinetta	Schneller, C. (1867, trans. 2007 by D. Ashliman). "Cattarinetta". *Märchen und Sagen aus Wälschtirol: Ein Beitrag zur deutschen Sagenkunde.*Innsbruck: Verlag der Wagner'schen Universitäts-Buchhandlung, pp. 8-9.
Latin	Ziolkowski, J. (1992) A fairy tale from before fairy tales: Egbert of Liege's "De puella a lupellis seruata" and the medieval background of "Little Red Riding Hood"

639
640
641 **List of characters**
642

1   Protagonist [0] girl [1] boy
2   Girl wears red hood: [0] absent [1] present
3   Who made red hood: [0] absent [1] mother [2] grandmother [3] godfather
4   Girl goes to visit relative: [0] absent [1] granny [2] aunt [3] mother
5   Relative is a witch: [0] absent [1] present [2] fairy]
6   Granny sick [0] absent [1] present
7   Girl told to fetch pan from relative: [0] absent [1] present

8   Girl told not to stay from path: [0] absent [1] present

9   Carries basket: [0] absent [1] present

10   Cargo: bread: [0] absent [1] present

11   Cargo: soup: [0] absent [1] present

12   Cargo: custard: [0] absent [1] present

13   Cargo: butter: [0] absent [1] present

14   Cargo: cakes: [0] absent [1] present

15   Cargo: eggs: [0] absent [1] present

16   Cargo: wine: [0] absent [1] present

17   Girl plays in forest: [0] absent [1] present

18   Girl eats the cargo: [0] absent [1] present

19   Villain is [0] ogre [1] wolf [2] werewolf [3] devil

20   Reconnaissance - villain finds out where the girl is going: [0] absent [1] present

21   Villain and girl take separate paths: [0] absent [1] pins vs needles [2] short vs long

22   Woodcutters are in the forest: [0] absent [1] present

23   Wolf impersonates girl: [0] absent [1] present

24   Grandmother gives instructions on opening door: [0] absent [1] present

25   Girl replaces cargo [0] absent [1] dung [2] nails

26   Monster eats granny: [0] absent [1] present

27   Monster dresses up in grannys clothes: [0] absent [1] present

28   Monster disguises voice: [0] absent [1] present

29   Girl eats remains of granny: [0] absent [1] present

30   Girl eats body parts: [0] absent [1] present [2] refuses

31   Girl eats granny teeth: [0] absent [1] present

32   Girl drinks blood: [0] absent [1] present [2] refuses

33   The girl is warned about the danger: [0] absent [1] by monster [2] by animals

34   Girl flees home boards up house: [0] absent [1] present

35   Monster stalks girl "I'm coming!": [0] absent [1] present

36   Wolf tells girl to take off clothes: [0] absent [1] present

37   Throws clothes into fire: [0] absent [1] present

38   Wolf tells girl to get into bed: [0] absent [1] present

39   Dialogue: [0] absent [1] present

40   My what! Head [0] absent [1] present

41   My what! Arms [0] absent [1] present

42   My what Feet [0] absent [1] present

43   My what! Legs [0] absent [1] present

44   My what! Ears [0] absent [1] present

45   My what! Teeth [0] absent [1] present

46   My what! Eyes [0] absent [1] present

47   My what! Nose [0] absent [1] present

48   My what! Hands [0] absent [1] present

49   My what! Mouth [0] absent [1] present

50   My what! Hairy [0] absent [1] present

51   Girl eaten: [0] absent [1] present

52   Girl cut out of stomach: [0] absent [1] present

53   Girl saved [0] absent] by [1] hunstman [2] woodcutters [3] father [4] mother [5] townsfolk [6] granny

54 Girl saved by magic cloak: [0] absent [1] present [2] magic wand

55 Girl tricks wolf: [0] absent [1] present

56 Wolf chases girl [0] no [1] to her house

57 Wolf killed: [0] absent [1] present

58 Wolf's stomach sewn up with stones inside

643

Matrix

644

645

646 [Character no. 1   10   20   30   40   50   ]
647 Latin   0130000009999999901000000000999000090099999999999909010000
648 Perrault   0121010010011000011211101110999000101101011100000100000000
649 RAndre   0111010010000001001120110111099900009010000111100009300010
650 DeWolfe   0121010010001110101121010011099900009010000111000009100010
651 Neill   0101000110001101101120010101099900009010000111100009300010
652 CupplesLeon   0111000010000000101101010001099900009010000110110009200010
653 Grimms   0121010110000101101120110110099900009010000101011011100011
654 Lusatia   0121010110000101101120110110099900009010000101011011100011
655 Goldenhood   0121000010000100001120000011099900010110100010001109610010
656 FintaNonna   0091001010000000000000001011110000101100010000111100011 00
657 Grandmother   0091000010000000021100001011101200111000000010110900110 0
658 Joisten   0101000010000100101110010100110110009011101010000109101110
659 RedCap   0101000010100000001101101011111100101100011000111100000 00
660 Catterinetta 0092101010000100010000001000099900010900999999999910000000
661 UncleWolf   0092101010000101011000001000099901109009999999999910000000
662 Serravalle   0092101010000101011000001000099901109009999999999911400100
663 ThreeGirls   0093010010000101001110002100099900009011000000000011500010
664 Legot   0091010009999999003120000101100120009110101011000009001110
665 Blade   1092000009999999001100110110100100009110001011000110000000
666 MillienA   0091000011000000001111000100110120011110010101001110000000
667 MillienB   0091000011000000021110001001101200111100001100000 9001100
668 MillienC   0091000011010000001110000100120200009110000110000109001100
669 Consigliere   0121200110000100002120100001000000009110101000001109620010
670 Moncorvo   01?101001000010000112000010100000000911000010100001 1100011

671

672 N.B. the value 9 represents a "gap" state for characters that were redundant or not relevant for a
673 particular tale. For example, if the girl did not carry a basket (character 9) then characters relating to the
674 contents of the basket (10-16) – which logically could not be present – were coded as gap characters

675

676

Appendix B. Description of the PhyloDAG method

Strimmer and Moulton (2000) proposed a likelihood-based method for comparing different phylogenetic hypotheses that correspond to directed acyclic graphs (DAGs). Each node in the graph corresponds to a taxon, either extant or hypothetical (unobserved). The edges in the DAG correspond to direct inheritance where the origin of the edge, the "parent", is the immediate ancestor and the end of the edge, the "child", is the offspring. Cases where a taxon has only one parent are modelled by using familiar sequence evolution models such as the Jukes-Cantor model. However, when a taxon has more than one parent, a different evolutionary model is assumed: each of the parent taxa is given a relative weight, and each character is inherited from a parent that is randomly chosen based on these weights. Inheritance from a parent follows the same model as in the case where there is only one edge pointing to the node in question.

Computing the likelihood of a DAG model, i.e., the probability that a given set of sequences is obtained as the outcome of the given DAG, is hard. Moulton and Strimmer proposed a random sampling technique to approximate the likelihood. Their technique eventually converges to the exact likelihood value but in practice it may take a large number of samples, and hence, a long time, before obtaining accuracy that is sufficient for comparing different DAGs.

We have developed an alternative approximation which is not based on random sampling but instead uses a technique called loopy belief propagation, see (Murphy, Weiss, & Jordan, 1999). It is not guaranteed to converge to the exact value but on the other hand, it is often significantly faster than random sampling. In our experiments (not shown here, see (Nguyen & Roos, in preparation)), it produces better accuracy than a number of different random sampling techniques with less computation time. We also extend the earlier method by Strimmer and Moulton by including a parameter learning step where the edge lengths that characterize the amount of evolutionary change along each edge in the network are learned from the data so that they need not be given as input to the PhyloDAG method.

In practice, the PhyloDAG method takes as input a set of sequences and a tree structure. It then considers all possible additional edges between any two nodes in the tree – including edges between two extant nodes, edges between an extant and an hypothetical node, and edges between two hypothetical nodes – in turn and evaluates the likelihood of the network where the edge in question is included in addition to the edges in the initial tree structure. The edge or the edges that improve the likelihood score the most are included in the output network. Often it is useful to also set an upper bound on the number of edges that are added so as to obtain a more easily interpreted network where only the most significant reticulation events are included. In the present work, we limited the number of additional edges to four to facilitate the interpretation of the models.

We used the Jukes-Cantor model, which can be directly extended to handle any other number of character states than four, for modeling the evolution of individual features and following Moulton and Strimmer, set the weigths on the parents to be uniform so that each parent taxon has the same influence on the dependent taxon.

Appendix C. Parametric bootstrap

Parametric bootstrapping for testing phylogenetic topologies, i.e., tree structures, was first suggested by (Huelsenbeck & Crandall, 1997). Our implementation is primary based on the later description by (Posada, 2003). The testing procedure of topology M_0 (null hypothesis) against topology M_1 (alternative hypothesis) can be briefly described as follows.

1. Estimate the parameters (edge lengths) in models M_1 and M_0 by maximum likelihood. Denote the maximum likelihood estimates (MLEs) by θ_1 and θ_0, respectively.
2. Calculate the log-likelihood ratio (LLR) $l(D|M_1, \theta_1) - l(D|M_0, \theta_0)$, where $l(D|M_1, \theta_1)$ and $l(D|M_0, \theta_0)$ are the log-likelihood of the data given structure M_1 and M_0 with MLE parameters respectively.
3. From structure M_0 with estimated parameters θ_0, draw $K=1000$ simulated data sets which all have the same size and missing data as the original data set.
4. For each simulated data set D_i, estimate parameters $\tilde{\theta}_1$ and $\tilde{\theta}_0$ for both structures, and calculate the LLR $l(D_i|M_1, \tilde{\theta}_1) - l(D_i|M_0, \tilde{\theta}_0)$. Use these to obtain an approximate distribution of the LLR between M_0 and M_1 under the null hypothesis M_0.
5. Let F be the number of time that the LLR on simulated datasets is bigger than the LLR on the original data in Step 2. If the quotient F/K (in this case $K=1000$) is smaller than a predefined threshold (0.05 or 0.01), the null hypothesis is rejected.

The intuition is that if the null hypothesis is true, then the simulated data sets in Step 4 are drawn from the same distribution as the observed data. This implies that the LLR based on the observed data, computed in Step 2, follows the same distribution as the LLR values for the simulated data in Step 4. Suppose now that the LLR for the observed data, which measures how much better model M_1 fits the obsered data than M_0, is higher than almost all of the simulated LLR values. By the above reasoning, this must be unlikely since the observed LLR value is supposed to be drawn from the same distribution as the simulated ones, and we are lead to reject the null hypothesis. It is obvious that such a test is valid in the sense that if the null hypothesis is true, it is unlikely to be rejected.

Appendix D. Additional results.

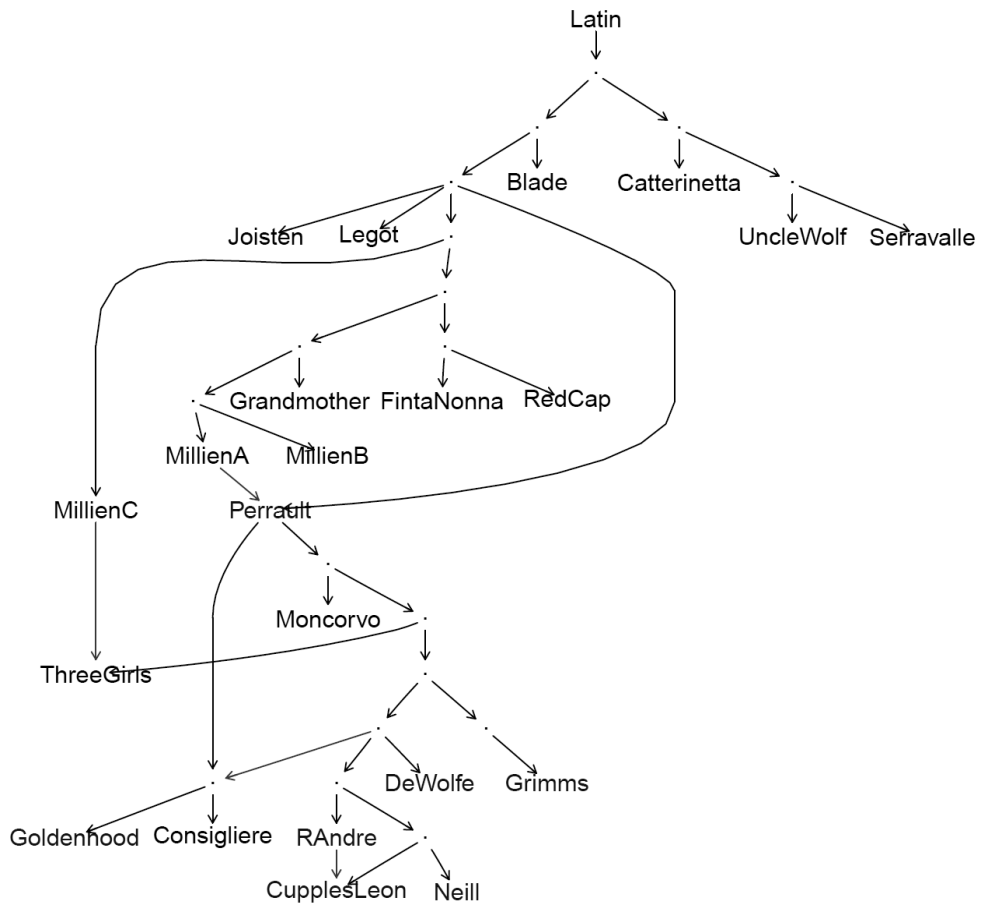Networks c (Fig. 4) and d (Fig. 5) are representative examples among the two main hypotheses: the oral origins hypothesis (network c) and the literary origins hypothesis (network d). Figures 6–14 show the rest of the networks for completeness.
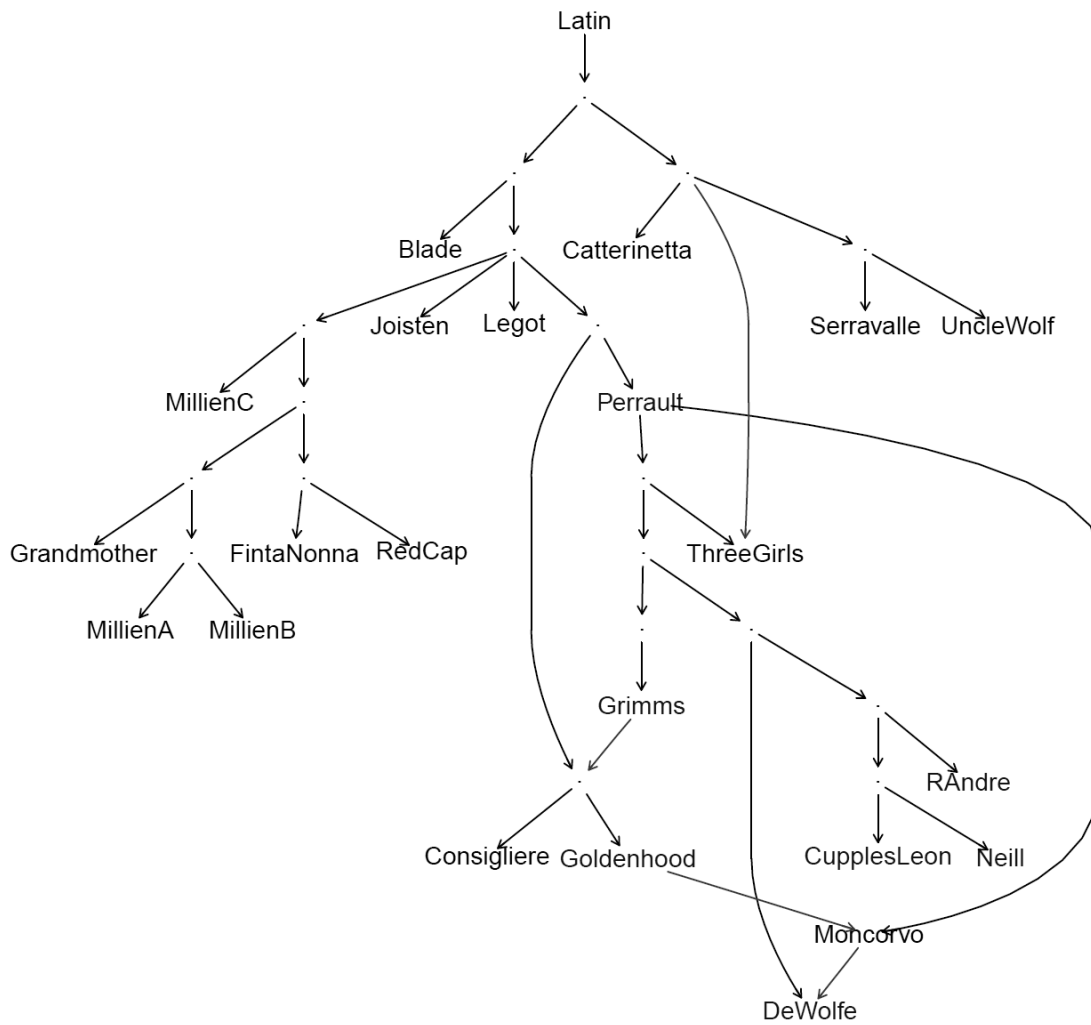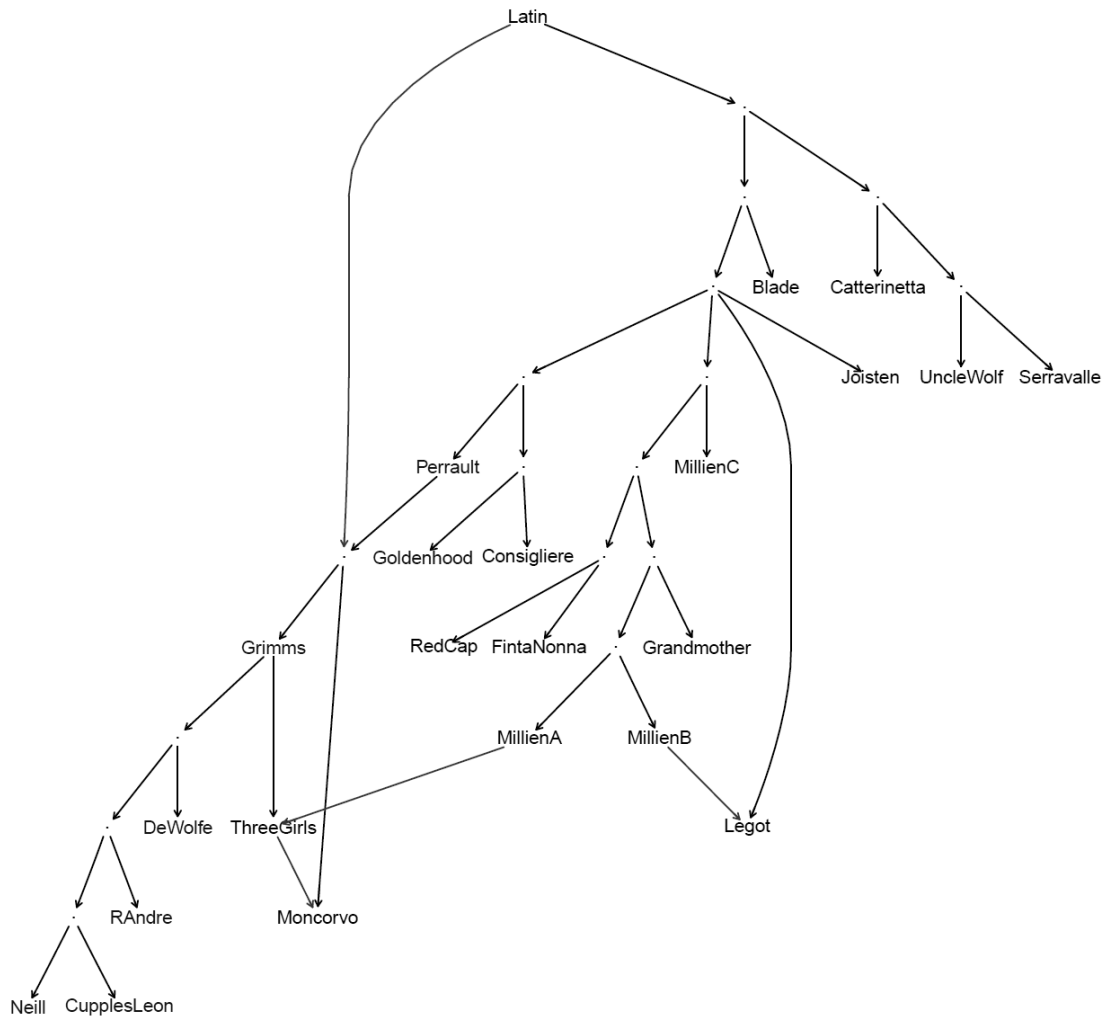
Fig. 6 PhyloDAG network *a*. Log-likelihood –875.6.

Fig. 7 PhyloDAG network *b*. Log-likelihood −862.3.

764
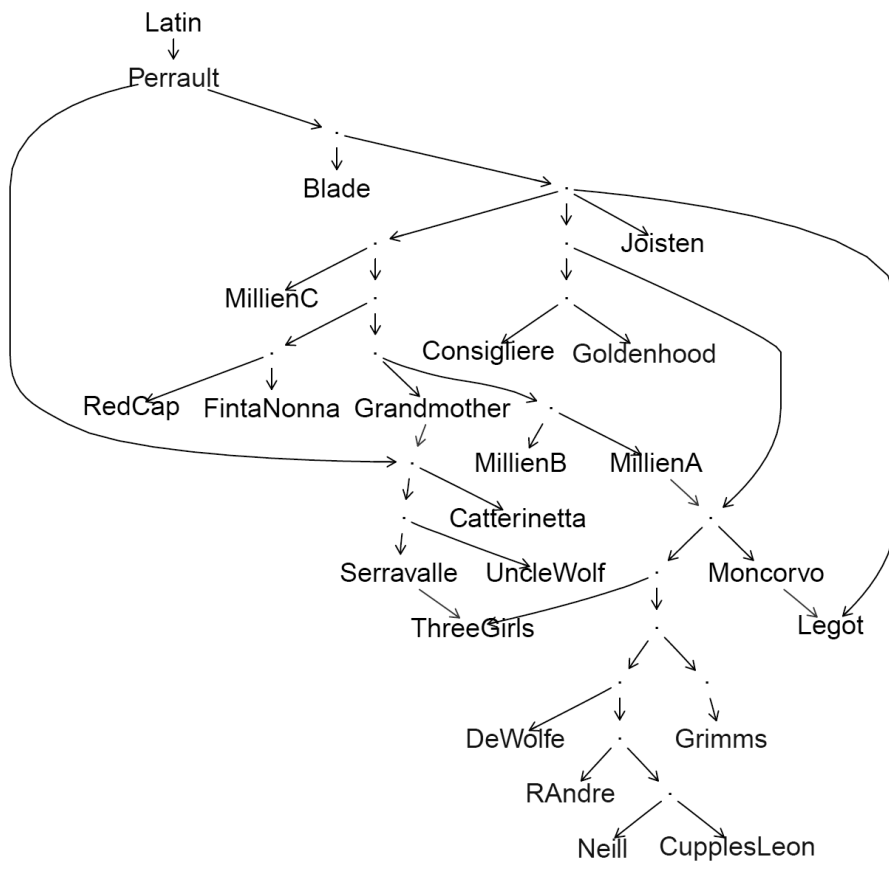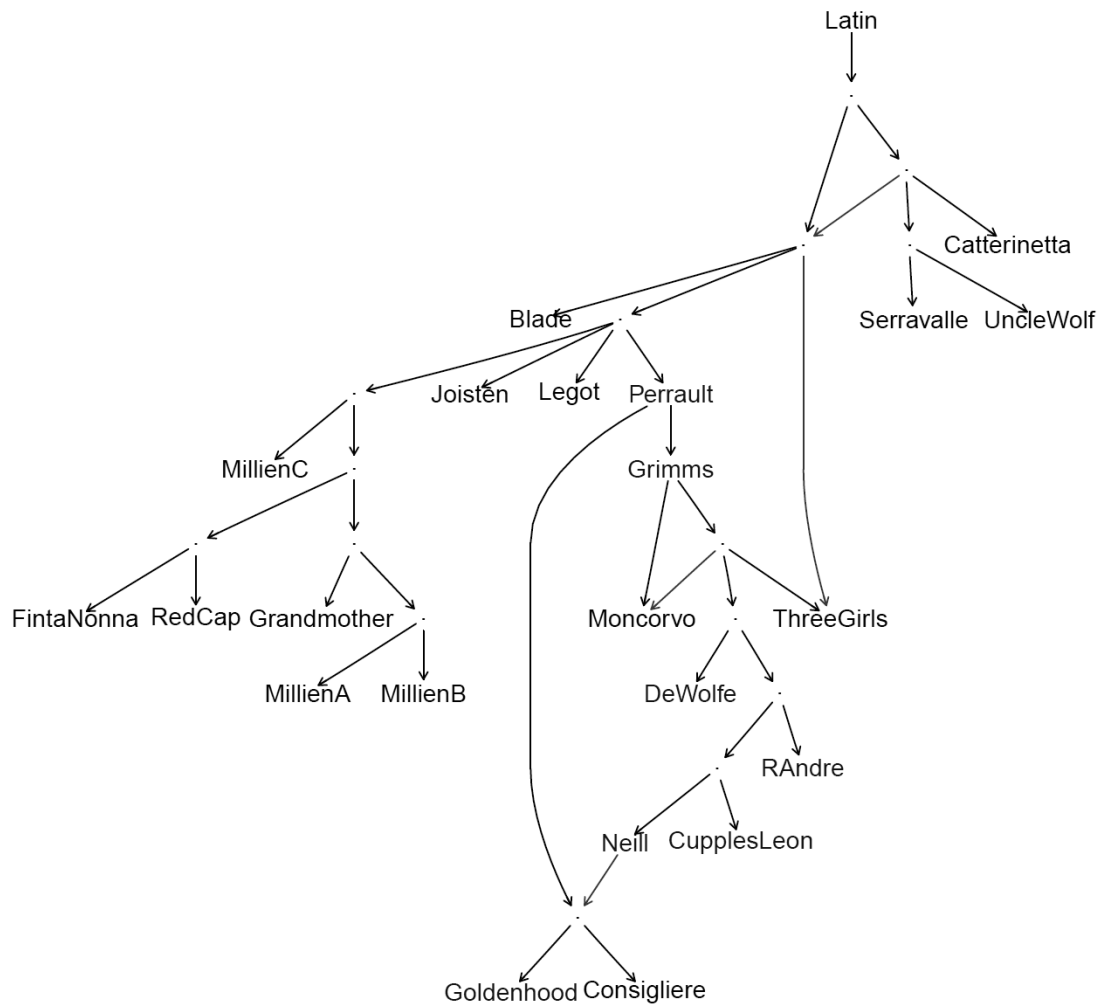765
766

Fig. 8 PhyloDAG network *e*. Log-likelihood −867.0.

767
768
769

Fig. 9 PhyloDAG network *f*. Log-likelihood −884.6.

770
771
772

Fig. 10 PhyloDAG network *g*. Log-likelihood −847.6.
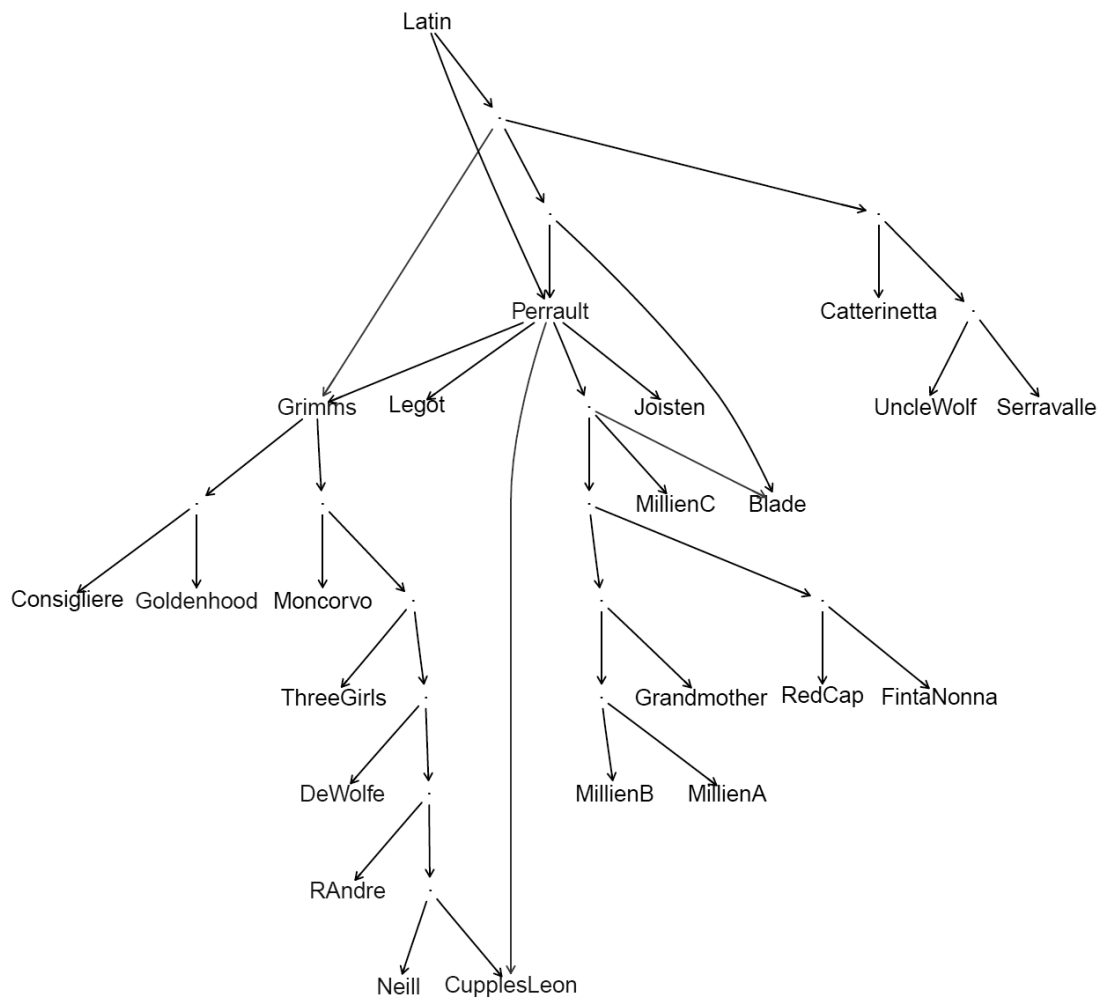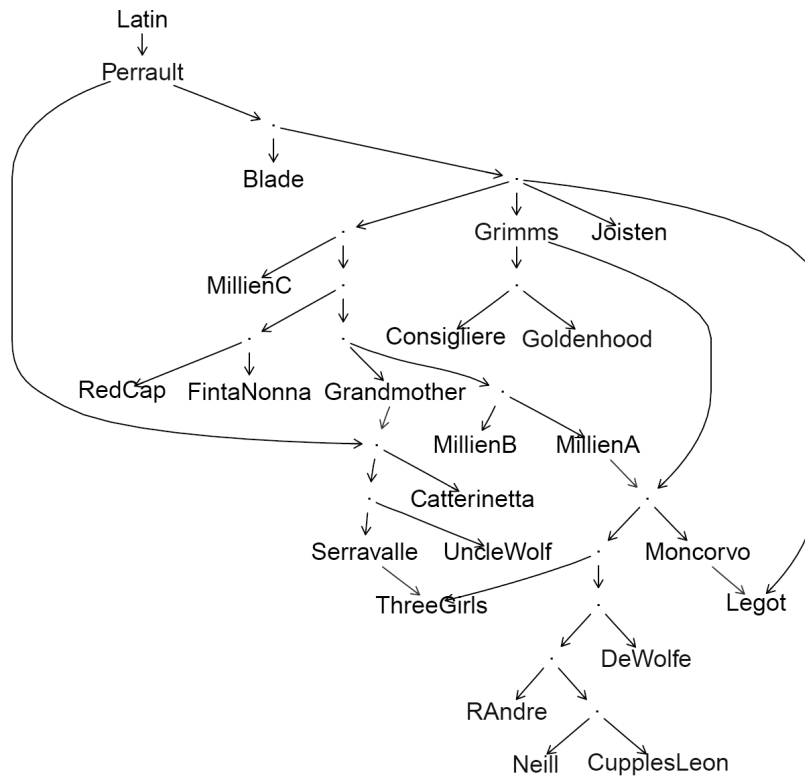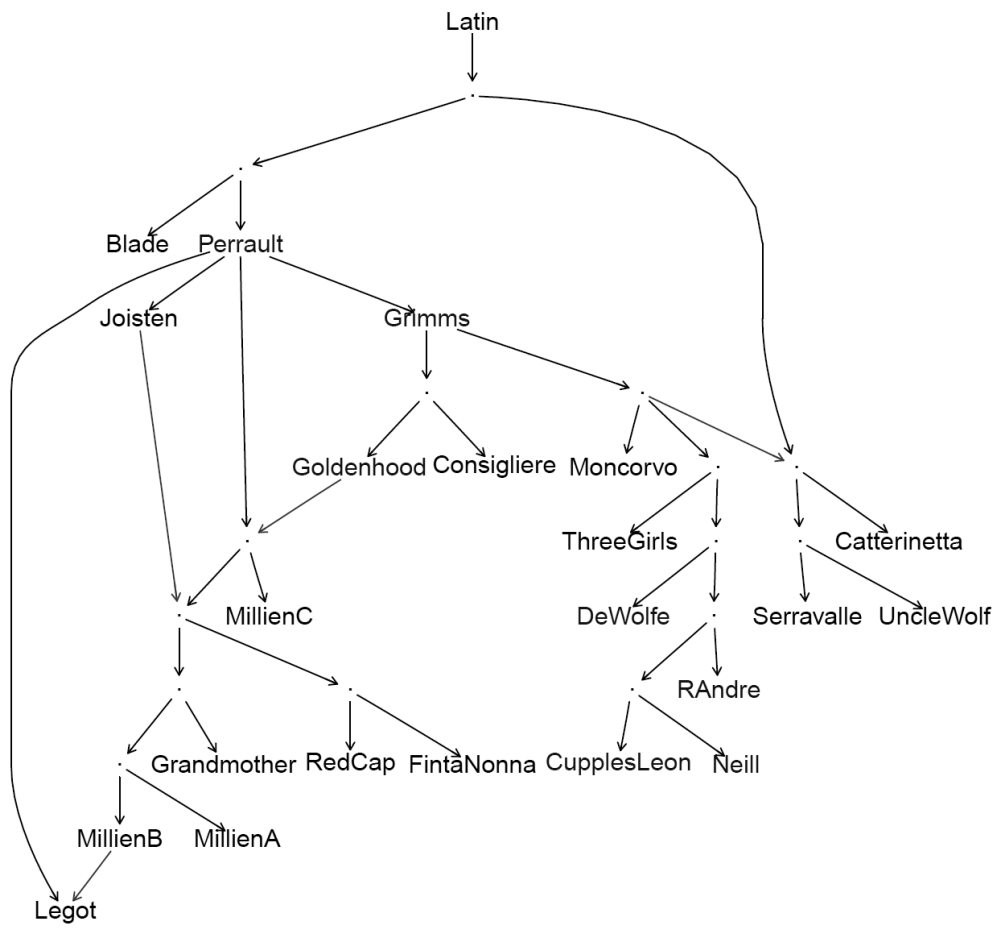
Fig. 11 PhyloDAG network *h*. Log-likelihood −896.76.

Fig. 12 PhyloDAG network *i*. Log-likelihood −897.32.

782
783
784

Fig. 13 PhyloDAG network *j*. Log-likelihood -870.13.

789
790   Fig. 14 PhyloDAG network k. Log-likelihood -870.87.
791
792