Systematic investigations of gene effects on both topologies and supports: an Echinococcus illustration

Christophe Guyeux, Stéphane Chrétien, Nathalie M Côté, Jacques Bahi

Systematic investigations of gene effects on both topologies and supports: an Echinococcus illustration

Christophe Guyeux[b,d,*], Stéphane Chrétien[c,d], Nathalie M.-L. Côté[a], Jacques M. Bahi[b,d]

[a] *Département d'Obstétrique & Gynécologie et Département de Microbiologie et Infectiologie, Faculté de Médecine et des Sciences de la Santé, Université de Sherbrooke et Centre de Recherche du CHUS, Sherbrooke, QC, Canada.*
[b] *FEMTO-ST Institute, UMR 6174 CNRS*
[c] *Laboratory of Mathematics, UMR 6623 CNRS*
[d] *University of Bourgogne Franche-Comté, Besançon, France*

Abstract

In this article we propose a high performance computing toolbox implementing efficient statistical methods for the study of phylogenies. This toolbox, which implements logit models and LASSO-type penalties, gives a way to better understand, measure, and compare the impact of each gene on a global phylogeny. As an application, we study the Echinoccocus phylogeny, which is often considered as a particularly difficult example. Mitochondrial and nuclear genomes (19 coding sequences) of 9 *Echinococcus* species are considered in order to investigate the molecular phylogeny of this genus. First, we check that the 19 gene trees lead to 19 totally different unsupported topologies (a topology is the sister relationship when both branch lengths and supports are ignored in a phylogenetic tree), while using the 19 genes as a whole are not sufficient for estimating the phylogeny. In order to circumvent this issue and understand the impact of the genes, we computed 43,796 trees using combinations ranging from 13 to 19 genes. By doing so, 15 topologies are obtained. Four particular topologies, appearing more robust and frequent, are then selected for more precise investigation. Refining further our statistical analysis, a particularly robust topology is extracted. We also carefully demonstrate the influence of nuclear genes on the likelihood of the phylogeny.

Keywords: Molecular phylogeny, Echinococcus, Statistical tests

*Corresponding author
Email addresses: christophe.guyeux@univ-fcomte.fr (Christophe Guyeux), stephane.chretien@univ-fcomte.fr (Stéphane Chrétien), jacques.bahi@univ-fcomte.fr (Jacques M. Bahi)

1. Introduction

Echinococcus (Cestoda: Taeniidae) species are parasitic organisms responsible for echinococcosis, a serious disease with fluid-filled cycle lesion that affects humans, livestock, and wildlife. Based on morphology and life cycle data, only *E. granulosus* (10 genotypes), *E. multilocularis*, *E. vogeli*, *E. oligarthra*, and *E. shiquicus* have been identified as true *Echinococcus* species [25, 19]. But, despite extensive research, the phylogenetic relationship between species of this genus remains unclear. The large amount of genetics data, recently obtained with complete mitochondrial (mt) genomes, can provide a key to better understand the phylogeny of the *Echinococcus* genus.

Genes from mt genomes are standard markers for phylogeny, as this genome presents interesting features for such types of analysis. Indeed, these genes are present in almost all animals and in only one copy. The organelle is maternally inherited and the genome does not recombine in most cases [5]. Moreover, animal cells contain many mitochondria, which also contain many copies of mt genomes. This feature can be useful when only minute amount of biological material is available for molecular analysis. Finally, coding sequences within this genome are well known: twenty-two tRNAs and 2 rRNA are encoded for mitochondrial translation machinery, while in most cases there are 13 genes coding for proteins of the respiratory chain (in some species, some genes have been lost during evolution, such as *atp8* within the Cestoda and Trematoda for instance [15]).

This is why phylogenetic analyses of animals are frequently performed by using a subset of mitochondrial coding sequences [11], while chloroplasts are often considered for plant phylogenies [4, 3, 2]. For instance, in tetrapods and mammals, the best results have been obtained with *nad5*, *nad4*, *nad2*, *cytb*, and *cox* [27]. The optimal size and the content of this subset of mt genes have however significant consequences: a smaller subset of sequences necessitates a lower cost for DNA extraction and sequencing, while demanding a lower computation time. Conversely, a larger set may be more representative, leading to a more accurate inference. Additionally, it is sometimes reported that, due to their own origin and history, the evolution of mt genomes may be different from the one of nucleus genomes, which may reflect more the species relationship. And various situations have appeared to be hard to solve by using only mitochondrial coding sequences. For instance, in the *Echinococcus* case, no combination of such mt genes has been able to produce a well supported tree, and so enlarging the set of sequences to nuclear genes has been considered as a way to estimate the phylogeny.

The problem in finding a relevant subset of sequences to infer a well supported phylogeny is due to the presence of "blurring" sequences that can increase uncertainties in some locations of the tree. At the origin of such blurring effects, explanations include homoplasy (spurious similarity due to convergence or reversion and not to common ancestry), stochastic errors, undetected paralogy, incomplete lineage sorting, horizontal gene transfers, or even hybridization. Due to such blurring genes, two different subset of sequences may lead to two

Species	Accession	reference
Echinococcus canadensis	NC_011121	[17]
Echinococcus equinus	NC_020374	[20]
Echinococcus felidis	NC_021144	[16]
Echinococcus granulosus	NC_008075	[14]
Echinococcus multilocularis	NC_000928	[18]
Echinococcus oligarthra	NC_009461	[17]
Echinococcus ortleppi	NC_011122	[17]
Echinococcus shiquicus	NC_009460	[17]
Echinococcus vogeli	NC_009462	[17]
Vesteria mustelae	NC_021143	[16]

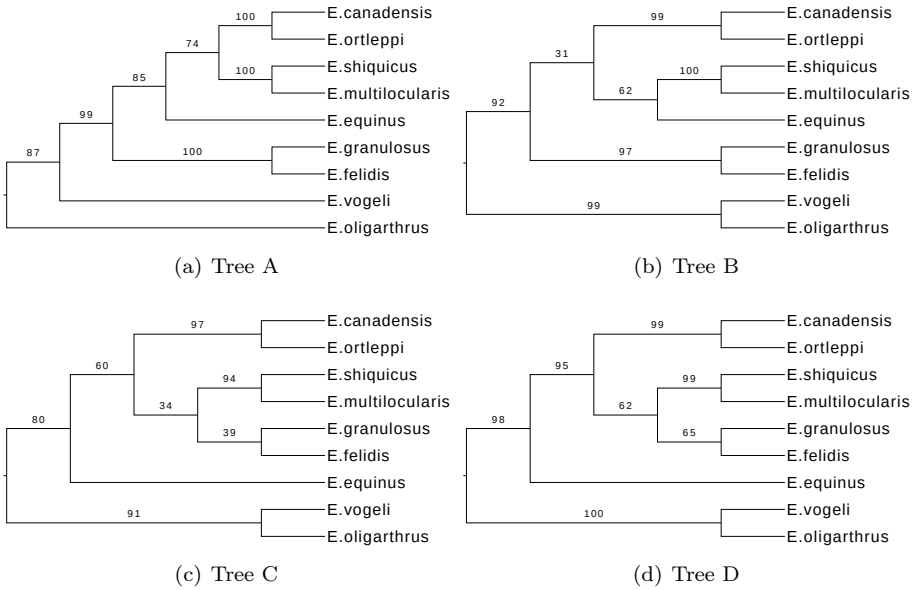
Table 1: Echinococcus species analyzed in this paper and the accession numbers of mitochondrial genomes. *V. mustelae* is an outgroup.

totally different phylogenetic trees. But, if we except the use of gene trees, there is a lack of tools that allow to reinforce the confidence put in a biomolecular phylogeny by understanding the impact of gene selection on a given phylogeny.

In this article, our aim is to present a way to obtain robust taxonomies for phylogenetic problems which are in particular hard to solve, by using a statistical approach (logit regression with potentially dummy variables and LASSO tests) and extensive computations on available genetic data. Our proposed methodology is able to detect blurring genes and to understand their effects on topologies and on bootstrap supports. This method is presented and illustrated by addressing a case that is well known to be difficult, namely the *Echinococcus* genus. The methodology encompasses a *de novo* annotation of each available mt genome that is completed with nuclear genes, the multi-alignment of each coding sequence, and the systematic investigations of trees that can be inferred using all possible large subsets of sequences. The computation of $43,796$ trees, corresponding to all the subsets from 13 to 19 genes, leads to the possibility to understand, by using advanced statistical techniques, the effects of each gene on both topologies and supports, and thus to choose the best phylogenetic representation of the considered set of species.

State-of-the-art

To date, 9 complete mitochondrial genomes of *Echinococcus* have been published, they are listed with their accession numbers in Table 1. These genomes or other nuclear genes have been used recently to update the molecular phylogeny of this genus. Indeed, 4 different phylogenetic trees obtained with biomolecular analyses can be found in the literature (Figure 1). Tree A has been inferred on the full mitogenome data set [17] while the three other *Echinococcus* trees based on nuclear genes have been proposed in [12]. Trees B and C have been inferred using maximum likelihood analysis (PAUP 4.0 software [24]) of *rpb2* (RNA polymerase II second largest subunit), *pepck* (phosphoenolpyruvate carboxykinase), and *pold* (DNA polymerase delta) genes (gDNA and exon data sets respectively, with 5008 and 4726 base pairs resp.). Finally, Tree D is the result of a Bayesian



(a) Tree A

(b) Tree B

(c) Tree C

(d) Tree D

Figure 1: Phylogenetic trees found in the literature (outgroups not represented): A. Phylogeny obtained by a maximum likelihood method on complete mitochondrial genomes [16]. B. Phylogram obtained by maximum likelihood method with nuclear genes (*rbp2*, *peck*, and *pold*) using gDNA. C. Phylogram obtained by maximum likelihood method with nuclear genes (*rbp2*, *peck*, and *pold*) using exons sequences. D. Phylogram obtained using Bayesian inference with nuclear genes (*rbp2*, *peck*, and *pold*) using sequence of exons. Note that C and D trees share the same topology.

inference using BEAST software [7] based on the exon data set detailed above. *V.mustelae* has been used as the outgroup for the tree B, C, and D.

A general topology with low bootstrap values of the *Echinococcus* phylogeny can be deduced from these trees: *E. vogeli* and *E. oligarthra* appear as sister species, and this is also the case for *E. canadensis* and *E. ortleppi*; and for *E. Granulosus* and *E. felidis*. Additionally, all these species plus *E. equinus* seem to constitute a clade sister to both *E. vogeli* and *E. oligarthra*. But obviously, these four trees are not well supported, and the following questions remain unsolved: what is the correct position of *E. equinus* ? Are *E. shiquicus* and *E. multilocularis* sister species of *E. canadensis* and *E. ortleppi*, or of *E. granulosus* and *E. felidis* ? And what is the real relation between *E. vogeli* and *E. oligarthra* ? Only a well-supported tree can provide clear responses.

2. Materials and methods

2.1. Alignment and annotations of coding sequences

To answer these questions, we have performed first Bayesian and maximum likelihood analyses on (1) the whole mitogenome, (2) its 12 protein-coding genes

Table 2: Gene tree topologies

gene	**topology (newick)**
atp6	(vogeli,((felidis,granulosus),(equinus,(multilocularis,((canadensis,ortleppi),(shiquicus,oligarthra)))));
cob	(multilocularis,((equinus,((felidis,granulosus),(oligarthra,vogeli))),(shiquicus,(canadensis,ortleppi))));
cox1	(oligarthra,(vogeli,((((granulosus,felidis),(canadensis,ortleppi)),(shiquicus,multilocularis)),equinus)));
cox2	(((((vogeli,(canadensis,ortleppi)),(shiquicus,(multilocularis,oligarthra))),equinus),felidis),granulosus);
cox3	(vogeli,(((granulosus,(oligarthra,felidis)),(canadensis,ortleppi)),((shiquicus,multilocularis),equinus)));
ef1a	(felidis,((canadensis,ortleppi),(((multilocularis,shiquicus),((oligarthra,vogeli),granulosus)),equinus)));
elp	(oligarthra,(((equinus,(shiquicus,multilocularis)),((canadensis,ortleppi),(felidis,granulosus))),vogeli));
nad1	(vogeli,(((granulosus,felidis),((equinus,(multilocularis,shiquicus)),(canadensis,ortleppi))),oligarthra));
nad2	(felidis,(granulosus,((((equinus,vogeli),(canadensis,ortleppi)),(multilocularis,shiquicus)),oligarthra)));
nad3	(oligarthra,(vogeli,(multilocularis,(((felidis,granulosus),(shiquicus,equinus)),(canadensis,ortleppi)))));
nad4	(vogeli,((((granulosus,felidis),(multilocularis,(shiquicus,(canadensis,ortleppi))))),equinus),oligarthra));
nad4l	(shiquicus,(((oligarthra,vogeli),multilocularis),((canadensis,ortleppi),(equinus,(felidis,granulosus)))));
nad5	(oligarthra,(vogeli,((((felidis,granulosus),(canadensis,ortleppi)),equinus),(shiquicus,multilocularis))));
nad6	(((equinus,(((felidis,(granulosus,oligarthra)),vogeli),(shiquicus,multilocularis))),ortleppi),canadensis);
pepck	((oligarthra,vogeli),((equinus,(shiquicus,multilocularis)),(granulosus,((canadensis,ortleppi),felidis))));
pold	(oligarthra,(((felidis,granulosus),(((shiquicus,multilocularis),(canadensis,ortleppi)),equinus)),vogeli));
rpb2	((((oligarthra,vogeli),(equinus,((multilocularis,shiquicus),(felidis,granulosus)))),canadensis),ortleppi);
rrnL	(oligarthra,((((canadensis,ortleppi),((shiquicus,multilocularis),equinus)),vogeli),(felidis,granulosus)));
rrnS	(multilocularis,(oligarthra,((((canadensis,ortleppi),(felidis,granulosus)),(vogeli,equinus)),shiquicus)));

merged after alignments, and (3) on all 19 available genes taken alone (leading to 19 gene trees). Theses analyses were realized using nucleotides and translated amino acid sequences. Tools used during these first analyses were:

- Muscle [8] for complete mitogenome alignments and T-Coffee [21] for gene alignments;
- NCBI annotations for the coding sequences in a first step, and then DOGMA [26] for further analyses;
- PhyloBayes [13] for Bayesian inference, PhyML [9] and RAxML [23] for maximum likelihood.

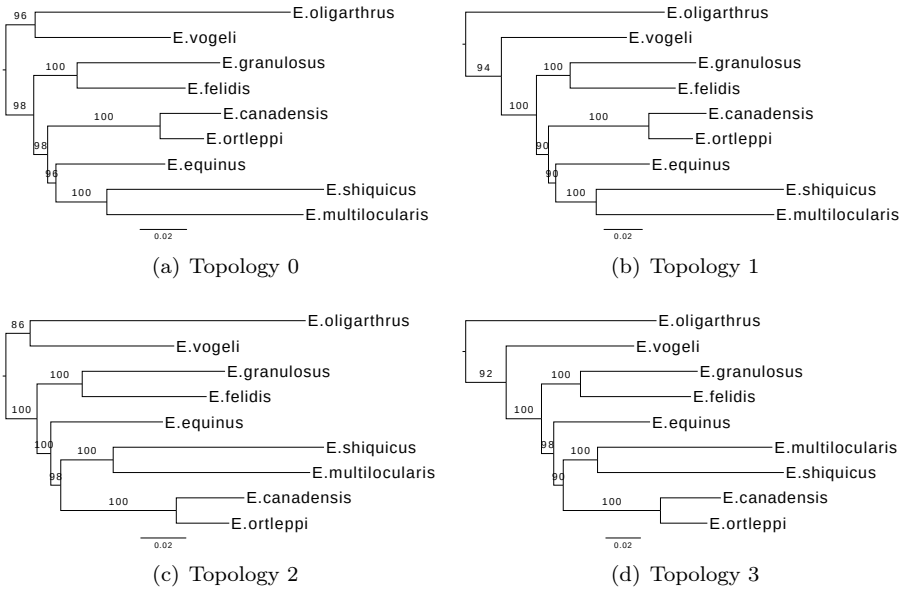
Contrary to the intuition, for every experiment, a problem with the support (i.e. support was less than 95% recovery) was found in at least one branch in each trees obtained. Obtained gene trees illustrate well the statistical hardness of this phylogeny: in Table 2, we can surprisingly see that 19 different topologies have been obtained when considering each coding sequence taken alone. Shared structures can be found between two given gene trees, but it is hard to deduce a consensus topology from their phylogenies, due to the large variability we obtained. Furthermore, each gene tree contains numerous problematic supports, as can be seen in the depicted trees of the supplementary material (see the appendix).

Based on these preliminary results, we are able to conclude that: (1) using coding sequences is better than using the whole mitogenome, (2) there are numerous inconsistencies in NCBI annotations (cf. the appendix), (3) the alignments obtained using T-Coffee lead to better supported trees than the ones using Muscle, (4) to separate the natural history of the genus from the individual history of each gene seems very difficult, when considering the large variability of gene trees, and (5) to increase the amount of genetic data leads to better supported trees. Having these results in mind, we our proposal consists in a general methodology for estimating a phylogeny by measuring gene effects. Application to the *Echinococcus* case will demonstrate the potential of our approach.

2.2. Methodological approach

To estimate the phylogeny of *Echinococcus* and to determine which genes do not support this phylogeny, a simple solution consists in considering all the available coding sequences (mt and nucleus ones) shared by these 10 species, and to investigate how the phylogeny is affected when using subsets of these sequences. Doing so will extend the first investigations of Hardman *et al.* [11], who studied the phylogeny of 5 *Taeniidae* based on the 12 individual mitochondrial coding genes.

Nineteen sequences have been extracted from each of the considered species: 5 nuclear genes, and 12 protein-coding sequences and 2 rDNA genes from the mitochondrial genomes. They are listed below.



(a) Topology 0

(b) Topology 1

(c) Topology 2

(d) Topology 3

Figure 2: The 4 trees with more than 300 occurrences, when considering 43,796 trees obtained with Algorithm 1 (*V. mustelae* is an outgroup)

- **Nuclear genes:** ezrin-radixin-moesinlike protein (*elp*), elongation factor 1 alpha (*ef1a*), RNA polymerase II second largest subunit (*rpb2*), phosphoenolpyruvate carboxykinase (*pepck*), DNA polymerase delta (*pold*).
- **Mitochondrial protein coding sequences:** ATP synthase 6 (*atp6*), cytochrome b (*cob*), cytochrome c oxydase 1 (*cox1*), *cox2*, *cox3*, NADH dehydrogenase subunit 1 (*nad1*), *nad2*, *nad3*, *nad4*, *nad4l*, *nad5*, *nad6* .
- **Mitochondrial rDNAs:** the large subunit rDNA (*rrnL*), and the small subunit rDNA (*rrnS*).

DOGMA has been used to annotate from scratch each up-to-date complete mitochondrial genome downloaded from NCBI [6] (accession numbers provided in Table 1). Default parameters of DOGMA have been selected, with identity cutoffs for amino acids equal to 60% and 80% for coding genes and rDNA genes. These thresholds have been reduced to 55% and 75% respectively for *V.mustelae*, due to a problem for the detection of *nad6* and *rrnL*. The e-value was equal to $1e-5$, and the number of blast hits to return has been set to 5, see Table 8 of the appendix section for a comparison with annotated genomes from NCBI.

Each of these 19 coding sequences has been aligned separately by using T-Coffee (M-Coffee mode, using 6 cores for multiprocessing). Then 43,796 trees have been constructed, corresponding to all the possible combinations of 13, 14, 15, ..., and 19 coding sequences among the 19 available ones ($\sum_{k=13}^{19} \binom{19}{k} = 43{,}796$), as described in Algorithm 1. The aim was to determine the most ro-

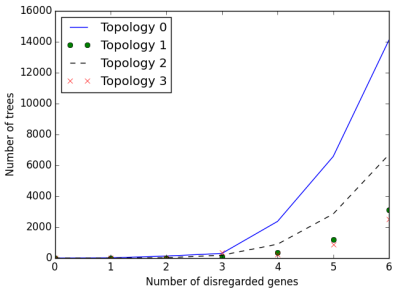
Topology	Lowest bootstrap	Number of occurrences	Average bootstrap	Discarded genes
0	96	23514	58	*cox1, elp, nad4, nad5, nad6, rrnS*
1	90	4803	54	*cob, ef1a, nad4, nad6, rpb2, rrnS*
2	86	10668	52	*cox3, nad2, nad3, pepck, rrnS*
3	90	4019	55	*cox3, ef1a, nad2, nad3, pepck, rpb2*
4	52	17	41	*cox1, nad1, nad3*
5	54	55	40	*cob, nad1, nad5, pepck, pold, rpb2*
6	59	113	41	*atp6, cox3, nad4, nad5, pepck, rrnS*
7	60	276	37	*cox1, elp, nad2, nad3, pepck, pold*
8	67	148	46	*cob, cox1, nad5, nad6, pold, rpb2*
9	55	123	41	*atp6, cob, nad2, nad6, pold, rrnL*
10	45	28	36	*cob, nad1, nad5, nad6, pold, rrnL*
11	53	25	42	*atp6, cox1, cox2, elp, nad3, pold*
12	44	3	39	*nad4, nad5, pepck, rpb2, rrnL, rrnS*
13	31	2	31	*ef1a, elp,nad1, nad4, pepck, pold*
14	34	2	28	*atp6, cox2, nad1, nad4l, pold, rrnL*

Table 3: Detail of obtained topologies. Lowest bootstraps, number of occurrences, and not included genes are provided.

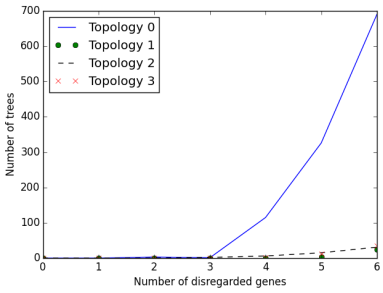
bust phylogenetic trees and the effects of each gene on topologies and supports. RAxML version 8.0.20 was used for maximum likelihood inference, with 3 distinct models/data partitions with joint branch length optimization at each computation, corresponding to the nuclear genes, the mitochondrial rDNA genes, and the mitochondrial protein-coding sequences [10]. All free model parameters have been estimated by RAxML for the GAMMA model of the rate heterogeneity and the ML estimate of the alpha-parameter. At each time, a maximum of 1000 non-parametric bootstrap inferences were executed, with MRE-based boots stopping criterion, and *V.mustelae* has been used as outgroup.

for *each gene g* **do**
  multi-align all the 10 sequences of g using T-coffee;
end
for *k=13,...,19* **do**
  for *each combination c of k genes* **do**
    concatenate the k alignments following the alphabetic order;
    find the best phylogenetic tree with RAxML;
    compute bootstraps with RAxML and MRE option;
    map the bootstraps on the best tree (RAxML);
    extract the list of bootstraps L(c) and the topology T(c);
    store (c, L(c), T(c));
  end
end

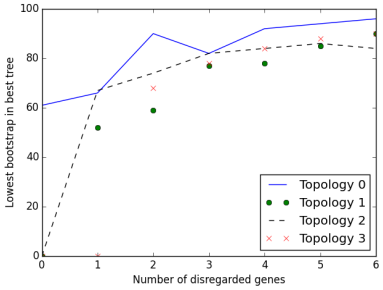
Algorithm 1: Pseudocode producing 43,796 phylogenetic trees



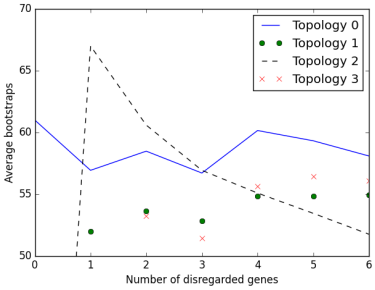
(a) Number of trees per topology.



(b) Number of trees whose lowest bootstrap is larger than 80.



(c) Lowest bootstrap in best trees.



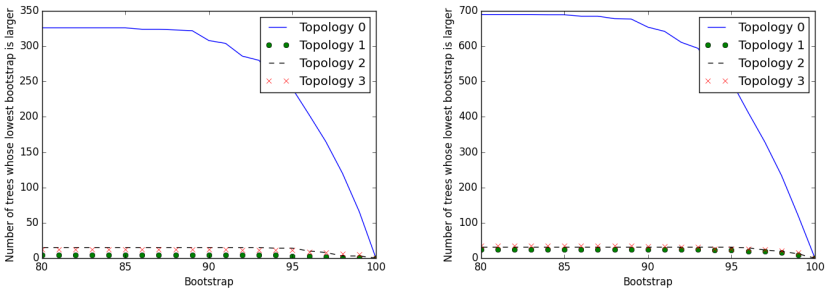
(d) Average value of lowest bootstraps

Figure 3: Comparison of the 4 best topologies, according to the number of discarded genes. A. The number of trees in each topology, according to the number of discarded genes. B. As in (A), but considering only trees whose supports are larger than 80. C. Minimal support in the best tree of each topology regarding the number of discarded genes. D. Averages of all minimal bootstraps in each tree of each topology.

3. Discussion and results

3.1. Results

Only 15 topologies have been obtained during our computations, among the 34,459,425 possible phylogenetic tree topologies with 9 species and 1 outgroup. Only 4 of these 15 topologies have a number of occurrences larger than 300, when considering the 43,796 obtained trees. These trees are depicted in Figure 2, while further information is provided in Table 3: for each topology, the lowest bootstrap of the best tree (that is, the lowest bootstrap of the tree that maximizes the minimum taken over all its bootstraps), the number of trees having this topology, the average minimal bootstrap value, and the list of genes that have been removed to obtain the best tree having this topology. We can remark that Topology 0 corresponds to Tree B of the literature (Figure 1), with more robust bootstraps. Topology 2 is similar to Tree A, while the Trees C and D have not been obtained during our computations.



(a) 5 discarded genes (b) 6 discarded genes

Figure 4: Comparison of the best topologies, according to the number of supported trees. A. The ordinate provides the number of trees in the topology whose lowest bootstrap is larger than the associated abscissa, for 5 removed genes. B. As in (A), but for 6 discarded genes.

The 4 best topologies represent 98.19% of the whole obtained trees, and they share most of their structure. And, at this stage, a first question raised in Section 1 can be answered: *E. shiquicus*, *E. multilocularis*, *E. canadensis*, and *E. ortleppi* are within the same clade, which is sister to the clade consisting of *E. granulosus* and *E. felidis*. For the correct position of *E. equinus*, Topologies 0 and 1 are opposed to Topologies 2 and 3, while Topologies 0 and 2 are in contradiction with the two other topologies for the relation between *E. vogeli* and *E. oligarthra*.

Several reasons lead us to consider the Topology 0 depicted in Figure 2(a) as the most probable one. Firstly, this is the most frequent topology, representing 53.69% of the trees, while the second one (Topology 2) represents only 24.40%. This topology remains the most frequent even when 1, 2, 4, 5, or 6 genes were removed. Surprisingly, Topology 3 is a bit more frequent than Topology 0 when 3 genes are discarded, see Figure 3(a). Secondly, this topology contains the most robust trees: the smallest bootstraps in Topology 0 are largest compared to the bootstraps of the other topologies. This property is preserved when 1 to 6 genes are removed from the dataset, see Figure 3(c). In particular, the best tree for Topology 0, depicted in Figure 2(a), outperforms the best tree of the other topologies, even when considering the smallest bootstrap or its average. Thirdly, the number of trees with no bootstrap under 80 is equal to 1,135 for Topology 0, and is equal to 29, 54, and 49 for Topologies 1, 2, and 3 respectively. This is illustrated in Figure 3(b) and this property is further investigated in Figure 4, which shows the number of trees with no bootstrap under a given threshold for each topology.

Let us remark now that we analyzed 14 mitochondrial, but only 5 nuclear genes: available data being what they are, our approach may preferably support trees based on mitochondrial genes. Additionally, the possibility of a systematic bias has been reported in General Time Reversible (GTR) maximum likelihood analyses of mitochondrial data. Trees based on specific nuclear genes can be

different from the overall species-level phylogenetic tree, which can be due to incomplete lineage sorting. Thus, we have to deal with potential artifacts due to the fact that mitochondrial trees are more represented than nuclear ones. In the next sections, the relationship between genes and tree topologies/supports has been investigated in detail, in order to find genes that behave differently from the other ones (for biological or methodological reasons), and to understand their differences regarding the phylogeny.

3.2. The frequency of occurrences of the genes

Using a simple Python script, the evenness of each gene in the 4 most frequent topologies has been evaluated. Since all the 43,796 trees have been constructed using a subset of the 19 available sequences, it is relevant to consider the frequency of occurrence of each of these sequences. Table 4 summarizes obtained results.

Topologies	0		1		2		3	
	number	rank	number	rank	number	rank	number	rank
atp6	16805	7	3684	9	7586	14	2932	13
cob	17074	5	4038	4	7467	15	2881	15
cox1	**20409**	**2**	2884	**18**	7601	13	1459	**19**
cox2	16926	6	3568	11	7609	12	2890	14
cox3	15310	16	3390	14	9789	**2**	3454	4
ef1a	*15117*	*17*	***4635***	***2***	*7765*	*8*	*3594*	*3*
elp	*16097*	*13*	3403	*13*	*8695*	*5*	*3006*	*10*
nad1	15622	15	3377	15	8886	4	3442	5
nad2	16500	11	3756	8	7742	9	3347	6
nad3	16758	8	3551	12	8654	6	2948	12
nad4	17222	3	3795	5	**6848**	**18**	2813	17
nad4L	16676	9	3760	7	7694	11	3054	9
nad5	**22335**	**1**	3203	16	7133	16	2776	**18**
nad6	17160	4	4299	3	6897	17	2844	16
pepck	***14657***	***19***	***2832***	***19***	***21336***	***1***	***3612***	***2***
pold	*16322*	*12*	*3675*	*10*	*7698*	*10*	*2959*	*11*
rpb2	***15052***	***18***	***4803***	***1***	***6795***	***19***	***12057***	***1***
rrnL	15890	14	3033	17	8922	3	3133	8
rrnS	16590	10	3787	6	7849	7	3210	7

Table 4: Number of times each gene were present to produce a tree having one of the 4 most relevant topologies.

Some genes, either over or under represented, seem to be associated to particular topologies. More precisely, the following points may be outlined.

1. The 3 least frequent genes of Topology 0 are 3 nuclear genes, while all mitochondrial genes are well ranked in that topology. Conversely, several mitochondrial genes are wrongly placed on the other topologies, while the first positions in Topologies 1, 2, and 3 are taken by nuclear genes. Based on mitochondrial data, Topology 0 seems to provide a better evaluation of the evolution, while *pepck*, *rpb2*, and *ef1a* modify this topology, leading to Topologies 1, 2, and 3.
2. The nuclear gene *pepck* moves *E. equinus* in the tree, as it is ranked 19 for Topologies 0 and 1, while it is respectively ranked 1 and 2 for Topologies 2 and 3.

	coef	std err	z	$P > \|z\|$	[95.0% Conf. Int.]
atp6	0.7162	0.036	20.067	0.000	[0.646, 0.786]
cob	1.3351	0.037	35.717	0.000	[1.262, 1.408]
cox1	2.0929	0.041	51.421	0.000	[2.013, 2.173]
cox2	0.9855	0.036	27.138	0.000	[0.914, 1.057]
cox3	-0.6574	0.035	-18.599	0.000	[-0.727, -0.588]
ef1a	-0.7374	0.035	-20.785	0.000	[-0.807, -0.668]
elp	-0.1997	0.035	-5.719	0.000	[-0.268, -0.131]
nad1	-0.5465	0.035	-15.529	0.000	[-0.615, -0.478]
nad2	-0.0333	0.035	-0.953	0.341	[-0.102, 0.035]
nad3	0.4243	0.035	12.050	0.000	[0.355, 0.493]
nad4	2.0052	0.040	49.812	0.000	[1.926, 2.084]
nad4L	0.1940	0.035	5.546	0.000	[0.125, 0.263]
nad5	2.6243	0.044	60.270	0.000	[2.539, 2.710]
nad6	1.7467	0.039	44.758	0.000	[1.670, 1.823]
pepck	-5.8472	0.060	-96.768	0.000	[-5.966, -5.729]
pold	-0.1390	0.035	-3.984	0.000	[-0.207, -0.071]
rpb2	-3.9145	0.046	-85.856	0.000	[-4.004, -3.825]
rrnL	-0.3557	0.035	-10.162	0.000	[-0.424, -0.287]
rrnS	0.0778	0.035	2.227	0.026	[0.009, 0.146]

Table 5: Dummy logit regression results for Topology 0

3. Similarly, *rpb2* changes the relationship between *E. vogeli* and *E. oligarthra*.

However, these claims need to be further investigated by a more rigorous statistical approach, which is the aim of the next sections.

3.3. Influence of genes on topology using Dummy logit model

To investigate more soundly the effects of each coding sequence on the species topology, 4 dummy binary choice logit models have been set up for each best topology using `scikit-learn` [22] module of Python language. The reference to the exogenous design is a $19 \times 43,796$ array, each row being a vector of 0's and 1's: a 0 in position i of row k means that, in the k-th tree computation, gene number i (in alphabetic order) were discarded, and conversely the gene was conserved if the coefficient is 1. The rows are thus the "observations" while the columns correspond to regressors. The 1-d endogenous response variable is a vector of size $43,796$, with a 1 in position k if and only if the Topology 0 has been produced with the choice of genes corresponding to the row number k in the exogenous design (resp. Topology 1, 2 or 3 in the three other binary choice logit models). The model has then been fitted using the maximum likelihood approach and a Newton-Raphson solver for performing the maximization. Convergence was obtained after 8 iterations, and the results of the Logit regression are summarized in Table 5.

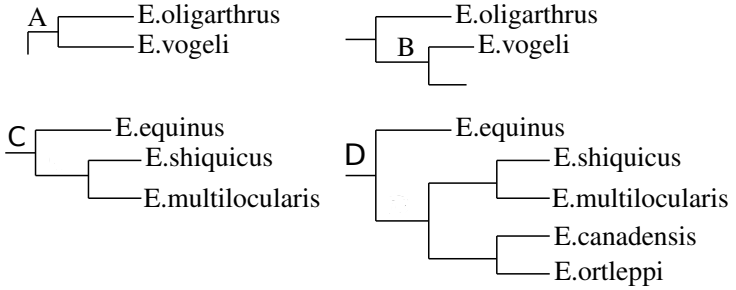


Figure 5: Location of focused bootstraps

The first conclusion that can be drawn from the investigation of the impact of each gene is that, except for the particular case of *nad2* and *rrnS* (see column $P > |z|$), all the coding sequences bring phylogenetic information. Additionally, all coding sequences reinforcing the choice of Topology 0 as the best topology are mitochondrial genes (see the largest values in *coef* column). Conversely, nuclear genes tend to participate in the rejection of this tree (lowest values). This trend is minimal for the genes *ef1a*, *elp*, and *pold*, but very important for genes *pepck* and *rpb2*. These last 2 genes have been used in [12] and our results could explain the limited robustness of the Topology 0 in Tree B, and how the position of *E. equinus* has changed, leading to the Tree C.

Another conclusion is that the negative effect of mitochondrial coding sequences on Topology 0 has a low impact. *Cob*, *cox1*, *nad4*, *nad5*, and *nad6* have a strong impact on this topology. All these findings are consistent with the frequency of occurrences of each gene in the choice of Topology 0: *nad5* has been included in 22,335 computations leading to this topology (94.99%), while only $14,657$ computations with *pepck* has led to this topology (62.33%), as described in Table 4.

Further investigations of the role of each sequence show that *rpb2* and *ef1a* are responsible for the inaccurate phylogenetic position of *E. vogeli* in the Topologies 1 and 3 (this species is considered as the sister species of *E. oligarthra* in these topologies). Their logit regression coefficients are among the lowest ones in Topologies 0 and 2, and among the largest ones in the two other topologies (see the Tables 9, 10, and 11 of supplementary data, which contain the results of the dummy logit regression test for Topologies 1, 2, and 3). Similarly, the inaccurate phylogenetic position of *E. equinus* in the Topologies 2 and 3 is mainly due to the gene *pepck*, whose coefficient is the lowest one in the Topologies 0 and 1, and almost the largest one in the Topologies 2 and 3. These points contribute to explain the results obtained by Knapp *et al.* in [12].

3.4. Regressing the bootstrap on the genes using the LASSO method

To better understand the effects of each gene on each topology, the 4 bootstrap values depicted in Figure 5 have been investigated with more detail . Bootstrap A corresponds to the sister relationship of *E. oligarthra* and *E. vogeli*

	Bootstraps							
	A		B		C		D	
	without	with	without	with	without	with	without	with
atp6	73.29	73.29	62.07	60.19	81.92	81.61	76.48	74.91
cob	72.14	75.56	61.22	62.19	81.42	83.64	75.90	77.27
cox1	72.05	75.46	61.67	60.90	81.35	83.09	76.84	70.18
cox2	72.87	74.18	61.83	60.77	81.26	83.39	76.20	75.87
cox3	73.75	72.03	62.31	58.96	81.34	82.70	76.20	75.97
ef1a	73.31	73.23	57.43	67.35	82.25	80.83	73.09	82.91
elp	73.57	72.61	61.92	60.36	81.74	82.03	76.27	75.75
nad1	73.75	72.06	62.38	58.61	82.17	81.24	76.87	74.37
nad2	73.03	73.93	61.30	62.25	81.99	81.47	76.13	76.14
nad3	73.11	73.71	61.94	60.38	81.46	82.73	76.45	75.15
nad4	71.86	75.88	60.22	64.29	81.69	82.59	75.96	77.16
nad4l	73.24	73.41	61.71	61.27	81.78	81.99	76.02	76.44
nad5	72.51	74.55	61.73	61.06	81.25	84.29	76.67	71.95
nad6	71.82	76.01	60.21	63.55	81.84	81.84	75.92	77.41
pepck	73.53	68.85	61.78	56.12	78.32	82.83	72.83	78.03
pold	74.29	70.80	62.82	57.93	85.04	73.49	78.24	69.81
rpb2	75.98	55.93	50.06	61.68	83.07	68.26	66.75	77.18
rrnl	74.24	70.84	62.72	55.48	83.21	79.36	78.22	70.53
rrns	73.03	73.93	61.35	62.12	80.53	85.09	74.77	79.58
total	73.29		61.58		81.84		76.13	

Table 6: Evolution of average bootstraps A, B, C, and D when considering or not each gene.

in the Topologies 0 and 2, and Bootstrap B corresponds to the basal position of *E. oligarthra*, considered as a sister species of all the other *Echinococcus* in Topologies 1 and 3.

With and without each gene, the average bootstrap in A (resp. B, C, and D) of all the trees that present this particular topology has been computed. To have an overall and global view of the effect of each gene on the bootstrap values, average computations have been preferred to minimum ones. Obtained results are presented in the Table 6. In the absence of gene *rpb2*, an average bootstrap of 75.98 (dropping to 55.93 when including *rpb2*) has been computed, pointing out the influence of genes *rpb2* and *ef1a* on the bootstraps A and B. Bootstraps C and D, for their part, are mainly affected by gene *pepck*.

To have a more rigorous validation of this type of influence and to determine which genes are associated with bootstraps of the different topologies, a Linear Model trained with L1 prior as regularizer (LASSO) test has been performed 4 times. By doing so, a gene that summarizes the evolution of the considered bootstrap, either by increasing it or by reducing it, has been identified. Then, by reducing the constant that multiplies the L1 term in the LASSO test, supplementary genes appear one by one, ordered by their ability to summarize the bootstrap value.

According to the bootstraps values C and D shown in Figure 5, two additional LASSO tests have been performed to determine the genes that are related to the position of *E. equinus* in the phylogenetic tree. They have been performed using `linear_model` class of `sklearn` [22] Python module, with inputted data being a Boolean array with 19 columns (one per gene presence) and a vector of bootstrap. The number of rows in the array corresponds to the number of trees having the topologies with this bootstrap (Topologies 0 and 2 for bootstrap A, and so on). The associated row in the vector provides the bootstrap in the tree obtained with this selection of genes.

Bootstrap	Genes ordered by influences
A	*pepck,nad6, nad4, nad5, cob, cox3, rrnL, nad1, elp, cox1, atp6, cox2, ef1a,pold, rpb2,nad3,nad2,rrnS,nad4l.*
B	*pepck,nad6, rpb2, cob,nad4,rrnL,nad5,cox3, nad1, elp, ef1a, cox2, atp6, cox1,nad4l,nad3, pold,rrnS, nad2*
C	*rpb2, cox1, ef1a,nad5,rrnL, nad6, cox2,cob, rrnS, nad3, nad2, nad4l, atp6, elp, nad1,pold, nad4, cox3, pepck.*
D	*rpb2, cox1, rrnL, pepck, elp, cox2, ef1a, nad1, cox3, nad3, nad5, atp6, pold, nad4l, nad2, rrnS, nad4, cob, nad6.*

Table 7: Genes with the largest impact regarding bootstraps A, B, C, and D (important ones are listed first).

Obtained results for each bootstrap are provided in Table 7. Using these results in combination with the previous ones, supplementary arguments are provided, that show that *pepck* is the gene breaking sister relationship between *E. oligarthra* and *E. vogeli*, while *rpb2* and *ef1a* tend to move the position of *E. equinus*. To some extent, it is possible to conclude that *pepck* has a negative impact on the position of *E. equinus* while *rpb2* has a role in the erroneous position of *E. vogeli* in Topologies 1 and 3.

Finally, the most likely phylogenetic tree for *Echinococcus* genus, resulting from the statistical analyses presented in previous sections, appears to be the Topology 0. As stated before, it is the most frequent tree obtained during our investigations of the influence of genes on the phylogeny of *Echinococcus*, and it is also the most robust tree.

4. Conclusion

Deep investigation of the molecular phylogeny of the genus *Echinococcus* has been performed in this paper. 19 coding sequences, taken from both mitochondrial and nuclear genomes, have been considered for maximum likelihood phylogenetic reconstruction. As the trees that we obtained were not as robust that we could expect, combinations from 13 to 19 genes have been further investigated. This analysis has generated $43,796$ trees that represent 15 topologies. The position of *E. equinus* and the sister relationship between *E. oligarthra* and *E. vogeli* led us to the selection of 4 specific topologies. Using the logit

model and the LASSO, Topology 0 has appeared as the most probable one. Finally, negative effects of several genes for that particular topology have been emphasized.

In future work, the phylogeny of the genus *Taenia* [1] will be explored using a similar approach, as some species seem to have discrepant phylogenetic positions. All the possible combinations of the 14 mitochondrial genes (12 protein-coding genes and 2 ribosomal RNAs) will be considered, leading to the production of $16,384$ phylogenetic trees. Their topologies will be compared, and the influence of each gene on these topologies will be rigorously measured, in order to determine the most probable phylogenetic tree of this species. Finally, the phylogeny of the class *Eucestoda* will be investigated using a similar approach.

All computations have been performed using the Mésocentre de Calcul de Franche-Comté facilities. Authors would like to thank Arnaud Mouly for helpful discussions, and the staffs of Asahikawa Medical University (Japan) and of other institutes that have kindly submitted their sequences to the GenBank database.

References

[1] H. Al-Nayyef, C. Guyeux, and J. M. Bahi. Taenia biomolecular phylogeny and the impact of mitochondrial genes on this latter. In *2015 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, pages 1–7, Aug 2015.

[2] Bassam AlKindy, Jean-François Couchot, Christophe Guyeux, Arnaud Mouly, Michel Salomon, and Jacques M. Bahi. Finding the core-genes of chloroplasts. *Journal of Bioscience, Biochemistery, and Bioinformatics*, 4(5):357–364, 2014. Journal version of ICBBS14 conference.

[3] Bassam Alkindy, Christophe Guyeux, Jean-François Couchot, Michel Salomon, and Jacques Bahi. Gene similarity-based approaches for determining core-genes of chloroplasts. In *BIBM14, IEEE Int. Conf. on Bioinformatics and Biomedicine*, Belfast, United Kingdom, nov 2014.

[4] Bassam Alkindy, Christophe Guyeux, Jean-François Couchot, Michel Salomon, Christian Parisod, and Jacques Bahi. Hybrid genetic algorithm and lasso test approach for inferring well supported phylogenetic trees based on subsets of chloroplastic core genes. In *Second International Conference, AlCoB 2015, Mexico City, Mexico, August 4-5, 2015, Proceedings*, volume 9199, pages 83–96, Mexico City, Mexico, aug 2015. Springer.

[5] J. William O. Ballard and Michael C. Whitlock. The incomplete natural history of mitochondria. *Molecular Ecology*, 13(4):729–744, 2004.

[6] D. A. Benson, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and E. W. Sayers. GenBank. *Nucleic Acids Res.*, 37(Database issue):26–31, Jan 2009.

[7] Alexei J Drummond and Andrew Rambaut. Beast: Bayesian evolutionary analysis by sampling trees. *BMC Evol Biol*, 7:214, 2007.

[8] R. C. Edgar. Muscle: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, 5(1), August 2004.

[9] S. Guindon, JF Dufayard, V. Lefort, M. Anisimova, W. Hordijk, and O. Gascuel. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of phyml 3.0. *Systematic Biology*, 59(3):307–321, 2010.

[10] Christophe Guyeux, Jean-Marc Nicod, Laurent Philippe, and Jacques Bahi. The study of unfoldable self-avoiding walks. application to protein structure prediction software. *JBCB, Journal of Bioinformatics and Computational Biology*, *(*):***–***, 2015. Available online.

[11] Michael Hardman and Lotta M. Hardman. Comparison of the phylogenetic performance of neodermatan mitochondrial protein-coding genes. *Zoologica Scripta*, 35(6):655–665, 2006.

[12] Jenny Knapp, Minoru Nakao, Tetsuya Yanagida, Munehiro Okamoto, Urmas Saarma, Antti Lavikainen, and Akira Ito. Phylogenetic relationships within echinococcus and taenia tapeworms (cestoda: Taeniidae): An inference from nuclear protein-coding genes. *Mol Phylogenet Evol*, 2011.

[13] Nicolas Lartillot, Thomas Lepage, and Samuel Blanquart. PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. *Bioinformatics*, 25(17):2286–2288, September 2009.

[14] T. H. Le, M. S. Pearson, D. Blair, N. Dai, L. H. Zhang, and D. P. McManus. Complete mitochondrial genomes confirm the distinctiveness of the horse-dog and sheep-dog strains of Echinococcus granulosus. *Parasitology*, 124(Pt 1):97–112, Jan 2002.

[15] Donald P McManus, Thanh Hoa Le, and David Blair. Genomics of parasitic flatworms. *International Journal for Parasitology*, 34(2):153 – 158, 2004. Annual Scientific Meeting of the Australian Society for Parasitology, Darwin, Carlton Hotel, The Esplande, 2003. Highlights.

[16] M. Nakao, A. Lavikainen, T. Iwaki, V. Haukisalmi, S. Konyaev, Y. Oku, M. Okamoto, and A. Ito. Molecular phylogeny of the genus Taenia (Cestoda: Taeniidae): proposals for the resurrection of Hydatigera Lamarck, 1816 and the creation of a new genus Versteria. *Int. J. Parasitol.*, 43(6):427–437, May 2013.

[17] M. Nakao, D. P. McManus, P. M. Schantz, P. S. Craig, and A. Ito. A molecular phylogeny of the genus echinococcus inferred from complete mitochondrial genomes. *Parasitology*, 134:713–722, 5 2007.

[18] M. Nakao, N. Yokoyama, Y. Sako, M. Fukunaga, and A. Ito. The complete mitochondrial DNA sequence of the cestode Echinococcus multilocularis (Cyclophyllidea: Taeniidae). *Mitochondrion*, 1(6):497–509, Oct 2002.

[19] Minoru Nakao, Antti Lavikainen, Tetsuya Yanagida, and Akira Ito. Phylogenetic systematics of the genus echinococcus (cestoda: Taeniidae). *International journal for parasitology*, 43(12-13):10171029, November 2013.

[20] Minoru Nakao, Tetsuya Yanagida, Sergey Konyaev, Antti Lavikainen, Valeriy A Odnokurtsev, Vladimir A Zaikov, and Akira Ito. Mitochondrial phylogeny of the genus echinococcus (cestoda: Taeniidae) with emphasis on relationships among echinococcus canadensis genotypes. *Parasitology*, 140(13):1625–1636, 2013.

[21] C. Notredame, D. G. Higgins, and J. Heringa. T-Coffee: A novel method for fast and accurate multiple sequence alignment. *Journal of molecular biology*, 302(1):205–217, September 2000.

[22] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[23] Ros Stamatakis and Thomas Ludwig. Raxml-omp: An efficient program for phylogenetic inference on smps. In *In Proc. of PaCT05*, pages 288–302, 2005.

[24] D. L. Swofford. PAUP*: phylogenetic analysis using parsimony, version 4.0b10. 2011.

[25] R. C. Thompson and D. P. McManus. Towards a taxonomic revision of the genus Echinococcus. *Trends Parasitol.*, 18(10):452–457, Oct 2002.

[26] Stacia K. Wyman, Robert K. Jansen, and Jeffrey L. Boore. Automatic annotation of organellar genomes with dogma. *Bioinformatics*, 20(17):3252–3255, 2004.

[27] R. Zardoya and A. Meyer. Phylogenetic performance of mitochondrial protein-coding genes in resolving relationships among vertebrates. *Molecular Biology and Evolution*, 13(7):933–942, 1996.

5. Appendices

Species	Gene	DOGMA location	NCBI location	Differences (bp)
E. canadensis	NAD4	4445-5650	4394-5650	51
E. equinus	NAD4	4339-5547	4291-5547	48
E. equinus	COX2	12475-13041	12475-13053	12
E. felidis	NAD4	4355-5557	4301-5557	54
E. felidis	COX2	12497-13069	12497-13075	6
E. granulosus	NAD4	2070-3275	2019-3275	51
E. granulosus	COX2	10182-10748	10182-10760	12
E. granulosus	NAD5	11763-13328	11763-13331	3
E. multilocularis	COX2	12602-13168	12602-13180	12
E. oligarthra	NAD4L	4216-4473	4201-4473	15
E. oligarthra	NAD4	4488-5696	4437-5696	51
E. oligarthra	NAD2	6419-7294	6413-7294	6
E. oligarthra	COX2	12659-13225	12659-13237	12
E. ortleppi	NAD4	4442-5647	4391-5647	51
E. ortleppi	COX2	12588-13154	12588-13163	9
E. shiquicus	NAD4L	4218-4472	4212-4475	9
E. shiquicus	NAD4	4490-5695	4439-5695	51
E. shiquicus	COX2	12664-13236	12664-13242	6
E. vogeli	NAD4	4470-5675	4419-5675	51
E. vogeli	COX1	9189-10793	9189-10782	12
E. vogeli	COX2	12623-13189	12623-13204	15
V. mustelae	NAD5	434-1990	434-1999	9
V. mustelae	NAD4L	4074-4325	4074-4331	6
V. mustelae	NAD4	4340-5551	4295-5551	45
V. mustelae	ATP6	5766-6269	5757-6269	9
V. mustelae	NAD2	6298-7149	6280-7152	21
V. mustelae	NAD3	8588-8905	8561-8905	18
V. mustelae	COX1	9041-10654	9041-10657	3
V. mustelae	COX2	12466-13038	12466-13041	3

Table 8: Comparaison between DOGMA and NCBI annotations

	coef	std err	z	$P > \|z\|$	[95.0% Conf. Int.]
atp6	0.3123	0.059	5.290	0.000	0.197, 0.428
cob	2.1253	0.066	32.418	0.000	1.997, 2.254
cox1	-2.5048	0.073	-34.238	0.000	-2.648, -2.361
cox2	-0.1443	0.059	-2.434	0.015	-0.261,-0.028
cox3	-0.4606	0.060	-7.686	0.000	-0.578, -0.343
ef1a	2.7846	0.071	39.332	0.000	2.646, 2.923
elp	-0.4456	0.060	-7.442	0.000	-0.563, -0.328
nad1	-0.5614	0.060	-9.326	0.000	-0.679, -0.443
nad2	0.8478	0.060	14.184	0.000	0.731, 0.965
nad3	-0.1777	0.059	-2.995	0.003	-0.294, -0.061
nad4	1.5521	0.062	24.899	0.000	1.430, 1.674
nad4l	0.8943	0.060	14.933	0.000	0.777, 1.012
nad5	-0.5913	0.060	-9.808	0.000	-0.710, -0.473
nad6	2.6145	0.069	37.742	0.000	2.479, 2.750
peck	-4.2902	0.102	-41.923	0.000	-4.491,-4.090
pold	0.2501	0.059	4.237	0.000	0.134, 0.366
rbp2	10.9685	0.314	34.888	0.000	10.352, 11.585
rrnl	-1.5005	0.065	-23.256	0.000	-1.627, -1.374
rrns	0.9407	0.060	15.675	0.000	0.823, 1.058

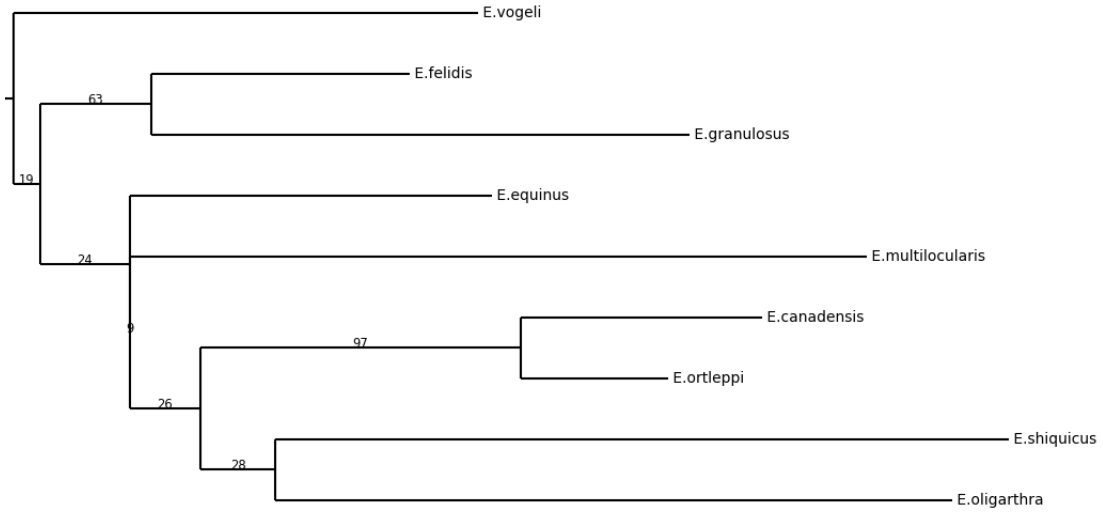
Table 9: Dummy logit regression results for Topology 1

	coef	std err	z	$P > \|z\|$	[95.0% Conf. Int.]
atp6	0.0119	0.040	0.301	0.763	-0.066, 0.089
cob	-1.3876	0.044	-31.288	0.000	-1.474, -1.301
cox1	0.1494	0.039	3.794	0.000	0.072, 0.227
cox2	0.1869	0.039	4.748	0.000	0.110, 0.264
cox3	1.6599	0.041	40.116	0.000	1.579, 1.741
ef1a	0.4144	0.039	10.554	0.000	0.337, 0.491
elp	1.1480	0.040	28.706	0.000	1.070, 1.226
nad1	1.5069	0.041	36.849	0.000	1.427, 1.587
nad2	0.4016	0.039	10.228	0.000	0.325, 0.479
nad3	0.6277	0.039	15.966	0.000	0.551, 0.705
nad4	-1.9148	0.048	-40.187	0.000	-2.008, -1.821
nad4L	0.2582	0.039	6.569	0.000	0.181, 0.335
nad5	-1.4657	0.045	-32.717	0.000	-1.553, -1.378
nad6	-1.8420	0.047	-39.056	0.000	-1.934, -1.750
peck	5.8118	0.063	92.599	0.000	5.689, 5.935
pold	0.2647	0.039	6.734	0.000	0.188, 0.342
rbp2	-3.7189	0.060	-61.680	0.000	-3.837, -3.601
rrnl	1.5806	0.041	38.438	0.000	1.500, 1.661
rrns	0.4384	0.039	11.166	0.000	0.361, 0.515

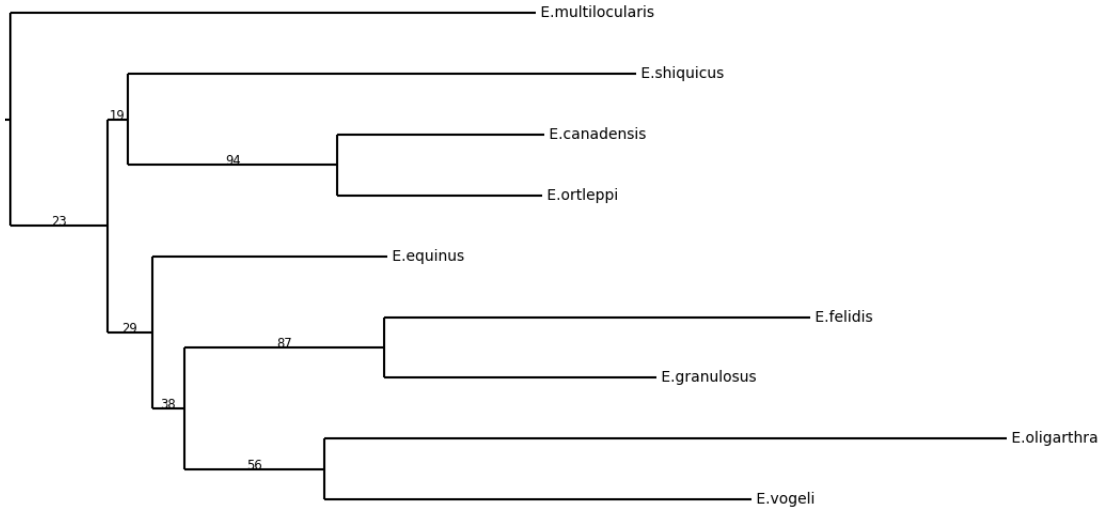
Table 10: Dummy logit regression results for Topology 2

	coef	std err	z	$P > \|z\|$	[95.0% Conf. Int.]
atp6	-0.0903	0.052	-1.752	0.080	-0.191, 0.011
cob	-0.7260	0.055	-13.105	0.000	-0.835, -0.617
cox1	-1.6773	0.065	-25.798	0.000	-1.805, -1.550
cox2	-0.3101	0.053	-5.893	0.000	-0.413, -0.207
cox3	0.7471	0.049	15.152	0.000	0.650, 0.844
ef1a	0.8376	0.049	17.006	0.000	0.741, 0.934
elp	0.3783	0.050	7.581	0.000	0.280, 0.476
nad1	0.6704	0.049	13.576	0.000	0.574, 0.767
nad2	0.5778	0.050	11.670	0.000	0.481, 0.675
nad3	0.1269	0.051	2.505	0.012	0.028, 0.226
nad4	-1.0815	0.058	-18.488	0.000	-1.196, -0.967
nad4l	0.4068	0.050	8.163	0.000	0.309, 0.504
nad5	-1.6045	0.064	-25.014	0.000	-1.730, -1.479
nad6	-1.0399	0.058	-17.899	0.000	-1.154, -0.926
peck	3.4841	0.060	57.968	0.000	3.366, 3.602
pold	0.2551	0.050	5.080	0.000	0.157, 0.354
rpb2	5.2340	0.077	68.151	0.000	5.083, 5.385
rrnl	0.4480	0.050	9.007	0.000	0.351, 0.546
rrns	0.5551	0.050	11.203	0.000	0.458, 0.652

Table 11: Dummy logit regression results for Topology 3

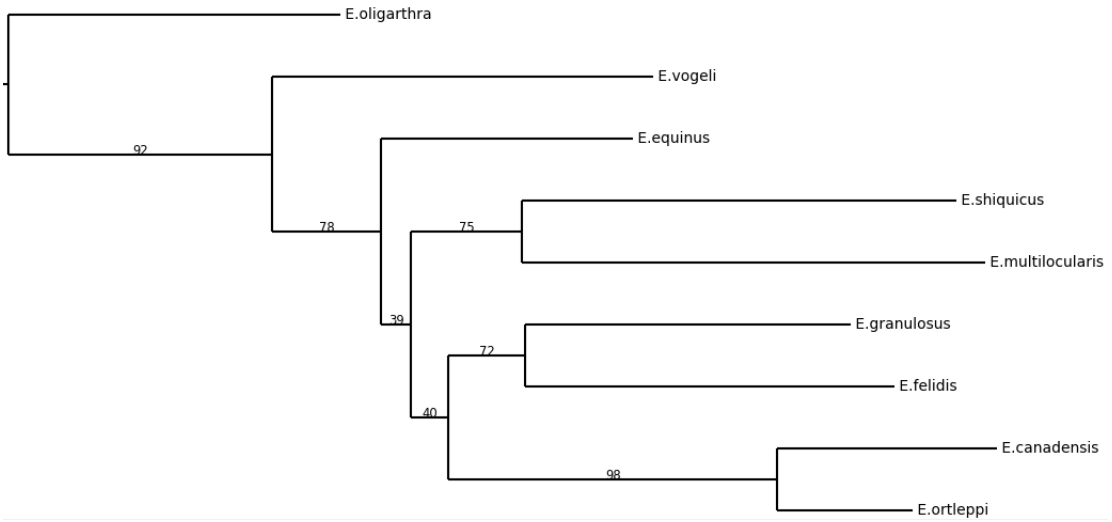


(a) atp6

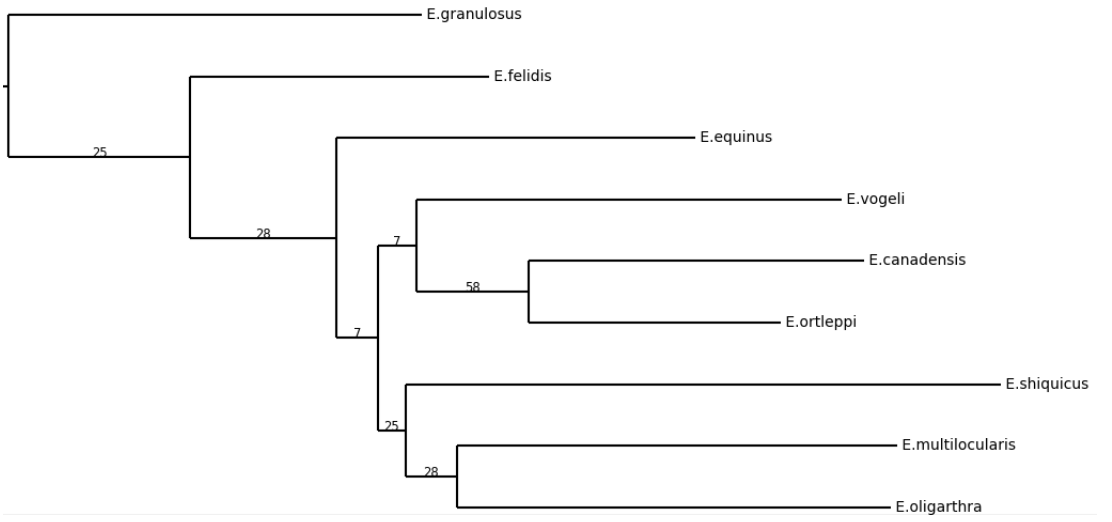


(b) cob

Figure 6: Gene trees of the 19 considered genes

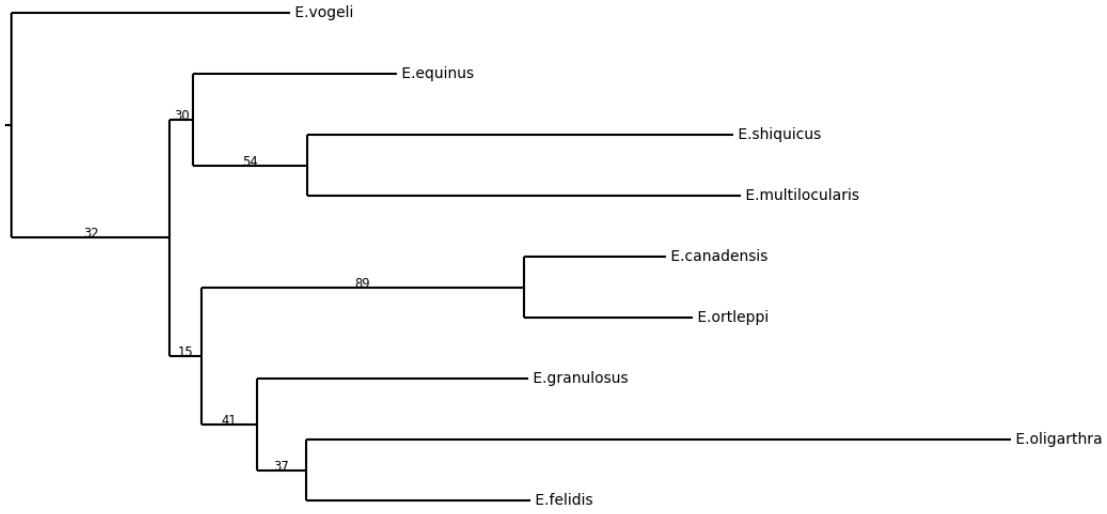


(a) cox1

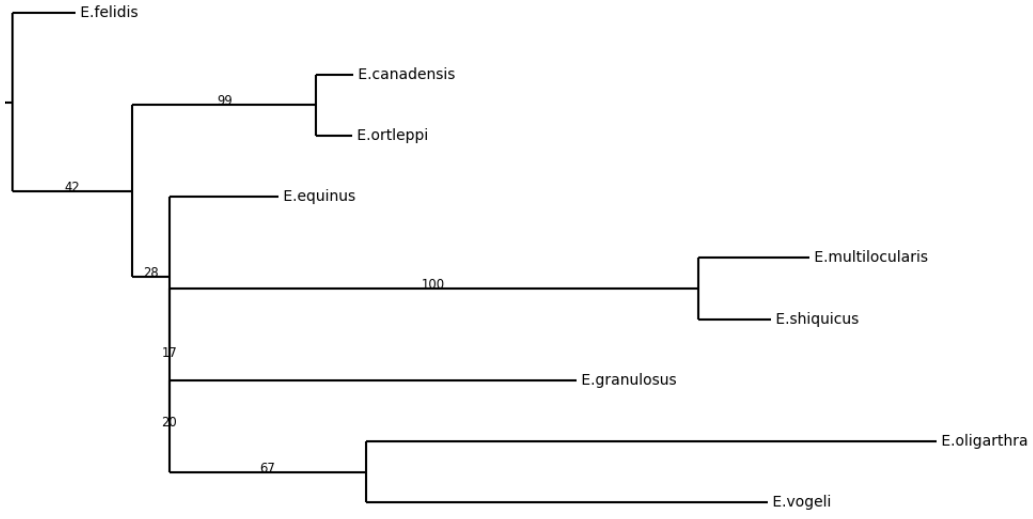


(b) cox2

Figure 7: Gene trees of the 19 considered genes (cont.)

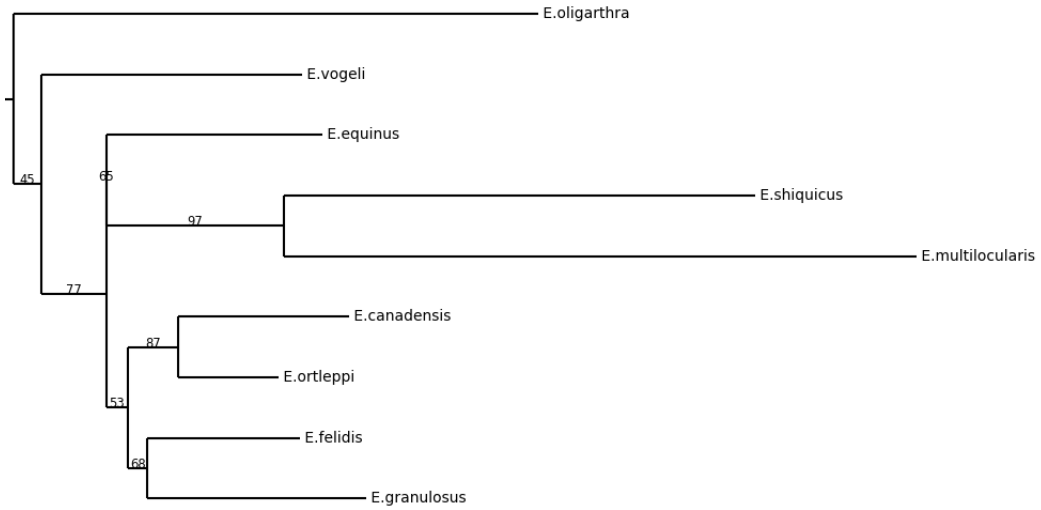


(a) cox3

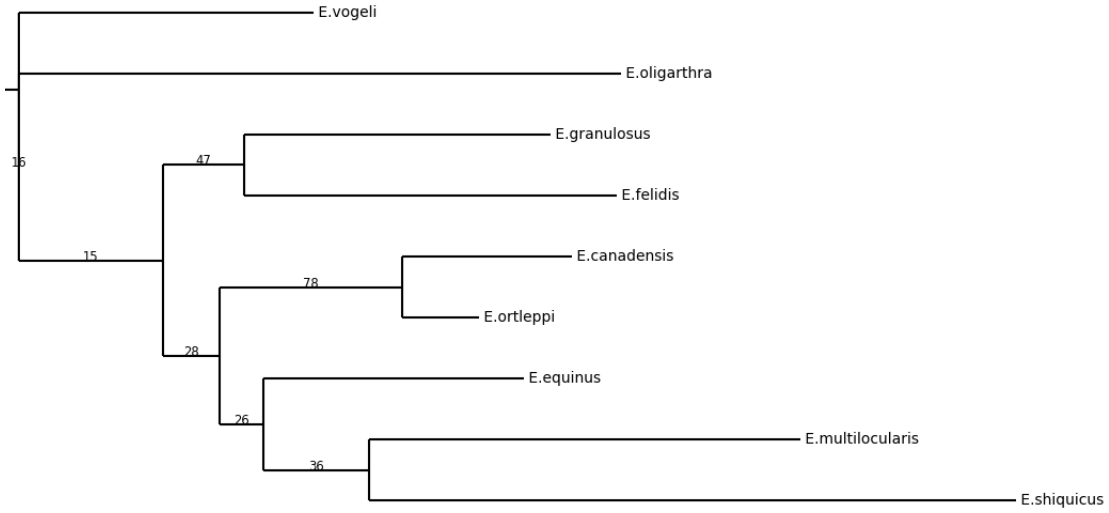


(b) ef1a

Figure 8: Gene trees of the 19 considered genes (cont.)

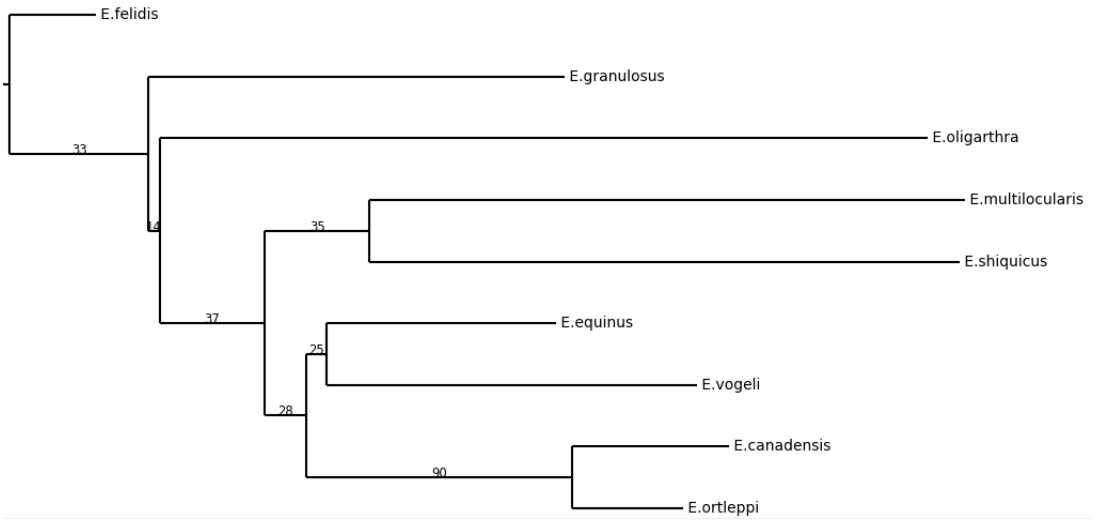


(a) elp

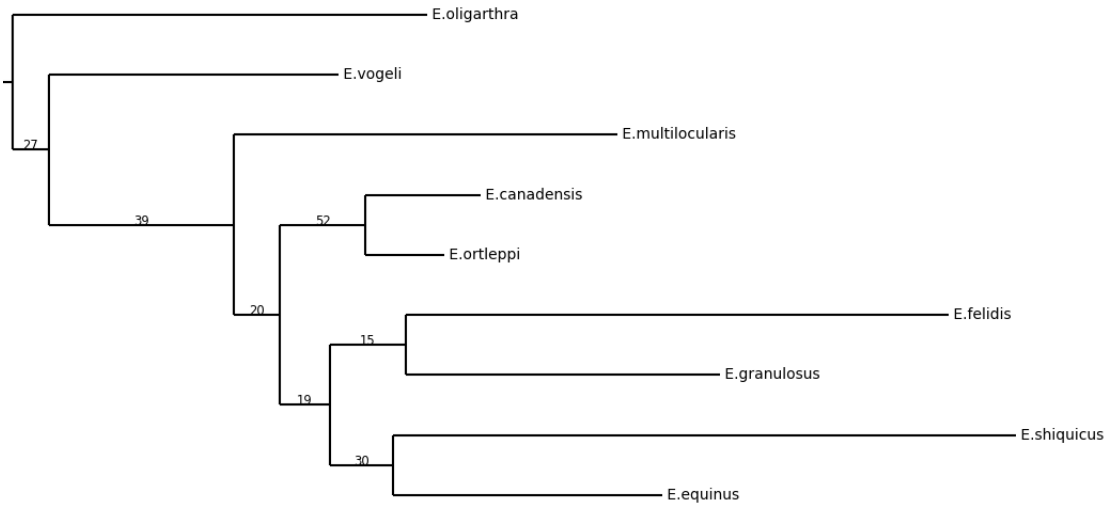


(b) nad1

Figure 9: Gene trees of the 19 considered genes (cont.)

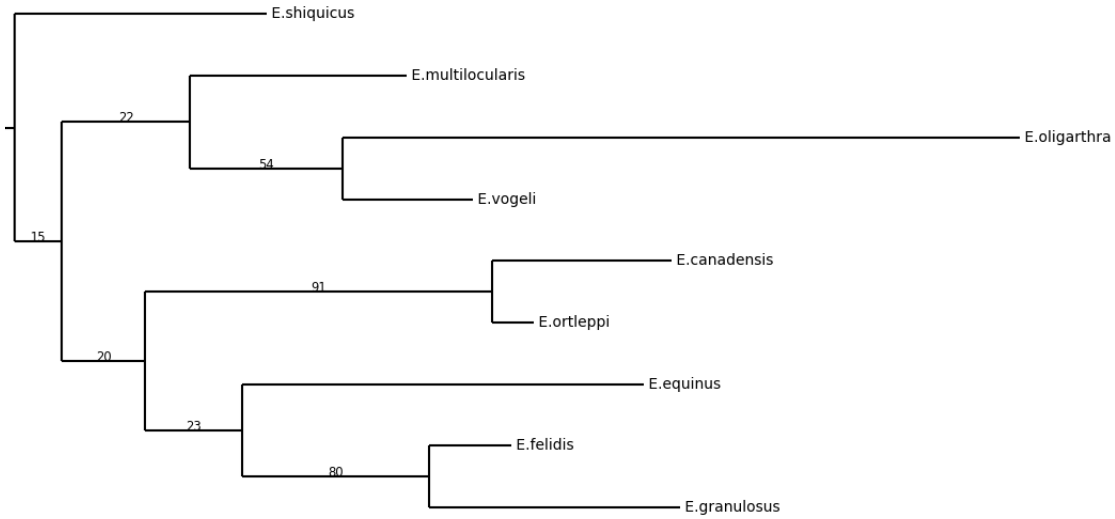


(a) nad2

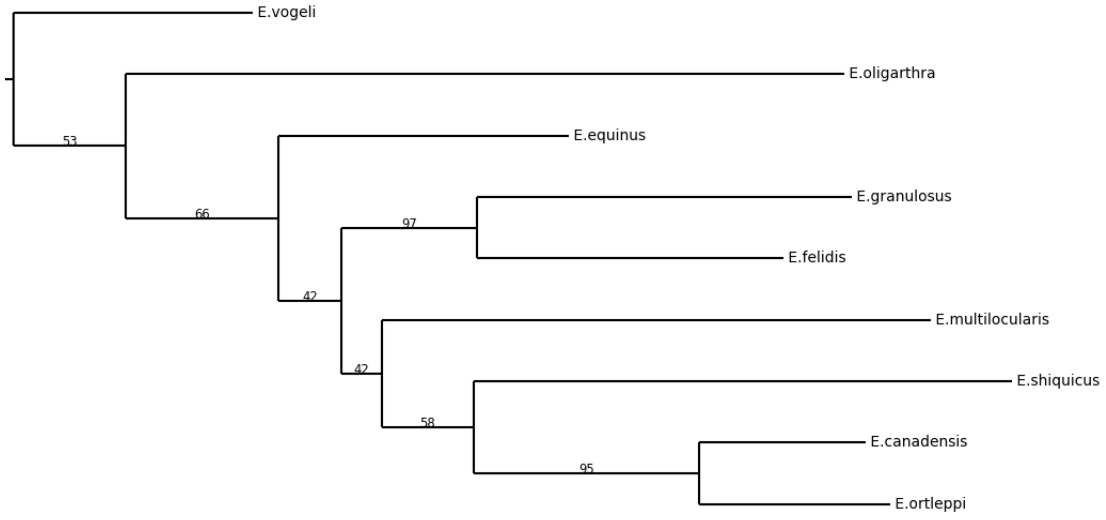


(b) nad3

Figure 10: Gene trees of the 19 considered genes (cont.)

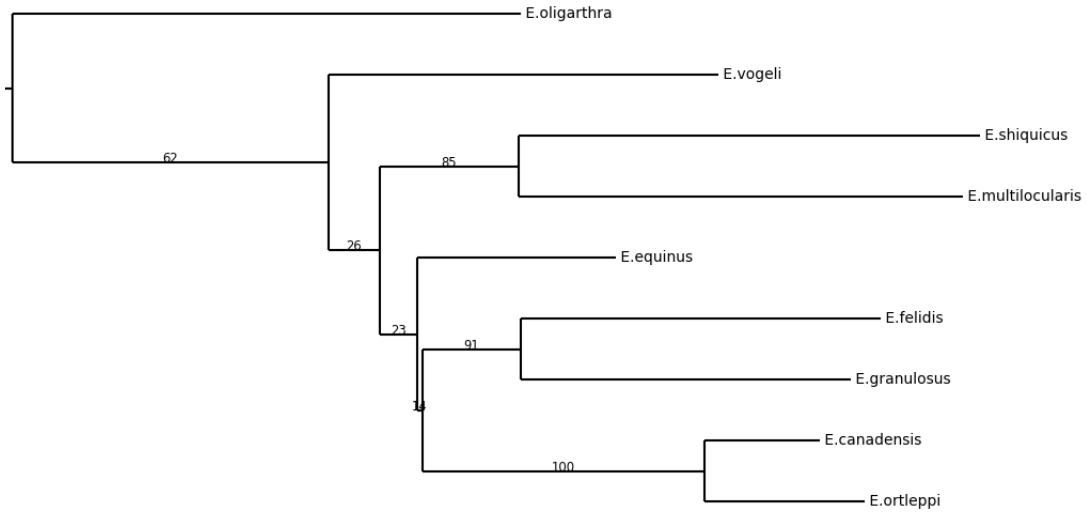


(a) nad4l

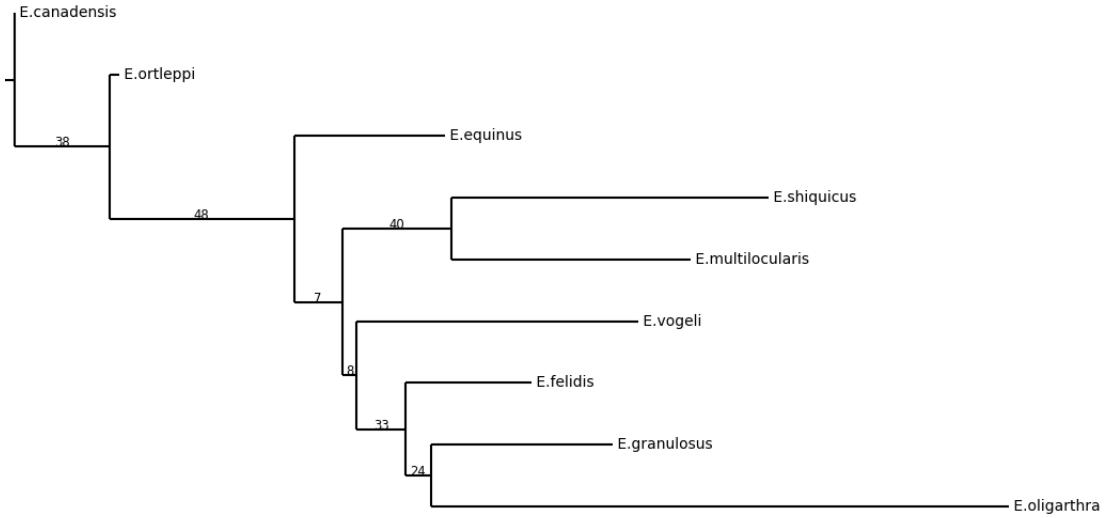


(b) nad4

Figure 11: Gene trees of the 19 considered genes (cont.)

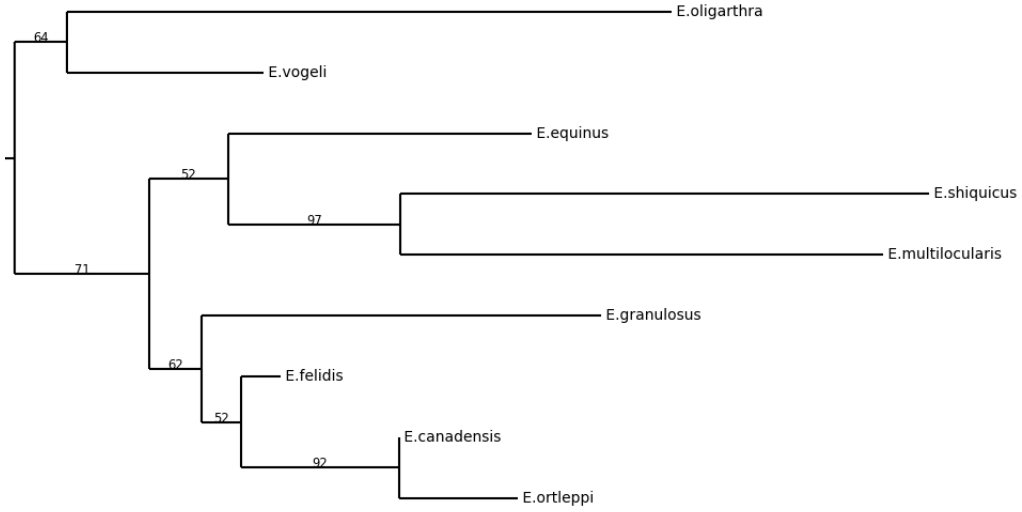


(a) nad5

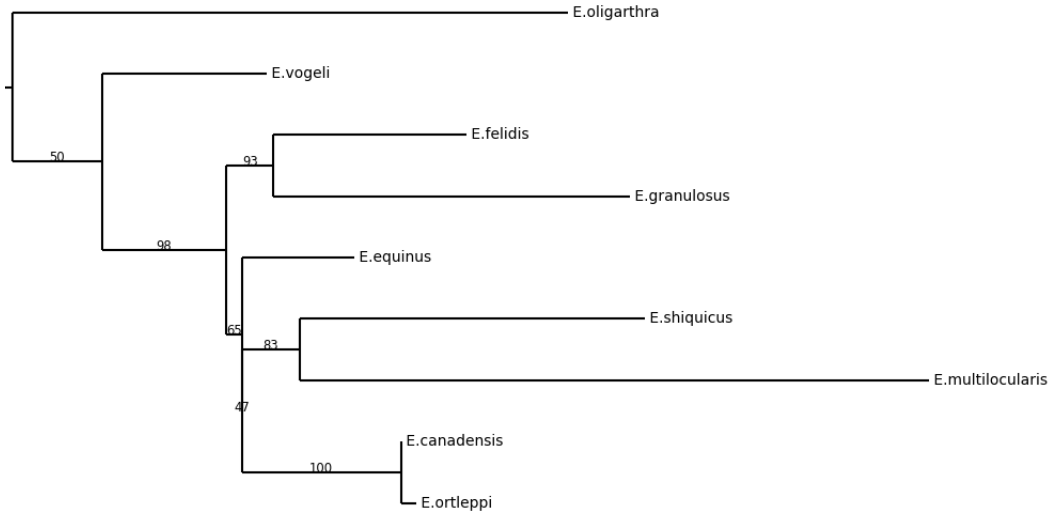


(b) nad6

Figure 12: Gene trees of the 19 considered genes (cont.)

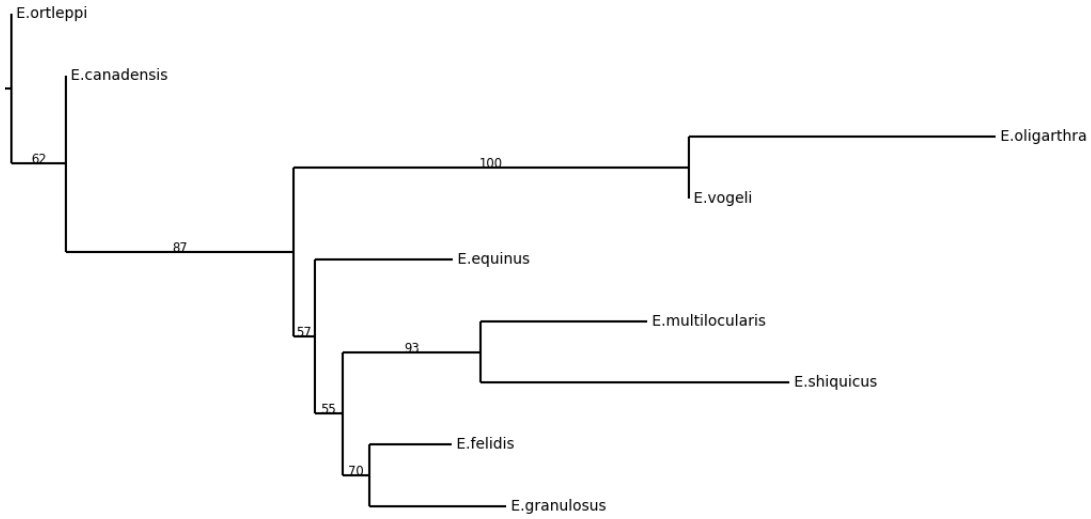


(a) pepck

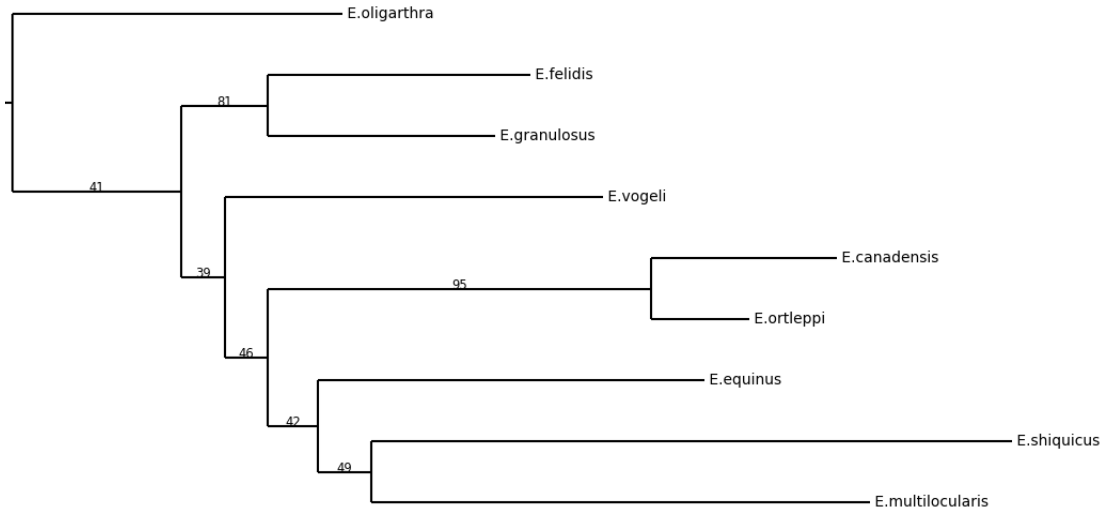


(b) pold

Figure 13: Gene trees of the 19 considered genes (cont.)

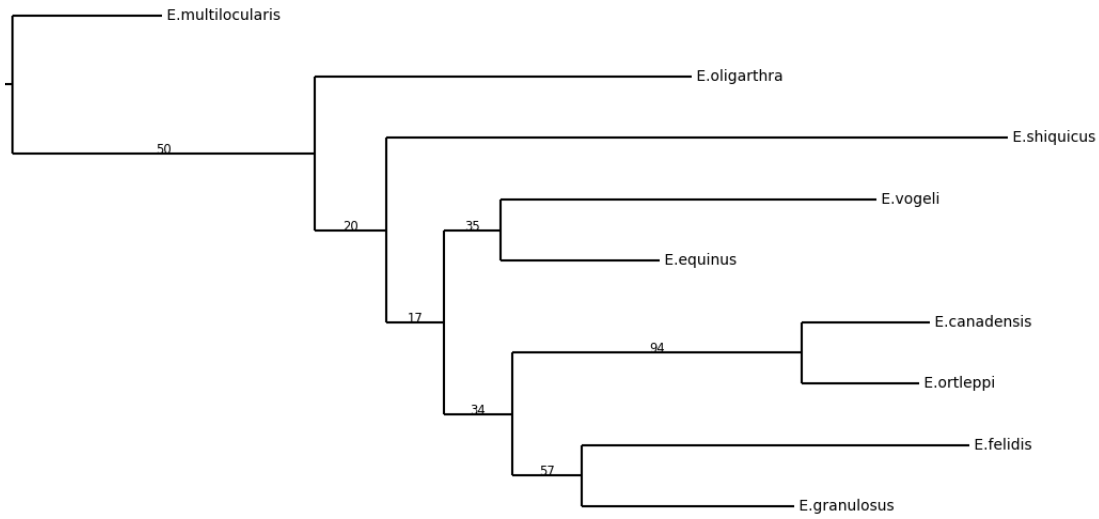


(a) rpb2



(b) rrnL

Figure 14: Gene trees of the 19 considered genes (cont.)



(a) rrnS

Figure 15: Gene trees of the 19 considered genes (end)