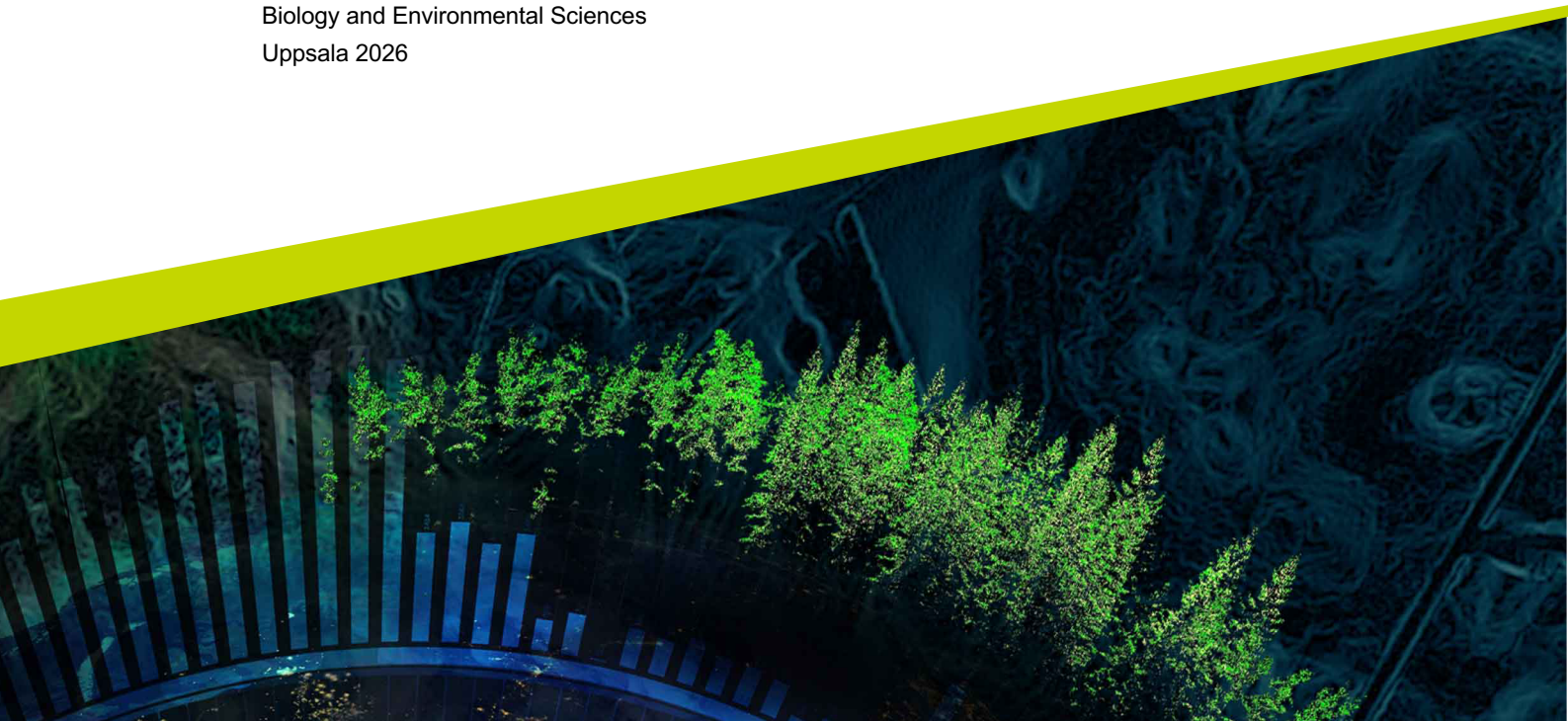
Lepidopteran Phylogenetics

The Largest Phylogenetic Tree of Butterflies and Moths Sequenced to Date

Josefin Westberg Larsson

Independent project • 15 credits
Swedish University of Agricultural Sciences, SLU
Department of Ecology
Biology and Environmental Sciences
Uppsala 2026



Lepidopteran Phylogenetics. The Largest Phylogenetic Tree of Butterflies and Moths Sequenced to Date

Fylogenin hos Lepidoptera. Det största fylogenetiska trädet över fjärilar som hittills sekvenserats.

Josefin Westberg Larsson

Supervisor:	**Nicolas Chazot, Swedish University of Agricultural Sciences, Department of Ecology**
Assistant supervisor:	Johanna Orsholm, Swedish University of Agricultural Sciences, Department of Ecology
Examiner:	Mariana Braga, Swedish University of Agricultural Sciences, Department of Ecology

Credits:	15
Level:	First cycle, G2E
Course title:	Independent Project in Biology
Course code:	EX0894
Programme/education:	Biology and Environmental Sciences
Course coordinating dept:	Department of Aquatic Sciences and Assessment
Place of publication:	Uppsala
Year of publication:	2026
Copyright:	All featured images are used with permission from the copyright owner.
Keywords:	Phylogeny, Lepidoptera, phylogenetic tree, phylogenetic diversity, BOLD barcodes, reference tree

Swedish University of Agricultural Sciences
Faculty of Natural Resources and Agricultural Sciences
Department of Ecology
Insect Ecology Unit

Abstract

Phylogenetic trees are used to illustrate and study evolutionary relationships between organisms or genes and are created using gene sequences from taxa. Barcode reference trees are used for studying ecological and evolutionary patterns, for classifying taxa or for measuring phylogenetic diversity. Although well-known, we still have a lot to learn about the Lepidoptera order of insects. In this study, I first created a genus-level reference tree based on 24 lepidopteran superfamilies that was then expanded into a BOLD-species supertree of Lepidoptera. The reference tree was created using maximum likelihood in IQ-tree, and the BOLD tree was made by adding BOLD taxa to the backbone, using phylogenetic placement and online tree inference (OTI). The resulting reference tree contained 5331 genera from 25 superfamilies, and the BOLD tree contained 54738 taxa. The quality of the tree was checked by measuring grouping of sequences according to their taxonomy and measuring how well sequences grouped to their reference genera. The sequences grouped relatively well according to taxonomy, indicating a successful tree inference, and the sequences placed themselves near their reference taxon. Accumulation curves of phylogenetic diversity (PD) for each superfamily were also made and the maximum PD and PD completeness estimated. The PD completeness ranged around 40-60 % indicating that there is still a lot of phylogenetic diversity to be discovered. To conclude, this tree provides a new overview of the current understanding of the Lepidopteran diversity and can be used for further identification of Lepidopteran taxa.

Keywords: phylogeny, lepidoptera, phylogenetic tree, phylogenetic diversity, BOLD barcodes, reference tree

Table of contents

1. Introduction

A phylogenetic tree is a graphical representation of evolutionary relationships between organisms or genes, and it can be used for classification, studying the evolution of genes and traits in morphology, ecology and biogeography as well as for conservation. The tips of the tree represent the existing taxa and the nodes between two branches in the tree represent the most recent common ancestor (MRCA) of those branches. The length of the branches between the nodes can represent the number of substitutions that divide them from one another, and the branching pattern makes out the topology of the tree. All taxa that share a branch is called a clade and has monophyletic origin, i.e. share a common ancestor. A tree can be rooted or unrooted. Rooting of a tree is essential for understanding the evolutionary relationships between the taxa, since the root represents the MRCA of all taxa in the tree.

A reference tree is a phylogenetic tree that is only built with named species with a known placement and evolutionary history. It can be used as the backbone for building on new, unknown taxa and further classification of unknown barcodes (Czech et al, 2018). A barcode is a short fraction of DNA sequenced from a specific gene region (Hajibabaei, 2007). The standardized barcode for animals is the COI gene, or Cytocrome oxidase subunit I. A reference tree can also be used in ecological analyses, such as for telling different evolutionary histories in community ecology or to describe biodiversity (Cavender-Bares, 2009).

Since all the tips of a rooted tree share a common ancestor, phylogenetic trees can be used to understand evolutionary processes over time, where the tips of the tree represent the present and the root the most ancient time in the past (Ricklefs, 2007). Each branch in the tree can either split into two branches (speciation) or be cut off (extinction), resulting over time in the diversity of species that we have today. One use of phylogenetic trees is thus to estimate rates of speciation and extinction, as well as these rates variation over time and among clades.

Because all species in a community differ from one another based on their divergence from their shared most recent common ancestor, phylogenies can also be a tool to understand the species assemblies of ecological communities (Webb et al, 2002). Analysing the phylogenetic relationships between species in a community, e.g. whether they are closely or distantly related from each other, can help us answer questions regarding ecology and community structure. The distribution and function of taxa can be understood better by integrating knowledge of their functional traits with their phylogenetic relationships (Cavender-Bares, 2009). Different traits are innovations over the evolutionary tree of life and therefore tend to be shared by species with common ancestry resulting

in them often having similar niches, and evolution can thus be considered to play a key role in the assembly of communities. Phylogeny can also be used for understanding biogeographical patterns such as historical movements of taxa across regions, habitats and biomes (Webb et al, 2002; Cavender-Bares, 2009).

A phylogenetic tree can be used as an alternative measure of diversity, by bringing more information than just species richness. This is called phylogenetic diversity (PD), and it is measured by summing over the branch lengths of sequences added to the tree and thereby accounts for the evolutionary relationships between taxa (Faith, 1992). For conservation biology, the concept of PD can be used based on the assumption that PD diversity is linked to feature diversity, since species with similar features often share a common ancestor (Faith, 2018). With this approach to conservation, one values the diversity of biological traits over the conservation of specific species, which can be beneficial because different features often correspond to different ecosystem functions.

If we want to estimate the phylogenetic diversity (PD) in a community, a useful tool is to make accumulation curves of the barcodes of the known, sampled taxa (Smith, 2009). The accumulation curves can be used to estimate the total diversity of a taxonomic group where a lot of species remain unsampled, which is useful information in ecology and conservation.

Identification and classification of unknown species can be done using genetic information such as barcodes. Barcodes can be used to quickly identify species and sort them into phylogenetic groups (Hajibabaei, 2007). Identification and classification of unknown species can be done using a reference tree, by finding the phylogenetic placement of the species. Phylogenetic placement is a set of methods to identify genetic sequences, called query sequences, by placing them in a fixed reference tree (Czech et al, 2022). By inferring the query sequences on the given set of reference sequences in the reference tree, phylogenetic information about the sequences is considered when identifying them (Barbera et al, 2018). Unknown species can thus be identified by finding the most likely phylogenetic placement of them in the reference tree.

Undoubtably, reference trees are very useful for a range of fields in biology, such as classification, evolution, ecology and conservation. The Lepidoptera order might be one of the most well-known insect orders and has been the object of many entomologists and common people's interest throughout history. They are estimated to have appeared about 300 Ma in the late carboniferous (Kawahara et al, 2019), and as one of the largest insect orders, with about 160000 known species, they have since spread their wings all over the Earth, evolving into a wide range of superfamilies and genera.

There are many existing studies of phylogeny among Lepidoptera. On one hand there are studies that tries to resolve the relationships between the major Lepidoptera clades (e.g. Kawahara et al, 20019) but containing a very tiny

fraction of total diversity. On the other hand, there are well sampled phylogenies (sometimes hundreds or thousands of species, see for example Chazot et al (2021)). Therefore, there is a need to synthesize into a single tree all these works.

Though well-studied, being such a large group, the phylogenetic relationships between many members of the order are still relatively unknown. Papilionoidea, i.e. butterflies, is probably one of the most well-studied superfamilies. For example, Chazot et al (2019) composed a genus-level phylogeny of Papilionoidea with 994 taxa. Another charismatic superfamily is Bombycoidea with some of the largest lepidopteran species and is also relatively well-studied (e.g. Hamilton et al, 2009). There is however a lack in the research of many Lepidopteran superfamilies, and these must be studied further to fully understand the diversity of the order Lepidoptera. With an increasing amount of genetic data available online, the opportunities to create high-level phylogenetic trees has become easier than ever. A reference tree for all Lepidoptera would be a useful tool for further classification and knowledge about ecological and evolutionary patterns of this order of insects.

The aim of this study was to construct a species-level reference tree containing almost all species of Lepidoptera sequenced to date. To make the tree, I first created a genus-level backbone with 5331 genera. Then I added all named Lepidoptera species sequences from BOLD, Barcode of Life Data Systems, which is a database with freely available DNA barcodes (Ratnasingham, 2024). This resulted in a species-level supertree containing 54738 taxa. The quality of the tree was tested by measuring how well sequences grouped according to their taxonomy and if they grouped to their reference taxon. The phylogenetic diversity sampled was finally estimated for each superfamily.

2. Methods

2.1 Collecting Genetic Data

For each known genus of Lepidoptera, accession codes of 1-7 genes, as many as were available for each genus, were collected from GenBank at the National Center for Biotechnology and Information (NCBI). The genes collected were Cytochrome oxidase subunit 1 (COI), Elongation factor 1 alpha (EF1a), Carbamoyl-phosphate synthetase (CAD), wingless (wgl), Cytosolic malate dehydrogenase (MDH), Glyceraldehyde-3-phosphate dehydrogenase (GAPDH) and Ribosomal protein S5 (RpS5). A small section of COI has been massively sequenced as it is the standard barcode of insects and must be included for using taxonomic assignment of unknown taxa. Therefore, if COI did not exist in GenBank for a specific genus, that genus was excluded altogether. Only known, scientifically described genera of Lepidoptera were used, meaning that if a genus was known genetically but had no name, it was excluded. If the same gene had been uploaded to GenBank several times for the same genus (e.g. for multiple species of the same genus) the accession code of the gene with the longest sequence was collected.

When all the accession codes were collected, the sequences were downloaded from GenBank using the function read.GenBank() in R (tidyverse-, ape- and readxl-packages), producing one file for each superfamily and gene. All superfamilies were finally compiled together by gene. In total, there were 25 superfamilies (Alucitoidea, Bombycoidea, Calliduliodea, Copromorphoidea, Cossoidea, Drepanoidea, Epermenioidea, Gelechioidea, Hyblaeoidea, Geometroidea, Immoidea, Noctuoidea, Mimallonoidae, Papilionoidea, Pterophoroidea, Pyraloidea, Schreckensteinioidea, Sesioidea, Simaethistioidea, Tineoidea, Tortricoidea, Uranioidea, Urodoidea, Yponomeutoidea and Zygaenoidea) and 5433 genera. Since all genera were sequenced for the COI gene, there was also a total number of 5433 sequences of the COI gene. For the other genes, wgl had 2390 sequences, RPS5 had 1877, CAD had 1698, MDH had 1625 and EF1A had 2789 collected sequences.

2.2 Alignment

The program MAFFT (v7.526) was used to align the sequences to make sure they were homologous and in the same position across genes and species. MAFFT is a command line program and the text file for each gene was used as input file and the output files were set as fasta-format. The "--Auto" parameter was added for all alignments, which switches the algorithm according to the data size (Katoh, 2013). For most alignments the default gap penalty of 1,53 was used. If the output alignment had a lot of gaps with this option, the gap penalty was changed with the "--op" parameter to 2.0.

The alignments were adjusted by hand by removing off looking sequences that aligned poorly with the rest of the sequences and gap only columns, and the insertions/deletions were adjusted to maintain the codon structure. If a species was removed from the COI-gene, the barcode, it was also removed from the other genes. The adjusted alignments were re-run in MAFFT to get a new alignment based on the adjustments. This was done repeatedly until no obvious rogue sequence could be identified in the alignment.

To check for long branches that could indicate misalignments or odd sequences, some simple gene trees were run using IQ-tree 2.4. Long branches were checked in the alignment and if the sequences looked weird compared to the rest of the alignment, e.g. due to lots of gaps or substitutions that could indicate a poorly made alignment, they were deleted from the alignment. The finished alignments for the different genes were concatenated into one big alignment including all genes. Now the dataset contained 5331 genera in total.

2.3 Creating the Backbone

A backbone tree was made with one known species per genus, to use as a reference tree for later addition of species. The backbone was made with the command line program IQ-tree 2.4 to create the base of the tree. IQ-tree uses maximum likelihood methods to infer the tree (Minh et al, 2020). Maximum likelihood involves searching for the optimal tree topology by examining different tree topologies and choosing the one with the highest likelihood. To root the tree, six outgroup species were added to the concatenation. The outgroup species were belonging to Trichoptera, the sister group of Lepidoptera.

Expecting that each gene is mutating at a different rate, the dataset was partitioned to let IQ-tree estimate the values of the substitution parameters for each gene separately. In the partition, the substitution model set for each gene was GTR+I+G, i.e. general time-reversable model with invariable site plus discrete gamma model.

A constraint for the tree topology was set in IQ-tree to fix the already known topology in the tree, using the "-G" parameter. That means that families and

superfamilies were forced to be monophyletic. The relationships between most superfamilies were constrained using Kawarahara et al. 2019. The superfamilies not included in Kawahara et al, 2019, were left unconstrained, leaving IQ-tree to find their positions on its own. These superfamilies were Copromorphoidea, Epermenioidea, Galacticoidea and Schreckensteinioidea. The outgroups were constrained as well.

2.4 Adding BOLD species

2.4.1 BOLD download and alignment

A BOLD download of all lepidoptera species was aligned using MAFFT v.7. Only the taxa with known species names were used, a total of 49883 taxa. First, all the BOLD species were aligned, with gap penalty set at 2.0 and adjusted manually to reduce the gaps. After the adjustment, the alignment still had a lot of gaps, so it was run again in MAFFT, this time with a gap penalty at 3.0. The BOLD alignment was aligned with the COI alignment of the backbone, creating the full alignment. Using R (Tidyverse, ape), the full alignment was trimmed to only have 658 bases per sequence. Sequences of less than 100 bases were removed altogether from the alignment. After the alignment the BOLD taxa had decreased to 49844 and the final dataset of the full alignment (BOLD+backbone COI) had 54738 sequences.

2.4.2 Phylogenetic Placements and Online Tree Inference

When working with trees of this size, common methods of tree building such as the one we used when creating the backbone, starts to get very computationally intensive. Therefore, we used a new method called Online Tree Inference (OTI), that uses phylogenetic placement to repeatedly ad species to an existing reference tree. Phylogenetic placement is a method that involves placing query sequences in a reference tree (Barbera et al, 2018). For this, the program EPA-ng was used, which first makes a preplacement of the query sequences by finding likely candidate branches for them and then determines the optimal placement of the queries via maximum likelihood. EPA-ng required a reference tree, the reference alignment, the COI substitution model and the full BOLD+backbone-alignment to run, so these were prepared in beforehand.

Using the new method Online Tree Inference, OTI, the phylogenetic inference was looped repeatedly over one query after another, adding fixed placements of

species to the backbone after each loop until all sequences were in the tree. The OTI was done on a supercomputer, using EPA-ng for the phylogenetic placement.

The COI substitution model, GTR + I + G, that was used when building the backbone, in which IQ-tree had estimated the value of the parameters, was specified for the program.

The BOLD+backbone-alignment, the reference tree, the reference alignment and the COI model were sent to the supercomputer. Because EPA-ng does not work with invariable sites, +I, in the substitution model, that parameter was removed. To reduce the risk of getting a tree topology that is not representable to reality due to chance, the OTI was run 10 times, giving 10 different tree topologies. For the following analyses however, only one of the trees was used.

2.5 Analyses

The finished trees were used for analyses. Three different analyses were done. As a sanity check of the tree building method, we tested if BOLD barcodes grouped by their taxonomy (genus name) rather than randomly. As a second sanity check, we tested if BOLD sequences grouped with their reference taxon or randomly in the tree. As a third analysis we measured how well the sampled sequences in the tree represent phylogenetic diversity.

2.5.1 Grouping of BOLD barcodes according to their taxonomy

To measure grouping in the tree, that is how well the taxa cluster according to their genera, pairwise distances of tips within versus between genera were calculated using R, ape-package, cophenetic()-function. Pairwise distances were calculated as the branch lengths between two tips in the tree. The branch length corresponds to the number of substitutions, meaning that the longer the branch length between two species, the more mutations separate them from each other. If taxa are grouping well in the tree, the pairwise distance within genera should therefore be smaller than the one between genera.

First, the reference genera from the backbone were removed from the BOLD tree, to make sure that all names of the tips were built similarly, and to reduce the risk of duplication of the species used for the reference since they are also among the BOLD species. This resulted in a reduced tree containing only the BOLD species and not their reference genera. The pairwise distances were calculated from the reduced tree and a pairwise distance matrix was made. The pairwise distances between each species of each genus were calculated, as well as the pairwise distances between the species of each genus and a species from a random genus. The results were plotted as histograms, one including the distances of all genera, as well as separate ones for each superfamily respectively.

2.5.2 Grouping of BOLD sequences to reference taxon compared with random taxon

To see if the phylogenetic distances between the BOLD species and their reference genera were smaller than the distances between the BOLD species and a random reference taxon, pairwise distances were calculated in R, ape-package. The distances between all BOLD species that had a matching genus reference and their respective reference genera, were compared to the distance between the BOLD species and a random reference genus. The distance from the root of the tree to every tip of the tree was first calculated and for each pair of sequences the most recent common ancestor (MRCA) between the grafted sequence and its corresponding reference genus was found. The pairwise distance was then calculated by summing over the root-to-tip distance for each tip and then subtracting the distance between the MRCA and the root times 2. The same procedure was done for the grafted sequence and a random reference genus. The calculated pairwise distances were plotted as a histogram.

2.5.3 Accumulation Curves

To measure how well the sampled sequences in the tree represent phylogenetic diversity (PD), accumulation curves were made using R, ape-package. At first, when adding samples to the tree, the accumulation rate is fast, as adding taxa in the beginning contributes to a lot of new phylogenetic diversity, making the curve steep in the beginning (Cardoso et al, 2014). The more taxa that are added, the less phylogenetic diversity is gained for each taxon since a lot of features have already been added, and the accumulation curve levels off with increasing number of taxa.

The BOLD tree was divided into superfamily subtrees, one for each superfamily, and the reference genera of the trees were removed, so they only contained the BOLD species. For each superfamily tree, one taxon was picked randomly and then one taxon was added at a time in a random order until recreating the complete phylogeny. The total branch length of each subtree was then calculated, so that the resulting curves show the accumulation of new phylogenetic diversity contributed by additional, random taxa.

The maximum phylogenetic diversity was predicted using an asymptotic fitted curve. This was done using the Michaelis-Menten model, that uses the observed curve to estimate when the curve reaches its asymptote, or maximum diversity (Cardoso et al, 2014). The predicted phylogenetic diversity was drawn over the observed accumulation curves to visualize the observed diversity versus the predicted maximum diversity. By fitting the model, we estimated the value of the

parameter in the Michaelis-Menten model that gives us the maximum phylogenetic diversity. By dividing the observed PD by the predicted PD, we get the PD completeness, which is the percentage of phylogenetic diversity that has been observed of the predicted maximum PD. This thus showed the proportion of the predicted maximum diversity that we compared with the observed diversity of the different superfamilies that are known and sequenced.

3. Results

3.1 Backbone and BOLD Super Trees

The resulting backbone tree contained 5331 tips, i.e. 5331 genera from 25 superfamilies (figure 1). The tips were constrained at family-level in their superfamilies except for the Copromorphoidea, Hyblaeoidea, Epermenioidea, Galacticoidea and Schreckensteinioidea, which are still unresolved regarding placement. In previous studies they have placed as follows: Copromorphoidea with Urodoidea or Epermenioidea. Hyblaeoidea with Pyraloidea. Epermenioidea as apoditrysian. Galacticoioidea as sister to Immoidea. Schreckensteinioidea near Urodoidea. In my reference tree, the Copromorphoidea placed with Gelechioidea, Pyraloidea, Papilionoidea and Urodoidea. Hyblaeoidea with Sesioidea. Epermenioidea placed with Calliduloidea, geometrioidea, Noctuoidea and Urodoidea. Galacticoidea grouped with Pyraloidea, and Schreckensteinioidea grouped with Zyganoidea. The BOLD tree contained 54738 tips of known BOLD species and the reference genera from the backbone tree (figure 2). The BOLD species placed themselves generally well according to genus.

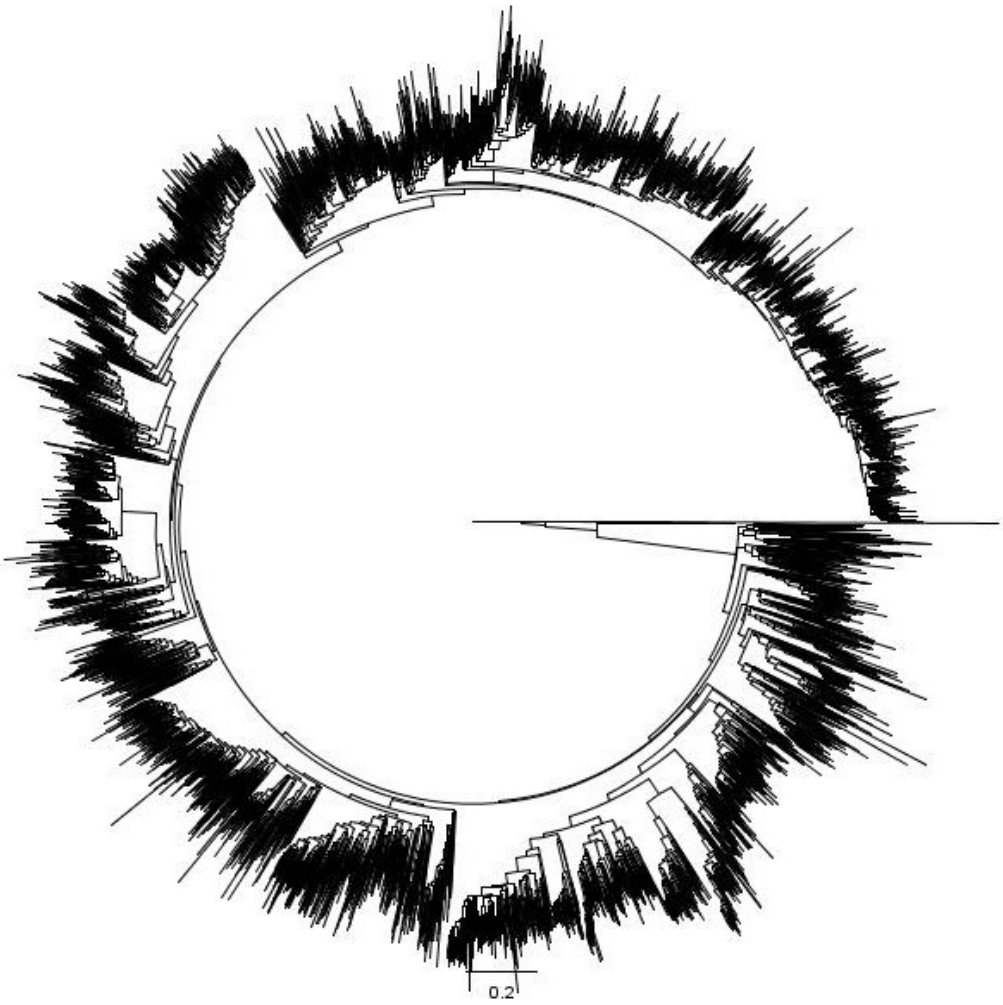


Figure 1. Genus-level backbone tree of Lepidoptera, containing the reference genera, 5331 tips.

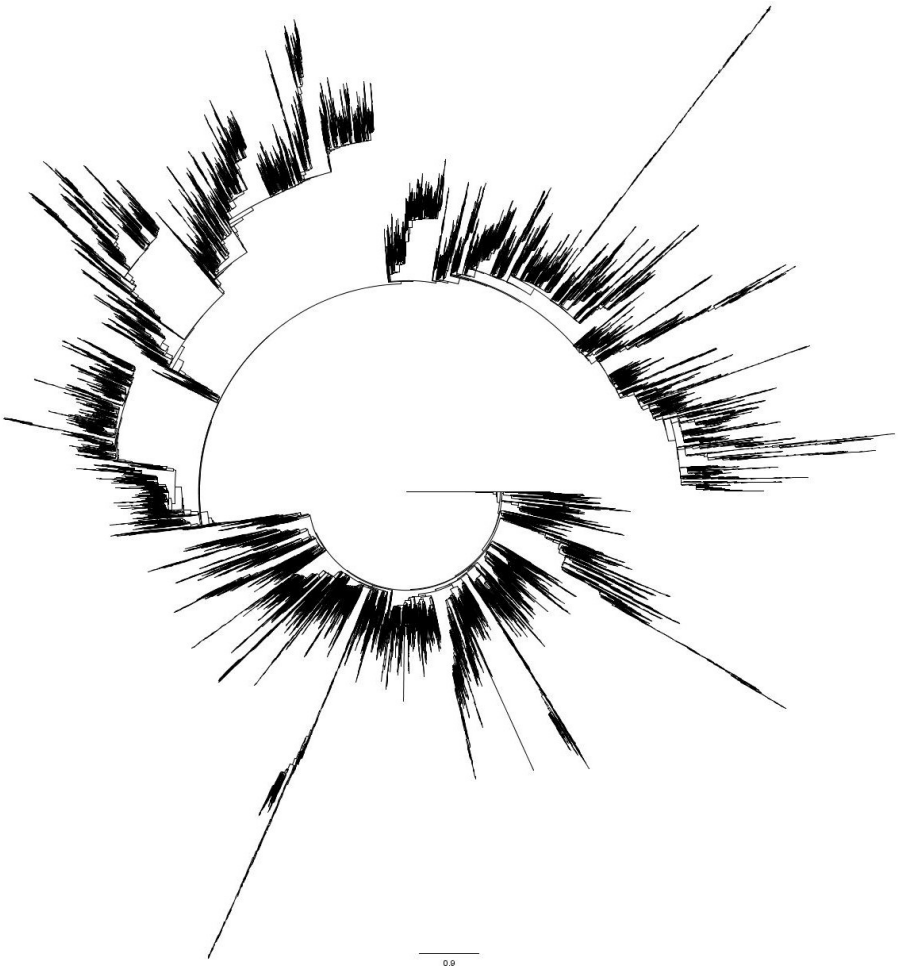


Figure 2. Species-level tree of Lepidoptera with BOLD species and reference genera, 54738 tips.

3.2 Are BOLD barcodes grouping according to their taxonomy or randomly?

The pairwise distance within each genus was generally smaller than the ones between different genera (figure 3). However, quite a lot of overlap was seen, where the distance was the same within versus between genera. Overall, the online tree inference successfully recovered phylogenetic signal with a tree matching the expected taxonomy.

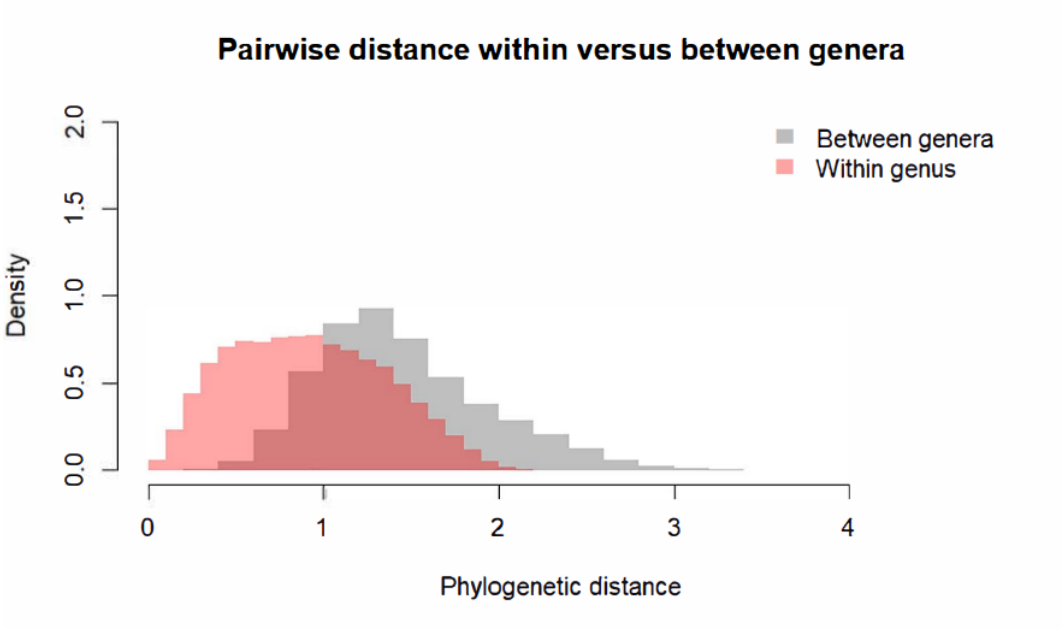


Figure 3. Pairwise distance within versus between genera. The pink colour shows the pairwise distance within genus, the grey colour the pairwise distance between genera and the darker pink is the overlap between the two. The X axis shows the phylogenetic distance and the Y axis the density of taxa.

3.3 Are BOLD sequences grouping with their reference taxon or randomly in the tree?

The phylogenetic distance between BOLD species and their reference genus was generally smaller than the distances between BOLD species and a random reference genus (figure 4). A small overlap is seen between the two. Generally, BOLD sequences are grouping well with their reference taxon as they are placed near their expected genus.

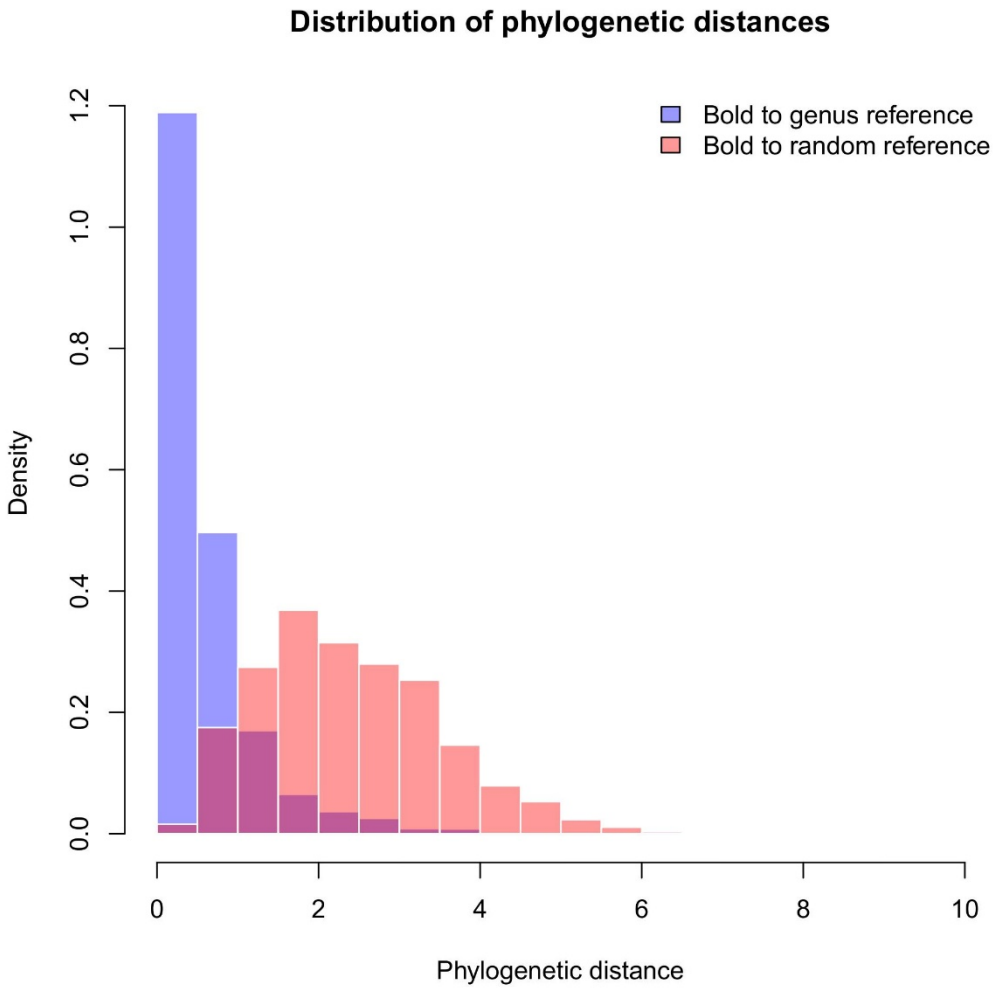


Figure 4. The distribution of phylogenetic distances between BOLD species and their reference genus (blue bars), compared to the distances between BOLD species and a random reference genus (pink bars). The purple colour is due to overlap between the two. The x-axis shows the phylogenetic distance (branch length) and the y-axis the density of taxa with that specific phylogenetic distance.

3.4 Observed vs fitted Accumulation Curves of Phylogenetic Diversity

The observed accumulation curves (full lines, Figure 5 and 6) are steep at first, as expected, and start to level off as we sample more species. The curves show that the superfamily with the highest observed phylogenetic diversity (PD) is the Noctuoidea and the one with the lowest observed PD is the Mimallonoidae. The

fitted curves (dashed lines, figure 5 and 6) show that for all superfamilies the current sampling remains far from the asymptotic (maximal) PD. For many of the superfamilies, the fitted curve starts levelling off a bit earlier than the observed curves, which could indicate an underestimation of the maximum phylogenetic diversity and the need for more appropriate models.

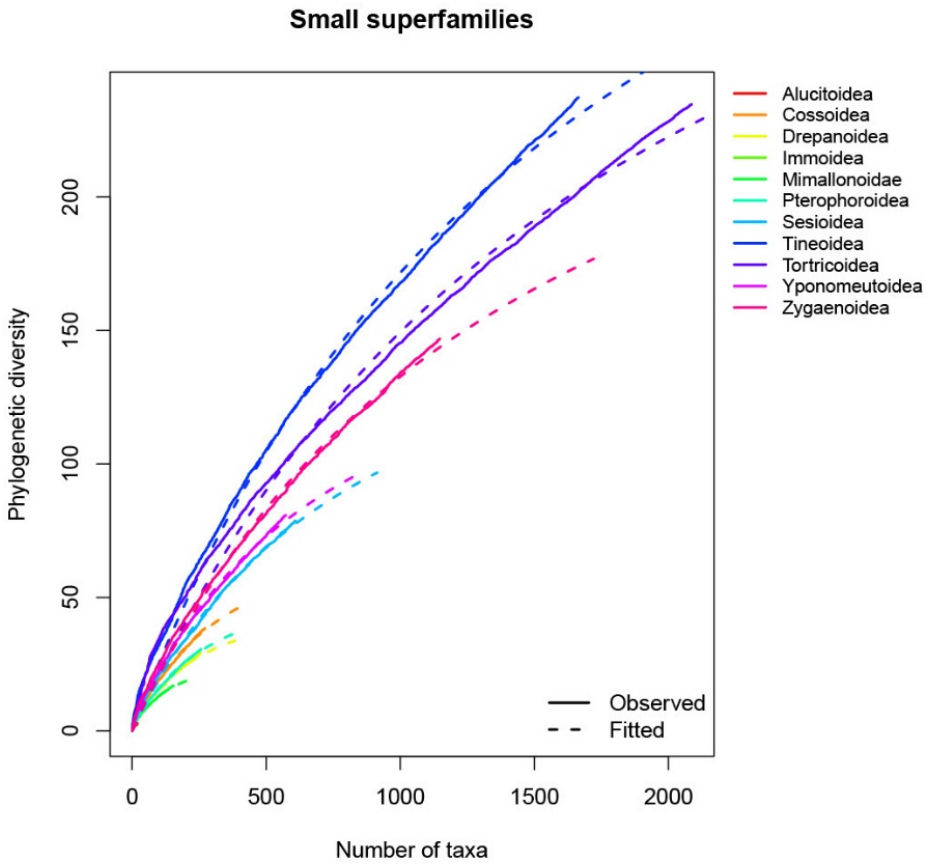


Figure 5. Accumulation curves of the small superfamilies of <3000 sampled species. The full lines are the observed phylogenetic diversity of the sequenced taxon in the tree. The dashed lines represent the predicted maximum phylogenetic diversity calculated by the Michaelis-Menten model.

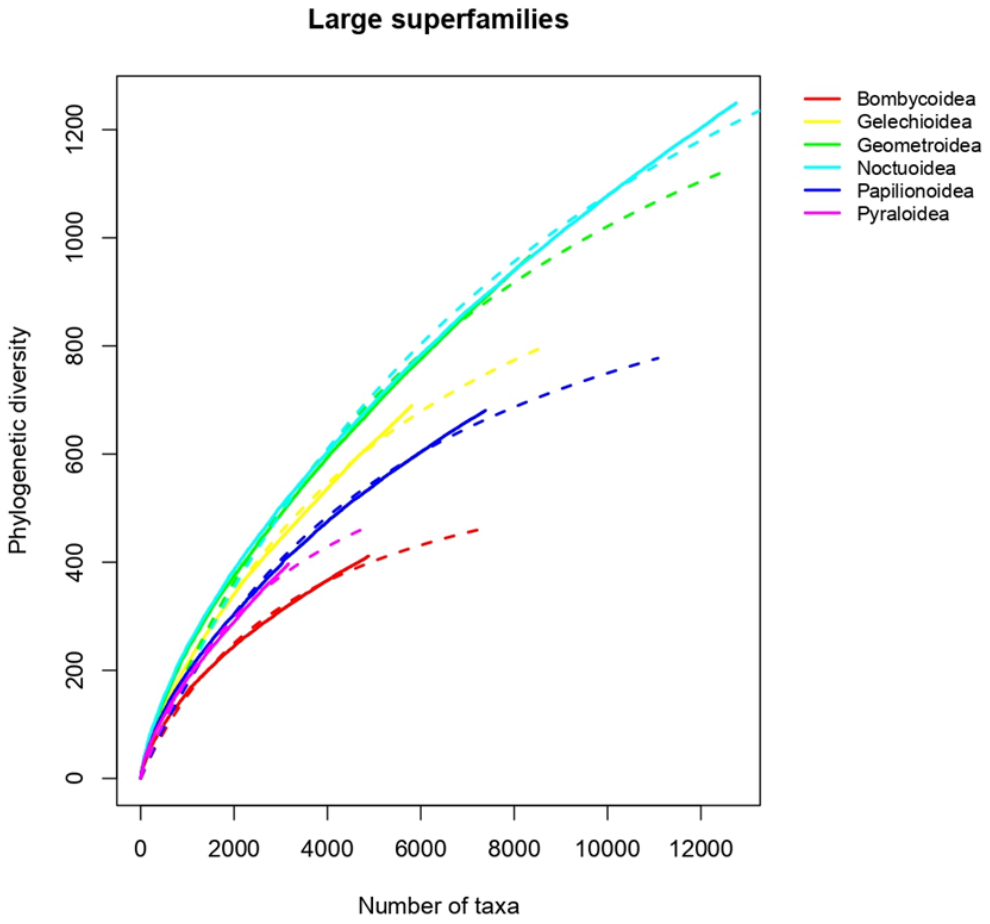


Figure 6. Accumulation curves of the large superfamilies, ≥ 3000 sampled species. The full lines are the observed phylogenetic diversity of the sequenced taxon in the tree. The dashed lines represent the predicted maximum phylogenetic diversity calculated by the Michaelis-Menten model.

3.4 Phylogenetic diversity sampling completeness

The PD completeness ranged between 40-60 % for all measured superfamilies (Figure 7). The Immoidea superfamily is empty because it was too small to measure. The most well-sampled superfamily is the Bombycoidea and the least well-sampled the Sesioidea.

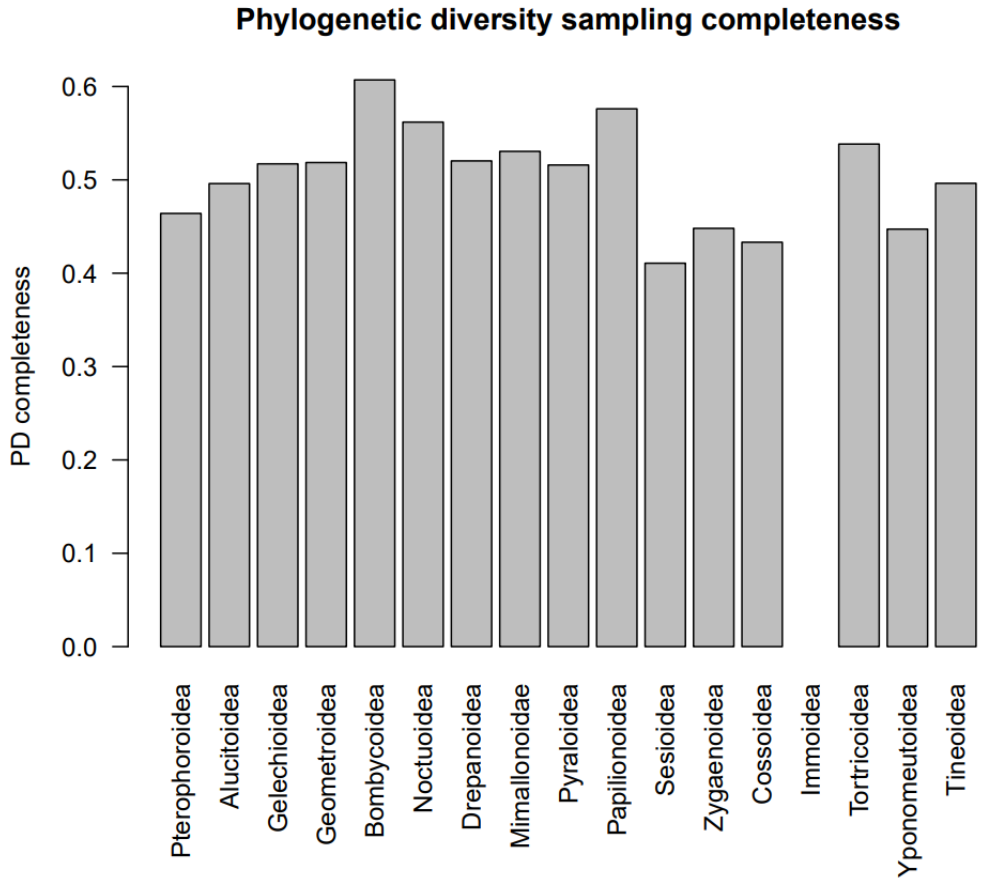


Figure 7. PD completeness for each superfamily. The completeness shows the percentage of the phylogenetic diversity that has been sampled, calculated as observed PD divided by predicted PD. Immoidea is empty due to it being a very small superfamily.

4. Discussion

The pairwise distance within versus between genera can be an indication of how well the sequences group according to their genera. If the pairwise distance within genera is smaller than the one between genera, it strongly indicates that the taxa are grouping well to their respective genera, which supports it being a monophyletic group. Therefore, a good tree should have little overlap within versus between genera. The result of the pairwise distances calculations show a separation between the distances within versus between genera, with tendency towards smaller distances within genus than between genera. This indicate a relatively good grouping according to taxonomy; however, a lot of overlap is still observed, see figure 3. There are at least three explanations for this large overlap. First, the online tree inference is a new method, which has not yet been thoroughly tested for its ability to recover the "true" tree. This could mean that the method we used to build the tree was not providing reliable results.

Second, it could be that some sequences have been classified wrongly in BOLD itself, and they cannot be recovered as monophyletic genera. Since the BOLD sequences were not restricted to be grouped with their reference genus, the placement of a species outside its genus clade could be correct, only that it had been wrongly classified to begin with. Previously, before the breakthrough of gene sequencing, classifications were dominantly made based on morphology rather than genetics, leading to some cases of misclassified taxa. Modern day methods of using genetic data to classify species give us a chance to correct previous wrongly classified species and put them in their true taxonomy.

A third explanation for the overlap is that the genera are different in size. Within a big genus, the distance between species could be relatively long due to a high number of mutations between two distantly related species. On the other hand, the distance between genera could be relatively short if the genus is small and the genus it is compared to is close on the tree. This could give cases where distances between genera are relatively short or within genus relatively long, causing an overlap of distances within versus between genera but no indicative of a low-quality tree.

The distribution of phylogenetic distances turned out to be good in terms of grouping of sequences to their reference genera. The BOLD sequences had a generally smaller phylogenetic distance to their reference genera than to another random reference genus, with only a small overlap, indicating that BOLD sequences placed themselves close to their reference genera, see figure 4.

The accumulation curves show how much phylogenetic diversity is gained by adding more species to the tree. The observed curves, i.e. the curves of the known, sequenced taxa, show steep curves at first, that levels off when adding more taxa. When we start sampling species in the tree, there is a high chance that a new

species represents an entire new clade unsampled before, thereby adding a lot of new phylogenetic diversity. With an increasing number of sampled species in the tree, this probability decreases since the species are added to groups already represented in the tree and these species only bring little new evolutionary history. An uncertainty about the shape of the observed curve is the fact that only one randomization of taxa added was done per superfamily, which brings a need to be careful with the estimates of maximum PD and PD accumulation. This could have given curves that do not really represent reality, as chance may have randomized the tip labels in an order that does not represent the average accumulation of phylogenetic diversity. To validate the results, we need to repeat the randomized addition of species multiple times. This was not done as part of this work due to the long computational time required for this. The PD completeness ranged between 40 to 60 %, which means that there is still a lot of phylogenetic diversity to be discovered, even among the most well-known and well-sampled superfamilies (Papilionoidea, Bombycoidea). The fitted curves started to level off a bit earlier than the observed curves for many of the superfamilies. This could be an indication that the predicted maximum PD has been underestimated by the fitted model, suggesting that the real maximum PD is higher.

To quickly check the quality of the predicted maximum PD, I looked up the total number of species for a few superfamilies (Papilionoidea, Bombycoidea and Pterophoidea) in COL, Catalogue of Life (Bánki, 2026). Catalogue of Life compiles all species that have been recorded, weather they are genetically sequenced or just morphologically known as species. According to COL, Papilionoidea has 21020 recorded species. The observed sequences used in the tree are 7313. This indicates that only about 35 % out of all recorded species have been sequenced. According to the PD completeness the Papilionoidea have about 58 % observed species. For Bombycoidea, the observed sequences were 4806. According to COL, there are 5709 species, which means that about 80 % have been sequenced. The PD completeness indicates that around 61 % have been observed out of the predicted maximum PD. Pterophoroidea have 1611 recorded species in COL and 254 observed in our tree, which is ca 16 % of the recorded species in COL. The predicted maximum PD, however, shows that around 47 % out of predicted maximum PD have been observed. Two out of the three superfamilies (Papilionoidea and Pterophoidea) I looked at, show a higher percentage of observed sequences according to the predicted maximum PD than to the recorded species on Catalogue of Life. This indicates that the value of the maximum PD is underestimated. One of the three superfamilies (Bombycoidea) showed a higher percentage observed species against the Catalogue of Life species than what the PD completeness indicated. That could mean that the

predicted maximum PD suggest that there is more species richness than we know and that there are many more species to be recorded.

To summarize, the quality of the tree building method can be seen as both good and bad. The first test, where I tested the grouping of BOLD barcodes according to taxonomy, showed a separation between the distance within versus between genera with a generally smaller distance within genus, which indicates a good grouping according to taxonomy. The big overlap, however, can be seen as a sign of bad grouping. The distribution of phylogenetic distances of BOLD sequences to their reference taxon compared to a random reference taxon, showed only a small overlap, indicating that the species had placed themselves well to their reference genus. The estimated maximum PD seem to be a bit underestimated when compared to the observed PD and the catalogue of life for some superfamilies, indicating that the measured PD completeness might have been too high for some superfamilies.

5. Conclusion

I produced the largest species-level tree of Lepidoptera, to my knowledge, which gives a new overview of the current understanding of Lepidopteran diversity and sequencing/barcoding effort, although some groups still need extra effort. Overall, this tree is a very good source of information for taxonomic identification of Lepidoptera using phylogenetic placement. This kind of reference tree is needed for further studying of the Lepidopteran order.

References

Bánki, O., Roskov, Y., Döring, M., Ower, G., Hernández Robles, D. R., Plata Corredor, C. A., Stjernegaard Jeppesen, T., Örn, A., Pape, T., Hobern, D., Garnett, S., Little, H., DeWalt, R. E., Miller, J., Orrell, T., Aalbu, R., Abbott, J., Abreu, C., Acero P, A., et al. (2026). Catalogue of Life (2026-05-15 XR). Catalogue of Life Foundation, Amsterdam, Netherlands. https://doi.org/10.48580/dgxsq

Barbera, Pierre, Alexey M. Kozlov, Lucas Czech, m.fl. ”EPA-Ng: Massively Parallel Evolutionary Placement of Genetic Sequences”. Systematic Biology 68, nr 2 (2019): 365–69. https://doi.org/10.1093/sysbio/syy054.

Cardoso, Pedro, François Rigal, Paulo A. V. Borges, och José C. Carvalho. ”A New Frontier in Biodiversity Inventory: A Proposal for Estimators of Phylogenetic and Functional Diversity”. Methods in Ecology and Evolution 5, nr 5 (2014): 452–61. https://doi.org/10.1111/2041-210X.12173.

Cavender-Bares, Jeannine, Kenneth H. Kozak, Paul V. A. Fine, och Steven W. Kembel. ”The Merging of Community Ecology and Phylogenetic Biology”. Ecology Letters 12, nr 7 (2009): 693–715. https://doi.org/10.1111/j.1461-0248.2009.01314.x.

Chazot, Nicolas, Fabien L. Condamine, Gytis Dudas, m.fl. ”Conserved Ancestral Tropical Niche but Different Continental Histories Explain the Latitudinal Diversity Gradient in Brush-Footed Butterflies”. Nature Communications 12, nr 1 (2021): 5717. https://doi.org/10.1038/s41467-021-25906-8.

Chazot, Nicolas, Niklas Wahlberg, André Victor Lucci Freitas, m.fl. ”Priors and Posteriors in Bayesian Timing of Divergence Analyses: The Age of Butterflies Revisited”. Systematic Biology 68, nr 5 (2019): 797–813. https://doi.org/10.1093/sysbio/syz002.

Czech, Lucas, Pierre Barbera, och Alexandros Stamatakis. ”Methods for Automatic Reference Trees and Multilevel Phylogenetic Placement”. Bioinformatics 35, nr 7 (2019): 1151–58. https://doi.org/10.1093/bioinformatics/bty767.

Czech, Lucas, Alexandros Stamatakis, Micah Dunthorn, och Pierre Barbera. ”Metagenomic Analysis Using Phylogenetic Placement—A Review of the First Decade”. Frontiers in Bioinformatics 2 (maj 2022): 871393. https://doi.org/10.3389/fbinf.2022.871393.

Faith, Daniel P. ”Conservation Evaluation and Phylogenetic Diversity”. Biological Conservation 61, nr 1 (1992): 1–10. https://doi.org/10.1016/0006-3207(92)91201-3.

Hajibabaei, Mehrdad, Gregory A. C. Singer, Paul D. N. Hebert, och Donal A. Hickey. ”DNA Barcoding: How It Complements Taxonomy, Molecular Phylogenetics and Population Genetics”. Trends in Genetics 23, nr 4 (2007): 167–72. https://doi.org/10.1016/j.tig.2007.02.001.

Hamilton, C. A., R. A. St Laurent, K. Dexter, m.fl. ”Phylogenomics Resolves Major Relationships and Reveals Significant Diversification Rate Shifts in the Evolution

of Silk Moths and Relatives". BMC Evolutionary Biology 19, nr 1 (2019): 182. https://doi.org/10.1186/s12862-019-1505-1.

Katoh, K., och D. M. Standley. "MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability". Molecular Biology and Evolution 30, nr 4 (2013): 772–80. https://doi.org/10.1093/molbev/mst010.

Kawahara, Akito Y., David Plotkin, Marianne Espeland, m.fl. "Phylogenomics Reveals the Evolutionary Timing and Pattern of Butterflies and Moths". Proceedings of the National Academy of Sciences 116, nr 45 (2019): 22657–63. https://doi.org/10.1073/pnas.1907847116.

Minh, Bui Quang, Heiko A. Schmidt, Olga Chernomor, m.fl. "IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era". Molecular Biology and Evolution 37, nr 5 (2020): 1530–34. https://doi.org/10.1093/molbev/msaa015.

Genbank, National Center for Biotechnology and Information (NCBI). GenBank. https://www.ncbi.nlm.nih.gov/genbank/ [Cited: 2026-06-03]

Ratnasingham S, Wei C, Chan D, Agda J, Agda J, Ballesteros-Mejia L, Ait Boutou H, El Bastami Z M, Ma E, Manjunath R, Rea D, Ho C, Telfer A, McKeowan J, Rahulan M, Steinke C, Dorsheimer J, Milton M, Hebert PDN (2024). BOLD v4: A Centralized Bioinformatics Platform for DNA-Based Biodiversity Data. In DNA Barcoding: Methods and Protocols, pp. 403–441. Chapter 26. New York, NY: Springer US, 2024.

Ricklefs, Robert E. "Estimating Diversification Rates from Phylogenetic Information". Trends in Ecology & Evolution 22, nr 11 (2007): 601–10. https://doi.org/10.1016/j.tree.2007.06.013.

Smith, M. Alex, Jose Fernandez-Triana, Rob Roughley, och Paul D. N. Hebert. "DNA Barcode Accumulation Curves for Understudied Taxa and Areas". Molecular Ecology Resources 9, nr s1 (2009): 208–16. https://doi.org/10.1111/j.1755-0998.2009.02646.x.

Webb, Campbell O., David D. Ackerly, Mark A. McPeek, och Michael J. Donoghue. "Phylogenies and Community Ecology". Annual Review of Ecology and Systematics 33, nr 1 (2002): 475–505. https://doi.org/10.1146/annurev.ecolsys.33.010802.150448.

Acknowledgements

I would like to thank my supervisor Nicolas Chazot for all the help I have received from you during this project and for all the things you have taught me along the way. I would also like to thank my assistant supervisor Johanna Orsholm for all the help you have given me, and for being the developer of the online tree inference. Without you guys this would have never been possible. Thank you!

,

Publishing and archiving

Approved students' theses at SLU can be published online. As a student you own the copyright to your work and in such cases, you need to approve the publication. In connection with your approval of publication, SLU will process your personal data (name) to make the work searchable on the internet. You can revoke your consent at any time by contacting the library.

Even if you choose not to publish the work or if you revoke your approval, the thesis will be archived digitally according to archive legislation.

You will find links to SLU's publication agreement and SLU's processing of personal data and your rights on this page:

- https://libanswers.slu.se/en/faq/228318

☒ YES, I, Josefin Westberg Larsson, have read and agree to the agreement for publication and the personal data processing that takes place in connection with this