2013 Rubenstein Research Fellows

Papilionoidea of the World: Evaluation and validation of EOL, BHL and GBIF data for Papilionidae, Pieridae and Riodinidae

JR Ferrer-Paris and AY Sánchez-Mercado
Centro de Estudios Botánicos y Agroforestales
Instituto Venezolano de Investigaciones Científicas

Report EOLR.r.2013.08
available at the PoW home page
Version of 24 de agosto de 2013

Abstract

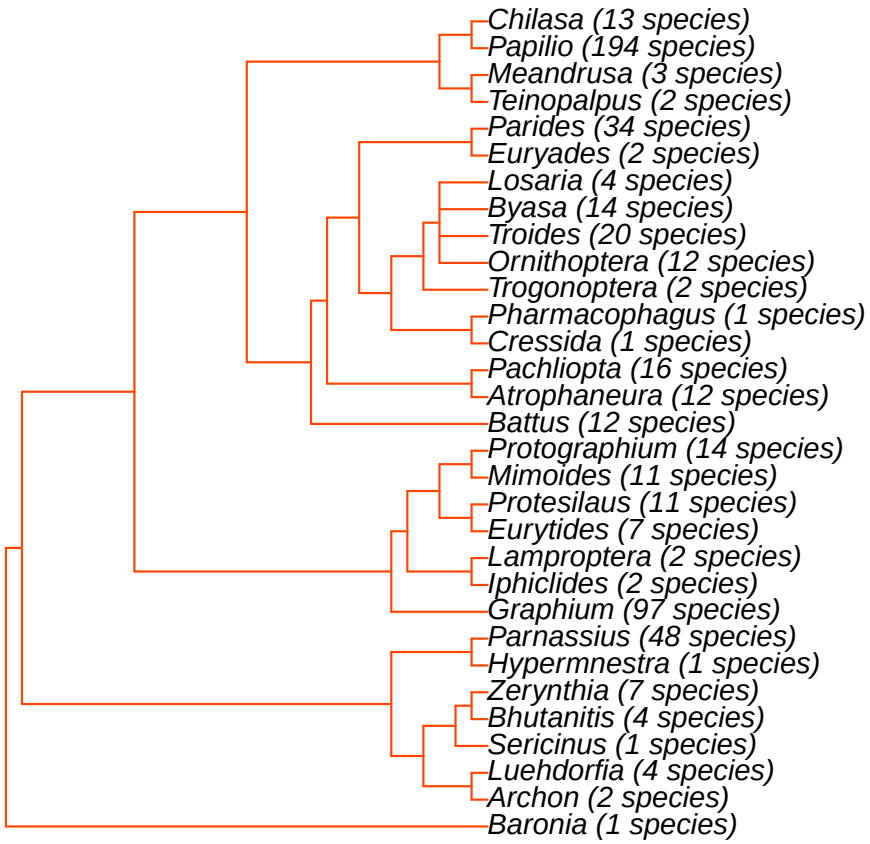
We evaluate the representativeness of three open sources of data for three families (Papilionidae, Pieridae and Riodinidae) that represent almost 20 % of the known species of butterflies (Papilionoidea). First we built taxonomic checklists and ordered the species lists according to the most probable phylogeny of each family. Checklists are based on the most updated and completed synonimic list and catalogues available in public sources, and phylogenies are based on the most recent studies for two families, and an approximated phylogeny for Riodinidae. For each species we retrieved all available text data objects from the *Encyclopedia of Life*, **EOL**; all pages from the *Biodiversity Heritage Library*, **BHL**; and distribution records from the *Global Biodiversity Information Facility*, **GBIF**. We then analyse the distribution of data objects, pages and records per species and the representativeness of each data source accross the phylogeny of each family.

We found that in general Papilionidae and Pieridae were better represented in all sources, but EOL had a more complete coverage and better representation of their genera. However, Riodinidae appears to be comparatively well represented in BHL, and this could be a source for improving information content in EOL.

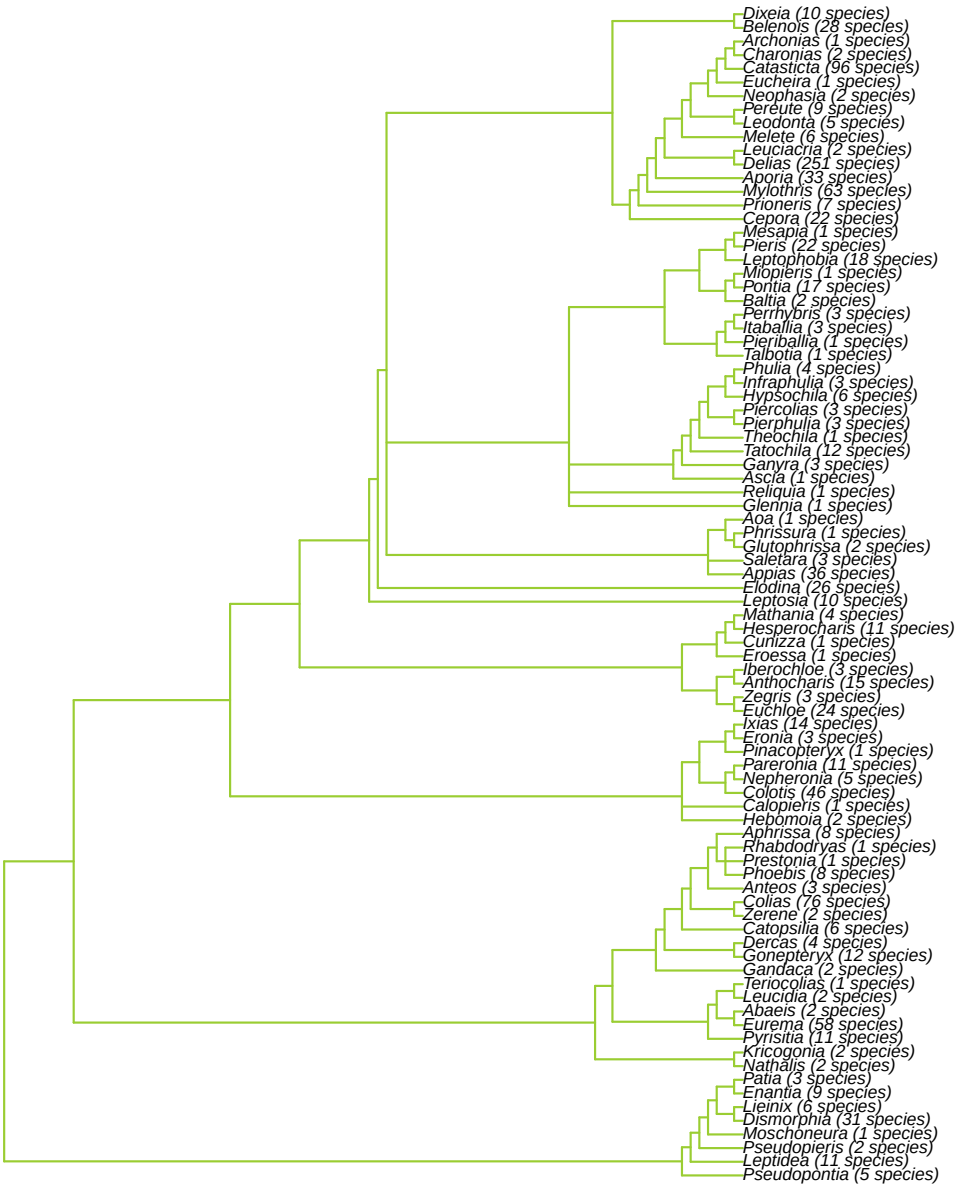
EOL contributors were usually complementary, with different regional and taxonomical focus, and their ranking was different for each family. BHL contributed several pages, but extracting information is a lengthty proccess and records tend to be very redundant. Both sources could provide complementary records for extending the current compilation of host-plant associations.

1. Checklist and phylogeny

First we built taxonomic checklists and ordered the species lists according to the most probable phylogeny of the Papilionidae, Pieridae and Riodinidae. These three families represent almost 20 % of the known species of Papilionoidea[6]. Details for this section are available at the PoW homepage under *A working checklist of butterfly species*.



Papilionidae is the most basal family in the Papilionoidea clade[8], and according to the most up-to-date checklist from the Gart/GloBIS project[7], there are 554 recognized species. We use a simplified phylogeny based on the recent work by Condamine et al.[5], which includes 31 genera. Papilio and Graphium are the most species rich genera within the Papilionidae.



Pieridae ist one of the best known families of Papilionoidea, it has an intermediate placement between the basal groups and the most derived families[8]. They include 1138 species. We drew a simplified phylogeny, which includes 86 genera, and is based on the tree of life webpage[4], which itself is based mostly on work by Braby and Trueman[1], Pieridae has five genera with more than 50 species: *Eurema*, *Mylothris*, *Colias*, *Catasticta* and *Delias*.

Riodinidae is unique among the Papilionoidea clade because it combines a high species richness with a restricted distribution, with up to 1410 species but more than 92 % of them restricted to the Neotropical region[6]. The Riodinidae is a sister clade of the Lycaenidae, and they are often considered the most derived groups of Papilionoidea[8]. The phylogeny of the family is not fully resolved, and a very basic cladogram was sketched from the information available at the Tree of Life project[3].

We use three metrics derived from phylogenetic community analysis[9;11] to measure the phylogenetic representativeness of data from each source.

Phylogenetic species richness, (PSR), is related to the number of taxa in a sample (SR), but accounting for the decrease of variance due to phylogenetic relatedness. *Phylogenetic species evenness*, (PSE), is a measure of phylogenetic variability that incorporate the effect of relative species abundance (here the effect of the number of data objects, pages or records). Higher values represent more similar abundances for all taxa, but the maximal value of one is only possible when the species considered are complete unrelated (star phylogeny). The *Mean Pairwise Distance*, (MPD), is the phylogenetic difference between two randomly taken individuals (here data objects, pages or records) from a sample.

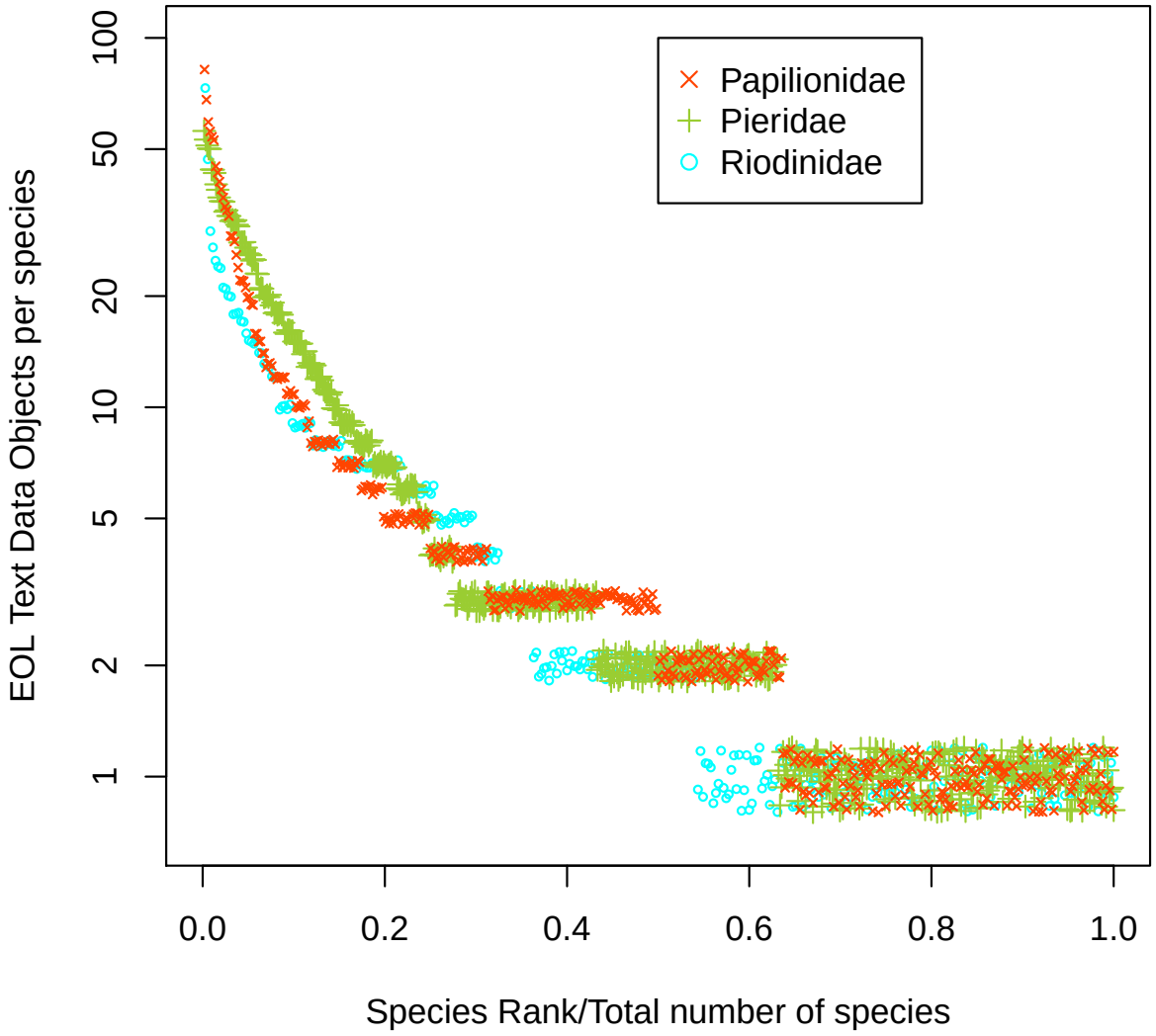
2. Encyclopedia of Life, EOL

We used the EOL API to retrieve information from The Encyclopedia of Life each species in our checklist. Details about the protocol used are available in the PoW home page under *EOL data search*. In this version we are assuming that the corresponding services are handling synonymies correctly, and thus we did not retrieve information for alternative names that might be present in some data sources. This would probably be desirable in the future.

In fact many names in our checklists returned matches for several taxon concepts within EOL hierarchies, but some names returned no match in EOL, or they returned a name match, but no text object.

The percentage of names matched was 100 % for Papilionidae, 91.7 % for Pieridae, and 69.9 % for Riodinidae, but the percentage of species with one or more text data objects was 88.1 % for Papilionidae, 53.5 % for Pieridae and 25.2 % for Riodinidae.

However, do they differ in the amount of information available per species?, It is usual to find a log-normal distribution of information accross species, with few abundant, common or biologically interesting species having much more available information, and a large number of poorly represented species. Thus we used a log-count ranked plot to compare the distribution of the number of EOL text data objects among species in each family. We further divided the ranks by the number of species with data to make the distributions comparables between families. Data points were slightly jittered to allow the visualization of several overlapping cases.

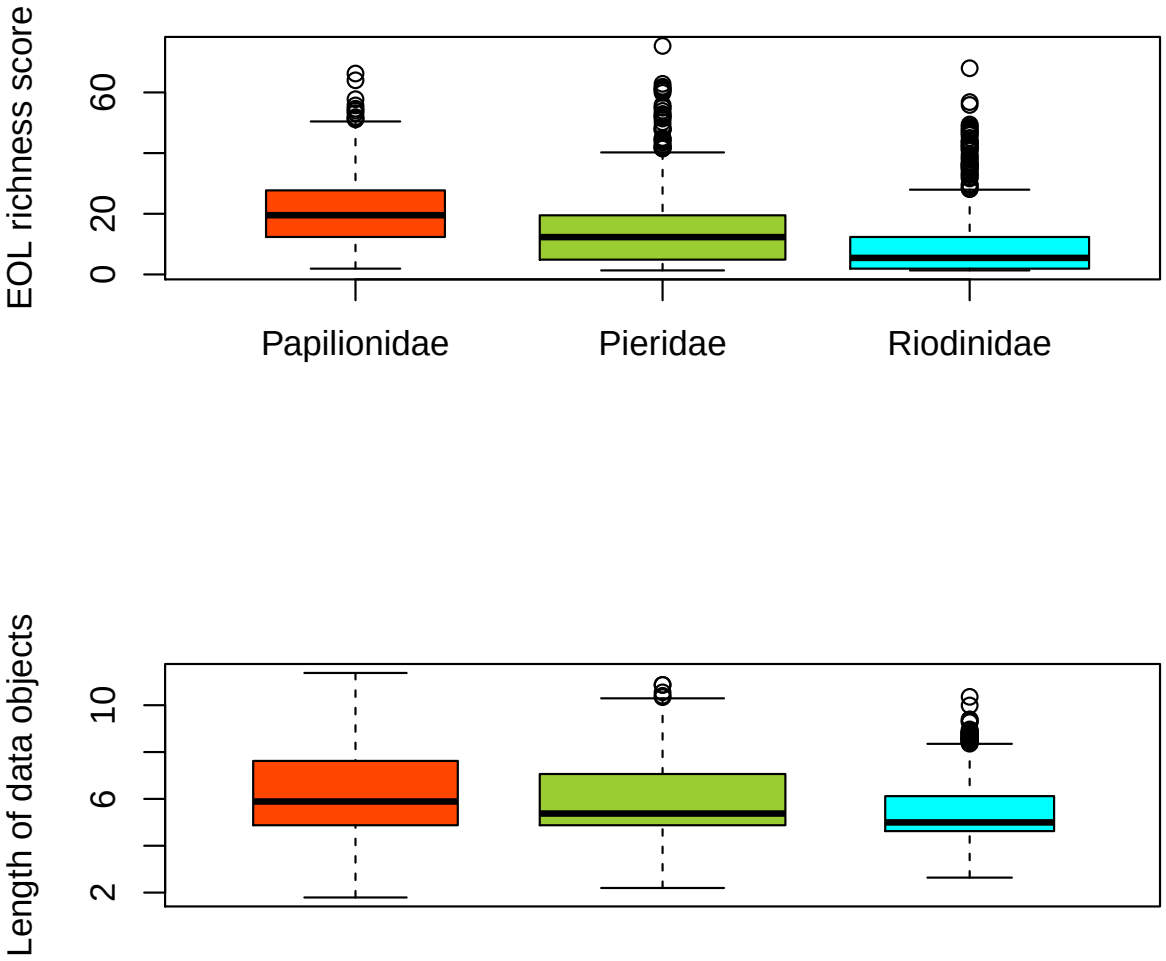


In fact few species in each family have more than 50 text data objects, and a large proportion, around 40 % of the species have only one data object. Among the first 20 % of the species, the Pieridae have more data objects than both the Papilionidae and the Riodinidae. For Papilionidae, only the first 10 % have a larger number of data objects than the Riodinidae, but after that, they have similar distribution.

2.1. Quality of available data

We further use two indirect measures of the quality of the data. EOL provides "richness scores" for the content associated with one species, it measures the diversity of content as the variety of sources, quantity of data, among other things. We also measure the size (log of

text length) of the data objects as a proxy for quality. Both metrics show that the content is not equal for all families: content about Papilionidae is slightly better, followed by Pieridae, and Riodinidae coming last.



2.2. Languages

Most data objects in EOL are in English, and some additional ones are in Spanish, with modest contributions in other languages. Spanish seems to be an important language for Riodinidae, probably because most species are distributed in Spanish speaking countries in Latinamerica. This emphasizes the importance of local knowledge to improve coverage of certain taxonomic groups.

```
> table(EOLo.pird$lan)

  de   en   es   fr
   5 2947  220   12

> table(EOLo.ppln$lan)

  en   es   fr
2069  130    8

> table(EOLo.riod$lan)

  en   es
1149  355
```

2.3. Agents

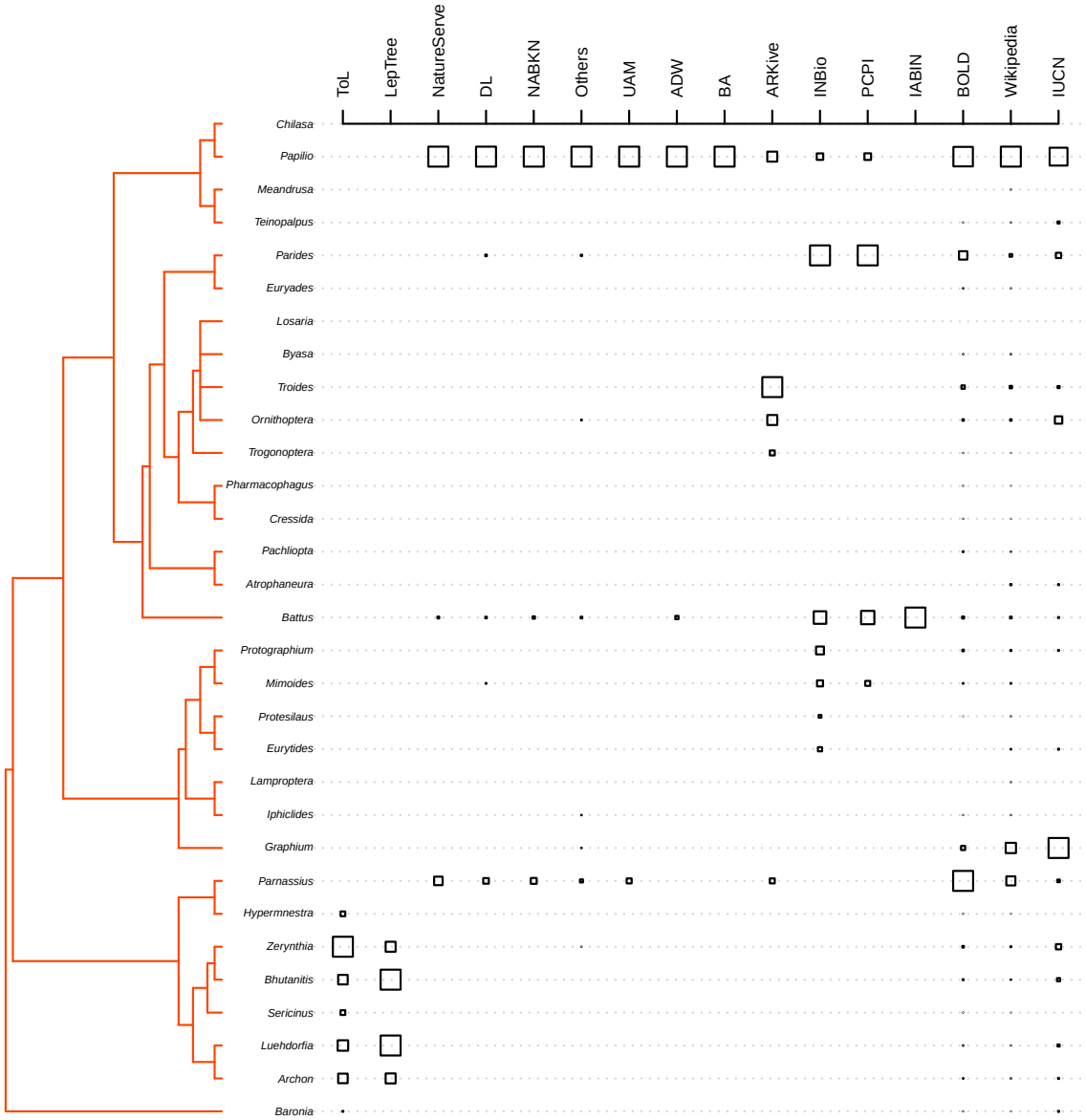
We also evaluated which sources are contributing more information to EOL for these three families. BOLD (Barcode of life) and NatureServe are the two most important sources, followed by Wikipedia. Together with INBio and IUCN, they can be considered to be the core contributors of text objects. The INBio is specially important for Riodinidae, and the main source for text objects in spanish. *Bibliotheca Alexandrina* (BA) and the *University of Alberta Museums* (UAM) provide an important amount of pages for Pieridae.

```
             IABIN LepTree PCPI ARKive DL ADW NABKN ToL UAM  BA Others IUCN
Papilionidae     1       6   27     70 37  99    48  63  70  12    102  141
Pieridae        19      12   36     49 60  37    79  38 126 186     86  180
Riodinidae       0      15   53      0 26   0    29  61   0   0     64   46
             INBio Wikipedia NatureServe BOLD
Papilionidae   129       424         432  550
Pieridae       201       484         797  794
Riodinidae     355        82         367  407
```

We map the contribution of the different sources (number of text objects) onto the phylogeny of each group.

2.3.1. Papilionidae

Some specialized sources overlap in the coverage of some genera, but differ from other specialized sources. Two sources (ToL and LepTree) are restricted to a single clade of Papilionidae, while a number of others, including NatureServe focus mostly on Papilio and Parnassius. INBio and PCPI focus on (mostly) neotropical taxa like Parides and Battus. BOLD, Wikipedia and IUCN have similar coverage, with strong representation of Papilio, but including some information on almost all genera, which is probably proportional to their species richness.



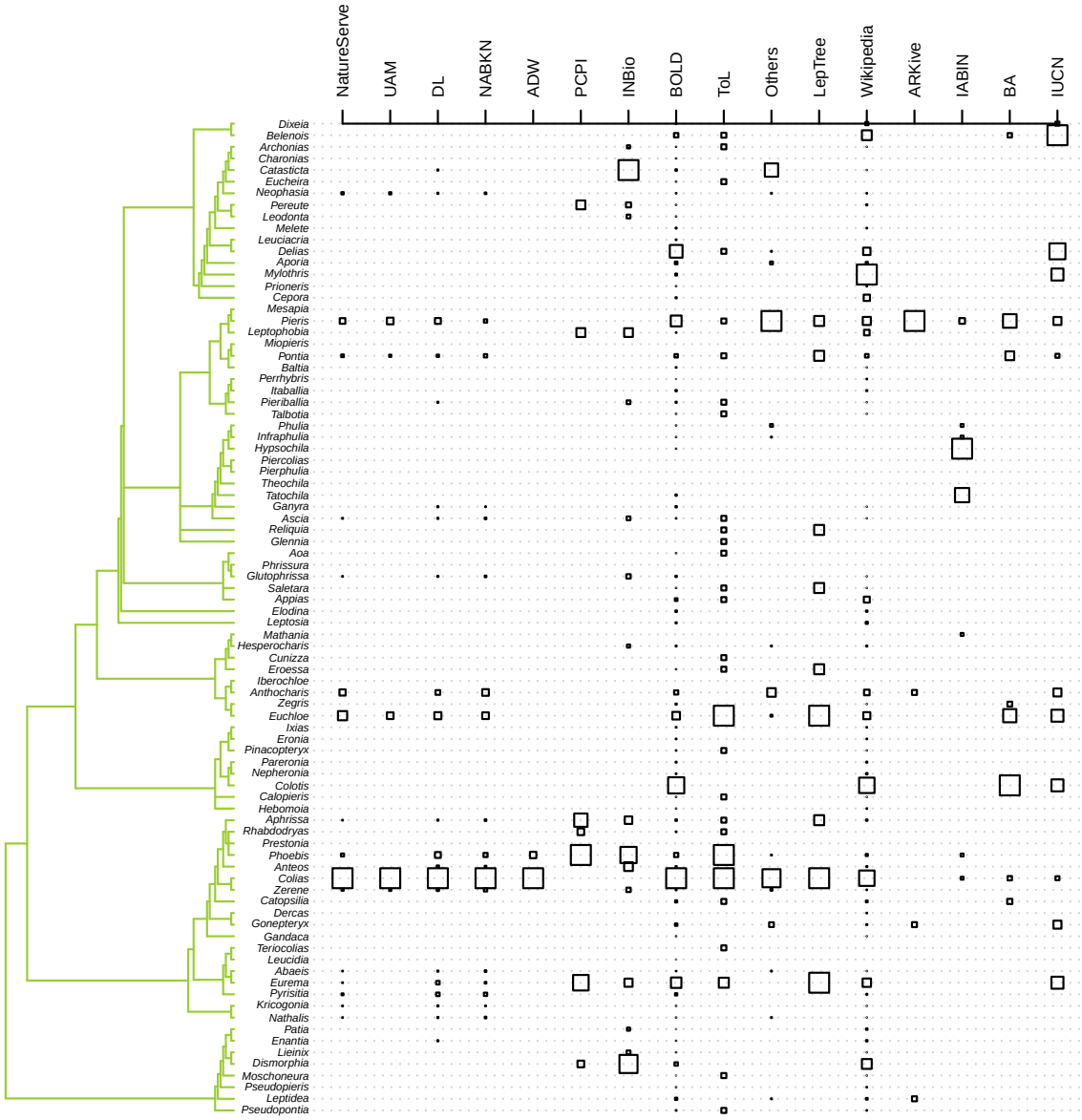
For Papilionidae the Wikipedia, BOLD and IUCN have a better representation of all phylogenetic groups, with high number of genera, and the highest values of phylogenetic richness, eveness and MPD. Only Natureserve achieves a similar level of PSE with a smaller number of included genera.

	Row.names	SR	PSR	vars	MPD	PSEs
16	Wikipedia	30	20.551724	0.05629527	1.1934652	0.61730959
4	BOLD	26	18.064000	0.24750553	1.2412774	0.64546422
8	IUCN	16	11.844444	0.48869352	1.2355381	0.65895366
12	Others	8	5.819048	0.44423507	0.7016019	0.40091540
14	ToL	7	2.666667	0.42545609	0.2426808	0.14156379
7	INBio	7	3.544444	0.42545609	0.8994331	0.52466932

```
5           DL  5  3.716667 0.38298279 0.8056489 0.50353056
2       ARKive  5  2.866667 0.38298279 0.6920000 0.43250000
13        PCPI  4  2.377778 0.36283315 0.7023320 0.46822131
9      LepTree  4  0.400000 0.36283315 0.1481481 0.09876543
11 NatureServe  3  2.433333 0.35167909 0.8590885 0.64431638
10       NABKN  3  2.433333 0.35167909 0.7557870 0.56684028
15         UAM  2  1.933333 0.38673962 0.6186667 0.61866667
1          ADW  2  1.000000 0.38673962 0.2428324 0.24283236
6        IABIN  1        NA         NA        NA         NA
3           BA  1        NA         NA        NA         NA
```

2.3.2. Pieridae

Overlap and differences in coverage of Pieridae seems to reflects different geographical focus from each source. NatureServe, UAM, DL and NABKN focus mostly on Nearctic genera like Colias, Euchloe and Pieris. PCPI and INBio focus on neotropical genera like Phoebis, Aphrissa, Dismorphia, Pereute and Catasticta. While IUCN covers mostly Afrotropical or Oriental groups like Belenois, Delias and Mylothris.



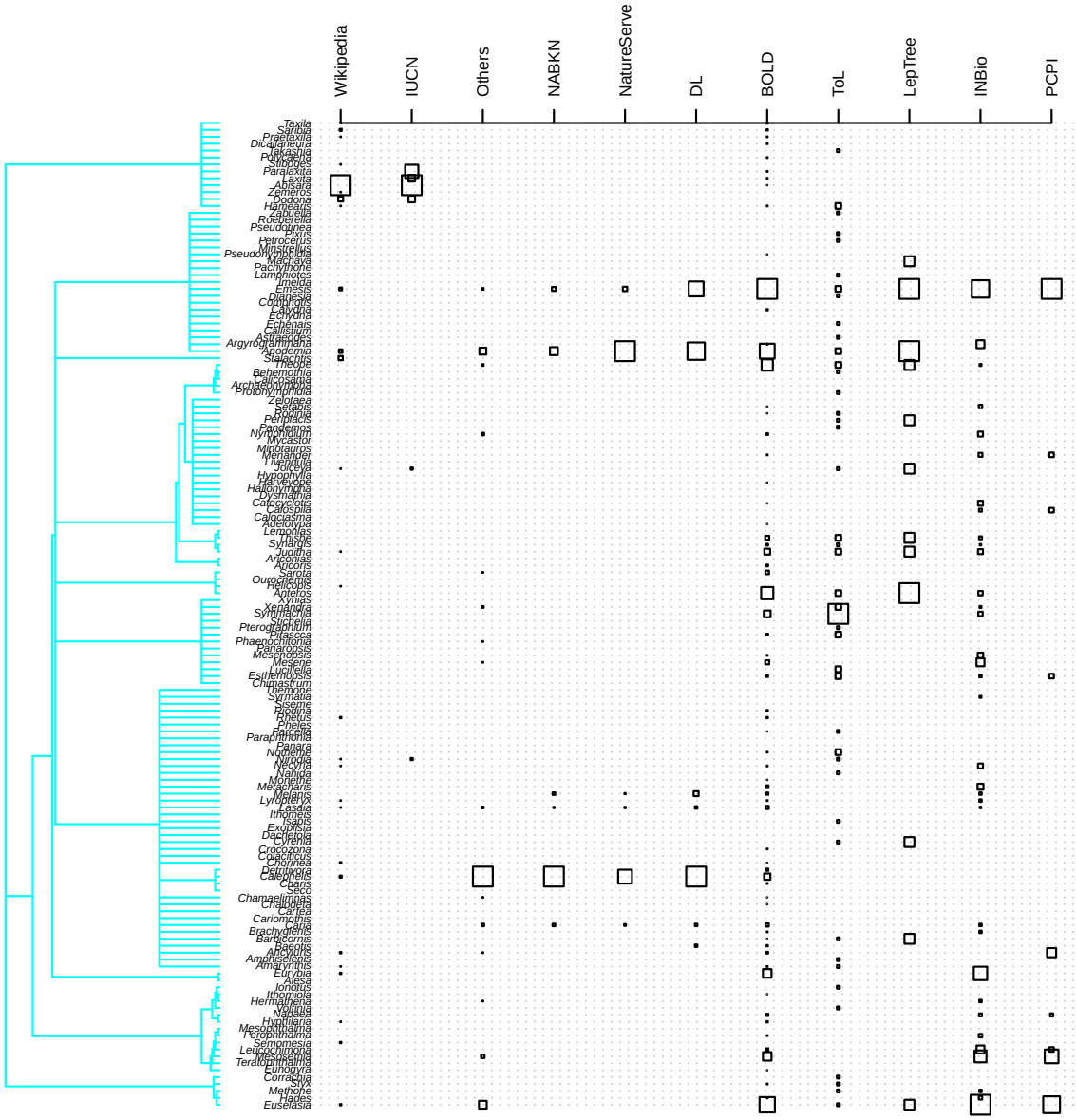
For Pieridae the Wikipedia, BOLD and ToL appear to have a better representation of all phylogenetic groups, and this is confirmed by the measures of phylogenetic richness, although phylogenetic eveness is higher for the contents of INBio and Arkive. All four of them have high values of Mean Phylogenetic Distance between represented taxa, while other sources like NABKN and Natureserve have a much lower MPD in spite of including a similar number of genera as INBio.

	Row.names	SR	PSR	vars	MPD	PSEs
4	BOLD	72	49.6347970	0.5843762	1.32970146	0.67421483
16	Wikipedia	61	43.4113725	0.8901062	1.32395147	0.67300867
14	ToL	28	19.1459695	0.9983407	1.31054261	0.67954061
5	DL	21	13.6776471	0.8661072	1.07558170	0.56468039

```
10       NABKN 18 10.7847751 0.7931116 1.04094178 0.55108682
12      Others 17 11.5264706 0.7666420 1.16562848 0.61924013
7        INBio 17 12.3441176 0.7666420 1.35440445 0.71952736
11 NatureServe 16  9.6501961 0.7391195 1.09583860 0.58444725
8          IUCN 12  7.7219251 0.6188461 1.12062745 0.61125134
9      LepTree  9  5.8970588 0.5189798 1.17352941 0.66011029
6        IABIN  8  4.3294118 0.4842998 0.56040411 0.32023092
3           BA  8  5.4453782 0.4842998 1.10714465 0.63265409
13       PCPI  7  4.7843137 0.4492748 0.96699346 0.56407952
15        UAM  6  4.0329412 0.4143679 1.03311547 0.61986928
2       ARKive  4  3.6078431 0.3498464 1.05546218 0.70364146
1          ADW  2  0.1411765 0.3499939 0.05197439 0.05197439
```

2.3.3. Riodinidae

Coverage of Riodinidae is very patchy for most sources. Wikipedia and IUCN focus mostly on Abisara and Dodona and have few or no content for the other genera. NatureServe, NABKN, DL and other sources have a good coverage of Calephelis and Apodemia, and four other genera. ToL has a focus on Symmachia, and few text objects in several genera, while BOLD, LepTree and INBio focus on Emesis, Euselasia and Apodemia, but do provide several data objects for other genera.



For Riodinidae BOLD appears to be the best source of data, with high number of genera and high values of PSR, MPD and PSE. INBio provide data for only half of the genera that BOLD includes, and thus have around half the PSR, but the values of MPD and PSE are similar, indicating that it is a representative sample of the phylogeny of the group. Similarly PCPI seems to be a good source, providing balanced amounts of information (high PSE) for few, but representative taxa (low PSR but high MPD).

	Row.names	SR	PSR	vars	MPD	PSEs
1	BOLD	67	50.720230	0.5588292	1.5019117	0.7623340
10	ToL	43	31.908786	0.5046041	1.4067095	0.7201013
3	INBio	34	25.082373	0.4560316	1.5070140	0.7763405
11	Wikipedia	27	21.078548	0.4078392	1.3269013	0.6889680

```
8       Others 16 11.936150 0.3149316 1.2959191 0.6911568
5      LepTree 12  8.143406 0.2771046 1.2441941 0.6786513
9         PCPI  9  7.082746 0.2486930 1.4427770 0.8115621
2           DL  7  3.544601 0.2311493 0.8842820 0.5158312
7  NatureServe  6  3.188732 0.2236469 0.8229701 0.4937820
6        NABKN  6  3.188732 0.2236469 0.7804257 0.4682554
4         IUCN  6  3.709859 0.2236469 0.4211507 0.2526904
```

2.4. Extracting hostplants records from EOL data objects

Some details of the data extraction protocol for EOL data objects are available in the PoW home page under *EOL data validation*.

We found that only a small proportion of text data objects are dedicated exclusively butterfly hostplants associations. In fact only 3.6 % of the data objects for Papilionidae refer to *Trophic strategy*, *Hostplant*, *Associations* or *Foodplant* in their title, similarly only 3 % for Pieridae and just 1.1 % for Riodinidae.

After searching the complete body of the data object for a list of different keywords associated with hostplant records, we found that those percentages increased:21.4 % for Papilionidae,18.1 % for Pieridae and 14.3 % for Riodinidae.

Some keywords like *Egg* or *Plant* were matched frequently for all families. The list of keywords for Papilionidae:

```
Attracted to          Egg     Fabaceae      Feeding      Feed on     Feeds on
          12          126            6           41          163           28
   Foodplant    Hospedera         Host   Host Plant       Larvae       Larval
          65           15          175           88          204           78
      Ovipos        Plant       Planta
          10          276           26
```

For Pieridae there were some additional matches when searching for the scientific or common name of common plant families like *legume* or *grasses*:

```
     Aliment Attracted to          Egg     Fabaceae      Feeding      Feed on
           5          11          107           27           20          220
    Feeds on    Foodplant      grasses    Hospedera         Host   Host Plant
          15          58            3           30          177           94
      Larvae       Larval       Legume       Ovipos        Plant       Planta
         262          71           27            9          302           36
     Poaceae
           1
```

and for Riodinidae more matches were found with spanish keywords like *Planta* and *Hospedera* :

```
   Aliment        Egg   Fabaceae    Feeding    Feed on   Feeds on  Foodplant
         4         20          1          3         13          2         32
```

```
   grasses  Hospedera       Host Host Plant     Larvae     Larval      Plant
         2         63         47         33         43         20        161
    Planta
        69
```

We performed a manual validation of all text data objects for the family Pieridae, and listed all data objects with hostplant associations records. We then compared the results of manual validation and simple keyword matching:

```
                keywordMatch
manualValidation FALSE TRUE
           FALSE  2593  152
           TRUE     14  425
```

The simple keyword match has a relatively high false positive rate of 0.263 but a very low false negative rate of 0.005. This means that the search for keywords is very effective for filtering out uninformative data objects, but additional steps are required for detecting false positives, but the amount of objects to be evaluated is manageable for manual inspection.

We found a total of 1319 hostplant records for 257 species of Pieridae in 425 data objects. These records refer to aproximatedly 742 plant taxa, including at least 146 species, but many records have not been fully validated and those numbers are ought to increase.

For comparison, our previous compilation[6], based on online databases and literature records, reported 4728 records for 443 species of the family Pieridae, and include 747 plant especies, 338 plant genera and 71 plant families. The search in EOL returned records for 28 butterfly species that were not represented in our compilation, and presumable several new plant species reports for the species already represented in it.

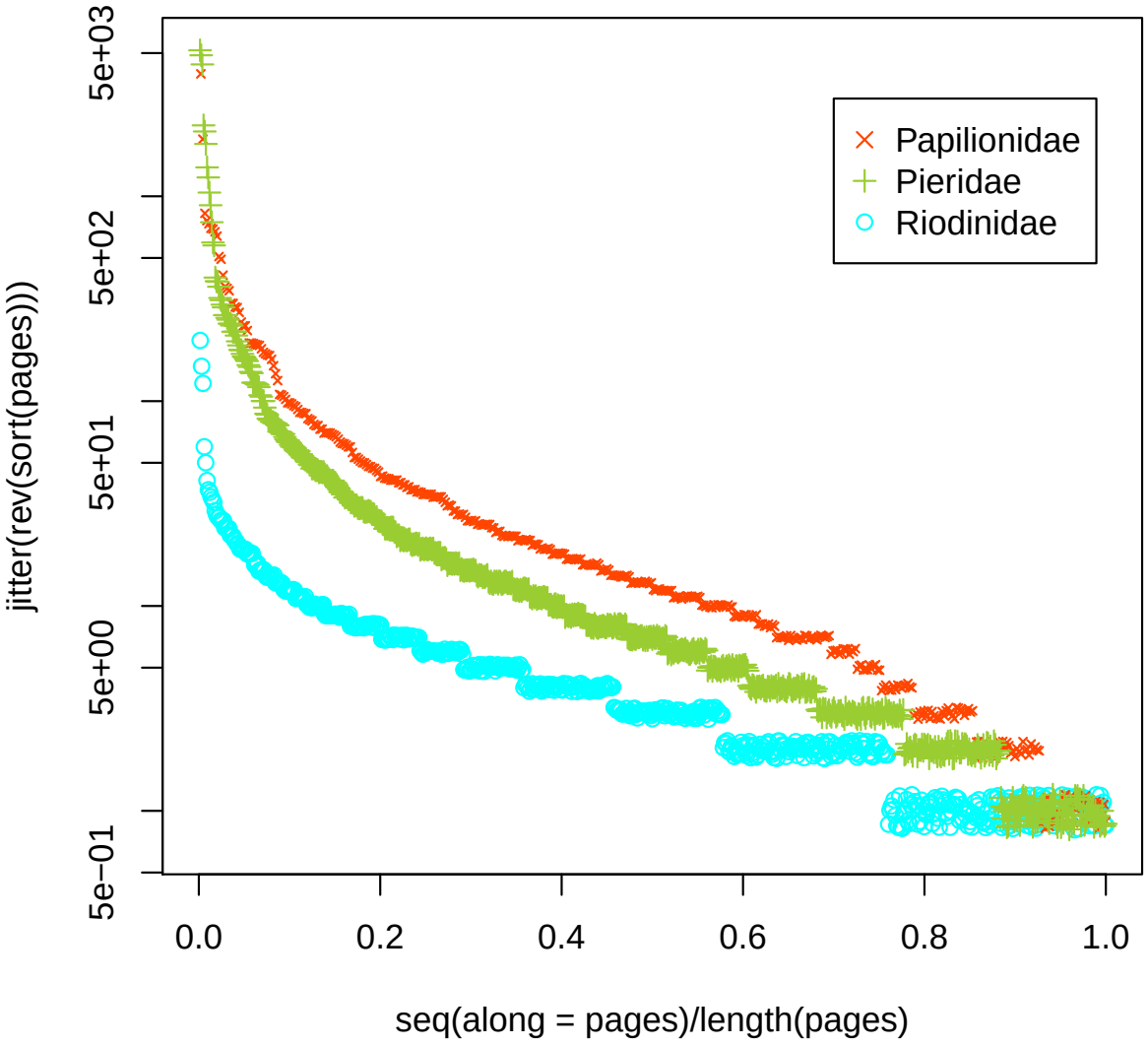
3. Biodiversity Heritage Library, BHL

We used the BHL API to retrieve information from the Biodiversity Heritage Library for each species in our checklist. Details about the search protocol used are available in the PoW home page under *BHL data search*. In this version we are assuming that the corresponding services are handling synonyms correctly, and thus we did not retrieve information for alternative names that might be present in some data sources. This would probably be desirable in the future.

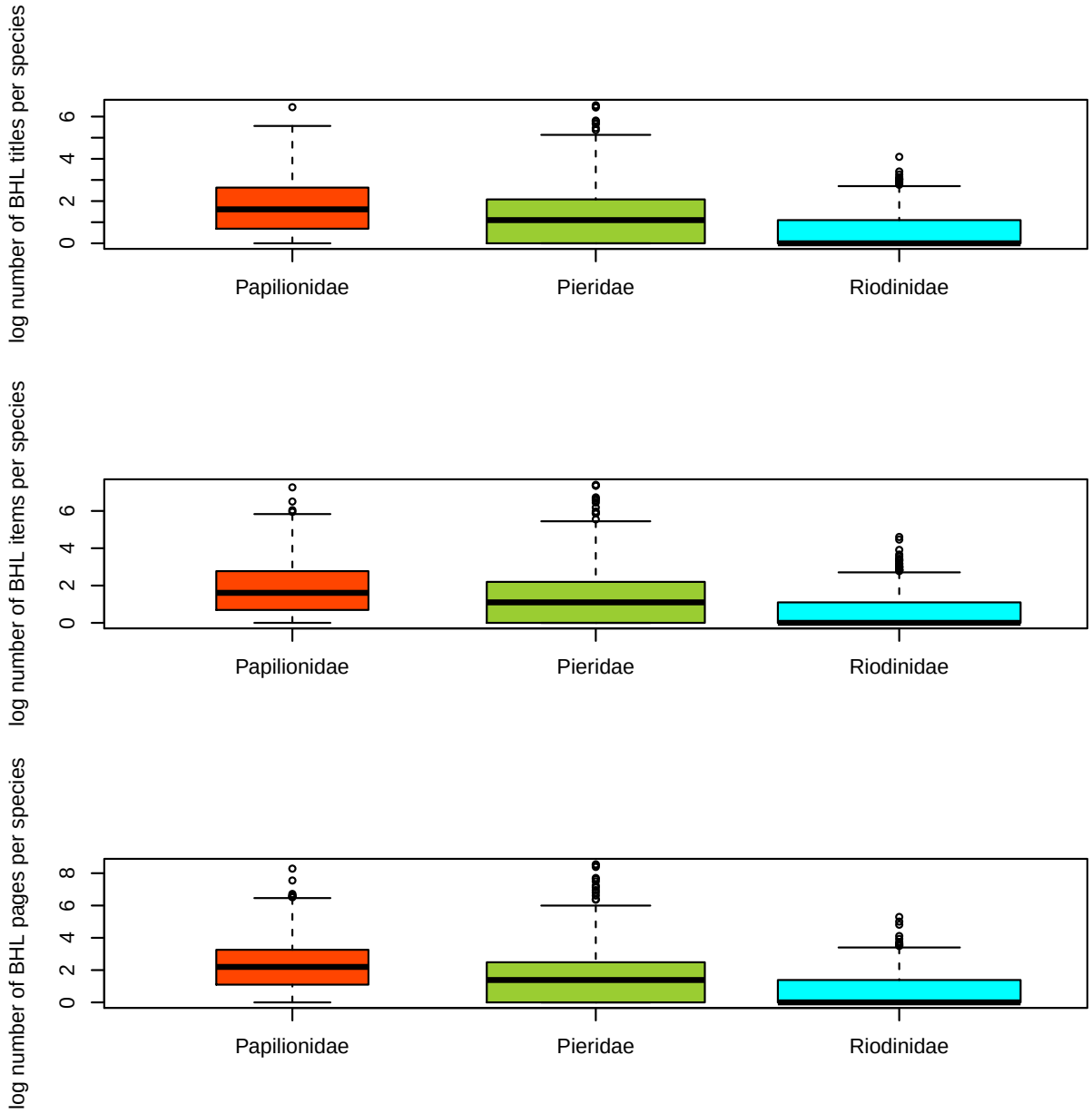
For 81 % of the species of Papilionidae 67.7 % of Pieridae, and 47.1 % of Riodinidae we found matches in BHL.

However the distribution of number of matched pages per species is strongly skewed, with very few species having more than 1000 matches and almost 300 species with 5 or fewer matches.

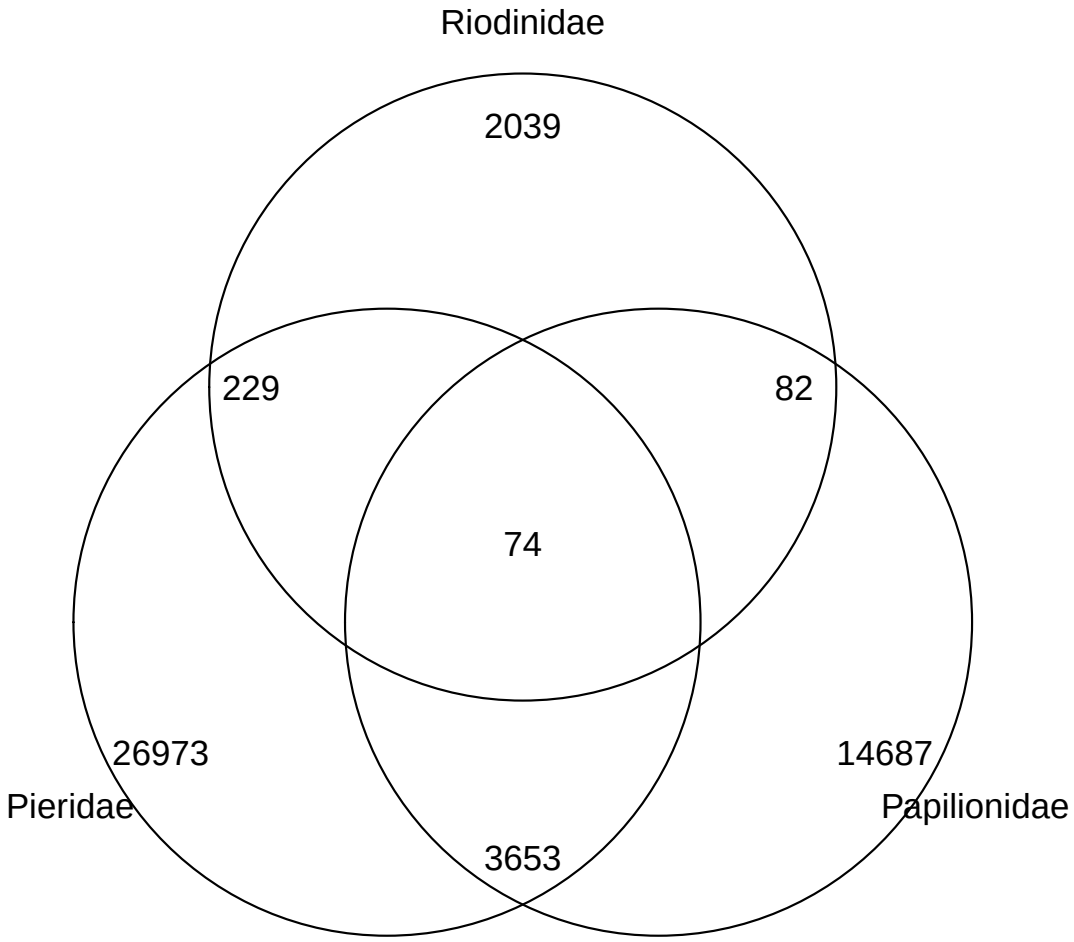
Few species among the top 5 % of both Papilionidae and Pieridae are found in 500 or more pages in BHL, but in general Papilionidae tend to have more matches than Pieridae. Riodinidae have fewer matches than any of the other two.



The pattern is similar for the number of titles, items and pages in BHL, with more matches for Papilionidae, then Pieridae and less for Riodinidae.



Taking all families together this represents a total of 47737 different pages. From these, around 13913 pages refer to two or more butterfly species from these families, but there is relatively little overlap between families. A good fraction of pages refers to Papilionidae and Pieridae, but few refer to species from all three families:



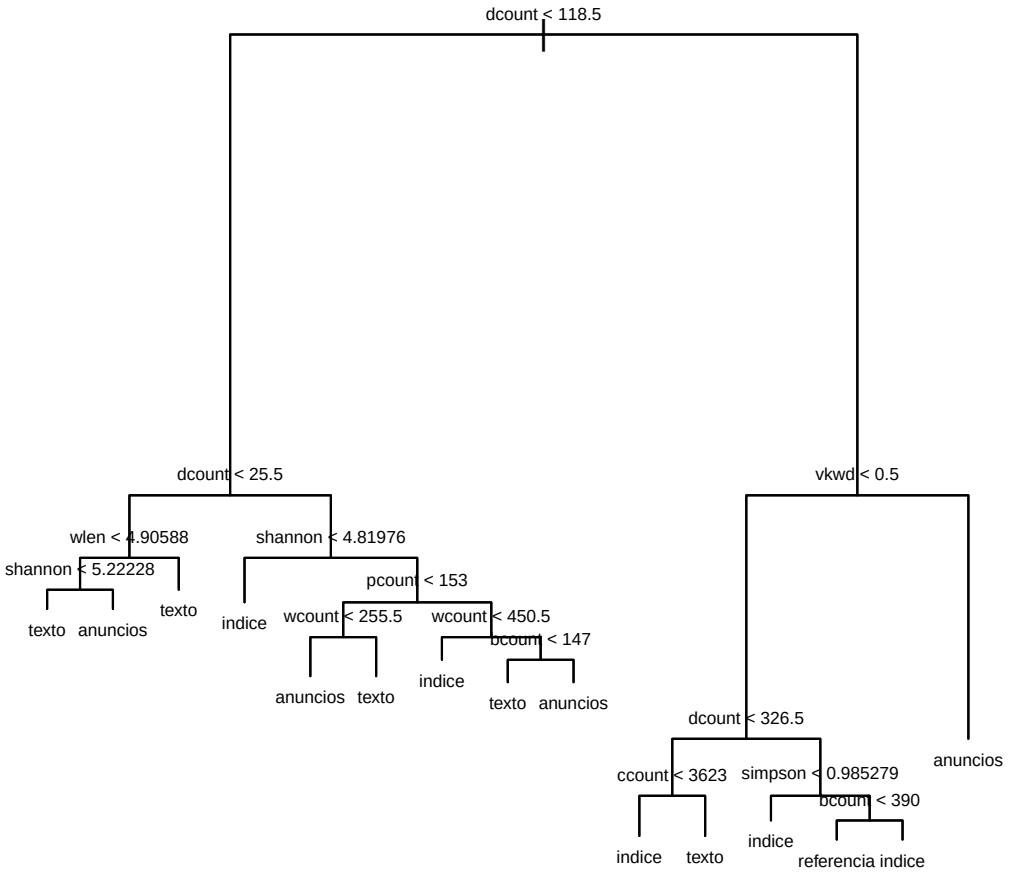
3.1. Classification of page types

A preliminary evaluation showed that not all pages were equally informative. Several matches refer to index pages, bibliography or comercial activities, such as specimen exchange and sales ("ads" pages), however these are not categorized within BHL. Thus we used our own manual evaluation of a sample of 794 pages to build a classification tree based on some text metrics like the number of digits, line breaks, alphabetic characters, number and diversity of words, etc.

The selected pages represented 412 text pages, 129 ads pages, 204 index pages, and 49 reference lists. We read the OCR of each page and calculate the text-metrics. Then we fit a

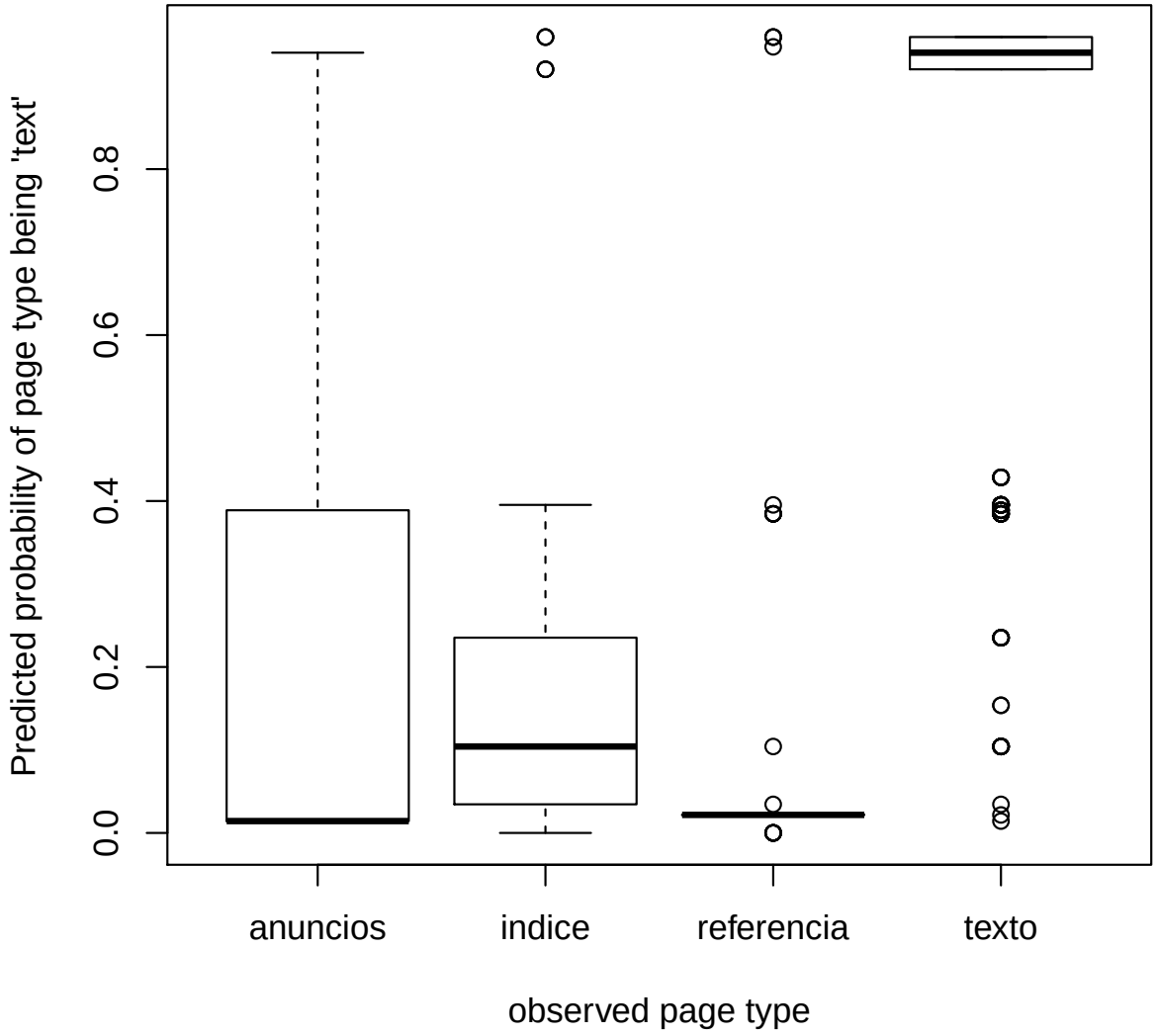
classification tree[2]:

```
Classification tree:
tree(formula = tipo ~ wcount + ucount + shannon + simpson + ccount +
    bcount + pcount + vkwd + wlen + wvar + dcount, data = pids)
Variables actually used in tree construction:
[1] "dcount"  "wlen"    "shannon" "pcount"  "wcount"  "bcount"  "vkwd"
[8] "ccount"  "simpson"
Number of terminal nodes:  15
Residual mean deviance:  0.9684 = 754.4 / 779
Misclassification error rate: 0.1763 = 140 / 794
```



The error rate is high, but it refers mostly to misclassification between references, ads and index pages, while most text objects are correctly classified:

observed	predicted anuncios	indice	referencia	texto
anuncios	93	18	0	18
indice	6	172	15	11
referencia	0	10	30	9
texto	16	36	1	359



We then applied this classification tree to all the OCR pages with matches for the three families, and we found that about half of them were predicted to be real text pages: 51.3 %

for Pieridae, 56 % for Pieridae and 52.2 % for Riodinidae. The application of this criterion greatly reduces the need for manual screaning of pages in search for useful content.

3.2. Extracting hostplants records from BHL pages

Some details of the data extraction protocol for BHL are available in the PoW home page under *BHL data validation*.

We searched the downloaded ocr pages for a list of different keywords associated with hostplant records, a great proportion of the matched pages did not refer to text pages but ads, indices and bibliography pages.

For Papilionidae only 22 % of the pages are classified as text pages with matches:

```
             is.text.page
keyword.match FALSE TRUE
        FALSE  3881 5161
        TRUE   4683 3862
```

For Pieridae up to 28.6 % of the pages are classified as text pages with matches:

```
             is.text.page
keyword.match FALSE TRUE
        FALSE  6131 8420
        TRUE   7356 8767
```

For Riodinidae 13.9 % of the pages are classified as text pages with matches:

```
             is.text.page
keyword.match FALSE TRUE
        FALSE   819  930
        TRUE    339  336
```

Up to 24 of the 26 keywords used were matched at least once. *Egg*, *Eier*, *Feeding*, *Larvae*, *Plant* and *Raupen* were the most common keywords, in general keywords in german were matched very often due to the high amount of BHL pages in that language.

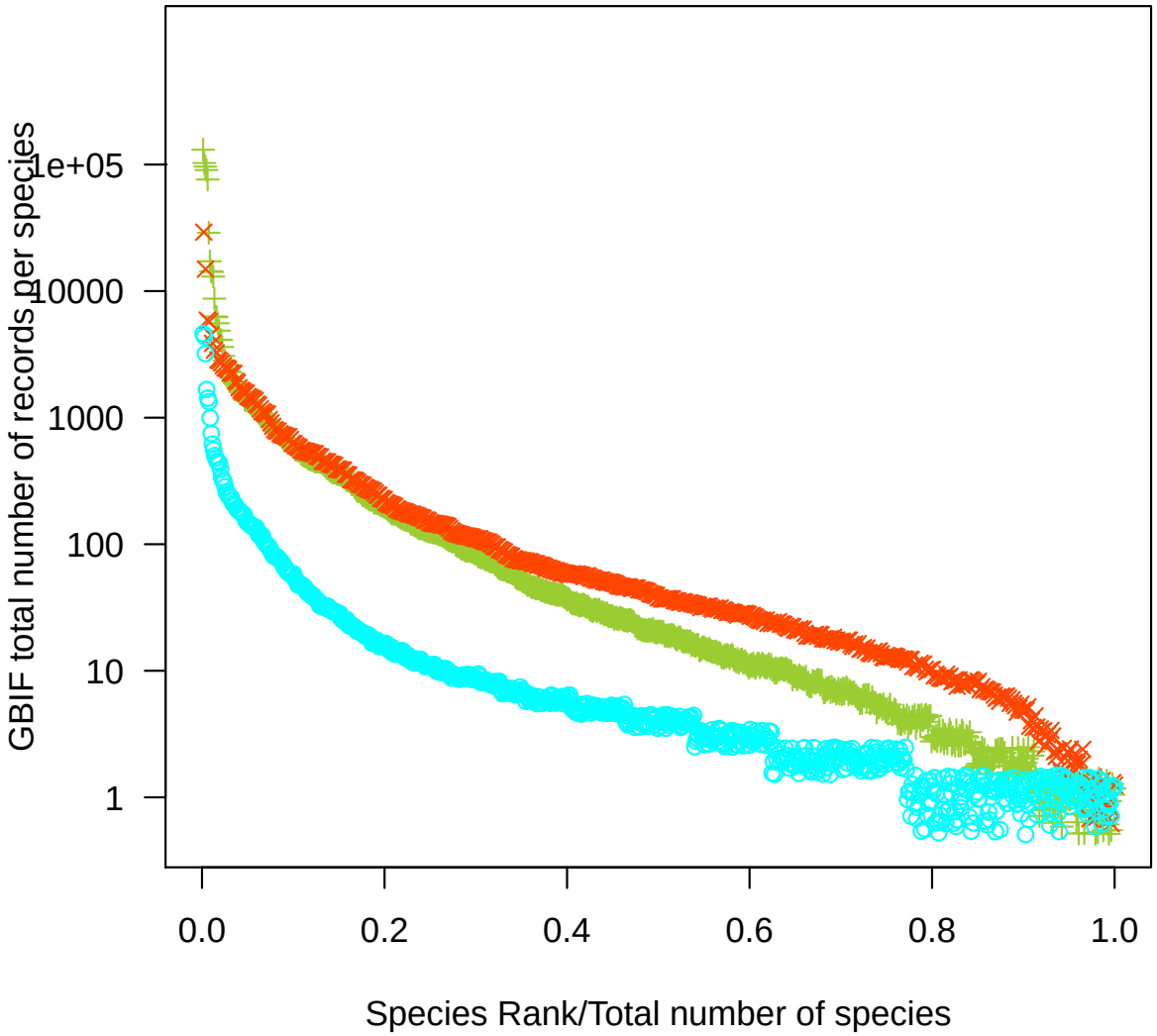
```
         keyword Papilionidae Pieridae Riodinidae
1        Aliment           35      117          2
2  Attracted to            73      139         10
3            Egg         1846     3783        151
4           Eier         1811     2479         65
5       Fabaceae            3       20          7
6        Feeding          959     2076        103
7        Feed on          287      334         23
8       Feeds on          365      341         37
9      Foodplant          483      701         75
10       grasses          109      263          9
```

11	Hospedera	0	2	0
12	Host	751	1334	79
13	Host Plant	179	170	18
14	Larvae	2281	5596	223
15	Larval	825	1469	95
16	Legume	9	100	2
17	Nahrung	165	401	8
18	Ovipos	535	813	63
19	Pflanze	686	1372	23
20	Plant	3428	6644	296
21	Planta	544	1054	39
22	Poaceae	4	6	1
23	Raupen	2025	2879	66
24	Recurso	18	22	0

4. Global Biodiversity Information Facility, GBIF

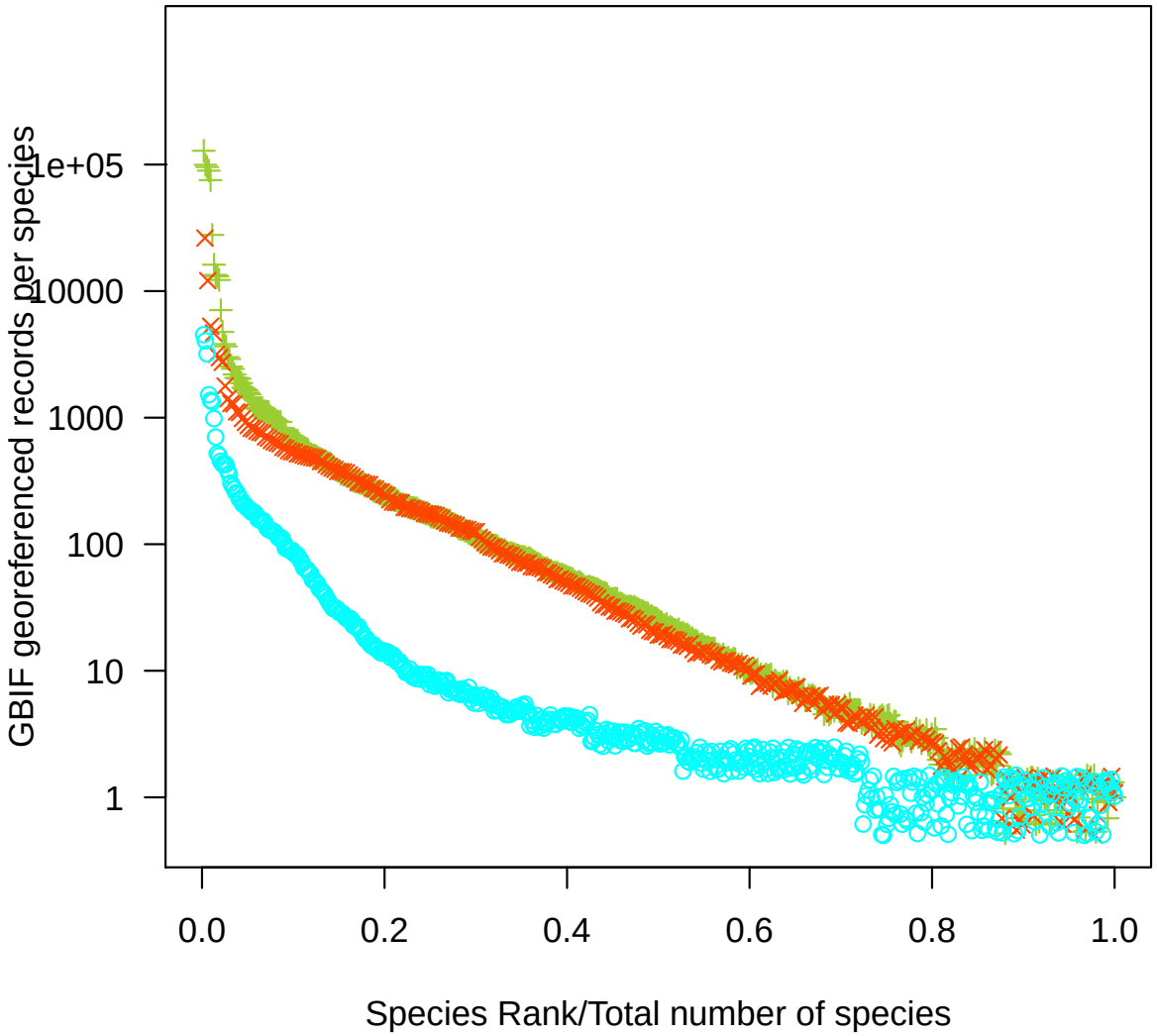
We used the `gbif` function in *R* package *dismo*[10] to retrieve information from the Global Biodiversity Information Facility for each species in our checklist. Details about the protocol used are available in the PoW home page under *GBIF data search*. In this version we are assuming that the corresponding services are handling synonyms correctly, and thus we did not retrieve information for alternative names that might be present in some data sources. This would probably be desirable in the future.

We found distribution records for 93.1 % of the species of Papilionidae 74.7 % of Pieridae, and 56 % of Riodinidae.



For the top 30 % of the species of Pieridae and Papilionidae the number of record per species was similar, but for the lower 70 %, Papilionidae had a higher number of records. However, for georeferencied records the differences were minimal. Both families were better represented than Riodinidae in both total number of records and georeferenced records.

Georeferenced records with normalized ranks between 0 and 1: no real difference between Pieridae and Papilionidae.

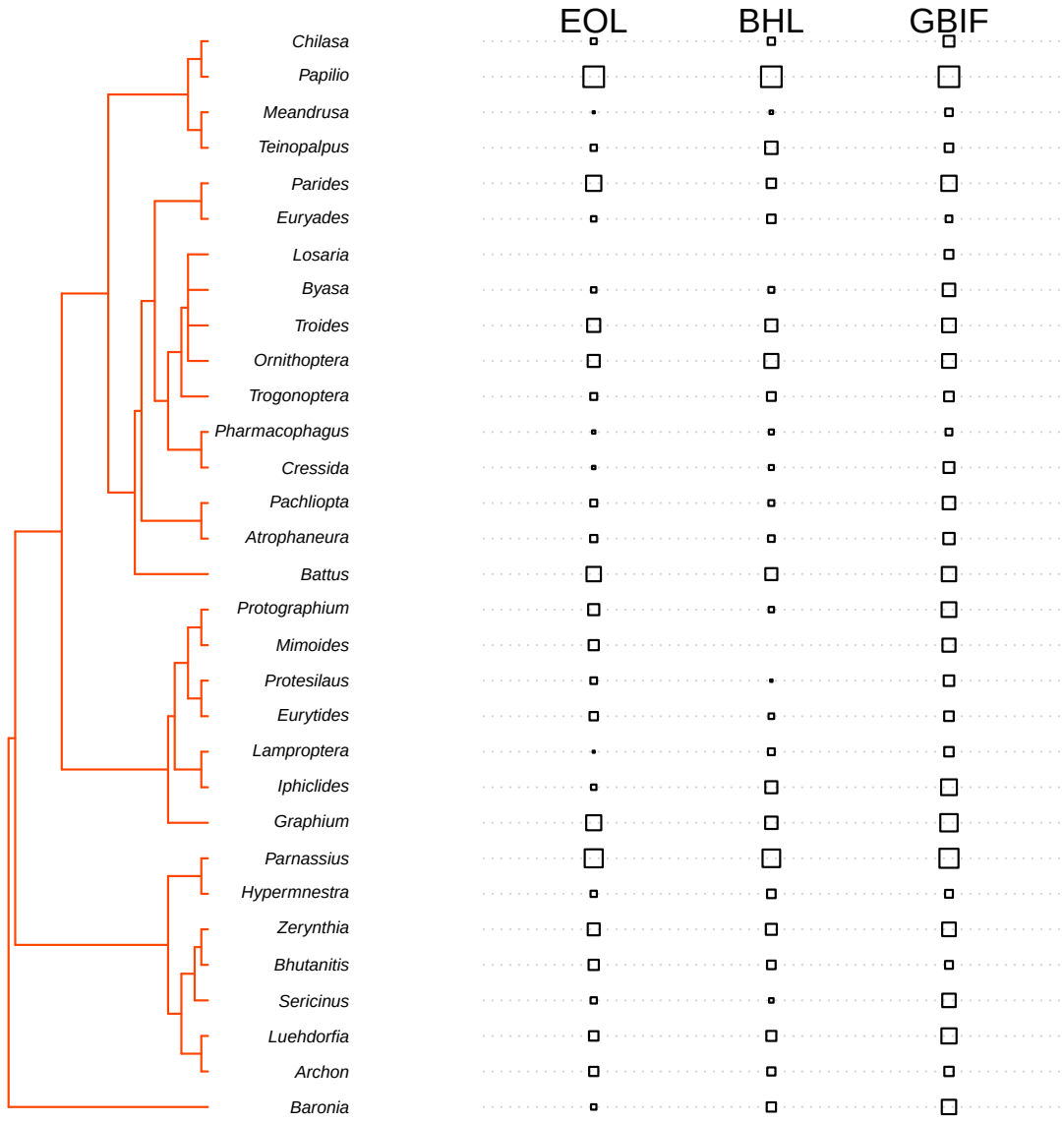


5. Comparing EOL, BHL and GBIF

We mapped the data available from all three sources onto the phylogenies of each family.

5.1. Papilionidae

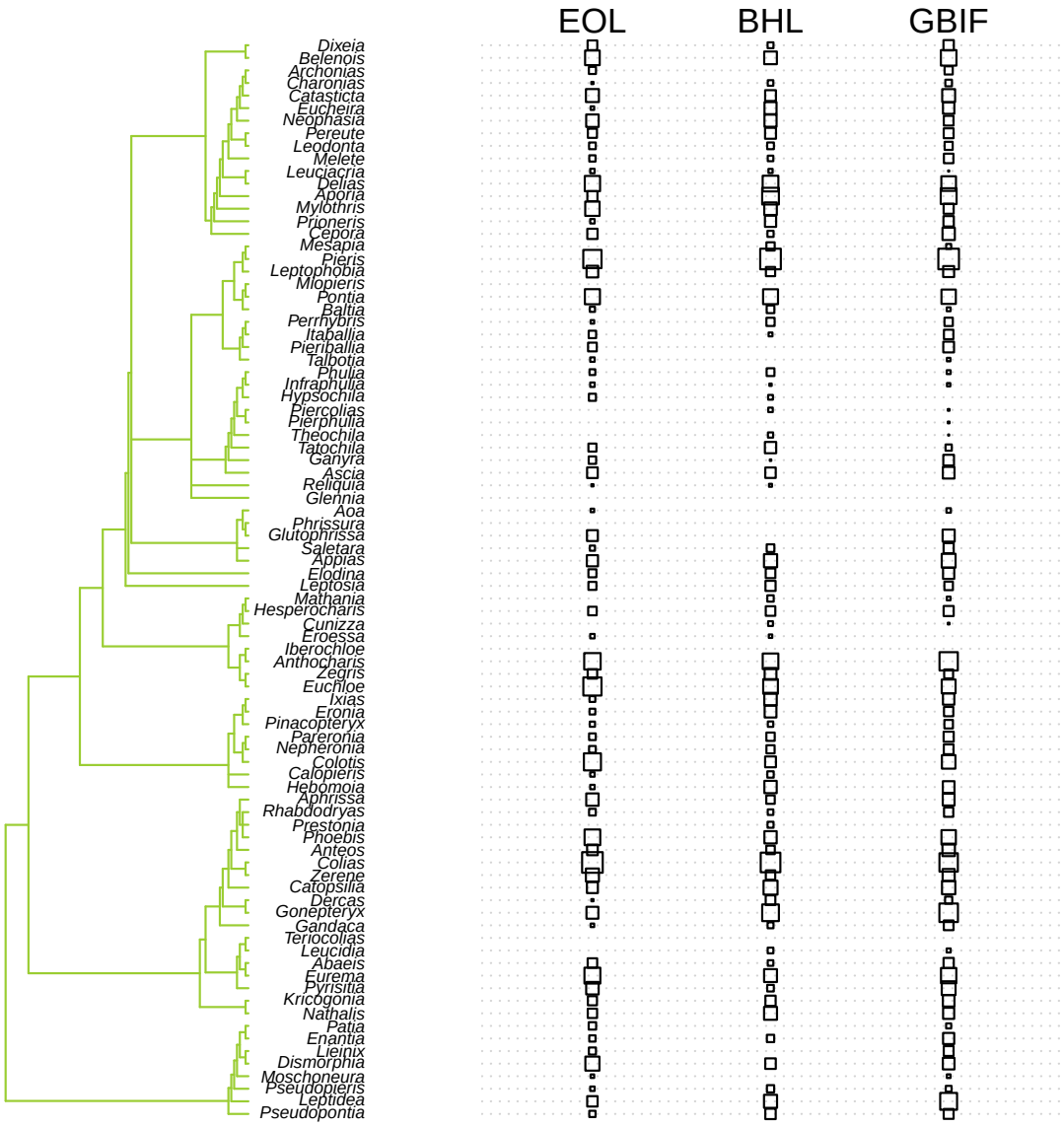
All genera were represented in GBIF, and most of them in the other two sources. GBIF and EOL had similar values of MPD and eveness, but BHL pages seems to be less evenly distributed accross the phylogeny of Papilionidae.



	SR	PSR	vars	MPD	PSE
EOL	30	20.55172	0.05629527	1.1489631	0.5942912
BHL	29	19.89048	0.10919200	0.7894366	0.4088154
GBIF	31	21.02000	0.00000000	1.1341995	0.5860031

5.2. Pieridae

Most genera in Pieridae are represented in all three sources, but fifteen are only represented in two or one source. GBIF records had slightly higher number of taxa and phylogenetic richness, while EOL data objects had higher values of MPD and phylogenetic eveness.



	SR	PSR	vars	MPD	PSE
EOL	78	53.69442	0.3607123	1.299560	0.6582189
BHL	76	51.56047	0.4397305	1.146857	0.5810741
GBIF	80	54.47625	0.2772303	1.130151	0.5722285

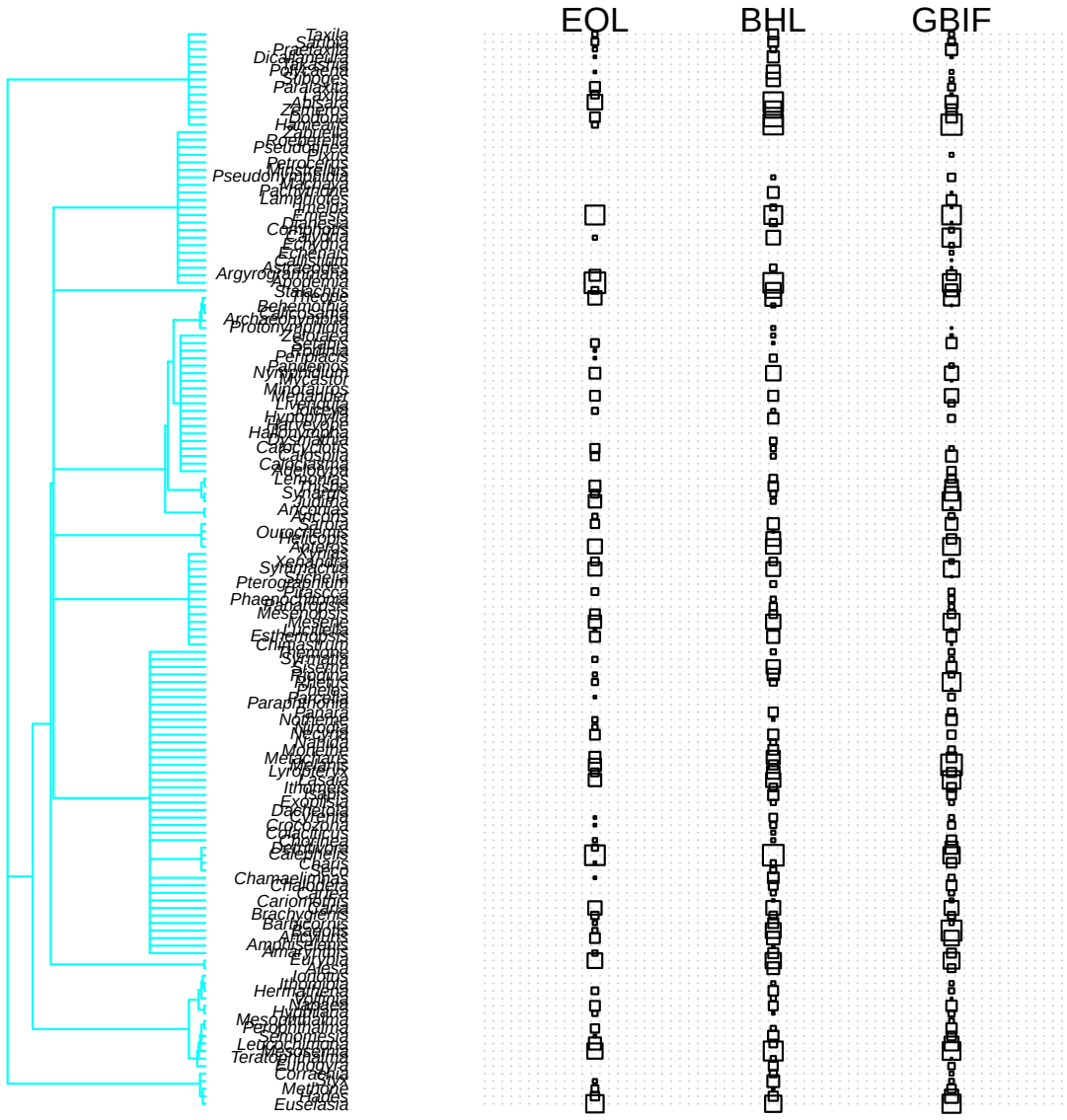
5.3. Riodinidae

Several genera of Riodinidae remain unrepresented in all three sources:

```
> rownames(riod.df)[rowSums(riod.df[,1:3]>0)==0]
```

```
[1] "Dachetola"     "Hallonympha"   "Minotauros"    "Archaeonympha"
[5] "Calicosama"    "Minstrellus"   "Pseudotinea"
```

GBIF has a higher number of taxa and phylogenetic richness, but BHL has higher values of MPD and phylogenetic eveness.




```
      SR      PSR      vars      MPD       PSE
EOL  103 78.88014 0.4328075 1.459399 0.7368535
BHL  109 81.93375 0.3875440 1.544174 0.7792361
GBIF 124 93.14142 0.2440355 1.443802 0.7277700
```

6. Comparing sources of hostplant records

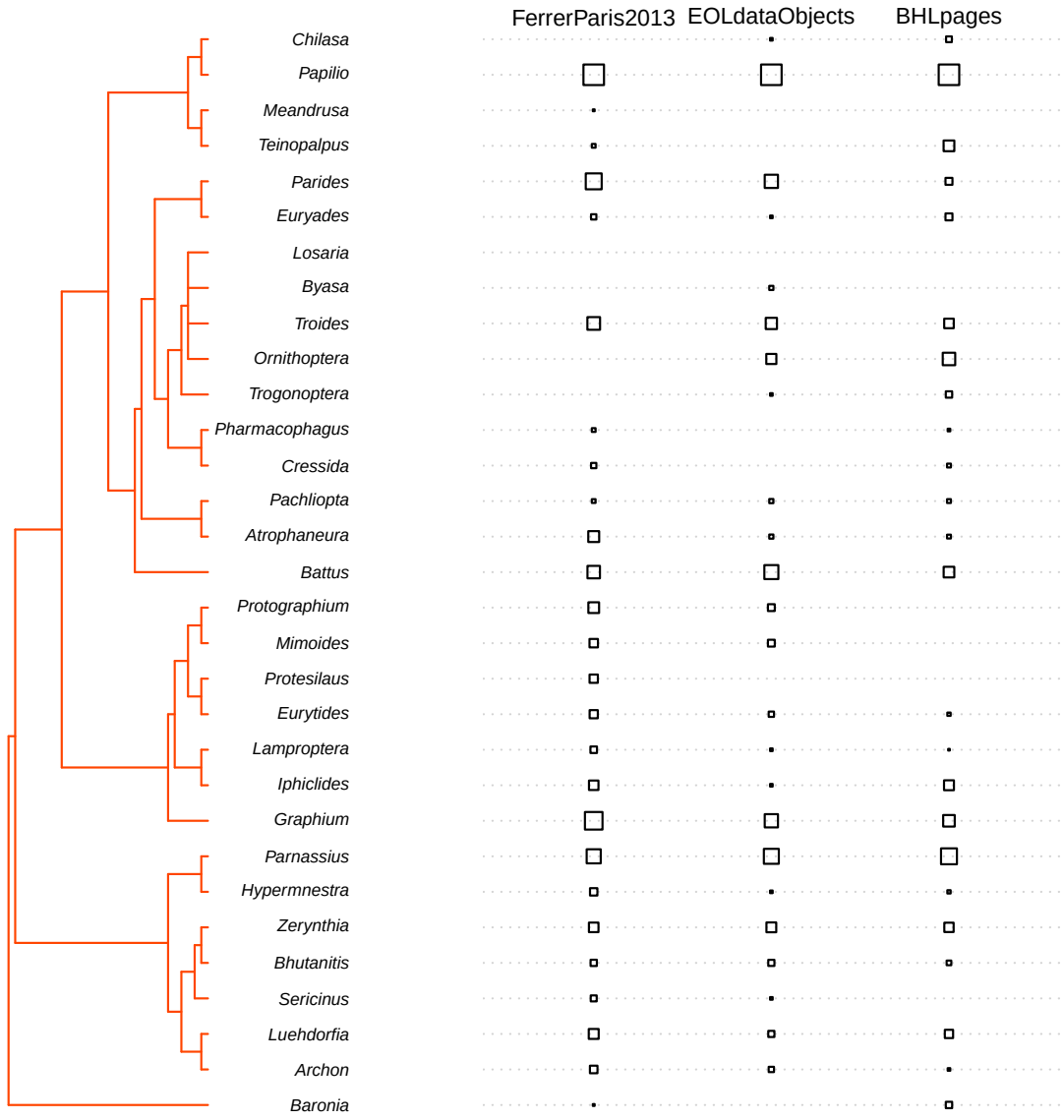
Finally, we compared the available information about hostplant associations. Here we use the data base compiled previously[6] to summarize current knowledge about the group and compare it with the number of EOL data objects selected by simple keyword matching, and the number of BHL pages that were classified as text pages and selected by simple keyword matching.

6.1. Papilionidae

The previous compilation has hostplant records for 27 genera, but EOL and BHL appear to have information for the following additional genera:

```
[1] "Trogonoptera" "Byasa"        "Chilasa"
```

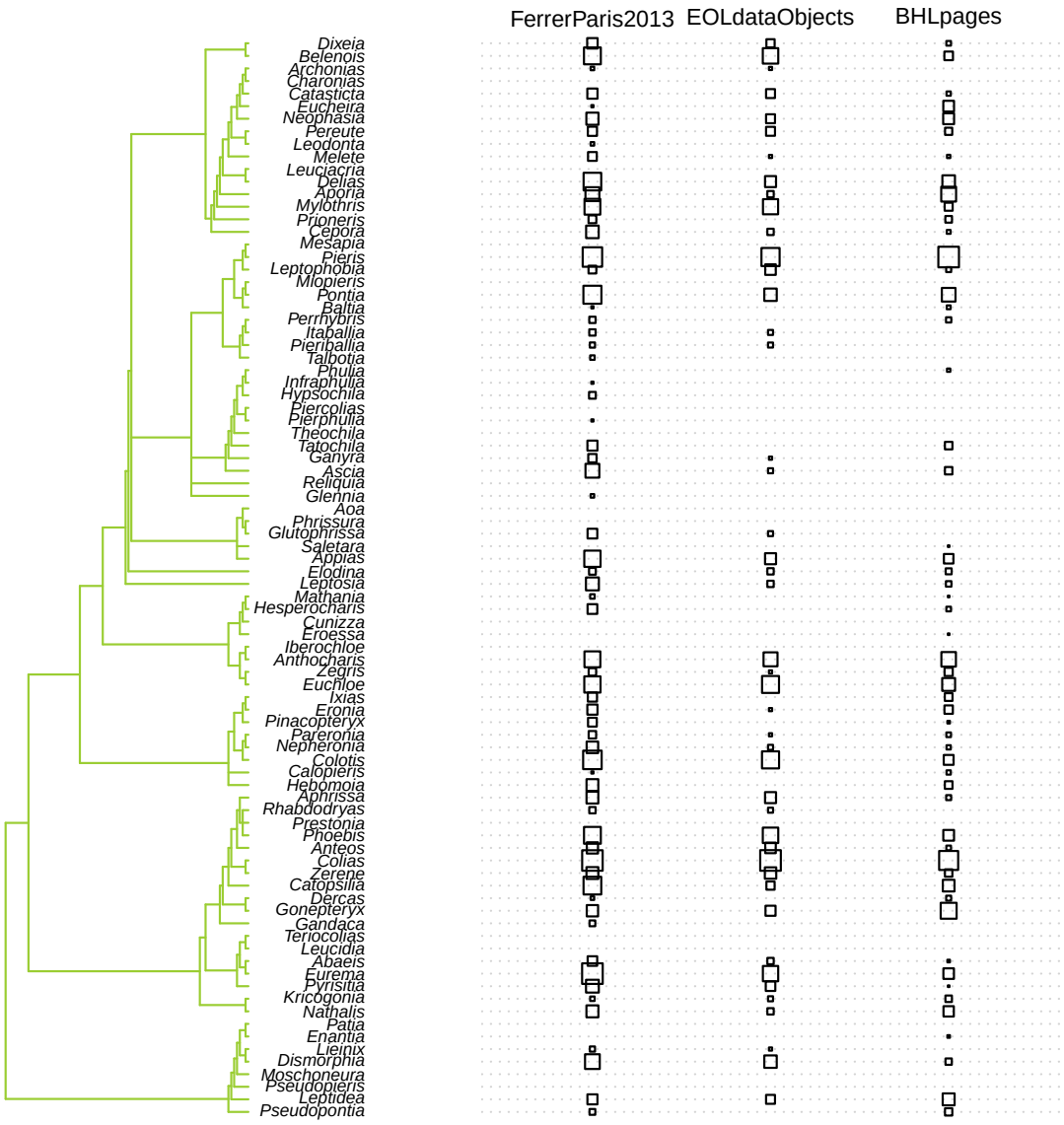
These additions would represent an almost complete coverage of the group.



6.2. Pieridae

The previous compilation of hostplant records appear to be more complete (including 71 genera) than either EOL or BHL, which include less taxa and seem to be more patchy. But even so, these two sources appear to have information for some additional genera:

```
[1] "Saletara"   "Piercolias" "Phulia"     "Mesapia"    "Charonias"
```



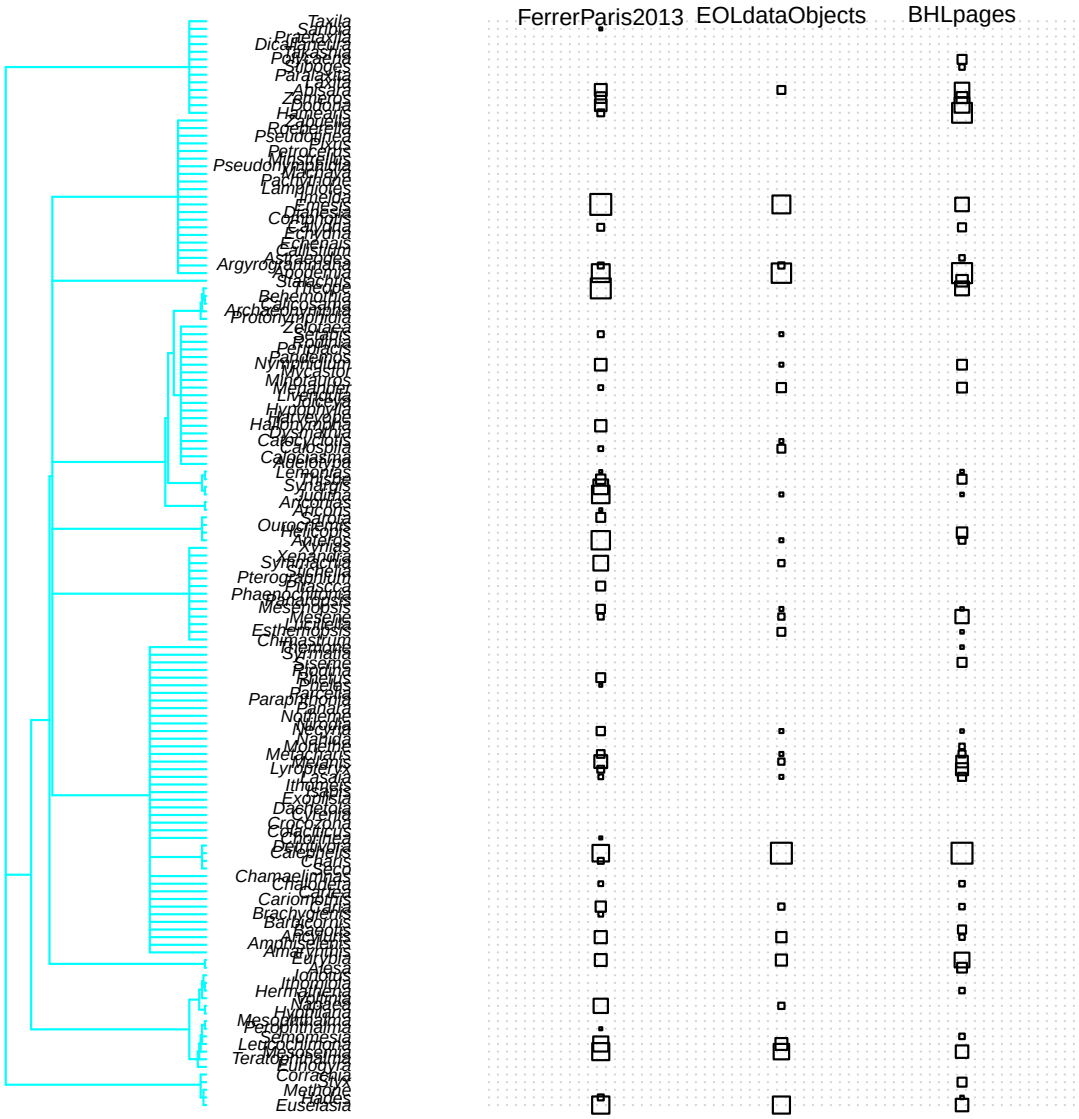
6.3. Riodinidae

The previous compilation has a poor coverage of the family Riodinidae, with records for only 58 of the 143 known genera. Although neither EOL or BHL have a throughout coverage of this family, together they could contribute new records for several additional genera:

```
 [1] "Methone"        "Styx"           "Teratophthalma" "Semomesia"
 [5] "Hermathena"     "Alesa"          "Amphiselenis"   "Baeotis"
 [9] "Isapis"         "Monethe"        "Nahida"         "Riodina"
[13] "Siseme"         "Themone"        "Xenandra"       "Helicopis"
[17] "Catocyclotis"   "Hypophylla"     "Joiceya"        "Behemothia"
```

```
[21] "Astraeodes"     "Dianesia"       "Imelda"         "Stiboges"
[25] "Dicallaneura"   "Praetaxila"
```

This would represent an increase of 44.8 %, but several important gaps still remain.



7. Conclusions

In EOL the content for Papilionidae species was richer, but slightly more text data object per species were found for Pieridae. Some contributing organizations provide the majority of the text data objects, but the coverage for each family was different. Contributions from local sources was important to extent the knowledge of regional faunas, specially those providing

content in spanish for Neotropical species. Searching for Hostplant associations in EOL is helped by the search of keywords, although the false positive rate is relatively high. However the access to concrete data objects allows fast manual validation of most records.

BHL provides a large number of matches for the butterfly species names, specially for Papilionidae species, although the total number of pages was larger for Pieridae. However, classifying useful content is more difficult. We found almost half of the matches refer to indices, reference list or bibliographies and comercial pages in older journals. We further narrowed search of hostplant associations by searching for keywords. This resulted in the selection of a large number of pages for Pieridae, and fewer for Papilionidae an Riodinidae. Manual validation of this information is very slow due to the high amount of redundant information.

GBIF data coverage was almost complete for Papilionidae, although the number of records per species is similar as for Pieridae.

For Papilionidae and Pieridae EOL provided more complete coverage and better representation of taxa than BHL or GBIF, but for Riodinidae BHL seems to be better, and could be a source of information for improving coverage in EOL.

Both EOL and BHL could be very useful sources to extend the current compilation of hostplant records, specially for Pieridae and Riodinidae.

Referencias

[1] M. F. Braby and J. W. H Trueman. Evolution of larval host plant associations and adaptive radiation in pierid butterflies. *Journal of Evolutionary Biology*, 19(5):1677–1690, 2006.

[2] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees*. Wadsworth, 1984.

[3] Andrew V. Z. Brower. *Riodinidae* Grote 1895. metalmarks. http://tolweb.org/Riodinidae/12174/2008.01.01 in The Tree of Life Web Project, http://tolweb.org/, 2008.

[4] Andrew V. Z. Brower. *Pieridae* Swainson 1820. http://tolweb.org/Pieridae/12176/2009.11.15 in The Tree of Life Web Project, http://tolweb.org/, 2009.

[5] F.L. Condamine, F. A. H. Sperling, N. Wahlberg, J Rasplus, and G.J. Kergoat. What causes latitudinal gradients in species diversity? evolutionary processes and ecological constraints on swallowtail biodiversity. *Ecology Letters*, 15:267–277, 2012. doi: 10.1111/j.1461-0248.2011.01737.x.

[6] José R. Ferrer-Paris, Ada Sánchez-Mercado, Ángel L. Viloria, and John Donaldson. Congruence and diversity of butterfly-host plant associations at higher taxonomic levels. *PLoS ONE*, 8(5):e63570, 05 2013. doi: 10.1371/journal.pone.0063570. URL http://dx.doi.org/10.1371%2Fjournal.pone.0063570.

[7] C. L. Haeuser, J. Holstein, and A. Steiner. The Global Butterfly Information System. http://www.globis.insects-online.de, 2005. (last update on 14.04.2011, accessed in march 2013).

[8] Maria Heikkilä, Lauri Kaila, Marko Mutanen, Carlos Peña, and Niklas Wahlberg. Cretaceous origin and repeated tertiary diversification of the redefined butterflies. *Proceedings of the Royal Society B: Biological Sciences*, 279(1731):1093–1099, 2012. doi: 10.1098/rspb.2011.1430. URL http://rspb.royalsocietypublishing.org/content/279/1731/1093.abstract.

[9] M.R. Helmus, T.J. Bland, C.K. Williams, and Ives A.R. Phylogenetic measures of biodiversity. *American Naturalist*, pages E68–E83, 2007.

[10] Robert J. Hijmans, Steven Phillips, John Leathwick, and Jane Elith. *dismo: Species distribution modeling*, 2012. URL http://CRAN.R-project.org/package=dismo. R package version 0.7-17.

[11] C. Webb, D. Ackerly, M. McPeek, and M. Donoghue. Phylogenies and community ecology. *Annual Review of Ecology and Systematics*, 33:475–505, 2002.