info phylogeny construction

The development of the phylogeny

This directory includes two phylogenetic trees that were developmented based on the phylogeny "ALLMB" published by Smith & Brown 2018. Taxa that are not included in "ALLMB" were added to their genus root if any of their congeneric taxa are included in "ALLMB". Otherwise, they were added to their family root if any taxon of the same faimly is included in "ALLMB".

The phylogeny dataset

The phylogeny of the taxa with accepted names in TPL (n = 326101) was saved as the data file "TPL.phylo.Rdata", while that of WCVP database was saved as "WCVP.phylo.Rdata". When using them, just load them with the R base function load().

how the taxa are matched or added to ALLMB

The dataset "df.TPL.matching.Rdata" and "df.WCVP.matching.Rdata" include the information of how each taxon was matched or added to ALLMB. The two datasets have the same columns.

- family: the family of the taxa
- genus: the genus of the taxa
- Taxon: the accepted name of the taxa in TPL or WCVP
- Taxon: replace the empty space in the Taxonomic name with "_"
- match.type: how the taxa were added to ALLMB. These could be "directly.matched", "added.to.genus.root", "added.to.family.root" or "added.to.root.of.closest.family". The number of taxa that were to added.to.root.of.closest.family is very small (n < 10 in both databases). The family of these taxa are not included in ALLMB, so I added them to the root of their closest family by the information obtained from published literature.
- tip.label.in.the.phylogeny: indicating the tip.label of the corresponding Taxon. In most cases, they are the same as Taxon_. In only very few cases (n=12 for TPL and n=2 for WCVP), there are special characters in the Taxon name, I used tip.labels that are different from Taxon_, so the encoding problem can be overcomed. These special cases could be identified by the note column, when note == "special character in taxonomic name"
- note: see above.

special notes for usage of hybrid symbol.

Both TPL and WCVP have a messy usage of hybrid symbol, by a mixed usage of "x" and "×". I unified this by using the lower case "x" for all hybrid symbols. So when you have a hybrid taxon, make sure you have also used "x" for hybrid symbol.

summary of species matching and adding

TPL

```
load(file = "df.TPL.matching.Rdata")
gridExtra::grid.table(dplyr::count(df.TPL.matching,match.type))
```

	match.type	n
1	added.to.family.root	6676
2	added.to.genus.root	20383
3	added.to.root.of.closest.family	5
4	directly.matched	299037

WCVP

```
load(file = "df.WCVP.matching.Rdata")
gridExtra::grid.table(dplyr::count(df.WCVP.matching,match.type))
```

	match.type	n
1	added.to.family.root	6341
2	added.to.genus.root	86060
3	added.to.root.of.closest.family	7
4	directly.matched	292739