**Notes on the pilot with Lythraceae**

In the end I chose Lythraceae for the pilot because: (1) it’s not a huge family, so the process should be faster. It’s also not too small in the sense that we wouldn’t find variation or the problems in trait scoring that we will see for other families.

# TREE

For the first step, I got the original tree named myrtales\_dated\_final.tre and ran the function tidyng.Myrtales.tree to keep only one species per genus. This function is based on the table tips\_to\_drop\_final.csv that Eve and I edited. We should double check to be sure we’re keeping all tips we want to keep. Note also that the tidyng.Myrtales.tree will drop the epithet of the species and keep only the genus name. This is also to emphasize that our working unit is genus, not the species sampled in the tree. Those species sampled in the tree are merely a representation of the genus (in other words, it does not matter for this analysis which species was sampled). This is important to keep in mind for when we go to the part where we analyse traits.

I saved this pruned tree as “myrtales\_pruned.tre”. Then, I loaded it back to extract the clade corresponding to the Lythraceae family for the pilot (27 tips). I saved the Lythraceae tree as “pilot\_tree.tre” in the folder “pilot” where all the dataset relevant to the pilot analysis will be stored.

# DISTRIBUTION DATA

I then logged in GBIF and searched for “Lythraceae” to download all points assigned to this family there. I chose Lythraceae for the pilot with the hopes that it would be a small dataset, but I forgot that it’s also a cosmopolitan family with a lot of records from Europe and North America (a common bias in these dataset – these areas tend to be over collected in relation to the rest of the world). GBIF gave me a dataset of 887,084 occurrences for Lythraceae, which is a substantial dataset (~400megabytes). The problem with working with these big datasets of points is usually that they are very heavy and all analyses take a while, but let’s move on.

I then used my GBIF account to ask to download a “simple csv” file containing all Lythraceae points. GBIF took a few minutes to prepare it and sent me an email when the points were ready to download. I then went to my “downloads” part of my GBIF account and download the massive zip file with the records. Important to note here that I always take notes of the citation that they generate because GBIF is particular about citing the dataset:

GBIF.org (23 March 2022) GBIF Occurrence Download https://doi.org/10.15468/dl.6y9wvm

After downloading it, I will clean the points using the information from WCVP and TWGD with the function FilterWCVP that I prepared earlier. I then save the cleaned points as a new file “pilot\_cleaned\_points.csv”. We removed 36126 points here.

# RANGE AND CLIMATIC ANALYSES

# TRAITS

Looking at the trait dataset now. The trait dataset was sent by Shirley Graham. Honestly, I think a problem is that we often think about important characters as those used in taxonomy, but for diversification analyses other traits can be more important.

Let’s use this pilot with Lythraceae to see what would be the ideal dataset to analyse:

“We combine a novel dataset of five reproductive biology traits

* pollination syndrome,
* flower size,
* fruit type,
* seed size,
* seed number,
* **life form and**
* environmental niche estimates produced using WCVP data, and a
* dated phylogenetic tree of the order, based on phylogenomic data. “

Let’s first get the prevalent life form for each genus based on WCVP and adjusted using the table from S.Graham. There are two challenges here: the first is that the classification used by WCVP is not the most intuitive (see pdf attached for a chapter that explains them) and there are a lot of different categories. Ideally, we will want to reduce the number of categories by lumping some, depending on what we think is the most relevant classification to answer our questions. The second is that there are also a lot of NAs that we will probably want to score manually after we decide how to best discretize the WCVP categories.

Let’s tackle the first problem first: what do we want to answer with life form data? I realize, as I write this and double check the abstract we submitted to NP, that we were not super clear about the role of life form, though there are several hypotheses that correlate life form with diversification and rates of range expansion (e.g. annuals and herbaceous plants would have higher rates of both – ask me references later). So let’s go back to the Givnish et al. paper where we took some of our hypotheses from and see what they did. At a quick look, it looks like they were interested in:

“Growth form: Herbs, graminoids, shrubs, trees, vines

Dominant life history: Annuals, biennials, perennials “

Let’s see if we get a good representation of these categories in our dataset. To do this, I’m manually adding two columns to the pilot\_most\_common\_life\_form.csv table where I’m going to use the information from WCVP and the notes sent by Shirley Graham to score growth form and dominant life history. According to the Raunkiaer's Classification (see pdf) and the link from WCVP, only Therophyte are annuals. Many annuals are also herbaceous, so lets \*preliminary\* mark them as so in the second scoring.

According to <https://wcsp.science.kew.org/about.do#lifeforms> phanerophytes are trees and nanophanerophyte are shrubs, so let’s score them as such. Oh, but checking the table again we see that there are some genera marked as “Nanophanerophyte or phanerophyte”! What to do in these cases? Im conservatively marking those as shrubs (all of them will be shrubs at some point), but we may have to discuss this. Chameophytes seem to be subshrubs, but let’s mark them also as shrubs. Hemycriptophytes are herbaceous. Koehneria is marked as NA (WCVP couldn’t find data). What to do with them? Let’s see if S Graham notes gives us some tips. Five marked as “unmatch taxonomist and WCVP” out of 27 is not bad. It shows that WCVP is getting it right most of the time. The main inconsistent seems to be on whether one classifies those shrubs as herbs or herbs as shrubs. Let’s move on with this classification, though we may want to ask the specialist opinions when we have the scoring finalized.

“We combine a novel dataset of five reproductive biology traits

* pollination syndrome,
* flower size,
* fruit type,
* seed size,
* seed number,
* **life form and (check)**
* environmental niche estimates produced using WCVP data, and a
* dated phylogenetic tree of the order, based on phylogenomic data. “

For seed number, let’s try a continuous and a discrete simplification. Remember that, if we go with the discrete, we may want to perform some sensitivity analyses.

Pollination syndrome: this trait is more similar to life form in the sense that is discrete, but often difficult to choose one value for the whole genus. Again, we have three problems: how to categorize (i.e. which categories to use), how to deal with infrageneric variation, and the NAs (unknown values).

To think about how to categorize, let’s think about what we want to answer. In my head, I would think that the most clear separation is between abiotic pollination (boring flowers, no resources, often dioecious), insects (generalists, medium investment in resources) and vertebrates (usually heavy investment on resources), but a lot of pollination biologists would disagree with this (i.e. it’s a strong simplification). Let’s see how other people with similar problems dealt with this: both Givnish et al. 2021 and Hernandez-Hernandez and Wiens (2020) divided merely on “abiotic and biotic fertilization” (by the way, the latter’s explanation on how they scored traits is very relevant for our work, see their appendix which Im now also uploading to our drive). We can also try scoring in that way in a second scoring.

To choose one per genus we have to (1) go with the prevalent (most species); when the prevalent is unknown, we can try a sensitivity analysis where we score in different ways.

Bottom line: we can write that the traits were scored by taxonomists, but it also required a lot of thought to standardize the traits afterwards. For this pilot, we got:

Growth form, life history, seed size (discrete, based on mean for genus), seed size (continuous, based on mean for genus), seed number (discrete, based on mean for genus), seed number (continuous, based on mean for genus), pollination syndrome (separating biotic in insects and vertebrates), pollination syndrome (biotic and abiotic), flower diameter (discrete, based on mean for genus), flower diameter (continuous, based on mean for genus),