**Notes on the pilot with Lythraceae**

In the end I chose Lythraceae for the pilot because it’s not a huge family (so running some codes should be fast). 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 before. 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 each genus (in other words, it does not matter for this analysis which species was sampled, or the traits of the species sampled).** This is important to keep in mind for when we go to the part where we analyze 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.

# Further notes on cleaning Myrtales:

I had some trouble because the WCVP does not have distribution data of many species. In the way I coded the functions to clean the data, I was not able to flag those properly.

One thing that I just thought is that we could see also if the niche breath is correlated with diversification rates, so to see if the lineages that diversified more are those that were more flexible or those who were merely exploring a particular niche with a larger area.

(In the end I used the points that I cleaned for Myrtales because I had to change the filters a lot for them to work, so what I said above is kind of out of date)

**RANGE AND CLIMATIC ANALYSES**

I used some codes that I already had to get some climate niche and habitat data. I don’t know how good those are, so I did this just for the pilot since Sam is going to provide a better version of this dataset.

**TRAITS**

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

In the abstract, we said that we would focus on:

* pollination syndrome,
* flower size,
* fruit type,
* seed size,
* seed number,
* life form

Let’s first get the prevalent life form for each genus based on the WCVP and adjusted the categorization using the table sent by S.Graham. There are two challenges here: the first is that the classification used by the WCVP is not the most intuitive (see pdf added to the drive for a chapter by Raunkiaer that explains them) and there are a lot of different categories. Ideally, we want to reduce the number of categories by lumping some, depending on what we think is the most important to answer our questions. The second challenge is that there are also a lot of NAs that we will probably have 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 in our questions, 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 for 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 there. 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 can get a good representation of these categories in our dataset using the WCVP data. 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 Therophytes as annuals and herbaceous for life history and growth form. According to the <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 both “Nanophanerophyte or phanerophyte”. What do we do in these cases? As of now, I’m conservatively marking those as shrubs (all of them will be shrubs at some point their lives), but we may have to discuss this. Chameophytes seem to be subshrubs, but let’s mark them also as shrubs to reduce the number of categories. 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 on what the life form of this genus is. I also marked which ones seem to differ between Shirley’s and WCVP classification and found 5 mismatches. Five marked as “unmatch taxonomist and WCVP” out of 27 is not bad. It shows that WCVP is getting these categories right most of the time. The main inconsistency 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.

Let’s now move on to the other traits for which we will use the trait dataset sent by Shirley Graham to score. For the traits that have continuous properties (seed number, seed size, flower diameter), let’s try two types of scoring: a continuous one, based on the mean of the trait and a discrete simplification into three categories (something like small, medium, and large). Remember that, if we go with the discrete, we may want to perform some sensitivity analyses (i.e. try changing a little bit the threshold between categories to see if our results stand). Here, I think it would be really helpful if we had a “mean” for the genus, a minimum (smallest value for the genus) and a maximum (maximum value for the genus) too.

Pollination syndrome: this trait is more similar to life form in the sense that it is discrete, but it’s often difficult to choose one value for the whole genus. Again, we have similar problems: how to categorize it (i.e. which categories to use), how to deal with infrageneric variation, and the NAs (unknown values). To think about how to categorize pollination syndrome, let’s think about what we want to answer with this data. We want to know how much the plant relies on animals for pollination, but I think it would be also cool to have a proxy for how much they invest in producing flowers. In my head, I would think that the clearest 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 I know a lot of pollination biologists would disagree with this categorization (i.e. it’s a strong simplification).

Let’s see how other people with similar questions dealt with this: both Givnish et al. 2021 and Hernandez-Hernandez and Wiens (2020) divided pollination syndrome merely as “abiotic and biotic fertilization” (by the way, Hernandez-Hernandez and Wiens’s explanation on how they scored traits for each family of angiosperm is very relevant for our work, see their appendix which I’m 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 is: we can write that the traits were scored by taxonomists, but it also required a lot of thought to standardize the traits afterwards. We may have to send it back to them asking if it’s alright.

**Some notes from the results of this pilot:**

In Lythraceae, larger flowers are correlated with fleshy fruits. These are also the flowers with more stamens and ovules. Curiously, the opposite pattern is observed in Melastomataceae and Myrtaceae, where larger flowers are found in groups with dry fruits.

In general, I think it would be much better if we had a mean, minimum and maximum values for the traits that are more continuous and use them as continuous in the analyses. In this way we would be able to also talk about flexibility and lability, and I think this may be an information that the taxonomists know from the top of their heads. For pollination biology, it looks like dividing between insects and vertebrates would be good, but I think this is not always possible, so maybe only abiotic or biotic based on syndrome for the whole genus sounds good.