**From:** **Joshua E Introne** jeintron@syr.edu
**Subject:** Re: fly data
**Date:** February 10, 2025 at 6:29 AM
**To:** Antonio Gomez rgomez02@syr.edu
**Cc:** Scott S Pitnick sspitnic@syr.edu, Zeeshan Syed zeeshan.syed@liu.se, Stephen Dorus sdorus@syr.edu

Hi y'all.  Sorry for the long delay!  I had a flight today though, and so spent a little time with this.   This is just preliminary fiddling around, but thought I'd share just a few early plots for now. Right now I'm just exploring UMAP and looking at how your data correlates with patristic distance.  Just to make things uber clear, I'm using the term patristic distance to quantify distance on the phylogenetic tree, and will use the term "feature distance" to speak of distance in the embedded feature space.

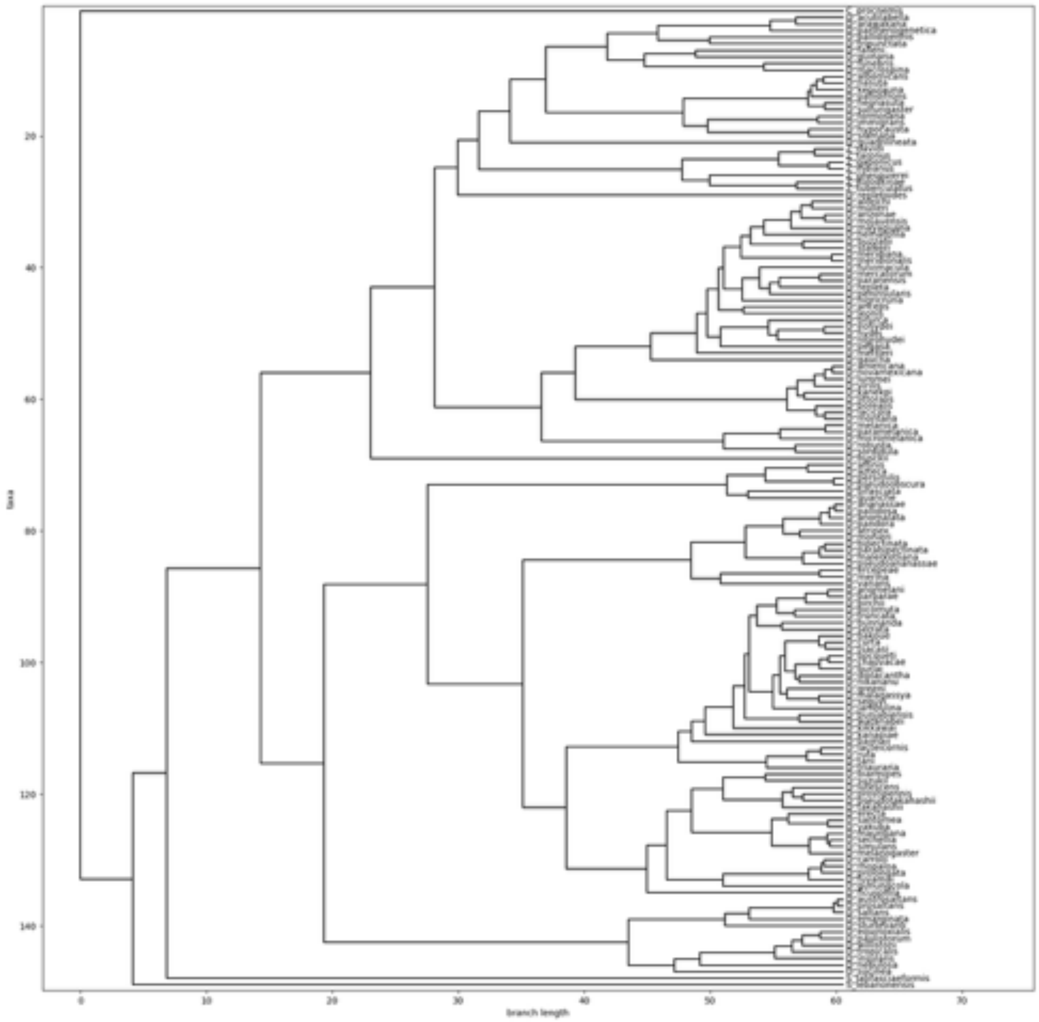
Note a couple of caveats here:

1) there is one strange value in the `insemination_rxn` column of 0A.  I've replaced that with a 0, but perhaps that means something specific?
2) I'm not accounting for sd at alll
3) I've scaled all features and I'm imputing missing values with KNN - some of your columns are missing quite a bit of data, so this could mess things up.  In particular, this is explicitly *wrong* for binary columns.  I'm aware of this and will fix, but honestly, it's probably not a huge difference.

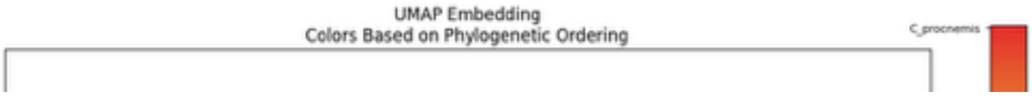
Also note - UMAP is stochastic and sensitive to parameters, and I've not done much in the way of robustness checking here.  However, over several tests, these results seem representative.

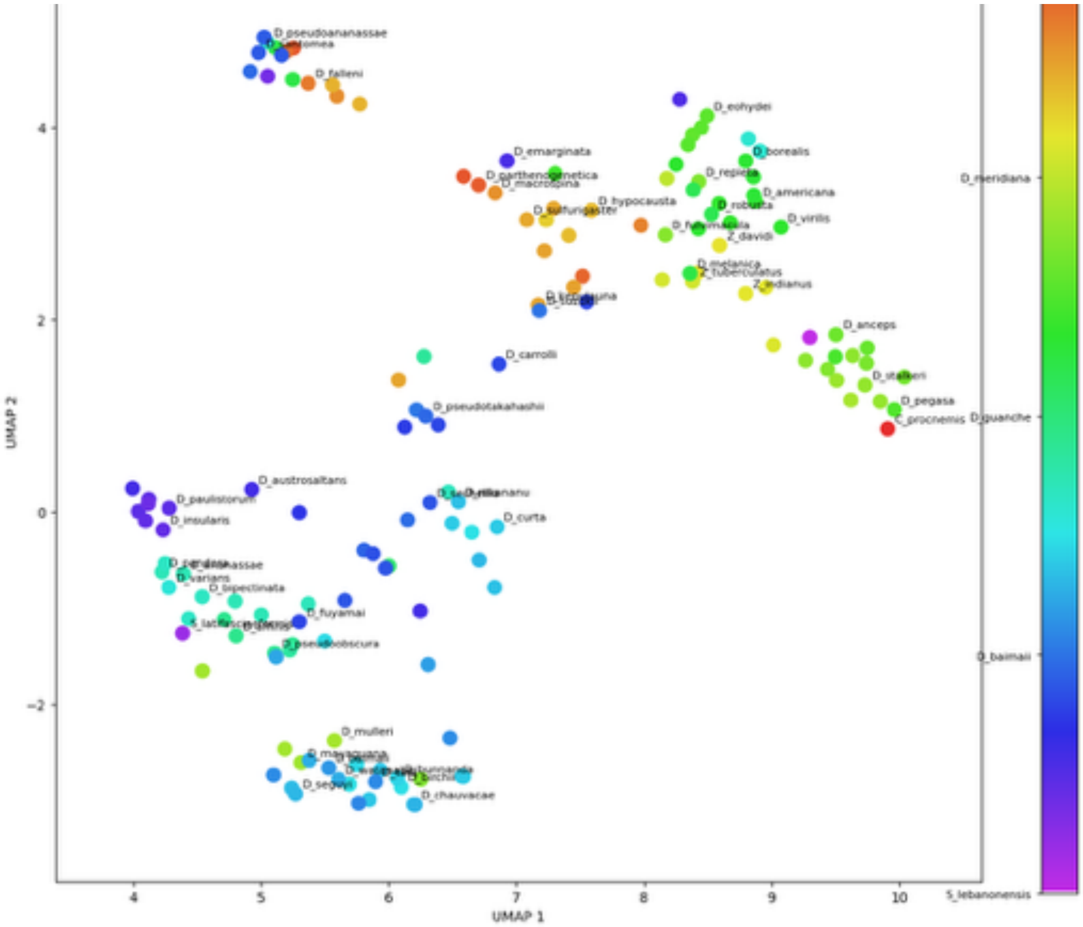
I don't know if you'll get the attached images inline, but you'll (mostly) be able to tell which is which from the plot titles.

So, here's your phylogenetic data, which I'm sure you know well :-)



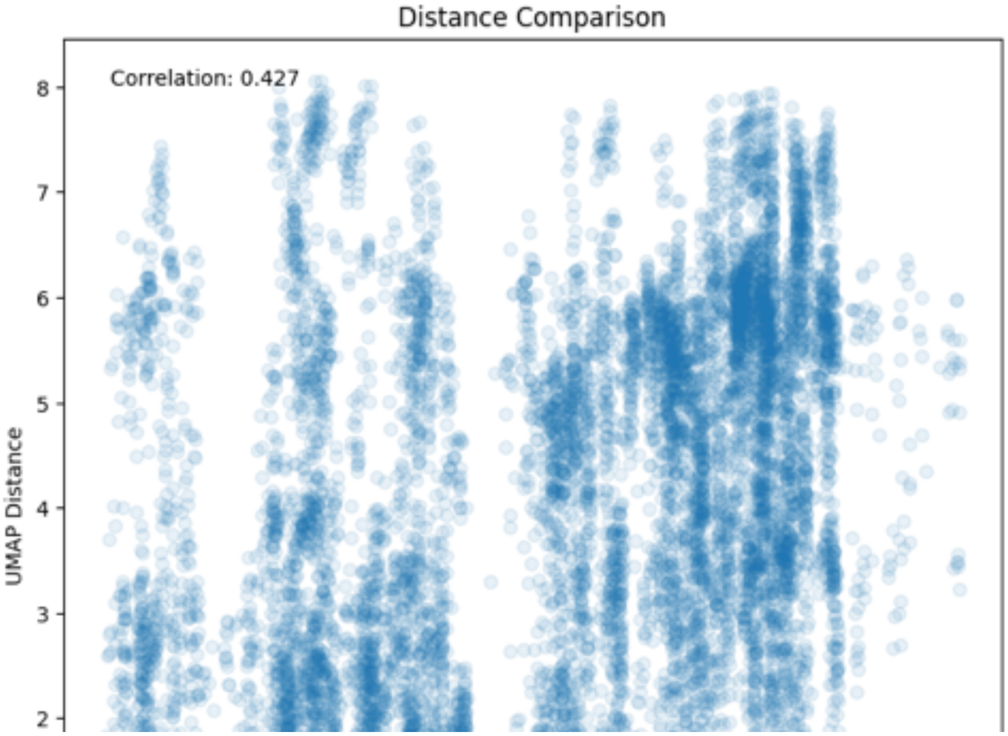
I've mapped the distances between adjacent species in this plot to hue, so we can get a visual sense of patristic distance, and used this to color the UMAP embedding. I've only labeled some of the species here to cut down on visual noise.  I can improve this, but this is a first cut.

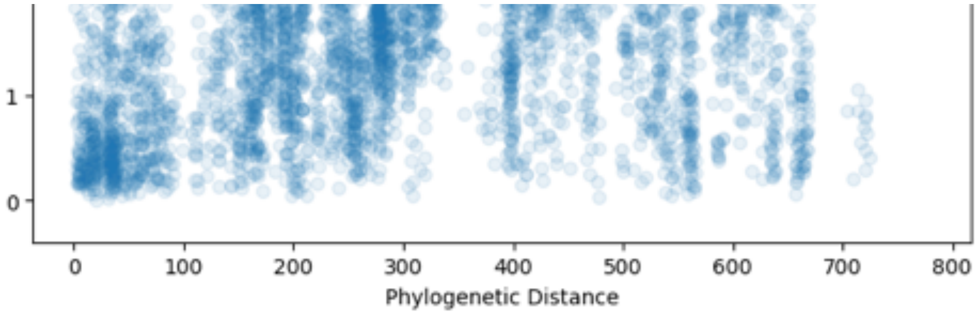




There a decent correlation here between feature distance and patristic distance, but there are definitely some oddballs out there - for instance, that group in the upper left.  Depending on parameters, the correlation between patristic distance and feature distance (using *cosine* rather than euclidean distance here, which shouldn't make a heck of a lot of difference) is roughly .4 - .5.

Here's a plot comparing patristic distance with feature distance:

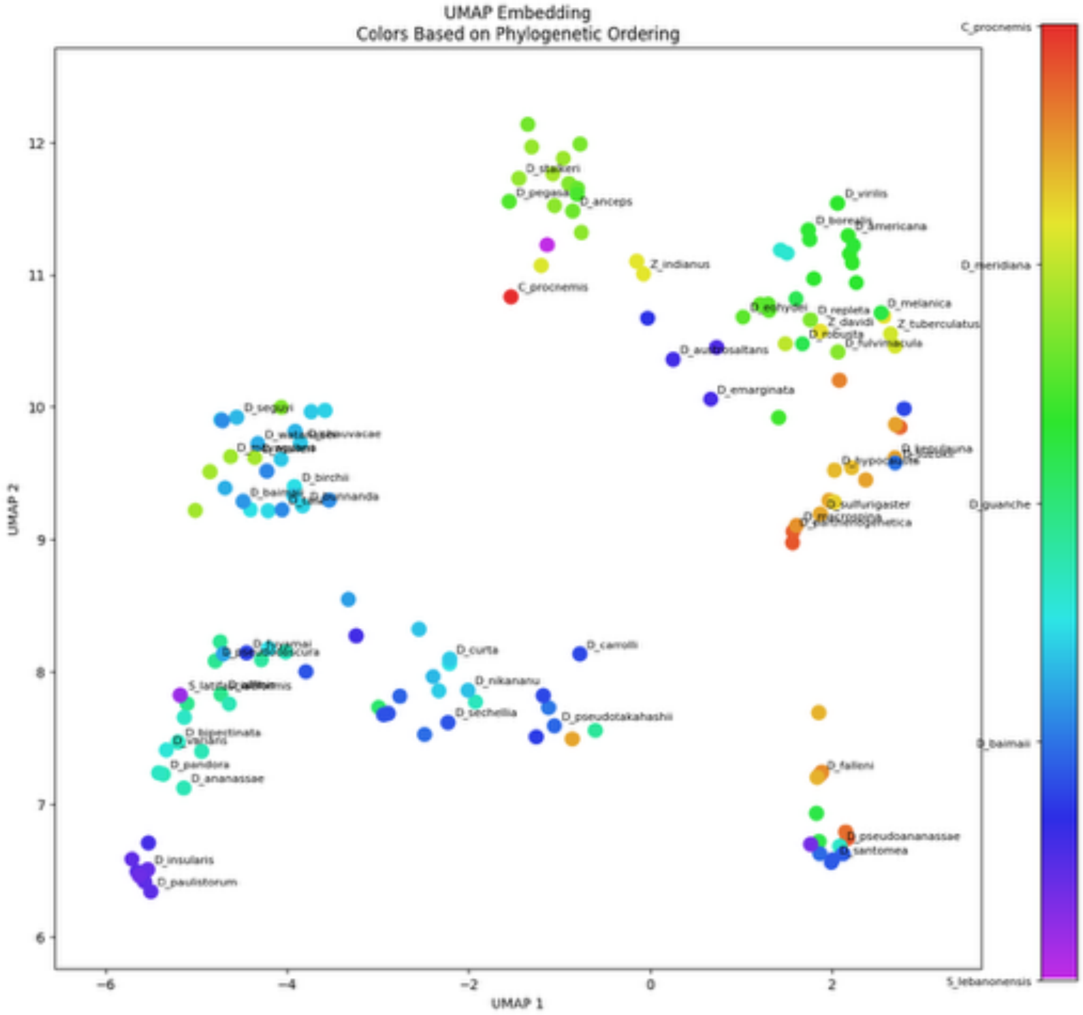




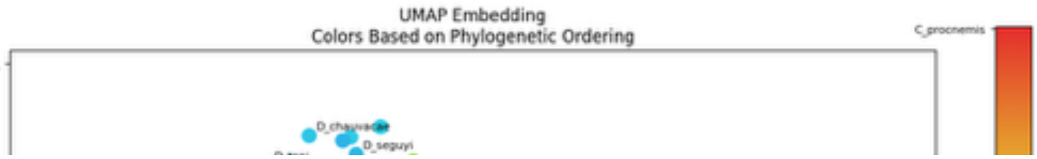
R (correlation) is not necessarily a great metric here, but as a rough cut, this is a reasonably strong.  I can give you better estimates and consider other metrics.

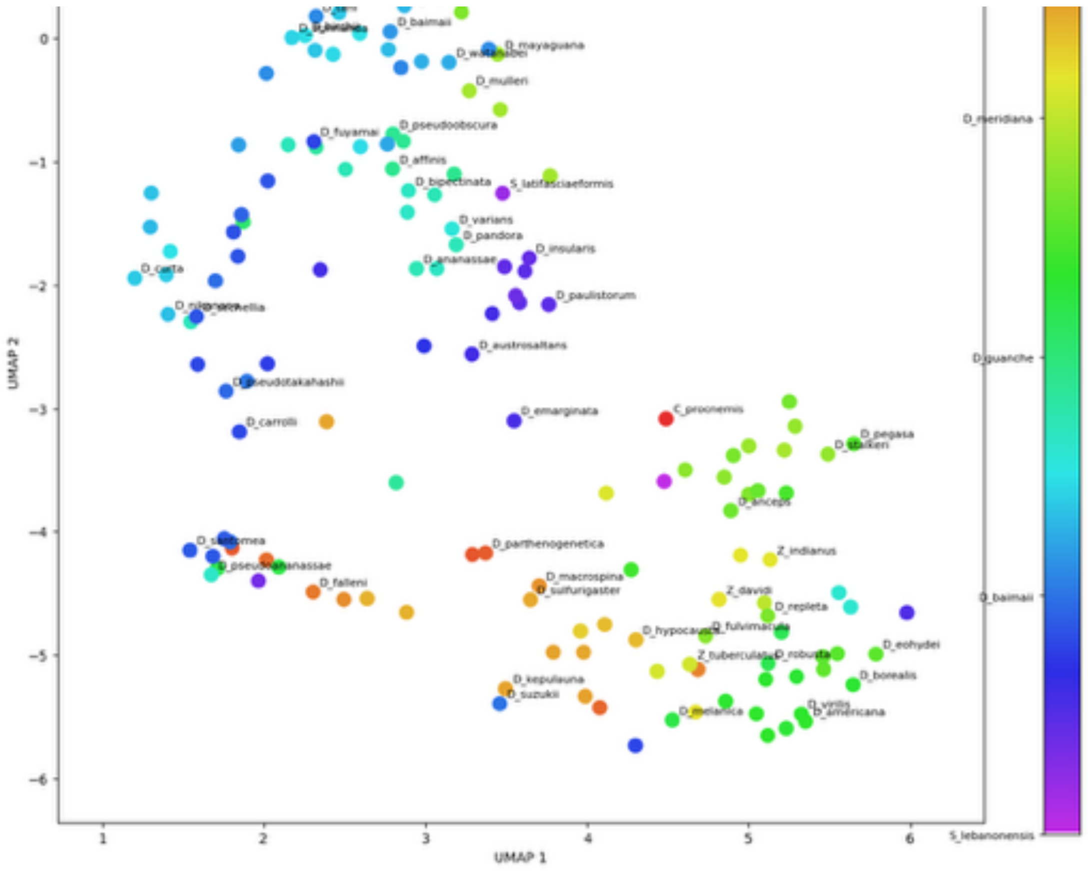
Just for a little more eye candy, here are a few more plots at different UMAP settings and reporting R in each case.  FWIW, I'm varying what might be thought of as “global fidelity” - this is roughly how far out from each point UMAP looks when trying to figure out spatial positioning.  This is subtle though - global fidelity may seem like a good thing, but it can ruin a mapping by disrupting *local* distances.  For a really great, interactive tutorial on UMAP, have a look here:  [https://pair-code.github.io/understanding-umap/](https://pair-code.github.io/understanding-umap/).  Check out the mammoth.  I’m playing with the “nearest_neighbors” parameter in the following.

R = .37; nearest neighbors = 25

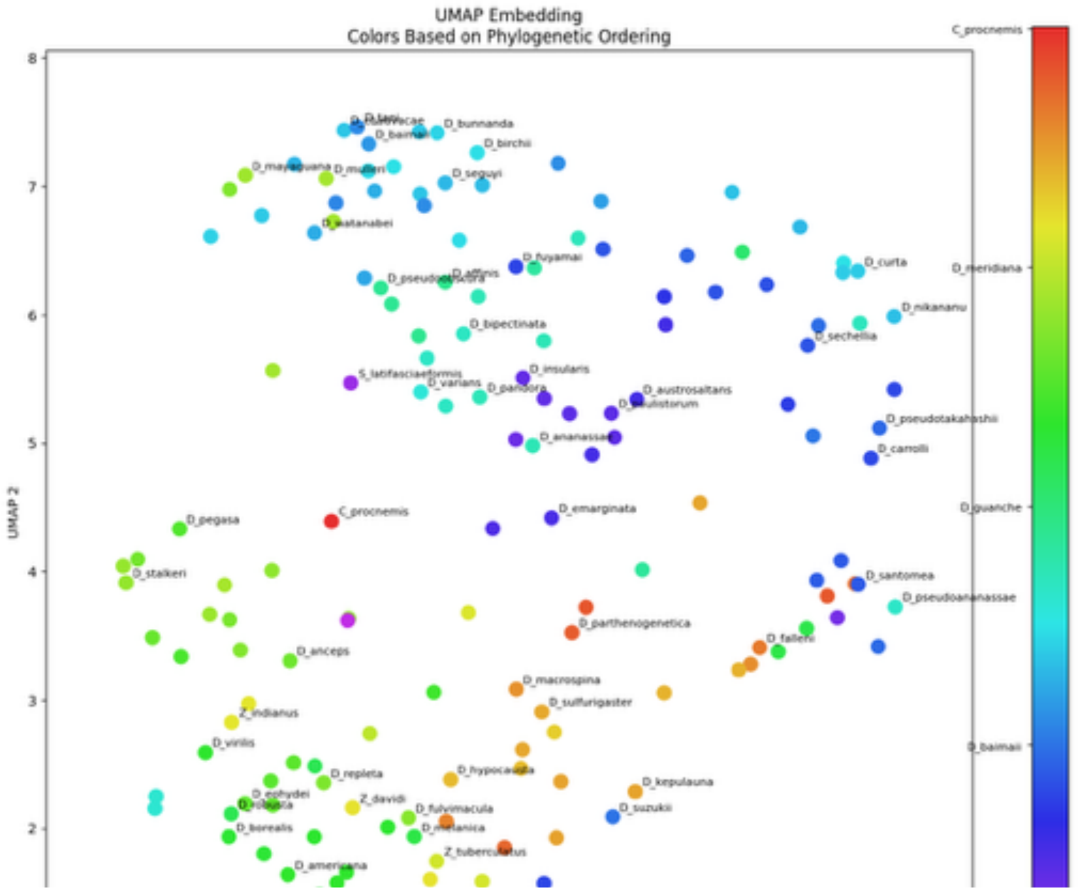


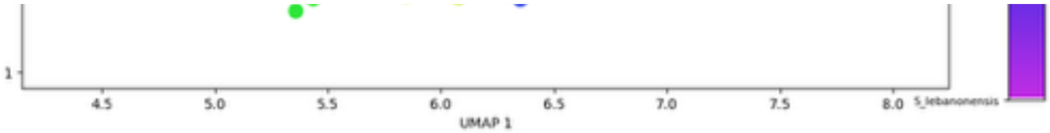
R = .49; ; nearest neighbors = 50





R = .47; ; nearest neighbors = 100





You’ll observe that regardless of parameters, the rough layout of things doesn’t change all that much.  I can cluster these results up for you if you want and report inferred categories, but I want to do a better job with imputation and perhaps incorporate SD if I can figure out a way.

I’ll continue to fiddle, but welcome any thoughts / requests!  I will explore bringing patristic distances into UMAP - it’s not that hard to do I think. I’ll also cluster *features* to see if there are natural groups.

All the best,
Josh

> On Jan 13, 2025, at 3:04 PM, Antonio Gomez <rgomez02@syr.edu> wrote:
>
> You rock Josh! Good catch Dude! I just dumped the whole DEP on Josh. 😅 I also forgot to point out that all the trait columns with "se" refer to species std error for particular traits. Our N per species is small.
>
> Btw, whenever you are ready for it, it may be very helpful to examine simulated matrices of various dimensions under different evo models. Very easy to do.
>
> So stoked you guys!
> Toño
>
> ---
>
> **From:** Joshua E Introne <jeintron@syr.edu>
> **Sent:** Monday, January 13, 2025 12:38 PM
> **To:** Scott S Pitnick <sspitnic@syr.edu>
> **Cc:** Zeeshan Syed <zeeshan.syed@liu.se>; Stephen Dorus <sdorus@syr.edu>; Antonio Gomez <rgomez02@syr.edu>
> **Subject:** Re: fly data
>
> Ok, thanks (esp. for that chapter, Toño!  Eager to check it out), and not a problem.  This is helpful to start with;  everything must wait till I get my paper out the door on Weds, but after that it won’t take long at all to process whatever you throw at me; I’m pretty decent with data wrangling so probably can get preliminary analyses out the door quite quickly even if the data is imperfect.
>
> Very psyched to play with this data and collaborate with you all.
>
> All the best,
> j
>
> Josh Introne: https://ischool.syr.edu/joshua-introne/
> C4 Lab: https://c4-lab.github.io/
> Recent Work: Measuring Belief Dynamics on Twitter
>
>> On Jan 13, 2025, at 12:16 PM, Scott S Pitnick <sspitnic@syr.edu> wrote:
>>
>> Hey Josh,
>>
>> For now, whatever time you devote to thinking about how to explore syndromes, I suggest limiting it to approaches rather than actually playing with data.  I was just looking over the DEP data file that Toño sent to you and realize that it needs a ton of selective cleaning up before it can be useful.  Also, I don’t recall us discussing continuous versus discrete traits and whether both can be included.  Most of the data/traits are continuous but a few (e.g., whether or not there is a copulatory plug) are discete.  I think those are most trivial and so perhaps best for us to remove them from the file you work with.
>>
>> Toño and I will soon work on and then get something more useful to you.
>>
>> Fyi – I’ve also just reached out to our colleagues at Cornell to try to move forward on the environmental selection/resource ecology front.
>>
>> Scott
>>
>> ---

> **From:** Antonio Gomez <rgomez02@syr.edu>
> **Date:** Wednesday, January 8, 2025 at 7:55 PM
> **To:** Joshua E Introne <jeintron@syr.edu>
> **Cc:** Zeeshan Syed <zeeshan.syed@liu.se>, Stephen Dorus <sdorus@syr.edu>, Scott S Pitnick <sspitnic@syr.edu>
> **Subject:** fly data
>
> Hey Josh,
>
> Great meeting you and super fun chatting. 2 hours is actually a pretty short meeting for us. ;) Glad you stuck around for the lox and bagels. Thanks Scott (and clan Pitnick) for hosting.
>
> I'm sharing the Drosophila & kin tree and trait files before I forget. The tree file is in newick format (like the kind Steve mentioned). FYI it's one kind of file format for trees. It's a basic text file with species names and numbers indicating branch lengths in million of years. Lots of programs should be able to read it (eg R), and you can examine it with a basic text editor or with a tree visualization software like Mesquite or Figtree. The trait file is a standard csv file with species names for rows and traits for columns. We can talk about what all the traits mean, but maybe for now it's ok to just give you some stuff to play around with. You had asked tons of great questions about trees, and I get the sense you might enjoy reading more about them. I'm attaching a nice book chapter about the kind of data we are working with and the kind of statistical approaches that we have thus far developed. It's intended for a very general audience and focuses on the basics. However, he also has a nice section on the "future of the phylogenetic comparative method", and I gotta say it has a lot of application to what we talked about today.  In case it wasn't already painfully obvious, I'd be happy to chat more about this stuff.
>
> Looking forward to our next think tank session…ideally with Zee on the conference room monitor.
>
> Later,
> toño