Analysis plan for Lepidoptera rates and diversity project

Overview:

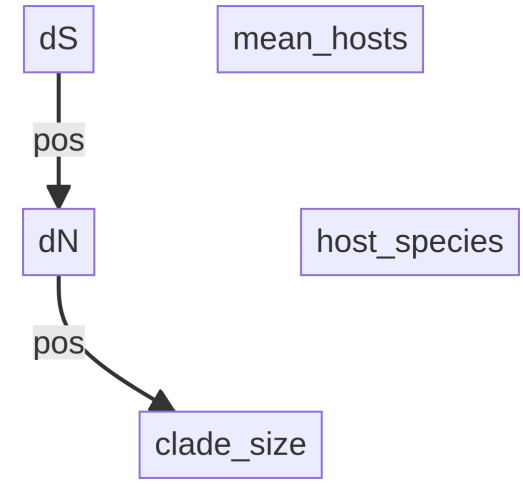
- I propose a path analysis for each of the genera, family, and major lineage level datasets to simultaneously test:
  - The direction of the relationship between mutation and speciation
  - mode of speciation (osculate, musical chairs, escape and radiate)
  - If results differ for each dataset, we will learn about overarching and proximal drivers of diversification (genera, major lineage, family levels)
- Key changes in my plan include:
  - Reducing host variables to `host_species` (the number of host species) and `hosts_mean` (the mean number of hosts)

Why reduce host variables?

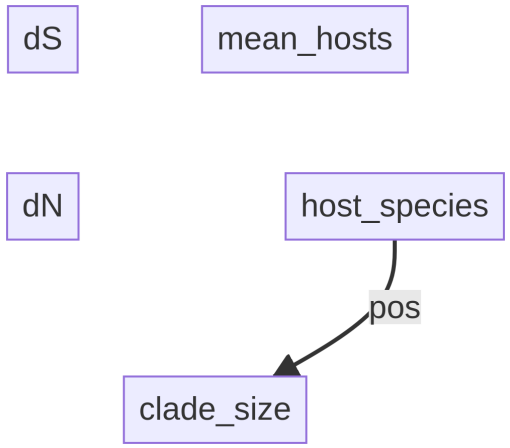
1. `host_species`, `host_families`, and `mean_hosts` are too correlated. 'Multicolinearity' is a weakness of path analysis and it will prevent convergence in estimation.
   - `host_families` could substitute for host_species, but it is necessarily a function of `host_species` (`host_families` >= `host_species`), so it would be extremely cumbersome to include both at once here. The escape and radiate model was the basis of including `host_families`, however I think we could recover a similar signal with `host_species`
2. Likewise, there are too many measures of generalism which will introduce the same issued. I propose we retain `mean_hosts` as the key measure of generalism.
   - Host phylogenetic diversity would be ideal, but contrasts too zero-inflated at the genera level.

- The proportion of generalists in a clade (proportion with >= 1 host) is too restrictive. E.g. a Koala is the epitome of specialisation, but would be a generalist here because it eats more than one kind of eucalypt!

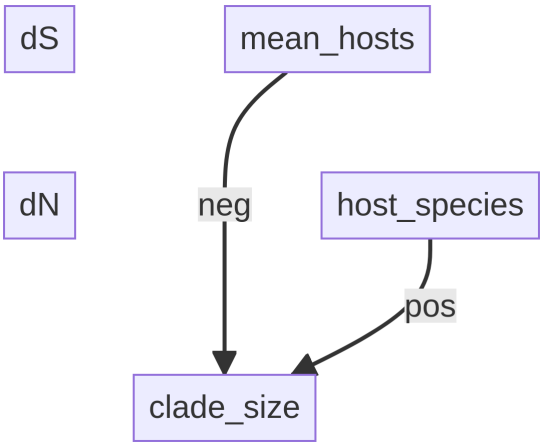
The proposed path analysis will relate to each speciation hypothesis as in the following diagram. The plan is for a nicer version of this to become part of a graphical methods in the paper.



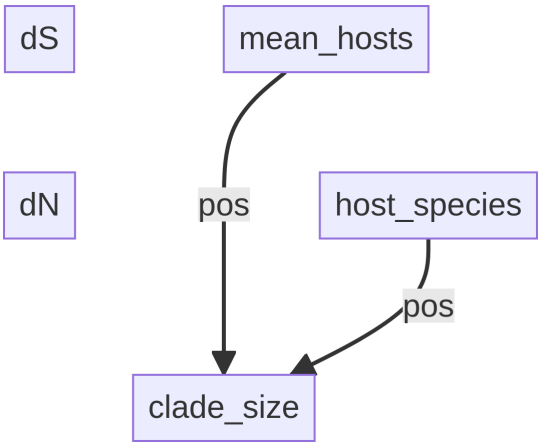
(a) Incompatibility due to isolation



(b) Escape and radiate



(c) Musical chairs



(d) Oscillate

Figure 1: Path analysis outcomes that would support each speciation and diversification hypothesis. **(a)** Incompatability due to isolation should show an association between substitution and clade size due to hybrid incompatability. **(b)** In escape and radiate, adaptation to a new 'niche' of hosts drives speciation, so more host species correlate with more diversity. **(c)** Musical chairs refers to taxa iteratively competing for specific hosts, hence a positive association with the number of host species and negative association with generalism. **(d)** Oscillate shouls show a positive association with generalism and the number of host species as generalists beget specialists in an oscillatory manner.

In the following, I include examples of what the results figues will look like with dummy data, and assess normality in the chosen (log-transformed) variables to support path analysis (Or

maybe poisson regression, Xia?).

Example results figure

The key results figure will look like the following (random data presented). I think a heatmap would be a simpler solution too.

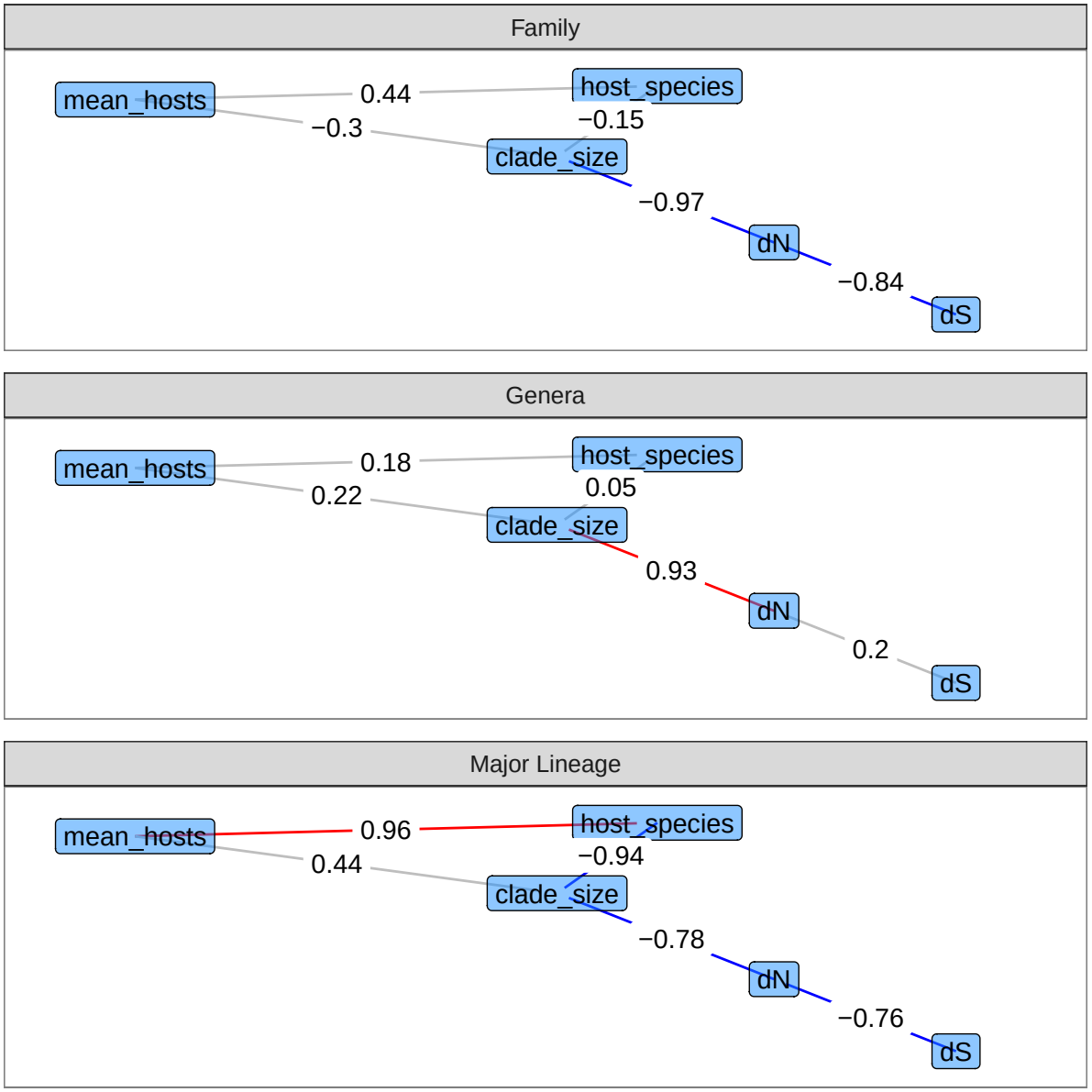


Figure 2: Path diagrams for each dataset. Edges are shown if significant and coloured by the sign of correlation. NB - These are made with random data for demonstration here! No formal analysis of the data has been done!

Normality of the chosen variables

Here I include what would become a supplementary figure assessing normality for the variables of choice. NB, this is using Andrew's original data. Some counts, such as host species, will

change slightly when we revise the counting scheme (as per discussion with Lindell).

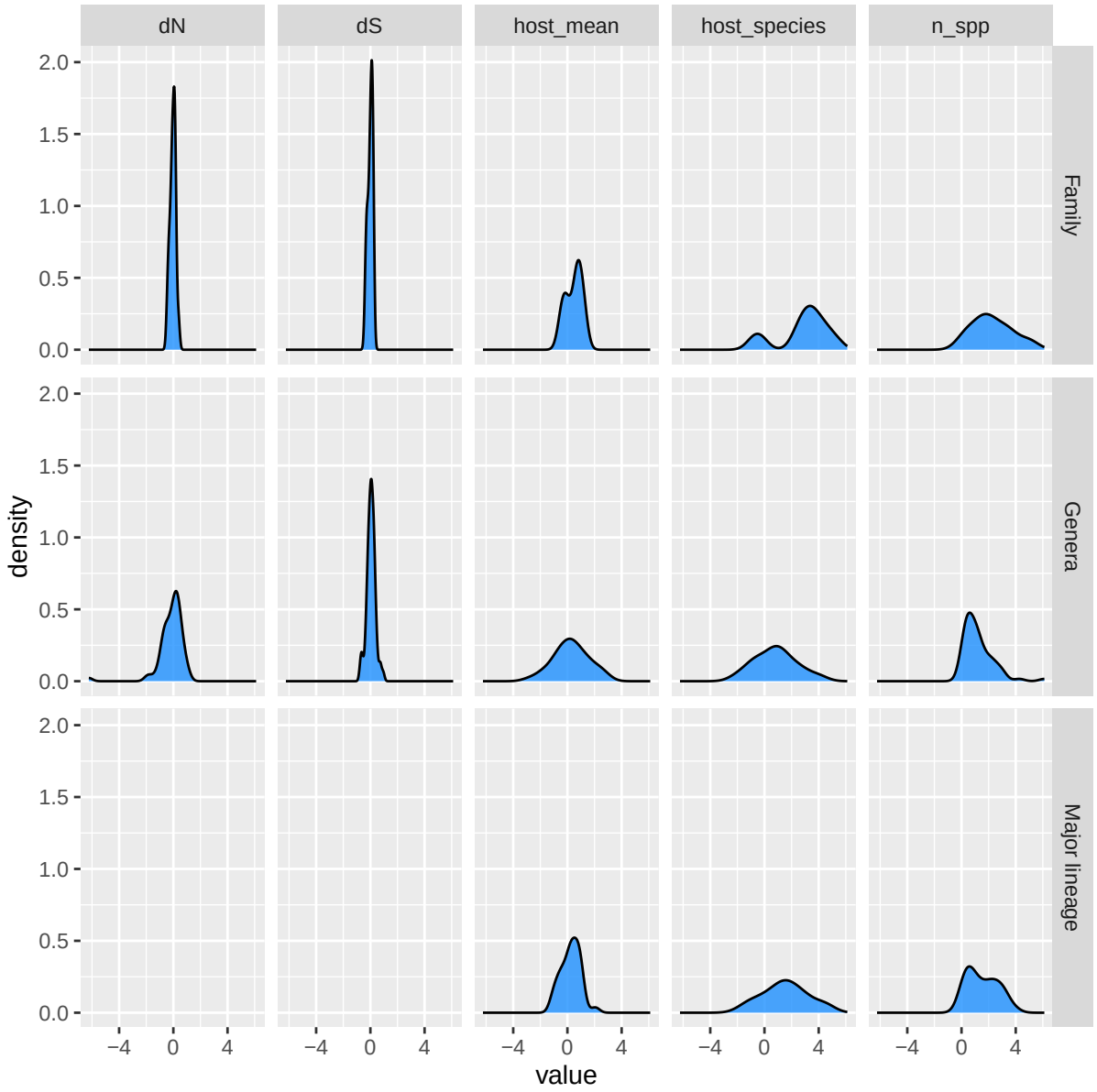


Figure 3: Histograms of log-transformed variable in each dataset as per Andrew's original data. Normality appears to be a reasonable assumption to proceed with path analysis with regular regression for most variables. `host_species` in the Family dataset may be an exception. `n_spp` is the same as the `clade_size` variable above