There are seven data files in this repository. 5 of these files are input data that contain the groups of pollinating animals, and 2 are output files that contain sets of likely pollinating species in either the 2017 version of the Catalogue of Life or the PREDICTS database. Fields for each of these datafiles are described below. For instructions on how to generate the output files please see the README at the base of this repo:

inputs:

1. '01_global_polllinating_confirmation_manual-edit.csv' is the original direct evidence csv aggregated from the output of the Ecography paper (i.e. any animal genus found in a pollination related abstract), edited following my initial check of abstracts. Each of the columns are described below:

- 'aggregated.scientific_name.i.' refers to each of the unique genera found in the initial scrape
- 'unique_class' refers to the taxonomic class of each genus
- 'unique_order' refers to the taxonomic order of each genus
- 'unique_family' refers to the taxonomic family of each genus
- 'unique_loc' refers to the set of unique Scopus IDs in which each genus was found
- 'DOI_count' refers to the number of abstracts in which that genus was found
- 'unique_year' refers to the set of unique years of the abstracts in which that genus was found
- 'unique_name' refers to the set of unique countries in which that genus was found
- 'unique_level' refers to the set of unique levels at which that taxonomic record was matched with the Catalogue of Life (see Ecography paper)
- 'Pollinating evidence/reference' is a binomial variable created on my initial pass through the abstracts for each genus, in which I assigned a Yes/No to indicate whether that genus is a likely pollinator
- 'Pollinating confidence' refers to the direct level of confidence assigned for each 'Y' above (see Nat Comms paper)

2. '02_confirmed_pollinating_families_edit' is the list of unique families with some evidence of pollination from the above direct evidence check, edited following my check of each family. Each of the columns are described below:

- 'class' refers to the taxonomic class of each unique family with some direct evidence of pollination
- 'order' refers to the taxonomic order of each unique family with some direct evidence of pollination
- 'family' refers to the taxonomic family of each unique family with some direct evidence of pollination
- 'species' refers to the number of species in total that family
- 'scraped' refers to the number of unique species of that family found in the initial taxonomic scrape
- 'prop' refers to the proportion of species in that family with evidence (i.e. 'scraped'/'species')
- 'family_checked' indicates that JM has check that family
- 'extrapolated' indicates whether that family (or a tribe/subfamily within that family) was extrapolated as pollinating

3. '03_clade_extrapolation.csv' is the set of taxa extrapolated as pollinators, prioritised through the initial text-analysis. Any taxa here assigned a value of direct confidence is one for which I found direct evidence at the genus level whilst searching for evidence to extrapolate. Each of the columns are described below:

- 'class' refers to the taxonomic class of each extrapolated group
- 'order' refers to the taxonomic order of each extrapolated group

- 'family' refers to the taxonomic family of each extrapolated group
- 'clade' refers to the taxonoic group extrapolated
- 'clade_rank' refers to the taxonomic level of the extrapolated group
- 'confidence' refers to the confidence at which that group was extrapolated (see Nat Comms paper for more details)
- 'additional_citations' refers to any additional references used in making the decision to extrapolate

4. '04_clade_extrapolation_non_text-analysis.csv' is the set of taxa extrapolated on the basis of Wardhaugh (2015), that didn't appear in the text analysis. Columns here take the same format as above.

5. '05_non_family-genus_species-list.csv' is the set of taxonomic names for groups that were not extrapolated at the family level (i.e. they cannot be easily merged with the PREDICTS database, so instead I searched for all the generic names and merged these instead). Each of the columns are described below:

- 'class' refers to the taxonomic class of that extrapolated non family group
- 'order' refers to the taxonomic order of that extrapolated non family group
- 'family' refers to the taxonomic family of that extrapolated non family group
- 'clade' refers to the extrapolated non-family group
- 'clade_rank' refers to the level at which that group was extrapolated
- 'genus' refers to the names of the genera within those non-family groups, used to match against PREDICTS
- 'confidence' refers to the confidence at which that non-family group was extrapolated.

outputs:

6. '06_COL_compiled_pollinators.rds' is the list of likely pollinating species returned after merging the above files with the 2017 version of the Catalogue of Life. Each of the columns are described below:

- 'scientific_name' refers to the species name of the likely pollinator
- 'Class' refers to the taxonomic class of that likely pollinating species
- 'Order' refers to the taxonomic order of that likely pollinating species
- 'Family' refers to the taxonomic family of that likely pollinating species
- 'genus' refers to the taxonomic genus of that likely pollinating species
- 'clade_rank' refers to the taxonomic rank of extrapolation for that likely pollinating species (if extrapolated, blank if not)
- 'confidence' refers to the confidence of pollination for that likely pollinating species (1-4 direct, 5.1-5.4 extrapolated)
- 'subfamily/tribe' refers to taxonomic subfamily/tribe of that likely pollinating species (if extrapolated at that level, blank if not)
- 'gen_prop' refers to an additional level of confidence (the number of scraped genera for that extrapolated group divided by the total genera for that group in the COL i.e. for what proportion of that group is there direct evidence)
- 'DOI_prop' refers to an additional level of confidence (the number of papers found for that extrapolated group divided by the total number of papers)
- 'add_conf' is a function of both 'gen_prop' and 'DOI_prop' (gen_prop * DOI_prop)
- 'fact_conf' refers to add_conf converted to a percentile of four levels, 'a' being highest confidence d = "0-25", c = "25-50", b = "50-75", a = "75-100

7. ‘PREDICTS_database.rds’ is a copy of the 2016 release of the PREDICTS database (see here https://data.nhm.ac.uk/dataset/the-2016-release-of-the-predicts-database for the full metadata and downloads in other formats)