nature portfolio

Corresponding author(s): Dr Jack Oyston

Last updated by author(s): Feb 17, 2022

# Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

## Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

| n/a | Confirmed | |
|---|---|---|
| ☐ | ☒ | The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement |
| ☐ | ☒ | A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly |
| ☐ | ☒ | The statistical test(s) used AND whether they are one- or two-sided<br>*Only common tests should be described solely by name; describe more complex techniques in the Methods section.* |
| ☐ | ☒ | A description of all covariates tested |
| ☐ | ☒ | A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons |
| ☐ | ☒ | A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals) |
| ☐ | ☒ | For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted<br>*Give P values as exact values whenever suitable.* |
| ☒ | ☐ | For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings |
| ☒ | ☐ | For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes |
| ☒ | ☐ | Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated |

*Our web collection on statistics for biologists contains articles on many of the points above.*

## Software and code

Policy information about availability of computer code

| | |
|---|---|
| Data collection | No software was used in data collection. |
| Data analysis | Statistical analyses were conducted in R (version 4.1.0) using RStudio (version 1.2.5033), using the following packages: dplyr, reshape2 and tidyverse. Figures were generated using the packages ggplot2, see and cowplot in R as well as Inkscape (version 1.1). |

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

## Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The data that support the findings of this study are available from the corresponding authors upon reasonable request.

# Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☐ Behavioural & social sciences ☒ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

# Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

| | |
|---|---|
| Study description | Comparison of species richness (number of species) in 356 pairs of polyploid and non-polyploid sister clades. Average species richness (means and medians) of the polyploid and non-polyploid samples were reported and differences in species richness compared using both two-tailed paired Wilcoxon signed rank tests as well as sign tests. This was done for the whole dataset as well as subsets based on taxonomic rank (genera only, families only, all ranks above genus) and taxonomic group (e.g. all animals, all plants, invertebrates, angiosperms, tetrapods). |
| Research sample | The sample consists of species richness estimates for 356 pairs of polyploid and non-polyploid clades identified from the existing literature. Of these 356 clade pairs, 321 of were genera. Of the remaining 35 pairs, 28 were pairs of families. While we focused our efforts on obtaining as representive sample as possible over the widest range of clades across the tree of life, certain groups are much better represented in the literature than others. 153 of the clades analysed were animals, while the remaining 203 were plants. Vertebrates compose the majority of animal clades (91 pairs), while within plants the majority of clade pairs are angiosperms (128 pairs). For a full breakdown of sample sizes within each group of organisms tests see Table 1 of the manuscript. |
| Sampling strategy | Biological publications spanning the years 1950 to 2018 were searched for known cases of polyploid taxa, using the names of major clades and “polyploidy” as keywords. n cases where this was ambiguous or contentious taxa known to be polyploid were compared with their closest known relatives. If polyploidy was found to occur in related taxa, then the least inclusive taxonomic ranking would be used to define the clade containing that taxon and the related taxa as the polyploid clade. Phylogenies in the published literature spanning 1975 to 2018 were used to identify the putative sister clades of our compiled list of polyploid clades. In cases where multiple phylogenies were found, the most recently published one was preferred, unless an older phylogeny resolved more polytomies, in which case that phylogeny was preferred. Sampling effort was focused on sampling as broadly across the tree of life as possible and obtaining sufficient sample sizes for statistical analysis for as many different taxonomic groups as possible, maximising the number of clades at the rank of class and above represented in the dataset over accurately representing the proportion of polyploid taxa in different groups. As a result, more research time was spent searching and analysing the literature of groups for which polyploidy was poorly or rarely documented than groups where polyploidy was widespread and well known. Known instances of artificial polyploidy or endopolyploidy were discounted. |
| Data collection | The number of species in each clade was estimated using online biodiversity databases, primarily using the Integrated Taxonomic Information System for animals and The Plant List for plants. A full list of source databases for species richness estimates of each group is provided in the methods. For a minority of clades diversity estimates were also made using the source publications, which are listed for each clade. Only accepted species names were counted, known synonymies and taxa of unresolved status were not included. Extinct taxa assigned to the clades in the dataset were included in species counts using the Fossilworks portal of the Paleontology Database, with clade names and “fossil” as keywords. Where possible, fossil data from the Paleontology Database was checked with data in source publications for those taxa. |
| Timing and spatial scale | Data were collected and analysed primarily between October 2017 and February 2019. While spatial scale is not relevant for the datasets and analyses presented here, clade pairs were identified and species richness estimated at the level of genus and above only. |
| Data exclusions | No data were excluded from statistical analyses of the whole dataset, although separate analyses were carried out on partitions within the dataset (e.g. all genera only, only animal clades etc.). |
| Reproducibility | A list of source papers, online resources and databases used to identify clade pairs and to generate species richness estimates for each group are provided in the methods section and supplementary materials. All input data and output values from all statistical analyses performed and all R scripts and details of analyses performed are provided in the supplementary material or available from the corresponding author on request. |
| Randomization | Data were not randomised as in each case we wished to compare the relative species richness of polyploid and non-polyploid sister clade pairs. Comparing sister clades provides the best possible test and set of controls for these kinds of analyses, as it controls for divergence time and many ecological and morphological properties of groups. |
| Blinding | Blinding was not possible or desirable in this study as in each case it was required to identify the specific taxa to form meaningful comparisons across sister clade pairs and to identify broader taxonomic groups in order to sample and test as broadly and evenly as possible across a wide range of groups. |

Did the study involve field work? ☐ Yes ☒ No

# Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

## Materials & experimental systems

| n/a | Involved in the study |
|---|---|
| ☒ | ☐ Antibodies |
| ☒ | ☐ Eukaryotic cell lines |
| ☒ | ☐ Palaeontology and archaeology |
| ☒ | ☐ Animals and other organisms |
| ☒ | ☐ Human research participants |
| ☒ | ☐ Clinical data |
| ☒ | ☐ Dual use research of concern |

## Methods

| n/a | Involved in the study |
|---|---|
| ☒ | ☐ ChIP-seq |
| ☒ | ☐ Flow cytometry |
| ☒ | ☐ MRI-based neuroimaging |