SCIENTIFIC DATA

OPEN

DATA DESCRIPTOR

LCVP, The Leipzig catalogue of vascular plants, a new taxonomic reference list for all known vascular plants

Martin Freiberg[1 ✉], Marten Winter [ID][2], Alessandro Gentile [ID][2], Alexander Zizka [ID][2,3], Alexandra Nora Muellner-Riehl [ID][2,4], Alexandra Weigelt [ID][1,2] & Christian Wirth[1,2,4,5]

The lack of comprehensive and standardized taxonomic reference information is an impediment for robust plant research, e.g. in systematics, biogeography or macroecology. Here we provide an updated and much improved reference list of 1,315,562 scientific names for all described vascular plant species globally. The Leipzig Catalogue of Vascular Plants (LCVP; version 1.0.3) contains 351,180 accepted species names (plus 6,160 natural hybrids), within 13,460 genera, 564 families and 84 orders. The LCVP a) contains more information on the taxonomic status of global plant names than any other similar resource, and b) significantly improves the reliability of our knowledge by e.g. resolving the taxonomic status of ~181,000 names compared to The Plant List, the up to date most commonly used plant name resource. We used ~4,500 publications, existing relevant databases and available studies on molecular phylogenetics to construct a robust reference backbone. For easy access and integration into automated data processing pipelines, we provide an 'R'-package (*lcvplants*) with the LCVP.

Background & Summary

Due to substantial progress in the last decade in improving plant taxonomy with phylogenetic findings, an updated global taxonomic reference list was urgently required. To date, the most commonly used reference list of vascular plant names is *The Plant List* (TPL, http://www.theplantlist.org/), hosted by the Royal Botanic Gardens, Kew. TPL contains 1,166,054 vascular plant names, including 308,397 accepted names, 304,419 of them angiosperms. ~760,000 names of TPL are synonyms, including 244,017 unresolved names. The here presented Leipzig Catalogue of Vascular Plants (LCVP) updates significantly the global knowledge of plant names not only compared to TPL (see Table 1) and thus is a major improvement for global plant research. It is based on existing databases (see Online-only Table 1) and an additional 4,500 publications (see the full literature package consisting of three different files as part of the publicly available LCVP data set at https://idata.idiv.de/ddm/Data/ShowData/1806 and Step 2 below for more details), which helped to clarify the status of plant names (i.e. accepted, synonym, taxonomic placement; see Methods). In the end, 4,059 publications provided relevant and robust additional information, e.g. changes in names and/or their status. A guiding principle during the compilation of the LCVP was to avoid polyphyletic genera, which are frequent in TPL, either by splitting genera (e.g. separating *Goeppertia* from *Calathea*) or fusing them (e.g. *Stapelia* and *Duvalia* in *Ceropegia*). However, we did not recombine any species name in the LCVP and in cases of unclear phylogenetic position of genera, we used the conservative (i.e. existing) name.

Taxonomists, ecologists and conservation biologists often work with many species (names) and cannot keep pace with the rapid progress in (plant) systematics, boosted by molecular phylogenetic methods[1]. These

[1]Systematic Botany and Functional Biodiversity / Botanical Garden, Institute of Biology, Leipzig University, Johannisallee 21-23, 04103, Leipzig, Germany. [2]German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, sDiv – Synthesis Centre, Deutscher Platz 5e, Leipzig, 04103, Germany. [3]Naturalis Biodiversity Center, Darwinweg 2, 2333, CR, Leiden, The Netherlands. [4]Molecular Evolution and Plant Systematics & Herbarium (LZ), Institute of Biology, Leipzig University, Johannisallee 21-23, 04103, Leipzig, Germany. [5]Max Planck Institute for Biogeochemistry, PO Box 100164, 07701, Jena, Germany. ✉e-mail: freiberg@uni-leipzig.de

	TPL	LCVP
Plant names	**1,166,054**	**1,315,562**
Accepted species names (excl. infraspecific taxa)	**308,397**	**351,180**
Synonyms	**760,744**	**846,279**
Accepted infraspecific names	20,719	48,257
Accepted Genera	12,660	13,460
Accepted Families	473	564
All Genera	22,830	32,986
Unresolved names	244,017	63,072
TPL - LCVP comparison	**Number of records**	
Identical names	841,151	
Additional names in LCVP not being in TPL	149,462	
Families in LCVP not being in TPL	96	
Families in TPL not being in LCVP	8	
Genera in LCVP not being in TPL	2,764	
Genera in TPL not being in LCVP	1,954	
Resolved in LCVP - unresolved in TPL	172,354	
Accepted in LCVP - Synonym in TPL	60,321	
Synonym in LCVP - Accepted in TPL	177,570	
Different authors in LCVP	27,824	
Different synonym in LCVP	64,703	
Different orthography in LCVP	4,962	

Table 1. Summary of information content in the LCVP and name differences between LCVP and TPL.

researchers often rely on taxonomic reference lists as tools to translate taxa names to accepted species names via accepted synonyms.

Comprehensive taxonomic lists, such as the LCVP[2], are essential to standardize names in databases compiled from various sources, relying on a robust 'translation' of species names into one scheme. The TRY database of functional plant traits (TRY[3]; www.try-db.org) is one of the most prominent examples containing trait information for about 150,000 vascular plant species. Other global databases using plant name reference lists focus on plant co-occurrence patterns, such as sPlot containing about 1,1 million vegetation surveys ([4]~55,000 species), or use any plant species occurrence information, such as the Global Biodiversity Information Facility (~315,000 vascular plant species; www.gbif.org), of the Botanical Information and Ecology Network (BIEN[5]: ~348,000). The Global Inventory of Floras and Traits (GIFT[6]: ~268,000; http://gift.uni-goettingen.de/home) or the inventory of the Global Naturalized Alien Flora (GloNAF[7]~14,000; glonaf.org) focus on plant distribution information from regional floras or floristic inventories.

Generally, such databases were compiled from heterogeneous data sources varying in time of publication and place of origin. The underlying sources may be primary or secondary literature - using work of scientists with excellent to no plant taxonomic background, thus combining data with various degrees of complexity and uncertainty. The merging of these databases works via species identities and thus depends on the use of accepted species names. These databases typically tap phylogenetic information contained in taxonomic references lists via available tools supporting automated matching and error checking (i.e. taxon scrubbing). There is a variety of R packages (e.g. taxonstand[8]; taxize[9]; RBIEN[10]) or online tools (e.g. Global Name Resolver http://resolver.globalnames.org/ or the Taxonomic Name Resolution Service[11] http://tnrs.iplantcollaborative.org/TNRSapp.html) supporting researchers to check their taxonomic information (see[12] for a review on some of those tools). However, most of these tools rely on TPL as a reference list, which has not been updated for almost a decade and originated in a time when phylogenetic information on many genera did not exist.

Global taxonomic name databases are useful in their own right, and jointly create synergies that have transformed ecology into a synthetic and global science, and can help identifying knowledge gaps[13]. For example, functional biogeography combines information on community composition, plant species distribution and functional traits of the component species to make inferences on determinants of global trait distribution[14]. While there is high potential for exciting research using up-to-date taxonomic information, it can be only as good as the input data and the ability of the user to understand the advantage and shortcoming of the data coming from those resources. For example, missing taxonomic background often leads to neglecting the importance of citing authors of names and inevitably leads to inconsistencies when data from different sources are matched. LCVP[2] shows that when matching plant taxonomic names without author names, results could have up to 10% mismatches (i.e. ~10% of all LCVP plant taxa names are identical but ultimately refer to different accepted plant taxa).

Methods

The creation of the LCVP involved three major steps. (1) We did a thorough search of available and relevant plant taxonomic databases (Online-only Table 1) to collate a raw data table of existing plant names (see Step 1: Producing the raw data table). This table included many contradictory opinions in taxonomic placement of species. (2) Based on additional information in ~4,500 publications and the reliability, timeliness and quality of

relevant scientific evidence in this literature we, decided for each name, whether that name is in LCVP accepted, synonymous or unresolved (see for more details Step 2: Decision making). Additionally, we harmonized and corrected taxonomic names orthographically. (3) We implemented the LCVP in an R package (LCVP) which is accessible under a MIT license from GitHub (https://github.com/idiv-biodiversity/LCVP) and will ensure a coherent versioning of the list and future updates. Furthermore, we provide a utility function to use LCVP for taxonomic name resolution (lcvplants), which is also available under the same license from GitHub (https://github.com/idiv-biodiversity/lcvplants).

Step 1: Producing the raw data table.   TPL provided the core of the raw data table for published vascular plant names, primarily supplemented by the International Plant Names Index (IPNI, https://www.ipni.org/). IPNI provides a list of published names and their source, but does not provide any information on accepted or synonymous names. We used additional major and minor databases (see Online-only Table 1 and http://www.ville-ge.ch/musinfo/bd/cjb/africa/recherche.php?; http://gentian.rutgers.edu/classNEW123.htm; http://botany.si.edu/gesneriaceae/checklist/result.cfm; http://www.systax.org/; https://parasiticplants.siu.edu/ListParasites.htm; https://floramalesiana.org/new/; http://www.plantsoftheworldonline.org/; www.catalogueoflife.org/annual-checklist/2019; http://www.omnisterra.com/bot/cp_home.cgi; https://rbg-web2.rbge.org.uk/diptero/diptax.html; https://compositae.landcareresearch.co.nz/; http://www.cvh.ac.cn/cvh6/view/index.php; http://cichorieae.e-taxonomy.net/portal/; http://ww2.bgbm.org/EuroPlusMed/query.asp; http://floradobrasil.jbrj.gov.br/reflora/listaBrasil/ConsultaPublicaUC/ConsultaPublicaUC.do#CondicaoTaxonCP; http://www.melastomataceae.net/MELnames/; https://plants.usda.gov/java/; https://collections.nmnh.si.edu/search/botany/; http://posa.sanbi.org/sanbi/Explore; http://palmweb.org/) which we have chosen based on their availability, on our expert judgement on comprehensiveness, and whether they contained information if taxa names are accepted or not (see Online-only Table 1 for a table of used databases). All additional names and potential synonyms found in those databases were incorporated in the raw data table.

Step 2: Decision making.   The raw data table with more than two million entries of plant taxa names contained a high number of orthographic errors, inconsistencies and contradictory opinions concerning the status of the names. A rough guideline for the acceptance of names was a subjective assignment of quality and reliability to the source. Generally, changes were only applied when the authors of the respective publications were clearly suggesting those changes. We ascribed a higher reliability rank (e.g. for conflicting information) usually to the most recent publications. Additionally, when conflicting information appeared we usually used information from publications with a) a more thorough literature section and b) a more comprehensive synonymy history than to those without. A complete synonymy history should include and properly cite not only the latest accepted taxon, but also the depending taxonomic history of all names connected to this taxon (e.g. if it is a recombined taxon) with all homonymic (i.e. species epitheton is the same) and heteronymic (i.e. genus name is the same) synonyms. Since phylogenies based on morphological data alone are prone to homoplasy, only phylogenetic studies that made taxonomical decisions also based on molecular data were taken into account. We did not create new species name combinations. In case of conflicting evidence on the phylogenetic placement or species name, due to e.g. different methods to build phylogenetic trees, species names were marked "comb.ined." following the basionym author.

The following examples illustrate how we treated name changes: The genus *Dracaena* and *Sansevieria* are closely related[15], where *Sansevieria* seems to be clearly nested within *Dracaena*, but the differences between both genera are continuous. Lu *et al.*[15] separated the Hawaiian species of *Dracaena* in a new genus *Chrysodracon*, but did not recombine *Sansevieria* with *Dracaena* yet. The presented argumentation and data in[15] were thorough and comprehensive and thus we accepted the authors arguments, kept *Sansevieria* and *Dracaena* as distinct genera and separated the Hawaiian species of *Dracaena* in the new genus *Chrysodracon*. In another case Borchsenius *et al.*[16] showed that *Calathea* in the traditional description was polyphyletic. In order to keep *Ischnosiphon* and *Monotagma* as distinct genera, being the sister clade to a smaller *Calathea* clade including the type species, the larger clade of *Calathea* was put into the then resurrected genus *Goeppertia*. The argumentation and presentation in[16] was robustly based on a molecular phylogeny producing well supported clades. As a consequence, we accepted the recombination of the much larger clade as suggested in[16].

We also applied changes to the spelling of species names. Generally, we recommend to check the species names prior to automated list treatments, following the guidelines given in[17] and the rules of the current version of the International Code of Nomenclature for algae, fungi, and plants (Shenzhen Code[18]). We followed the Shenzen Code using standardized orthography of epitheta across genera and families, e.g. *warscewiczii* (neither *warscewitzii* nor *warszewiczii*). Only upper cases from 'A' to 'Z', lower cases from 'a' to 'z' and the hyphen '-' should be used in scientific names, special characters are not valid and to be avoided (Isoëtes- > Isoetes, Köberlinia - > Koeberlinia). Authors were given in their short form as provided by IPNI. For further standardization and easier use in automated workflows, we omitted spaces within author names (C. F. W. Meissn., C.F. W. Meissn., C. F.W. Meissn. C. F. W.Meissn. - > C.F.W.Meissn.; Balf. f.- > Balf.f.). We linked names published by two authors with the '&' sign (e.g. *Primula minor* Balf.f. & Kingdon-Ward). Names published by three and more authors were restricted to the first authors followed by '& al.' (e.g. *Limonium irtaense* P.P.Ferrer, A.Navarro, P.Pérez, R.Roselló, Rosselló, M.Rosato & E.Laguna - > *Limonium irtaense* P.P.Ferrer & al.). This refers to the recommendation of the Shenzhen Code, Art. 46 c. We tried to include only natural hybrids (i.e. no cultivars; based on expert judgement of LCVP authors) in the LCVP. Since hybrids were not the focus of the LCVP, we only marked them with '_x', either following the genus name or the epitheton to recognize them as such, but we did not give any parent taxa information.

In most cases, we adopted the names used by the taxonomic expert (i.e. reference author who is usually a person with a publication record within a certain taxonomic group). However, there are many taxa belonging to

genera or species which have not been phylogenetically analyzed yet. For those, we adapted the most frequently used taxon name from the recent literature. Despite a major effort, there are still names, which we could not resolve.

As part of the LCVP data package we also provide at https://idata.idiv.de/ddm/Data/ShowData/1806 three different files related to the used literature that we used to decide upon species names to create LCVP. We provide a complete bibliography (as.bib file and as full text pdf) of all ~4,500 literature references ordered by plant families. We focused on literature published from 1994 onwards, when molecular phylogenies became widespread[19,20]. The third file is a table directly matching >104,000 individual taxa and literature, used to inform the applied name changes for the respective taxa.

Step 3: Implementation in R. Besides providing LCVP as downloadable text table[2] with this article, we also provide LCVP as R package for easy integration with analyses pipelines. Due to the large size of the data we provide a pure data package, LCVP, and a separate tool package, lcvplants, with a fuzzy matching algorithm for taxonomic name resolution. Both can be downloaded and installed via github. The LCVP data package solely contains three files: the dataset of plant names and their taxonomic status, a package of the literature references used to compile the list (consisting of three files) and a meta data description file. The lcvplants package contains one user-level function to perform a fast fuzzy matching for taxonomic name resolution using the LCVP data[2]. This taxonomic names resolution is implemented in a user-friendly way, and can be done with few lines of code (see https://idiv-biodiversity.github.io/lcvplants/articles/taxonomic_resolution_using_lcplants.html for a tutorial):

```
```
# install LCVP and lcvplants from GitHub
install.packages("devtools")
library(devtools)
devtools::install_github("idiv-biodiversity/LCVP")
devtools::install_github("idiv-biodiversity/lcvplants")
# load the package
library(lcvplants)
# run analyses
LCVP("Hibiscus vitifolius")
""
```

*Input data.* For taxonomic name resolution an individual name or a vector of names can be provided. There are no limits on the number of names submitted at a time, but we recommend to submit less than 5000 names at a time to ensure a reasonable computation time. For the input data, following the International Code of Nomenclature for algae, fungi, and plants (Shenzhen Code: https://www.iapt-taxon.org/nomen/main.php), genus, epithet, infraspecies rank, infraspecies name and authorities need to be separated by spaces (e.g. *Draba mollissima* var. *kusnezowii* N.Busch). Special characters (such as ü, á, ø, etc.) are only allowed for the authority names. Infraspecific names have to be preceded by their rank (e.g. "subsp.", "var.", "forma", "ssp.", "f.", "subvar.", "subf."). The genus name and the epitheton need to be provieded; the infraspecific ranks and authority names are optional for better results. If the genus or the epitheton are composed of two words, they have to be separated by a hyphen (e.g. *Hibiscus rosa-sinensis* L.). Hybrid names use the characters '_x' at the end of the genus and epithet name (e.g. *Spartocytisus*_x *filipes* Webb & Berthel., *Lycopodium habereri*_x House) annotations in other formats such as 'x' or 'x_' before the names are automatically changed into the required format. The commonly used special Unicode Character 'x' (U + 00D7) for indicating hybrids is not accepted (e.g. *Crassocephalum* x *picridifolium*).

*Fuzzy matching.* The lcvplants package performs a string comparison between the user-submitted names and LCVP using a fuzzy matching algorithm to solve orthographic errors. The fuzzy matching algorithm can be applied to the genus name, the epitheton, the infraspecific names and the authority (see Online-only Table 2 for a description of the options for customization), and runs in the following order:

(1) Submitted name standardization. The submitted name is standardized into parts using a space as delimiter: The genus level (first word) and the epitheton (second word). If there are more than three words in the submitted name and the third word is any of: "subsp.", "var.", "forma", "ssp.", "f.", "subvar." or "subf." the fourth term will be recognized as the infraspecies name. Otherwise all the words after the epitheton will be recognized as authority description.
(2) Genus resolution with a user-specified threshold of allowed mismatches (i.e. the number of letters that can disagree between submitted and matched name).
(3) Epitheton resolution. If a match for the submitted genus name is found, a similar matching will be done to find the correct epitheton.
(4) Infraspecific name and authority resolution. If genus and epitheton resolution were successful, the fuzzy matching will be applied also for infraspecific names and authority names (if supplied).
(5) The results for all submitted names will be combined into the output table and the results will be returned by the function and printed to the screen.

*Output data.* The output is a data.frame of the submitted and matched taxon names with additional information on the taxonomic status. If the option 'save' is turned active (Save = TRUE), the output will additionally be saved in a comma-separated file (.csv) in the working directory or the path specified with the 'out_path' option. The

following list describes the columns of the output table. If a name could not be resolved, in the LCVP the respective row in the output data.frame is empty except for the 'Submitted_Name' and the 'Score' field, which gives detail information in which parts of the name could not be matched. See Online-only Table 3 for a description of the output fields.

## Data Records

LCVP[2] contains 1,315,562 vascular plant names with 351,180 accepted species names (405,687 including infraspecific taxa) and 846,279 synonyms (Table 1). The accepted species in LCVP belong to 13,460 genera, 564 families, and 84 orders. LCVP significantly reduced the number of unresolved plant names by ~181,000 to ~63,000 (5%) taxa compared to TPL (Table 1).

## Technical Validation

We tested whether all synonyms lead to an accepted name or another synonym. One major issue with TPL is the high amount of unresolved names. A link to another name sometimes is another synonym leading to unresolved loops. LCVP only links to accepted names, not to the taxonomic predecessor. If taxon A is synonym to taxon B and it turned out, that taxon B is synonym to taxon C, the accepted name given for taxon A is taxon C, not B. We treated invalid names as synonyms and assigned them to their appropriate accepted name.

Most of the still unresolved species names in LCVP were originally published in the 19th century. There is a high probability that the majority of them are synonyms, e.g. because of historic transfer errors from one publication to the other. An extraordinarily high amount of unresolved names can be found in Asteraceae (in *Hieracium* 5,781 out of 19,300 names are unresolved, *Senecio* 682 out of 6,684, *Cirsium* 353 out of 2,162), Rosaceae (*Rubus* 4,005 out of 10,199, *Rosa* 2,298 out of 5,965, *Prunus* 509 out of 2,072, *Potentilla* 724 out of 3,954, *Crataegus* 720 out of 2,717, *Pyrus* 373 out of 1,199), Salicaceae (*Salix* 610 out of 3,867), Araceae (*Anthurium* 582 out of 2,261), and Geraniaceae (*Pelargonium* 962 out of 1,846).

**Comparison to TPL.** Due to the improved name resolution and increased name information in general in LCVP compared to TPL, any work flow including taxonomic harmonization of plant names, will very likely yield more robust and reliable results for e.g. species richness patterns and matches between different data sources. For an easier comparison between LCVP and TPL, LCVP includes information whether taxa name entries are identical, differ in the cross-reference to a synonym, differ only orthographically either by the name or the author, or whether a name is new in the LCVP and not present in TPL. This unique information makes it possible for the users of TPL to update their names according to the LCVP, because all differences are clearly stated in the column 'status' of the LCVP.

Kew Gardens´ research effort to standardize plant names recently focuses on their new flagship program, Plants of the World Online (POWO, http://www.plantsoftheworldonline.org/), which includes a new taxonomic reference backbone (Alan Paton from Kew Gardens, pers. comm. July 2019). Given that this is becoming the successor of TPL (see http://www.plantsoftheworldonline.org/about) we also compared the available POWO list with LCVP (POWO access date: November 2018; directly provided by Kew). With ~335,000 accepted species names and ~458,000 names of vascular plants marked as synonyms in this POWO version, LCVP contains also significantly more species name information than POWO (this comparison includes only vascular plants and excludes infraspecific taxa since LCVP covers only vascular plants and this POWO version does not include taxa below species level).

TPL and the tested POWO version cover all plants, LCVP only vascular plants. With the current information we have, LCVP contains more information about vascular plant names (e.g. more resolved names, more accepted species, more synonyms) than TPL and POWO. A user is more likely to resolve a given vascular plant name with LCVP than with the given versions of TPL and POWO. Any future updated versions of LCVP and POWO will change these numbers and might strengthen different purposes of use for each reference list, and could ideally lead to a harmonized global backbone if applicable. LCVP covers also infraspecific names which are not covered in the tested POWO version. The information in LCVP to which genus a species belongs and/or thus which accepted name should be used, is based on taxonomic, but also on most recent phylogenetic (i.e. mainly genetic) information. TPL was not updated for many years, and is mainly based on taxonomic information (i.e. not molecular phylogenies). With respect to usability of LCVP, we do see advantages compared to the POWO version we tested, which to our knowledge does not offer an R package nor any other functionality of (half)automatic name checking or any fuzzy name matching functions.

## Code availability

The LCVP generally consists of (1) the LCVP itself, available as R data package (version 1.0.3 as of July 2020) and as tab-delimited textfile file and (2) the R-package lcvplants. The LCVP version 1.0.3 is available in both Microsoft Excel and text formats in the iDiv data portal (https://idata.idiv.de/ddm/Data/ShowData/1806; https://doi.org/10.25829/idiv.1806-40-3009). A developmental version of the LCVP and the lcvplants package are publicly available via GitHub (https://github.com/idiv-biodiversity/lcvplants). We will constantly update the LCVP and plan to release a new version once every second to third year. We plan to closely collaborate with plant synonymy services and tools like e.g. BIEN, GNR, R packages taxonstand and taxize, to include LCVP as reference option. Requests for integrating LCVP can be made via the projects GitHub (https://github.com/idiv-biodiversity/LCVP/issues).

## References

1. Rouhan, G. & Gaudeul, M. Plant Taxonomy: A Historical Perspective, Current Challenges, and Perspectives. *Methods Mol Biol* **1115**, 1–37 (2014).
2. Freiberg, M. *et al.* Leipzig Catalogue of Vascular Plants (LCVP) and R Package to Query Data. *German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig* https://doi.org/10.25829/idiv.1806-40-3009 (2019).
3. Kattge, J. *et al.* TRY - a global database of plant traits. *Global Change Biology* **17**, 2905–2935 (2011).
4. Bruelheide, H. *et al.* sPlot - A new tool for global vegetation analyses. *J Veg Sci* **30**, 161–186 (2019).
5. Enquist, B., Condit, R., Peet, R., Schildhauer, M. & Thiers, B. Cyberinfrastructure for an integrated botanical information network to investigate the ecological impacts of global climate change on plant biodiversity. Preprint at https://peerj.com/preprints/2615/ (2016).
6. Weigelt, P., König, C. & Kreft, H. GIFT – A Global Inventory of Floras and Traits for macroecology and biogeography. *Journal of Biogeography* **47**, 16–43 (2019).
7. van Kleunen, M. *et al.* The Global Naturalized Alien Flora (GloNAF) database. *Ecology* **100**, e02542 (2019).
8. Cayuela, L., Granzow-de la Cerda, I., Albuquerque, F. S. & Golicher, D. J. TAXONSTAND: An R package for species names standardisation in vegetation databases. *Methods Ecol Evol* **3**, 1078–1083 (2012).
9. Chamberlain, S. & Szöcs, E. taxize: taxonomic search and retrieval in R. *F1000Research* **2**, 191 (2013).
10. Maitner, B. S. *et al.* The BIEN R package: A tool to access the Botanical Information and Ecology Network (BIEN) database. *Methods Ecol Evol* **9**, 373–379 (2018).
11. Boyle, B. *et al.* The taxonomic name resolution service: an online tool for automated standardization of plant names. *Bmc Bioinformatics* **14**, 16 (2013).
12. Wagner, V. A review of software tools for spell-checking taxon names in vegetation databases. *J Veg Sci* **27**, 1323–1327 (2016).
13. Cornwell, W. K., Pearse, W. D., Dalrymple, R. L. & Zanne, A. E. What we (don't) know about global plant diversity. *Ecography* **0** (2019).
14. Bruelheide, H. *et al.* Global trait-environment relationships of plant communities. *Nat Ecol Evol* **2**, 1906–1917 (2018).
15. Lu, P.-L. & Morden, C. W. Phylogenetic Relationships among Dracaenoid Genera (Asparagaceae: Nolinoideae) Inferred from Chloroplast DNA Loci. *Systematic Botany* **39**, 90–104, 115 (2014).
16. Borchsenius, F., Suárez, L. S. S. & Prince, L. M. Molecular Phylogeny and Redefined Generic Limits of *Calathea* (Marantaceae). *Systematic Botany* **37**, 620–635, 616 (2012).
17. Stearn, W. T. *Botanical Latin*. (Timber Press, 2004).
18. Turland, N. J. *et al. International Code of Nomenclature for algae, fungi, and plants (Shenzhen Code) adopted by the Nineteenth International Botanical Congress Shenzhen, China, July 2017*. (Koeltz Botanical Books, 2018).
19. Soltis, P. S. & Soltis, D. E. In *Evolutionary Biology* (eds. Max, K. Hecht, Ross, J. Macintyre, & Michael, T. Clegg) Ch. Plant Molecular Systematics - Inferences of Phylogeny and Evolutionary Processes 139–194 (Springer US, 1995).
20. Wolf, P. G., Soltis, P. S. & Soltis, D. E. Phylogenetic-Relationships of Dennstaedtioid Ferns - Evidence from Rbcl Sequences. *Mol Phylogenet Evol* **3**, 383–392 (1994).
21. Aedo, C. In *Species 2000 & ITIS Catalogue of Life, 2018 Annual Checklist* (eds Y. Roskov *et al.*) (Species 2000: Naturalis, 2018).
22. Calonje, M., Stanberg, L. & Stevenson, D. In *Species 2000 & ITIS Catalogue of Life, 2018 Annual Checklist* (eds Y. Roskov *et al.*) (Species 2000: Naturalis, 2015).
23. Culham, A. & Yesson, C. In *Species 2000 & ITIS Catalogue of Life, 2018 Annual Checklist* (eds Y. Roskov *et al.*) (Species 2000: Naturalis, 2018).
24. Farjon, A., Gardner, M. & Thomas, P. In *Species 2000 & ITIS Catalogue of Life, 2018 Annual Checklist* (eds Y. Roskov *et al.*) (Species 2000: Naturalis, 2018).
25. Govaerts R. (ed.) In *Species 2000 & ITIS Catalogue of Life, 2018 Annual Checklist* (eds Y. Roskov *et al.*) (Species 2000, 2018).
26. Hassler, M. In *Species 2000 & ITIS Catalogue of Life, 2018 Annual Checklist* (eds Y. Roskov *et al.*) (Species 2000: Naturalis, 2018).
27. Kiefer, M. *et al.* BrassiBase: Introduction to a Novel Knowledge Database on Brassicaceae Evolution. *Plant and Cell Physiology* **55**, e3–e3 (2013).
28. Koch, M. A., German, D. A., Kiefer, M. & Franzke, A. Database Taxonomics as Key to Modern Plant Biology. *Trends in Plant Science* **23**, 4–6 (2018).
29. Maslin, B. In *Species 2000 & ITIS Catalogue of Life, 2018 Annual Checklist* (eds Y. Roskov *et al.*) (Species 2000: Naturalis, 2018).
30. Rainer H. & Chatrou L.W. (eds). In *Species 2000 & ITIS Catalogue of Life, 2018 Annual Checklist* (eds Y. Roskov *et al.*) (Species 2000: Naturalis, 2018).
31. Roskov, Y., Zarucchi, J., Novoselova, M. & Bisby F. (eds). In *Species 2000 & ITIS Catalogue of Life, 2018 Annual Checklist* (eds Y. Roskov *et al.*) (Species 2000: Naturalis, 2018).
32. Vattakaven, T. *et al.* India Biodiversity Portal: An integrated, interactive and participatory biodiversity informatics platform. *Biodiversity Data Journal* **4**, e10279 (2016).
33. Warwick, S. I., Francis, A. & Al-Shehbaz, I. A. In *Species 2000 & ITIS Catalogue of Life, 2018 Annual Checklist* (eds Y. Roskov *et al.*) (Species 2000: Naturalis, 2018).

## Acknowledgements

## Author contributions

M.F. compiled the LCVP. A.G., M.W. & A.Z. designed the R packages. A.G. & A.Z. implemented the R functions based on discussions with M.W. & M.F. M.W., M.F. & A.Z. compiled the drafts of the data paper. All authors contributed to the writing of the manuscript.

## Competing interests

The authors declare no competing interests.

## Additional information

**Correspondence** and requests for materials should be addressed to M.F.

**Reprints and permissions information** is available at www.nature.com/reprints.

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.