Data in Brief 42 (2022) 108090

Contents lists available at ScienceDirect

Data in Brief

journal homepage: www.elsevier.com/locate/dib

Data Article

rbcL gene in global molecular data repository

Conrad Omonhinmin[*], Chinedu Onuselogu

Department of Biological Science Biotechnology Cluster, College of Science and Technology, Covenant University, Canaan land Ota, Ogun State, Nigeria

ARTICLE INFO

Article history:
Received 12 August 2021
Revised 10 February 2022
Accepted 17 March 2022
Available online 23 March 2022

Dataset link: rbcL Gene in Global Molecular Data Repository (Original data)

Keywords:
rbcL gene
Evolutionary
Biogeography
Phylogeny
Molecular repository

ABSTRACT

The ribulose-bisphosphate carboxylase (*rbcL*) gene sequence data in the molecular data repository has been increasing significantly, over the years with contributions from different parts of the world. The abundance of the gene has enhanced its applications in several ways. Bulk records were obtained from National Center for Biotechnology Information (NCBI) GenBank database using the entrez efetch utility as implemented in the Biopython package version 1.77. Records corresponding to the following keywords "*rbcL* AND plants [filter] AND biomol_genomic [PROP] AND is _nuccore [filter]" were created. Generated records were cleaned and then further analysed using the code file in the supplementary materials. Country information was obtained by searching reference information for matches to countries present in the pycountry package. Where no match was found, null was returned. This data article contains information about the plant family and species whose *rbcL* gene sequence has been deposited on the NCBI and regions of the world that has contributed to the *rbcL* repository growth. This data can be used to analyse the intra and inter family relatedness of plant and compare with existing relationships the molecular characterization of plants, evolutionary relationship studies, inferring biogeography origin of plant.

[*] Corresponding author.
E-mail address: conrad.omonhinmin@covenantuniversity.edu.ng (C. Omonhinmin).

https://doi.org/10.1016/j.dib.2022.108090

Specifications Table

Subject	Biological sciences
Specific subject area	Molecular phylogenetics, Phylogeny and Evolution
Type of data	Text, Table, Chart, Figure
How data were acquired	Biopython package version 1.77. was used to retrieve the *rbcL* gene sequence data from the NCBI GenBank. The written code used for retrieving the data from the NCBI GenBank can be assessed in the supplementary materials.
Data format	Raw, Analysed and Filtered.
Description of data collection	Bulk data were obtained from NCBI GenBank database using the entrez efetch utility as implemented in the Biopython package version 1.77. Datasets that do not have the matching words *rbcL*, Plant and DNA were filtered off from the data leaving behind data with the keywords *rbcL*, plant and DNA.
Data source location	The data was obtained from the NCBI GenBank database.
Data accessibility	With the article.
Repository name	Mendeley Data
Data identification number	10.17632/wdmtpnwsrn.1
Direct link to the dataset:	http://www.rbcLGeneinGlobalMolecularDataRepository.com

Values of the Data

- This data present information of plant species, phylum, and family for which *rbcL* gene sequence have been deposited on NCBI GenBank.
- Molecular systematics can use the data to renew the relatedness of plants both within and between families as well as compare with existing relationships.
- This data is useful in the following field: molecular characterization of plants, evolutionary relationship studies, inferring biogeography, origin of plant, codon bias usage profile, protein structure analysis, ecological preference studies.
- This data can be used to determine the pattern of growth of *rbcL* gene sequence from different regions in the molecular repository.
- This data shows the least explore plant species and the need for exploitation.

1. Data Description

The data in this article gives an overview of the total number of plant species, families, with *rbcL* gene sequence in the GenBank and the regions that has contributed to the growth of the *rbcL* sequence in the repository. The sequence data of the *rbcL* gene are used for renewal of phylogenies among the seed plants [1]. The *rbcL* gene is preferred among other plant genes for phylogenetic studies due to its slower rate of evolutionary changes and the lowest divergence among the plastid genes in flowering plants [2,3]. [4] described the suitability of the gene for solving intergeneric and interspecific relationship and no difficulties of alignment. Some of the applications of the gene in the molecular investigations of plant species include: tracing of the molecular origin of plants [5], the biogeography origin of plants [6]. The datasets used, in the study was collected as a secondary data and the Bio python code written for data collection can be assessed as Supplementary data, the *rbcL* gene data used was obtained from the first report till 2020. Fig. 1, shows the most studied plant families with *rbcL* gene on the GenBank. Fig. 2; shows plant phyla with *rbcL* gene sequence and the extent to which the sequences have been utilized for *rbcL* related studies. The continents with *rbcL* sequence submission and the percentage of contribution to the GenBank is represented in Figs 3, 4 and 5 shows countries with higher *rbcL* sequence submission on GenBank. The plant species and other species with *rbcL* gene sequence can be assessed in the supplementary materials.

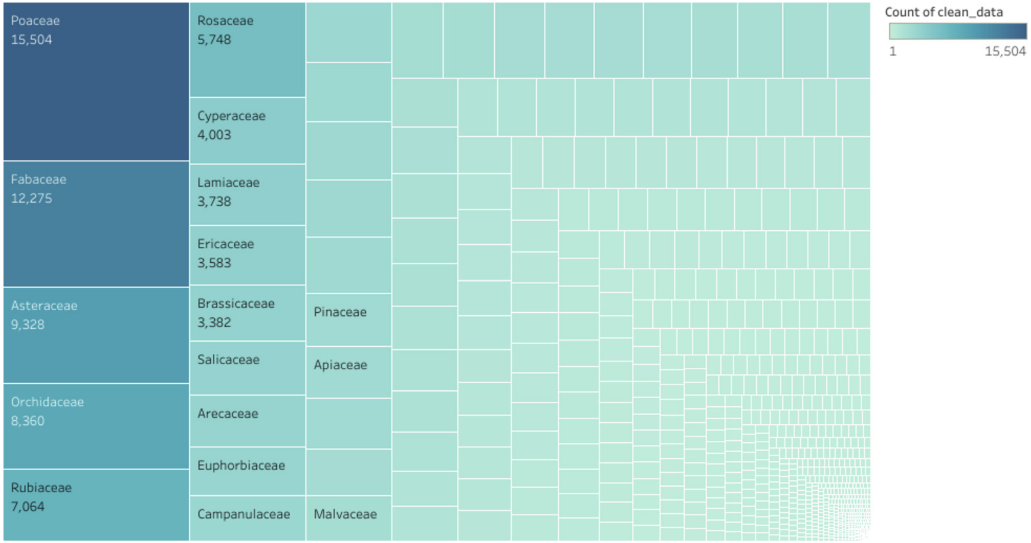

Fig. 1. Most studied plant families with *rbcL* gene sequence in GenBank.

*The numbers indicate the number of species in each family with *rbcL* gene deposited on NCBI GenBank.

*NB: The study discovered a total number of 808 plant families with *rbcL* gene sequence submitted on the NCBI GenBank making it difficult to include all the families in the tree map in Fig. 1, hence the plant families with the most *rbcL* gene submission are mentioned in Fig. 1.

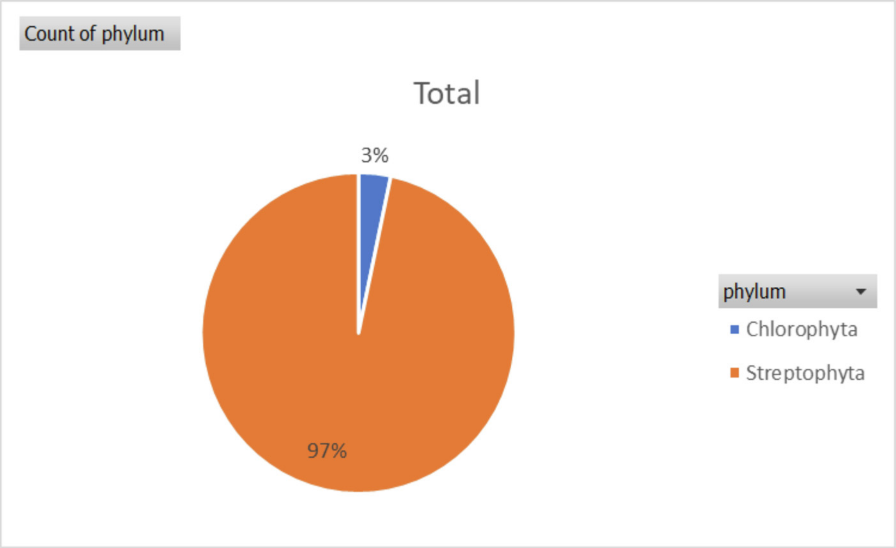

Fig. 2. Percentage of plant phyla with *rbcL* gene data deposited on GenBank.

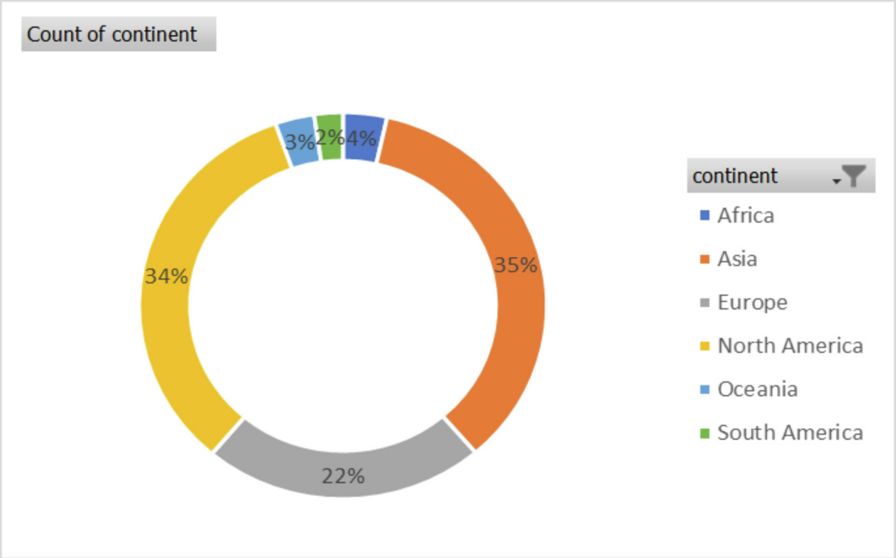

Fig. 3. Percentage of *rbcL* sequences contribution from different regions.

2. Experimental Design, Materials and Methods

Bulk records were obtained for NCBI GenBank database using the entrez efetch utility as implemented in the Biopython package version 1.77. Records corresponding to the following keywords “*rbcL* AND plants[filter] AND biomol_genomic[PROP] AND is_nuccore[filter]” were obtained. Obtained records were cleaned and then further analysed using the codes files in the supplementary material. Country information was obtained by searching reference information

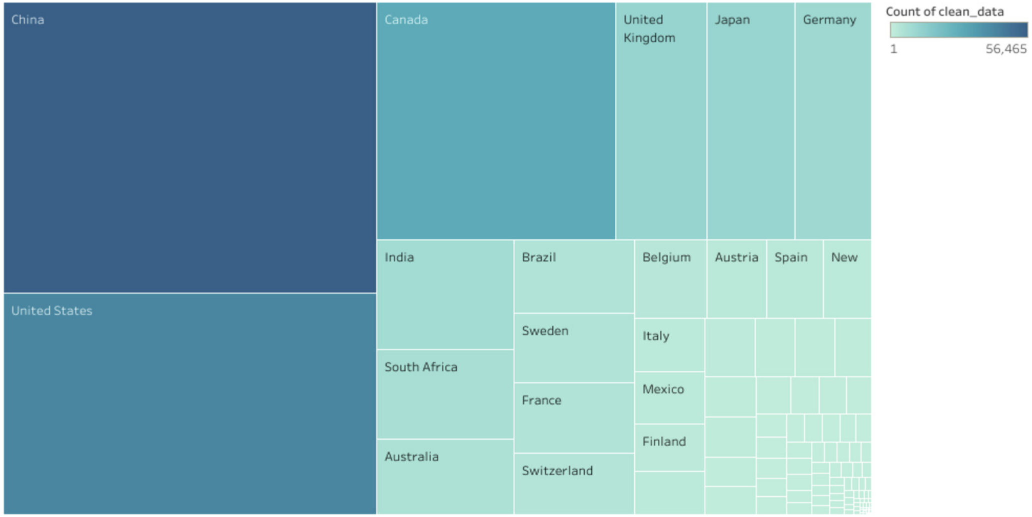

Fig. 4. Countries with higher submissions of *rbcL* sequences on the GenBank repository.

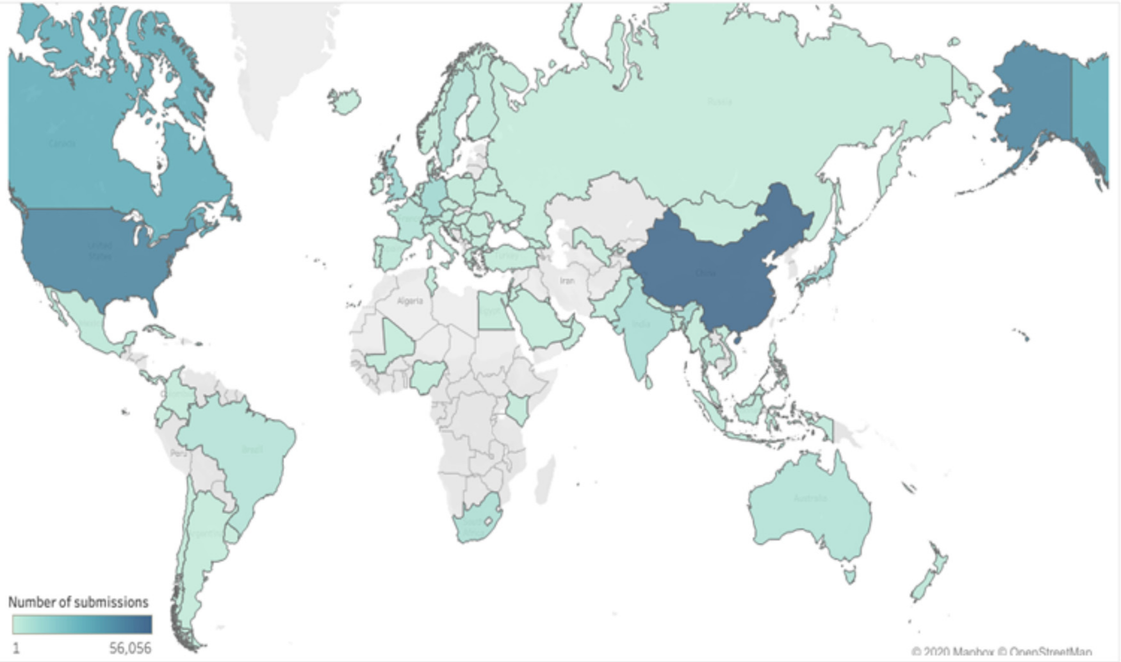

Fig. 5. Map showing global concentration of *rbcL* sequence contribution to GenBank repository.
* Regions with dark blue has higher contribution of *rbcL* gene sequence on the NCBI GenBank.

for matches to countries present in the pycountry package. Where no match was found, null was returned.

CRedit Author Statement

Conrad Omonhinmin: Conceptualization, Methodology, Validation and Supervision; **Chinedu Onuselogu:** Data curation, Investigation, software, Reviewing and Editing, Writing-Original draft preparation.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships which have or could be perceived to have influenced the work reported in this article.

Data Availability

rbcL Gene in Global Molecular Data Repository (Original data) (Mendeley Data).

Acknowledgments

The authors grateful to the Covenant University Centre for Research Innovation and Discovery (CUCRID) for the Publication Funding and Mr Bode Onile-ere of Biological Science, department, Covenant University for the assistance in data acquisition from the NCBI GenBank.

Supplementary Materials

Supplementary material associated with this article can be found in the online version at doi:10.1016/j.dib.2022.108090.

References

[1] A.M. Hashim, A. Alatawi, F.M. Altaf, S.H. Qari, M.E. Elhady, G.H. Osman, H.H. Abouseadaa, Phylogenetic relationships and DNA barcoding of nine endangered medicinal plant species endemic to Saint Katherine protectorate, Saudi J. Biol. Sci. 28 (3) (2021) 1919–1930.

[2] W.J. Kress, K.J. Wurdack, E.A. Zimmer, L.A. Weigt, D.H. Janzen, Use of DNA barcodes to identify flowering plants, Proc. Natl. Acad. Sci. 102 (23) (2005) 8369–8374.

[3] I. Ahmed, P.J. Biggs, P.J. Matthews, L.J. Collins, M.D. Hendy, P.J. Lockhart, Mutational dynamics of aroid chloroplast genomes, Genome Biol. Evol. 4 (12) (2012) 1316–1323.

[4] B.U. Reddy, Cladistic analyses of a few members of Cucurbitaceae using *rbcL* nucleotide and amino acid sequences, Int. J. Bioinf. Res. 1 (2009) 58–64.

[5] D. Soltis, P. Soltis, P. Endress, M. Chase, S. Manchester, W. Judd, E. Mavrodiev, Phylogeny and Evolution of the Angiosperms: Revised and Updated Edition, University of Chicago Press, 2018.

[6] A. Marion, A. Sfriso, C. Andreoli, I. Moro, The presence of exotic *Hypnea flexicaulis* (Rhodophyta) in the Mediterranean Sea as indicated by morphology, *rbcL* and cox1analyses, Aquat. Bot. 95 (1) (2011) 55–58.