Botanical Sciences 97 (4): 754-760. 2019
DOI: 10.17129/botsci.2226

Received: February 26, 2019, accepted: June 29, 2019
On line first: December 17, 2019

Taxonomy and Floristics/Taxonomía y Florística

Datataxa: a new script to extract metadata sequence information from GenBank, the Flora of Bajío as a case study

Datataxa: un nuevo script para extraer la información de los metadatos de secuencias de GenBank: la Flora del Bajío como un caso de estudio

Eduardo Ruiz-Sanchez[1,2], Carlos Alonso Maya-Lastra[3*], Victor W. Steinmann[4], Sergio Zamudio[5], Eleazar Carranza[6], Rosa María Murillo[5], and Jerzy Rzedowski[7]

[1] Departamento de Botánica y Zoología, Centro Universitario de Ciencias Biológicas y Agropecuarias, Universidad de Guadalajara. Zapopan, Jalisco, Mexico.
[2] Laboratorio Nacional de Identificación y Caracterización Vegetal (LaniVeg), Zapopan, Jalisco, Mexico.
[3]Department of Ecology, Evolution, and Environmental Biology, Columbia University, New York, New York USA.
[4]Facultad de Ciencias Naturales, Universidad Autónoma de Querétaro, Campus Juriquilla, Juriquilla, Querétaro Mexico.
[5]Investigador Independiente. Avellanos 4, Fraccionamiento Los Nogales,Pátzcuaro, Michoacán 61608 Mexico.
[6] Instituto de Investigación de Zonas Desérticas, Universidad Autónoma de San Luis Potosí, San Luis Potosí, Mexico.
[7]Centro Regional del Bajío, Instituto de Ecología, A.C., Pátzcuaro, Michoacán, Mexico.

*Corresponding author: camayal@gmail.com

Abstract

Background: GenBank is a public repository that houses millions of nucleotide sequences. Several software have been developed to extract information stored in GenBank. However, none of them are useful to extract and organize GenBank accession based on metadata. We developed a new script called Datataxa, which works to mine GenBank information. The checklist of the Flora del Bajío y de Regiones Adyacentes (FBRA) was used as a case study to apply our script.

Questions: How many species occurring in the FBRA have records in GenBank? What percentage of those records have been used for phylogenetic, phylogeographic, phylogenomic, barcoding, genetic diversity, and biogeographic studies?

Methods: Datataxa was written in AutoIt Scripting Language in order to facilitate the extraction of information from GenBank. This information was classified in six study categories. A checklist of species published fascicles of FBRA was used as study case to apply our new script, and the previous categories were applied to the FBRA species list.

Results: The script allowed us to search for meta information, like publication titles, for 2,558 species that were included in the FBRA. Of these, 1,575 had a least one record in GenBank. A total of 1,322 species were used in phylogenetic studies, followed by barcoding studies (326) and biogeographic studies (298). Phylogenomic (41), phylogeographic (34), and diversity studies (34) were the least represented.

Conclusions: Datataxa was useful for mining metadata sequence information from GenBank and can be used with any list of species to get the GenBank accessions' metadata.

Key words: API, checklist, entrez, floristic treatment, GenBank, vascular plants.

Resumen

Antecedentes: GenBank es un repositorio público de millones de secuencias nucleotídicas. Se han desarrollado varios programas para extraer la información almacenada en GenBank. Ninguno de ellos es útil para extraer y organizar información de los metadatos de las entradas de GenBank. Desarrollamos un nuevo script llamado Datataxa, que extrae metainformación de Genbank. El listado de la Flora del Bajío y de Regiones Adyacentes (FBRA) fue utilizado como caso de estudio, para probar nuestro script.

Pregunta: ¿Cuántas especies de la FBRA tienen registros en GenBank? y ¿Qué porcentaje de esos registros se han utilizado en estudios de filogenética, filogeografía, filogenómica, código de barras, diversidad genética y biogeografía?.

Métodos: Datataxa está escrito en lenguaje AutoIt Scripting Language para facilitar la extracción de información de GenBank. La información extraída de GenBank fue clasificada en seis categorías. La lista preliminar de especies de la FBRA fue utilizada como caso de estudio para aplicar nuestro script. Estas categorías fueron aplicadas a la lista de especies de la FBRA.

Resultados: El script nos permitió extraer y organizar la información de lo metadatos, como los títulos de publicación de 2,558 especies que están incluidas en la FBRA, 1,575 de esas especies tienen registros en GenBank. 1,322 fueron de estudios filogenéticos, seguido de código de barras (326) y biogeografía (298). Filogenómica (41), filogeografía (34) y diversidad genética (34), tuvieron menos representación.

Conclusiones: Datataxa trabajó muy bien extrayendo los metadatos de las secuencias de Genbank. Datataxa puede ser utilizado en cualquier lista de especies para extraer los metadatos de GenBank.

Palabras clave: API, checklist, entrez, GenBank, plantas vasculares, tratamiento florístico.

GenBank is a public database of nucleotide sequences and supporting bibliographic and biological annotations (Benson *et al*. 2006), from which it is free to download information. The number of sequences deposited in GenBank is massive. Release 230 (February 2019) contains 303.7 tn nucleotides in 212.2 M sequences, plus another 4.1 tn nucleotides in 945 M sequences of whole genome shotgun sequencing (https://www.ncbi.nlm.nih.gov/genbank/statistics/ accessed February 20, 2019). Additionally, GenBank also contains a taxonomic section for all taxa with sequences in the database. At present, GenBank has 577,801 sequenced taxa, of which 177,982 belong to the Viridiplantae lineage (https://www.ncbi.nlm.nih.gov/Taxonomy/taxonomyhome.html/index.cgi?chapter=STATISTICS&uncultured=hide&unspecified=hide accessed February 20, 2019).

Molecular DNA sequences are systematically deposited in GenBank, and they have been and continue to be used mainly to reconstruct the phylogenetic history of taxa (Soltis *et al*. 2004, Smith & Brown 2018). They are also employed for barcoding, a method allowing species-level identification using short DNA sequences (Hebert *et al*. 2003, Hajibabaei *et al*. 2007). The amount of sequence data available for use in barcode, biogeographic, genetic diversity, phylogenetic, phylogenomic, and phylogeographic analyses has been continuously increasing due to advances in DNA sequencing techniques (Sanderson & Driskell 2003).

One of the signs of progress is that genome sequencing has been used to reconstruct the Tree of Life (Soltis *et al*. 2004, Delsuc *et al*. 2005, Soltis *et al*. 2018). Genome sequences obtained from several methods of genome reduction are able to add thousands of more pair bases than traditional Sanger methods (McCormack *et al*. 2013). The use of genomic data coupled with phylogenetic principles has resulted in a new field of research termed phylogenomics (Eisen & Fraser 2003).

Even in the phylogenomic era, the floristic knowledge of ecoregions or particular geographical regions described in floras, checklists, and other publications, provides us with information about the global vascular plant flora. This floristic knowledge is used, for example, to understand global patterns of plant diversity and propose strategies for its conservation (Kier *et al*. 2005).

Several software have been developed to mine the massive amount of information stored in GenBank. Among those programs are PhyLoTA Browser (Sanderson *et al*. 2008), phylotaR (Bennett *et al*. 2018), and restez (https://www.rdocumentation.org/packages/restez/versions/1.0.0). However, until now none of these are useful for extracting GenBank accession metadata like the type of study, journal, or involved institution. Such information in some cases can be useful to focus research efforts in a particular field or report the importance of institutions or journals in communication, contribution and impact related with nucleotide sequences. Also, it can provide decision-making information for use by governmental or institutional research and conservation efforts.

In this paper, we focus on mining GenBank data to parse metadata from the information including in the accession. To help us to mine GenBank information, we developed a new script called Datataxa. As a case study, we used the published floristic knowledge of the Flora del Bajío y de Regiones Adyacentes (FBRA), an ongoing floristic project (Calderón de Rzedowski & Rzedowski 1991). Applying our Metadata script, we want to know how many species present in the FBRA have records in GenBank and what percentage of those records have been used for phylogenetic, phylogeographic, phylogenomic, barcoding, genetic diversity, and biogeographic studies. We selected those six categories because they are the most widely used in the metadata of GenBank access numbers. However, users can download different categories or expressions contained in the metadata.

Material and methods

Extracting the information from GenBank. A new script called Datataxa (Maya-Lastra 2019), that facilitates the extraction of information from GenBank, was written in AutoIt Scripting Language (Bennett 2015). For the extraction of information, an interface was created that recurrently does the following procedures with all names in the main list and then sorts the result into a readable table. Initially, the program searches for the correct spelling and taxonomic synonyms using the Espell utility in the Entrez Application Programming Interface (API, Maglott *et al*. 2005). API refers to an interface that allows the inclusion of functions from different applications in new scripts and software in order to utilize of previously developed technologies. In our case, the script makes a connection with the GenBank database through Entrez API. Subsequently, it searches using the Esearch utility in the same API, looking for the GenInfo Identifier (GI) accession number in the Nucleotide GenBank database associated with each species. For each GI accession number, the following information was retrieved using the Efetch utility: nucleotide length, definitions (title found in GenBank), paper titles, journal names, and creation date (GI accessions for each species in Appendix S1 see Supplemental Data with this article). The time required for the extraction of metadata from a list of 6,000 species is commonly six hours, but it can vary depending on the number of accessions found for each species, times and query limits of GenBank (since December 2018, the limitation is 3-10 requests per second), and internet connection speed, to mention a few factors. Other tests showed no apparent limit of the number of species included in the search. To facilitate long time execution, the script automatically saves progress in case of an interruption. In such instances, when it is running again, it restarts from the last saved point. The script can be downloaded at no cost from https://github.com/camayal/Datataxa; detailed instructions on how to use it can also be found in the Github repository. The only input file the user needs to provide is a plain text file with no heading, one species per line, and the generic and specific epithets separated by a "+". This script is only intended to work with species, and only this functionality was tested. However, it is not limited, and in theory will work for other taxonomic ranks, like genera or families. The output are two different comma separated value (CSV) files, one for each phase, extraction and metasearch.

Classification and analysis using search terms. The entire classification was divided into six categories to determine which species are included in the different types of research. For each category, a list of generalized terms (Table 1) was used to include variations like plural and singular; for diversity studies, a couple of terms were always used, as described in Table 1. Finally, the generalized terms were designed to search in English, Spanish, and Portuguese, and all searches were executed as case-insensitive. The search was performed in the paper titles where each sequence was used (Appendix S2; see Supplemental Data within this article).

Floristic study. Flora del Bajío y de Regiones Adyacentes (FBRA) is one of the most active ongoing floristic projects in Mexico. It was initiated in 1985 by Jerzy Rzedowski and Graciela Calderón, with the aim of publishing an inventory of the wild vascular plants of the region. After five years of intensive collecting to obtain herbarium specimens, the formal preparation and publication of the Flora started in 1991 (Calderón de Rzedowski & Rzedowski 1991). The vascular plant diversity calculated for this region consists of ca. 5,500 species, 1,205 genera, and 185 families (Calderón de Rzedowski & Rzedowski 1991), thus representing 23.5 % of the total Mexican plant diversity. To date, 203 fascicles have been published. Some species included in the FBRA also have been used in phylogenetic, phylogeographic, phylogenomic, barcoding, genetic diversity, and biogeographic studies, in addition to many other studies.

The study area. The geographical delimitation of the Flora del Bajío y de Regiones Adyacentes adheres to Calderón de Rzedowski & Rzedowski (1991) and corresponds to the traditional Bajío area of central Mexico and some adjacent regions. It includes the entire states Guanajuato and Querétaro, in addition to the northern portion of Michoacán. In total, it covers 60,171 km^2 (Figure 1), of which 49.8 % has been transformed or fragmented by anthropogenic activities (Suárez-Mota *et al*. 2015). Three physiographic provinces have shaped the study area: the Trans-Mexican Volcanic Belt, the Mexican Plateau, and the Sierra Madre Oriental (Ferrusquia-Villafranca 1993). The Trans-Mexican Volcanic Belt is a volcanic region that extends east-west across much of the study area (Gómez-Tuena *et al*. 2007, Ferrari *et al*. 2012). The Mexican plateau occurs only in a small north-west portion of the study area, whereas the Sierra Madre Oriental occupies the northeast portion (Ferrusquia-Villafranca 1993). Topographic diversity and ample climatic variation favor the development of both temperate and tropical vegetation in this region (Rzedowski 1978, 1991, Calderón de Rzedowski & Rzedowski 1991). Among the temperate vegetation present are pine forests, fir forests, oak forests, pine-oak forests, and montane cloud forests. The tropical vegetation is represented by tropical dry forests, subdeciduous tropical forests, and xerophytic scrubs (Rzedowski 1978).

Data obtainment and database curation. In order to use data that is as accurate and current as possible, we included only the 2,558 species previously published in treatments of the FBRA. Although not a complete representation of the diversity of the region, it constitutes nearly half of the species present and is a sufficiently representative sample of the total flora. All 203 published floristic treatments of the FBRA were reviewed in order to prepare a list of families, genera, and species (Appendix S3; see Supplemental Data with this article). Searches of taxonomic and nomenclatural updates were conducted using the "Taxonomy name/id Status Report Page" (NCBI Resources Coordinators 2016). All names on the main list that are no longer in use were removed. It is worth mentioning that the biggest families (Asteraceae, Fabaceae, Orchidaceae, and Poaceae) have only been partially treated, due the large number of species belonging to each of them.

Results

After taxonomic and nomenclatural updates of the names on the list, we obtained a total of 2,558 species, 799 genera, and 186 families for the region (Appendix S3; see Supplemental Data with this article). Of the species, 1,575 have a least one accession number in the GenBank database (Table 2; Appendix S1, S3), that represents 61.57 % of the total species studied in FBRA from the region. A total of 1,322 of the 1,575 species have been used in phylogenetic studies, and this figure represents 83.94 % of the species in GenBank and 51.68 % of the total species treated in FBRA (Table 2). Barcoding studies represent the second place in Genbank with

Table 1. List of generalized terms used in the classification of the obtained information; an asterisk (*) means zero or more characters different to whitespace (tabulation, space, or line breaks). For the category of diversity studies, this is the combination of diver* and all terms in curly brackets.

Category	Terms
Phylogenetic studies	phylogen*, filogen*, monop*, monof*, systemat*, sistemat*, retationsh*, relacio*
Phylogeographic studies	phylogeog*, filogeog*
Phylogenomics analysis	phylogenom*, genome-scale, "plastid genome", filogenóm*
Barcoding works	barcod*, barra*
Diversity studies	diver* {geneti*, pop*, pobl*}
Biogeography	biogeog*

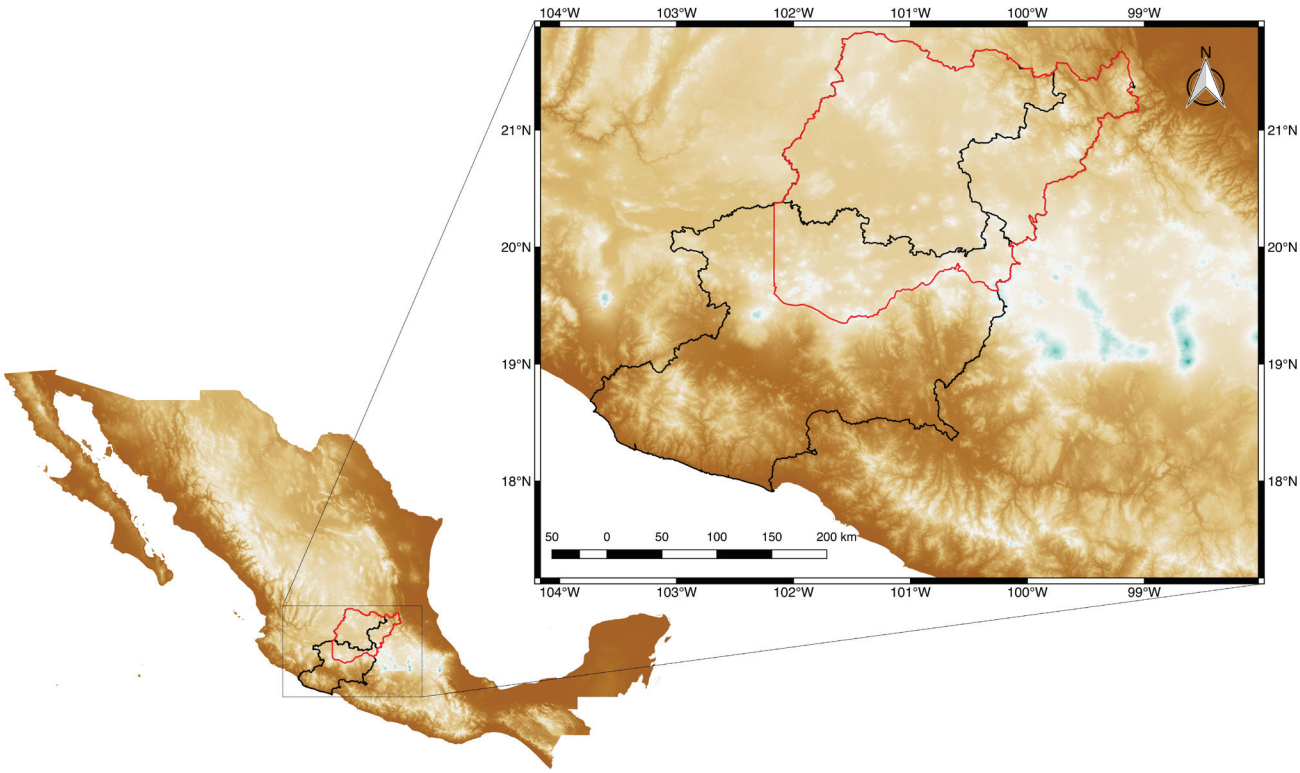

Figure 1. Map showing the area of the Flora del Bajío y de Regiones Adyacentes (represented by the polygon in red).

326 species, followed by biogeographic studies with 298, 41 for phylogenomic studies, 34 for phylogeographic studies, and finally only 34 species for diversity studies (Table 2).

In conjunction with the floristic treatments, one genus, 173 species, four subspecies, and five varieties have been described as new to science from the region (Appendix S4; see Supplemental Data with this article). Of these, only 26 species have been included in phylogenetic studies, seven in biogeographic studies, and one in a barcoding analysis. This value (19.8 %) is significantly lower than the overall percentage of species included in each study.

Of all the families studied in the FBRA, only eight lack species from the region that have sequences in GenBank, these being: Theaceae (five spp.), Ebenaceae and Eriocaulaceae (three spp.), Mayacaceae, Platanaceae, Pterostemonaceae, Thymelaeaceae, and Xyridaceae (with one species each). On the other hand, 77 families have a GenBank accession for all species that occur in the region. Crassulaceae and Malvaceae are two speciose families that have 48 and 46 % of the species with at least one entry in GenBank. Some other families lack a good representation of species with regard to the total included in GenBank. These are Aristolochiaceae (1 of 7 species), Dioscoreaceae (3 of 12), Geraniaceae (2 of 14), Linaceae (2 of 9), Phrymaceae (1 of 5), and Sterculiaceae (4 of 16) (Appendix S5; see Supplemental Data with this article).

Table 2. The total number of species treated in the FBRA and species with GenBank records organized by number, kind of study, and percentage.

Kind of study	Number of species in a study category	Percentage of the number of species with a GenBank accession	Percentage of the entire sample of the FBRA
Phylogenetic	1322	83.94 %	51.68 %
Phylogeographic	34	2.16 %	1.33 %
Phylogenomics	41	2.60 %	1.60 %
Barcoding	326	20.70 %	12.74 %
Diversity	34	2.16 %	1.33 %
Biogeography	298	18.92 %	11.65 %

Discussion

The extensive API system integrated in GenBank makes this repository not only a public platform to store data but also an important source to obtain data. In this case, an open source script was designed to extract the information and organize it in a way that demonstrates the importance and representation of the species found in floristic projects like FBRA. However, it can also be applied to other floristic projects or different sets of taxa of interest. This script mainly uses the Entrez system to access information other than the DNA sequences, which separates it from other programs and scripts that automatically extract data from GenBank. For our purposes, the focus was on the title of the cited paper for each sequence in order to infer the use of the data in the different disciplines analyzed. A future possibility could be explored using other information like abstract and keywords (retrieved from other public databases), by employing deep learning techniques that refine the search and classification.

The use of molecular-based phylogenies has improved our understanding of taxonomic relationships (Ulloa-Ulloa *et al*. 2017). Phylogenetic and phylogeographic studies, based on molecular DNA sequences, are sometimes necessary for the quality and accuracy of floristic work. In the case of FBRA, the correct taxonomic position of *Velascoa recondita* Calderón & Rzed. (Crossosomataceae), a newly described genus and species with unusual floral morphology endemic to the FBRA region (Calderón de Rzedowski & Rzedowski 1997), was only possible through a phylogenetic analysis using the chloroplast gene *rcb*L (Sosa & Chase 2003). The phylogenetic analysis of the Crossosomataceae found that *V. recondita* is the sister species to *Apacheria chiricahuensis* C.T. Mason, another monotypic genus endemic to Arizona and New Mexico (Mason 1975, Sosa & Chase 2003).

On the other hand, floristic studies are an essential tool for any evolutionary study (Funk 2006), being sometimes the only updated repository where morphological descriptions and identification keys can be found. Such studies support the correct application of names to samples in an evolutionary study (Palmer & Richardson 2012). For this reason, a complete descriptive Flora can influence the results and quality of the phylogenetic hypotheses.

For the FBRA, currently 1575 species have sequence records in GenBank, making that study an important repository for nomenclature, morphological descriptions and in some cases keys to their proper identification.

Despite the importance of having a better understanding of land plant phylogeny and evolution, at a regional level we do not even know the total number of species present in a small area like that covered by the FBRA. Furthermore, although we understand the phylogenetic relationships of the orders and families of Angiosperms, for the most of the 416 families recognized by APG *et al*. (2016), we do not know their internal (genus and species level) phylogenetic relationships.

For the FBRA, having 51.68 % of species included in a phylogenetic analysis can be an important state of the knowledge of the evolutionary relationships at a regional level. Currently, half of the species documented in the FBRA lack a tested evolutionary relationship. This impact can be being mitigated by the high percentage of genera included in some phylogenetic analysis, with least 84.35 % of the genera having been found in the GenBank repository. Nevertheless, genera like *Velascoa* show that the proper inclusion in an evolutionary study can change the understanding of their evolutionary history and taxonomy.

The low values in barcoding studies for species from the region show a potential concern about the correct interpretation and implementation of these kind of techniques in a broad study. In the region, only 12.74 % of the species were included in a barcoding study. With next-generation sequencing costs in constant decline, sequencing projects nowadays are taxonomically broad (Chan & Ragan 2013). For example, the objective of the 1KP initiative is to sequence complete plant genomes (https://sites.google.com/a/ualberta.ca/onekp/ accessed February 20, 2019). However, we are still very far from sequencing the complete genomes of the 383,671 vascular species in the world (Lughadha *et al*. 2016), showing a worrying panorama, where the number species sequenced is currently increasing, whereas the number of species treated in a floristic study is decreasing. It is necessary to include species never studied before to try to have a complete and robust evolutionary history of vascular plants. Our study region includes 2,558 species that have been formally included in the FBRA, and our results indicate that only 41 species (Table 2) have been used for phylogenomic studies. Our study case is an example of the impact of floristic projects for a particular region where most of its species have been used in phylogenetic studies, but it reveals that the number of species used for phylogenomic studies continues to be very low.

The goal of the Flora del Bajío y de Regiones Adyacentes is to contribute to our knowledge of the species of the area and to provide morphological descriptions and keys to the correct identification of these species. Additionally, it aims to document the total number of species for the region, their distribution, and the level of endemism.

It could be possible that in the upcoming years, all the floristic knowledge generated worldwide could be deposited on the Biodiversity CyberBank an initiative proposed by Wen *et al*. (2015), and it will be free to use, as is GenBank today. This information will be essential for phylogenetic, biogeographic, phylogeographic, phylogenomic, and diversity studies, but also crucial for policy makers involved in the conservation strategies of the flora of a particular region. It is expected that as time passes, the use of the Flora in various studies will increase, and although this is not yet reflected in the statistics, it will be clearer when the Flora del Bajío y de Regiones Adyacentes has been completed.

Of all families included in this metadata extraction, Crassulaceae and Malvaceae are two interesting groups that need to have their sampling efforts increased. Despite the other largest families having included a high percentage of species in GenBank, these two families have the fewest species with an entry in GenBank.

Some other low represented families like Aristolochiaceae, Dioscoreaceae, Geraniaceae, Linaceae, Phrymaceae, and Sterculiaceae *sensu* FBRA also require more effort in future molecular evolutionary studies to try to have a better sample.

In conclusion, our new script worked well to download metadata sequence information from GenBank. Although the script was developed and tested with a checklist of plants from the Flora of the Bajío region, it can be used with any list of species to answer various questions involving the metadata that usually are present but unused in accessions.

Acknowledgements

We thank Mario Suárez Mota for providing the shapefile (vector file) of the Flora del Bajío y de Regiones Adyacentes polygon. CONACyT-Laboratorio Nacional de Identificación y Caracterización Vegetal (LANIVEG) is also acknowledged.

Supplementary Material

Appendix S1. List of the species with Genbank accession by type of study ("X" indicates found in that particular study, whereas "-" indicates not found).
Appendix S2. List of species with GenBank accession by type of study; the order of the species corresponds to the order of the published treatments.
Appendix S3. List of the species treated in the published FBRA treatments.
Appendix S4. List of new taxa described from research conducted for the FBRA.
Appendix S5. Number of species per family by category.

Literature cited

Bennett J. 2015. AutoIt Script Homepage. <http://www.autoitscript.com/site/> (Accessed 5 May 2018).

Bennett D, Hettling H, Silvestro D, Zizka A, Bacon C, Faurby S, Antonelli A. 2018. phylotaR: An automated pipeline for retrieving orthologous DNA sequences from GenBank in R. *Life*, **8:** E20. DOI: https://doi.org/10.3390/life8020020

Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL. 2006. GenBank. Nucleic acids research, **36**: D16-D20. DOI: https://doi.org/10.1093/nar/gkj157

Calderón de Rzedwoski G, Rzedowski J. 1991. *Flora del Bajío y de Regiones Adyacentes*. Fascículo complementario I. Instituto de Ecología, A.C., Centro Regional del Bajío. Pátzcuaro, México. 14 pp.

Calderón de Rzedoski G, Rzedowski J. 1997. *Velascoa* (Crossosomataceae), un género nuevo de la Sierra Madre Oriental de México. *Acta Botanica Mexicana* **39**: 53-59. DOI: https://doi.org/10.21829/abm39.1997.776

Chan CX, Ragan MA. 2013. Next-generation phylogenomics. *Biology Direct* **8**: 1-6. DOI: https://doi.org/10.1186/1745-6150-8-3

Delsuc F, Brinkmann H, Philippe H. 2005. Phylogenomics and the reconstruction of the tree of life. *Nature Reviews Genetics* **6**: 361-375. DOI: https://doi.org/10.1038/nrg1603

Eisen JA, Fraser CM. 2003. Phylogenomics: intersection of evolution and genomics. *Science* **300**: 1706-1707. DOI: https://doi.org/10.1126/science.1086292

Ferrari L, Orozco-Esquivel T, Manea V, Manea M. 2012. The dynamic history of the Trans-Mexican Volcanic Belt and the Mexico subduction zone. *Tectonophysics* **522-523**: 122-149. DOI: https://doi.org/10.1016/j.tecto.2011.09.018

Ferrusquia-Villafranca I. 1993. Geology of Mexico: a synopsis. Biological diversity of Mexico: origins and distribution. *In*: Ramamoorthy TP, Bye R, Lot A, Fa J, eds. *Biological Diversity of Mexico: Origins and Distribution.* New York: Oxford University Press, 3-107. ISBN-13: 978-0195066746; DOI: https://doi.org/10.1007/BF02908211

Funk VA. 2006. Floras: a model for biodiversity studies or a thing of the past? *Taxon*, **55**: 581-588.

Gómez-Tuena A, Orozco-Esquivel MT, Ferrari L. 2007. Igneous petrogenesis of the Trans-Mexican Volcanic Belt. *Geological Society of America Special Paper* **422**: 129-181. DOI: https://doi.org/10.1130/2007.2422(05)

Hajibabaei M, Singer GA, Hebert PD, Hickey DA. 2007. DNA barcoding: how it complements taxonomy, molecular phylogenetics and population genetics. *Trends in Genetics* **23**: 167-172. DOI: https://doi.org/10.1016/j.tig.2007.02.001

Hebert PD, Cywinska NA, Ball SL, deWaard JR. 2003. Biological identifications through DNA barcodes. *Proceedings of the Royal Society of London B* **270**: 313-321. DOI: https://doi.org/10.1098/rspb.2002.2218

Kier G, Mutke J, Dinerstein E, Ricketts TH, Küper W, Kreft H, Barthlott W. 2005. Global patterns of plant diversity and floristic knowledge. *Journal of Biogeography* **32**: 1107-1116. DOI: https://doi.org/10.1111/j.1365-2699.2005.01272.x

Lughadha EN, Govaerts R, Belyaeva I, Black N, Lindon H, Allkin R, Magill RE, Nicolson N. 2016. Counting counts: revised estimates of numbers of accepted species of flowering plants, seed plants, vascular plants and land plants with a review of other recent estimates. *Phytotaxa* **272**: 82-88. DOI: http://dx.doi.org/10.11646/phytotaxa.272.1.5

Maglott D, Ostell J, Pruitt KD, Tatusova T. 2005. Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Research* **33**: D54-D58. DOI: https://doi.org/10.1093/nar/gki031

Mason CT. 1975. *Apacheria chiricahuensis*: a new genus and species from Arizona. *Madroño* **23**: 105-108.

Maya-Lastra CA. 2019. Datataxa v.U. Available from: <https://github.com/camayal/Datataxa> (accessed January 21, 2019)

McCormack JE, Hird SM, Zellmer AJ, Carstens BC, Brumfield RT. 2013. Applications of next-generation sequencing to phylogeography and phylogenetics. *Molecular Phylogenetics and Evolution* **66**: 526-538. DOI: https://doi.org/10.1016/j.ympev.2011.12.007

NCBI Resource Coordinators. 2016. Database resources of the national center for biotechnology information. *Nucleic Acids Research* **44**: D7-D19. DOI: https://doi.org/10.1093/nar/gkv1290

Palmer MW, Richardson JC. 2012. Biodiversity data in the information age: Do 21st century floras make the grade? *Castanea* **77**: 46-59. DOI: https://doi.org/10.2179/11-035

Rzedowski J. 1978. *Vegetación de México*. México, D.F.: Limusa.

Rzedowski J. 1991. Diversidad y orígenes de la flora fanerogámica de México. *Acta Botanica Mexicana* **14**: 3-21. https://doi.org/10.21829/abm14.1991.611

Sanderson MJ, Driskell AC. 2003. The challenge of constructing large phylogenetic trees. *Trends in Plant Science* **8**: 374-379. DOI: https://doi.org/10.1016/S1360-1385(03)00165-1

Sanderson MJ, Boss D, Chen D, Cranston KA, Wehe A. 2008. The PhyLoTA Browser: processing GenBank for molecular phylogenetics research. *Systematic Biology* **57**: 335-346. DOI: https://doi.org/10.1080/10635150802158688

Smith SA, Brown JW. 2018. Constructing a broadly inclusive seed plant phylogeny. *American Journal of Botany* **105**: 302-314. DOI: https://doi.org/10.1002/ajb2.1019

Soltis DE, Albert VA, Savolainen V, Hilu K, Qiu YL, Chase MW, Farris JS, Stefanović S, Rice DW, Palmer JD, Soltis PD. 2004. Genome-scale data, angiosperm relationships, and 'ending incongruence': a cautionary tale in phylogenetics. *Trends in Plant Science* **9**: 477-483. DOI: https://doi.org/10.1016/j.tplants.2004.08.008

Soltis DE, Moore MJ, Sessa EB, Smith SA, Soltis PS. 2018. Using and navigating the plant tree of life. *American Journal of Botany* **105**: 287-290. DOI: https://doi.org/10.1002/ajb2.1071

Sosa V, Chase MW. 2003. Phylogenetics of Crossosomataceae based on *rbc*L sequence data. *Systematic Botany* **28**: 96-105. https://www.jstor.org/stable/3093940

Suárez-Mota ME, Villaseñor JL, López-Mata L. 2015. La región del Bajío, México y la conservación de su diversidad florística. *Revista Mexicana de Biodiversidad* **86**: 799-808. DOI: http://dx.doi.org/10.1016/j.rmb.2015.06.001

The Angiosperm Phylogeny Group (APG), Chase MW, Christenhusz MJM, Fay MF, Byng JW, Judd WS, Soltis DE, Mabberley DJ, Sennikov AN, Soltis PS, Stevens PF. 2016. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG IV. *Botanical Journal of the Linnean Society* **181**: 1-20. DOI: https://doi.org/10.1111/boj.12385

Ulloa-Ulloa C, Acevedo-Rodríguez P, Beck S, Belgrano MJ, Bernal R, Berry PE, Brako, Celis M, Davidse G, Forzza RC, Gradstein SR, Kokche O, León B, León-Yánez S, Magil RE, Neil DA, Nee M, Rave PH, Stimmel H, Strong MT, Villaseñor JL, Zarucchi JL, Zuluoaga FO, Jørgensen PM. 2017. An integrated assessment of the vascular plant species of the Americas. *Science* **358**: 1614-1617. DOI: https://doi.org/10.1126/science.aao0398

Wen J, Ickert-Bond SM, Appelhans MS, Dorr LJ, Funk VA. 2015. Collections-based systematics: Opportunities and outlook for 2050. *Journal of Systematics and Evolution* **53**(6), 477-488. DOI: https://doi.org/10.1111/jse.12181

Associated editor: Ivón Ramírez-Morillo

Author contributions: ERS (https://orcid.org/0000-0002-7981-4490): conceived the idea, wrote the paper, and reviewed drafts of the paper. CAML (https://orcid.org/0000-0002-0550-3331): wrote the script, analyzed the data following the methodology, wrote the paper, and reviewed drafts of the paper. VWS: curated the database and reviewed drafts of the paper. SZ: produced the list of species, organized and curated the database, and reviewed drafts of the paper. EC: curated the database and reviewed drafts of the paper. RMM: produced the list of species, organized the database, and reviewed drafts of the paper. JR: curated the database and reviewed drafts of the paper.