MR. BRIAN SALVIN MAITNER (Orcid ID : 0000-0002-2118-9880)

DR. NAIA MORUETA-HOLME (Orcid ID : 0000-0002-0776-4092)

DR. STEPHEN SMITH (Orcid ID : 0000-0003-2035-9531)

DR. CYRILLE VIOLLE (Orcid ID : 0000-0002-2471-9226)

Article type : Application

Handling Editor: Dr. Sean McMahon

Corresponding author email id: bmaitner@gmail.com

Title:

The BIEN R package: A tool to access the Botanical Information and Ecology Network (BIEN) Database

Authors:

Brian S. Maitner[1], Brad Boyle[1], Nathan Casler[2], Rick Condit[3], John Donoghue II[1], Sandra M. Durán[1], Daniel Guaderrama[1], Cody E Hinchliff[4], Peter M. Jørgensen[5], Nathan J.B. Kraft[6], Brian McGill[7], Cory Merow[8], Naia Morueta-Holme[9], Robert K. Peet[10], Brody Sandel[11], Mark Schildhauer[12], Stephen A. Smith[4], Jens-Christian Svenning[13], Barbara Thiers[14], Cyrille Violle[15], Susan Wiser[16] and Brian J. Enquist[1,17]

Author Affiliations:

This is the author manuscript accepted for publication and has undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1111/2041-210X.12861

This article is protected by copyright. All rights reserved

Author Manuscript

[1]Department of Ecology and Evolutionary Biology, University of Arizona,
Tucson, AZ 85721, USA;
[2]National Center for Supercomputing Applications, University of Illinois Urbana-Champaign,
Urbana, IL 61801, USA;
[3]Smithsonian Tropical Research Institute, Center for Tropical Forest Science, Global Forest
Observatory Network, Unit 9100 Box 0948, DPO AA 34002, Panama;
[4]Department of Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, MI
48109, USA;
[5]Missouri Botanical Garden, P.O. Box 299, St. Louis, MO 63166-0299, USA;
[6]Department of Ecology and Evolutionary Biology, University of Los Angeles, Los Angeles, CA
90095, USA;
[7]School of Biology and Ecology, University of Maine, Orono, ME 04469, USA;
[8]Department of Ecology and Evolutionary Biology, Yale University, New Haven, CT 06520,
USA;
[9]Department of Integrative Biology, University of California, Berkeley, CA 94720, USA;
[10]Department of Biology, University of North Carolina, Chapel Hill, NC 27599-3280, USA;
[11]Department of Biology, Santa Clara University. Santa Clara, CA 95057, USA;
[12]National Center for Ecological Analysis and Synthesis, Santa Barbara, CA, USA;
[13]Section for Ecoinformatics & Biodiversity, Department of Bioscience, Aarhus University, Ny
Munkegade 114, DK-8000 Aarhus C, Denmark;
[14]William and Lynda Steere Herbarium at the New York Botanical Garden, Bronx, NY, USA;
[15]Centre d'Ecologie Fonctionnelle et Evolutive (UMR 5175), CNRS – Université de Montpellier
– Université Paul Valéry Montpellier, EPHE, Montpellier, France;
[16]Landcare Research, Lincoln 7608, New Zealand;
[17]The Santa Fe Institute, USA, 1399 Hyde Park Rd, Santa Fe, NM 87501, USA;

Tweetable Abstract:

BIEN for R: Global plant occurrences, trait data, plot data. Range maps and phylogenies for most New World plants.

Abstract:

This article is protected by copyright. All rights reserved

Author Manuscript

1. There is an urgent need for large-scale botanical data to improve our understanding of community assembly, coexistence, biogeography, evolution, and many other fundamental biological processes. Understanding these processes is critical for predicting and handling human-biodiversity interactions and global change dynamics such as food and energy security, ecosystem services, climate change, and species invasions.

2. The Botanical Information and Ecology Network (BIEN) database comprises an unprecedented wealth of cleaned and standardized botanical data, containing roughly 81 million occurrence records from ~375,000 species, ~915,000 trait observations across 28 traits from ~93,000 species, and co-occurrence records from 110,000 ecological plots globally, as well as 100,000 range maps and 100 replicated phylogenies (each containing 81,274 species) for New World species. Here, we describe an R package that provides easy access to these data.

3. The BIEN R package allows users to access the multiple types of data in the BIEN database. Functions in this package query the BIEN database by turning user inputs into optimized PostgreSQL functions. Function names follow a convention designed to make it easy to understand what each function does. We have also developed a protocol for providing customized citations and herbarium acknowledgements for data downloaded through the BIEN R package.

4. The development of the BIEN database represents a significant achievement in biological data integration, cleaning and standardization. Likewise, the BIEN R package represents an important tool for open science that makes the BIEN database freely and easily accessible to everyone.

Keywords: R, plants, range maps, presence, occurrence, phylogeny, community plot, functional traits, biodiversity, herbarium records, ecoinformatics

Introduction:

Understanding the distribution and diversity of organisms relies on the availability of large quantities of standardized, robust, and comprehensive biological data. Furthermore, developing

This article is protected by copyright. All rights reserved

Author Manuscript

ecological theories that are both general and predictive also necessitates high-quality data. There is an especially urgent need for large quantities of botanical data across differing spatial scales. Despite the crucial importance of plants for ecosystem functioning and services, plants have lacked comprehensive data sets similar to those of less-speciose clades such as birds and mammals, which have trait (Myhrvold *et al.* 2015), phylogenetic (Bininda-Emonds *et al.* 2007; Jetz *et al.* 2012; Faurby & Svenning 2015) and range information (BirdLife International 2013; IUCN 2013) available for most species. Critical issues with economic implications, such as food and energy security, climate change, species invasions and ecosystem services, require an understanding of plant ecology and evolution globally, and an ability to predict how these factors will respond to global change.

In the era of Big Data and open science, the informatics challenge is to develop workflows that can standardize and provide access to disparate and heterogeneous data sources. Increasingly, the synthetic questions biologists are asking require computationally intensive integration between these numerous and diverse data sources. However, the existence and accessibility of data is only part of the issue. Before data can be integrated, they must be assessed, cleaned, and standardized.

The Botanical Information and Ecology Network (BIEN; http://bien.nceas.ucsb.edu/bien/) database is an unprecedented collection of integrated and standardized New World botanical data (Enquist *et al.* 2016). This has resulted from the efforts of more than 50 scientists from 20 institutions across the globe. The BIEN database is constructed from a workflow that performs three essential tasks. First, it integrates multiple data types (e.g. trait data, occurrence records, ecological plot data, Table 1) from numerous sources (Supplementary Tables S1-3). Second, utilizing open source tools, it assesses and corrects taxonomic errors, standardizes taxonomy, validates geographic information, and standardizes data formats. Third, it generates species-level distribution models and phylogenies. The current version of the BIEN database (BIEN 3.4 as of Jan 15th 2017) contains 81,108,996 occurrence records that stem from 378,502 species from across the globe. While the data are global in extent (Figures 1 & 2), data integration efforts have focused on the New World (North and South America, Table 2).

The BIEN database differs from other biodiversity databases in the breadth of data types it contains (Table 1), the integration of these data, the degree of validation and standardization the data goes through, and the amount of additional metadata added. Most other databases focus

This article is protected by copyright. All rights reserved

Author Manuscript

on a single data type, such as TRY (Kattge *et al.* 2011) with trait data and GBIF ('Global Biodiversity Information Facility' 2012) with occurrence records, where BIEN integrates multiple data types, including occurrence, plot, trait, and phylogenetic data. Data within the BIEN database are also put through standardization and validation procedures, thereby increasing data quality and ease of use. For example, BIEN occurrence data undergo additional taxonomic and spatial validations that aren't used by GBIF ('Global Biodiversity Information Facility' 2012), and unlike Traitbank (Schulz 2016), BIEN trait units and names have been standardized.

The wealth of information contained in the BIEN database is housed at the National Center for Ecological Analysis and Synthesis (NCEAS) in Santa Barbara, CA. The database is open-access and can either be accessed by a graphical user interface (GUI) at BIENdata.org or through the newly developed BIEN package for R that we introduce here (R Core Development Team 2015). The BIENdata.org website offers a user-friendly way to visualize species occurrence data and range maps and download range maps, occurrences and trait data for a single species at a time. Users who are interested in downloading data for many taxa or geographic locations and accessing differing combinations of data will find the BIEN R package a more flexible approach.

Data Cleaning and Integration Workflow

As described by Enquist et al. (in prep., 2016), the generation of the BIEN database consists of a linked workflow that (i) standardizes taxonomy by correcting spelling of species names and updating synonyms to currently accepted names via the Taxonomic Name Resolution Service or TNRS (Boyle *et al.* 2013); (ii) detects and flags observations with erroneous geographic coordinates; and (iii) flags cultivars and non-native species via the Native Species Resolver (http://bien.nceas.ucsb.edu/bien/tools/nsr/). Coordinates were flagged as erroneous if they fell outside the specified political regions, if the latitude was exactly 0 or 90 degrees, the longitude was exactly 0 or 180 degrees, or if the point fell in the ocean. The detection of cultivars and non-native records relies on native species lists, which are not available throughout the New World, so this filtration is imperfect.

Range maps were built for each species using a method determined by the number of observations of that species. A species with a single record was assigned a range that included

This article is protected by copyright. All rights reserved

Author Manuscript

only the 100km^2 cell where it was found. Ranges for species with 2-3 records were rectangular bounding boxes with the limits set by the minimum and maximum latitude and longitude of all occurrences. Ranges for species with 4-9 records were built with convex hulls (the minimum-fitting polygon that encompasses all occurrences of that species). For species with >9 records, we built species distribution models using the Maxent algorithm (Phillips *et al.* 2006). Only one occurrence record per cell (in cases of multiple records) was used for Maxent model building. Maxent model building generally followed the recommendations outlined in (Merow *et al.* 2013, 2014). Model settings were chosen to balance overfitting, which underestimates range sizes, with underfitting, which results in excessively smooth models that over predict range size. Only linear, quadratic, and product features were used and regularization was set at the default value.

Environmental predictors for SDMs were obtained from the WorldClim current (1960-1990) climate data at 10 arc-minute resolution (Hijmans *et al.* 2005) and resampled to a 10 km resolution. Predictors included mean annual temperature, mean diurnal temperature range, annual precipitation, precipitation seasonality, precipitation in warmest quarter/ (precipitation in warmest quarter + precipitation in coldest quarter), and five spatial eigenvectors (De Marco *et al.* 2008). The spatial eigenvectors essentially captured large-scale regional differences in occurrence and primarily served as broad-scale dispersal limitation of species ranges, limiting predictions far in geographic space from presence locations (Blach-Overgaard *et al.* 2010).

Maxent's continuous predictions were converted to binary presence/absence predictions by choosing a threshold based on the 75th percentile of the cumulative output.

We constructed a phylogeny of 18,641 species based on a standardized list of New World species and the gene regions atpB-rbcL, ndhF, psbA, psbA-psbH, rbcL, and trnT-trnL-trnF marker regions using the software PHLAWD (Smith *et al.* 2009). The phylogeny was constructed with RAxML (7.3.0; Stamatakis 2006) with an unconstrained ML search and divergence times were estimated using penalized likelihood and the treePL software package (Smith & O'Meara 2012). Additional details on the methodology used to extract these data from GenBank and align them are presented in Hinchliff and Smith (2014). Onto this phylogeny, we grafted the additional taxa from the BIEN dataset using taxonomy (genus membership) as a guide for the remaining ~72,000 species. This grafting was repeated to create a set of 100 phylogenies to account for uncertainty in placement of species without genetic information.

This article is protected by copyright. All rights reserved

Author Manuscript

Additional information on the BIEN phylogenies is available online (http://bien.nceas.ucsb.edu/bien/biendata/bien-2/phylogeny/).

BIEN R package

A common hurdle for researchers is accessing the specific subset of data they need for their study. Particularly for large databases with multiple data types, data access can involve complex queries in database languages that are unfamiliar to many researchers. The BIEN package for R (stable version available on CRAN and development version at https://github.com/bmaitner/RBIEN) provides a set of tools to access and work with the BIEN database that allows users without a knowledge of PostgreSQL or the BIEN database structure to easily access the data within BIEN.

The functions within the BIEN package fall into nine main categories:

I. Species lists,
II. Occurrence records (species observations from specimens and plots),
III. Plot data (species occurrence and abundance from plots),
IV. Stem data (stem counts and measurements from plots),
V. Trait data,
VI. Species range maps,
VII. Taxonomic information,
VIII. Phylogenetic information,
IX. Associated metadata

We follow a function naming convention where each function begins with the prefix "`BIEN_`" followed by a designation of which data type it accesses (eg "`trait_`" or "`ranges_`") followed by a description of what the specific function does. For example, the function `BIEN_occurrence_species()` is used to download geographic information system (GIS) occurrence data for a given species. Detailed descriptions of each function family, and examples of common usages can be found in Supplementary Document S4. A complete list of functions with abbreviated descriptions is available in tabular format with `vignette("BIEN")`, and a

This article is protected by copyright. All rights reserved

Author Manuscript

tutorial is available with `vignette("BIEN_tutorial")`. Data usage caveats can be found in Supplementary Table S5.

Data license

The data and derived products are distributed via CC-BY-NC-ND Creative Commons license (https://creativecommons.org/licenses/by-nc-nd/4.0/). This allows redistribution and re-use of a licensed work on the conditions that the creator is appropriately credited, that use is noncommercial, and that users do not distribute derivative works. Users interested in uses of these data and products outside the scope of this license are encouraged to contact the BIEN working group (bien.working.group@gmail.com) to discuss alternative licensing options.

Authorship guidelines

Any publications using data obtained from BIEN should cite the original publications and acknowledge data providers when possible. Studies utilizing large amounts of BIEN data should seek solutions that ultimately lead to an entry in the Web of Science so that data providers and original publications are given credit. Normal ethics apply to co-authorship of scientific publications. Users of data stored in BIEN should consider inviting the original data contributor to be a co-author(s) of any resultant publications if that contributor's data are a major portion of the dataset analyzed, or if a data contributor makes a significant contribution to the analysis of the data or to the interpretation of results. Similarly, studies utilizing significant fractions of BIEN data and derived data products (species lists, geographic ranges, etc.) for their research are encouraged (but NOT required) to contact the BIEN working group, who may be able to provide insight into the use of BIEN data and expected updates to the database. In no circumstances should authorship be attributed to data contributors, individually or collectively, without their explicit consent.

Future directions

The BIEN database represents an unprecedented source of global botanical data that currently focuses on the New World. The BIEN database, workflow, and R package are under continuous development. We plan on expanding the BIEN database to increase coverage of other continents as well as incorporating additional types of trait data. Another priority is to

This article is protected by copyright. All rights reserved

Author Manuscript

develop an API to allow easier access to the BIEN database by other programs. As development proceeds, we will add functionality to allow users to contribute their own data and range maps. Accessing these data will allow studies at an unprecedented scale with more complete, high quality data sets than smaller scale initiatives might achieve on their own. This is an important step toward improving our understanding of global patterns of biodiversity and developing predictive ecological theory at the largest of scales.

Authors' contributions:

B.S.M, C.M, B.B. and B.J.E. wrote the manuscript. B.S.M. led development of the R package with assistance by B.B., C.M., D.G. and N.C. All authors worked on development of the BIEN database and workflow, contributed feedback on the manuscript, and approved the publication.

Acknowledgements:

We thank Irena Šímová for input on the manuscript, Scott Chamberlain for valuable comments on R programming, our beta-testers for their input, and our CRAN reviewers for their cheerful comments and assistance. We thank Scott Chamberlain, Robert Freckleton, and an anonymous reviewer for their helpful comments on both the manuscript and the R package. N.M.-H. acknowledges support by a Villum Foundation Postdoctoral Fellowship. We gratefully acknowledge the Canada Foundation for Innovation for funding the infrastructure for Canadensys. This work was conducted as a part of the Botanical Information and Ecology Network (BIEN) Working Group (PIs BJE, Richard Condit, BB, SD, RKP) supported by the National Center for Ecological Analysis and Synthesis (funded by NSF Grant #EF-0553768), the University of California, Santa Barbara, and the State of California. The BIEN Working Group was also supported by iPlant (NSF #DBI-0735191). CV was supported by the French Foundation for Research on Biodiversity (FRB; www.fondationbiodiversite.fr) in the context of the CESAB project "Assembling, analysing and sharing data on plant functional diversity to understand the effects of biodiversity on ecosystem functioning: a case study with French Permanent Grasslands" (DIVGRASS), by a Marie Curie International Outgoing Fellowship within the 7th European Community Framework Program (DiversiTraits project, no. 221060) and by the European Research Council (ERC) Starting Grant Project "Ecophysiological and biophysical constraints on domestication in crop plants" (Grant ERC-StG-2014-639706- CONSTRAINTS).

This article is protected by copyright. All rights reserved

Author Manuscript

JCS was supported by the European Research Council (ERC-2012-StG-310886-HISTFUNC). SKW was supported by Core funding for Crown Research Institutes from the New Zealand Ministry of Business, Innovation and Employment's Science and Innovation Group. We would like to acknowledge the herbaria that contributed data to this work: A, AAH, AAS, AAU, ABH, ACAD, ACOR, AD, ADW, AFS, AHUC, AIMS, AJOU, AK, AKPM, ALCB, ALT, ALTA, ALU, AMD, AMES, AMNH, AMO, ANA, ANGU, ANSM, ANSP, ANUC, ARAN, ARC, ARIZ, ARM, AS, ASDM, ASU, ATCC, AUG, AUT, B, BA, BAA, BAB, BACP, BAF, BAFC, BAI, BAJ, BAL, BARC, BAS, BBB, BBS, BC, BCF, BCMEX, BCN, BCRU, BEREA, BG, BH, BHCB, BHO, BILAS, BIO, BISH, BLA, BM, BO, BOCH, BOG, BOL, BOLV, BONN, BOUM, BPI, BR, BRA, BREM, BRI, BRIT, BRIU, BRLU, BRM, BSB, BSIP, BSN, BTN, BUL, BULU, BUT, C, CALI, CAMU, CAN, CANB, CANL, CAS, CAY, CBG, CBM, CBS, CEN, CEPEC, CESJ, CGE, CHAM, CHAP, CHAPA, CHI, CHL, CHR, CHRB, CIB, CICY, CIIDIR, CIMI, CINC, CIQR, CLEMS, CLF, CM, CMC, CMMEX, CNHM, CNPO, CO, COA, COAH, COCA, CODAGEM, COFC, COL, COLO, CONC, CORD, CP, CPAP, CPUN, CR, CRAI, CRP, CS, CSU, CTES, CTESN, CU, CUVC, CUZ, CVRD, CWU, DAO, DAOM, DAV, DBN, DES, DLF, DMNH, DMU, DNA, DR, DS, DUKE, DUSS, E, EA, EAC, EBH, EBUM, ECH, ECU, EIF, EIU, EKY, EM, EMMA, ENCB, ENS, ERA, ESA, ESS, F, FAA, FAU, FB, FBCS, FCME, FCO, FCQ, FEN, FH, FHO, FI, FLAS, FLOR, FM, FR, FRP, FTG, FUEL, FURB, G, GB, GDA, GDAC, GE, GENT, GEO, GES, GH, GI, GJO, GLM, GMNHJ, GOET, GUA, GZU, H, HA, HAC, HAJB, HAL, HAM, HAMAB, HAO, HAS, HAST, HASU, HAW, HB, HBG, HBR, HCIB, HEID, HGI, HIP, HKU, HNHM, HNT, HO, HPL, HRCB, HRP, HSS, HSU, HU, HUA, HUAA, HUAL, HUAZ, HUEFS, HUEM, HUFU, HUSA, HUT, HXBH, HYO, IA, IAA, IAC, IAL, IAN, IB, IBGE, IBUG, ICEL, ICN, IEB, IFO, ILL, ILLS, IMSSM, INB, INEGI, INIF, INM, INPA, IPA, IPRN, ISC, ISL, ISTC, ISU, ITCV, ITMH, IZAC, IZTA, JACA, JBAG, JE, JEPS, JOE, JUA, JYV, K, KANU, KIEL, KMN, KMNH, KOELN, KOR, KPM, KSC, KSTC, KSU, KTU, KU, KUO, KYO, L, LA, LAE, LAF, LAM, LCR, LD, LE, LEB, LEMA, LG, LI, LIL, LINN, LISE, LISI, LISU, LKHD, LL, LM, LOJA, LOMA, LP, LPAG, LPB, LPD, LPS, LSU, LTR, LY, LYJB, LZ, M, MA, MAF, MAIC, MAK, MAN, MARY, MASS, MB, MBK, MBM, MBML, MCM, MCN, MCNS, MEL, MEN, MERL, MEXU, MFA, MFU, MG, MGC, MICH, MIL, MIN, MISS, MJG, MMMN, MNHM, MNHN, MO, MOL, MOR, MOSS, MPU, MPUC, MRSN, MSB, MSC, MSE, MSTR, MSUN, MT, MTMG, MU, MUB, MUCV, MVFA,

This article is protected by copyright. All rights reserved

Author Manuscript

MVFQ, MVJB, MVM, MY, N, NA, NCSC, NCU, ND, NE, NEB, NHM, NHMC, NHT, NLH, NLU, NMB, NMC, NMCR, NMNL, NMR, NMSU, NMW, NO, NOU, NRCC, NSPM, NSW, NT, NUM, NWOSU, NY, O, OC, OCLA, ODU, OHN, OKL, OKLA, OMA, OS, OSA, OSC, OSH, OSN, OULU, OWU, OXF, P, PACA, PAR, PE, PEL, PENN, PERTH, PEUFR, PFC, PH, PI, PKDC, PLAT, PMA, PMNH, PNH, POLL, POM, PORT, PR, PRC, PRE, PTBG, PVNH, PY, QCA, QCNE, QFA, QM, QMEX, QRS, QUE, R, RAS, RB, RBR, REG, RENO, RFA, RIOC, RM, RNG, ROST, RPM, RSA, RYU, S, SALA, SAM, SAN, SANT, SAPS, SASK, SBT, SD, SEL, SEV, SF, SFSU, SGO, SI, SIM, SING, SIU, SJRP, SLPM, SMB, SMDB, SMF, SNM, SOM, SP, SPF, SPSF, SQF, SRFA, STL, STU, SUVA, SVG, SZU, TAES, TAI, TAIF, TAMU, TAN, TEF, TENN, TEPB, TEX, TFC, TFM, TI, TKPM, TNS, TO, TRA, TRH, TROM, TRT, TRTC, TRTE, TRTS, TS, TSM, TTRS, TU, TULS, TUR, U, UADY, UAM, UAMIZ, UARK, UAS, UAT, UB, UBA, UBC, UC, UCAM, UCBG, UCR, UEC, UESC, UFG, UFMA, UFMT, UFP, UFRJ, UFRN, UFS, UGDA, UH, UI, UJAT, ULM, ULS, UME, UMO, UNA, UNB, UNCC, UNEX, UNL, UNM, UNR, UNSL, UPCB, UPEI, UPNA, UPNG, UPS, US, USAS, USJ, USM, USNC, USON, USP, USZ, UT, UTC, UTEP, UTMC, UV, UVIC, UVSC, UWO, V, VA, VAL, VALD, VDB, VEN, VM, VMSL, VT, W, WAG, WAT, WELT, WFU, WII, WIN, WIS, WMNH, WOH, WRSL, WS, WTU, WU, XAL, Y, YA, YAM, YU, Z, ZMT, ZSS, ZT.

Data Accessibility:

The BIEN package is available for download from CRAN at: https://CRAN.R-project.org/package=BIEN . The development version of the package is available at: https://github.com/bmaitner/RBIEN .

References

Bininda-Emonds, O.R.P., Cardillo, M., Jones, K.E., MacPhee, R.D.E., Beck, R.M.D., Grenyer, R., Price, S. a., Vos, R. a., Gittleman, J.L. & Purvis, A. (2007). The delayed rise of present-day mammals. *Nature*, **446**, 507–512.

BirdLife International. (2013). IUCN Red List for birds. *http://www.birdlife.org*.

This article is protected by copyright. All rights reserved

Author Manuscript

Blach-Overgaard, A., Svenning, J.-C., Dransfield, J., Greve, M. & Balslev, H. (2010). Determinants of palm species distributions across Africa: the relative roles of climate, non-climatic environmental factors, and spatial constraints. *Ecography*, **33**, 380–391.

Boyle, B., Brad, B., Nicole, H., Zhenyuan, L., Garay, J.A.R., Dmitry, M., Tony, R., Naim, M., Narro, M.L., Piel, W.H., Mckay, S.J., Sonya, L., Chris, F., Peet, R.K. & Enquist, B.J. (2013). The taxonomic name resolution service: an online tool for automated standardization of plant names. *BMC Bioinformatics*, **14**, 16.

De Marco, P., Diniz-Filho, J.A.F. & Bini, L.M. (2008). Spatial analysis improves species distribution modelling during range expansion. *Biology Letters*, **4**, 577–580.

Enquist, B.J., Condit, R., Peet, R.K., Schildhauer, M. & Thiers, B.M. (2016). *Cyberinfrastructure for an integrated botanical information network to investigate the ecological impacts of global climate change on plant biodiversity*. PeerJ Preprints.

Enquist, B.J., Sandel, B., Boyle, B., Svenning, J.-C., McGill, B.J., Donoghue, J.C., Hinchliff, C.E., Jørgensen, P.M., Kraft, N.J.B., Marcuse-Kubitza, A., Merow, C., Morueta-Holme, N., Peet, R.K., Schildhauer, M., Spencer, N., Regetz, J., Šímová, I., Smith, S.A., Thiers, B., Violle, C., Wiser, S.K., Andelman, S., Casler, N., Condit, R., Dolins, S., Guaderrama, D., Maitner, B.S., Narro, M.L., Ott, J.E., Phillips, O., Sloat, L.L. & ter Steege, H. (in prep.). Botanical big data shows that plant diversity in the New World is driven by climatic-linked differences in evolutionary rates and biotic exclusion.

Faurby, S. & Svenning, J.-C. (2015). Historic and prehistoric human-driven extinctions have reshaped global mammal diversity patterns. *Diversity and Distributions*, **21**, 1155–1166.

Global Biodiversity Information Facility. (2012). URL http://www.gbif.org/ [accessed 11 December 2012]

Hijmans, R.J., Cameron, S., Parra, J., Jones, P.G. & Jarvis, A. (2005). WorldClim, version 1.3. *University of California, Berkeley*.

Hinchliff, C.E. & Smith, S.A. (2014). Some limitations of public sequence data for phylogenetic inference (in plants). *PloS one*, **9**, e98986.

This article is protected by copyright. All rights reserved

Author Manuscript

IUCN. (2013). IUCN Red List of Threatened Species. *Version 2013.2*. URL IUCNredlist.org

Jetz, W., Thomas, G.H.H., Joy, J.B.B., Hartmann, K. & Mooers, a. O.O. (2012). The global diversity of birds in space and tim. *Nature*, **491**, 1–5.

Kattge, J., Diaz, S., Lavorel, S., Prentice, I.C., Leadley, P., Bönisch, G., Garnier, E., Westoby, M., Reich, P.B., Wright, I.J. & Others. (2011). TRY--a global database of plant traits. *Global change biology*, **17**, 2905–2935.

Merow, C., Cory, M. & Silander, J.A. (2014). A comparison of Maxlike and Maxent for modelling species distributions. *Methods in Ecology and Evolution*, **5**, 215–225.

Merow, C., Cory, M., Smith, M.J. & Silander, J.A. (2013). A practical guide to MaxEnt for modeling species' distributions: what it does, and why inputs and settings matter. *Ecography*, **36**, 1058–1069.

Myhrvold, N.P., Baldridge, E., Chan, B., Sivam, D., Freeman, D.L. & Ernest, S.K.M. (2015). An amniote life-history database to perform comparative analyses with birds, mammals, and reptiles. *Ecology*, **96**, 3109–3109.

Phillips, S.J., Anderson, R.P. & Schapire, R.E. (2006). Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, **190**, 231–259.

R Core Development Team. (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing. URL https://www.R-project.org/

Schulz, K. (2016). TraitBank: An open digital repository for organism traits. *2016 International Congress of Entomology*. Entomological Society of America.

Smith, S.A., Beaulieu, J.M. & Donoghue, M.J. (2009). Mega-phylogeny approach for comparative biology: an alternative to supertree and supermatrix approaches. *BMC evolutionary biology*, **9**, 37.

Smith, S.A. & O'Meara, B.C. (2012). treePL: divergence time estimation using penalized likelihood for large phylogenies. *Bioinformatics* , **28**, 2689–2690.

This article is protected by copyright. All rights reserved

Author Manuscript

385 Stamatakis, A. (2006). RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with
386 thousands of taxa and mixed models. *Bioinformatics* , **22**, 2688–2690.

Author Manuscript

Table 1. Summary of Data Types in BIEN 3.4

	Count	Number of Species
Occurrence Records	81,108,996	378,502
Ecological Plots Observations	17,244,737	16,216
Trait Records	916,988	93,251
Taxonomic Records	539,781	378,577
Range Maps (see text for details)	98,829	98,829

This article is protected by copyright. All rights reserved

Phylogenies **(see text for details)**	100	81,274

Author Manuscript

This article is protected by copyright. All rights reserved

433 **Table 2. Summary of the number of occurrences within BIEN 3.4**

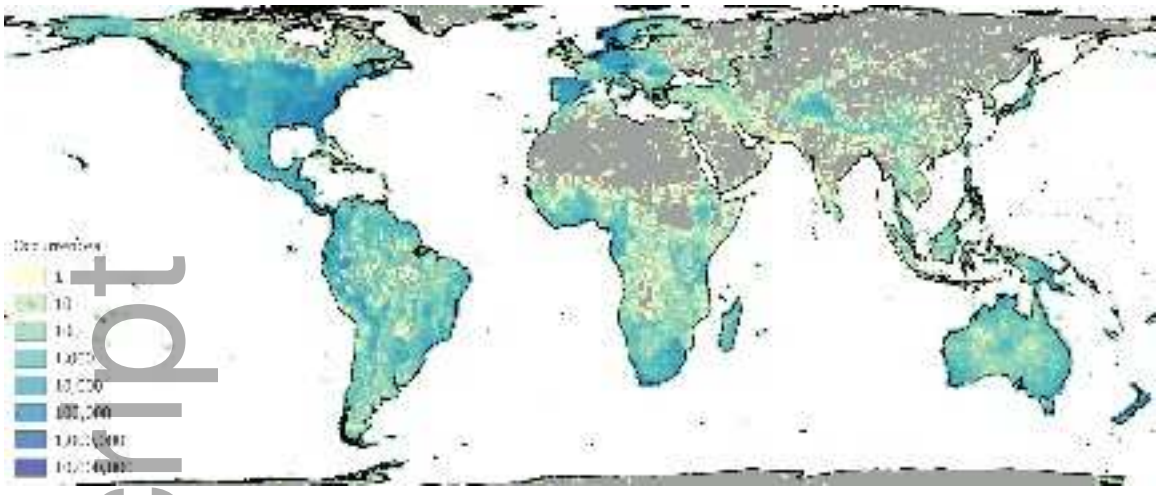
434

	Number observed within BIEN 3.4 (Global)	Number observed within BIEN 3.4 (New World)
Total Observations	81,108,996	28,311,777
Geographically validated Observations	46,432,247	22,462,543
Specimens	63,864,259	13,365,277
Plot observations	17,244,737	14,946,500
Plots	113,146	107,619
Species	378,502	117,808

435

Author Manuscript

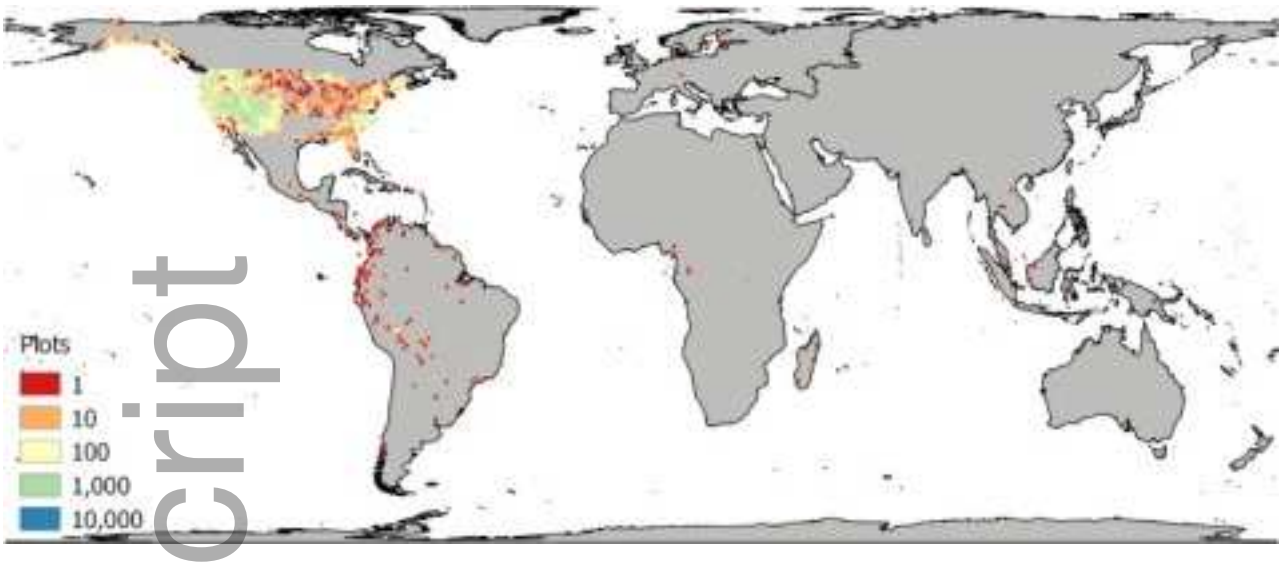
This article is protected by copyright. All rights reserved



mee3_12861_f1.png

Author Manuscript

This article is protected by copyright. All rights reserved



mee3_12861_f2.png

Author Manuscript

This article is protected by copyright. All rights reserved