Systematic Botany (2016), 41(3): pp. 529–539

DOI 10.1600/036364416X692505
Date of publication August 26, 2016

Specimens at the Center: An Informatics Workflow and Toolkit for Specimen-level analysis of Public DNA database data

The Global *Carex* Group

Kasey K. Pham,[1,2,27] Marlene Hahn,[1] Kate Lueders,[1] Bethany H. Brown,[1,3] Leo P. Bruederle,[4] Jeremy J. Bruhl,[5] Kyong-Sook Chung,[6] Nathan J. Derieg,[4,7] Marcial Escudero,[8] Bruce A. Ford,[9] Sebastian Gebauer,[10] Berit Gehrke,[11] Matthias H. Hoffmann,[10] Takuji Hoshino,[12] Pedro Jiménez-Mejías,[13] Jongduk Jung,[14] Sangtae Kim,[15] Modesto Luceño,[16] Enrique Maguilla,[16] Santiago Martín-Bravo,[16] Robert F. C. Naczi,[17] Anton A. Reznicek,[18] Eric H. Roalson,[13] David A. Simpson,[19] Julian R. Starr,[20] Tamara Villaverde,[16] Marcia J. Waterway,[21] Karen L. Wilson,[22] Okihito Yano,[23,24] Shuren Zhang,[25] and Andrew L. Hipp[1,26,27]

[1]The Morton Arboretum, Lisle, Illinois 60532, U.S.A.
[2]Michigan State University, East Lansing, Michigan 48824, U.S.A.
[4]University of Colorado Denver, Denver, Colorado 80217, U.S.A.
[5]University of New England, Armidale NSW 2351, Australia
[6]Jungwon University, Chungbuk 367-805, Korea
[8]Department of Plant Biology and Ecology, University of Seville, 41004 Seville, Spain
[9]University of Manitoba, Winnipeg, Manitoba R3T 2N2, Canada
[10]Martin-Luther-Universität Halle-Wittenberg, D-06108 Halle (Saale), Germany
[11]Johannes Gutenberg-Universität Mainz, 44122 Mainz, Germany
[12]Okayama University of Science, Okayama 700-0005, Japan
[13]Washington State University, Pullman, Washington 99164, U.S.A.
[14]Ajou University, Suwon 16499, South Korea
[15]Sungshin Women's University, Seoul 136-742, South Korea
[16]Universidad Pablo de Olavide, 41004 Seville, Spain
[17]The New York Botanical Garden, Bronx, New York 10458, U.S.A.
[18]University of Michigan, Ann Arbor, Michigan 48109, U.S.A.
[19]Royal Botanic Gardens, Kew, Richmond, Surrey TW9 3AB, United Kingdom
[20]University of Ottawa, Ottawa, Ontario K1N 6N5, Canada
[21]McGill University, Macdonald Campus, Ste-Anne-de-Bellevue, QC, H9X 3V9, Canada
[22]Royal Botanic Garden Sydney, Sydney NSW 2000, Australia
[23]The University Museum, University of Tokyo, Tokyo 113-0033, Japan
[25]Beijing Institute of Botany, the Chinese Academy of Sciences, Beijing 100093, China
[26]The Field Museum, Chicago, Illinois 60605, U.S.A.
[3]Current address: Ball Horticultural Company, West Chicago, Illinois 60185, U.S.A.
[7]Current address: University of California, Santa Barbara, California 93106, U.S.A.
[24]Current address: 12 Okayama University of Science, Japan
[27]Authors for correspondence (ahipp@mortonarb.org, kase.khanh.pham@gmail.com)

Communicating Editor: James Smith

***Abstract*—**Major public DNA databases — NCBI GenBank, the DNA DataBank of Japan (DDBJ), and the European Molecular Biology Laboratory (EMBL) — are invaluable biodiversity libraries. Systematists and other biodiversity scientists commonly mine these databases for sequence data to use in phylogenetic studies, but such studies generally use only the taxonomic identity of the sequenced tissue, not the specimen identity. Thus studies that use DNA supermatrices to construct phylogenetic trees with species at the tips typically do not take advantage of the fact that for many individuals in the public DNA databases, several DNA regions have been sampled; and for many species, two or more individuals have been sampled. Thus these studies typically do not make full use of the multigene datasets in public DNA databases to test species coherence and select optimal sequences to represent a species. In this study, we introduce a set of tools developed in the R programming language to construct individual-based trees from NCBI GenBank data and present a set of trees for the genus *Carex* (Cyperaceae) constructed using these methods. For the more than 770 species for which we found sequence data, our approach recovered an average of 1.85 gene regions per specimen, up to seven for some specimens, and more than 450 species represented by two or more specimens. Depending on the subset of genes analyzed, we found up to 42% of species monophyletic. We introduce a simple tree statistic—the Taxonomic Disparity Index (TDI)—to assist in curating specimen-level datasets and provide code for selecting maximally informative (or, conversely, minimally misleading) sequences as species exemplars. While tailored to the *Carex* dataset, the approach and code presented in this paper can readily be generalized to constructing individual-level trees from large amounts of data for any species group.

Keywords*—Carex*, Cyperaceae, phylogenetic workflow, specimen-level data, supermatrix, taxon disparity index (TDI).

Specimen-level data are at the heart of revisionary taxonomy, but much synthetic work in systematics has focused on development of species-level tools for phylogenetics (e.g. supertree and supermatrix approaches, and gene tree – species tree reconciliation) and monography (e.g. Scratchpads [Smith et al. 2012] and Encyclopedia of Life [Parr et al. 2014]). In the collections community, great strides have been made in databasing, georeferencing, and aggregating specimen data

(e.g. BioGeomancer [Guralnick et al. 2006], BRAHMS [http://herbaria.plants.ox.ac.uk/bol/brahms/], GBIF IPT [Robertson et al. 2014], Symbiota [Gries et al. 2014]), and in tracking specimen duplicates and reconciling annotations across these (e.g. BiSciCol; http://biscicol.blogspot.com/p/home.html). However, the tools needed to extract specimen-level data from major public sequence databases—NCBI GenBank (Benson et al. 2007), the DNA DataBank of Japan (DDBJ), and the European Molecular Biology Laboratory (EMBL)—are not in place. Hereafter we will refer only to NCBI GenBank in this paper, with the understanding that the issues are general across all three databases.

Why aggregate data to the specimen level instead of just focusing on species-level data? There are at least two potential advantages to discriminating specimens in NCBI data. The first is that aggregating data to the specimen level allows us, at least in principle, to test species boundaries under a phylogenetic or genealogical species concept (De Queiroz, 2007; Hausdorf 2011). In the conventional approach to utilizing NCBI data (e.g. Edwards and Smith 2010; Hinchliff and Roalson 2013), which we will refer to in this paper as the aggregate-to-species approach, phylogenetic tips are left at the species level, so the resulting trees offer no information on species boundaries. Related to this is the potential discovery of cryptic taxa, which may lie hidden in NCBI data but undiscovered in downstream analyses that only aggregate data to species. The second advantage is that our approach affords the researcher greater power to detect misidentified specimens and lab errors that would be all but invisible in the aggregate-to-species approach, without prior expectations about where a given species should fall in the tree.

The genus *Carex* L. (Cyperaceae) is a particularly challenging group from which to develop a specimen-level NCBI resource. Under the broad circumscription of the genus to include the previously segregated genera *Cymophyllus* Mack., *Kobresia* Willd., *Schoenoxiphium* Nees, and *Uncinia* Pers. (Global *Carex* Group 2015), *Carex* comprises ca. 2000 species, spanning six continents and a remarkably wide range of terrestrial and aquatic ecosystems (Hipp 1998; Reznicek 1990, 1993; Escudero et al. 2012). Four or five major clades have been well characterized in the genus in many studies, but relationships among these clades and fine-scale relationships within clades remain in flux (Waterway and Starr 2007; Waterway et al. 2009; Starr and Ford 2009; Global *Carex* Group 2015). A more recent study (Starr et al. 2015) utilizes broader sampling of southeast Asian taxa and deeper DNA sampling (4400 base pairs) to further refine this understanding, suggesting that *Carex* comprises two major alliances and seven major clades. While there have been numerous subgenus- and section-level phylogenetic studies in *Carex* (Starr et al. 1999; Hendrichs et al. 2004a, b; Roalson and Friar 2004; Ford et al. 2006, 2012; Hipp et al. 2006; Escudero et al. 2008; Dragon and Barrington 2009; Escudero et al. 2009; Jiménez-Mejías et al. 2011; Shekhovtsov et al. 2012; Derieg et al. 2013; Gebauer et al. 2014; Yano et al. 2014; Maguilla et al. 2015; Molina et al. 2015), there has not been a recent effort to summarize these in a single synthetic paper focused on the genus. The most inclusive Cyperaceae phylogeny to date is from analysis of a Cyperaceae supermatrix (Hinchliff and Roalson 2013). This study confirms many of the higher-level relationships identified by others in previous studies but does not provide strong phylogenetic conclusions at fine scales. However, to date, no one has constructed a supermatrix multigene tree to the individual level for *Carex*. Presumably, this lack of individual-level trees can be attributed to the difficulty of parsing others' data to determine the individual specimens that are the source of the sequences. This problem becomes especially tricky when one considers that multiple studies may use the same individual specimen.

In this study, we provide (1) a set of informatics tools for annotating NCBI data to voucher, working around the limited data structures provided by NCBI for linking sequences to voucher, and for exploring resulting phylogenies for apparent non-monophyly of species or other taxa that may be due to lab error, taxonomic misidentification, or the need for taxonomic revision; and (2) a case study in *Carex*. We also provide a relatively small number of annotations of the NCBI database based on re-inspection of data by Global *Carex* Group (GCG) members who are coauthors on this paper; this is presented in the supplement to this paper for *Carex* researchers who may have used data in the past and would like to know which identifications have been annotated.

Materials and Methods

Downloading and parsing NCBI nucleotide records—The NCBI GenBank database was queried on 18 March 2015 for all nucleotide sequences for which the organism field contained the string "Carex," "Cymophyllus," "Kobresia," "Schoenoxiphium," or "Uncinia." Data were exported as NCBI INSDSeq XML (http://www.insdc.org/). Each sequence was maintained, as in GenBank, as a separate data record, irrespective of the source of genetic data; each data record, then, comprises sequence data plus a variety of metadata: information on the locus sequenced, taxonomic classification of the organism from which the DNA was extracted, and various forms of information on the identity of the specimen sequenced, including collector, collector number, geographic locality, "isolate" and "clone" (in quotes here because these terms are applied differently by different researchers), and other information distinguishing among sources of DNA sequence data. This information was used to identify each individual plant specimen that provided one or multiple sequences in the data set. The XML data were parsed into a flat file using the XML (Lang et al. 2015) and morton packages (https://github.com/andrew-hipp/morton) in R versions 2.15.3 ('Security Blanket') through 3.2.5 ('Very, Very Secure Dishes'; R Core Team 2015).

Linking nucleotide records to individual specimens—The parsed data table (Supplemental Table S1) was then analyzed to create a unique voucher or specimen code, a label unique to individual plant specimens. The NCBI database was not designed to hold specimen-level data that could be consistently compared across collections, researchers, research studies, or even different genes sampled within a single study. There is a field for voucher, which is often used. However, collector names, number, and collection (museum, herbarium) codes and numbers are not consistently indicated. As a consequence, we had to infer from the metadata associated with each sequence record what the specimen was, and create a code for specimens that could be associated across studies and genes. Collector names, locations, and numbers pertaining to the individual who collected the specimen, as well as collection names and numbers pertaining to the museum or collection housing the specimen, were parsed out of the specimen_voucher field using the parse.specimen function in morton. The parsing rules we used are hard-coded in the parse.specimen function. Even after parsing, however, substantial manual cleanup was required to address inconsistencies in collector and collection names and placement of collector and collection data within different fields. Following manual cleanup of the individual fields, the voucher names were automatically generated by concatenating five fields, if present: (1) collector name, (2) collector number, (3) isolate number, (4) collection (herbarium) name, and (5) collection (herbarium) accession number. Collector name was the only field required to contain information; if information associated with the collector of the plant sample was missing, the author name (s) for the paper in which the sequence was published was used instead and indicated using the label "AUTHOR." Spaces and punctuation were removed and all characters were changed to lowercase in vouchers to minimize spurious differences among records. Plant names were not included in the voucher name, due to the relatively transient nature of classifications and the fact that identification changes made on specimens are often not

communicated to NCBI. Over the 22 yr of sequence deposits represented in our sampling (Fig. 1), changes in taxonomic names and determinations on individual specimens make it likely that at least some individual specimens may have two or more scientific names associated with them. These names may or may not be synonyms; we did not attempt to distinguish the difference in the current study, though this could easily be done. All missing information was omitted from the voucher labels.

Following automated generation of voucher labels, the voucher labels were inspected visually and manually cleaned to differentiate those that were not fully differentiated in the previous step due to missing specimen data or variation in collection information (in some cases collection information was included for a given specimen in one study but not in another). When specimen-specific metadata for sequences were missing, multiple individuals were often erroneously grouped under a single voucher, usually by the primary author of the paper in which these individuals were first published. Our approach was to aggregate individuals to species within a research study if no specimen data were provided, under the assumption that multiple exemplars of a given species within a study are likely to derive from a single specimen, provided duplicate sequences are not present in the study. This assumption may of course result at times in incorrect attribution of specimen identity. When taxa lacking specimen data could not be unambiguously matched to a DNA region within a study, those individuals were excluded from analysis.

An advantage of linking specimens to vouchers became apparent with inspection of ITS and *trn*L–*trn*F data. Each of these regions was in many cases sequenced as two separate gene regions (e.g. ITS1 as one region, ITS2 as a separate region) and was therefore present in the data set as halves of a whole gene region. We concatenated any split sequences. Any gap in a sequence left unsequenced — for example, a gap for 5.8S between the first and second internal transcribed spacer (ITS) regions — was filled with Ns. Starr et al. (1999), the seminal study in the use of ITS in *Carex*, was used to estimate how many base pairs should be inserted for the 5.8S region, and 10 base pairs were removed from that value to avoid introducing spurious gaps during multiple alignment. For *trn*L–*trn*F, we inserted gaps by inspection and by reference to a publication in a more distantly related taxon (Bayer and Starr 1998).

***Multiple alignment*—** After cleaning the data set, each gene region was written to a separate FASTA file to be aligned. FASTA files were exported from the INSDSeq XML data with a label indicating the taxon, the collector, and an arbitrary number referencing the voucher. Only specimens attributable to a single organism were exported — any voucher identified to different taxa for different sequences was excluded — and a translation table was exported to relate the FASTA files to the original NCBI metadata. Only the top 12 most represented DNA regions (Fig. 2) were written to FASTA files for inclusion in the final tree. This decision was made based on the observation that the less-represented regions have few individuals common with other regions and are mostly from highly taxon-specific studies, making their inclusion in the final tree problematic; sequences that did come from individuals sequenced for other loci would not be likely to resolve correctly in the phylogenetic trees inferred because there would be an equal likelihood that they would be located in multiple places on the tree. Multiple alignment was performed using MUSCLE v.8 (Edgar 2004a, b). The resulting multiple alignment files were adjusted manually, and sequences that did not align properly were flagged for reverse-complementation or deletion. After editing flagged sequences, multiple alignments were re-aligned using MUSCLE and readjusted manually. Problematic sequences were removed, and for *trn*L-*trn*F, data were aligned by four major clades — outgroups + *Siderostictae* clade; Vignea clade; core *Carex* clade; and Caricoid clade — prior to profile-to-profile realignment in MUSCLE. Sequence alignments were trimmed lightly to get rid of ragged ends.

***Concatenated data matrices*—** We exported six datasets for analysis, with nicknames indicated in bold:

- **5-region:** all individuals containing any one of the most commonly sampled five regions

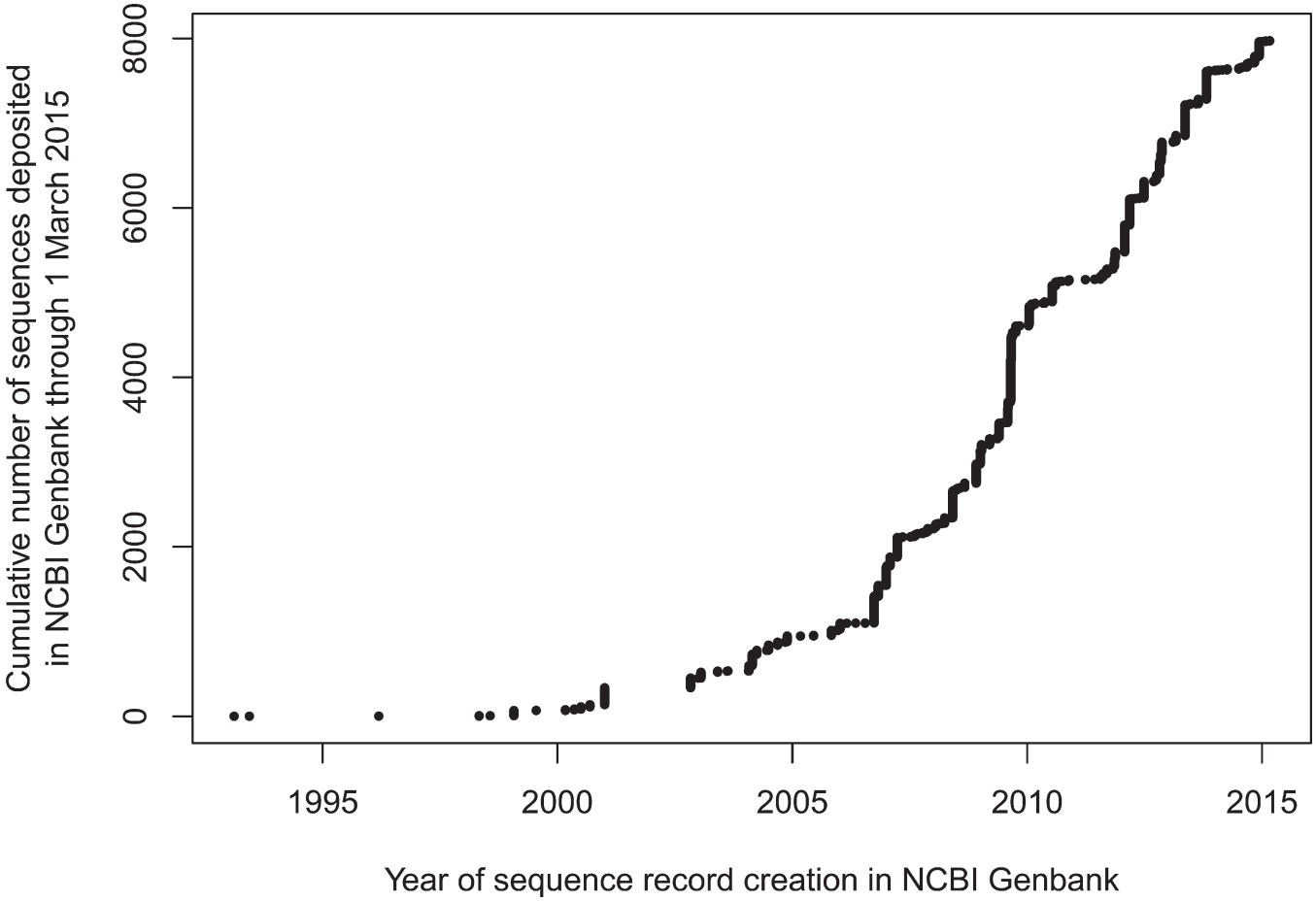

FIG. 1. GenBank sequence accumulation, 1991 to present. Cumulative sequencing progress through March 2015. Date of sequence deposit was taken from data provided by GenBank.

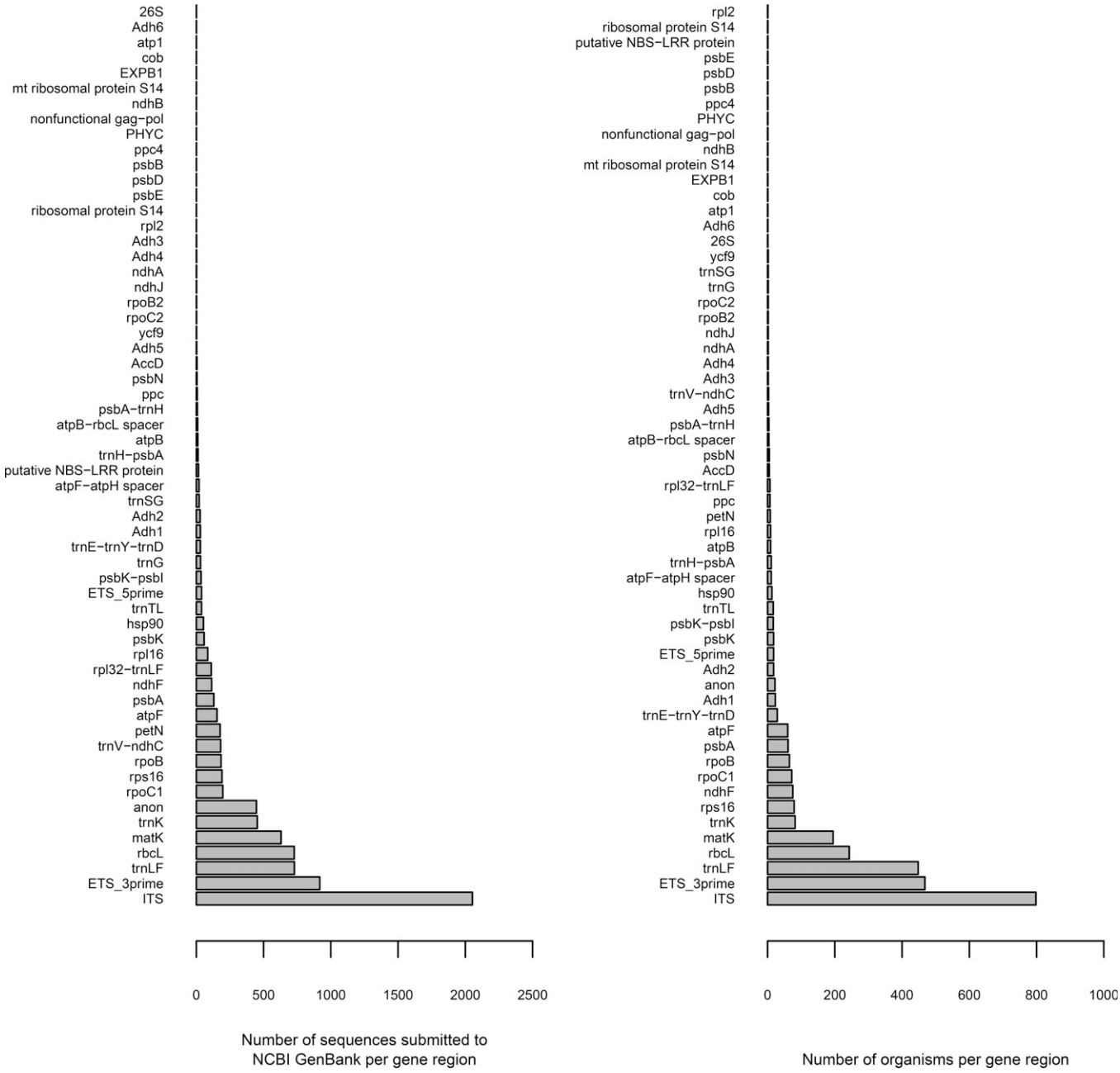

FIG. 2. Gene sampling barplot. Gene regions were categorized based on the more heterogeneous DNA regions descriptions provided in NCBI GenBank.

- **12-region:** all individuals containing any one of the most commonly sampled 12 regions
- **ITS scaffold:** the top 12 DNA regions for all individuals that had been sampled for ITS; thus all individuals have ITS, and many have additional gene regions as well
- **ITS+ETS:** ITS and ETS for those individuals that have both ITS and ETS (no missing data)
- **ITS only, ITS+ETS subset:** ITS for those individuals that have *both* ITS and ETS
- **ETS only, ITS+ETS subset:** ETS for those individuals that have *both* ITS and ETS

Nicknames indicated above are used throughout this manuscript.

Phylogenetic analyses — All analyses presented were conducted in RAxML v. 8 (Stamatakis 2014) using 50–100 fast bootstrap searches followed by slow ML search and optimization (this is the '-f a' option commonly used in RAxML). Trees in all analyses presented were rooted by the outgroups *Eriophorum vaginatum* L., *Scirpus polystachyus* F.Muell., *Trichophorum alpinum* (L.) Pers., and *T. cespitosum* (L.) Hartm. These outgroups were available for the ITS, ETS, *trn*L-*trn*F, *mat*K and *rbc*L datasets; for outgroups, no effort was made to match individuals. The following analyses were undertaken:

Comparison of datasets — To examine the impact of the number of regions included on the phylogenetic reconstruction, pairwise tree distances calculated using Penny and Hendy's (1985) tree bipartition metric, ignoring branch lengths, were calculated among all trees in the 5-region bootstrap treeset, the 12-region bootstrap treeset, and the ITS scaffold bootstrap treeset, as well as the ML tree for each of these datasets using the dist.topo function in the ape package (Paradis et al. 2004). All bootstrap and ML trees were pruned down to the taxa shared by all individuals shared among the three datasets prior to calculating pairwise tree distances. It was important to leave all taxa in the data matrices and prune after phylogenetic analysis for this portion of the study so that we could assess whether increasing both taxon sampling and percent missing data alters our understanding of relationships around a core group of tips for which we have maximal data. Multidimensional scaling as implemented in the vegan package (Oksanen et al. 2015) was used to visualize tree dissimilarities in two dimensions.

Effects of aggregating to individual: ITS vs. ETS—To evaluate the effects of aggregating data to the individual level, we estimated the ML tree for ITS+ETS; ITS only, ITS+ETS subset; and ETS only, ITS+ETS subset. Nodes on the ITS and ETS bootstrap consensus trees (bipartitions trees in RAxML) were matched back to equivalent nodes on the ITS+ETS tree using the phytools package (Revell 2012); equivalent nodes are defined as nodes on the rooted tree that have identical sets of descendants. Bootstraps for equivalent nodes were plotted for comparison using morton.

Taxon disparity index—We evaluated monophyly of species on our combined datasets by flagging tips based on species names (as reported in NCBI GenBank, ignoring synonymy, names below the species level, and possible misidentifications) and calculating the distribution of a taxon disparity index (TDI), which we introduce in this study. The TDI is defined here as the number of tips in the most restricted clade that includes all individuals of a given species label minus the number of tips of that species label. Unlike the consistency index (CI), the TDI has the desirable property of increasing as even a single outlier increases in phylogenetic distance from the core of the species; whereas a species label might have a CI 0.5 no matter how phylogenetically distant its tips are, TDI increases as species labels that violate monophyly increase in phylogenetic distance from one another. The taxon disparity index might also be formulated as the number of additional steps or the decrease in likelihood required to make all tips with a given species label monophyletic, using a paired-sites test such as the Shimodaira and Hasegawa (1999) test to compare the unconstrained tree from one in which all tips of a particular species label are monophyletic. However, both of these alternatives are computationally much more demanding, as they would require additional tree optimization steps, and it is not clear that they would appreciably alter our interpretation of how pruning tips or concatenating data affect monophyly of species labels. Note that we use "monophyly of species labels" deliberately in this context, because we are not distinguishing in this analysis among non-monophyly due to taxonomic issues (e.g. morphological species not reconciled with genealogical species), nomenclatural discrepancies, specimen misidentifications, or lab error. Note, too, that this analysis differs fundamentally from rogue taxon analysis, which investigates how consistent the placement of tips is among trees in a confidence set (e.g. bootstrap set).

Results

NCBI data, specimen assignment, and matrices—A total of 7994 sequence records encompassing 58 named DNA regions (after manual cleanup by the senior author; Fig. 2) and an additional set of anonymous regions (including SSR loci) were downloaded. Regions were drawn from plastid (37 regions), mitochondrial (three regions), nuclear (13), and nuclear ribosomal (five) DNA. Sequences were deposited in NCBI GenBank between 1993 and 2015 (Fig. 1). In the end, only 68 of 3909 vouchers identified had two taxonomic names associated with them, and none had more than two. Vouchers have an average of 1.85 +/− 0.95 gene regions sequenced per individual and range from one to seven gene regions each (Figs. 3, 4). The top 12 individual gene matrices range from 149 individuals (*atp*F and *trn*V-*ndh*C) to 1766 individuals (ITS), and in data completeness from < 5% gaps (*trn*V-*ndh*C) to more than 20% (ITS, ETS 3′ end, *trn*L-*trn*F, *trn*K, *rps*16; Table S2), with an average of 24% missing data. Much of the missing data was due to uncleaned ragged ends, and some from gaps that might be eliminated by aligning first within clades, then doing profile-to-profile alignment among clades (cf. Global *Carex* Group 2016, this issue). While we did this for the *trn*L-*trn*F data, we found little need to do so for the other matrices, though such an approach could be automated readily using an iterative phylogenetic analysis / multiple alignment approach. The combined 5-region and 12-region data matrices are 2827 individuals × 6228 bp and 3266 individuals × 13106 bp respectively. They represent 773 and 775 named species and comprise 83.5% and 91.1% gap characters respectively. The ITS scaffold matrix, like the ITS matrix itself, comprises 680 species, but a much higher proportion of missing data (90.8% gaps, as with the 12-region dataset; Supplemental Fig. S3). Supplemental tables and figures, including all matrices, are available from the Dryad Digital Repository at http://dx.doi.org/10.5061/dryad.6tn70d.

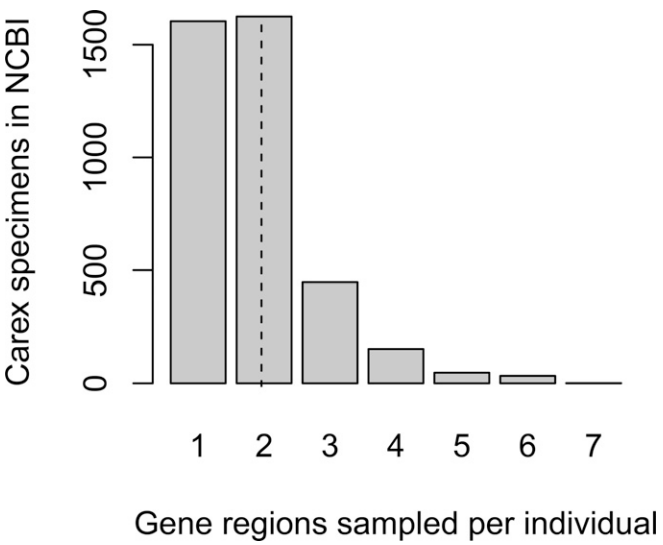

FIG. 3. DNA regions barplot. After aggregating data to individual, the number of gene regions sampled per individual was calculated.

Phylogenetic analyses and comparison of datasets—Each of the matrices enumerated under "Concatenated Data Matrices" (Methods) was analyzed individually as described in Methods (Fig. S4A–F). Individual genes were analyzed separately in preliminary analyses for this in the course of cleaning the alignments, but separate analyses are not shown here. A tree from analysis of the 5-region matrix was pruned to make a separate tree for each of 20 GCG coauthors (in a few cases, combinations of authors) who had sequence data deposited in NCBI GenBank with their name as a lead or associated author. Each of the authors had an opportunity to review the tips associated with their names and give feedback to the communicating authors, including redetermination of identifications. This resulted in removal of 40 tips and redetermination of 15 vouchers.

All analyses recovered the same major clades found in prior studies of the genus: the *Siderostictae*, core *Carex*, *Vignea*, and Caricoid clades (Global *Carex* Group 2015 and references therein; Figs. 4, 5). The placement of these clades is not strongly supported, fitting with the uncertainty reported in prior studies (Starr and Ford 2009, Fig. 3 presents a good review), but the overall topology recovered in the 5-region and 12-region trees is essentially the same (Figs. 4, 5), with some within-clade differences particularly in the core *Carex* clade that are relatively minor compared to some of the within-clade rearrangements between the 5-region and ITS-scaffold analyses (Fig. S5).

In the ordination of bootstrap trees for the 12-region, 5-region, and ITS scaffold datasets, pruned down to taxa common to all three datasets, there is strong overlap between the 12-region and ITS scaffold bootstrap sets, and minimal overlap between these and the 5-region bootstrap set (Fig. 6). The distances among the three ML trees are substantially lower than any other distance in the dataset (497–595 branches defining bipartitions that differ among the three ML trees,

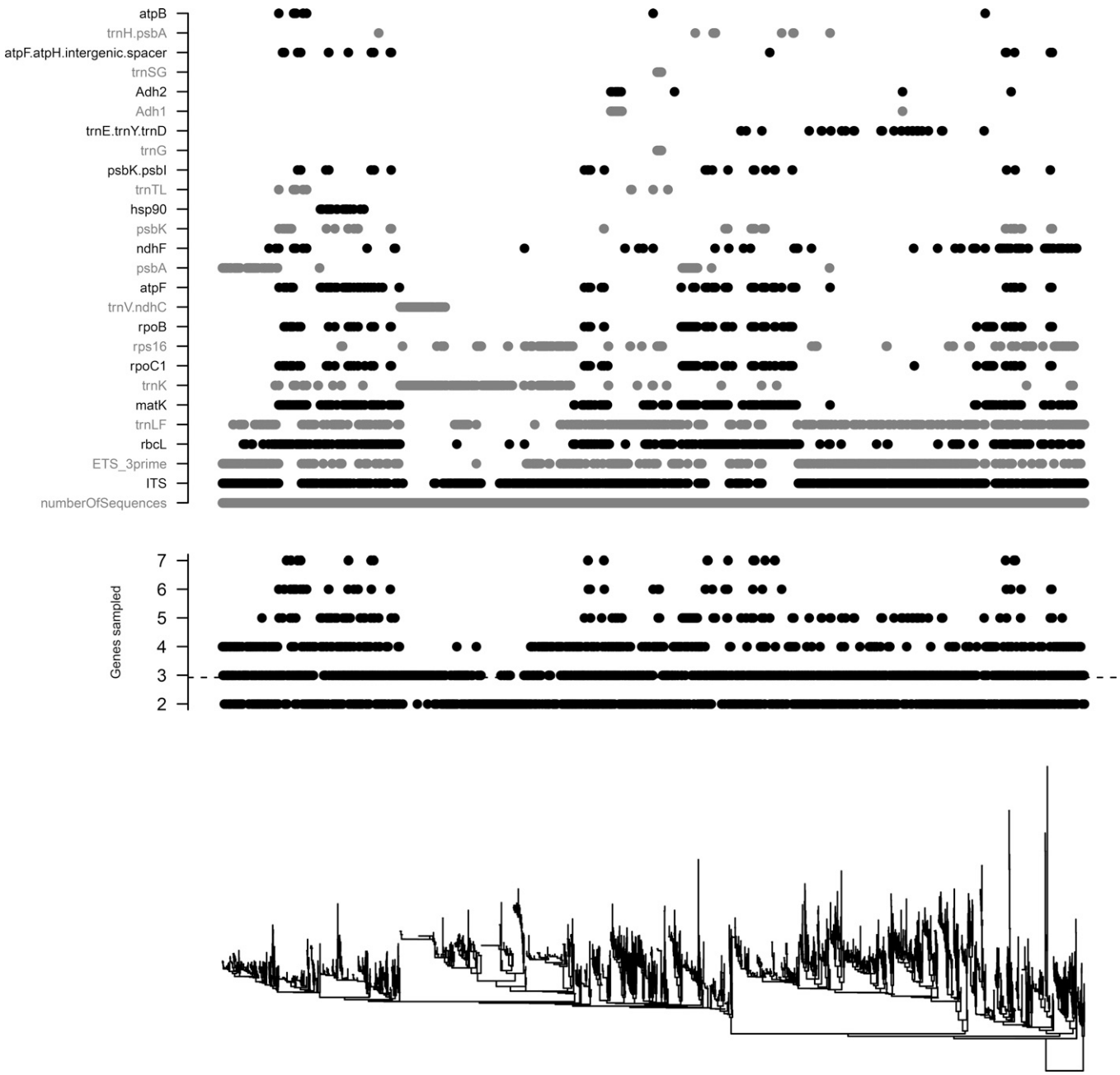

FIG. 4. Maximum likelihood phylogeny from the 12-region concatenated dataset, plotted against DNA region coverage. The 12-region phylogeny (lowest panel) is aligned with data for each tip on number of gene regions sampled (middle panel), and what particular gene regions are sampled (upper panel). The number of gene regions sampled is not cumulative: an individual will have only one point on the plot in the middle panel. Colors (black and gray in print) are only to aid in reading across the figure. Gene regions were included only if they had been sampled in a minimum of 10 individuals.

in comparison to differences of 835–1171 bipartitions for every pairwise comparison involving a bootstrap tree). Thus, the ML trees fall close together in the ordination (Fig. 6), surrounded by the bootstrap trees, rather than each ML tree in its own cloud of bootstrap trees.

Aggregating to individual—Aggregating data to specimen results in a net increase in clade support over not aggregating at all (Fig. 7). Only ITS and ETS were chosen for this investigation because a relatively large number of *Carex* individuals in NCBI have been sampled for both DNA regions, but we presume the benefits of aggregating to individual may extend beyond ITS and ETS. Looking across data matrices, we find that most species exhibit relatively low taxonomic disparity (TDI <100), with only a small number ranging to a TDI of 3000 or more (Figs. 8, S6), encompassing uncertainty across nearly the entire tree. There are islands of disparity that correspond to increasingly deep phylogenetic bipartitions that a species label might transgress (e.g. all taxa that are errantly split between the *Vignea* and core *Carex* clades will have a very similar and large disparity index, irrespective of how many individuals have that label). In linear regression of taxonomic disparity against number of taxa (Fig. S7), number of taxa is a weak predictor of TDI for all datasets (for the 12-gene and 5-gene datasets, $r^2 = 0.10–0.12$, $p < 0.001$; in all other datasets, the regression coefficients are not significantly different from 0), and we thus consider it a poor corrector to

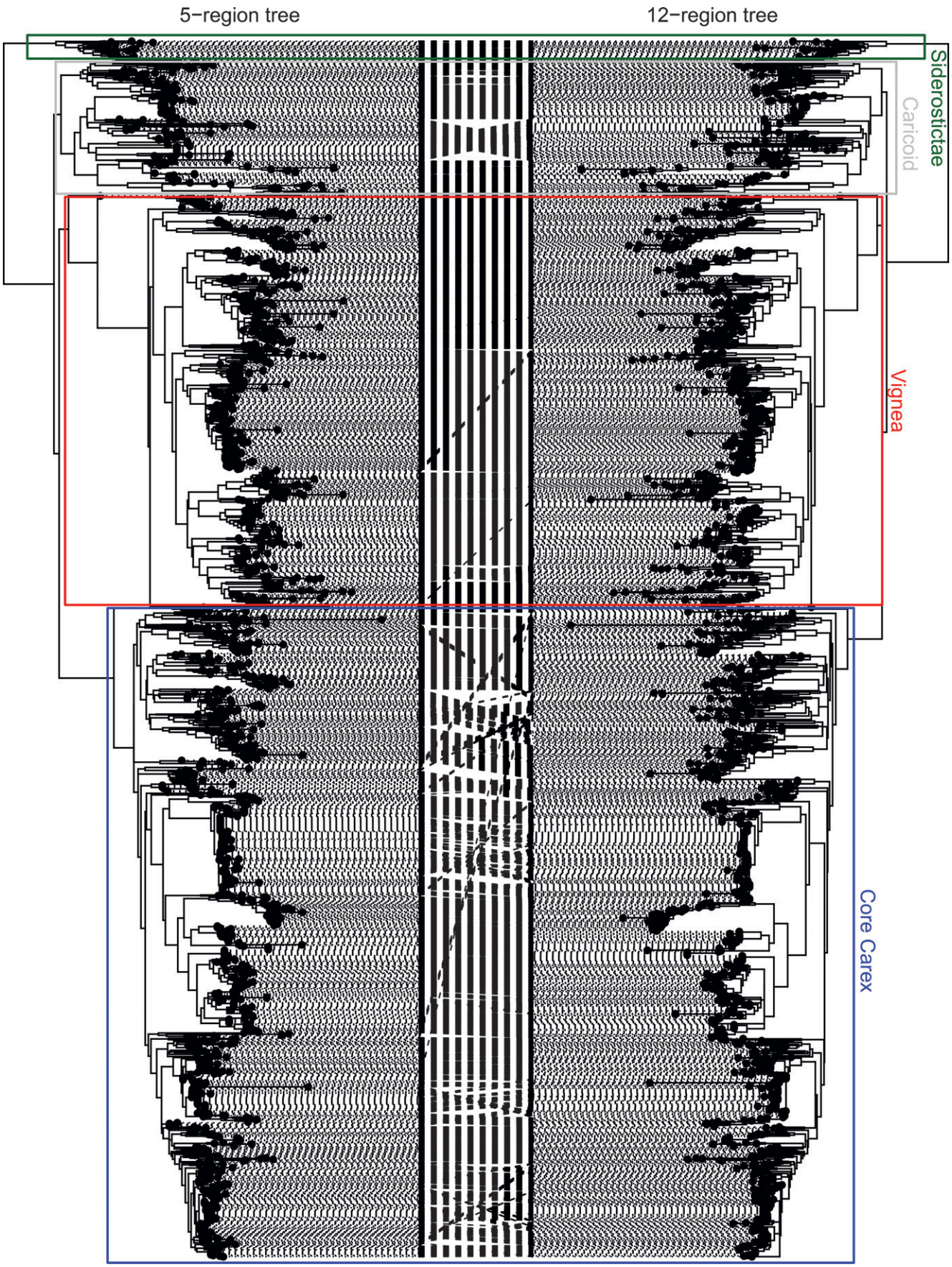

FIG. 5. 5-region phylogeny compared with 12-region phylogeny. Maximum likelihood phylogenies from the 5-region data matrix and the 12-region data matrix were pruned to shared species and plotted using the cophylo function in phytools. Lines connecting the phylogenetic trees represent shared taxa and indicate topological similarities and differences. Colored boxes indicate the major *Carex* clades.

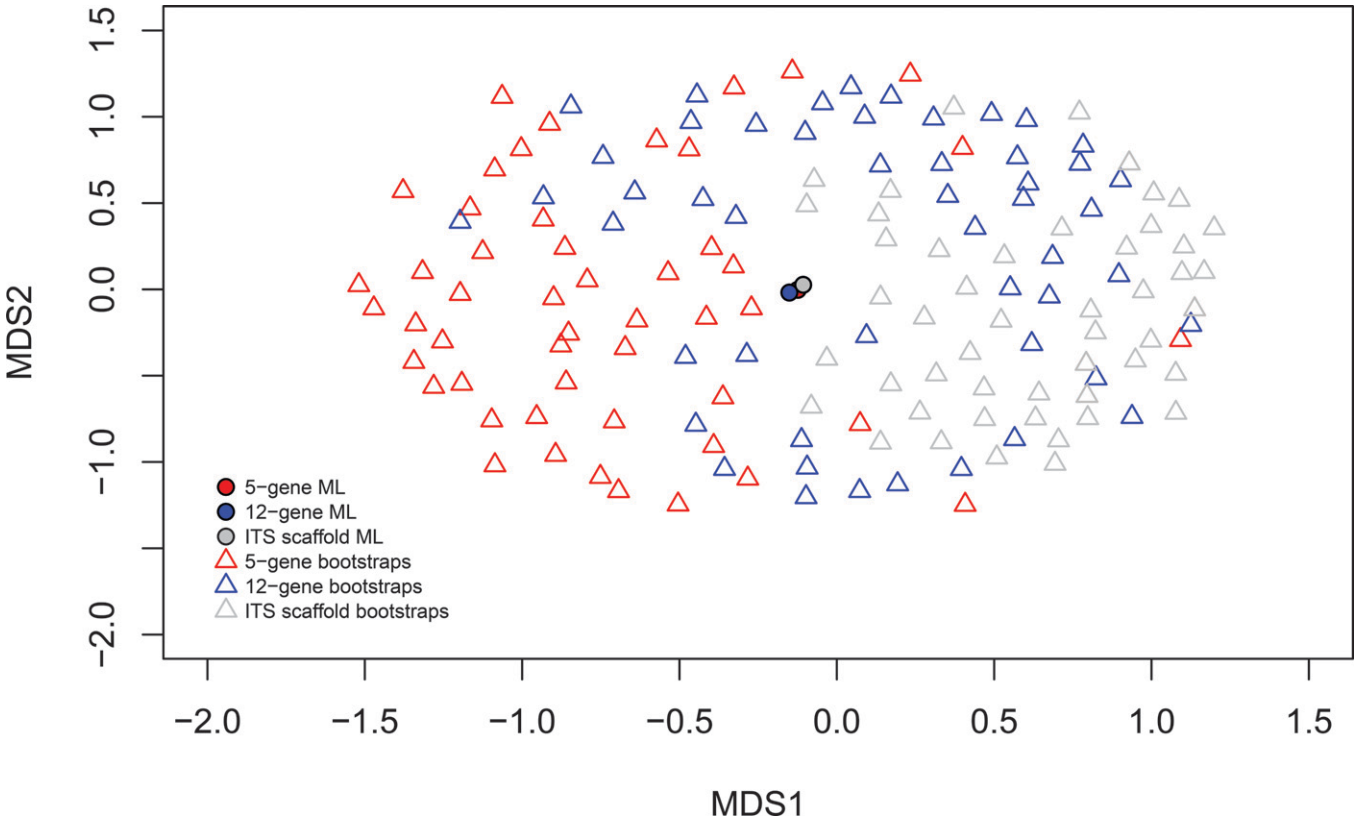

FIG. 6. Ordination of 12-region bootstrap and ML trees, 5-region bootstrap and ML trees, and ITS scaffold bootstrap and ML trees. Pairwise topological distances were calculated between all trees using Penny and Hendy's (1985) tree distance. Ordination was performed using multidimensional scaling in the vegan package of R.

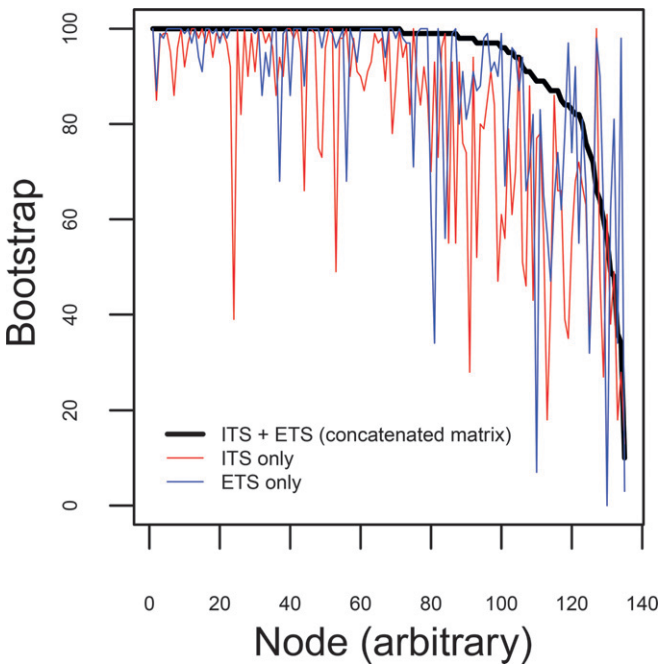

FIG. 7. Bootstrap support: ITS, ETS, and ITS + ETS concatenated. All individuals for which both ITS and ETS had been sampled were exported in three datasets: ITS, ETS, and a concatenated dataset. Maximum likelihood bootstraps for the two individual analyses were mapped back to the comparable nodes in the concatenated matrix analysis; comparable nodes were defined as nodes that had the same descendants in all three trees. Bootstrap in the concatenated matrix was at or above the maximum bootstrap for ITS and ETS for almost all nodes.

normalize taxonomic disparity. We recommend using TDI in a largely heuristic manner, accompanied by manual inspection of the relevant trees to detect potential misidentifications, gene sampling issues, or taxonomically complicated species (cf. Global *Carex* Group 2016). Gross misidentifications could easily be weeded out in further analyses, and tools to do so are provided in the morton package.

Of 769 species labels in the 5-region tree, 441 have two or more individuals; of those, 22% of those species labels are monophyletic. In the 12-region tree, 452 of 771 species labels have two or more individuals, 20% of which are monophyletic. In the ITS scaffold tree, 345 of 676 species labels have two or more individuals, and 42% of these named taxa are monophyletic; an additional 31% have a TDI of 10 or lower (Fig. 8; Table S8A–C).

DISCUSSION

In this paper, we introduce a practical approach to aggregating NCBI data to the individual or specimen level rather than the species level and provide a toolkit (https://github.com/andrew-hipp/morton) to assist in the informatics. Our efforts to identify individual specimens from the database were fairly successful despite the fact that the pieces needed to identify specimens—collector name and number, collection location and accession number, isolate, clone—are not fully atomized or normalized in NCBI and have not always been entered in consistent ways. We have automated numerous steps in extracting these elements, and with additional manual manipulation of the data we arrived at an average of 1.85 +/− 0.95 DNA regions per specimen. We estimate that

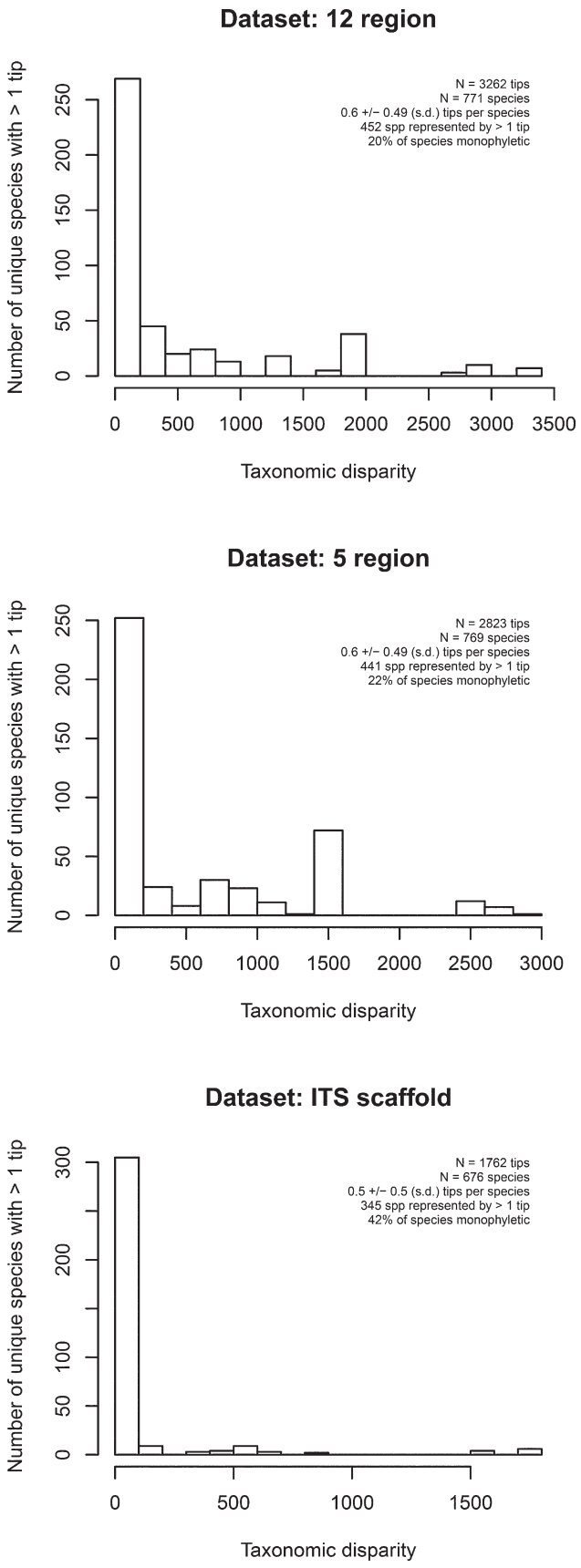

FIG. 8. Taxonomic disparity: 5-region, 12-region, and ITS scaffold datasets. Taxonomic disparity by species, calculated in this study as the difference between the smallest clade that includes all individuals labelled as a given species and the number of individuals with that species label, summarized for three data matrices.

we spent ca. 100 person-hours manually curating specimen IDs for the 7994 sequence records, a relatively small time investment for a dataset of this magnitude.

Success at identifying individual specimens relied on proper and thorough entry of specimen data, especially voucher information on collection and collector number. The amount of information necessary to distinguish individuals varied from author to author based on the size of their study. In several studies, the only identifying information provided for a sequence was taxonomic identification and the author of the paper in which the sequence was published. In such cases, our automated parsing was unable to distinguish individuals, so manual examination of the data was necessary to tease out individuals. With no further specimen information, we generally assumed that all sequences for a given species in a given study came from a single individual, provided that a single sequence was provided per DNA region for that individual.

Our approach offers at least three benefits over the status quo of aggregating to species, ignoring specimens: (1) assessing species monophyly and reliable identification of tips; (2) allowing redeterminations of specimens to inform otherwise unlinked sequence records; and (3) improving species-level phylogenies through more informed concatenation of sequence data and selection of representative individuals.

Assessing species monophyly, and allowing redeterminations of specimens to inform multiple sequence records—The low taxonomic disparity index (TDI) demonstrated by the majority of species provides important data on the validity of those taxa as phylogenetic species, the concordance between morphological classification and phylogenetic classification, and the validity of NCBI data. In the ITS scaffold tree, for example, 40% of 327 species with more than one tip had a TDI of 0, and 71% had a TDI of 10 or lower (Fig. 8; Table S8A–C). It seems unlikely that frequent violations of species monophyly in this study reflect a widespread misunderstanding of species boundaries among practicing caricologists, who are an exceptionally attentive group of taxonomists. Rather, such violations are likely to more often represent misidentifications that would be missed in a traditional data-harvesting approach to phylogenetic analysis from NCBI data, or phylogenetic error due to gene sampling issues or sequencing noise. This argues for taking an approach such as the one presented here as a means to selecting which individual to represent a species in downstream analyses of NCBI data.

Some small values for TDI, however, probably bespeak true non-monophyly of species, discordance between taxonomic species and phylogeny of the genes studied. For example, the non-monophyly of *C. deweyana* Schwein. (TDI = 4, 4 individuals sampled) in the ITS scaffold tree may reflect the fact that two varieties of *C. deweyana* — *C. deweyana* var. *deweyana* and *C. deweyana* var. *senanensis* (Ohwi) T. Koyama — are perhaps better thought of as separate species than as varieties of a single species (Ford et al. 2006). This type of discordance is likely to result in relatively small TDI, as discordance between phylogenetic and taxonomic species will more often involve fine-scale relationships than transgressions of deep phylogenetic divergences. The misclassification of an individual into the wrong species — simple misidentification — or lab errors among close relatives will also cause an increase in TDI and may be difficult to differentiate from discordance between phylogenetic and taxonomic species without additional study. Large TDI values are likely due to lab error or profound errors of identification: 40 species names in the

ITS scaffold dataset had a TDI of 100 or greater (Fig. 8), including such distinctive species as *C. aurea* Nutt. (TDI = 1535, N = 3) and *C. gibba* Wahlenb. (TDI = 575, N = 21). Of these remarkable examples, one was the subject of an earlier study of divergent paralogues (*C. gibba*), and is thus known to include problematic sequences (King and Roalson 2008); and one appears to be a misidentification based on habitat affinities (a putative *C. aurea* growing in a gravel beach, genotyping as *C. glareosa*; GenBank Accession JN999020, GI:359389285). In such cases, the TDI can be a useful tool for quickly identifying problematic specimens that are probably best removed from analysis.

Improving species-level phylogenies through better-informed selection of individuals—For species-level phylogenies, rogue-taxon analysis is useful for identifying tips that are phylogenetically unstable due to poor sequencing of loci or genealogical discordance among loci (Aberer et al. 2013). However, rogue taxon analyses cannot help identify whether a given individual is the best representative of the taxon to which it has been ascribed. Our approach, by contrast, can be used to recognize potentially misidentified individuals by noting tips that are outliers with respect to other tips of the same name. In the examples presented above, rank-ordered TDI could be used to guide manual inspection of sequence data to remove individuals that are clearly misplaced or taxa for which placement is ambiguous, based on conflict among placement of individuals. Individuals could then be chosen to minimize missing data and the impact of the individual on apparent monophyly of the species it represents. This approach is complementary to rogue taxon analysis but fundamentally different from it. Rather than aiming at phylogenetic stability and removing problematic branches based on their movement among trees in a bootstrap set, the approach we describe aims at selecting individuals that best represent their species and are sequenced for the largest number of loci.

Generalizability of our approach—We have run our specimen-parsing scripts on NCBI sequence downloads for *Euphorbia* L. (Euphorbiaceae) and *Quercus* L. (Fagaceae) to validate that our approach has the potential to work with other datasets. After automated parsing, an individual researcher will still need to work through the data table to clean up collector names, collector numbers, collection names, and accession numbers. This portion of the work is perhaps best shared between a worker who is not familiar with the group and a more experienced researcher who knows many of the collectors and collections. Additionally, it will probably be quickest for an experienced researcher to bin the heterogeneous labels for a given locus to a single locus name; in our dataset, for example, the ITS regions were housed under "contains 18S ribosomal RNA, internal transcribed spacer 1, 5.8S ribosomal RNA, internal transcribed spacer 2, and 28S ribosomal RNA," "contains internal transcribed spacer 1, 5.8S ribosomal RNA, internal transcribed spacer 2," and 13 other unique names. Rather than writing a rule for assigning locus identity, it is probably easier for an experienced researcher to judge what labels represent the same locus and provide the unified label for this locus. Again, the task is not difficult. In our dataset, there were only 160 unique DNA region descriptions for nearly 8,000 sequence records, and these were easily binned to locus in about an hour. After parsing and cleanup of the collection data and gene region names, concatenation of separated loci (e.g. each of ITS and *trn*L-*trn*F when they are sequenced and submitted to NCBI as separate pieces), the mapping of individuals to gene regions, the production of graphical representations of the data, and the generation of summary statistics can be easily automated to facilitate data exploration using scripts provided in the morton package or modifications thereof.

Conclusions and next steps—The *Carex* dataset constructed for this study is one of the most inclusive to date for this large genus, comprising analyzable datasets between ca. 670 and 790 *Carex* species (ITS scaffold and 12 / 5-region datasets respectively). It is also the only NCBI-based supermatrix study of which we are aware that puts specimens rather than taxa at the center, and it may thus serve as a model for analyses of other large taxa. The study recovers the four major *Carex* clades—core *Carex*, *Siderostictae*, *Vignea*, and the Caricoid clade—and demonstrates that there is some topological variation among datasets within clades (Fig. 5). The study also supports the monophyly of approximately 345 morphological species based on the ITS scaffold dataset, 441 based on the 5-region dataset (Fig. 8; Table S8A–C). While the data are not consistently coded for geographic origin of sample, of those monophyletic species with three or more tips, 93% were collected by two or more different collectors, suggesting that this apparent monophyly is not likely to be due simply to collections made from the same population. The approach we present here may thus serve in investigations of species monophyly and genetic coherence beyond single studies.

Our study suggests the need for global databases that integrate specimen and DNA sequence data. While specimen databases have not fully lived up to the dream of data flow between collections and from users to curators, there has been a lot of improvement: users of the Symbiota system can, for example, share annotations among collections and readily incorporate feedback from scientists around the globe. As a next step, we would like to see direct links between sequence data deposited in public DNA databases and specimen data housed in the world's herbaria and zoological museums. Thoughtful integration of these databases or protocols for communication among them would facilitate downstream use of specimen-level sequence data. This in turn would propel taxonomic enterprises worldwide, allowing systematists to annotate sequence data in the same way they annotate specimens, and at the same time.

Acknowledgments. We thank James Smith for his help in organizing this proceedings, and American Society of Plant Taxonomists and the Botany 2015 organizing committee for supporting the symposium in which the Global *Carex* Group papers in this issue of *Systematic Botany* were presented: "Ecological diversification and niche evolution in the rate zone's largest genus: *Carex*." Chuck Cannon and two anonymous reviewers provided feedback on a draft of this manuscript. Funding for this work was provided by the National Science Foundation (Award #1255901 to ALH and MJW and Award #1256033 to EHR), including an REU supplement that supported KKP's work.

Literature Cited

Aberer, A. J., D. Krompass, and A. Stamatakis. 2013. Pruning rogue taxa improves phylogenetic accuracy: an efficient algorithm and webservice. *Systematic Biology* 62: 162–166.

Bayer, R. J. and J. R. Starr. 1998. Tribal Phylogeny of the Asteraceae based on two non-coding chloroplast sequences, the *trn*L intron and the *trn*L/*trn*F intergenic spacer. *Annals of the Missouri Botanical Garden* 85: 242–256.

Benson, D. A., I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and D. L. Wheeler. 2007. GenBank. *Nucleic Acids Research* 35(suppl 1): D21–D25.

Derieg, N. J., S. J. Weil, A. A. Reznicek, and L. P. Bruederle. 2013. *Carex viridistellata* (Cyperaceae), a rare new species of the prairie peninsula. *Systematic Botany* 38: 82–91.

De Queiroz, K. 2007. Species concepts and species delimitation. *Systematic Biology* 56: 879–886.

Dragon, J. A. and D. S. Barrington. 2009. Systematics of the *Carex aquatilis* and *C. lenticularis* lineages: Geographically and ecologically divergent sister clades of *Carex* section *Phacocystis* (Cyperaceae). *American Journal of Botany* 96: 1896–1906.

Edgar, R. C. 2004a. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* 32: 1792–1797.

Edgar, R. C. 2004b. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 5: 113.

Edwards, E. J. and S. A. Smith. 2010. Phylogenetic analyses reveal the shady history of C4 grasses. *Proceedings of the National Academy of Sciences USA* 107: 2532–2537.

Escudero, M., V. Valcarcel, P. Vargas, and M. Luceño. 2008. Evolution in *Carex* L. sect. *Spirostachyae* (Cyperaceae): A molecular and cytogenetic approach. *Organisms, Diversity & Evolution* 7: 271–291.

Escudero, M., V. Valcarcel, P. Vargas, and M. Luceño. 2009. Significance of ecological vicariance and long-distance dispersal in the diversification of *Carex* sect. *Spirostachyae* (Cyperaceae). *American Journal of Botany* 96: 2100–2114.

Escudero, M., A. L. Hipp, T. F. Hansen, K. L. Voje, and M. Luceño. 2012. Selection and inertia in the evolution of holocentric chromosomes in sedges (*Carex*, Cyperaceae). *The New Phytologist* 195: 237–247.

Ford, B. A., H. Ghazvini, R. F. C. Naczi, and J. R. Starr. 2012. Phylogeny of *Carex* subg. *Vignea* (Cyperaceae) based on amplified fragment length polymorphism and nrDNA data. *Systematic Botany* 37: 913–925.

Ford, B. A., M. Iranpour, R. F. C. Naczi, J. R. Starr, and C. A. Jerome. 2006. Phylogeny of *Carex* subg. *Vignea* (Cyperaceae) based on noncoding nrDNA sequence data. *Systematic Botany* 31: 70–82.

Gebauer, S., J. R. Starr, and M. Hoffmann. 2014. Parallel and convergent diversification in two northern hemispheric species-rich *Carex* lineages (Cyperaceae). *Organisms, Diversity & Evolution* 14: 247–258.

Global *Carex* Group. 2015. Making *Carex* monophyletic (Cyperaceae, tribe Cariceae): a new broader circumscription. *Botanical Journal of the Linnean Society* 179: 1–42.

Global *Carex* Group. 2016. Megaphylogenetic specimen-level approaches to the *Carex* (Cyperaceae) phylogeny using ITS, ETS, and *matK* sequences: Implications for classification. *Systematic Botany* 41: 500–518.

Gries, C., E. Gilbert, and N. Franz. 2014. Symbiota – A virtual platform for creating voucher-based biodiversity information communities. *Biodiversity Data Journal* 2: e1114, doi: 10.3897/BDJ.2.e1114.

Guralnick, R. P., J. Wieczorek, R. Beaman, and R. J. Hijmans. 2006. BioGeomancer: automated georeferencing to map the world's biodiversity data. *PLoS Biology* 4: e381, doi: 10.1371/journal.pbio.0040381.

Hausdorf, B. 2011. Progress toward a general species concept. *Evolution* 65: 923–931.

Hendrichs, M., S. Michalski, D. Begerow, F. Oberwinkler, and F. H. Hellwig. 2004a. Phylogenetic relationships in *Carex*, subgenus *Vignea* (Cyperaceae), based on ITS sequences. *Plant Systematics and Evolution* 246: 109–125.

Hendrichs, M., F. Oberwinkler, D. Begerow, and R. Bauer. 2004b. *Carex*, subgenus Carex (Cyperaceae) – A phylogenetic approach using ITS sequences. *Plant Systematics and Evolution* 246: 89–107.

Hinchliff, C. E. and E. H. Roalson. 2013. Using supermatrices for phylogenetic inquiry: an example using the sedges. *Systematic Biology* 62: 205–219.

Hipp, A. L. 1998. A checklist of carices for prairies, savannas and oak woodlands of southern Wisconsin. *Transactions of the Wisconsin Academy of Sciences Arts and Letters* 86: 77–99.

Hipp, A. L., A. A. Reznicek, P. E. Rothrock, and J. A. Weber. 2006. Phylogeny and classification of *Carex* section *Ovales* (Cyperaceae). *International Journal of Plant Sciences* 167: 1029–1048.

Jiménez-Mejías, P., M. Escudero, S. Guerra-Cárdenas, K. A. Lye, and M. Luceño. 2011. Taxonomic delimitation and drivers of speciation in the Ibero-North African *Carex* sect. *Phacocystis* river-shore group (Cyperaceae). *American Journal of Botany* 98: 1855–1867.

King, M. G. and E. H. Roalson. 2008. Exploring evolutionary dynamics of nrDNA in *Carex* subgenus *Vignea* (Cyperaceae). *Systematic Botany* 33: 514–524.

Lang, D. T. the CRAN Team. 2015. XML: Tools for Parsing and Generating XML Within R and S-Plus. R package version 3.98–1.3. http://CRAN.R-project.org/package=XML.

Maguilla, E., M. Escudero, M. J. Waterway, A. L. Hipp, and M. Luceño. 2015. Phylogeny, systematics and trait evolution of *Carex* section *Glareosae*. *American Journal of Botany* 102: 1128–1144.

Molina, A., K.-S. Chung, and A. L. Hipp. 2015. Molecular and morphological perspectives on the circumscription of *Carex* section *Heleoglochin* (Cyperaceae). *Plant Systematics and Evolution* 301: 2419–2439.

Oksanen, J., F. Guillaume Blanchet, R. Kindt, P. Legendre, P. R. Minchin, R. B. O'Hara, G. L. Simpson, P. Solymos, M. Henry, H. Stevens, and H. Wagner. 2015. vegan: Community Ecology Package. R package version 2.3–1. https://CRAN.R-project.org/package=vegan.

Paradis, E., J. Claude, and K. Strimmer. 2004. APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* 20: 289–290.

Parr, C. S., N. Wilson, P. Leary, K. S. Schulz, K. Lans, L. Walley, J. A. Hammock, A. Goddard, J. Rice, M. Studer, J. T. G. Holmes, and R. J. Corrigan Jr. 2014. The Encyclopedia of Life v2: providing global access to knowledge about life on Earth. *Biodiversity Data Journal* 2: e1079, doi: 10.3897/BDJ.2.e1079.

Penny, D. and M. D. Hendy. 1985. The use of tree comparison metrics. *Systematic Zoology* 34: 75–82.

R Core Team. 2015. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Revell, L. J. 2012. phytools: An R package for phylogenetic comparative biology (and other things). *Methods in Ecology and Evolution* 3: 217–223.

Reznicek, A. A. 1990. Evolution in sedges (*Carex*, Cyperaceae). *Canadian Journal of Botany* 68: 1409–1432.

Reznicek, A. A. 1993. A revision of the Mexican members of genus *Carex*, section *Ovales* (Cyperaceae). *Contributions from the University of Michigan Herbarium* 19: 97–136.

Roalson, E. H. and E. A. Friar. 2004. Phylogenetic relationships and biogeographic patterns in North American members of *Carex* section *Acrocystis* (Cyperaceae) using nrDNA ITS and ETS sequence data. *Plant Systematics and Evolution* 243: 175–187.

Robertson, T., M. Döring, R. Guralnick, D. Bloom, J. Wieczorek, K. Braak, J. Otegui, L. Russell, and P. Desmet. 2014. The GBIF Integrated Publishing Toolkit: facilitating the efficient publishing of biodiversity data on the internet. *PLoS One* 9: e102623, doi: 10.1371/journal.pone.0102623.

Shekhovtsov, S. V., I. N. Shekhovtsova, and S. E. Peltek. 2012. Phylogeny of Siberian species of *Carex* sect. *Vesicariae* based on nuclear and plastid markers. *Nordic Journal of Botany* 30: 343–351.

Shimodaira, H. and M. Hasegawa. 1999. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. *Molecular Biology and Evolution* 16: 1114–1116.

Smith, V. S., S. Rycroft, B. Scott, E. Baker, L. Livermore, A. Heaton, K. Bouton, D. N. Koureas, and D. Roberts. 2012. Scratchpads 2.0: a virtual research environment infrastructure for biodiversity data. Accessed at http://scratchpads.eu on 2015–12–18.

Stamatakis, A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30: 1312–1313.

Starr, J. R. and B. A. Ford. 2009. Phylogeny and evolution in Cariceae (Cyperaceae): current knowledge and future directions. *Botanical Review* 75: 110–137.

Starr, J. R., R. J. Bayer, and B. A. Ford. 1999. The phylogenetic position of *Carex* section *Phyllostachys* and its implications for phylogeny and subgeneric circumscription in *Carex* (Cyperaceae). *American Journal of Botany* 86: 563–577.

Starr, J. R., F. H. Janzen, and B. A. Ford. 2015. Three new, early diverging *Carex* (Cariceae, Cyperaceae) lineages from East and Southeast Asia with important evolutionary and biogeographic implications. *Molecular Phylogenetics and Evolution* 88: 105–120.

Waterway, M. J., T. Hoshino, and T. Masaki. 2009. Phylogeny, species richness, and ecological specialization in Cyperaceae tribe Cariceae. *Botanical Review* 75: 138–159.

Waterway, M. J. and J. R. Starr. 2007. Phylogenetic relationships in tribe Cariceae (Cyperaceae) based on nested analyses of three molecular data sets. *Aliso* 23: 165–192.

Yano, O., H. Ikeda, X.-F. Jin, and T. Hoshino. 2014. Phylogeny and chromosomal variations in East Asian *Carex, Siderostictae* group (Cyperaceae), based on DNA sequences and cytological data. *Journal of Plant Research* 127: 99–107.