Documenting distributions of hyperdiverse groups (terrestrial invertebrates) through grid-based spatial sampling, high-throughput sequencing, and occupancy modeling

# Author Information

**Matthew L. Bowser,\* John M. Morton**

**M.L. Bowser** USFWS Kenai National Wildlife Refuge, Soldotna, Alaska 99669

**J.M. Morton** USFWS Kenai National Wildlife Refuge, Soldotna, Alaska 99669 (retired)

\*Corresponding author: matt\_bowser@fws.gov

# Introduction

Alaska National Wildlife Refuges have been given mandates to conserve natural biodiversity, habitats, biological integrity, and environmental health (96th Congress [1980](#ref-96th_Congress_1980); 105th Congress [1997](#ref-105th_Congress_1997)). Woodward and Beever ([2010](#ref-Woodward_Beever_2010)) noted that, under these broad purposes of National Wildlife Refuges in Alaska to protect natural landscapes and entire ecosystems, they must develop methods that can successfully monitor biodiversity.

Morton et al. ([2009](#ref-Morton_et_al_2009)) determined that the best way to measure biodiversity at the scale of Alaska Refuges was to document occurrences and co-occurrences of many species on a set of points distributed widely across the landscape. This kind of sampling scheme would yield species distributions and assemblages and, if repeated over time, could be used to assess whether Alaska Refuges were fulfilling their conservation purposes. Magness et al. ([2008](#ref-Magness_et_al_2008)) demonstrated the utility of this sampling design for generating species distribution data that could serve this purpose.

A key deficiency identified by Morton et al. ([2009](#ref-Morton_et_al_2009)) was an inability to feasibly monitor most of biodiversity due to the high cost and lengthy time required to obtain identifications for species in hyperdiverse groups, a well-known problem that has been referred to as the taxonomic impediment (Taylor [1983](#ref-Taylor_1983)). Since that time, high-throughput sequencing (HTS) methods have emerged as a means to obtain many identfications from environmental samples. Multiple studies (Gibson et al. [2015](#ref-Gibson_et_al_2015); Hajibabaei et al. [2016](#ref-Hajibabaei_et_al_2016); Bowser et al. [2017](#ref-Bowser_et_al_2017), [2020](#ref-Bowser_et_al_2020); Bush et al. [2019](#ref-Bush_2019)) have demonstrated that assemblages of diverse taxa can be efficiently identified using these methods.

The second deficiency recognized by Morton et al. ([2009](#ref-Morton_et_al_2009)) and Magness et al. ([2008](#ref-Magness_et_al_2008)) was the inability to account for imperfect detection. When a species is present but there is a chance that it will not be detected, this may lead to false negatives (i.e., recording an absence where a species is present), leading to bias in the resulting estimates of species distributions. MacKenzie et al. ([2002](#ref-MacKenzie_et_al_2002), [2006](#ref-MacKenzie_et_al_2006)) presented occupancy modeling methods that can yield unbiased estimates of species occurrence and distributions. These models explicitly account for imperfect detection through spatially or temporally repeated sampling.

The biological inventory work of Bowser et al. ([2020](#ref-Bowser_et_al_2020)) was designed in part to provide the kind of occurrence data required for monitoring distributions of multiple species (i.e., monitoring assemblages of species) and accounting for imperfect detection by spatial subsampling. Our intent in this paper is to test the usefulness of the methods of Bowser et al. ([2020](#ref-Bowser_et_al_2020)) for delivering unbiased estimates of species occurrence and distributions of hyperdiverse taxa.

# Materials and Methods

In our detailed presentation of our methods and even our thought process for why methods were chosen, we are seeking to follow the recommendations of Devarajan et al. ([2020](#ref-Devarajan_et_al_2020)).

## Study area

The study area, described by Bowser et al. ([2020](#ref-Bowser_et_al_2020)), was a 938 ha portion of the Slikok Creek watershed on the Kenai National Wildlife Refuge (KNWR) with a bounding box from 60.44° to 60.47° latitude and from -151.10° to -151.03° longitude.

Based on data from the U.S. National Land Cover Database (Homer et al. [2015](#ref-Homer_et_al_2015)), the land area of this study area was coverd mostly by mixed forest (29%), evergreen forest (22%), emergent herbaceous wetlands (13%), deciduous forest (12%), woody wetlands (11%), and open water (10%). Common trees in well-drained areas were white spruce (*Picea glauca* (Moench) Voss), Alaska birch (*Betula pendula* subsp. *mandshurica* (Regel) Ashburner & McAll.), quaking aspen (*Populus tremuloides* Michx.), and cottonwood (*Populus* × *hastata* Dode). Black spruce (*Picea mariana* Britton, Sterns & Poggenb.) was the dominant tree in forested wetlands. Herbaceous wetlands were characterized by mosses, sedges, and low shrubs. Shallow, eutrophic lakes (Headquarters Lake, Nordic Lake, and smaller, unnamed lakes) accounted for most of the open water in the study area.

This study area was chosen because it could be conveniently sampled from our office at the Kenai National Wildlife Refuge, which lies at 60.4647°N, 151.0735°W within the Slikok Creek watershed and because it could serve well as an analog of a real conservation unit. Like most conservation units and like the Kenai National Wildlife Refuge as a whole, this study area includes varied habitat types; arbitrary, artificial, and ecologically open boundaries that have no meaning to wildlife; and a range of levels of human disturbance from fairly remote, undisturbed areas to developed land.

Sampling design, field methods, and identification methods were detailed by Bowser et al. ([2020](#ref-Bowser_et_al_2020)), but they are summarized here.

## Sampling design

We chose a systematic sampling design because of its simplicity and because could be used to obtained unbiased estimates representing the entire study area. A grid with 500 m spacing between points was chosen by using the coordinates of the centroids of the 250 m pixels from the Alaska eMODIS product (Jenkerson et al. [2010](#ref-Jenkerson_2010)), choosing every other centroid to make a grid of sites having 500 m spacing. The resulting sample frame consisted of 40 terrestrial sites.

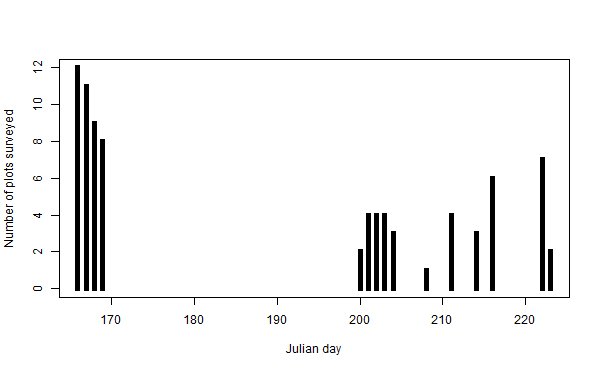
## Field methods

Sampling sites were permanently marked by driving 122 cm long, 13 mm diameter SunGUARD Smart Stake™ fibreglass rods into the ground, then labelling them with aluminium tags. During the survey period, sites were also temporarily marked with high-visibility forestry flagging tape.

Sweep net samples of terrestrial arthropods were collected in two sampling periods: from 14 to 17 June 2016 and from from 18 July to 9 August 2016. During each sampling period, two samples were collected at each plot for a total of 160 sweep net samples (40 plots × 2 samples/plot × 2 visits/plot).

The timing of the first sampling window was determend by concurrent sampling of birds using variable circular plot methods. This meant that all sampling took place in the morning. Two field crews of two people were available for full work days so that all 40 plots could be visited in a short, 4 day time window.

During the second sampling window, invertebrates were sampled concurrently with vegetation sampling. Even though it would have been ideal to complete the second sampling in a short time window, this was not accomplished because only a single, two person crew was available; vegetation sampling took longer than bird surveys; and because other time demands on surveyors’ time meant that not every day could be devoted to this project. This resulted in a much longer sampling window of 23 days.

  
Number of plots surveyed per day.

Arthropods were sampled within a 100 m2, 5.64 m radius, circular plot using the center stake as plot center. We split the plot into two semicircular subplots divided along the north-south axis. Each subplot was independently sweep-netted, such that the entire area was swept from the ground surface up to a height of roughly 2 m. No defined pattern of sweeping was enforced, but we ensured that all substrates and macrohabitats within reach were swept once within a time limit of 5 min per sample. We used a BioQuip™ model 7112CP 30.5 cm diameter net with a BioQuip™ model 7312AA 30.5 cm extension handle and a BioQuip™ model 7112CPA net bag with a mesh size of approximately 8 × 9 meshes/mm.

All specimens were collected into a single Nalgene® model 2104-0008 wide-mouth 250 ml bottle containing UniGard -100 propylene glycol antifreeze.

## Laboratory methods

Invertebrates and fragments of invertebrates in the sweep net samples were separated from debris by hand under a stereomicroscope.

Due to budget limitations, we could only process 125 of the 160 sweep net samples. We considered that there was a trade-off between choosing as many sites as possible, which would be optimal for the inventory objectives of the study to document the occurrence of as many species as possible, and choosing as many samples per site as possible, which would have been better for obtaining estimates of imperfect detection and occupancy. Because we knew so little about the diversity of terrestrial invertebrates that could be encountered, we chose to make documenting as many species as possible our priority over optimization for obtaining occupancy metrics.

We selected all 80 samples taken from the east side of each plot (40 plots × 1 sample/plot × 2 visits/plot). To choose 45 samples from the remaining 80, we selected plots spatially. First, we chose 20 samples from plots at 1 km spacing (10 plots × 2 visits/plot), then we chose 25 of 26 samples from another 13 plots that were maximally distant from these 10 plots (13 plots × 2 visits/plot). These 45 samples from west plot halves were intended to be used for estimating occupancy metrics.

Sweep net samples were shipped to RTL Genomics, Lubbock, Texas (<http://rtlgenomics.com>) for extraction and DNA sequencing steps. For details of sequencing and identifications obtained through a high-throughput sequencing analysis pipeline, see Bowser et al. ([2020](#ref-Bowser_et_al_2020)).

## Molecular identifications

Invertebrate sequences that could not be confidently assigned to described species were assigned to BOLD Barcode Index Numbers (BINs, Ratnasingham and Hebert [2013](#ref-Ratnasingham_Hebert_2013)) if possible. Sequences that could be assigned to neither species nor BINs were given provisional names including labels of the molecular operational taxonomic units (MOTUs, Blaxter et al. [2005](#ref-Blaxter_et_al_2005)), e.g. “*Liriomyza* sp. SlikokOtu253”. For our purposes we considered all of these entities to be species.

We sought to follow the guidelines of Penev et al. ([2017](#ref-Penev_et_al_2017)) for publication of biodiversity data. Species occurrence data have been made available via Arctos (<https://arctosdb.org/>), where they are associated together via an Arctos project (<http://arctos.database.museum/project/10002227>). These occurrence data on Arctos are also provided to the Global Biodiversity Information Facility (<https://www.gbif.org/>).

In the time since Bowser et al. ([2020](#ref-Bowser_et_al_2020)) was published, a handfull of identifications of the records on Arctos have been improved based on new data that became available in the reference databases (Ratnasingham and Hebert [2007](#ref-Ratnasingham_Hebert_2007); Clark et al. [2016](#ref-Clark_et_al_2016)).

A summary of identfications from all HTS occurrences from sweep net samples obtained by Bowser et al. ([2020](#ref-Bowser_et_al_2020)) were downloaded from Arctos on 12 November 2020 (saved search URI: <https://arctos.database.museum/saved/2020-11-12-1157_Slikok_metabarcoding_taxa_list>), yielding a list of 975 unique identifications. For unidentified MOTUs where 10 or more occurrences were observed, we attempted to improve identifications by submitting the sequences to BOLD’s ID Engine (Ratnasingham and Hebert [2007](#ref-Ratnasingham_Hebert_2007)) and BLAST (Altschul et al. [1990](#ref-Altschul_et_al_1990)) searches. We obtained no new identifications.

The 2,375 occurrences were downloaded on 12 November 2020 (saved search URI: <https://arctos.database.museum/saved/2020-11-12-1400_Slikok_project_metabarcoding_occurrences>).

## Data analysis

Data processing and analysis was carried out using R version 4.0.2 (R Core Team [2020](#ref-R_Core_Team_2020)); the packages knitr (Xie [2014](#ref-Xie2014), [2015](#ref-Xie2015), [2020](#ref-Xie2020)) and reshape (Wickham [2007](#ref-Wickham_2007)), reshape2 (Wickham [2007](#ref-Wickham_2007)), and rjags (Plummer [2019](#ref-Plummer_2019)); and JAGS version 4.3.0 (Plummer [2003](#ref-Plummer_2003)).

We began with data exploration.

### Occupancy model assumptions

We considered multispecies occupancy model assumptions as presented by Devarajan et al. ([2020](#ref-Devarajan_et_al_2020)).

A challenge of applying occupancy models to ephemeral, highly vagile animals like winged insects is that assumptions of geographic and demographic closure can seldom be satisfied. The population sizes of many insects can change rapidly over a short time, for example with aphids. The life stages and abundances of some insects are highly seasonal, in some species with adults appearing over a brief seasonal time window. In addition, many species are winged, phoretic on winged species, or can be wind disperesed by ballooning. These species can move freely into and out of the study area on a time scale of hours and they can move onto or out of plots in a matter of seconds.

We were aware that many species in our study area would be unlikely to be detected within our two temporal sampling windows. For example, the stonefly *Utacapnia columbiana* (Claassen, 1924) is abundant in large, glacial streams of the area. The adults emerge in large numbers in March through May and disperse over our study area, where they can be found on vegetation, but the adults were gone by the time we began sampling on 14 June.

Because our “revisits” at each sampling event were spatial replicates taken within 10 min. of each other, there was some degree of closure between the two revisits. Also, ours were removal methods where individuals collected in one sample would not be available for subsequent sampling, but it seems reasonable that one 50 m2 sweep net sample would not remove enough of the invertebrate population to affect the next sampling event one month or more later.

Because we assume that geographic and geographic closure would not apply to at least many terrestrial invertebrate species, it may be appropriate to use staggered-entry occupancy models (Kendall et al. [2013](#ref-Kendall_et_al_2013)).

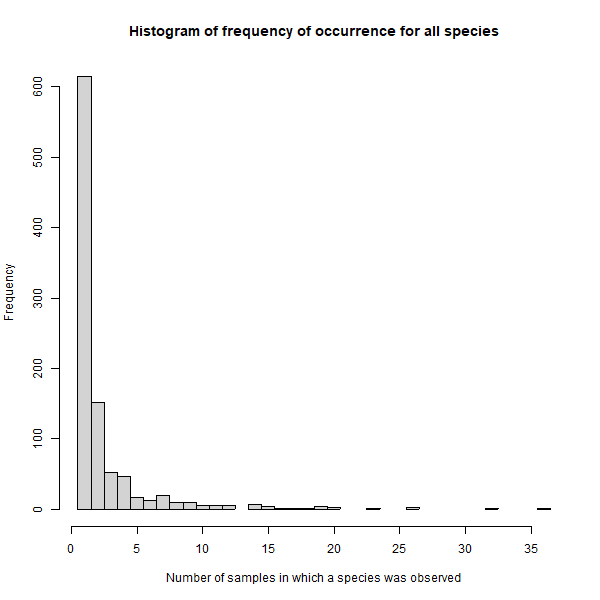
### Occupancy models

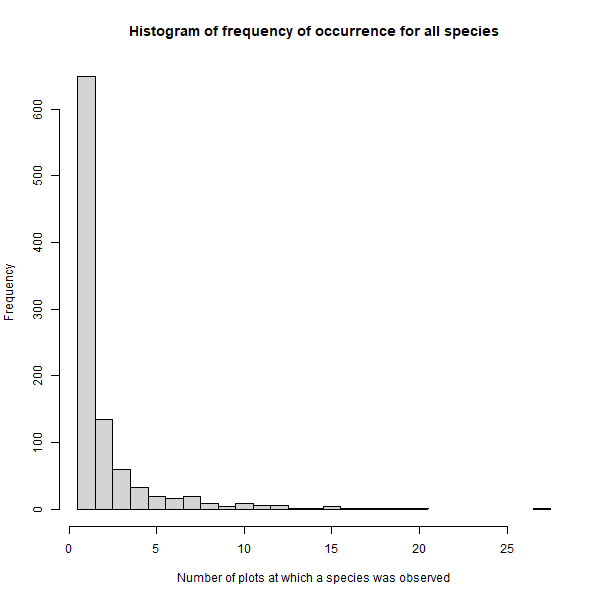
We first tried running a simple multi-species, multi-season occupancy model provided by Joseph ([2013](#ref-Joseph_2013)).

# Results

## Data exploration

Out of the 975 species observed, 615 species (63% of species) were detected in only one of the 125 samples and 649 species (67% of species) were detected at a single plot. The most commonly observed species was the mosquito *Ochlerotatus communis*, observed in 36 samples (29% of samples) and at 27 plots (68% of plots).

  
Histogram of numbers of samples in which each species was observed.

  
Histogram of numbers of plots at which each species was observed.

# References

105th Congress 1997. 111 Stat. 1252 - National Wildlife Refuge System Improvement Act of 1997. Pages 1252–1260 *in* United States Statutes at Large, Volume 111, 105th Congress, 1st Session. [Online] Available: <https://www.govinfo.gov/app/details/STATUTE-111/STATUTE-111-Pg1252>.

96th Congress 1980. Alaska National Interest Lands Conservation Act. *in* United States Statutes at Large, Volume 94, 96th Congress, 2nd Session. U.S. Government Printing Office. [Online] Available: <https://www.govinfo.gov/app/details/STATUTE-94/STATUTE-94-Pg2371>.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. 1990. Basic local alignment search tool. Journal of Molecular Biology **215**: 403–410. doi:[10.1016/S0022-2836(05)80360-2](https://doi.org/10.1016/S0022-2836(05)80360-2).

Blaxter, M., Mann, J., Chapman, T., Thomas, F., Whitton, C., Floyd, R., and Abebe, E. 2005. Defining operational taxonomic units using DNA barcode data. Philosophical Transactions of the Royal Society B **360**: 1935–1943. doi:[10.1098/rstb.2005.1725](https://doi.org/10.1098/rstb.2005.1725).

Bowser, M.L., Brassfield, R., Dziergowski, A., Eskelin, T., Hester, J., Magness, D.R., McInnis, M., Melvin, T., Morton, J.M., and Stone, J. 2020. Towards conserving natural diversity: A biotic inventory by observations, specimens, DNA barcoding and high-throughput sequencing methods. Biodiversity Data Journal **8**: e50124. doi:[10.3897/BDJ.8.e50124](https://doi.org/10.3897/BDJ.8.e50124).

Bowser, M.L., Morton, J., Hanson, J., Magness, D., and Okuly, M. 2017. Arthropod and oligochaete assemblages from grasslands of the southern Kenai Peninsula, Alaska. Biodiversity Data Journal **5**: e10792. doi:[10.3897/BDJ.5.e10792](https://doi.org/10.3897/BDJ.5.e10792).

Bush, A., Compson, Z.G., Monk, W.A., Porter, T.M., Steeves, R., Emilson, E., Gagne, N., Hajibabaei, M., Roy, M., and Baird, D.J. 2019. Studying ecosystems with DNA metabarcoding: Lessons from biomonitoring of aquatic macroinvertebrates. Frontiers in Ecology and Evolution **7**: 434. doi:[10.3389/fevo.2019.00434](https://doi.org/10.3389/fevo.2019.00434).

Clark, K., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., and Sayers, E.W. 2016. GenBank. Nucleic Acids Research **44**: D67–D72. doi:[10.1093/nar/gkv1276](https://doi.org/10.1093/nar/gkv1276).

Devarajan, K., Morelli, T.L., and Tenan, S. 2020. Multi-species occupancy models: Review, roadmap, and recommendations. Ecography **43**: 1612–1624. doi:[10.1111/ecog.04957](https://doi.org/10.1111/ecog.04957).

Gibson, J.F., Shokralla, S., Curry, C., Baird, D.J., Monk, W.A., King, I., and Hajibabaei, M. 2015. Large-scale biomonitoring of remote and threatened ecosystems via High-Throughput Sequencing. PLOS One **10**: e0138432. doi:[10.1371/journal.pone.0138432](https://doi.org/10.1371/journal.pone.0138432).

Hajibabaei, M., Baird, D.J., Fahner, N.A., Beiko, R., and Golding, G.B. 2016. A new way to contemplate Darwin’s tangled bank: How DNA barcodes are reconnecting biodiversity science and biomonitoring. Philosophical Transactions of the Royal Society of London B: Biological Sciences **371**: 20150330. doi:[10.1098/rstb.2015.0330](https://doi.org/10.1098/rstb.2015.0330).

Homer, C.G., Dewitz, J.A., Yang, L., Jin, S., Danielson, P., Xian, G., Coulston, J., Herold, N.D., Wickham, J.D., and Megown, K. 2015. Completion of the 2011 National Land Cover Database for the conterminous United States – representing a decade of land cover change information. Photogrammetric Engineering and Remote Sensing **81**: 345–354. [Online] Available: <https://www.asprs.org/a/publications/pers/2015journals/PERS_May_2015/HTML/index.html#345>.

Jenkerson, C.B., Maiersperger, T., and Schmidt, G. 2010. eMODIS: A user-friendly data source. Open-File Report, U.S. Geological Survey, Reston, Virginia. [Online] Available: <http://pubs.usgs.gov/of/2010/1055/pdf/OF2010-1055.pdf>.

Joseph, M.B. 2013. Dynamic community occupancy modeling with R and JAGS. [Online] Available: <https://web.archive.org/web/20150929072630/http://www.r-bloggers.com/dynamic-community-occupancy-modeling-with-r-and-jags/>.

Kendall, W.L., Hines, J.E., Nichols, J.D., and Grant, E.H.C. 2013. Relaxing the closure assumption in occupancy models: Staggered arrival and departure times. Ecology **94**: 610–617. [Online] Available: <http://www.jstor.org/stable/23436264> [2020 Dec. 7].

MacKenzie, D.I., Nichols, J.D., Lachman, G.B., Droege, S., Royle, J.A., and Langtimm, C.A. 2002. Estimating site occupancy rates when detection probabilities are less than one. Ecology **83**: 2248–2255. doi:[10.1890/0012-9658(2002)083[2248:ESORWD]2.0.CO;2](https://doi.org/10.1890/0012-9658(2002)083%5B2248:ESORWD%5D2.0.CO;2).

MacKenzie, D.I., Nichols, J.D., Royle, J.A., Pollock, K.H., Bailey, L.L., and Hines, J.E. 2006. Occupancy Estimation and Modeling. Elsevier, New York.

Magness, D.R., Huettmann, F., and Morton, J.M. 2008. Using Random Forests to provide predicted species distribution maps as a metric for ecological inventory & monitoring programs. Pages 209–229 *in* T.G. Smolinski, M.G. Milanova, and A.-E. Hassanien, eds. Applications of Computational Intelligence in Biology: Current Trends and Open Problems. Springer-Verlag, Berlin. doi:[10.1007/978-3-540-78534-7\_9](https://doi.org/10.1007/978-3-540-78534-7_9).

Morton, J.M., Bowser, M.L., Berg, E., Magness, D., and Eskelin, T. 2009. Long Term Ecological Monitoring Program on the Kenai National Wildlife Refuge, Alaska: An FIA adjunct inventory. Pages 1–17 *in* W. McWilliams, G. Moisen, and R. Czaplewski, eds. 2008 Forest Inventory and Analysis (FIA) Symposium; October 21-23, 2008; Park City, UT. U.S. Department of Agriculture, Forest Service, Rocky Mountain Research Station, Fort Collins, Colorado. [Online] Available: <http://www.treesearch.fs.fed.us/pubs/33332>.

Penev, L., Mietchen, D., Chavan, V.S., Hagedorn, G., Smith, V.S., Shotton, D., Tuama, É.Ó., Senderov, V., Georgiev, T., Stoev, P., Groom, Q.J., Remsen, D., and Edmunds, S.C. 2017. Strategies and guidelines for scholarly publishing of biodiversity data. Research Ideas and Outcomes **3**: e12431. doi:[10.3897/rio.3.e12431](https://doi.org/10.3897/rio.3.e12431).

Plummer, M. 2003. JAGS: A program for analysis of bayesian graphical models using gibbs sampling. Pages 1–10 *in* K. Hornik, F. Leisch, and A. Zeileis, eds. Proceedings of the 3rd international workshop on distributed statistical computing. Vienna, Austria. [Online] Available: <https://www.r-project.org/conferences/DSC-2003/Proceedings/Plummer.pdf>.

Plummer, M. 2019. Rjags: Bayesian graphical models using mcmc. [Online] Available: <https://CRAN.R-project.org/package=rjags>.

Ratnasingham, S., and Hebert, P.D.N. 2007. BOLD: The Barcode of Life Data System (www.barcodinglife.org). Molecular Ecology Notes **7**: 355–364. doi:[10.1111/j.1471-8286.2007.01678.x](https://doi.org/10.1111/j.1471-8286.2007.01678.x).

Ratnasingham, S., and Hebert, P.D.N. 2013. A DNA-based registry for all animal species: The Barcode Index Number (BIN) system. PLoS ONE **8**: e66213. doi:[10.1371/journal.pone.0066213](https://doi.org/10.1371/journal.pone.0066213).

R Core Team 2020. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. [Online] Available: <https://www.R-project.org/>.

Taylor, R.W. 1983. Descriptive taxonomy: Past, present, and future. Pages 93–134 *in* E. Highley and R.W. Taylor, eds. Australian Systematic Entomology: A Bicentenary Perspective. CSIRO, Canberra, Australia. [Online] Available: <http://hdl.handle.net/102.100.100/286670?index=1>.

Wickham, H. 2007. Reshaping data with the reshape package. Journal of Statistical Software **21**: 1–20. [Online] Available: <http://www.jstatsoft.org/v21/i12/>.

Woodward, A., and Beever, E.A. 2010. Framework for ecological monitoring on lands of Alaska National Wildlife Refuges and their partners. Report, Reston, Virginia. doi:[10.3133/ofr20101300](https://doi.org/10.3133/ofr20101300).

Xie, Y. 2014. Knitr: A comprehensive tool for reproducible research in R. *in* V. Stodden, F. Leisch, and R.D. Peng, eds. Implementing reproducible computational research. Chapman; Hall/CRC. [Online] Available: <http://www.crcpress.com/product/isbn/9781466561595>.

Xie, Y. 2015. Dynamic documents with R and knitr. 2nd editions. Chapman; Hall/CRC, Boca Raton, Florida. [Online] Available: <https://yihui.org/knitr/>.

Xie, Y. 2020. Knitr: A general-purpose package for dynamic report generation in r. [Online] Available: <https://yihui.org/knitr/>.

# Figures and Figure Captions