rgbif: R client for working with GBIF species occurrence data

Scott Chamberlain*,[a]

[a] *University of California, Berkeley, CA, USA*

Abstract

1. xxx
2. xxx
3. xxx
4. xxxx

Introduction

Perhaps the most fundamental element in many fields of ecology is the inividual. How many individuals of each species in a given location forms the basis for many sub-fields of ecology and evolution. Some research questions necessitate collecting new data, while others can easily take advantage of existing data. In fact, some ecology fields are built largely on existing data, e.g., macro-ecology (Brown, 1995; Beck et al., 2012).

Data on individuals, including which species, and where they're found, can be used for a large number of research questions. Biodiversity records have been used for a suite of other use cases: validating habitat suitability models with real occurrence data (Ficetola et al., 2014); ancestral range reconstruction (Ferretti et al., 2015; María Mendoza et al., 2015); development of invasive species watch lists (Faulkner et al., 2014); evaluate risk of invasive species spread (Febbraro et al., 2013); and effects of climate change on future biodiversity (Brown et al., 2015).

In addition to wide utility, this data is important for conservation. Biodiversity loss is one of the greatest challenges of our time (Pimm et al., 2014). Some have called this the sixth great mass extinction (Ceballos et al., 2015). Given this challenge there is a great need for data on specimen records, whether collected from live sightings in the field or specimens in museums.

There are many online services that collect and maintain specimen records. However, Global Biodiversity Information Facility (hereafter, GBIF, http://www.gbif.org/) is the largest collection of biodiversity

*Corresponding author
Email address: scott(at)ropensci.org (Scott Chamberlain)

records globally, currently with 580 million records, 1.6 million taxa, 15,000 datasets from 770 publishers (figures collected on 2015-10-04). Many large biodiversity warehouses such as iNaturalist (http://www.inaturalist.org/), VertNet (http://vertnet.org/), and USGS's Biodiversity Information Serving Our Nation (BISON; http://bison.usgs.ornl.gov/) all feed into GBIF.

Herein, we describe the `rgbif` library (Chamberlain et al.) for working with GBIF data in the R programming environment (R Core Team, 2014). R is a widely used language in academia, and in non-profit and private sectors. Importantly, R makes it easy to do all of the steps of the research process, including data management, data manipulation and cleaning, statistics, and vizualization. Thus, an R client for getting GBIF data is a powerful tool to facilitate reproducible research.

The rgbif package

The `rgbif` package is completely written in R, uses an MIT license to maximize use everywhere. `rgbif` is developed publicly on GitHub at https://github.com/ropensci/rgbif, where development versions of the package can be installed, and bugs and feature requests reported. Stable versions of `rgbif` can be installed from CRAN, the distribution network for R packages. `rgbif` is part of the rOpenSci project, a developer network making R software to facilitate reproducible research.

Package interface

`rgbif` is designed following the GBIF Application Programming Interface, or API. The GBIF API has four major components: registry, taxonomic names, occurrences, and maps. We ignore maps in `rgbif` as it is concerned with generating maps primarily for web applications. `rgbif` has a suite of functions dealing with each of registry, taxonomic names, and occurrences - we'll go through each in turn describing design and example usage.

GBIF feedback loop

With each request `rgbif` makes to GBIF's API, we send request headers that tell GBIF that the request is coming from `rgbif`, including what version of the package. This will help GBIF know what proportion of requests are coming from this package, and therefore R, as most requests likely will come from `rgbif`.

Registry

The GBIF reqistry API services are spread across four sets of functions:

- Datasets
- Installations
- Networks
- Nodes
- Organizations

Datasets are owned by organizations. Organizations are endorsed by nodes to share datasets with GBIF. Datasets are published through institutions, which may be hosted at another organization. A network is a group of datasets (managed by GBIF).

Datasets

Dataset functions include search, dataset metadata retrieval, and dataset metrics. Searching for datasets is an important part of the discovery process. One can search for datasets on the GBIF web portal. However, programmatic searching using this package is much more powerful. Identifying datasets appropriate for a research question is helpful as you can get metadata for each dataset, and track down dataset specific problems, if any.

The `dataset_search()` function is one way to search for datasets. Here, we search for the query term "oregon", which finds any datasets that have terms matching that term.

```
res <- dataset_search(query = "oregon")
res$data$datasetTitle[1:10]
#>  [1] "SDNHM Birds Collection"
#>  [2] "CM Birds Collection"
#>  [3] "condoncollection"
#>  [4] "Taxonomy in Flux Checklist"
#>  [5] "Wool carder bees of the genus Anthidium in the Western Hemisphere"
#>  [6] "Bryophyte Collection - University of Washington Herbarium (WTU)"
#>  [7] "University of British Columbia Herbarium (UBC) - Bryophytes Collection"
#>  [8] "UWFC Ichthyology Collection"
```

```
#>  [9] "Lichen Collection - University of Washington Herbarium (WTU)"
#> [10] "UWBM Mammalogy Collection"
```

Also, check out `datasets()` and `dataset_suggest()` for searching for datasets.

Dataset metrics. Dataset metrics are another useful way of digging in to figure out what datasets you may want to use. One drawback is that these data are only available for datasets of type *checklist*, but there are quite a lot of them (2483).

Here, we seaerch for dataset metrics for a single dataset, with uuid `ec93a739-1681-4b04-b62f-3a687127a17f`, a checklist of the ants (Hymenoptera: Formicidae) of the World.

```
res <- dataset_metrics(uuid='ec93a739-1681-4b04-b62f-3a687127a17f')
data.frame(rank = names(res$countByRank),
           count = unname(unlist(res$countByRank)))
```

rank	count
SPECIES	13710
SUBSPECIES	3234
GENUS	726
TRIBE	53
SUBFAMILY	20
FAMILY	2
KINGDOM	1
PHYLUM	1
CLASS	1
ORDER	1

Networks, nodes, and installations

Networks, nodes and installations are at a higher level of organization above datasets, but can be useful if you want to explore data from given organizations. Here, we search for the first 10 GBIF networks, returning just the title field.

```
networks(limit=10)$data$title
#>  [1] "GBIF Backbone Sources"
#>  [2] "Canadensys"
#>  [3] "Southwest Collections of Arthropods Network (SCAN)"
#>  [4] "VertNet"
#>  [5] "Dryad"
#>  [6] "GBIF Network"
#>  [7] "The Knowledge Network for Biocomplexity (KNB) "
#>  [8] "Online Zoological Collections of Australian Museums (OZCAM)"
#>  [9] "Catalogue of Life"
#> [10] "Ocean Biogeographic Information System (OBIS)"
```

Taxonomic names

The GBIF taxonomic names API services are spread across five functions:

- Search GBIF name backbone - `name_backbone()`
- Search across all checklists - `name_lookup()`
- Quick name lookup - `name_suggest()`
- Name usage of a name according to a checklist - `name_usage()`
- GBIF name parser - `parsenames()`

The goal of these name functions is often to settle on a taxonomic name known to GBIF's database. This serves two purposes: 1) when referring to a taxonomic name, you can point to a URI on the internet, and 2) you can search for metadata on that species, and occurrences of that species in GBIF.

Taxonomic names are particularly tricky. Many different organizations have their own unique codes for the same taxonomic names, and some taxonomic groups have preferred sources for the definitive names for that group.

When searching for occurrences (see below) you can search by taxonomic name (and other filters, e.g., taxonomic rank), but you're probably better off figuring out the taxonomic key in the GBIF backbone taxonomy, and using that to search for occurrences. The `taxonkey` parameter in the GBIF occurrences API expects a GBIF backbone taxon key.

GBIF Backbone

The GBIF backbone taxonomy is used in GBIF to have a consistent way to refer to taxonomic names throughout their services. The backbone has 4410899 unique names and 2497114 species names. The backbone taxonomy is also a dataset with key `d7dddbf4-2cf0-4f39-9b2a-bb099caae36c` (http://www.gbif.org/dataset/d7dddbf4-2cf0-4f39-9b2a-bb099caae36c).

We can search the backbone taxonomy with the function `name_backbone()`. Here, we're searching for the name *Poa*, restricting to genera, and the family *Poaceae*.

```
res <- name_backbone(name='Poa', rank='genus', family='Poaceae')
res[c('usageKey', 'kingdom')]
#> $usageKey
#> [1] 2704173
#>
#> $kingdom
#> [1] "Plantae"
```

Name searching

One of the quickest ways to search for names is using `name_suggest()`, which does a very quick search and returns minimal data. Here, we're searching for the query tem *Pum*, and we get back a bunch of names:

```
name_suggest(q='Pum', limit = 6)
```

key	canonicalName	rank
3269133	Pumilus	GENUS
4407849	Pumilinura	GENUS
4323990	Pumiliopsis	GENUS
4324083	Pumiliopes	GENUS
4161281	Pumilea	GENUS
4312370	Pumilopagurus	GENUS

Occurrences

GBIF provides two ways to get occurrence data: through the `/occurrence/search` route (see `occ_search`), or via the `/occurrence/download` route (many functions, see below). `occ_search()` is the main funtion for the search route, and is more appropriate for smaller data, while `occ_download*()` functions are more appropriate for larger data requests.

Large is of course a subjective term. When you hit a "large dataset" will depend primarily on the size of the your data request. GBIF imposes for any given search a limit of 200,000 records in the search service, after which point you can't download any more records for that search. However, you can download more records for different searches.

We think the search service is still quite useful for many people even given the 200,000 limit. For those that need more data, we have created a similar interface in the `download_*()` functions, that should be easy to use. Users should take note that using the download service has a few extra steps to get data into R, but is straight-forward.

Download API

The download API syntax is similar to the occurrence search API in that the same parameters are used, but the way in which the query is defined is different. For example, in the download API you can do greater than searches (i.e., `latitude > 50`), whereas you can not do that in the occurrence search API. Thus, we can't make the query interace exactly the same for both search and download functions.

Using the download service can be as few as three steps: 1) Request data via a search; 2) Download data; 3) Import data into R.

Request data download given a query. Here, we search for the taxon key `3119195`, which is the key for *Helianthus annuus* (http://www.gbif.org/species/3119195).

```
occ_download('taxonKey = 3119195')
#> <<gbif download>>
#>   Username: xxxx
#>   E-mail: xxxx
#>   Download key: 0000840-150615163101818
```

You can check on when the download is ready using the functions `occ_download_list()` and `occ_download_meta()`. When it's ready use `occ_download_get()` to download the dataset to your computer.

```
(res <- occ_download_get("0000840-150615163101818", overwrite = TRUE))
#> <<gbif downloaded get>>
#>   Path: ./0000840-150615163101818.zip
#>   File size: 3.19 MB
```

What's printed out above is a very brief summary of what was downloaded, the path to the file, and its size (in human readable form).

Next, read the data in to R using the function `occ_download_import()`.

```
library("dplyr")
dat <- occ_download_import(res)
dat %>%
  select(gbifID, decimalLatitude, decimalLongitude)
#>       gbifID decimalLatitude decimalLongitude
#> 1  725767384        61.01005         24.41740
#> 2  725767447        59.82923         23.13550
#> 3  725767450        60.38505         25.17449
#> 4  725767513        68.37648         23.51963
#> 5  725767546        67.19203         24.85820
#> 6  725767579        60.21607         24.67412
#> 7  725767609        66.49260         25.70471
#> 8  725767645        61.36634         24.76218
#> 9  725767678        62.29174         27.96500
#> 10 725767681        60.28615         22.38489
#> ..       ...              ...              ...
```

Downloaded data format. The downloaded dataset from GBIF is actually a Darwin Core Archive (DwC-A), an internationally recognized biodiversity informatics standard (http://rs.tdwg.org/dwc/). The DwC-A downloaded is a compressed folder with a number of files, including metadata, citations for

each of the datasets included in the download, and the data itself, in separate files for each dataset as well as one single `.txt` file. In `occ_download_import()`, we simply fetch data from the `.txt` file. If you want to dig into the metadata, citations, etc., it is easily accessible from the folder on your computer.

Search API

The search API follows the GBIF API and is broken down into the following functions:

- Get a single numeric count of occurrenes - `occ_count()`
- Search for occurrences - `occ_search()`
- Get occurrences by occurrence identifier - `occ_get()`
- Get occurrence metadata - `occ_metadata()`

The main search work-horse is `occ_search()`. This function allows very flexible search definitions. In addition, this function does paging internally, making it such that the user does not have worry about the 300 records per request limit - but of course we can't go over the 200,000 maximum limit.

...

Cleaning data. GBIF provides optional data issues with each occurrence record. These issues fall into many different pre-defined classes, covering issues with taxonomic names, geographic data, and more (see `occ_issues_lookup()` to find out more information on GBIF issues; and the same data on GBIF's development site).

`occ_issues()` provides a way to easily filter data downloaded via `occ_search()` based on GBIF issues.

```
out <- occ_search(issue='DEPTH_UNLIKELY', limit = 500)
NROW(out)
#> [1] 4
out %>% occ_issues(-cudc) %>% .$data %>% NROW
#> [1] 2
```

Use cases

Ecological niche modelling

In this example, we plot actual occurrence data for *Bradypus* species against a single predictor variable, BIO1 (annual mean temperature). This is only ont step in a species distribution modelling nworkflow.

This example can be done using BISON data as well with our rbison package.

Load libraries

```
library("rgbif")
library("dismo")
library("maptools")
library("plyr")
```

Raster files

Make a list of files that are installed with the dismo package, then create a rasterStack from these

```
files <- list.files(paste(system.file(package = "dismo"), "/ex", sep = ""),
                    "grd", full.names = TRUE)
predictors <- stack(files)
```

Get world boundaries

```
data(wrld_simpl)
```

Get GBIF data using the rOpenSci package rgbif

```
nn <- name_lookup("bradypus*", rank = "species")
nn <- na.omit(unique(nn$data$nubKey))
df <- occ_search(taxonKey = nn, hasCoordinate = TRUE, limit = 500)
df <- df[ sapply(df, function(x) class(x$data)) %in% "data.frame" ]
df <- ldply(lapply(df, "[[", "data"))
df2 <- df[,c('decimalLongitude','decimalLatitude')]
```

Plot

(1) Add raster data, (2) Add political boundaries, (3) Add the points (occurrences)

```
plot(predictors, 1)
plot(wrld_simpl, add = TRUE)
points(df2, col = "blue")
```

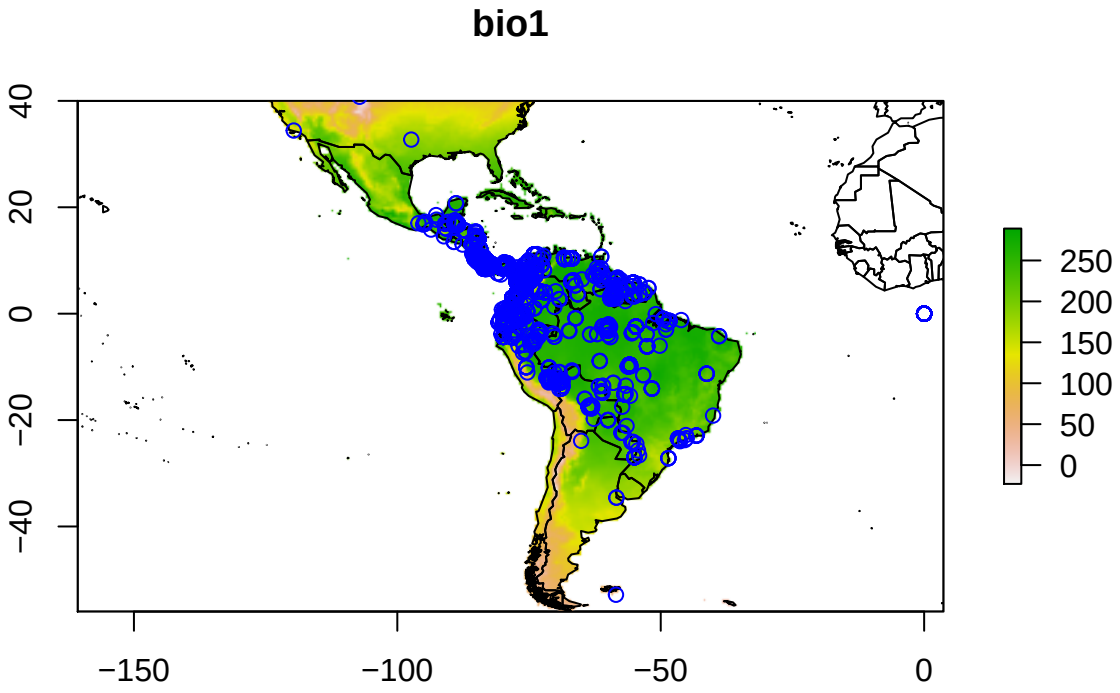

Figure 1:

Biodiversity in big cities

In this example, we collect specimen records across different cities using GBIF data from the `rgbif` package.

Load libraries

```
library("rgbif")
library("ggplot2")
```

```
library("plyr")
library("RCurl")
library("RColorBrewer")
```

Get bounding boxes for some cites

Bounding lat/long data is from here.

```
rawdat <- getURL('https://raw.githubusercontent.com/amyxzhang/boundingbox-cities/master/boundbox
dat <- read.table(text = rawdat, header = FALSE, sep="\t", col.names=c("city","minlat","maxlon",
dat <- data.frame(city=dat$city, minlon=dat$minlon, minlat=dat$minlat, maxlon=dat$maxlon, maxlat
```

```
getdata <- function(x){
  coords <- as.numeric(x[c('minlon','minlat','maxlon','maxlat')])
  num <- occ_search(geometry = coords)$meta$count
  data.frame(city=x['city'], richness=num, stringsAsFactors = FALSE)
}
```

```
out <- apply(dat, 1, getdata)
```

Merge to original table

```
out <- merge(dat, ldply(out), by="city")
```

Add centroids from bounding boxes

```
out <- transform(out, lat = (minlat+maxlat)/2, lon = (minlon+maxlon)/2)
```

Plot data

```
mapp <- map_data('world')
ggplot(mapp, aes(long, lat)) +
  geom_polygon(aes(group=group), fill="white", alpha=0, color="black", size=0.4) +
  geom_point(data=out, aes(lon, lat, color=richness), size=5, alpha=0.8) +
```

```
  scale_color_continuous(low = "#60E1EE", high = "#0404C8") +
  labs(x="", y="") +
  theme_grey(base_size=14) +
  theme(legend.position = "bottom", legend.key = element_blank()) +
  guides(color = guide_legend(keywidth = 2))
```

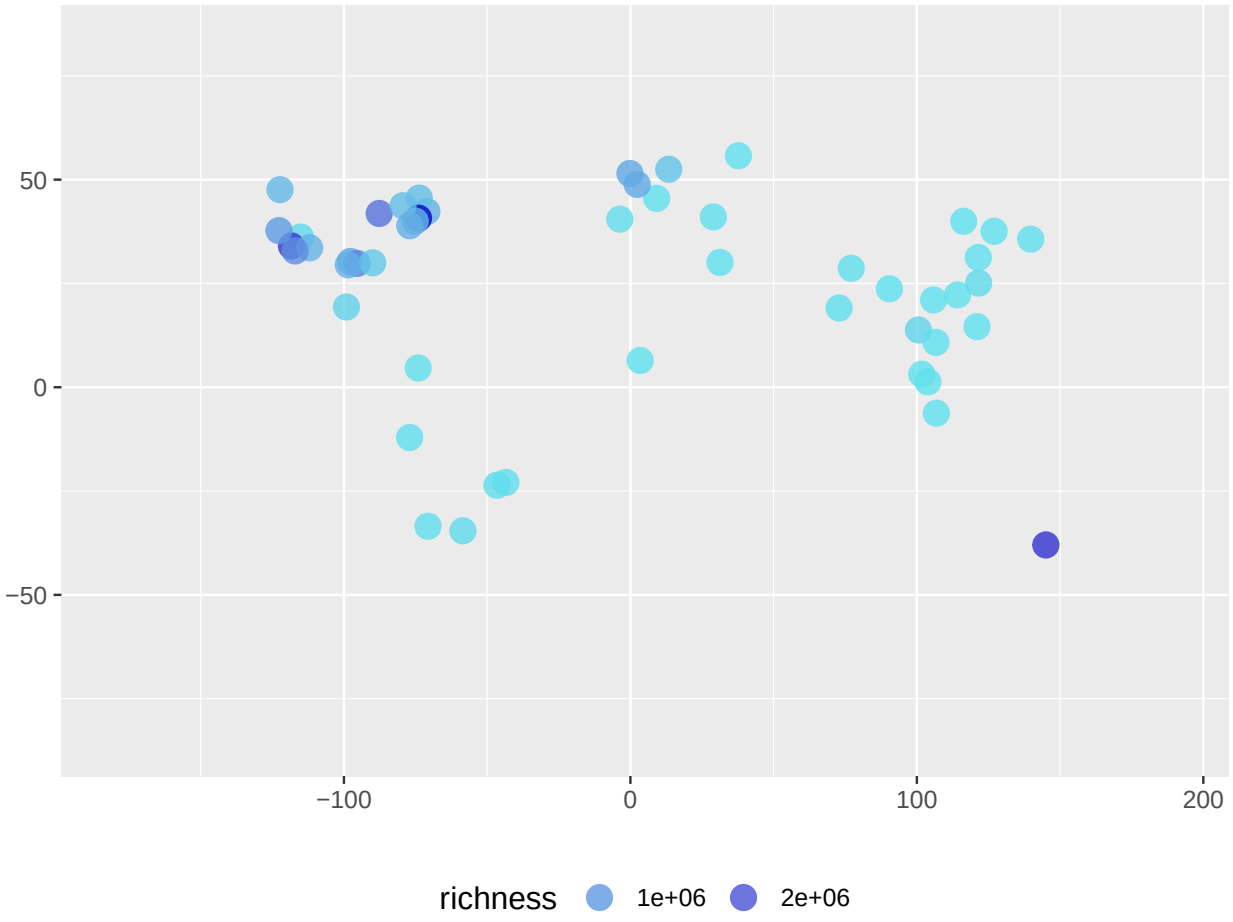

Figure 2:

Valley oak occurrence data comparison

This example comes from Antonio J. Perez-Luque who shared his plot on Twitter. Antonio compared the occurrences of Valley Oak (*Quercus lobata*) from GBIF to the distribution of the same species from the Atlas of US Trees.

Load libraries

```
library('rgbif')
library('raster')
library('sp')
library('maptools')
library('rgeos')
library('scales')
```

Get GBIF Data for Quercus lobata

```
keyQl <- name_backbone(name='Quercus lobata', kingdom='plants')$speciesKey
dat.Ql <- occ_search(taxonKey=keyQl, return='data', limit=50000)
```

Get Distribution map of Q. lobata Atlas of US Trees (Little, E.)

From http://esp.cr.usgs.gov/data/little/. And save shapefile in same directory

```
url <- 'http://esp.cr.usgs.gov/data/little/querloba.zip'
tmp <- tempdir()
download.file(url, destfile = "~/querloba.zip")
unzip("~/querloba.zip", exdir = tmp)
ql <- readShapePoly(file.path(tmp, "querloba.shp"))
```

Get Elevation data of US

```
alt.USA <- getData('alt', country = 'USA')
```

Create Hillshade of US

```
alt.USA <- alt.USA[[1]]
slope.USA <- terrain(alt.USA, opt = 'slope')
aspect.USA <- terrain(alt.USA, opt = 'aspect')
hill.USA <- hillShade(slope.USA, aspect.USA, angle = 45, direction = 315)
```

Plot map

```r
plot(hill.USA, col = grey(0:100/100), legend = FALSE, xlim = c(-125, -116), ylim = c(32, 42), ma
# add shape from Atlas of US Trees
plot(ql, add = TRUE, col = alpha("white", 0.6), border = FALSE)
# add Gbif presence points
points(dat.Ql$decimalLongitude, dat.Ql$decimalLatitude, cex = .7, pch = 19, col = alpha("darkgre
legend(x = -121, y = 40.5, "GBIF Data", pch = 19, col = 'darkgreen', bty = 'n', pt.cex = 1, cex =
legend(x = -121, y = 41.5, "Atlas of United States Trees \n (Little, E. 1971)", pt.cex = 1.5, ce
```

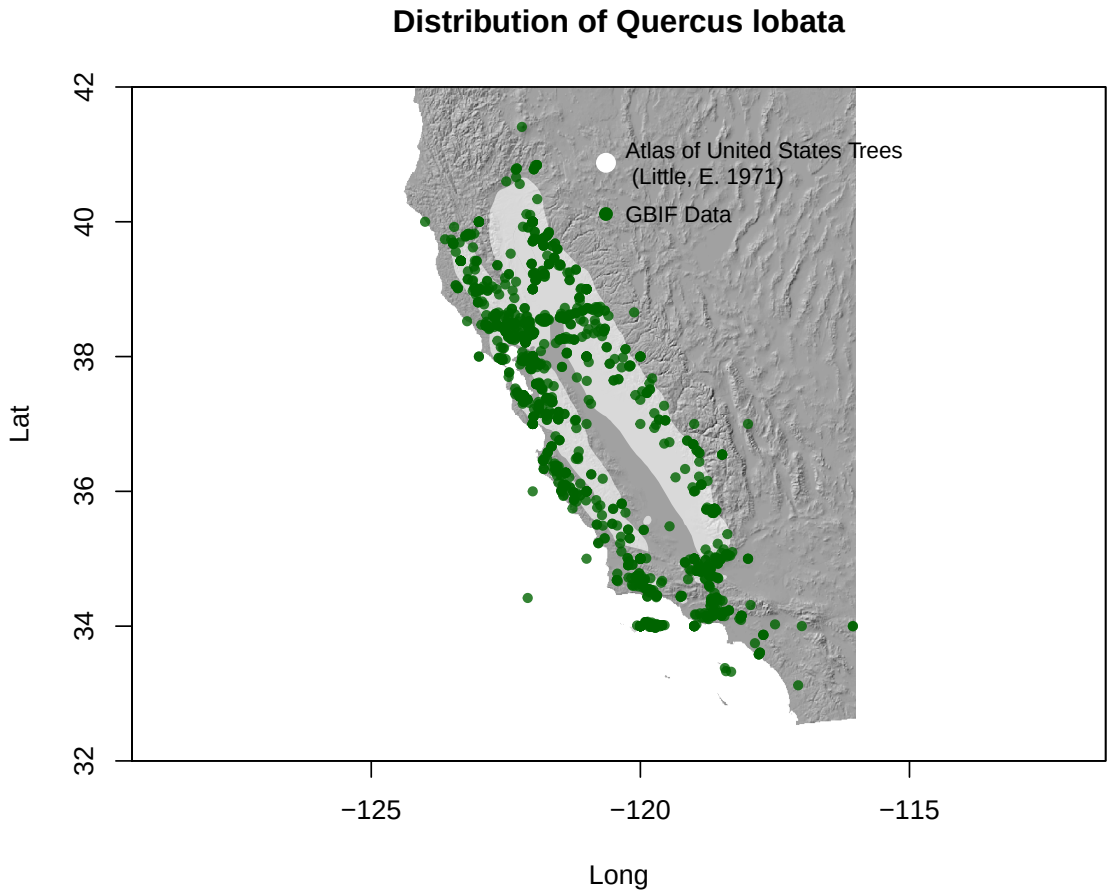

Figure 3:

Conclusions and future directions

- pt 1

- pt 2
- pt 3
- pt 4

Acknowledgements

This project was supported in part by the Alfred P Sloan Foundation (Grant 2013-6-22).

Data Accessibility

All scripts and data used in this paper can be found in the permanent data archive Zenodo under the digital object identifier (DOI). This DOI corresponds to a snapshot of the GitHub repository at github.com/sckott/msrgbif. Software can be found at github.com/ropensci/rgbif, under the open and permissive MIT license.

References

Beck J., Ballesteros-Mejia L., Buchmann CM., Dengler J., Fritz SA., Gruber B., Hof C., Jansen F., Knapp S., Kreft H., Schneider A-K., Winter M., Dormann CF. 2012. Whats on the horizon for macroecology? *Ecography* 35:673–683.

Brown JH. 1995. *Macroecology*. University of Chicago Press.

Brown KA., Parks KE., Bethell CA., Johnson SE., Mulligan M. 2015. Predicting plant diversity patterns in madagascar: Understanding the effects of climate and land cover change in a biodiversity hotspot. *PLOS ONE* 10:e0122721.

Ceballos G., Ehrlich PR., Barnosky AD., Garcia A., Pringle RM., Palmer TM. 2015. Accelerated modern human-induced species losses: Entering the sixth mass extinction. *Science Advances* 1:e1400253–e1400253.

Chamberlain S., Ram K., Barve V., Mcglinn D. *Rgbif: Interface to the global 'biodiversity' information facility 'aPI'*.

Faulkner KT., Robertson MP., Rouget M., Wilson JR. 2014. A simple, rapid methodology for developing invasive species watch lists. *Biological Conservation* 179:25–32.

Febbraro MD., Lurz PWW., Genovesi P., Maiorano L., Girardello M., Bertolino S. 2013. The use of climatic niches in screening procedures for introduced species to evaluate risk of spread: A case with the american eastern grey squirrel. *PLoS ONE* 8:e66559.

Ferretti F., Verd GM., Seret B., Šprem JS., Micheli F. 2015. Falling through the cracks: The fading history of a large iconic predator. *Fish and Fisheries*:n/a–n/a.

Ficetola GF., Rondinini C., Bonardi A., Baisero D., Padoa-Schioppa E. 2014. Habitat availability for amphibians and extinction threat: A global analysis. *Diversity and Distributions* 21:302–311.

María Mendoza., Ospina OE., Cárdenas-Henao H., García-R JC. 2015. A likelihood inference of historical biogeography in the world's most diverse terrestrial vertebrate genus: Diversification of direct-developing frogs (craugastoridae: Pristimantis) across the neotropics. *Molecular Phylogenetics and Evolution* 85:50–58.

Pimm SL., Jenkins CN., Abell R., Brooks TM., Gittleman JL., Joppa LN., Raven PH., Roberts CM., Sexton JO. 2014. The biodiversity of species and their rates of extinction, distribution, and protection. *Science* 344:1246752–1246752.

R Core Team. 2014. *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.