# Bird museum data (BMD)

*Alexander Florez Rodriguez*
*Center for Macroecology, Evolution and Climate*

*4/3/2017*

## Map of genetic diversity of Aves

We need georeferenced DNA sequences to map the genetic diversity of any taxa. The lack of a central repository with this type of data propels us to implement multiple approaches to retrieve georeferenced DNA sequences. First, when available, we obtain georeferences from genetic repositories (e.g. GenBank, https://www.ncbi.nlm.nih.gov/genbank/). Second, if coordinates are not available, we use GeoNames (http://www.geonames.org/) to estimate coordinates from locality descriptions in genetic repositories. Third, if neither coordinates nor localities are available, we use non-geographical information associated to the sequences (e.g. voucher number) to cross-reference with other databases (e.g. museums and GBIF). Here I describe the avian mitochondrial DNA dataset, and how to obtain coordinates using voucher information.

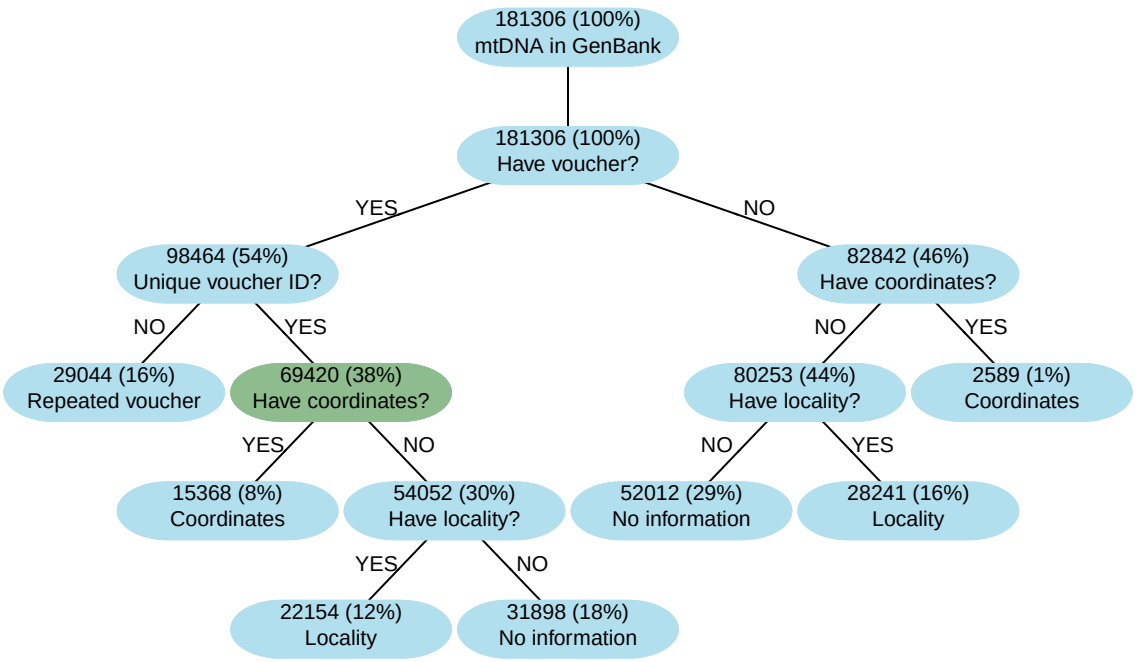
We obtained all avian mitochondrial sequences in GenBank (n = 181306). Then, we used binary trees (see below) to subset (nested nodes) the dataset by the type of information available. In each node the number of sequences and the percentage of the total database of each subset are shown. For example, the GenBank dataset can be divided in sequences with and without voucher information, for Aves, 54% (98646 sequences) of the sequences have voucher information. The GenBank binary tree shows that 38% (69420) of the avian dataset have unique vouchers (green node). This subset can be used to retrieve geographical information from open access biodiversity data (e.g. GBIF).

### GenBank binary tree

```
db <- "/Users/afr/Desktop/A/Postdoc/Birds_museum_data/BMD_exploratory/Data/coordinates.temp"
# Read the database
BMD_raw <- read.delim(db, header = F, stringsAsFactors = F)
# Add names to the columns
colnames(BMD_raw) <- c("ID", "Species", "Coordinates", "Location", "Voucher","Isolate", "Haplotype")
# node_names is a table with one colum indicating the name of the nodes

node_names <- read.delim("node_names.txt", header=T, sep = "\t")
# fromto_edge indicate the connections among nodes
fromto_edge <- read.delim("fchart_fromto.txt", header=F)
#### libraries and internal functions ####

library(diagram)
library(schoolmath)
library(shape)
make.flowtable_GenB <- function(DB_raw, node_names_tbl){
        tmp_flow <- node_names_tbl
        tmp_flow$Pos <- c(1,2,5,11,15,17,21,23,27,29,31,33,39,41)
        # Total number of mtDNA sequences in GenBank
        tmp_flow$Value[1] <- dim(DB_raw)[1]
        # Total number of mtDNA sequences in GenBank
        tmp_flow$Value[2] <- dim(DB_raw)[1]
        # Sequences with voucher information
```

```
        BMD_all_voucher <- DB_raw[DB_raw$Voucher != "voucher_is_not_available",]
        tmp_flow$Value[3] <- dim(BMD_all_voucher)[1]
        # Sequences WITHOUT voucher information
        BMD_Wo_voucher <- DB_raw[DB_raw$Voucher == "voucher_is_not_available",]
        tmp_flow$Value[4] <- dim(BMD_Wo_voucher)[1]
        # Sequences with voucher information, how many have unique voucher IDs
        BMD_unq_voucher <- BMD_all_voucher[-which(duplicated(BMD_all_voucher$Voucher)),]
        tmp_flow$Value[6] <- dim(BMD_unq_voucher)[1]
        # Sequences with voucher information, how many have REPEATED voucher IDs
        BMD_rep_voucher <- BMD_all_voucher[which(duplicated(BMD_all_voucher$Voucher)),]
        tmp_flow$Value[5] <- dim(BMD_rep_voucher)[1]
        # Sequences WITHOUT voucher information, neither coordinates
        BMD_Wo_Vo_NO_coor <- BMD_Wo_voucher[BMD_Wo_voucher$Coordinates == "coordinates_are_not_available
        tmp_flow$Value[7] <- dim(BMD_Wo_Vo_NO_coor)[1]
        # Sequences WITHOUT voucher information, but with coordinates
        BMD_Wo_Vo_coor <- BMD_Wo_voucher[BMD_Wo_voucher$Coordinates != "coordinates_are_not_available",]
        tmp_flow$Value[8] <- dim(BMD_Wo_Vo_coor)[1]
        # Sequences with voucher, and coordinates
        BMD_unq_vo_coo <- BMD_unq_voucher[BMD_unq_voucher$Coordinates != "coordinates_are_not_available
        tmp_flow$Value[9] <-  dim(BMD_unq_vo_coo)[1]
        # Sequences with voucher, WITHOUT coordinates, but with locality
        BMD_unq_vo_NO_coo <- BMD_unq_voucher[BMD_unq_voucher$Coordinates == "coordinates_are_not_availa
        tmp_flow$Value[10] <-  dim(BMD_unq_vo_NO_coo)[1]
        # Sequences WITHOUT voucher, no coordinates, NOR with localities
        BMD_Wo_Vo_NOcoo_NO_loc <- BMD_Wo_Vo_NO_coor[BMD_Wo_Vo_NO_coor$Location == "location is not avai
        tmp_flow$Value[11] <-  dim(BMD_Wo_Vo_NOcoo_NO_loc)[1]
        # Sequences WITHOUT voucher, no coordinates, BUT with localities
        BMD_Wo_Vo_NOcoo_loc <- BMD_Wo_Vo_NO_coor[BMD_Wo_Vo_NO_coor$Location != "location is not availab
        tmp_flow$Value[12] <-  dim(BMD_Wo_Vo_NOcoo_loc)[1]
        # Sequences with unique voucher, no coordinates, BUT with localities
        BMD_unq_vo_NO_coo_loc <- BMD_unq_vo_NO_coo[BMD_unq_vo_NO_coo$Location != "location is not availa
        tmp_flow$Value[13] <-  dim(BMD_unq_vo_NO_coo_loc)[1]
        # Sequences with unique voucher, BUT no coordinates, nor localities
        BMD_unq_vo_NO_coo_NO_loc <- BMD_unq_vo_NO_coo[BMD_unq_vo_NO_coo$Location == "location is not ava
        tmp_flow$Value[14] <-  dim(BMD_unq_vo_NO_coo_NO_loc)[1]
        tmp_flow$Percent <- round(tmp_flow$Value * 100 / tmp_flow$Value[1], digits = 0)
        tmp_flow
}
BMD.fchart.nodes <- function(flow_table){
        elpos <- coordinates(c(1, 1, rep(11, 4)))
        for(A in seq_along(flow_table[,1])){
                tmp_pos <- flow_table$Pos[A]
                textround (elpos[tmp_pos,], 0.048, 0.04,lab = "",
                           box.col = ifelse(tmp_pos == 17,"#8FBC8F", "#B2DFEE"), shadow.col = NULL,
                           lcol = ifelse(tmp_pos == 17,"#8FBC8F", "#B2DFEE"))
        }
}
BMD.fchart.backbone <- function(fromto_edge, boolean_lab){
        par(mar = c(4, 0, 0, 0), oma=c(0,0,0,0))
        openplotmat()
        elpos <- coordinates(c(1,1, rep(11, 4)))
        fromto <- fromto_edge[,c(1,2)]
        nr <- nrow(fromto)
```

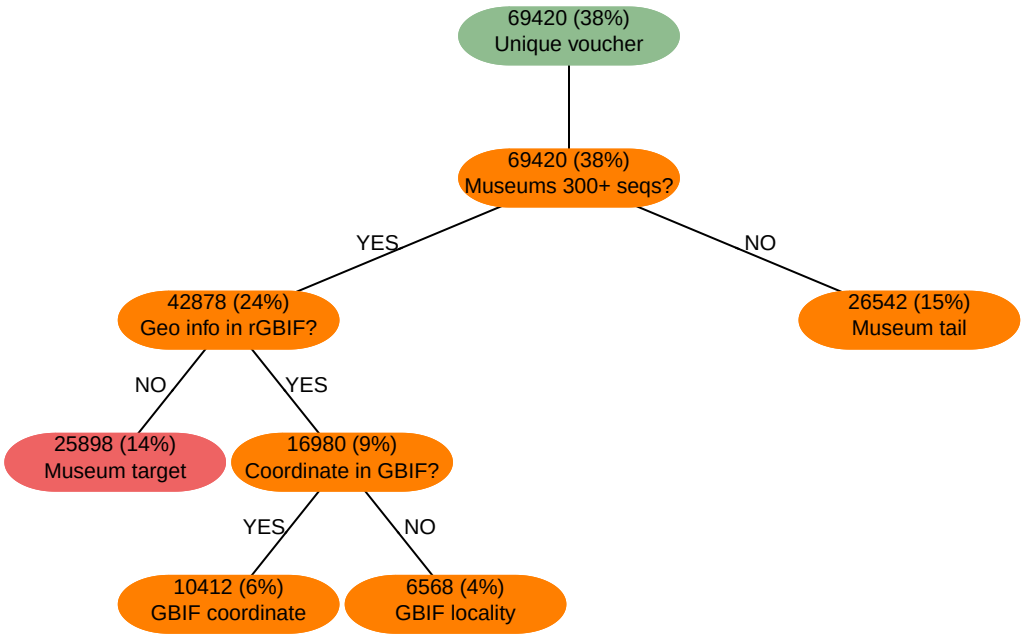
```
        arrpos <- matrix(ncol = 2, nrow = nr)
        for (i in 1:nr) {
                arrpos[i, ] <- straightarrow (to = elpos[fromto[i, 2], ], from = elpos[fromto[i, 1], ],
                                              lwd = 1, arr.length = 0)
        }
        for(pos in seq_along(arrpos[,1])){
                if (is.even(pos)){
                        text(arrpos[pos,1] - 0.017, arrpos[pos,2] + 0.01, boolean_lab[pos], cex=.7)
                }else{
                        text(arrpos[pos,1] + 0.017, arrpos[pos,2] + 0.01, boolean_lab[pos], cex=.7)
                }
        }
}
BMD.fchart.text <- function(flow_table){
        elpos <- coordinates(c(1, 1, rep(11, 4)))
        for(A in seq_along(flow_table[,1])){
                tmp_pos <- flow_table$Pos[A]
                tmp_label <- flow_table$Label[A]
                tmp_value <- flow_table$Value[A]
                tmp_percent <- flow_table$Percent[A]
                tmp_paste <- paste(tmp_value, " (", tmp_percent, "%)", sep = "")
                textplain(elpos[tmp_pos,], adj=c(0.5,1),lab = tmp_label, cex=0.7)
                textplain(elpos[tmp_pos,], adj=c(0.5,-0.7),lab = tmp_paste, cex=0.7)
        }
}

#### the script ####
#pdf(file = "flowchart_genbank.pdf")
flow_nodes <- make.flowtable_GenB(BMD_raw,node_names)
BMD.fchart.backbone(fromto_edge, fromto_edge[,3])
BMD.fchart.nodes(flow_nodes)
BMD.fchart.text(flow_nodes)
```



```
#dev.off()
```

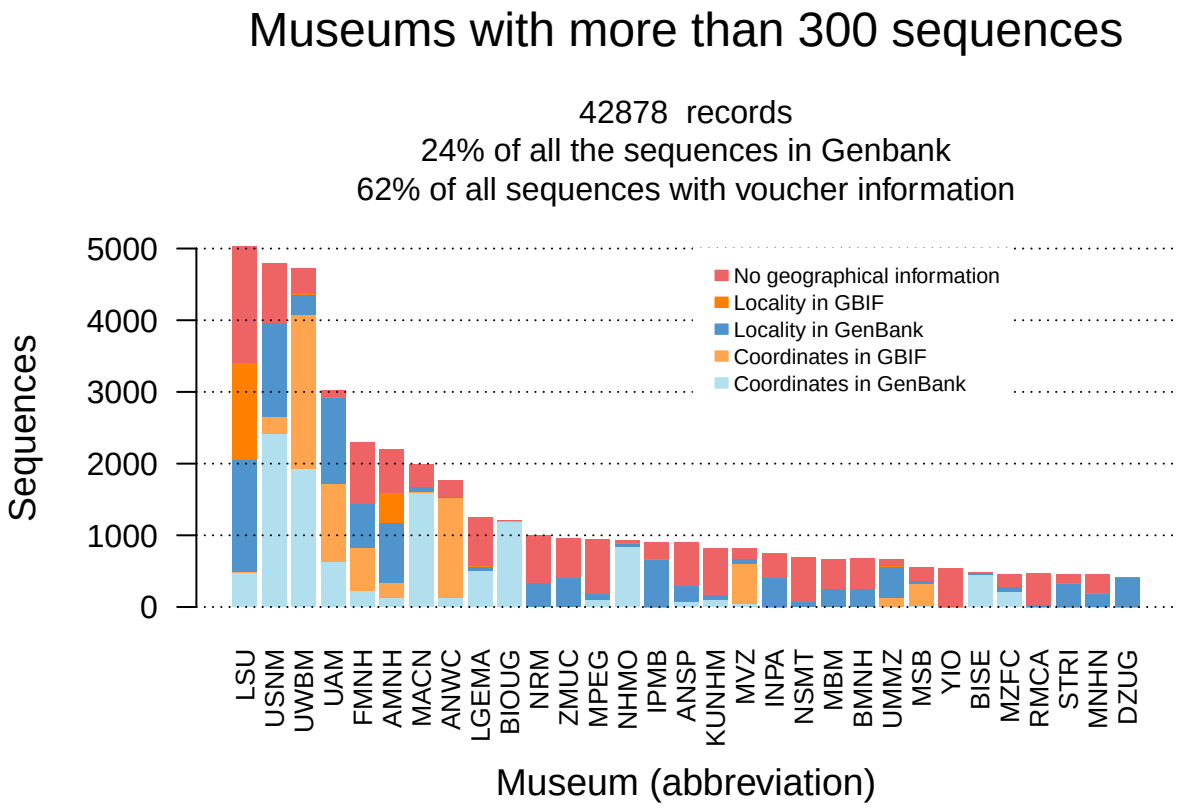
Similarly to the GenBank binary tree, the GBIF binary tree (see below) shows the hierarchy of each one of the nodes (subsets) and the percentage relative to the total dataset. The GBIF binary tree describes the amount of sequences for which is possible to obtain geographical information from GBIF, this using the voucher information as a link between GenBank and GBIF. To obtain the geographical information we used the package rgbif in R. After using the three approaches described in the first paragraph, we show that 25898 (14%, red node) sequences have voucher information, but lack geographical information (e.i. coordinates or localities). For this subset (red node) the geographical information could be retrieved using museums' databases.

## GBIF binary tree




## Museums' data

We used the voucher information available in GenBank to associate each one of the sequences with their respective museum (see barplot below). Subsequently, for each museum we quantified: 1) sequences with coordinates in GenBank (lighter blue), 2) sequences with coordinates in GBIF (lighter orange), 3) sequences with locality description in GenBank (darker blue), 4) sequences with locality description in GBIF (darker orange), and 5) sequences with voucher, but lacking geographical information (red). Each bar represents a museum (abbreviation).



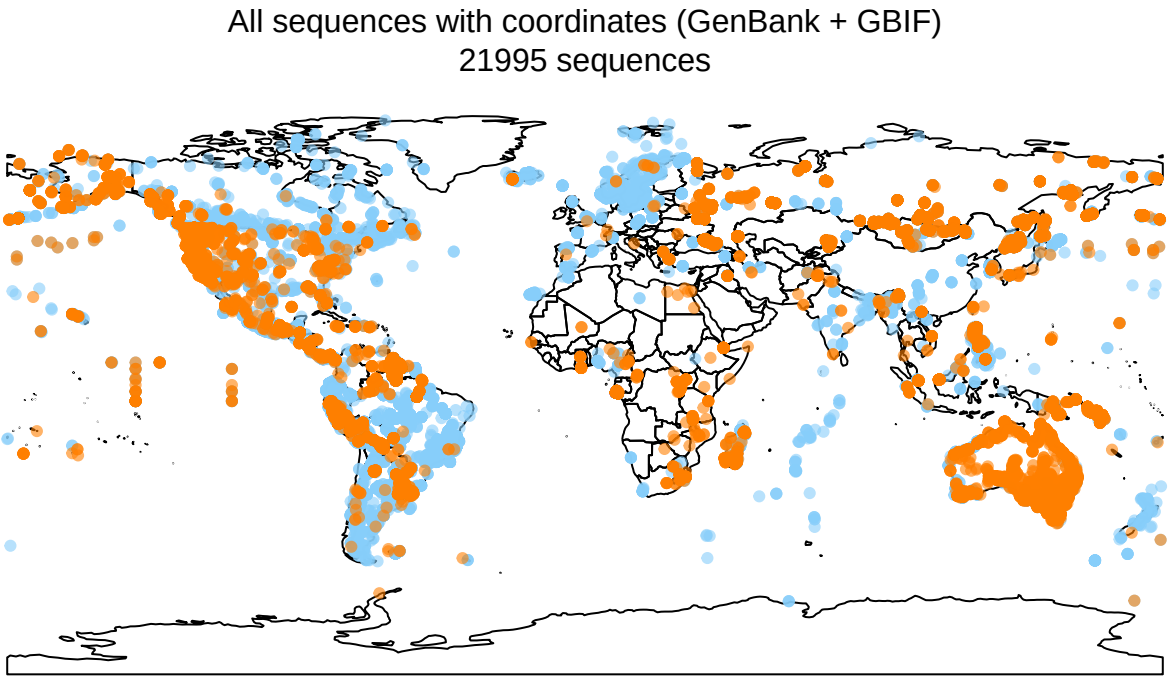

### Museum full names

| Abbreviation | Full.name |
|---|---|
| LSU | Louisiana State University Museum of Natural Science |
| USNM | Smithsonian Institution National Museum of Natural History |
| UWBM | Burke Museum |
| UAM | University of Alaska Museum |
| FMNH | The Field Museum |
| AMNH | American Museum of Natural History |
| MACN | Museo Argentino de Ciencias Naturales |
| ANWC | Australian National Wildlife Collection |
| LGEMA | Laboratorio de Genetica e Evolucao Molecular de Aves |
| BIOUG | Centre for Biodiversity Genomics |
| NRM | Naturhistoriska riksmuseet |
| ZMUC | Zoological Museum University of Copenhagen |
| MPEG | Museu Paraense Emilio Goeldi |
| NHMO | Natural History Museum Oslo University |
| IPMB | Institute of Plant and Microbial Biology |
| ANSP | Academy of Natural Sciences of Drexel University |
| KUNHM | Kansas University Biodiversity Institute & Natural History Museum |
| MVZ | The Museum of Vertebrate Zoology at Berkeley |
| INPA | National Institute of Amazon Researches |
| NSMT | National Museum of Nature and Science,Tokyo |
| MBM | Museu Botanico Municipal (Curitiba) |
| BMNH | Natural History Museum London |

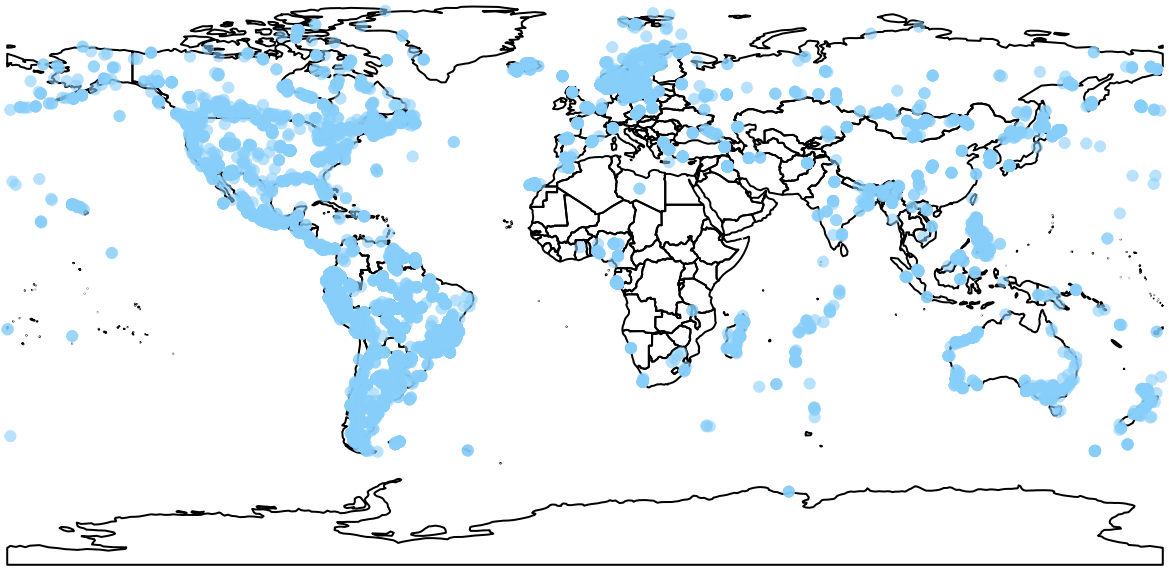
| Abbreviation | Full.name |
|---|---|
| UMMZ | University of Michigan Museum of Zoology |
| MSB | Museum of Southwestern Biology The University of New Mexico |
| YIO | Yamashina Institute for Ornithology |
| BISE | Biodiversity Information System for Europe |
| MZFC | UNAM - Coleccion Ornitologica del Museo de Zoologia |
| RMCA | Royal Museum for Central Africa Tervuren Belgium |
| STRI | Smithsonian Tropical Research Institute |
| MNHN | Museum national dHistoire naturelle |
| DZUG | Zoology University of Gothenburg |

## Maps for the sequences

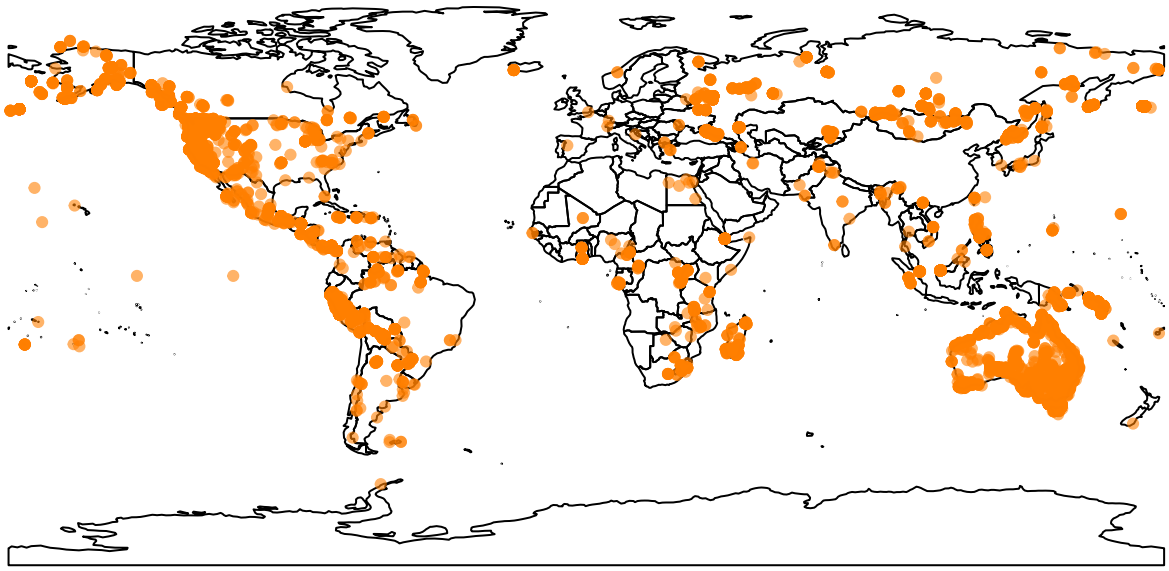
We mapped all the sequences for which coordinates were obtained from GenBank and GBIF. This data do not include the sequences with locality descriptions. Blue and orange dots represent GenBank and GBIF coordinates respectively.



Sequences with coordinates ONLY in GenBank
11597 sequences

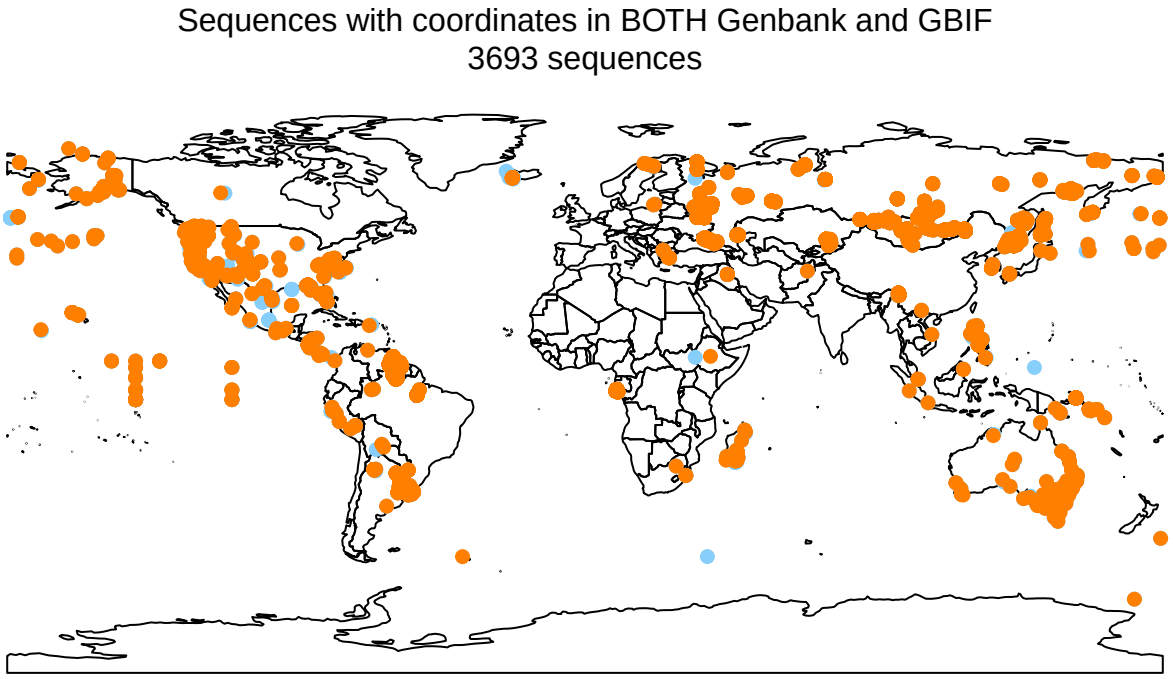


Sequences with coordinates ONLY in GBIF
6712 sequences



We were able to simultaneously retrieve coordinates from both GenBank and GBIF for 3693 sequences. Assuming that coordinates in GenBank and GBIF are correct, we would expect that the map with points from both GenBank and GBIF will overlap completely, and that the absolut paired distance between GenBank and GBIF coordinates is equal to zero. We found that most of the points overlap and that in general the paired distance between GenBank and GBIF coordinates is smaller than 0.015 decimal degrees. However,

for 405 sequences the absolut distance between Genbank and GBIF coordinates range from 0.017 to 360 decimal degrees, indicating that aproximately 11% of the sequences have incongruent coordinates in different repositories.




## Mean distance between GenBank and GBIF coordinates