

Bird museum data (BMD)

Alexander Florez Rodriguez

Center for Macroecology, Evolution and Climate

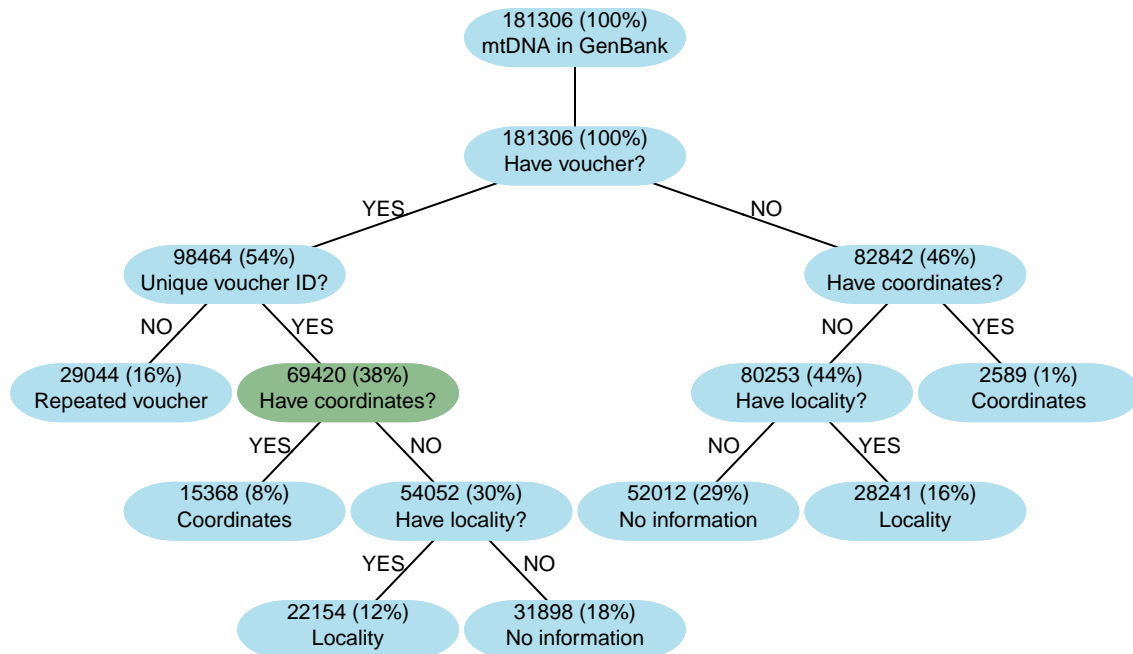
4/3/2017

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.

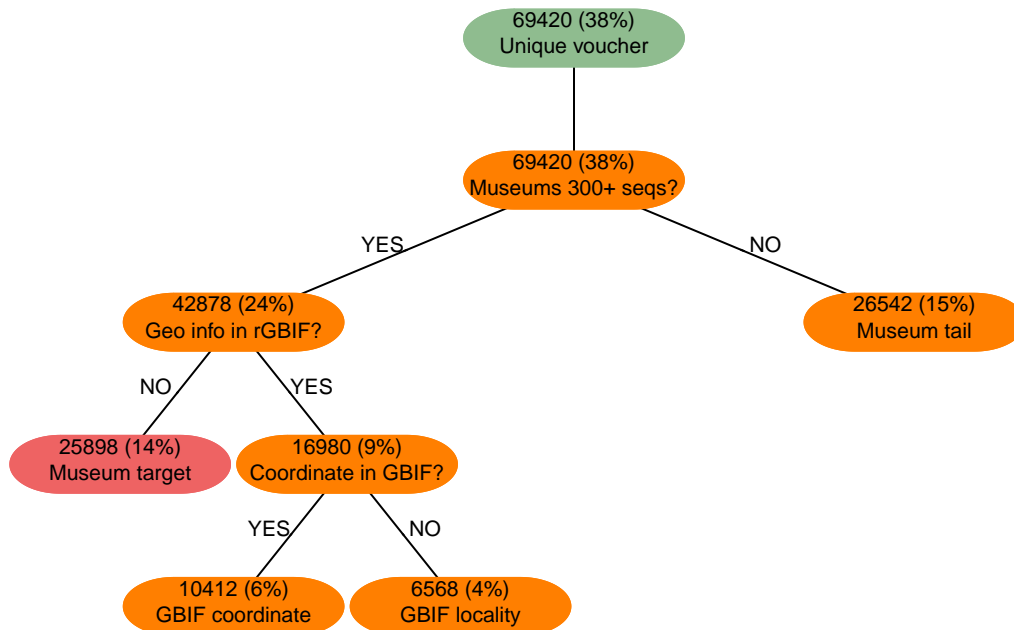
We obtained all avian mitochondrial sequences in GenBank ($n = 181306$). Then, we used binary trees (see below) to subset (nested nodes) the dataset by the type of information available. In each node the number of sequences and the percentage of the total database of each subset are shown. For example, 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



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.

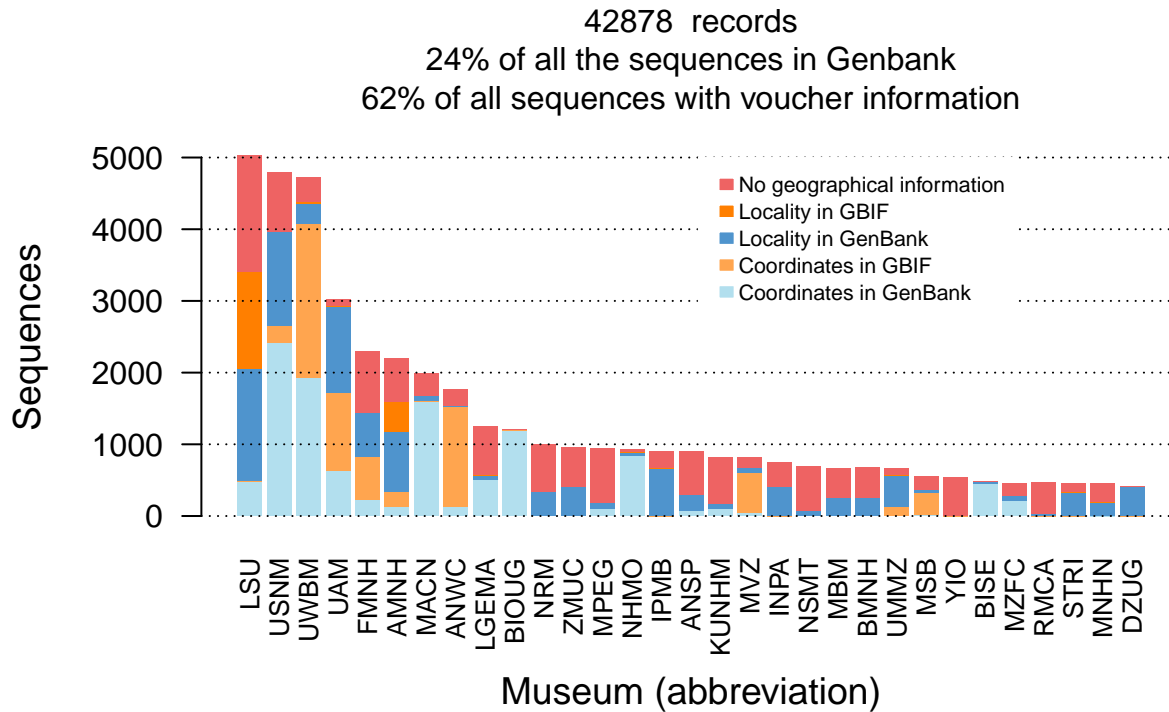
GBIF binary tree



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).

Museums with more than 300 sequences



Museum full names

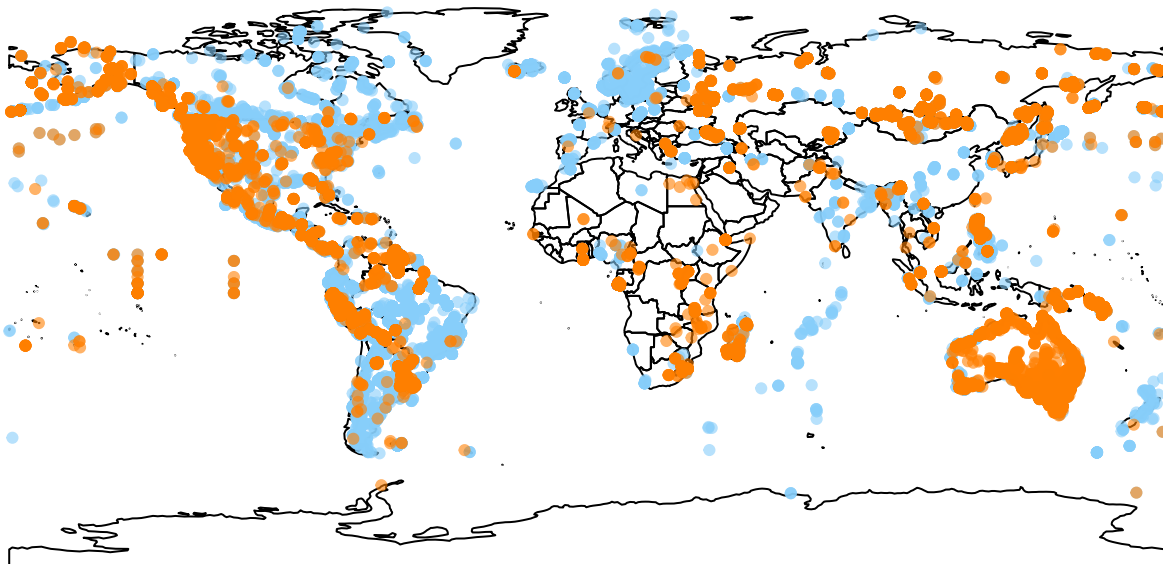
Abbreviation	Full.name
LSU	Louisiana State University Museum of Natural Science
USNM	Smithsonian Institution National Museum of Natural History
UWBM	Burke Museum
UAM	University of Alaska Museum
FMNH	The Field Museum
AMNH	American Museum of Natural History
MACN	Museo Argentino de Ciencias Naturales
ANWC	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 d'Histoire naturelle
DZUG	Zoology University of Gothenburg

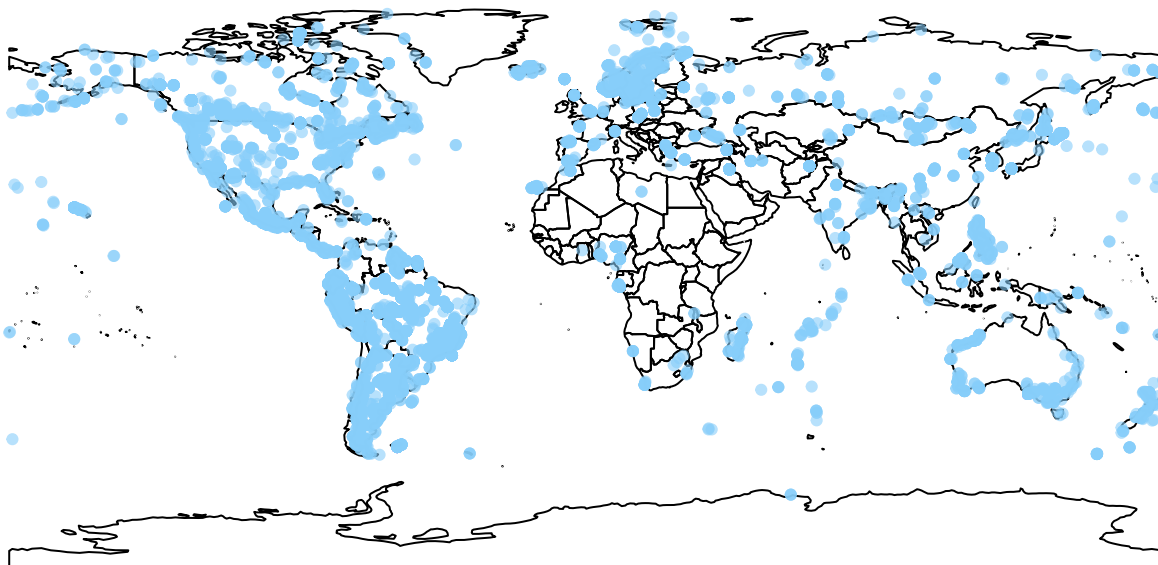
Maps for the sequences

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.

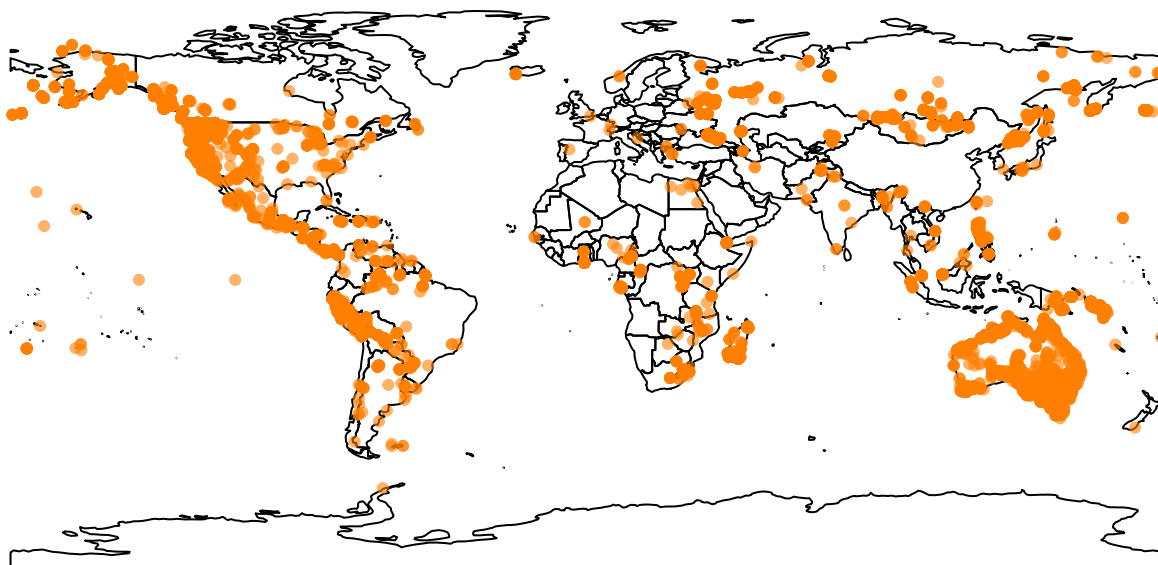
All sequences with coordinates (GenBank + GBIF)
21995 sequences



Sequences with coordinates ONLY in GenBank
11597 sequences



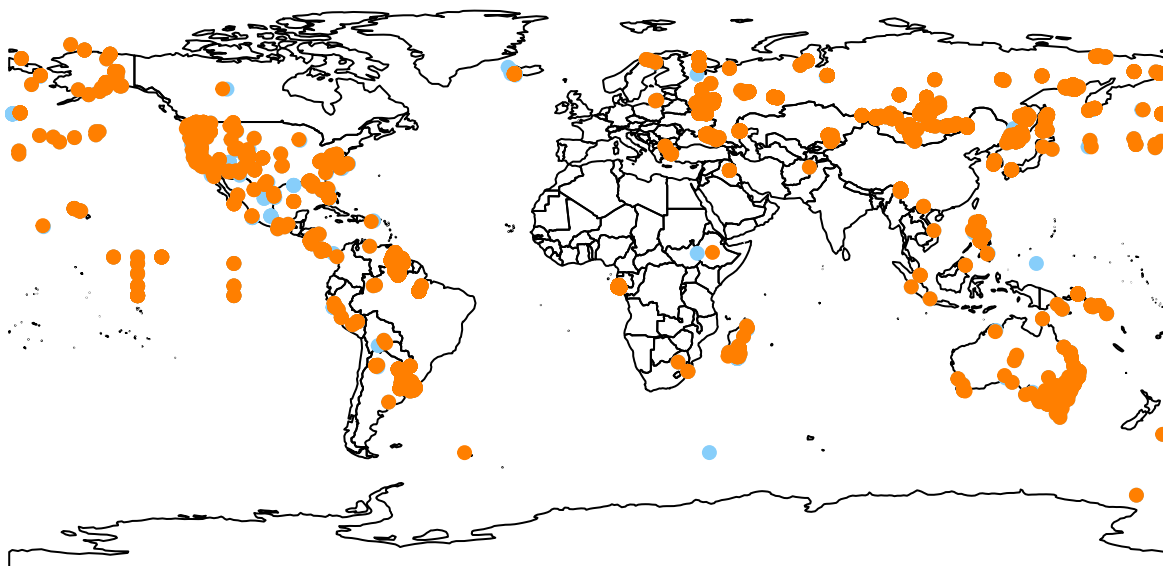
Sequences with coordinates ONLY in GBIF
6712 sequences



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

Sequences with coordinates in BOTH Genbank and GBIF
3693 sequences



Mean distance between GenBank and GBIF coordinates

