Paolo Gratton, Silvio Marta, Gaëlle Bocksberger, Marten Winter, Emiliano Trucchi & Hjalmar Kühl

**A world of sequences. Can we use georeferenced nucleotide databases for an Automated Phylogeography?**

Journal of Biogeography

**Appendix S1.** Supplementary Methods

*Obtaining sequence metadata from GenBank*

We downloaded authoritative lists of Tetrapod species from class-specific taxonomy databases available online (Amphibians: IUCN red lists [www.iucnredlist.org/technical-documents/spatial-data#amphibians](http://www.iucnredlist.org/technical-documents/spatial-data#amphibians) - downloaded on Dec 8 2014, 6312 species; Birds: IOC World Bird List <http://www.worldbirdnames.org/ioc-lists/> - downloaded on Nov 24 2014, 10695 species; Mammals Wilson & Reeder’s Mammal Species of the World: [www.departments.bucknell.edu/biology/resources/msw3/](http://www.departments.bucknell.edu/biology/resources/msw3/) - downloaded on Nov 21 2014, 5416 species; Reptiles: The Reptiles Database [www.reptile-database.org/data/](http://www.reptile-database.org/data/) - released on Dec 1 2014, Downloaded on Dec 8 2014, 10119 species). We then translated Linnean binomials into NCBI species taxonomy identifiers (taxonIDs) by submitting the lists of species to the NCBI Taxonomy name/id Status Report Page ([www.ncbi.nlm.nih.gov/Taxonomy/TaxIdentifier/tax\_identifier.cgi](http://www.ncbi.nlm.nih.gov/Taxonomy/TaxIdentifier/tax_identifier.cgi)) and obtained genus taxonIDs using the parent() R function (package CHNOSZ, Dick, 2008). The translation resulted into 21262 unique species-level taxonIDs (amphibians: 3839; birds 7755; mammals 3964; reptiles 5704). We used species taxonIDs to search the NCBI GenBank (http://www.ncbi.nlm.nih.gov/genbank/) and collaborative databases adhering to the INSDC (International Nucleotide Sequence Database Collaboration: NCBI, DDBJ, EMBL) using a custom python script (including functions from the Bio.Entrez module, Biopython package, Cock et al., 2009). In our search, we applied the ‘DDBJ/EMBL/GenBank (INSDC) with limits’ filter, which excludes Sequence-Tagged Sites (STS), Third-Party Annotation sequences (TPA), High-Throughput Genomic sequences (HTG), Whole Genome Shotgun (WGS) and patents, and we further excluded transcriptome, mRNA, Transcriptome Shotgun Assembly (TSA) and genome scaffolds data not yet filtered out (search string: ‘txid<taxonID>[orgn] AND ddbj embl genbank with limits[filt] NOT transcriptome[All Fields] NOT mRNA[filt] NOT TSA[All Fields] NOT scaffold[All Fields]’).

For each accession that matched our search criteria, we stored the metadata contained in the ‘accession’, ‘definition’, ‘organism', 'speciesID', 'genusID', 'organelle', 'gene', 'CDS\_product', 'tRNA', 'rRNA', 'haplotype', 'isolate', 'pop\_variant', 'country', 'lat\_lon', 'ref\_pubmed', 'ref\_journal', 'ref\_authors' and 'ref\_titles' fields. In addition, we retrieved the annotation for the mtDNA d-loop (control region) from the ‘D-loop’ field, the ‘miscellaneous’ field or the ‘definition’ field. Sequences longer than 20,000 bps were ignored. Humans (*Homo sapiens*) and a few highly synanthropic or domesticated species (*Mus musculus*, *Rattus rattus*, *Rattus norvegicus*, *Chrysochloris asiatica*, *Oryctolagus cuniculus*, *Felis catus, Felis silvestris*, *Canis lupus*, *Mustela putorius*, *Sus scrofa*, *Bos taurus*, *Ovis aries*, *Capra hircus*, *Equus caballus* and *Gallus gallus*) were excluded from our search. The complete list of searched taxa is available in Supplementary Table S1. A limit of 10000 retained sequences per species was applied, only affecting the reptilian species *Agkistrodon contortrix*, for which 14612 microsatellite sequences from a 454 shotgun (Castoe et al., 2010) were retrieved in our search.

*Geocoding of GenBank data*

We assigned geographic coordinates (geocoded) the subset of ‘informative’ accessions using a custom pipeline written in R and relying on geographic information from two publicly available databases: Global Administrative Areas (GADM, [www.gadm.org](http://www.gadm.org), version 2.0 January 2012, containing georeferenced polygons with the boundaries of 556049 administrative units worldwide) and GeoNames ([www.geonames.org](http://www.geonames.org), downloaded on Feb 4 2015, containing point geographic coordinates for 9863141 placenames worldwide).

Our pipeline started by isolating the country name from the rest of the text string (substring before the first ‘:’) and then removing most non-toponomastic information by eliminating all-lowercase words, many generic terms that do not usually form part of the toponym (e.g. ‘Province’, ‘State’, etc.) and several recurrent expressions (e.g. ‘25 Mi. N of’). ‘Country’ fields that did not contain a valid country name (i.e. a name that could be unambiguously referred to a country listed in the GADM or GeoNames databases) were excluded. The resulting strings were then splitted by the most common separators found in GenBank ‘country’ fields (more precisely: ‘)’ were removed before splitting on ‘(‘, ‘,’, ’:’, ’;’, ‘/’ and ‘\_’). Each substring was searched for countrywise in the GADM and GeoNames databases. We recorded whether the substring appeared as such in the ‘adm1’ to ‘adm5’ (names of the administrative units from level 1 to level 5) and ‘varnames adm1’ to ‘varnames adm4’ (variant names from level 1 to level 4) fields of the GADM database. When a substring matched more than one administrative levels, we chose to conservatively consider the higher level match (unless the placename occurred more than once: e.g. ‘New York, New York’ would match the city within the state). Any substring not matched in the GADM database was further searched for in the GeoNames database (‘name’, ‘asciiname’ and ‘alternate names’ fields). If at least one match was found in GADM for the same record, the search was limited to those GeoNames placenames whose geographic coordinates fell within the extent of the polygon corresponding to the smallest level matched administrative unit. Searches in GeoNames started by looking for exact matches between the searched substring and the entire content of the field (or subfield in the case of ‘alternate names’). If no exact match was found, we searched for partial matches (whether the searched substring was contained in the field). When multiple matches occurred, we calculated distance matrices among all matched locations. If the maximum distance among matched locations was smaller than 100 km, we assigned the substring to the centroid of matched locations, if it was larger, it was not recorded. Further details about the search algorithm are given in Supplementary Figure S1.

**Supplementary Figure S1.** Flow-chart of the geocoding algorithm.

