Journal of Biogeography (J. Biogeogr.) (2016)

METHODOLOGICAL APPLICATION

A world of sequences: can we use georeferenced nucleotide databases for a robust automated phylogeography?

Paolo Gratton[1]*, Silvio Marta[2], Gaëlle Bocksberger[1], Marten Winter[3], Emiliano Trucchi[4] and Hjalmar Kühl[1]

[1]*Department of Primatology, Max Planck Institute for Evolutionary Anthropology, Deutscher Platz 6, 04103, Leipzig, Germany,* [2]*Dipartimento di Biologia, Università di Roma "Tor Vergata", Via della Ricerca Scientifica 1, 00133 Rome, Italy,* [3]*German Center for Integrative Biodiversity Research (iDiv), Deutscher Platz 5, 04103 Leipzig, Germany,* [4]*Department for Botany and Biodiversity Research, University of Vienna, Rennweg 14, A-1030 Vienna, Austria*

*Correspondence: Paolo Gratton, Department of Primatology, Max Planck Institute for Evolutionary Anthropology, Deutscher Platz 6, 04103 Leipzig, Germany.
E-mail: paolo_gratton@eva.mpg.de

ABSTRACT

Aim Comparative phylogeography across a large number of species allows investigating community-level processes at regional and continental scales. An effective approach to such studies would involve automatic retrieval of georeferenced sequence data from nucleotide databases (a first step towards an 'automated phylogeography'). It remains unclear if, despite repeated calls, georeferencing of nucleotide databases has increased in frequency, and if accumulated data allow for broad applications based on automated retrieval of sequence data and associated geographical information. Here, we investigated geographical information available in NCBI GenBank accessions for tetrapods, exploring temporal and geographical patterns in georeferencing, and quantifying data available for automated phylogeography.

Location Global.

Methods We developed Python and R scripts to (1) download metadata from GenBank (1,125,514 accessions, > 20,000 species); (2) geocode accessions from associated metadata; (3) map originally georeferenced and geocoded accessions and plot their frequency against time; (4) assess the size of intraspecific sets of homologous sequences and compare their geographical extent with species ranges, thus evaluating their potential for phylogeographical analyses.

Results Only 6.2% of surveyed tetrapod GenBank submissions reported geographical coordinates, without increase in recent years. Our geocoding raised georeferenced accessions to 15.1%. The geographical distribution of georeferenced accessions is patchy, and especially sparse in economically underdeveloped areas. Automatically retrievable informative data sets covering most of the range are available for very few species of wide-ranging tetrapods.

Main conclusions Although geocoding offers a partial solution to the scarcity of direct georeferencing, the amount of data potentially useful for automated phylogeography is still limited. Strong underrepresentation of hard-to-access areas suggests that sampling logistics represent a main hindrance to global data availability. We propose that, besides enhancing georeferencing of genetic data, future research agendas should focus on collaborative efforts to sample genetic diversity in biodiversity-rich tropical areas.

Keywords
biodiversity, comparative phylogeography, DNA barcoding, GenBank, geodata, georeferencing, metadata

INTRODUCTION

To what extent do species in a regional assemblage share long-term demographic trends? What is the relative role of abiotic and biotic factors in shaping these trends? Ultimately, how much do local and regional communities represent stable, identifiable entities through time and space? Answers to these questions, and more, could emerge from a

Journal of Biogeography

comparative analysis of the spatial distribution of genetic diversity of species, that is, from comparative phylogeography (Bermingham & Moritz, 1998; Arbogast & Kenagy, 2001; Hewitt, 2004; Hickerson *et al.*, 2010). Molecular genetic data are probably the most important source of information about demographic histories (e.g. Drummond *et al.*, 2005; Schiffels & Durbin, 2014), and the analysis of spatial patterns of genetic diversity (and of gene genealogies in particular, i.e. phylogeography) has become the tool of choice to infer range dynamics (e.g. Hewitt, 2000).

As the spatial distribution of biodiversity emerges from the sum of range dynamics of independently evolving lineages (species), statistical comparisons of phylogeographical patterns across species may play a paramount role in revealing how current patterns of biodiversity were shaped by dispersal, diversification and competition (Marske *et al.*, 2013). However, most work published so far in the field of comparative phylogeography consists of two broad categories of studies, each with their own limitations. On the one hand, original research papers, although often relying on well-designed sampling schemes and rigorous statistical approaches, usually cover relatively small taxonomic or geographical scales (e.g. Carnaval *et al.*, 2009; Moritz *et al.*, 2009). On the other hand, review articles taking a larger scale approach typically consist of descriptive accounts of published results (e.g. Lorenzen *et al.*, 2012), while statistically rigorous meta-analyses remain rare and hindered by the challenge of comparing results obtained from different methods and study designs (Dawson, 2014). While the scope of original research papers tends to be obviously constrained by the amount of data that can be produced within a single research project, the weakness of the meta-analysis and review approaches often depends on analysing interpretations, or very simple summaries, of published data (not always with support from a solid statistical analysis), rather than the data themselves (e.g. Pyron & Burbrink, 2010; Shafer *et al.*, 2010). Recent, elegant efforts towards statistically robust meta-analyses have been published. However, they are often restricted to relatively simple spatial systems and/or implied the tiresome manual compilation of data sets from only those studies that reported the appropriate summary statistics (e.g. Riginos *et al.*, 2011, 2014; Paz-Vinas *et al.*, 2015).

For comparative phylogeography to mature as a research field, large amounts of primary data should be readily available to be analysed within a consistent framework. According to this view, an approach to comparative phylogeography involving automated retrieval of sequence data from nucleotide databases has the potential to bridge the gap between 'review/meta-analysis' and 'single study' approaches. This might represent a first step towards an 'automated phylogeography', that is, a set of tools for fully automated, self-updating analyses of available published data (for an example of a recent effort towards automated analyses in the field of phylogenetics, see Antonelli *et al.*, 2014). These might include web-based interfaces providing, for example, estimates of demographic trends for all species of mammals with available genetic data within a queried area, or maps of local intraspecific genealogical depth for a custom set of taxa. As most biological journals require that published nucleotide sequence data are deposited in INSDC databanks (International Nucleotide Sequence Database Collaboration: NCBI, DDBJ, EMBL, see Nakamura *et al.*, 2013), these represent the main resource for any research application based on DNA data retrieval. In 2005, NCBI GenBank introduced special fields for latitude and longitude, which allows for an immediate connection of nucleotide sequence data to the geographical location where samples were collected and opening the way for efficient automated retrieval of phylogeographically informative data. More recently, many journals have been asking that genetic data sets (usually, but not necessarily, including associated metadata) are made available as supplementary materials, or released in data repositories (e.g. datadryad.org). While this represented an important improvement towards reproducibility of analyses (see e.g. Mesirov, 2010), the diversity of sources and data formats makes this form of data storage rather unpractical as a resource for large-scale data analyses (see Hardisty & Roberts, 2013; Tedersoo *et al.*, 2015). Very recently, Weissenbacher *et al.* (2015) reported an attempt to automatically link geographical information contained in published papers to databases accessions. However, the precision of such methods is still limited, and improvements in their efficiency ultimately depend on interfacing information extracted from papers with metadata in the primary nucleotide databases (Weissenbacher *et al.*, 2015). Therefore, direct georeferencing of DNA sequences deposited in INSDC databases, such as the NCBI GenBank, appears as the most promising resource for a global-scale comparative phylogeography, and an exceptionally valuable resource for biogeography in general. Key aspects of biogeography, such as the estimation of global patterns of speciation/extinction rates, for example, would hugely benefit from readily available intraspecific georeferenced sequence data. Analyses of the geographical distribution of intraspecific genetic diversity can indicate where each species has persisted longest over time, a possible proxy for the region where species originated (see Losos & Glor, 2003, for an argument on why regular phylogenetic inference may not be sufficient for this task). Moreover, georeferenced sequence data might also serve as a powerful archive of biodiversity in the face of global change, allowing, for example, to track medium-term changes in allele frequencies and population structure.

An appeal for generalized georeferencing of biodiversity related data (and GenBank sequences in particular) was launched in a *Nature* editorial article in 2008 (Anonymous, 2008; but see also Guralnick *et al.*, 2007), and was warmly welcomed by a substantial group of researchers and institutions (Anonymous, 2008) and reiterated, for example, in a more recent letter to *Science* (Marques *et al.*, 2013). Since then, efforts were made to integrate nucleotide data and geographical information, for example, with the recent

publication of the 'Geographically tagged INSDC sequences' into the GBIF portal (www.gbif.org). Another, very important, contribution to the construction of a publicly available framework linking biogeographical and nucleotide data came with the establishment, in 2004, of the Consortium for the Barcode of Life (CBOL). CBOL has made a substantial move in this direction by requiring that the geographical origin of samples is recorded in the Barcode Of Life Data system (BOLD). Earlier attempts at estimating the amount of INSDC data with sufficient level of geographical information were reported by Scotch *et al.* (2011) on a subset of nucleotide sequences from RNA viruses and by Ryberg *et al.* (2008) and Tedersoo *et al.* (2011) on mycorrhizal fungi. However, these studies (1) focused on taxa with peculiar biogeographical features (dispersal of most well-known RNA viruses is generally human-mediated) and/or relatively limited taxonomic resolution and (2) did not explore temporal patterns in georeferenced sequence data and the completeness of geographical coverage for each species. Therefore, it remains unclear how much the calls for an integration of genetic and biogeographical data have been heeded, how much they translated into an increased frequency of nucleotide data georeferencing, and how much the accumulated data allow for a comparative phylogeography based on the automated retrieval of large multispecies data sets.

While the importance of linking genetic and geographical data was being brought to the attention of the scientific community, the next-Generation sequencing revolution (NGS; Ekblom & Galindo, 2011) created new opportunities for a global description of spatial genetic structures across population and species. On the one hand, NGS techniques can be used to sequence many copies of a few genetic markers (e.g. via amplicon sequencing: e.g. Bybee *et al.*, 2011; Wielstra *et al.*, 2014) that are then stored in standard databases like the GenBank. On the other hand, the more typical NGS application is the production of millions of genome-wide short sequence reads. Dedicated platforms, as the Short Reads Archive (SRA, www.ncbi.nlm.nih.gov/sra) allow storage and publicly sharing of huge amounts of data. As with non-NGS INSDC databanks, SRA can store georeferencing information about each sample, but this is not required in order to upload sequence data. Moreover, analyses including several species/taxa would imply prohibitive computational loads, if raw data from several different studies must be processed and compared for each taxon. Last, and possibly more importantly, availability of NGS-generated genomic data is still restricted to a relatively small number of species, usually with a small number of specimens on a limited geographical extent, so that the potential of this kind of data for comparative phylogeography and for tracking genetic diversity across space and time is still relatively limited.

In this study, therefore, we quantified georeferencing across INSDC accessions available through the NCBI GenBank, excluding NGS-dedicated platforms. We explored temporal trends (is georeferencing of sequence data becoming more/less common with time?), geographical patterns (is georeferencing more common in some regions than others?) and distribution across different sets of publications and genomic regions (which classes of genetic markers and research topics are contributing more georeferenced data?). Moreover, we estimated the amount of INSDC sequence data that can potentially be used for an automated approach to large-scale comparative phylogeography (i.e. an approach that avoids the time-consuming step of manually retrieving geographical information from the original publications). For our survey, we focused on terrestrial vertebrates (tetrapods), a large taxon for which abundant distributional and ecological data are available.

To this end, we (1) searched and stored INSDC accessions metadata for a list of 32,542 tetrapod species; (2) developed a custom geocoding pipeline to retrieve geographical coordinates from the GenBank textual 'country' field; (3) analysed temporal patterns in INSDC georeferencing across different subsets of data; (4) examined the geographical distribution of georeferenced accessions; (5) clustered georeferenced sequence by sequence similarity within species and evaluated the geographical coverage of alignable sequence clusters to quantify the amount of alignable data sets that could be automatically retrieved and used in phylogeographical studies (i.e. usable in an automated phylogeography).

METHODS

Obtaining sequence metadata from GenBank

We downloaded authoritative lists of 32,542 tetrapod species from class-specific taxonomy databases available online (see Appendix S1 in Supporting Information for details). 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 and obtained 21,262 unique species-level taxonIDs. We used species taxonIDs to search the NCBI GenBank and collaborative databases adhering to the INSDC using a custom Python script, filtering out RNA sequences, sequences from whole genome shotguns and other data categories (see Appendix S1 for details). GenBank searches were performed separately for each class from 21 November 2014 (mammals) to 9 December 2014 (reptiles).

For each accession that matched our search criteria, we stored the associated metadata (see Supplementary Methods for a list of stored fields). Humans (*Homo sapiens*) and a few highly synanthropic or domesticated species (listed in Appendix S1) were excluded from our search. The complete list of searched taxa is available in Appendix S2.

Georeferencing and geocoding of GenBank data

We considered as 'originally georeferenced', those accessions that contained unambiguous geographical coordinates in the 'lat_lon' field in one of the following formats: decimal degrees NSEW; decimal degrees with '–' to indicate S or W; degrees with decimal minutes. Although direct reporting of

geographical coordinates is obviously the most convenient way to make geographical information available, many GenBank accessions do contain information on the geographical origin of the sequenced samples in the 'country' field, which may contain very different levels of detail about the sampling location (see Scotch *et al.*, 2011). We considered as potentially 'informative' those accessions whose 'country' field contained some additional information besides the country itself (i.e. contained at least one ':' character – the almost universal separator that follows country names in GenBank 'country' fields – followed by a space, a number or a letter).

In order to estimate the amount of geographical information that could be automatically retrieved from GenBank metadata, we attempted to geocode (assign geographical coordinates based on a textual description) the subset of 'informative' accessions using a custom pipeline written in R and relying on geographical information from two publicly available databases: Global Administrative Areas (GADM) and GeoNames. Our geocoding pipeline is described in more detail in Appendix S1. In short, we started by isolating the country name from the rest of the text string in each unique 'country' field and removing most text not referring to placenames. The resulting character strings were then split at common separators and each substring was searched for in the subset of GADM and GeoNames databases relative to the country of interest. We recorded whether each substring appeared as such in the names (and variant names) of the administrative units from level 1 to level 5 in the GADM database. When a substring matched more than one administrative levels, we conservatively chose to consider the highest level match. Any substring not matching in the GADM database was searched for (as an exact match or partial match) in the GeoNames database. If at least one match was found in GADM for the same record, the substring was only searched among those GeoNames placenames whose geographical coordinates fell within the extent of the polygon corresponding to the matched GADM administrative unit. 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, no coordinates were assigned. Accessions that were unambiguously assigned to an administrative unit smaller than 10,000 km^2 or to a GeoNames toponym not representing an administrative feature were finally considered as successfully geocoded. In order to check the precision of our geocoding, we calculated the distance between our retrieved coordinates and the coordinates provided by the authors for those successfully geocoded unique 'country' fields that were originally georeferenced.

Taxonomic, temporal and spatial patterns in GenBank georeferencing

We first assessed the relative frequency of original georeferencing across GenBank accessions for each tetrapod class. We then analysed the temporal variation in the relative frequency of georeferencing across GenBank published submissions (here a 'submission' was defined as the combination of publishing year, authors list, journal and title extracted from the GenBank metadata) in the global data set and in four classes of publications: publications whose accessions are mostly (> 75%) mitochondrial DNA (mtDNA); publications with a reference to 'phylogeography', 'geographic structure', or similar in the title (title contains the string 'geograph'); publications with reference to DNA barcoding in the title (title contains the string 'barcod'); publications whose sequences are all deposited in the BOLD database (http://www.boldsystems.org/searched 29 January 2015), not counting BOLD sequences that were 'mined from GenBank'. We also explored the spatial distribution of georeferenced (both original and geocoded) GenBank data by counting the number of GenBank accessions and the number of species with at least one sequence on a global 200 × 200 km grid, and, for each class, we assessed the correlation between the number of originally georeferenced sequences in any map cell and tetrapod species richness in the same cell. Species richness layers were obtained by summing species ranges on the same 200 × 200 km global grid (an approach similar to Jenkins *et al.*, 2013). Polygons of species ranges were obtained by IUCN (www.iucnredlist.org, mammals and amphibians) and BirdLife (www.birdlife.org, birds). We did not perform the same comparison for reptiles as, while IUCN and BirdLife spatial data are taxonomically comprehensive for mammals, amphibians and birds, no similar complete data set is currently available for reptiles. Last, we estimated the number of INSDC accessions and the relative frequency of georeferencing by country, based on unambiguous country names in the 'country' field.

Evaluating data availability for automated phylogeography

To be potentially informative for phylogeography, a set of sequence data must (1) contain a number of gene copies from the same taxon (e.g. from the same species or cluster of closely related species); (2) represent a substantial portion of the range of the taxon of interest. In order to evaluate the amount of data potentially useful for automated phylogeography we (1) clustered our set of georeferenced GenBank accessions by sequence similarity and assessed the number of sequences in each cluster within each species (2) measured the spatial coverage of such within-species clusters relative to the distribution range of the species.

We obtained spatial data (shapefiles) for species ranges from the IUCN website (www.iucnredlist.org/technical-documents/spatial-data, downloaded 11 November 2014; amphibians: 6312 species; birds: 10,254 species; mammals: 5291 species, excluding marine mammals; reptiles: 3903 species). Therefore, we limited this analysis to those georeferenced GenBank accessions whose organism name

(field 'organism') could be unambiguously matched to a Linnean binomial listed in the IUCN spatial data. More specifically, we first searched for exact matches between genus names appearing in the INSDC accession (first word of the field 'organism') and the genus name of the Linnean binomial as reported in the IUCN shapefiles. To account for typing errors, non-matching genus names were then searched for partial matches, and the type of match was recorded ('exact match', 'partial match', 'genus not present'). Once assigned a (putative) genus name, we searched for the second term of the field 'organism' within the species names of the selected genus. Again, to account for typing errors and changes in species name/gender (e.g. *Actitis macularia* versus *Actitis macularius*), non-matching species names were searched for partial matches. Finally, all partial matches were manually checked for concordance, and 23 binomials with apparent mismatch were excluded.

Selected accessions were clustered by sequence similarity within each genus, using a custom R script calling for the fast sequence clustering algorithm uclust (Edgar, 2010; implemented in UCLUSTQ 1.2.22). Sequences shorter than 200 bp (1846 accessions) were excluded from clustering, as well as sequences longer than 5000 bp (233 accessions). The latter choice aimed at avoiding clustering of unrelated sequences to a few, very long sequences (e.g. complete mtDNA: our data set contained 136 georeferenced complete mtDNA sequences; maximum for a single species = 6). We set the minimum identity threshold for clustering at 0.80, and run uclust with the –optimal option. The latter choice prompts a clustering algorithm ensuring that each sequence is clustered together with its best match, and minimizes the number of recovered clusters given the chosen threshold. All sequence clusters that contained at least 20 sequences from the same species were retained as potentially interesting.

To evaluate the spatial coverage of each cluster relative to the species range, we first rasterized the shapefiles for each range to a *c.* 100 × 100 km grid (Behrmann cylindrical equal area projection), we then partitioned the resulting cells in 20 clusters of neighbouring cells using a k-means clustering (R function kmeans() {stats}with 1000 random starts and 1000 iterations) and, last, we counted the number of such clusters (corresponding to 20 Voronoi cells) that contained at least one sequence. All species whose ranges intersected < 20 cells (roughly 200,000 km^2) were excluded from this analysis.

RESULTS

Obtaining sequence metadata from GenBank

Our search returned metadata for 1,125,514 accessions (see Fig. 1 for the distribution across classes), with 26,077 unique taxonIDs at the specific or subspecific level, representing 20,810 Linnean binomials (the exact number of species may vary due to taxonomic instability) and 4201 unique taxonIDs at the genus level.

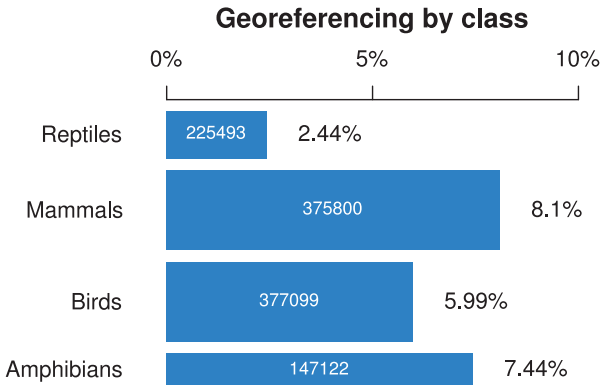

Figure 1 Proportion of originally georeferenced data in a selection of 1,125,514 tetrapods INSDC accessions. Length of bars and black numbers besides bars represent the relative frequency of georeferencing across accessions for a given class. Bar depths and white number within bars represent the total number of accessions for each class. Full color version available online.

Georeferencing and geocoding of GenBank data

The geographical coordinates ('lat_lon') field was filled for 69,782 accession (6.20% of all retrieved accessions), with 161 of them containing non-numeric strings. In the following, we consider as 'originally georeferenced' the 69,483 accessions (6.17%) containing consistently interpretable latitude and longitude in one of the three most common formats (decimal degrees NSEW: 68,855 accessions; decimal degrees with '–' to indicate S or W: 421 accessions; degrees with decimal minutes: 207 accessions). We note that, as we did not check for consistency of coordinates at this stage (e.g. we did not check that a point for a terrestrial animal did not fall in the ocean), it is not guaranteed that all of the originally georeferenced accessions are also correctly georeferenced.

The 'country' field was populated in 379,200 accessions (33.7%), while 377,357 (33.5%) contained unambiguous country names (ambiguous names include, for example, names referring to dissolved political entities, such as 'Yugoslavia'). A total of 270,476 (24.0%) of them were considered as 'informative' (39,813 unique 'country' fields, with 39,627 of them containing a valid country name). According to our criteria, 23,238 unique 'country' fields (58.6% of the unique searched fields) were successfully geocoded (i.e. they were unambiguously assigned to an administrative unit smaller than 10,000 km^2 or to a point placename). Our geocoding thus successfully assigned coordinates to 101,082 accessions, thus raising the total amount of georeferenced accessions to 170,565 (15.1% of the total retained accession). Among the successfully geocoded accessions, 4624 were originally georeferenced, and were used to check geocoding precision. We found that assigned coordinates were within 100 km of the original coordinates in 90.3% of cases, and within 200 km in 94.6% of cases.

Taxonomic, temporal and spatial patterns in INSDC georeferencing

The relative frequency of original georeferencing varied across classes, being markedly lower in reptiles than in amphibians, birds or mammals (Fig. 1).

The analysed set of tetrapods GenBank accessions has been accumulating quasi-exponentially in the last 25 years (Fig. 2a). Georeferenced accessions increased rapidly for *c.* 5 years after the introduction of the 'lat_lon' field in 2005 (the few georeferenced accessions pre-dating 2005 result from post-publishing additions), but, at least in the last 5–6 years, their growth rate did not exceed that of the total data set. Georeferenced+geocoded accessions accumulated at a steeper rate than the total data before the year 2000, but then flattened out too. Temporal trends in the frequency of submissions whose accessions are mostly georeferenced is shown in Fig. 2(b). The trend for the total data set reflects the accession-wise growth pattern in Fig. 2(a), with an increase in 2005–2010 and subsequent flattening out, at a rather low frequency *c.* 5–6%. Mitochondrial DNA sequences (mtDNA) have been the workhorse for phylogeographical inference since the late 1980s, therefore, it would be reasonable to expect that papers employing mtDNA as the main marker (i.e. > 75% of related accessions) display a markedly higher frequency of georeferencing. Figure 2(b) shows that this is not the case, as the values for mtDNA-based submissions are just slightly above the total data set in the original+geocoded data, and essentially average in the originally georeferenced data. Even more unexpectedly, papers whose title contains explicit reference to geography also do not differ markedly from the total data set considered in this study (Fig. 2b). As the Barcode of Life database (BOLD) strongly recommends georeferencing for all accessions deposited there, it is not surprising that the few submissions whose sequences are (mostly) deposited in BOLD have a very high georeferencing frequency (Fig. 2b). Georeferencing frequency across papers referring to 'barcode' in the title is lower, but still markedly higher than in the global data set (Fig. 2b). The georeferencing frequency of both barcode-related sets of submissions apparently decreases with time, and particularly so for submissions referring to DNA barcode in their title (Fig. 2b). Nonetheless, BOLD accessions, which represented only 3.4% of the analysed data, contributed a large share of the total georeferenced sequences (20.2%), and about half (47.3%) of the originally georeferenced accessions.

Originally georeferenced sequences are sparsely distributed, with large areas of land lacking any data (especially in socio-economically underdeveloped and/or scarcely inhabited regions) (Fig. 3a). Our geocoding substantially increased the amount of mappable data, but did not alter the general pattern (Fig. 3b). The same considerations apply for the number of species with mappable sequence data (Fig. 4a,b). Moreover, data from the different taxonomic classes do not follow

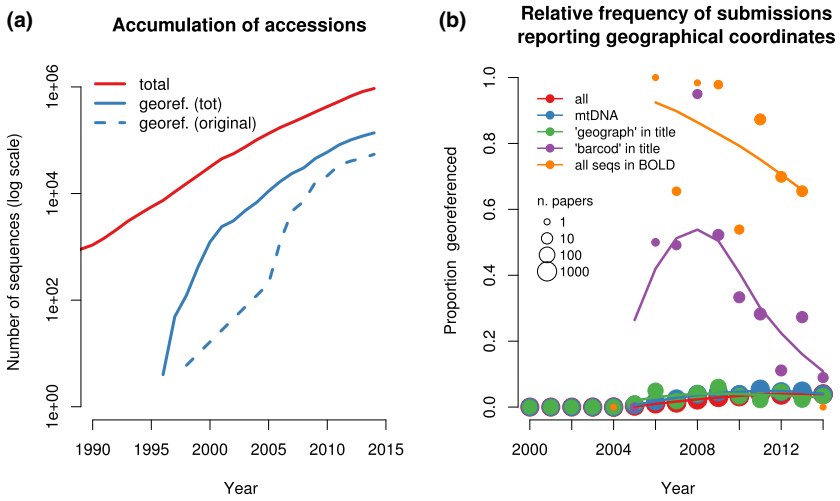

Figure 2 Accumulation of georeferenced GenBank accessions and frequency of submissions with mostly georeferenced data. (a) Growth of total and georeferenced NCBI GenBank accessions. Cumulative distribution of a selection of 934,556 published tetrapods accessions (red, continuous line) by year of publication. Cumulative distribution of accessions reporting an unambiguously interpretable 'lat_lon' field (originally georeferenced: blue, dashed line) and accessions that were georeferenced, including geocoded accessions (originally georeferenced + geocoded: blue, continuous line). Data are presented on a logarithmic vertical scale. (b) Relative frequency of submission that reported geographical coordinates for most of their accessions across time in a selection of 13,144 unique published submissions to GenBank containing tetrapods sequences. The proportion of published submissions whose sequences are georeferenced for more than 50% is plotted against the publication year for the total data set (red) and four subsets of data: submissions containing mostly (> 75%) mtDNA sequences (N = 6187, blue); submission containing a reference to geography (string 'geograph') in publication title (N = 1347, green); submissions containing a reference to DNA barocoding (string 'barcod') in publication title (N = 71, purple); submissions whose sequences are all deposited in the BOLD (Barcode Of Life Database) systems (N = 39, orange). The area of circles is proportional to the logarithm of the sample size for each year. Lines are LOESS trends since 2005 (special 'lat_lon' field introduced in the GenBank). Full color version available online.

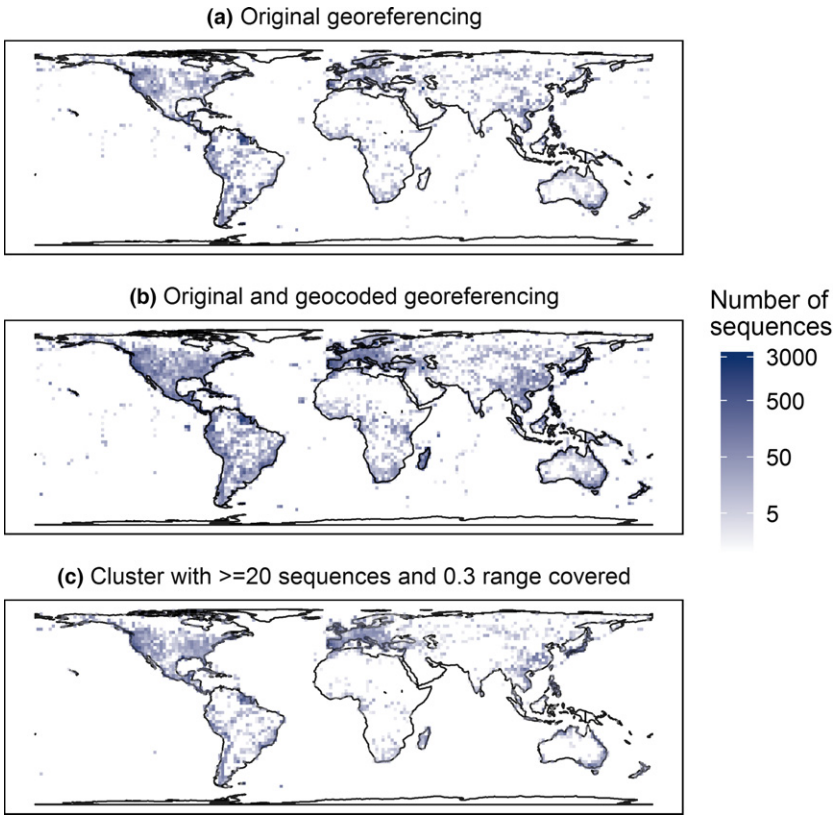

Figure 3 Spatial distribution of georeferenced sequences in a selection of 1,125,514 tetrapods GenBank accessions. Maps show the number of sequences georeferenced to each cell of a 200 × 200 km grid. (a) Originally georeferenced sequences; (b) total georeferenced sequences (original + geocoded); (c) sequences belonging to intraspecific clusters with at least 20 sequences and whose geographical extent covers at least 30% of the species range (see text for details). The colour scale is proportional to the logarithm of the number of sequences. All maps are in the Lambert cylindrical equal area projection. Full color version available online.

a very consistent pattern, with low to modest correlation coefficients (see Appendix S3). In all classes, we found weak (though significant) correlations between the number of sequences with original georeferencing to a given grid cell (see Fig. 3a) and the number of species predicted in the same cell (Pearson's r, amphibians: 0.10, $P = 0.015$; birds: 0.23, $P < 0.001$; mammals 0.28, $P < 0.001$), and between the number of species with sequences (originally) georeferenced to a given cell (see Fig. 4a) and the number of species predicted in the cell (Pearson's r, amphibians: 0.28, birds: 0.22, mammals: 0.32; all $P < 0.001$).

The number of INSDC accessions by country and the relative frequency of originally georeferenced data are shown in Fig. 5. Most African countries have very few data, and there is no strong regional pattern in the relative frequency of georeferencing, as above average (red) and below average (blue) countries are scattered across every continent.

Evaluating data availability for automated phylogeography

A total of 1327 clusters in 761 species and 435 genera (summing up to 86,122 accessions) contained at least 20 sequences from the same species and were retained as potentially interesting.

The distribution of within-species clusters containing at least 20 sequences according to the number of sequences in the cluster and the proportion of the species range covered are illustrated in Fig. 6 and Table 1. Very few intraspecific sets of automatically georeferenceable GenBank sequences contain a large number of sequences and provide a satisfactory coverage of the species range (Table 1). In fact, a reliable, range-wide description of intraspecific phylogeographical patterns seems to be possible only for a handful of tetrapod species. Even when we considered a very permissive threshold (30% of the species range covered with at least one sequence), the spatial distribution of the potentially usable clusters is highly dispersed, with very few areas covered by more than a few clusters (Figs 4c & 5c). Again considering a conservative threshold at 30% of the range, we can track the enrichment of certain categories of data along the process from georeferencing to assessment of potential for phylogeography. Mitochondrial DNA accounts for 70.1% of the retained sequences, increasing from 47.4% in the total data and 67.7% in the georeferenced (original + geocoded) data. Publications with a reference to geography in the title contribute 28.7% of the retained data (15.8% in the total data and 17.1% in the georeferenced set). BOLD data represent 21.1% of retained data (3.4% in the total data and 20.3% in the georeferenced data set).

DISCUSSION

Our analysis of GenBank (and associated INSDC databanks) accessions for over 20,000 tetrapod species showed that, despite a quasi-exponential growth of the absolute number of accessions that directly report geographical coordinates, such georeferenced accessions still represent a very small

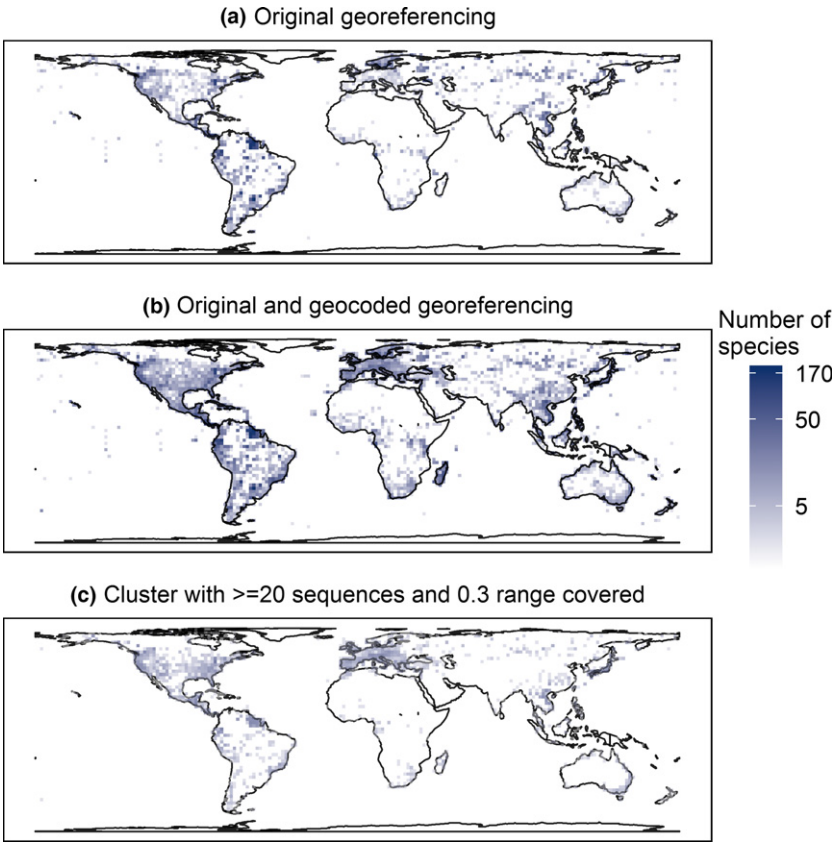

Figure 4 Spatial distribution of species with georeferenced sequences in a selection of 1,125,514 tetrapods GenBank accessions. Maps show the number of species with at least one sequence georeferenced to each cell of a 200 × 200 km grid. (a) Originally georeferenced sequences; (b) total georeferenced sequences (original + geocoded); (c) sequences belonging to intraspecific clusters with at least 20 sequences and whose geographical extent covers at least 30% of the species range (see text for details). The colour scale is proportional to the logarithm of the number of species. All maps are in the Lambert cylindrical equal area projection. Full color version available online.

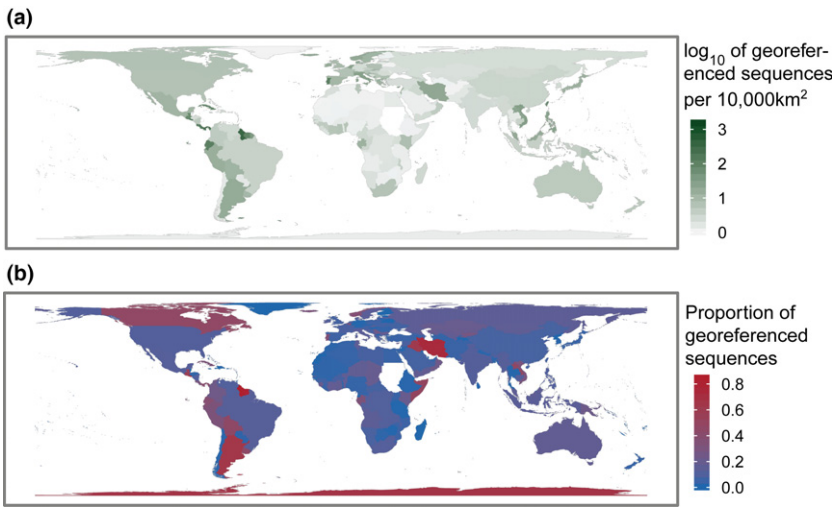

Figure 5 Georeferencing by country in a selection of 1,125,514 tetrapods GenBank accessions. In Fig. 5(a) colour intensity is proportional to the number of GenBank accessions that can be traced to each country (unambiguous country name appearing in the 'country' field) divided by the country area (log scale). In Fig. 5(b), colour scale indicates the proportion of originally georeferenced sequences among those that can be tracked to each country (white indicates no data). All maps are in the Lambert cylindrical equal area projection. Full color version available online.

fraction (6.2%) of the available sequence data for terrestrial vertebrates. Moreover, and very importantly, despite the widespread recognition of the potential importance of linking genes to geography (e.g. Field, 2008; Marques *et al.*, 2013), the relative frequency of georeferencing across GenBank submissions has not been increasing in recent years. Although our INSDC search did not target any specific genomic region, and considered all kind of publications, including reports from physiology studies or gene-expression experiments, this relative scarcity is not essentially due to the prevalence of submissions dealing with topics not related with biogeography. In fact, we found that the frequency of georeferencing is also very low across mitochondrial sequences (largely employed in phylogeography, and representing 47.5% of the accessions we surveyed) and among papers with a reference to geographyin their title (18.9% of the published accessions we surveyed). The general paucity of georeferenced data may well explain their patchy global distribution (Figs 3 & 4) and lack of consistency across taxonomic classes (see Appendix S3). Indeed, a large share of the available georeferenced accessions originate from a limited number of local initiatives (e.g. DNA barcoding of national

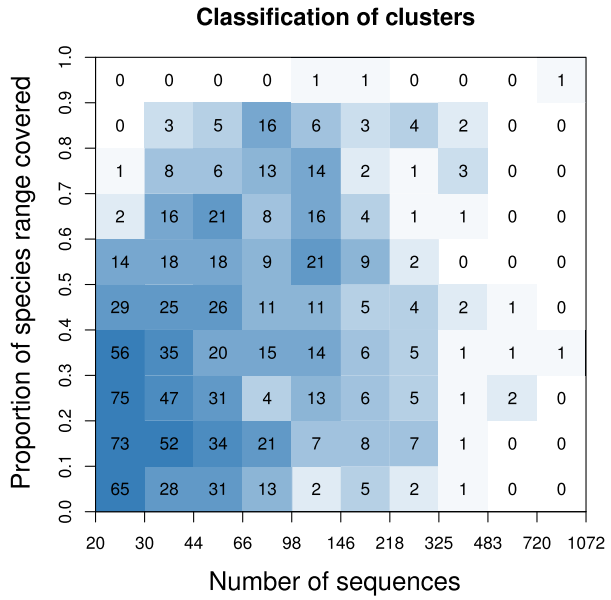

Figure 6 Classification of intraspecific sets (clusters) of putatively homologous GenBank sequences according to the number of sequences and the proportion of species range covered (spatial coverage). The number of clusters in each category is indicated, with colour scale proportional to the logarithm of the number of clusters. Full color version available online.

Table 1 Number of potentially informative intraspecific clusters contained in a globally surveyed selection of 1,125,514 tetrapods INSDC accessions (see text for details) according to the number of sequences contained and the proportion of the species range covered (sequences: condition regarding the number of sequences contained in the cluster; range: condition regarding the proportion of the species range covered by the sequences contained in the cluster; clusters: number of clusters satisfying the conditions for sequences and range; species and genera: number of species and genera represented in the clusters that satisfy the conditions for sequences and range).

Sequences	Range	Clusters	Species	Genera
≥ 20	≥ 0.3	611	332	221
≥ 50	≥ 0.3	308	165	123
≥ 100	≥ 0.3	155	89	66
≥ 20	≥ 0.5	299	140	107
≥ 50	≥ 0.5	202	91	75
≥ 100	≥ 0.5	104	52	44
≥ 20	≥ 0.8	65	29	25
≥ 50	≥ 0.8	61	27	24
≥ 100	≥ 0.8	28	17	15

faunal assemblages, Clare *et al.*, 2007). Differences in accessibility among regions, together with other geopolitical and socio-economic factors, might influence the absolute distribution of georeferenced data. Indeed, scarcely populated circumpolar and dry regions have very little data, while data are very dense throughout the highly developed areas of North America and Europe, and are extremely scarce across vast economically underdeveloped areas of Africa and Asia (Figs 4, 5 & 6a). On the other hand, the relative frequency of georeferencing across countries does not show any interpretable pattern, as countries with high and low relative georeferencing are spread across every continent (Fig. 5). Interestingly, our findings about the relative frequency of INSDC data georeferencing and their geographical distribution in tetrapods are qualitatively very similar to previous reports from very different groups of organisms such as mycorrhizal fungi (Tedersoo *et al.*, 2011) and RNA viruses (Scotch *et al.*, 2011). This suggests that the trends we observed for tetrapods might not be characteristic to a special taxonomic group, but rather represent general features of INSDC databases.

The reasons why so many INSDC data are submitted without explicit georeferencing cannot be assessed through an analysis like ours. Sensible hypotheses fall into a few general categories: (1) genuine lack of precise geographical information; (2) unwillingness to reveal sensitive data (e.g. for samples from threatened species or populations); (3) lack of interest and awareness about the potential importance of direct georeferencing of data deposited in nucleotide databases for large-scale reanalysis of sequence data (in fact, publishing the relevant geographical information in appendices and supplementary tables is not equivalent to easy access) and/or lack of time: researchers are often pressed for time and tend to skip accurate annotation (see Ryberg *et al.*, 2008). Regarding the last points, the only viable solution lies in efforts to educate the growing community of biogeographers and ecologists about the need to provide readily accessible georeferenced data, to which we hope our study, may contribute. Moreover, databank policies might actively encourage best-practices in metadata enrichment, for example by asking submitters to explain why some fields are left empty. Importantly, INSDC databanks could take action to open themselves to verifiable third-party annotation, which are currently not encouraged (see Pennisi, 2008). On the other hand, we suggest that neither lack of precision or perceived sensitivity of the data should refrain researchers from making georeferenced data as available as possible. In fact, geographical coordinates may be often precisely inferred *a posteriori* even when sampling was carried out without positioning devices and, even when submitted at a relatively low level of precision, may still prove useful for many applications, provided that precision estimates are provided. The Barcode Of Life Data system (BOLD) proposes a good example in this respect, in containing a special field ('coord_accuracy') to this end. Unfortunately, though, very few BOLD accessions actually use it (only 0.02% of 56,866 georeferenced tetrapod data excluding humans and sequences mined from GenBank). The BOLD system, while containing a proportionally low number of the GenBank accessions considered in this study (38,435 accessions on over one million), contributes significantly to the available georeferenced data (> 20%). Our results thus highlight the importance of this initiative towards the creation of a global archive of genetic diversity. However, we note that this contribution is

somewhat limited by the relatively large proportion of unreleased sequences (28.0% of 63,480 tetrapod data in BOLD).

We also showed that a relatively simple geocoding algorithm may allow to largely increase (*c.* 3-fold) the amount of nucleotide sequence data that can be placed on a map with an accuracy (< 10,000 km^2 or 100 km radius) that would allow acceptably precise description of phylogeographical patterns at a continental scale (Figs 2, 4 & 5).

Lastly, our classification of intraspecific data sets based on the number of samples and spatial coverage of the species range, clearly indicates that the number and geographical distribution of data sets potentially amenable to automated phylogeographical analyses (Figs 4c, 5c & 6) is still insufficient for a satisfactory description of global patterns of genetic diversity. Although it is not possible to have an exact estimate of the amount of potentially informative data that is not available due to lack of georeferencing, our analyses allow for a more educated guess, whose result is rather sobering. As we managed to obtain some georeferencing for little less than one-sixth of the considered data (15.1%), we can project that, if all INSDC data that we have considered in this study were georeferenced, the availability would be moved up by no more than a factor of 6. From Table 1 we see that such an upgrade would bring the number of species with at least 100 sequences from at least 80% of the range to *c.* 100. Although this would seem promising, Figs 4(c) and 5(c) suggest that, as we did not find a consistent geographical pattern in the relative frequency of georeferencing among countries (Fig. 6), continental scale comparative phylogeography involving more than very few species would likely only be possible for North America and Europe, while, for other continents, the amount of accumulated data would be barely sufficient for exploratory analyses.

This rather sobering picture might change quickly, though, if the next wave of biogeographical and phylogeographical studies will be able to seize the opportunities opened by NGS techniques, which produce millions of sequences at very low cost. In our survey, we did not consider most NGS-generated genomic data, mostly because the availability of genomic data through taxonomic and geographical space is still particularly limited. Another key practical concern is the computational loads needed to analyse raw data from several different studies for comparative purposes. Storing available information as sets of processed data might represent a promising alternative. NGS archives (like the SRA) could encourage the submission of summary tables using formats like the variant call format (VCF; Danecek *et al.*, 2011). This standard format can be used to summarize genome-wide diversity as SNPs, SSR, insertion/deletion, sorted along a reference genome or according to a *de novo* assembled set of contigs, from different sources of data (whole-genome sequencing, SNP-chip, RADseq). However, the integration of different data sets into effective analytical pipelines may still be challenging, because genomic data produced by different studies can be highly heterogeneous. We expect that novel, computationally affordable, approaches capable of exploiting the exponentially growing amount of genomic data will be developed soon, and that a standard set of genome-level, multiple 'barcode' markers might be eventually defined.

Our results show that, at least for terrestrial vertebrates, geographically biased sampling represents the main limiting step to provide a global reach to genetic diversity assessments. We suggest that future research agendas may focus on (1) widespread georeferencing of newly produced data, possibly involving more careful quality control by agencies managing nucleotide databases and more direct linking between different databases (e.g. adding a GBIF voucher to GenBank sequences, when available); (2) concerted efforts towards annotation of existing databases (see, e.g. Nilsson *et al.*, 2014) (3) use of NGS techniques to produce extensive data on multiple individuals of several species; (4) coordination of large-scale collaborative efforts for biodiversity sampling, with a special attention to tropical biodiversity-rich areas in general, and to Africa in particular.

ACKNOWLEDGEMENTS

This study was mainly supported by the Max Planck Society through the Pan African Programme (panafrican.eva.mpg.de). The authors are grateful to Michael Hofreiter for his valuable support during the development of this study, to Katharine Ann Marske and an anonymous referee for their exceptionally constructive comments and their active contribution to improve the first draft of the manuscript and to Dylan Craven for revising language and writing style.

REFERENCES

Anonymous (2008) A place for everything: more researchers must record the latitude and longitude of their data. *Nature*, **453**, 2. doi:10.1038/453002a.

Antonelli, A., Condamine, F.L., Hettling, H., Nilsson, K., Nilsson, R.H., Oxelman, B., Sanderson, M.J., Sauquet, H., Scharn, R., Silvestro, D., Töpel, M. & Vos, R.A. (2014) SUPERSMART: ecology and evolution in the era of big data. *PeerJ PrePrints*, **2**, e501v1. Available at: https://dx. doi.org/10.7287/peerj.preprints.501v1. Last accessed 1 June 2016.

Arbogast, B.S. & Kenagy, G.J. (2001) Comparative phylogeography as an integrative approach to historical biogeography. *Journal of Biogeography*, **28**, 819–825.

Bermingham, E. & Moritz, C. (1998) Comparative phylogeography: concepts and applications. *Molecular Ecology*, **7**, 367–369.

Bybee, S.M., Bracken-Grissom, H., Haynes, B.D., Hermansen, R.A., Byers, R.L., Clement, M.J., Udall, J.A., Wilcox, E.R. & Crandall, K.A. (2011) Targeted amplicon sequencing (TAS): a scalable next-gen approach to multilocus, multitaxa phylogenetics. *Genome Biology and Evolution*, **3**, 1312–1323.

Carnaval, A.C., Hickerson, M.J., Haddad, C.F., Rodrigues, M.T. & Moritz, C. (2009) Stability predicts genetic

diversity in the Brazilian Atlantic forest hotspot. *Science*, **323**, 785–789.

Clare, E.L., Lim, B.K., Engstrom, M.D., Eger, J.L. & Hebert, P.D. (2007) DNA barcoding of Neotropical bats: species identification and discovery within Guyana. *Molecular Ecology Notes*, **7**, 184–190.

Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T., McVean, G. & Durbin, R. (2011) The variant call format and VCFtools. *Bioinformatics*, **27**, 2156–2158.

Dawson, M.N. (2014) Natural experiments and meta-analyses in comparative phylogeography. *Journal of Biogeography*, **41**, 52–65.

Drummond, A.J., Rambaut, A., Shapiro, B. & Pybus, O.G. (2005) Bayesian coalescent inference of past population dynamics from molecular sequences. *Molecular Biology and Evolution*, **22**, 1185–1192.

Edgar, R.C. (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, **26**, 2460–2461.

Ekblom, R. & Galindo, J. (2011) Applications of next generation sequencing in molecular ecology of non-model organisms. *Heredity*, **107**, 1–15.

Field, D. (2008) Working together to put molecules on the map. *Nature*, **453**, 978–978.

Guralnick, R.P., Hill, A.W. & Lane, M. (2007) Towards a collaborative, global infrastructure for biodiversity assessment. *Ecology Letters*, **10**, 663–672.

Hardisty, A. & Roberts, D. (2013) A decadal view of biodiversity informatics: challenges and priorities. *BMC Ecology*, **13**, 16.

Hewitt, G.M. (2000) The genetic legacy of the Quaternary ice ages. *Nature*, **405**, 907–913.

Hewitt, G.M. (2004) The structure of biodiversity–insights from molecular phylogeography. *Frontiers in Zoology*, **1**, 1–16.

Hickerson, M.J., Carstens, B.C., Cavender-Bares, J., Crandall, K.A., Graham, C.H., Johnson, J.B., Rissler, L., Victoriano, P.F. & Yoder, A.D. (2010) Phylogeography's past, present, and future: 10 years after. *Molecular Phylogenetics and Evolution*, **54**, 291–301.

Jenkins, C.N., Pimm, S.L. & Joppa, L.N. (2013) Global patterns of terrestrial vertebrate diversity and conservation. *Proceedings of the National Academy of Sciences USA*, **110**, E2602–E2610.

Lorenzen, E.D., Heller, R. & Siegismund, H.R. (2012) Comparative phylogeography of African savannah ungulates. *Molecular Ecology*, **21**, 3656–3670.

Losos, J.B. & Glor, R.E. (2003) Phylogenetic comparative methods and the geography of speciation. *Trends in Ecology and Evolution*, **18**, 220–227.

Marques, A.C., Maronna, M.M. & Collins, A.G. (2013) Putting GenBank data on the map. *Science*, **341**, 1341.

Marske, K.A., Rahbek, C. & Nogués-Bravo, D. (2013) Phylogeography: spanning the ecology-evolution continuum. *Ecography*, **36**, 1169–1181.

Mesirov, J.P. (2010) Computer science. Accessible reproducible research. *Science*, **327**, 415–416.

Moritz, C., Hoskin, C.J., MacKenzie, J.B., Phillips, B.L., Tonione, M., Silva, N., VanDerWal, J., Williams, S.E. & Graham, C.H. (2009) Identification and dynamics of a cryptic suture zone in tropical rainforest. *Proceedings of the Royal Society B: Biological Sciences*, **276**, 1235–1244.

Nakamura, Y., Cochrane, G. & Karsch-Mizrachi, I. (2013) The international nucleotide sequence database collaboration. *Nucleic Acids Research*, **41**, D21–D24.

Nilsson, R.H., Hyde, K.D., Pawłowska, J. *et al.* (2014) Improving ITS sequence data for identification of plant pathogenic fungi. *Fungal Diversity*, **67**, 11–19.

Paz-Vinas, I., Loot, G., Stevens, V.M. & Blanchet, S. (2015) Evolutionary processes driving spatial patterns of intraspecific genetic diversity in river ecosystems. *Molecular Ecology*, **24**, 4586–4604.

Pennisi, E. (2008) Proposal to 'wikify' GenBank meets stiff resistance. *Science*, **319**, 1598–1599.

Pyron, R.A. & Burbrink, F.T. (2010) Hard and soft allopatry: physically and ecologically mediated modes of geographic speciation. *Journal of Biogeography*, **37**, 2005–2015.

Riginos, C., Douglas, K.E., Jin, Y., Shanahan, D.F. & Treml, E.A. (2011) Effects of geography and life history traits on genetic differentiation in benthic marine fishes. *Ecography*, **34**, 566–575.

Riginos, C., Buckley, Y.M., Blomberg, S.P. & Treml, E.A. (2014) Dispersal capacity predicts both population genetic structure and species richness in reef fishes. *The American Naturalist*, **184**, 52–64.

Ryberg, M., Nilsson, R.H., Kristiansson, E., Töpel, M., Jacobsson, S. & Larsson, E. (2008) Mining metadata from unidentified ITS sequences in GenBank: a case study in *Inocybe* (Basidiomycota). *BMC Evolutionary Biology*, **8**, 50.

Schiffels, S. & Durbin, R. (2014) Inferring human population size and separation history from multiple genome sequences. *Nature Genetics*, **46**, 919–925.

Scotch, M., Sarkar, I.N., Mei, C., Leaman, R., Cheung, K.H., Ortiz, P., Singraur, A. & Gonzalez, G. (2011) Enhancing phylogeography by improving geographical information from GenBank. *Journal of Biomedical Informatics*, **44**, S44–S47.

Shafer, A., Cullingham, C.I., Cote, S.D. & Coltman, D.W. (2010) Of glaciers and refugia: a decade of study sheds new light on the phylogeography of northwestern North America. *Molecular Ecology*, **19**, 4589–4621.

Tedersoo, L., Abarenkov, K., Nilsson, R.H., Schussler, A., Grelet, G.A., Kohout, P., Oja, J., Bonito, G.M., Veldre, V., Jairus, T., Ryberg, M., Larsson, U. & Kõljalg, U. (2011) Tidying up international nucleotide sequence databases: ecological, geographical, and sequence quality annotation of ITS sequences of mycorrhizal fungi. *PLoS ONE*, **6**, e24904.

Tedersoo, L., Ramirez, K.S., Nilsson, R.H., Kaljuvee, A., Kõljalg, U. & Abarenkov, K. (2015) Standardizing metadata and taxonomic identification in metabarcoding studies. *GigaScience*, **4**, 1–4.

Weissenbacher, D., Tahsin, T., Beard, R., Figaro, M., Rivera, R., Scotch, M. & Gonzalez, G. (2015) Knowledge-driven

geospatial location resolution for phylogeographic models of virus migration. *Bioinformatics*, **31**, i348–i356.

Wielstra, B., Duijm, E., Lagler, P., Lammers, Y., Meilink, W.R.M., Ziermann, J.M. & Arntzen, J.W. (2014) Parallel tagged amplicon sequencing of transcriptome-based genetic markers for Triturus newts with the Ion Torrent next-generation sequencing platform. *Molecular Ecology Resources*, **14**, 1080–1089.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article:

Appendix S1 Methods.
Appendix S2 List of searched taxa.
Appendix S3 Class-wise geographical patterns of georeferencing.

DATA ACCESSIBILITY

R and Python scripts developed for and used in this study are available at https://github.com/paolo-gratton/Gratton_et_al_JBiogeogr_2016.

BIOSKETCHES

Paolo Gratton is a population geneticist and biogeographer, and a current PostDoc fellow at the Max Planck Institute for Evolutionary Anthropology, Leipzig. His research interests focus on the mechanisms through which biodiversity (at the level of genes, populations, species, and communities) is generated and maintained by spatial and temporal environmental variation.

Paolo Gratton, **Hjalmar Kühl** and **Gaëlle Bocksberger** are currently developing a research project centred on using published genetic data to evaluate evidences for community-level refugia in Sub-Saharan Africa.

Author contributions: P.G., G.B., H.K. and M.W. conceived the original idea for this study, which was further developed with the contribution of S.M. and E.T. S.M., P.G. and G.B. developed and performed the data analysis. P.G. led the writing, which was performed by all authors.

Editor: Alexandre Antonelli