University of Glasgow

BIOL4246P Investigative Honours Project

Supervisor: Professor Rod Page

Mapping the global distribution of phylogenetic diversity using DNA barcodes

Word count: 5776

2330339, BSc Hons Zoology
17/03/2021

Contents

Abstract:

Different methods of quantifying biodiversity have been applied to determine conservation priority for geographic regions. The phylogenetic diversity PD of an assemblage *G* is the sum of branch lengths in a phylogenetic tree which describes distance between individuals or taxa in *G*. A web application was developed to compute a sample-based PD index PD_N for areas across the globe using DNA barcodes from BOLD, to assess the feasibility of barcode-based PD methods. Spatial pattern of the index was analysed and compared to expected species number. PD_N was found to correlate with sampling effort (Likelihood ratio test: $\chi^2 = 15.439$, df = 1, p=0.00008523). It was concluded that further use of DNA barcodes in PD must correct for the significant spatial differences in sampling effort.

Introduction

Measuring biodiversity

Conservation biology originated the term “biodiversity” and has provided the context for popularizing the idea (Maclaurin and Sterelny, 2008). Consequently, the metrics developed for biodiversity assessment reflect conservationist goals (Lean and Maclaurin, 2006). Heightened extinction rate from anthropogenic causes (Ceballos et al., 2015) is closely associated with the accelerating destruction of globally unique habitats (Silva et al, 2021; Taubert et al., 2018); accordingly, biodiversity may be measured as a property of habitats and so studied in terms of spatial patterns (Kinlock et al., 2017).

A pressing concern for conservation is prioritisation of threatened habitats, to enable preservation of maximal diversity with limited resources (Weitzman, 1998). Choice of biodiversity measurement thus raises philosophical questions of what units we seek to measure, and why we value them (Faith, 2017). For application to this prioritisation problem, metrics of biodiversity must be thoroughly understood in their construction and interpretation (Faith, 2006).

A host of different metrics have been used to assess biological groups at differing levels, from Whittaker (1972) partitioning variation within and between environments (alpha, beta, and gamma diversity), to modern developments leveraging large genomic datasets in fields such as community ecology (Djurhuus et al., 2020) and epidemiology (van Dorp et al., 2020). This report considers an index of biodiversity based on Faith’s framework of phylogenetic diversity, PD (Faith, 1992), and seeks to assess its applicability to sets of barcode DNA sequences on a global scale.

Phylogeny-based measures of biodiversity

Phylogenetic diversity, and related indices, contrast with diversity measures based on presence and abundance of species or taxonomic groups. The latter treat taxa (typically species) as types with equal status—apart from relative abundance—so do not consider distinctness or relatedness. The same value of species richness may describe assemblages of species containing few or many higher-level taxonomic groups (figure 1).

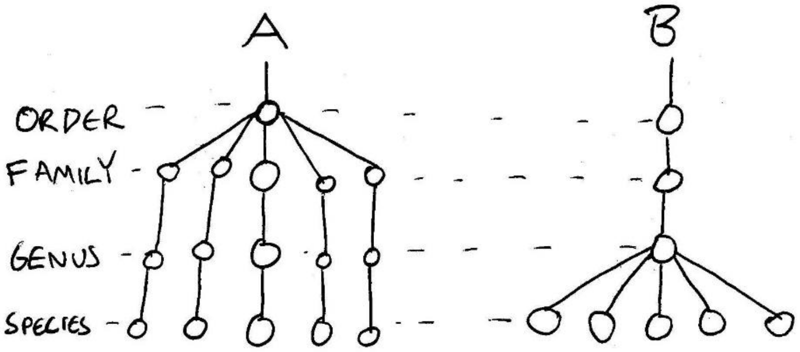

Figure 1. Demonstration of different cladistic structure under the same species richness and abundance. Assemblage A and B both have 5 species within one order, but assemblage A is intuitively more evolutionarily diverse as it includes 5 genera and 5 families compared with 1 in B.

Vane-Wright et al. (1991) recognized these shortcomings and sought to develop a measure which encoded cladistic relationships between species. This approach was extended by Faith (1992) to PD, which is applicable to any taxonomic level at which a phylogeny of features can be constructed. Faith (2017) characterized PD as a framework for quantifying distinctness at the taxonomic level in terms of lower-level units (features, which may be genetic characters). This framework is intended to reflect the future "option value" of genetic diversity for evolution and ecosystem services (Faith, 2018). Current human utilization of organisms depends on specific features, but future utility relies on the full diversity of features (Faith, 2015).

The conservationist argument for considering phylogenetic diversity may take a similarly anthropocentric view, or an ecological view. A decade before PD was formalised, the International Union for Conservation of Nature stated the importance for conserving genetic diversity for human activity as both investment and insurance (Anon., 1980), and related genetic distinctness to prioritization of both species and locations.

PD, as defined by Faith (1992), utilizes a phylogenetic tree which describes distance between taxa in terms of branch length. PD(A) for some group A within an overall taxon set T is the sum of branch lengths in the subtree connecting A (fig. 2), including the root of T (Faith, 2006).

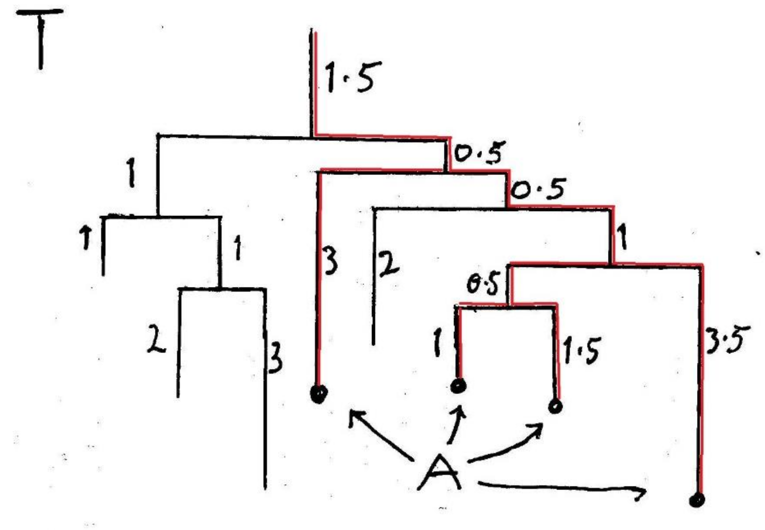

Figure 2. Phylogenetic tree indicating calculation of Faith's phylogenetic diversity PD. Branch lengths are in arbitrary units of distance such as evolutionary time. The subtree connecting A within T is outlined in red. The PD of A relative to T is 1.5 + 0.5 + 0.5 + 1 + 3.5 + 1.5 + 0.5 + 1 + 3 = 13. Note that the branch length of the root of T is included, even though it does not form part of the minimal spanning tree of A.

Faith (1992) initially considered phylogenies constructed from feature distance matrices, and in this case, PD will be lower than expected in an assemblage displaying homoplasy (shared features that do not reflect shared ancestry). Convergent evolution, such as due to strong environmental selection, obscures the evolutionary interpretation of PD.

However, the definition is typically extended to the use of phylogenetic trees incorporating models of character evolution over numerous loci (Lozupone and Knight, 2008), allowing a more rigorous evolutionary interpretation of PD as the total amount of evolution in A relative to T.

Such evolutionary measures of biodiversity are of interest for clarifying cases where taxonomic terms such as “subspecies” may refer to small or large divergences (Ryder, 1986), or where an apparently monotypic taxon may include a disputed number of diverse lineages (Hay et al., 2003). Programs such as Evolutionarily Distinct and Globally Endangered (EDGE) aim to prioritise conservation of evolutionary history (Isaac et al., 2007), since evolutionarily distinct taxa are subject to greater extinction risk than expected under a hypothesis of random extinction (Purvis et al., 2000). PD is more suitable than the EDGE method for ranking distinctness, as PD better accounts for closely related species (Kuntner et al., 2011).

DNA barcodes for phylogenetic diversity

If PD is computed from a predetermined tree, specimens must be reliably identified so that the subtree describing a given habitat is correct. This presents a practical problem shared with counting of species richness, especially when considering clades with a high diversity of hard-to-distinguish species (Cognato et al., 2020). Available taxonomic expertise, after a long decline, is insufficient to the task of describing biodiversity (Terlizzi et al., 2003).

Alternatively, phylogenetic trees for PD analysis may be constructed directly from sampled sequences. This avoids the “taxonomic bottleneck” (Kim and Byrne, 2006) and allows repeatable, empirical analysis linked to a specific sampling event. Ideal sequences for characterising genetic diversity should be homologous gene regions which are short enough to scale to large analysis yet possess a high ratio of phylogenetic signal to noise. DNA barcodes are such a form of sequence data, which have seen a rapid growth of interest over the last 18 years (DeSalle and Goldstein, 2019) and application in diverse contexts from taxonomy (Schindel and Miller, 2005) to forensic entomology (Koroiva et al., 2018) and monitoring of illegal wildlife trade (Chang et al., 2018).

The Barcode of Life Data System (BOLD) is a database and “informatics workbench” for the use of DNA barcodes (Ratnasingham and Hebert, 2007). BOLD supports the aims of the International Barcode of Life consortium (iBOL) in capturing the genetic diversity of the entire Eukaryote domain with barcodes. Some 1.3 million public records, most of which contain DNA barcodes, are available through a web portal.

The most sequenced barcode for animals is the 648 base pair 5’-3’ section of the cytochrome c oxidase I gene (COI). This mitochondrial barcode has a strong record of distinguishing species: 97.9% effectiveness in Lepidoptera (Hajibabaei et al., 2005) and 98.3% in Coleoptera (Pentinsaari et al., 2014). It is of interest to determine the suitability of COI barcodes for diversity metrics on a global scale.

A phylogenetic diversity index

Due to their increasing availability and utility for taxonomic resolution, DNA barcodes present an opportunity to generate phylogenies on an arbitrarily wide geographic scale. However, the computational hurdle of constructing accurate phylogenies from this growing dataset is considerable. The number of possible trees on N barcodes is semi-factorial in N (Dale and Moon, 1993), and in practice most phylogenetic methods are quadratic in N (Louca and Doebeli, 2018). In the interest of scaling PD to arbitrarily large barcode sets with a feasible computational load, the following approach utilizes subsampling to “rarefy” the barcodes from which the tree is calculated.

The PD index PD_N is defined as the sum of branch lengths of a phylogenetic tree constructed from N sequences randomly selected from the existing sampling of a given assemblage. This is analogous to the species index E(S_n) introduced by Hurlbert (1971) which counts the expected number of species in a sample of n individuals.

Important differences between PD_N and PD must be highlighted. PD_N is not determined with respect to a super-tree, so lacks a shared root length (compare figure 2). PD_N is sensitive to the relative abundance of types (groups at the chosen taxonomic level, e.g., species), since the branch ends are individuals, rather than unique types. PD_N is therefore sensitive to within-species diversity. As a result, PD_N is an estimator of sample PD which is downward-biased (producing a lower-than-expected diversity) for a sample where the number of types is greater than N. If the number of types in the sample is less than N, PD_N may be greater than PD due to including intra-type diversity.

Sample PD and PD_N are thus conceptually different estimates of the true PD in a population. As an abundance-sensitive measure, PD_N is related to the “effective PD” of the population (Chao et al., 2010). Effective PD is the phylogenetic equivalent of effective species number ${}^q\mathrm{D}_\mathrm{S}$ (Chao et al., 2014), the number of equally abundant species which would be as diverse as the observed assemblage. Effective PD and species numbers are families of statistics varying by a parameter q which determines their sensitivity to relative abundance.

Chao et al. (2015) developed formulae for the rarefaction and extrapolation of PD. These formulae may be used to adjust PD_N to an estimator $\mathrm{P\hat{D}}_{N+k}$ of the PD index for a different sample size, and so determine the diversity discovered by additional sampling efforts. However, since the undiscovered diversity for the assemblage cannot be bounded above, there is no estimator for the total real-world PD in terms of PD_N.

Given that observed species richness is strongly affected by sampling effort, obscuring potential spatial patterns of diversity (Colwell et al., 2004) and PD is expected to be similarly influenced (Chao et al., 2015), it is desirable for a PD index to show less influence of sampling effort. This project therefore aims to investigate the feasibility of deriving PD_N from existing barcodes and investigate its spatial properties in relation to sampling effort.

Project outline

The goals of this project are: (a) to establish whether DNA barcodes are suitable for measuring global phylogenetic diversity; (b) to compare a barcode-based index of phylogenetic diversity with the expected species diversity ${}^{q}\mathrm{D}_{\mathrm{S}}$. These aims were addressed through the development of a bioinformatics software pipeline to compute the index PD_N using DNA barcodes from BOLD, assess its sensitivity to number of available sequences, and calculate ${}^{q}\mathrm{D}_{\mathrm{S}}$ in the same locations.

Methods

Program components

An initial command-line program was developed to automate phylogenetic analysis of DNA barcodes. This was later expanded into a web-based application with the following components:

1. A script to compute PD_N for areas across the globe, able to run on a web server or local command-line,
2. A script to compute species diversity indices ${}^q\mathrm{D}_\mathrm{S}$ for each area,
3. A web page interfacing with components 1 and 2, able to display their output on a scatterplot and world map in a web browser.

All program scripts and data are publicly available on a GitHub repository (https://github.com/dermestid/bold-phylodiv-scripts).

PD calculation algorithm

Component 1 above is subdivided into the following steps:

- Division of globe into equal areas,
- Retrieval of observations (sequences and location data) from BOLD,
- Random subsampling of observations to reduce processing time,
- Building a tree for each subsample and obtaining its length.

These steps are displayed in figure 3 and detailed below.

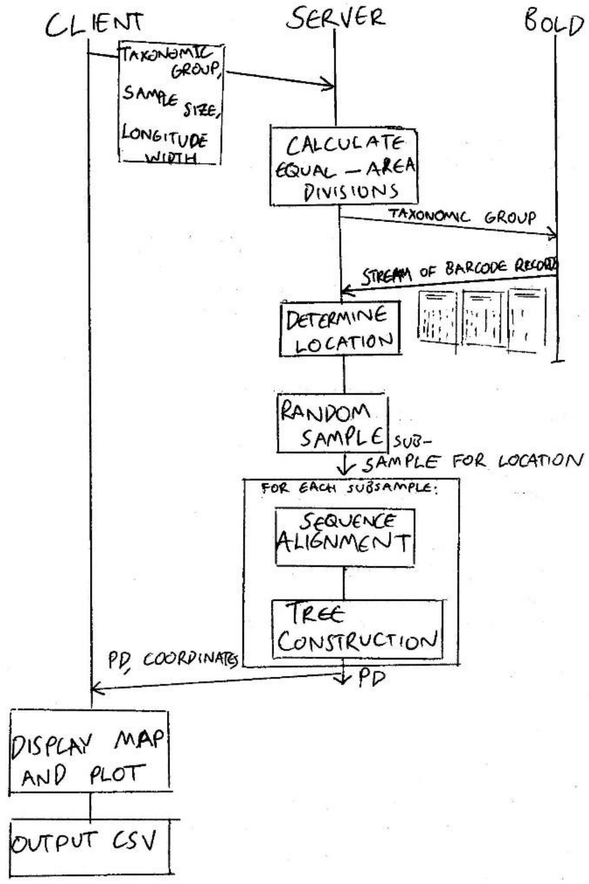

Figure 3. Schematic flowchart of component 1 of a program to compute PD. Client, Server and BOLD lines respectively indicate the processes that take place on a web browser, a web server running the PD calculation scripts, and the BOLD web portal. Not indicated here are the replication of this overall process to obtain an estimate of the error in PD.

Division of globe into equal areas

Equal area divisions are obtained using rectangles bounded by lines of latitude and longitude, where the longitudinal spacing is fixed, and the latitudinal spacing is varied. This process is the inverse of an area-preserving map of the sphere onto a rectangle (an equal-area cylindrical map projection: Miller, 1942).

The lines of latitude are obtained by $\sin^{-1}(h)$ where $h \in H$ are equally spaced divisions of the interval $[-1, 1]$. Note that the sphere divisions bounded by these lines of latitude and longitude are of equal surface area but may vary in terrestrial area (ratio of land to sea).

Retrieval of observations from BOLD

The BOLD website provides several interfaces for requesting data over HTTP. Since this API does not allow for requests bounded by latitude and longitude, the program requests all data for a given taxonomic group, then allocates data into areas while randomly sampling. Frequently, more than half of observations in BOLD are not adequately georeferenced (lacking latitude and longitude), so are discarded.

Several filtered BOLD datasets for taxonomic groups of interest are included as example data with the program. These datasets can be processed more quickly as they do not need to be retrieved from BOLD.

Random sampling

The user provides a sample size N and a random sample of barcodes in each area (which has at least N barcodes) is obtained from the stream of observations in BOLD. This sample must be fair in the sense that no barcode is more likely to be used than any other.

The algorithm used for sampling is an adaptation of “reservoir sampling” introduced by Vitter (1985). This has the advantage of minimising the storage footprint of the program: only $N \times M$ observations (where M is the number of areas available to sample) are held in memory, rather than the entire dataset.

Building a tree from DNA barcodes

Trees are built using an external program PAUP* (version 4.0a169; Swofford, 1998), executed using system calls from the PHP script. PAUP* requires input of aligned sequences, which is achieved using the external program ClustalW (version 2.0; Larkin et al., 2007).

Clustal takes an input file of barcodes in FASTA format (developed by Pearson and Lipman, 1988) and produces a multiple alignment. This process, consisting of iteratively adjusting sequences to line up homologous residues, is computationally expensive: complexity $\mathrm{O}(N^2)$ (Larkin et al., 2007).

The alignments are assembled into a single input file to PAUP*, which uses the neighbour-joining method (NJ) to assemble a single tree representing a likely evolutionary scenario. NJ is regarded as a fast alternative to other tree-building methods (Simonsen et al., 2008) although potentially lacking in accuracy compared to more sophisticated methods (Kapli et al., 2020) such as maximum-likelihood.

PAUP* produces output including a branch length matrix, from which PD_N can be immediately extracted as total tree length. Units of branch length (and so of PD) are nucleotide substitutions per site. This includes both change due to evolutionary time and the unknown rate of evolution, and the two cannot be easily separated without assuming a molecular clock (Bromham and Penny, 2003). This is not necessary for calculating diversity.

Computing species diversity

A second script (component 2) retrieves species data from the global biodiversity information facility GBIF (Telenius, 2011) for comparison to PD_N. The HTTP request to GBIF obtains a list of species and corresponding abundances for each coordinate-bounded area for which PD_N has been calculated.

Expected species diversity ${}^q\mathrm{D}_\mathrm{S}$ with q=1 is then calculated as the exponential of Shannon-Wiener entropy (Chao et al., 2010). This is chosen as the species comparison with PD_N because it incorporates relative abundance.

GBIF, as the largest initiative for open access to biodiversity data, collates data from numerous sources. The number of distribution records available through GBIF is greater than independent compilation can reasonably provide (Beck et al., 2013); however, a perceived lack of scrutiny, along with spatial bias, has prompted criticism of its use in biogeography (Beck et al., 2014). Nonetheless, as a large database growing through the input of diverse organizations and individuals (Costello et al., 2013), it is a suitable analogue to BOLD for comparing species occurrence to phylogenetic data.

Web application implementation and visualisation

Components 1 and 2 described above were implemented in the scripting language PHP (version 8.0.0, PHP Group, 2020: https://www.php.net/). Testing of this program was performed by hosting on a local Apache-based web server on Windows 10 with future web deployment on the Heroku cloud platform planned.

User input on a web form triggers an HTTP request to component 1, providing as arguments the taxonomic group, sample size, and division longitudinal width. The PHP script running on the server communicates its results (PD_N and BOLD observation count) to the client using server-sent events (SSE). Quantiles for PD_N at levels of 0.05, 0.3, 0.7 and 0.95 are calculated, separating PD_N into groups of highest and lowest 5%, middle 40%, and remaining 25% brackets. Each group is assigned a colour, which is used to indicate the PD_N of each area. These colours are plotted on a world map displayed in the web browser using JavaScript and SVG through the library D3.js (Bostock, 2015).

A request is then sent to component 2 to obtain data from GBIF. A scatterplot of ${}^{q}\mathrm{D}_{\mathrm{S}}$ against PD_N is displayed in the browser using D3.js, allowing an initial assessment of whether a relation exists between PD_N and species diversity. Quantiles for ${}^{q}\mathrm{D}_{\mathrm{S}}$ are calculated at 0.05, 0.3, 0.7 and 0.95, allowing the diversity for each area to be visualised on the map.

The browser script then sends further requests to component 1 to obtain new PD_N values. Map colours and scatterplot are adjusted to the mean of PD_N after multiple iterations, with a confidence interval of 1.96 standard errors of the mean displayed as an error bar around each point.

The overall process is iterated with different starting conditions of the equal-area division of the globe. An offset of latitude and/or longitude less than the size of one grid area is added to the coordinates of each area. These alternative iterations can be compared to determine if the choice of grid origin alters the overall distribution. This partly addresses the "modifiable areal unit problem": the possibility for statistical bias resulting from changes in zoning (Jelinski and Wu, 1996).

At any point of iteration, the user can download a data file in CSV format of the values and positions of each map area on the page, suitable for further statistical analysis.

Statistical analysis

Barcode data for phylum Arthropoda were accessed on 12th March 2021. Values for ${}^{1}\mathrm{D}_{\mathrm{S}}$ (${}^{q}\mathrm{D}_{\mathrm{S}}$ with q=1) and PD_{20} (PD_N with N = 20) for each geographical area were calculated using the web application and imported into RStudio (version 1.3.1093, RStudio Team, 2020) for statistical analysis using R (Core R Team, 2013).

Each map area was assigned a PD decile score δ_{P}, and species diversity decile score δ_{S}, defined as the ranking of the decile group (between 1 and 10) for PD_{20} and ${}^{q}\mathrm{D}_{\mathrm{S}}$ respectively. The decile difference Δ for each area was computed as ($\delta_{\mathrm{P}} - \delta_{\mathrm{S}}$). This statistic is 1 to 9 for areas with relatively high PD_{20} and relatively low ${}^{q}\mathrm{D}_{\mathrm{S}}$, -

1 to -9 for areas with relatively high $^{1}D_{S}$ and low PD_{20}, and closer to 0 the more similar are relative levels of PD_{20} and $^{1}D_{S}$.

Spatial autocorrelation

Spatial autocorrelation (SAC) was tested for the distribution of $^{1}D_{S}$, PD_{20} and Δ using Moran's I statistic. SAC can be interpreted as the degree to which values in space are dependent on nearby values. A positive result of an autocorrelation test first suggests that the values in question are non-randomly distributed and so cluster at a scale greater than the division areas. Secondly, SAC of two variables implies that inference of a relation between them is likely confounded by a third variable (De Knegt et al., 2010).

Moran's I is defined as a quotient of weighted sums of deviations across spatial units (Moran, 1950). The weighting matrix w_{ij} for I is chosen such that it describes spatial nearness or relation appropriately for the phenomenon of interest. Defining the mean point of area A as the mean of coordinates of barcodes sampled in A, and denoting the great circle distance in km between mean points of areas A_i and A_j as d_{ij}, the matrix w_{ij} was defined as

$w_{ij} = 0$ if $i = j$,

$w_{ij} = 0$ if $d_{ij} > 2500$,

$w_{ij} = 1$ if $d_{ij} < 200$,

$w_{ij} = 100 / d_{ij}$ if $2500 \geq d_{ij} \geq 200$.

The distance of 2500km is a boundary beyond which areas are deemed unrelated; this value is chosen because it is the approximate width of the largest non-polar terrestrial ecoregion as classified in the WWF Global 200 (Olson and Dinerstein, 1998).

In addition, the following standard weight matrices were computed for comparison: *k*-nearest-neighbours, which assigns weight 1/*k* to the *k* nearest points and 0 to others, with *k* = 10; and *d*-near-neighbours, which assigns weight 1/*m* to the *m* points within great circle distance *d* of a point, with *d* = 2500km.

Spatial autocorrelation is indicated if $I > E(I)$, $E(I) = -1 / (M - 1)$ where M is the number of areas. Values of I for PD_{20}, ${}^1D_S$ and Δ were tested using the function `moran.test` from the R package `spdep` v1.1-2, using a two-tailed test with the null hypothesis that observed $I = E(I)$. Results were confirmed using the permutation test function `moran.mc` with 999 permutations. The null hypotheses tested were:

- $H_{0,1}$: observed I for $PD_{20} = E(I)$
- $H_{0,2}$: observed I for ${}^1D_S = E(I)$
- $H_{0,3}$: observed I for $\Delta = E(I)$

Modelling of PD_{20}

The relationship of PD_{20} to ${}^1D_S$ was tested in a generalized linear model (GLM) framework in R to determine whether higher levels of ${}^1D_S$ are associated with higher levels of PD_{20}. In addition, the effect of sampling effort on PD_{20} and ${}^1D_S$ was analysed, where sampling effort for an area is defined as the number of observations in that area from BOLD and GBIF respectively. The following null hypotheses were outlined:

- $H_{0,4}$: there is no effect of ${}^1D_S$ on PD_{20}
- $H_{0,5}$: there is no effect of BOLD record count on PD_{20}
- $H_{0,6}$: there is no effect of GBIF observation count on ${}^1D_S$

These hypotheses were tested using a likelihood ratio test on the model including explanatory variable of $^{1}D_{S}$ or sampling effort against the null model. $^{1}D_{S}$ was modelled as Poisson-distributed and PD_{20} as normal-distributed.

Results

Summary of barcode data

The program was tested using BOLD data from taxonomic groups Anura, Aves, Mammalia, Malacostraca, Echinodermata, Chordata, and Arthropoda. Example results from phylum Arthropoda are presented below; this group is of interest due to the intensive sampling of certain arthropod taxa (particularly Lepidoptera), its high species diversity, and the relative lack of attention to insect biodiversity hotspots in conservation literature (Stork and Habel, 2014).

BOLD hosts 8,336,032 arthropod specimens with sequences (as of 16th March 2021), of which 6,556,817 are publicly available. 646,965 arthropod records were downloaded from BOLD on 12th March 2021 (number was limited by time and software failures). Of these, 553,813 (86%) were useable for program input, as they were georeferenced to latitude and longitude and free of database formatting errors. These georeferenced sequences were sourced from 221 different countries or oceanic regions.

Public arthropod sequences in BOLD represent 228,819 different species names. However, 3,673,754 records (56%) include no species name, and the “species” field of the remaining records may denote uncertain identifications (e.g., *Columba* cf. *oenas*) or identifications above species level (e.g., *Rattus sp.*).

Among used georeferenced sequences, 314,410 (57%) listed no species name, and a further 11,619 (2%) listed an incomplete or uncertain species name. 227,867 records (41% of total) were not identified to genus level.

Global distribution of PD

The 553,813 georeferenced barcode sequences were allocated to rectangles 9° of longitude wide and approximately 638,000 km^2 in area. All rectangles south of 68.96°N measured between 633,244 and 642,291 km^2 –a variation of 0.7% due to rounding error and variation in the radius of the earth. Rectangles north of 68.96°N were significantly smaller due to cutting off at the north pole, to a minimum of 212,666 km^2.

Across 9 different offsets calculated, 305 to 309 areas (38% of total surface) contained at least 20 barcodes and at least one species observation in GBIF. PD_{20} was calculated for each area, taking the mean of 5 iterations of sampling and PD_{20} calculation (figure 4). PD_{20} values over all offsets ranged between 0.346 and 2.472 substitutions per nucleotide, with a median value of 1.665. Figure 5 shows values from the globe division centred on (0, 0).

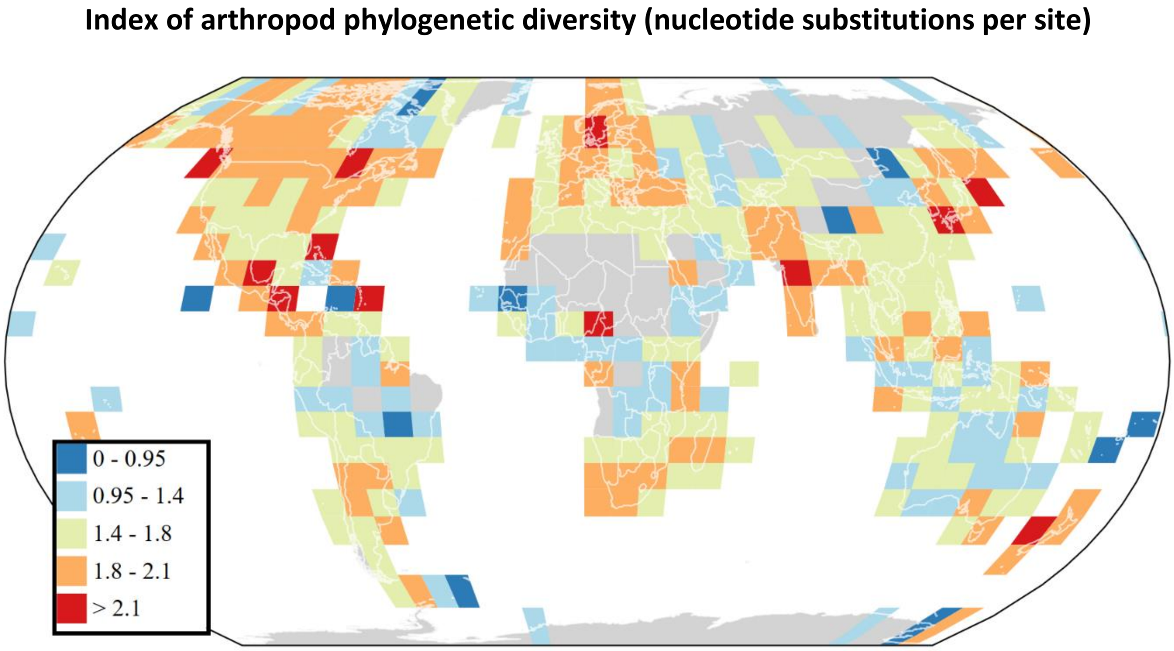
Index of arthropod phylogenetic diversity (nucleotide substitutions per site)

Figure 4. Global map of phylogenetic diversity index PD_{20} for phylum Arthropoda. 309 map areas of approx. 638,000 km^2 are coloured by PD_{20} value blue to green to red. Dark red and blue areas indicate the top and bottom 5% of PD_{10} values respectively. Light green areas represent the middle 40% of PD_{20} values. Units of PD_{20} are average nucleotide substitutions per site on branches of a phylogenetic tree on 20 random taxa.

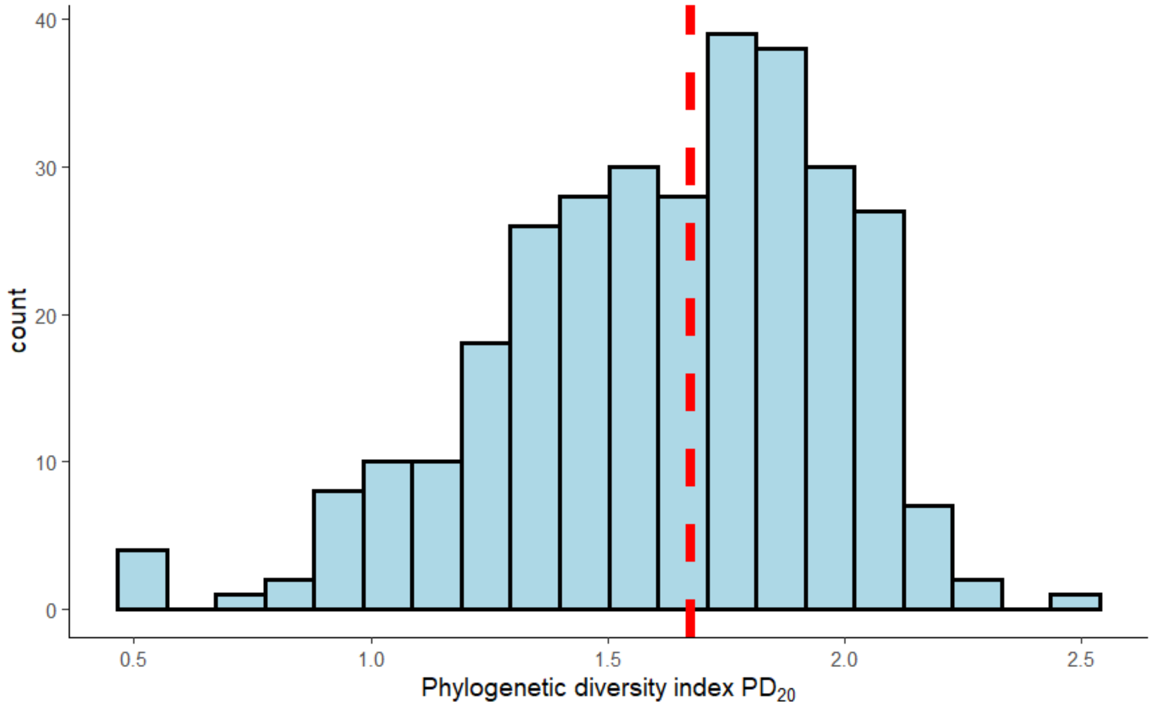

Figure 5. Histogram of phylogenetic diversity index PD_{20} for phylum Arthropoda over a single division of the globe into 309 areas of approx. 638,000 km^2. Red dashed line indicates the median value of 1.675.

Averaged across all offsets, 27% of the top 10% of areas for PD_{20} lay in the tropics with centre between ±23.43° of latitude (table 1). 25% of the top decile areas lay in subtropical latitudes with centre between 23.43° to 35°, or -35° to -23.43°.

Latitudinal region (outer boundary)	**Approximate surface area**	**Mean percentage of areas in top PD_{20} decile**	**Standard deviation of mean percentage**
Tropical (±23.43°)	2.02×10^8 km^2	26.9	6.3
Subtropical (±35°)	8.9×10^7 km^2	24.7	6.5
Temperate (±66°)	1.73×10^8 km^2	43.7	7.9
Polar	4.4×10^7 km^2	7.0	3.2

Table 1. Mean percentage of top-decile PD areas in each latitudinal region, over 9 different divisions of the globe.

Comparison of PD with species diversity

Species data was obtained from GBIF for 98.7% of areas with at least 20 barcodes (figure 6). Order-1 effective species number ${}^1D_S$ overall varied between 1.3 and 2194 species.

Over the 9 globe divisions, 15.1% of areas (s.d. 4.3%) in the top decile for ${}^1D_S$ also possessed PD_{20} values in the top decile.

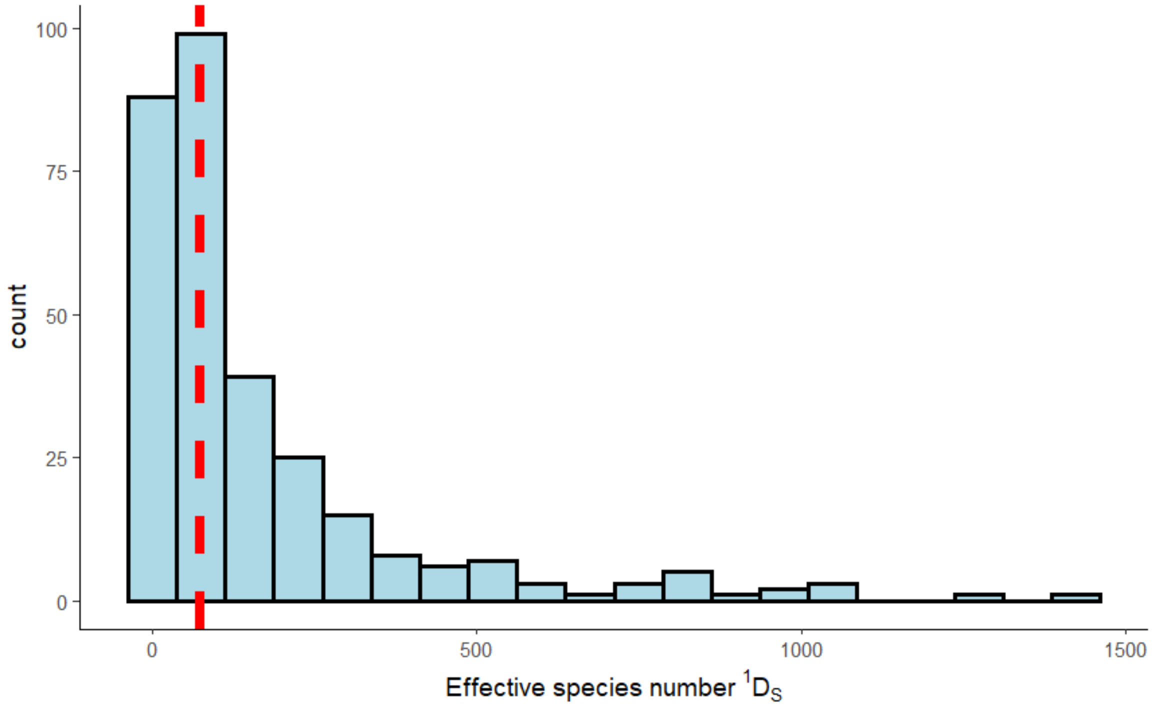

Figure 6. Histogram of arthropod effective species number ${}^1D_S$ (exponential of Shannon-Wiener entropy) observed in GBIF data, for a single division of the globe into 309 areas of approx. 638,000 km^2. Dashed red line indicates the median value of 73.42. In this division, ${}^1D_S$ varies between 3.2 and 2062 species. Two areas with ${}^1D_S$ > 1500 are excluded.

The decile difference Δ of PD_{20} and ${}^1D_S$ was calculated for each area. Overall mean and standard deviation of Δ and absolute value of Δ over 9 different globe divisions are shown in table 2. Standard deviation is uncorrected for spatial autocorrelation between different divisions.

Statistic	**Global mean**	**Standard deviation of mean**
Δ	0	0.071
Abs(Δ)	2.994	0.043

Table 2. Values of the decile difference Δ and its absolute value averaged across all areas in all divisions. Δ is defined as the difference of decile rank (1 to 10) between PD_{20} and ${}^1D_{S.}$ The global mean value of Δ is expected to be 0 since it balances out overall. Higher levels of Abs(Δ) indicate more frequent differences between phylogenetic and species diversity.

Comparison of spatial clustering of PD_{20} and species diversity

Moran's measure of spatial autocorrelation I was calculated for distribution of PD_{20} and ${}^1D_S$, over each different division. Additionally, the Moran's I for the decile difference Δ of

PD_{20} and $^{1}\mathrm{D}_{\mathrm{S}}$ was calculated for each division. Results from the division centred on (0,0) computed using inverse-distance weight matrix w_{ij} are summarised in table 3.

Statistic	**Moran's *I***	**Expected value E(*I*)**	**Variance of E(*I*)**	***p*(*I* = E(*I*))**
PD_{20}	0.1562	-0.0032	0.0005	< 0.001
$^{1}\mathrm{D}_{\mathrm{S}}$	0.6214	-0.0032	0.0005	< 0.001
Δ	0.2485	-0.0032	0.0005	< 0.001

Table 3. Observed and expected values, with corresponding p values, for Moran's *I* on three spatially distributed statistics: phylogenetic diversity index, expected species richness, and decile difference of the two. The weight matrix in all cases is an inverse-distance matrix which assigns between 0 and 1 for regions between 200 and 2500km distant. Values of *I* greater than E(*I*) indicate spatial autocorrelation.

Analysis of correlation between PD and species diversity

PD_{20} distribution relative to $^{1}\mathrm{D}_{\mathrm{S}}$ in one of 9 trials is shown in figure 7. A likelihood ratio test was performed to determine the significance of any effect of $^{1}\mathrm{D}_{\mathrm{S}}$ on PD_{20}; results are presented in table 4.

Model	Intercept	Slope	AIC	Log-likelihood	Pseudo-R^2
Null: $\mathrm{PD}_{20} \sim 1$	1.62*		229.65	-112.82	
$\mathrm{PD}_{20} \sim {}^{1}\mathrm{D}_{\mathrm{S}}$	1.58*	0.00026*	219.09	-106.55	0.040
These models are significantly different, and the more complicated one is preferred. Likelihood ratio test: χ^2 = 12.552, df = 1, *p*=0.0003958					

Table 4

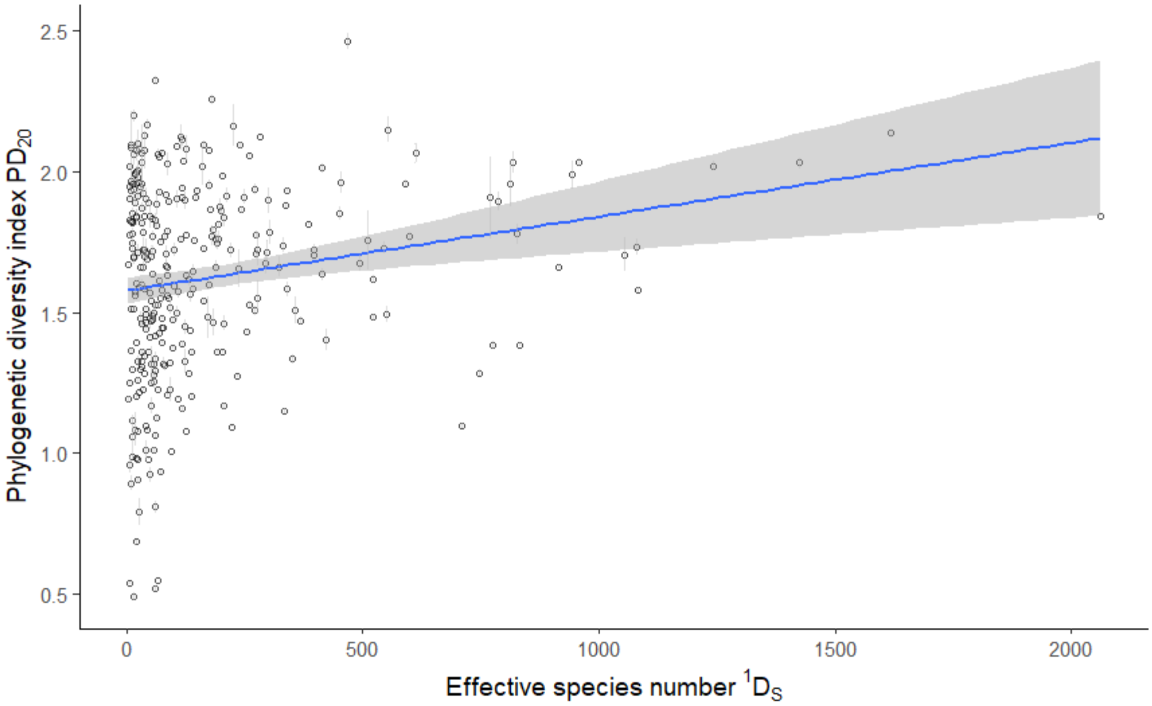

Figure 7

Effect of sampling effort on PD

PD_{20} relative to sampling effort (as measured by public sequences in BOLD) in one of 9 trials is shown in figure 8. A likelihood ratio test was performed to determine the significance of any effect of sampling effort on PD_{20}; results are presented in table 5.

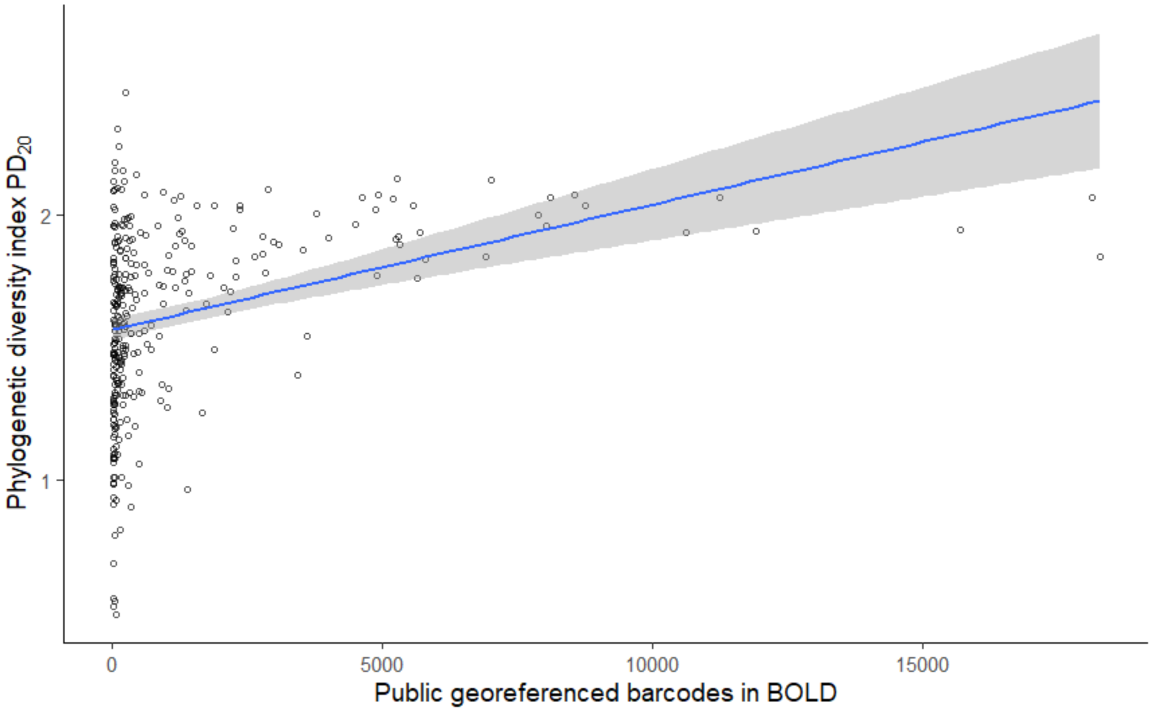

Figure 8

Model	Intercept	Slope	AIC	Log-likelihood	Pseudo-R^2
Null: $PD_{20} \sim 1$	1.62*		229.65	-112.82	
$PD_{20} \sim$ BOLD count	1.60*	0.0000109*	216.21	-105.10	0.049
These models are significantly different, and the more complicated one is preferred. Likelihood ratio test: χ^2 = 15.439, df = 1, *p*=0.00008523					

Table 5

Effect of sampling effort on species diversity

A likelihood ratio test was performed to determine the effect of sampling effort (defined as number of occurrences in GBIF) on ${}^1D_S$; results are presented in table 4.

Model	Intercept	Slope	AIC	Log-likelihood	Pseudo-R^2
Null: ${}^1D_S \sim 1$	5.16*		80059	-40028	
${}^1D_S \sim$ GBIF count	5.09*	8.93×10^{-8}*	72196	-36096	0.10
These models are significantly different, and the more complicated one is preferred. Likelihood ratio test: χ^2 = 7865.68 df = 1, $p < 10^{-10}$ However, residual deviance is more than twice the degrees of freedom: overdispersion.					

Table 6

Discussion

Interpretation of results

Availability of barcode data

The rate of georeferencing in arthropod data allowed a high proportion (86%) of downloaded barcodes to be used; this compares favourably with chordate data, where only 44% of downloaded barcodes were georeferenced.

By contrast, there is a huge data gap in identification: specimens with unknown genus or species made up 41% and 56-59% of available data. The potential utility of barcodes in quantifying diversity of these "dark taxa" is a major motivation for their study.

Distribution of PD_{20}

The maps of arthropod PD_{20} (figure 4) did not show the pattern that is expected of diversity in a high-level taxon. In particular, the general gradient of increasing biodiversity towards low latitudes (Kinlock et al., 2017) was not demonstrated here, and was almost reversed (table 1). The most speciose class within Arthropoda, insects, show a history of diversification linked to evolution in vascular plants (Jermy, 1984; Stork and Habel, 2013); plant species diversity is distinctly higher in tropical regions (Barthlott et al., 2007), in contrast to observed PD_{20} which was disproportionately higher in temperate regions. Therefore, it is unlikely that this observed PD_{20} demonstrates some real biogeographic pattern specific to Arthropoda.

Results of Moran's I test (table 3) indicated that PD_{20} was nonrandomly distributed and positively spatially autocorrelated ($H_{0,1}$ should be rejected), and thus clustered. However, PD_{20} did not cluster solely in expected biodiversity hotspots (Myers et al., 2000). The countries in which top-5% values of PD_{20} consistently appeared are mostly developed economies with sequencing labs which generate a significant proportion of the barcodes in BOLD. Within Europe, these are more often northern and central European countries than Mediterranean (Gaytán et al., 2020).

At a sub-national level, observed PD_{20} further contradicted expected biodiversity distribution. In 5/9 divisions, hotspots within the USA were indicated in northern states but not in Hawaii, California, or Florida, which are regarded as biodiversity hotspots (Soltis and Soltis, 2016).

Comparison of PD_{20} with effective species number

A high value of Moran's I calculated for ${}^{1}D_{S}$ (table 3) may indicate that species diversity is clustered in fewer hotspots than PD. The mean of $\mathrm{Abs}(\Delta)$ and I for Δ suggest that differences of PD_{20} and ${}^{1}D_{S}$ are heterogeneously clumped. Null hypotheses $H_{0,2}$ and $H_{0,3}$ should be rejected. The linear model (table 4) shows support for rejecting $H_{0,4}$ and affirming a significant correlation between ${}^{1}D_{S}$ and PD_{20}. However, the wide spread of PD_{20} values (figure 7) and the low pseudo-R^2 value in this model make this a somewhat uncertain conclusion. Since these variables are both spatially autocorrelated, and thus technically violating assumptions of independence of points, there is insufficient evidence to decide whether these two measures of diversity are correlated.

Relation of PD_{20} and species diversity to sampling effort

The model summarised in table 5 should be preferred over the null model by likelihood ratio test, so $H_{0,5}$ should be rejected. This provides a clear explanation for apparently counterfactual distribution of PD_{20}: the measure is strongly affected by sampling effort. This clearly justifies studies of PD rarefaction (Chao et al., 2015) because uncorrected subsampling measures reflect only the fraction of biodiversity that has been well sampled.

Accuracy and feasibility of barcode-based diversity indices

Accuracy of the index PD_N

Isaac et al (2007) stated that a valid biodiversity measure must possess robustness to uncertainty and must capture some aspect of biodiversity. The foregoing results suggest that PD_{20} does not satisfy the latter condition, as it does not rank regions by any useful aspect of biodiversity. A full assessment of the performance of PD_N should determine whether the taxonomic skew in sample availability plays a role in biasing PD_N towards more sampled regions.

Practical PD indices should satisfy some quality of complementarity such that they capture the endemism of distinct taxa (Faith, 2010). This is not the case for PD_N as defined here because each area is considered independent of the global scale. It is therefore possible that a relatively ubiquitous but evolutionarily distinct taxon may skew the PD of several regions, despite not being at threat; or, a region with a high PD may consist of taxa well represented individually in other regions. This limits the usefulness of the index for conservation prioritisation. It is best considered as a “concentration” of phylogenetic diversity in each region.

Computational performance of the web application

A rigorous assessment of the methods and their implementation in the program described above is beyond the scope of this report. In addition, the empirical speed of the tool in online use is not yet known, as the program has only been tested on a personal computer, rather than the target of a dedicated web server.

However, informal observation and testing revealed several bottlenecks which limit use of this application for rapid diversity assessment.

The BOLD API for retrieving sequences operated at rates less than 100 kilobytes/s, an unacceptably slow transfer considering that larger taxonomic groups may present hundreds of megabytes of data. This cannot be remedied by improvements to any part of the web application, so was avoided during testing by downloading the required data from BOLD in advance of their use. For assessing biodiversity in arbitrary taxonomic groups, the speed of access to barcode data must be considerably improved.

Sequence alignment was the slowest phase of processing in tests. The task of aligning N sequences for each of hundreds of regions in multiple iterations was considerably parallelised by generating system calls to run separate ClustalW processes. On a personal computer with limited multiprocessing capabilities, this may result in reduced efficiency over more sequentially organised processing, due to the impact of context-switching between different tasks.

In future development, sequence alignment may be accelerated by offloading the task to a dedicated multithreaded alignment system, as is available at https://www.ebi.ac.uk/Tools/msa/clustalo/ using the modern alignment program Clustal Omega (Sievers and Higgins, 2018). Alternatively, the process of alignment may be avoided by using alignment-free methods of building sequence distance matrices such as comparison of *k*-mer frequency (reviewed in Zielezinski et al., 2017).

Utility of barcode-based PD for biodiversity assessment

Owen et al. (2019) observe that the value of a PD approach does not depend on its ability to predict functional diversity, although this is a potential correlate (Delgado-Baquerizo et al., 2016). PD seeks to quantify and promote conservation of evolutionary material: historical evolution, and potential material for future (Faith et al., 2010).

Conclusions

The results displayed above demonstrate flaws in the use of the index PD_{20} in combination with BOLD barcode data. While the robustness of PD_N has not been assessed here, it can be concluded that a correction is necessary to mitigate sampling effort. However, the methodology of the web application described above has potential for producing quick assessments of biodiversity differences, once adjusted. To the author’s knowledge, the only similar software tool seeking to map phylogenetic diversity is Biodiverse (Laffan et al., 2010), which has not yet integrated with the potential of DNA barcoding. In addition, it is to be hoped that increasing throughput of barcoding technology (Yang et al., 2020) and future citizen science projects (Chiovitti et al., 2019) will help to address the vast amount of yet undiscovered biodiversity.

References

Anon, 1980. World conservation strategy. Living resource conservation for sustainable development. *World conservation strategy. Living resource conservation for sustainable development.***,** 72 pp.-72 pp.

Barthlott, W., Hostert, A., Kier, G., Koper, W., Kreft, H., Mutke, J., Rafiqpoor, M. D. & Sommer, J. H. 2007. Geographic patterns of vascular plant diversity at continental to global scales. *Erdkunde,* 61**,** 305-315.

Beck, J., Boller, M., Erhardt, A. & Schwanghart, W. 2014. Spatial bias in the GBIF database and its effect on modeling species' geographic distributions. *Ecological Informatics,* 19**,** 10-15.

Bostock, M., 2015. Data-driven documents. 2015. *URl: http://d3js.org*.

Bromham, L. & Penny, D. 2003. The modern molecular clock. *Nature Reviews Genetics,* 4**,** 216-224.

Ceballos, G., Ehrlich, P. R., Barnosky, A. D., Garcia, A., Pringle, R. M. & Palmer, T. M. 2015. Accelerated modern human-induced species losses: Entering the sixth mass extinction. *Science Advances,* 1**,** 5.

Chang, C.-H., Dai, W.-Y., Chen, T.-Y., Lee, A.-H., Hou, H.-Y., Liu, S.-H. & Jang-Liaw, N.-H. 2018. DNA barcoding reveals CITES-listed species among Taiwanese government-seized chelonian specimens. *Genome,* 61**,** 615-624.

Chao, A., Chiu, C. H., Hsieh, T. C., Davis, T., Nipperess, D. A. & Faith, D. P. 2015. Rarefaction and extrapolation of phylogenetic diversity. *Methods in Ecology and Evolution,* 6**,** 380-388.

Chao, A., Chiu, C. H. & Jost, L. 2010. Phylogenetic diversity measures based on Hill numbers. *Philosophical Transactions of the Royal Society B-Biological Sciences,* 365**,** 3599-3609.

Chao, A. N., Chiu, C. H. & Jost, L. 2014. Unifying Species Diversity, Phylogenetic Diversity, Functional Diversity, and Related Similarity and Differentiation Measures Through Hill Numbers. *In:* Futuyma, D. J. (ed.) *Annual Review of Ecology, Evolution, and Systematics, Vol 45.* Palo Alto: Annual Reviews.

Chiovitti, A., Thorpe, F., Gorman, C., Cuxson, J. L., Robevska, G., Szwed, C., Duncan, J. C., Vanyai, H. K., Cross, J., Siemering, K. R. & Sumner, J. 2019. A citizen science model for implementing statewide educational DNA barcoding. *Plos One,* 14.

Cognato, A. I., Sari, G., Smith, S. M., Beaver, R. A., Li, Y., Hulcr, J., Jordal, B. H., Kajimura, H., Lin, C. S., Pham, T. H., Singh, S. & Sittichaya, W. 2020. The Essential Role of Taxonomic Expertise in the Creation of DNA Databases for the Identification and Delimitation of Southeast Asian Ambrosia Beetle Species (Curculionidae: Scolytinae: Xyleborini). *Frontiers in Ecology and Evolution,* 8**,** 17.

Colwell, R. K., Mao, C. X. & Chang, J. 2004. Interpolating, extrapolating, and comparing incidence-based species accumulation curves. *Ecology,* 85**,** 2717-2727.

Core R Team, 2013. R: A language and environment for statistical computing.

Costello, M. J., Michener, W. K., Gahegan, M., Zhang, Z.-Q. & Bourne, P. E. 2013. Biodiversity data should be published, cited, and peer reviewed. *Trends in Ecology & Evolution,* 28**,** 454-461.

Dale, M.R.T. and Moon, J.W., 1993. The permuted analogues of three Catalan sets. *Journal of statistical planning and inference*, *34*(1), pp.75-87.

De Knegt, H. J., Van Langevelde, F., Coughenour, M. B., Skidmore, A. K., De Boer, W. F., Heitkonig, I. M. A., Knox, N. M., Slotow, R., Van Der Waal, C. & Prins, H. H. T. 2010. Spatial autocorrelation and the scaling of species-environment relationships. *Ecology,* 91**,** 2455-2465.

Delgado-Baquerizo, M., Maestre, F. T., Reich, P. B., Jeffries, T. C., Gaitan, J. J., Encinar, D., Berdugo, M., Campbell, C. D. & Singh, B. K. 2016. Microbial diversity drives multifunctionality in terrestrial ecosystems. *Nature Communications,* 7**,** 8.

Desalle, R. & Goldstein, P. 2019. Review and Interpretation of Trends in DNA Barcoding. *Frontiers in Ecology and Evolution,* 7**,** 11.

Djurhuus, A., Closek, C. J., Kelly, R. P., Pitz, K. J., Michisaki, R. P., Starks, H. A., Walz, K. R., Andruszkiewicz, E. A., Olesin, E., Hubbard, K., Montes, E., Otis, D., Muller-Karger, F. E., Chavez, F. P., Boehm, A. B. & Breitbart, M. 2020. Environmental DNA reveals seasonal shifts and potential interactions in a marine community. *Nature Communications,* 11**,** 9.

Faith, D. P. 1992. CONSERVATION EVALUATION AND PHYLOGENETIC DIVERSITY. *Biological Conservation,* 61**,** 1-10.

Faith, D. P. 2006. The Role of the Phylogenetic Diversity Measure, PD, in Bio-informatics: Getting the Definition Right. *Evolutionary Bioinformatics,* 2**,** 277-283.

Faith, D. P. 2015. Phylogenetic diversity, functional trait diversity and extinction: avoiding tipping points and worst-case losses. *Philosophical Transactions of the Royal Society B-Biological Sciences,* 370.

Faith, D. P. 2017. A GENERAL MODEL FOR BIODIVERSITY AND ITS VALUE. *Routledge Handbook of Philosophy of Biodiversity***,** 69-85.

Faith, D. P., Magallon, S., Hendry, A. P., Conti, E., Yahara, T. & Donoghue, M. J. 2010. Evosystem services: an evolutionary perspective on the links between biodiversity and human well-being. *Current Opinion in Environmental Sustainability,* 2**,** 66-74.

Faith, D.P., 2018. Phylogenetic diversity and conservation evaluation: perspectives on multiple values, indices, and scales of application. In *Phylogenetic diversity* (pp. 1-26). Springer, Cham.

Gaytan, A., Bergsten, J., Canelo, T., Perez-Izquierdo, C., Santoro, M. & Bonal, R. 2020. DNA Barcoding and geographical scale effect: The problems of undersampling genetic diversity hotspots. *Ecology and Evolution,* 10**,** 10754-10772.

Hajibabaei, M., Janzen, D. H., Burns, J. M., Hallwachs, W. & Hebert, P. D. N. 2006. DNA barcodes distinguish species of tropical Lepidoptera. *Proceedings of the National Academy of Sciences of the United States of America,* 103**,** 968-971.

Hay, J. M., Daugherty, C. H., Cree, A. & Maxson, L. R. 2003. Low genetic divergence obscures phylogeny among populations of Sphenodon, remnant of an ancient reptile lineage. *Molecular Phylogenetics and Evolution,* 29**,** 1-19.

Hurlbert, S. H. 1971. NONCONCEPT OF SPECIES DIVERSITY - CRITIQUE AND ALTERNATIVE PARAMETERS. *Ecology,* 52**,** 577-+.

Isaac, N. J. B., Turvey, S. T., Collen, B., Waterman, C. & Baillie, J. E. M. 2007. Mammals on the EDGE: Conservation Priorities Based on Threat and Phylogeny. *Plos One,* 2**,** 7.

Jelinski, D. E. & Wu, J. G. 1996. The modifiable areal unit problem and implications for landscape ecology. *Landscape Ecology,* 11**,** 129-140.

Jermy, T. 1984. EVOLUTION OF INSECT HOST PLANT RELATIONSHIPS. *American Naturalist,* 124**,** 609-630.

Kapli, P., Yang, Z. H. & Telford, M. J. 2020. Phylogenetic tree building in the genomic age. *Nature Reviews Genetics,* 21**,** 428-444.

Kim, K. C. & Byrne, L. B. 2006. Biodiversity loss and the taxonomic bottleneck: emerging biodiversity science. *Ecological Research,* 21**,** 794-810.

Kinlock, N. L., Prowant, L., Herstoff, E. M., Foley, C. M., Akin-Fajiye, M., Bender, N., Umarani, M., Ryu, H. Y., Sen, B. & Gurevitch, J. 2018. Explaining global variation in the latitudinal diversity gradient: Meta-analysis confirms known patterns and uncovers new ones. *Global Ecology and Biogeography,* 27**,** 125-141.

Koroiva, R., De Souza, M. S., Roque, F. D. O. & Pepinelli, M. 2018. DNA Barcodes for Forensically Important Fly Species in Brazil. *Journal of Medical Entomology,* 55**,** 1055-1061.

Kuntner, M., May-Collado, L. J. & Agnarsson, I. 2011. Phylogeny and conservation priorities of afrotherian mammals (Afrotheria, Mammalia). *Zoologica Scripta,* 40**,** 1-15.

Laffan, S. W., Lubarsky, E. & Rosauer, D. F. 2010. Biodiverse, a tool for the spatial analysis of biological and related diversity. *Ecography,* 33**,** 643-647.

Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., Mcgettigan, P. A., Mcwilliam, H., Valentin, F., Wallace, I. M., Wilm, A., Lopez, R., Thompson, J. D., Gibson, T. J. & Higgins, D. G. 2007. Clustal W and clustal X version 2.0. *Bioinformatics,* 23**,** 2947-2948.

Lean, C. & Maclaurin, J. 2016. The Value of Phylogenetic Diversity. *Biodiversity Conservation and Phylogenetic Systematics: Preserving Our Evolutionary Heritage in an Extinction Crisis,* 14**,** 19-37.

Louca, S. & Doebeli, M. 2018. Efficient comparative phylogenetics on large trees. *Bioinformatics,* 34**,** 1053-1055.

Lozupone, C. A. & Knight, R. 2008. Species divergence and the measurement of microbial diversity. *Fems Microbiology Reviews,* 32**,** 557-578.

Maclaurin, J. & Sterelny, K. 2008. What is biodiversity? *What is biodiversity?.* i-xii, 1-217.

Miller, O. M. 1942. NOTES ON CYLINDRICAL WORLD MAP PROJECTIONS. *Geographical Review,* 32**,** 424-430.

Moran, P. A. P. 1950. NOTES ON CONTINUOUS STOCHASTIC PHENOMENA. *Biometrika,* 37**,** 17-23.

Myers, N., Mittermeier, R. A., Mittermeier, C. G., Da Fonseca, G. A. B. & Kent, J. 2000. Biodiversity hotspots for conservation priorities. *Nature,* 403**,** 853-858.

Olson, D. M. & Dinerstein, E. 1998. The global 200: A representation approach to conserving the Earth's most biologically valuable ecoregions. *Conservation Biology,* 12**,** 502-515.

Owen, N. R., Gumbs, R., Gray, C. L. & Faith, D. P. 2019. Global conservation of phylogenetic diversity captures more than just functional diversity. *Nature Communications,* 10**,** 3.

Pearson, W. R. & Lipman, D. J. 1988. IMPROVED TOOLS FOR BIOLOGICAL SEQUENCE COMPARISON. *Proceedings of the National Academy of Sciences of the United States of America,* 85**,** 2444-2448.

Pentinsaari, M., Hebert, P. D. N. & Mutanen, M. 2014. Barcoding Beetles: A Regional Survey of 1872 Species Reveals High Identification Success and Unusually Deep Interspecific Divergences. *Plos One,* 9**,** 8.

Purvis, A. & Hector, A. 2000. Getting the measure of biodiversity. *Nature,* 405**,** 212-219.

Ratnasingham, S. & Hebert, P. D. N. 2007. BOLD: The Barcode of Life Data System (www.barcodinglife.org). *Molecular Ecology Notes,* 7**,** 355-364.

RStudio Team (2020). RStudio: Integrated Development Environment for R. RStudio, PBC, Boston, MA URL: http://www.rstudio.com/.

Ryder, O. A. 1986. SPECIES CONSERVATION AND SYSTEMATICS - THE DILEMMA OF SUBSPECIES. *Trends in Ecology & Evolution,* 1**,** 9-10.

Schindel, D. E. & Miller, S. E. 2005. DNA barcoding a useful tool for taxonomists. *Nature,* 435**,** 17-17.

Sievers, F. & Higgins, D. G. 2018. Clustal Omega for making accurate alignments of many protein sequences. *Protein Science,* 27**,** 135-145.

Silva, C. H. L., Pessoa, A. C. M., Carvalho, N. S., Reis, J. B. C., Anderson, L. O. & Aragao, L. 2021. The Brazilian Amazon deforestation rate in 2020 is the greatest of the decade. *Nature Ecology & Evolution,* 5**,** 144-145.

Simonsen, M., Mailund, T. & Pedersen, C. N. S. 2008. Rapid Neighbour-Joining. *Algorithms in Bioinformatics, Wabi 2008,* 5251**,** 113-122.

Soltis, D. E. & Soltis, P. S. 2016. Mobilizing and integrating big data in studies of spatial and phylogenetic patterns of biodiversity. *Plant Diversity,* 38**,** 264-270.

Stork, N. E. & Habel, J. C. 2014. Can biodiversity hotspots protect more than tropical forest plants and vertebrates? *Journal of Biogeography,* 41**,** 421-428.

Taubert, F., Fischer, R., Groeneveld, J., Lehmann, S., Muller, M. S., Rodig, E., Wiegand, T. & Huth, A. 2018. Global patterns of tropical forest fragmentation. *Nature,* 554**,** 519-+.

Telenius, A. 2011. Biodiversity information goes public: GBIF at your service. *Nordic Journal of Botany,* 29**,** 378-381.

Terlizzi, A., Bevilacqua, S., Fraschetti, S. & Boero, F. 2003. Taxonomic sufficiency and the increasing insufficiency of taxonomic expertise. *Marine Pollution Bulletin,* 46**,** 556-561.

Van Dorp, L., Acman, M., Richard, D., Shaw, L. P., Ford, C. E., Ormond, L., Owen, C. J., Pang, J., Tan, C. C. S., Boshier, F. A. T., Ortiz, A. T. & Balloux, F. 2020. Emergence of genomic diversity and recurrent mutations in SARS-CoV-2. *Infection Genetics and Evolution,* 83**,** 9.

Vane-Wright, R. I., Humphries, C. J. & Williams, P. H. 1991. WHAT TO PROTECT - SYSTEMATICS AND THE AGONY OF CHOICE. *Biological Conservation,* 55**,** 235-254.

Vitter, J. S. 1985. RANDOM SAMPLING WITH A RESERVOIR. *Acm Transactions on Mathematical Software,* 11**,** 37-57.

Weitzman, M. L. 1998. The Noah's Ark Problem. *Econometrica,* 66**,** 1279-1298.

Whittaker, R. H. 1972. EVOLUTION AND MEASUREMENT OF SPECIES DIVERSITY. *Taxon,* 21**,** 213-251.

Yang, C. T., Zheng, Y. X., Tan, S. J., Meng, G. L., Rao, W., Yang, C. Q., Bourne, D. G., O'brien, P. A., Xu, J. Q., Liao, S., Chen, A., Chen, X. W., Jia, X. R., Zhang, A. B. & Liu, S. L. 2020. Efficient COI barcoding using high throughput single-end 400bp sequencing. *Bmc Genomics,* 21**,** 10.

Zielezinski, A., Vinga, S., Almeida, J. & Karlowski, W. M. 2017. Alignment-free sequence comparison: benefits, applications, and tools. *Genome Biology,* 18**,** 17.