Title: The impacts of humans on animal genetic diversity worldwide

Authors: Katie L. Millette[1,5,6,*], Vincent Fugère[1,2,5,6], Chloé Debyser[1,5,6], Ariel Greiner[3], Frédéric J. J. Chain[4] and Andrew Gonzalez[1,5].

[1]McGill University, Department of Biology
[2]Université du Québec à Montréal, Département des Sciences Biologiques
[3]University of Toronto, Department of Ecology & Evolutionary Biology
[4]University of Massachusetts at Lowell, Department of Biological Sciences
[5]Quebec Centre for Biodiversity Science
[6]These authors contributed equally.

Author emails: katie.millette@mail.mcgill.ca; vincent.fugere@mail.mcgill.ca; chloe.debyser@mail.mcgill.ca; ariel.greiner@mail.utoronto.ca; frederic_chain@uml.edu; andrew.gonzalez@mcgill.ca

Running title: Human impacts on genetic diversity

Keywords: intraspecific genetic diversity; land use; spatial scale; cytochrome *c* oxidase subunit I (COI); Anthropocene; nucleotide diversity

Type of article: Letters

Words in abstract (151); main text (6,025); number of references (71); number of figures (4); number of tables (1); number of text boxes (0).

***Corresponding author:** Katie Millette, 1205 Docteur Penfield, Department of Biology, McGill University, Montreal QC, H3A 1B1, Canada; phone 1-514-398-1622; fax 1-514-398-5069.

Statement of authorship: All authors contributed to the idea and design of the project. KLM and FJJC contributed to sequence data collection and analysis. VF led statistical analyses. CD led spatial analyses. AGr collected taxonomic information. KLM drafted the manuscript. VF and CD made the figures. AGo supervised the project. All authors contributed to data interpretation and writing of the manuscript.

Data accessibility statement: All data used in this manuscript are available online in open access, as described in the Methods section. Formatted datasets used in statistical analyses will

be archived on an online repository upon manuscript acceptance. Data manipulation and analysis code can be found at https://github.com/VFugere/GenDivProject.

Abstract

Human impacts on genetic diversity are poorly understood yet critical to understanding the evolutionary capacity of the world's biodiversity. We combined global estimates of land use and human density with 175,247 mitochondrial sequences to assess human impacts on the intraspecific genetic diversity of 17,082 species of birds, fishes, insects, and mammals. We determined drivers of diversity across geographical space, over time, and at four spatial scales. In both spatial and time series analyses, the magnitude and sign of human impacts depended on taxon, and effects were generally small. Weak latitudinal diversity gradients were also found in birds, fishes, and mammals. Geographic distance was a strong driver of genetic diversity among sequences at large spatial scales. Our analyses provide a first worldwide picture of human impacts on animal genetic diversity and reveals critical gaps in taxonomic, geographic, and temporal data coverage, and calls for increased global efforts to systematically monitor genetic diversity.

Introduction

Intraspecific genetic diversity, a measure of the genetic variation within populations, is a fundamental dimension of biodiversity. Intraspecific genetic diversity is a reflection of both past and current evolutionary bottlenecks, as well as an indicator of a population's potential for adaptation to future stressors (Hewitt 2000; Reed & Frankham 2003; Frankham 2005; Bijlsma & Loeschcke 2012). Understanding the drivers of genetic diversity change worldwide, across taxonomic groups, is of great interest to ecologists and conservation biologists (Hughes *et al.* 2008; Pereira *et al.* 2013; Mimura *et al.* 2017; Paz-Vinas Ivan *et al.* 2018). Humans are now acting as an evolutionary force, modifying rates of extinction and colonization, but also altering the intraspecific genetic diversity of plants and animals around the world (Palumbi 2001; Alberti 2015; Thomas 2015; Schlaepfer *et al.* 2018). To date, no global assessment of temporal trends in genetic diversity has been conducted, nor have human impacts on such trends been quantified.

Theory predicts that human activities can affect intraspecific genetic diversity via demographic and evolutionary mechanisms (Kimura 1968, 1983; King & Jukes 1969). Depending on how human disturbances alter selection, drift, gene flow, and mutation rates, intraspecific genetic diversity may decrease, increase, or remain unchanged over time (DiBattista 2008). For example, disturbances like habitat fragmentation and excess harvesting can reduce diversity due to sustained selection, decreased gene flow linked to population isolation, and chronic inbreeding associated with reduced population sizes (Banks *et al.* 2013). Alternatively, human disturbances can maintain or increase genetic diversity through time, for example by magnifying temporal variation in selection, increasing mutation rates (e.g. mutagenic pollutants), or creating environments which favour hybridization and heterozygote advantage (Dubrova *et al.* 1996; Ellegren *et al.* 1997; Bickham *et al.* 2000; Crispo *et al.* 2011). Over time and across geographic space, these outcomes can accumulate within populations such that intraspecific genetic diversity reflects a complex combination of past and present evolutionary processes that we are only beginning to investigate at the global scale.

Trends in intraspecific genetic diversity are expected to be scale-dependent as are trends in other dimensions of biodiversity (McGill *et al.* 2015; Jarzyna & Jetz 2018; Schlaepfer *et al.* 2018). Human disturbances occur at different scales (e.g., habitat loss and fragmentation) but the impacts on genetic diversity should be strongest at the scales that impact the size and fitness of individual populations and migration among them within metapopulations. To detect those

impacts it is imperative that assessments are conducted and repeated at local scales to capture population trends. Since rates of population extirpation exceed those of species extinction (Hughes *et al.* 1997), assessments at the population level are also most appropriate for biodiversity management strategies (Ceballos *et al.* 2017; Lawrence *et al.* 2019). In terms of genetic diversity, assessments conducted at spatial scales larger than the geographic extent of interbreeding populations will inflate diversity values, particularly when individuals from distant and genetically divergent groups are aggregated. The sampling of individuals in this manner may confound our ability to understand genetic diversity patterns and their relationships with anthropogenic drivers.

A recent global assessment of mammal and amphibian genetic diversity, analyzing the large number of mitochondrial sequences deposited in public repository, found evidence of reduced genetic diversity in human-impacted regions (Miraldo *et al.* 2016). However, this analysis focused on one broad spatial scale of analysis unlikely to reflect population-level processes and human impacts, and used a simple categorical classification of human impacts based on anthropogenic biomes ('anthromes'; Ellis & Ramankutty 2008). The study also failed to account for the decay of genetic similarity with increasing distance between sequences, a potentially strong driver of genetic diversity (Gratton *et al.* 2017) which may itself be affected by human disturbances (e.g., via changes in land use heterogeneity). Finally, in the analysis by Miraldo *et al.* (2016), sequences were grouped together irrespective of collection year, which could misrepresent patterns if important temporal trends in genetic diversity exist within populations, or if land use in sampled areas differed between the year of sequence collection and the year for which the global anthrome map (Ellis & Ramankutty 2008) was generated.

Here, we also take advantage of the large number of mitochondrial sequences in public repositories to conduct the first large-scale assessment of human impacts on animal intraspecific genetic diversity, taking into account spatial autocorrelation, temporal trends, and the scale-dependence of results. We overlay time-referenced mitochondrial cytochrome *c* oxidase subunit I (COI) sequence data (Moritz *et al.* 1987) for four animal classes (birds, inland and coastal bony fishes, insects, and mammals) on year-specific, worldwide, and continuous estimates of human density and land use (the raw estimates on which anthrome classifications are based). We calculated the mean pairwise dissimilarity among all sequences from a population to estimate nucleotide diversity ($\hat{\pi}$), a measure of population genetic diversity, and examine its distribution

in time, geographical space, and across major human impact gradients. We used spatial and time series analyses to test the following three predictions: 1) genetic diversity is lower in populations exposed to intensive land use and high human population density, and this relationship is strongest at small spatial scales (i.e., in populations that only include sequences in close geographic proximity); 2) genetic diversity is declining over time, but only consistently in human-impacted populations, otherwise we expect no systematic trend; and 3) based on previous findings (Miraldo *et al.* 2016; Gratton *et al.* 2017), we also expect a latitudinal gradient in genetic diversity, at least in mammals.

Material and methods

Sequence and human impact data

R version 3.5.0 (R Core Team 2018) was used for all data manipulations and analyses described below. To determine human impacts on intraspecific genetic diversity, we combined a large dataset of georeferenced animal genetic sequences with global estimates of land use and human population density. Mitochondrial cytochrome *c* oxidase subunit I (mtDNA COI) sequences for birds (Aves), fishes (Actinopterygii), insects (Insecta), and mammals (Mammalia) were downloaded from the National Center for Biotechnology Information (NCBI) 'GenBank' (Benson *et al.* 2013) and from the 'Barcode of Life Data Systems'(BOLD; Ratnasingham & Hebert 2007) in April 2017. Sequences from GenBank were retrieved with the Entrez Utilities Unix Command Line, while for BOLD we used the application platform interface. Only sequences with documented geographic coordinates and sampling dates available in the databases were downloaded. Species and year-specific sequence alignments were then performed using default parameters in MAFFT (Katoh & Standley 2013). Sequences with ambiguous taxonomic assignment (e.g., species name containing '.spp', 'var', etc.) were excluded from the analysis, as were rare sequences with collection dates prior to 1980. We also downloaded sequences from plants (markers: ITS, MatK, RbcL), amphibians (COI), and molluscs (COI), but the proportion of sequences with collection dates and coordinates was too low for a reliable analysis of temporal trends.

Global human population and land use estimates were obtained from the most extensive and up-to-date version of the 'History Database of the Global Environment' (HYDE 3.2; Klein Goldewijk *et al.* 2017) for years 1980-2016. HYDE 3.2 provides land use and human population

density estimates for all land masses divided into 5 arc-minute grid cells. Data are available for every decade from 1980 to 2000, and for every year after 2000. HYDE 3.2 variables included in our analyses were: maximum land area (km^2/grid cell), human population counts (inhabitants/grid cell), as well as four classes of intensive land use: cropland, pasture, converted rangeland, and built-up area (km^2/grid cell). See Klein Goldewijk *et al.* (2017) for a detailed description of these variables. HYDE 3.2 datasets for the time-period 1980-2016 were converted to raster data structures using the R package 'raster' (Hijmans 2015). Cropland area, pasture area, converted rangeland area, and built-up area were divided by grid cell land area to estimate the proportions of each cell consisting of each respective category. We then computed two grid cell-specific human impact variables: (1) 'human population density' (inhabitants/km^2), i.e. human population counts divided by grid cell land area, and (2) 'land use intensity' (ranging from 0 to 1), i.e. the summed proportions of cropland, pasture, converted rangeland, and built-up area in a grid cell, giving the proportion of land under intensive use.

Using sequence-specific geographic coordinates, all COI sequences were then overlaid on the HYDE 3.2 map closest in time to the year of sequence collection. All sequences were attributed a human population density and a land use intensity value based on the grid cell in which they fell. Sequences sampled between 2000 and 2016 were assigned year-specific human density and land use intensity values; however, due to the absence of yearly HYDE 3.2 data for years 1980-1999, sequences were assigned values computed from the most recent map, i.e. year 1980 for sequences collected between 1980 and 1989 and year 1990 for sequences collected between 1990 and 1999. Sequences from grid cells composed exclusively of water were excluded, as were sequences falling exactly on the border between two grid cells with (potentially) different land use categories. However, we included data from aquatic animals (e.g. bony fishes) found in grid cells that contained some land, reasoning that land use can impact inland and coastal waters and thus the animals found in these environments (Stoms *et al.* 2005; Dudgeon *et al.* 2006). All data were processed using the 'tidyverse' collection of R packages (Wickham & RStudio 2017).

To visualize our dataset, we examined the distribution of COI sequences with respect to geography, taxonomy, time, and human impact variables. We first generated a global map showing the number of sequences falling into each HYDE (5') grid cell, using R packages 'latticeExtra' (Sarkar & Andrews 2016) and 'rworldmap' (South 2011). The number of

sequences per year were tallied for each animal class (birds, fishes, insects, and mammals). We also determined whether sequences were under levels of human influence representative of global averages by comparing the distribution of human density and land use intensity values associated with 1) all HYDE 3.2 grid cells, worldwide and pooled across all years, and 2) the time and place of sequence collections.

For each class, we quantified the proportion of global taxa represented in our sequence dataset. We first retrieved genus, family, and order-level classification for all species using taxonomic information from the NCBI and the 'Integrated Taxonomic Information System' (ITIS) databases accessed through the R package 'taxize' (Chamberlain & Szöcs 2013). Supplemental information regarding Actinopterygii order classification was obtained from 'Fishbase' (Froese & Pauly 2018; *www.fishbase.org*) using 'rfishbase' (Boettiger *et al.* 2012). Taxonomic information for species missing full classifications in NCBI, ITIS, and Fishbase (such as obsolete, or synonymous names) was retrieved from BOLD itself. This order was chosen to ensure that the classification was as current as possible. The total number of genera, families, and orders in each class was then obtained from the 'Catalogue of Life' 2018 database (Roskov *et al.* 2018; *www.catalogueoflife.org/annual-checklist/2018/*) in taxize and compared to the distribution of taxa in our dataset. This provided a rough estimate of the proportion of taxa included in our dataset; exact proportions cannot be calculated due to unresolved taxonomy and discrepancies among taxonomic references (Guralnick *et al.* 2007). The proportion of taxonomic representation obtained for Actinopterygii orders (95%) is likely an over-estimate due to large discrepancies in the number of orders present in different taxonomic databases; for example, some orders represented in BOLD/NCBI (and thus in our dataset) are not present in either Fishbase, nor the Catalogue of Life, and vice versa. Proportions for other classes and taxonomic levels should be more accurate as mismatches across databases were limited.

Calculating genetic diversity and human impacts at the population scale

Intraspecific genetic diversity should be computed at the population scale, aggregating sequences into groups of potentially interbreeding individuals. To split species-specific sequences into populations located in distinct geographical areas, we grouped all sequences from a given species across years and mapped their distribution in space. We then used a simple agglomerative (bottom-up) hierarchical clustering algorithm to group sequences based on spatial

proximity, following the ‘single linkage’ clustering rule (Sneath 1957). In this clustering method, sequences are grouped together as long as the maximum geographical distance among all sequences in the group does not exceed a maximum distance chosen *a priori* (x km; see below for chosen values). For example, with x = 10 km, two sequences A and B would be grouped in the same population if they are separated by 5 km, but a third sequence C would be placed in its own population if it is located more than 10 km from both A and B. We implemented this clustering method with the R function ‘hclust’ (method = ‘single’), using a distance matrix of ‘Great Circle’ distances among sequences. Great circle distances represent geographic distances among objects along the surface of an ellipsoid; in this case a WGS84 ellipsoid used by the R function ‘spDists’ (package: ‘sp’; Pebesma *et al.* 2018).

To assess the scale-dependence of our results, we manipulated ‘x’ in the clustering algorithm, using four different minimum spatial distances for the creation of a new cluster/population: 10, 100, 1000, and 10,000 km (henceforth: ‘scales’). Depending on the species, at the largest spatial scale (10,000 km), sequences spanning an entire continent may be grouped into one population, likely aggregating sequences from isolated and genetically divergent populations. Likewise, at the smallest spatial scale (10 km), sequences from one (true) population may be split into separate populations if no individuals were sampled in the middle of the population’s range. A perhaps better alternative would be to use species-specific dispersal range information to group sequences into populations, however such information was not available for the majority of species included in our analysis. Moreover, important spatial biases in the distribution of COI sequences suggest that population ranges were rarely sampled in their entirety, and the mean great circle distance among sequences was 0 for ~ 50 % of species in our dataset, and for other populations distances were usually < 100 km (Fig. S1). Therefore, with the data at hand, the best compromise to incorporate sequence sampling autocorrelation and scale-dependence in our analysis was to examine whether our inferences were altered by the choice of clustering scale criterion.

Following this procedure, each sequence was attributed a unique population ID at each spatial scale (e.g. common redpoll, *Acanthis flammea*, in population ‘1’ at scale = 10km, with a mean geographical location of 52.4° latitude and 4.54° longitude). Pairwise nucleotide differences were calculated among all sequences in a population with > 50% sequence overlap as in Miraldo *et al.* (2016), using adapted Julia script available at https://github.com/mkborregaard/.

We then calculated the mean pairwise dissimilarity among population-specific sequence comparisons to estimate nucleotide diversity ($\hat{\pi}$), a measure of population genetic diversity (henceforth: 'diversity'; (Nei & Li 1979; Tajima 1993). Since at least two sequences are required to calculate diversity, populations with one sequence were not considered. Diversity estimates with extreme values 10 standard deviations greater than the mean of all estimates were discarded. For populations with multiple years of data, separate $\hat{\pi}$ values were computed for each year to assess temporal trends in diversity. It should be noted that spatial scale affects the number of sequences and populations in the dataset (Table 1), as the spatial scale of sequence aggregation influences both the number of units into which species-specific sequences are split, and the probability that a sequence is rejected for being alone in its population.

For each population and at each scale, we calculated the geographical centroid of the population (latitude and longitude), the mean great circle distance among sequences contributing to each $\hat{\pi}$ estimate, and the mean land use intensity and human population density of HYDE 3.2 grid cells from which the sequences originated. We also estimated the variance in human density and land use intensity among sequences, reasoning that populations covering more heterogeneous land areas could potentially have higher genetic diversity. However, population-level land use heterogeneity was strongly collinear with mean geographic distance among sequences, as populations with a large spatial extent were more likely to contain various types of land uses. We therefore excluded land use heterogeneity variables from the analysis, preferring to focus on the effect of geographic distance, a strong (and well-known) predictor of genetic diversity (Gratton *et al.* 2017).

Spatial analyses

We examined the drivers of COI nucleotide diversity with generalized additive mixed models (GAMMs) fitted with the 'bam' function in the R package 'mgcv' (Wood 2017). Models were constructed separately for each animal class and spatial scale of analysis. All GAMMs used a Tweedie error structure (with a log link function) to account for the zero-inflated, positive, and right-skewed distribution of diversity values. Because nucleotide diversity is likely better estimated with an increasing number of pairwise sequence comparisons, we used as model weights the log-transformed number of sequences contributing to a diversity estimate, divided by the mean log-transformed number of sequences of all diversity estimates belonging to a given

class and spatial scale. To map diversity, we first fitted a GAMM with the following predictor variables: mean geographic (great circle) distance among sequences (a smooth term), latitude and longitude (a Gaussian process smooth to account for spatial autocorrelation), and year of sequence collection, a random effect which could influence diversity estimates if, for example, the error rate of sequencing technology has changed over time. The structure of this model was: diversity ~ s(distance, k=10, bs='tp') + s(lat, long, bs='gp', k=50) + s(year, bs= 're', k =5, m=1). Following Gratton *et al.* (2017), fitted values from this model were used to construct global maps of 'smoothed' COI nucleotide diversity, after removing the confounding effects of spatial autocorrelation and mean spatial distance among sequences in a population (setting it to 0 km). Diversity values in 2 x 2° grid cells (a spatial resolution facilitating visualization) were predicted for all combinations of latitudes and longitudes with COI data available. Predicted values were obtained from GAMMs fitted at the 1000 km scale, which corresponded best to the grain size of the maps.

We then fitted another GAMM including additional predictors (smooth terms) hypothesized to influence genetic diversity: absolute value of latitude (in degrees), mean human population density, and mean land use intensity. The structure of this model was: diversity ~ s(distance, k=10, bs='tp') + s(lat, long, bs='gp', k=50) + s(year, bs= 're', k =5, m=1) +s(|latitude|, k=8, bs='tp') + s(human.density, k=8, bs='tp') + s(land.use, k=8, bs='tp'). All predictors were log(1+x)-transformed and then rescaled from 0 to 1 to facilitate model fitting. A few species contributed more than one population to our dataset, such that some data points might not be independent (Table 1: columns 'populations' vs. 'species'). However, given the large number of species in the dataset, and that most species had a single data point, we were computationally unable to fit thousands of random smooths and include 'species' as a random effect. Instead, in this spatial analysis, we simply retained a single observation per species (the $\hat{\pi}$ estimate with the greatest number of sequence comparisons, considered most reliable). A condensed summary of model predictions from all GAMMs is shown in the main text, while supplementary figures S2 to S5 provide additional model information (confidence intervals around predictions, F and p values, and adjusted R^2). When showing the predicted relationship between diversity and one of the smooth terms in the model, values for all other predictors were set to their median value in the taxon- and scale-specific dataset used to fit the GAMM.

All models were validated with plots of residuals against fitted values and predictor variables, with autocorrelation functions (for temporal autocorrelation), and with variograms and maps of residuals (for spatial autocorrelation). We also ensured that the choice of basis dimension (k) was sufficiently high using the 'gam.check' function in mgcv. Multicollinearity was assessed with variance inflation factors, which were below 3 for all variables in all models. Finally, we conducted a sensitivity analysis to determine whether the low number of sequences contributing to most diversity estimates influenced our results. Indeed, the median number of sequences compared across all diversity estimates was 3 (range = 2–1434 sequences; Fig. S6). Using sequence number as weights in the GAMMs partly accounts for this problem; however, we also re-ran all analyses using a stricter data inclusion criterion, whereby all diversity estimates with fewer than 5 sequences were discarded. Given the dataset, this is the strictest criterion we could apply while keeping the same GAMM structure and fitting models for all combinations of animal classes and spatial scales. Results of GAMMs only including diversity estimates with > 5 sequences are presented in Figures S7-S11.

Time series analyses

We investigated temporal trends in the subset of populations that were sampled repeatedly over time (4 or more years of data). This analysis was conducted exclusively at the 1000 km scale because smaller scales produced very few time series (e.g., < 10 series for some taxa), whilst at the largest scale (10,000 km), very distant individuals could be sampled in different years, thus conflating spatial and temporal influences on diversity. We chose 4 years as the minimum number of years and retained all diversity estimates regardless of their number of sequences because stricter thresholds led to too few time series for reliable analysis. These data inclusion criteria led to 909 time series from 873 species. With these, we first estimated Mann-Kendall trend coefficients (τ), a non-parametric test of monotonic (directional) trend based on the Kendall rank correlation. Statistically-significant coefficients below and above 0 respectively indicate decreasing and increasing monotonic trends of COI nucleotide diversity within a population.

We then fitted one GAMM per taxon, again using a Tweedie error structure and the relative log number of sequences as model weights, and re-scaling all predictor variables from 0 to 1. The model structure was similar to the spatial GAMMs described above, with the inclusion

of a smooth term (thin plate spline) for ‘year’, a factor-smooth interaction (random effect) for population, and two tensor-product interactions between ‘year’ and ‘human density’ or ‘land use intensity’. These tensor-product interactions allow temporal trends to vary along human impact gradients, allowing us to test the prediction that diversity is declining (faster) in impacted areas. The model structure was as follows: diversity ~ s(distance, k=8, bs='tp') + s(lat,long, bs='gp', k=50) + s(year, k=8, bs='tp') + s(human.density, k=8, bs='tp') + s(land.use, k=8, bs='tp') + ti(year, human.density, k=8) + ti(year, land.use, k=8) + s(year, population, bs='fs', k=5, m=1). Again, we could not fit a random effect for ‘species’ given that most species included a single time series; we thus retained a single series per species (the longest) to avoid non-independence. Models were validated as described above for spatial GAMMs, and the statistical-significance of smooth terms was assessed with the ‘summary.gam’ function in mgcv.

Results

Our dataset includes a total of 175,247 COI sequences sampled from 17,082 species of birds (Aves), inland and coastal bony fishes (Actinopterygii), insects (Insecta), and mammals (Mammalia; Table 1). The aggregation of species-specific sequences sampled in the same area resulted in a total of 17,124–27,588 'populations', depending on the spatial scale of analysis. Mapping the geographic location of sequences revealed a spatially heterogeneous pattern of sampling, with 70.2% of sequences originating from North America and Europe (Fig 1a). With respect to taxonomy, the dataset was dominated (74%) by insect sequences (Fig. 1b; Table 1). Moreover, for all classes, 1–3 speciose orders contributed a large proportion of sequences (Fig. 1b; see section ‘Supplementary Results’ in supporting information for additional information on taxonomic biases). Only a small proportion of the global number of species of birds, bony fishes, insects, and mammals was represented in our dataset, but this number increased significantly for higher taxonomic levels (families and orders), suggesting a phylogenetically-broad pattern of biodiversity sampling (Table 1). The number of COI sequences collected on any given year and deposited in GenBank or BOLD databases has increased for birds, fishes, and insects from 1980 to 2010 (Fig. 1c). In contrast, mammal sequence collection seems to have reached a peak and remained stable from approximately 1985 to 2005 (Fig 1c). All groups demonstrate a recent (~5 year) decline in the number of sequences, perhaps as a result of the lag between sequence collection and sequence availability in the databases (Fig. 1c). Finally, sequences in the database

originate from grid cells generally representative of the distribution of extensive heavy land use, however from areas with higher human population density than the mean of all grid cells across the world (Fig. 1d, e).

Fig. 2 shows global maps of smoothed COI nucleotide diversity where sequence data was available, after removing the confounding (positive) effect of spatial distance among sequences. A latitudinal gradient was apparent for birds, fishes, and mammals, but not insects (Fig. 2). Generalized additive mixed models (GAMMs) confirmed that the geographic distance among sequences in a population was a strong positive predictor of genetic diversity in all taxa (Fig. 3a). This effect was strongest at large spatial scales (> 1000 km), when potentially distant and divergent sequences are aggregated into populations. At small spatial scales (10 and 100 km), mean geographic distances are small, such that there is not enough variance in distance to have a significant effect on diversity (Fig. 3a). The absolute latitude of a population centroid was a negative but weak predictor of genetic diversity in birds, fishes, and mammals, but not insects (Fig. 3b). This effect was visible in 3 out of 4 scales in all three taxa with significant latitudinal gradients, although the scales at which this effect was significant varied among taxa (Fig. 3b). Moreover, for fishes, the significant trend was entirely driven by latitudes close to the equator, and model uncertainty was important at higher latitudes (Fig. S3). Finally, human impacts on genetic diversity were taxon-dependent and generally weak (Fig. 3c, d). The clearest effects were: 1) the negative impact of very high human population density on insect genetic diversity, apparent at all scales of analysis (at human densities > 5000 inhabitants/km^2, corresponding to a value of 0.75 on the scaled axis; Fig. 3c), and: 2) the negative impact of heavy land use on fish genetic diversity (when grid cells occupied by the animal population had 70% or more intensive land use), at all but the largest spatial scale (Fig. 3d).

Spatial GAMMs had low adjusted R^2 (< 0.2; Figs. S2-S5), indicating that despite the presence of some significant effects, the predictor variables included in this analysis (latitude, longitude, spatial distance among sequences, human impacts, and year of sequence collection) were not, collectively, strong drivers of COI nucleotide diversity. This could in part be due to the low precision of diversity estimates, many of which were based on a handful of sequence comparisons (Fig. S6).

When fitting spatial GAMMs with more stringent data inclusion criteria (using diversity estimates based on ≥ 5 sequences, giving 10 potential pairwise comparisons or more), we

confirmed the strong positive effect of geographic distance among sequences on genetic diversity, visible in all taxa at the two largest spatial scales (Fig. S7). However, negative effects of absolute latitude on diversity vanished in birds and fishes, while it remained visible in mammals (Fig. S7), suggesting that a clear latitudinal gradient is only present in this taxon. Negative impacts of human population density on insect genetic diversity and of land use on fish genetic diversity remained apparent at all scales in this analysis; in fact, these trends were strengthened by the exclusion of diversity estimates based on < 5 sequences (Fig. S7). For bird and fish GAMMs, the stricter data inclusion criterion generally improved model R^2 (Figs. S8-S11), although values remained low (maximum adjusted R^2 across all models = 0.43).

A Mann-Kendall trend analysis of 909 time series of COI nucleotide diversity in populations with at least 4 years of data indicated that most time series showed no significant monotonic trend in diversity, regardless of time series duration (Fig. 4a). In the 2.5% of time series with significant temporal trends, an equal number of time series showed positive vs. negative trends (Fig. 4a). GAMMs incorporating a smooth term for time and its interaction with human impact variables revealed no significant temporal trend in COI nucleotide diversity in birds, fishes, and mammals (Fig. 4b). In insects, a significant trend was visible (at $p < 0.01$), driven by a modest reduction in diversity from 2010 to 2016 (Fig. 4b). Tensor product interactions between year and human density or land use intensity revealed several significant interactions, showing that temporal trends in genetic diversity often varied along the two anthropogenic gradients–although effects were stronger for land use than for human population density (Fig. 4c, d). For example, in fishes, diversity declined in populations experiencing heavy land use while it remained stable in less-disturbed populations, confirming the result obtained in the spatial analysis (Fig. 4d). However, other interactions were complex and inconsistent with our predictions. For example, in mammals, genetic diversity pre-1990 was highest in areas with heavy land use, but then declined over time to the same mean value as diversity in areas with little land use, while in insects, temporal increases in diversity were greatest in areas with heavy land use (Fig. 4d).

Discussion

Analyzing worldwide patterns of intraspecific genetic diversity in birds, fishes, insects, and mammals across space, time, and anthropogenic gradients, we show that the geographic

distance among sequences was by far the strongest driver of diversity in the global dataset of animal mitochondrial sequences deposited on BOLD and GenBank (Fig. 3). This confirms the necessity to account for this factor when measuring genetic diversity (Gratton *et al.* 2017), especially when sequences are aggregated across large spatial scales. After accounting for the distance-decay of similarity, latitude still had an influence on the distribution of diversity for birds, fishes, and mammals. The elevated diversity documented in mammals at low latitudes (Miraldo *et al.* 2016; Gratton *et al.* 2017) therefore extends to our dataset of mammal sequences, as well as to birds and fishes (although in these two taxa, this result was uncertain and sensitive to rarefaction). We did not observe significant monotonic changes in genetic diversity; time series analysis revealed that the average temporal trend across populations was zero for all taxa but insects, the latter only showing a recent, weak, and uncertain decline in genetic diversity (Fig. 4a, b). However, some individual populations of fishes, insects, and mammals do demonstrate temporal changes in diversity, with an equal proportion of increasing and decreasing trends. Finally, the direction and magnitude of human impacts were taxon-dependent and not generally obvious, even at small spatial scales. As such, our initial hypotheses regarding human impacts were mostly unsupported, which also questions one of the key results of Miraldo *et al.* (2016).

Current estimates indicate that up to 70% of the Earth's surface has been modified by human activities, especially within the last century (Foley *et al.* 2005). Human impacts on the environment such as urbanization and land use intensification are known to influence intraspecific variation and species evolutionary parameters (DiBattista 2008; Hendry *et al.* 2008; Alberti *et al.* 2017; Fugère & Hendry 2018), and can have both negative and positive effects on species diversity (Pautasso 2007; McGill *et al.* 2015). It is therefore not surprising that we find complex, often weak, and always taxon-dependent human impacts on diversity–that is, it might be overly simplistic to always expect lower diversity in impacted areas. However, two of our findings, robust to rarefaction and clearly visible across spatial scales of analysis, warrant further attention given their potential importance for conservation. First, we found lower insect genetic diversity in areas with high human densities, which is consistent with the dramatic contemporary decline in insect abundance attributed in part to habitat loss and urbanization (Lister & Garcia 2018; Sánchez-Bayo & Wyckhuys 2019). Second, intensive land use had a negative impact on fish genetic diversity, a finding apparent in both spatial and time series analyses. Apart from

these two general trends, changes in the magnitude and sometimes direction of human impacts through time also reinforce the need for taxon and population-specific studies, and suggest that attempts at generalization will be challenging.

While our results do not robustly support the finding that intraspecific genetic diversity is lower in human-dominated areas (Miraldo *et al.* 2016), we believe that detrimental effects of human activity on global genetic diversity is still a potential outcome–but one that is difficult to detect because of data limitations and analytical shortcomings. For example, areas with heterogeneous land use may harbour higher genetic diversity than homogeneous landscapes; however, our models could not incorporate land use heterogeneity due to its strong collinearity with the mean pairwise geographic distances among sequences. Future analyses could perhaps solve this issue by employing statistical methods that are robust to collinearity (e.g. machine learning approaches). Another limitation is that a large proportion of sequences in our dataset originate from a single 'site' (set of coordinates), especially at small spatial scales. Without sufficient spatial variation among sequences, we cannot effectively test the importance of spatial scale and land use heterogeneity on genetic diversity and incorporate, for example, 'neighbour effects' of land use on diversity in adjacent areas. At large spatial scales (e.g. 10,000 km maximal geographic distance), sequences do vary in geographic location, but to such an extent that isolation by distance necessarily becomes the main determinant of genetic diversity. Finally, another important advancement for the future assessment of global genetic diversity trends would be to incorporate species traits, dispersal potential, and phylogeny, all of which could influence relationships between diversity and anthropogenic drivers (Wiens 1989; Levin 1992; Chave 2013; Schiebelhut & Dawson 2018). For example, recent analyses suggest that movement distances have been decreased in human modified landscapes (Tucker *et al.* 2018), which may be affecting population connectivity and thus genetic diversity.

We also acknowledge that the COI locus does not evolve under neutrality (Pentinsaari *et al.* 2016) and that sequence variation at this one mitochondrial locus does not reflect genomic intraspecific diversity, nor may it necessarily reflect anthropogenic pressures (Bazin *et al.* 2006). More appropriate genetic tools and metrics are known for measuring neutral intraspecific diversity (e.g., microsatellites, allelic diversity); however, there is currently no global database for these data. COI is thus one of few genes with abundant sequences in common databases (Porter & Hajibabaei 2018), with metadata readily available (e.g. spatial coordinates and year of

sequence collection), and with sequences available for a large number of species (likely due to the use of COI for species identification); hence the use of COI in a previous global assessment of genetic diversity (Miraldo *et al.* 2016). Despite this wealth of COI data with collection years, we still noted some taxonomic gaps in the database, e.g. a small number of Coleoptera sequences despite this order accounting for a large number of insect (and animal) species. As outlined in Supplementary Results, a few individual species contributed a disproportionately large number of sequences. Moreover, a majority of populations in our dataset were represented by < 10 sequences and/or a single year of data, which limits the precision of diversity estimates and constrains time series analyses. All of this suggests that data limitations still hinder our ability to assess global patterns of genetic diversity.

In conclusion, anthropogenic activity has complex, taxon-specific effects on intraspecific genetic diversity, such that one cannot simply conclude that genetic diversity is generally lower in human-impacted areas. There is a clear need to establish a global and systematic monitoring program to gather more data on intraspecific diversity (Mimura *et al.* 2017). For example, more time series of genetic diversity within individual populations are urgently needed. The lack of replication in time is a persistent problem in modern ecology (Estes *et al.* 2018) and evolutionary biology that is constraining our ability to make strong inferences about global patterns of biodiversity change (Gonzalez *et al.* 2016; White 2019). We also urge data collectors to upload metadata such as collection year and spatial coordinates when depositing sequences in databases–a remarkably large number of sequences in GenBank do not have a collection year (e.g. 95% of amphibian sequences, hence their exclusion from this analysis; Marques *et al.* 2013; Pope *et al.* 2015). Global monitoring of genetic diversity would improve our ability to detect change and attribute the causes of worldwide patterns of spatial and temporal variation in genetic diversity we report here.

Acknowledgements

This work resulted from insightful discussions among past and present Gonzalez lab members. We are grateful to Andreia Miraldo and Michael Borregaard for assistance with Julia scripts. We thank Navin Ramankutty, Erle Ellis, and Kees Klein Goldewijk for help accessing and interpreting the HYDE 3.2 dataset. Simon Joly, Arne Mooers, and Daniel Schoen contributed insightful comments on earlier versions of the manuscript. The authors acknowledge

support from the Natural Sciences and Engineering Research Council of Canada (NSERC) and the Quebec Centre for Biodiversity Science (QCBS). AG is supported by the Liber Ero Chair in Biodiversity Conservation and a Killam Fellowship.

Competing interests

The authors declare no competing interests.

References

Alberti, M. (2015). Eco-evolutionary dynamics in an urbanizing planet. *Trends Ecol. Evol.*, **30**, 114–126.

Alberti, M., Correa, C., Marzluff, J.M., Hendry, A.P., Palkovacs, E.P., Gotanda, K.M., *et al.* (2017). Global urban signatures of phenotypic change in animal and plant populations. *PNAS*, **114**, 8951–8956.

Banks, S.C., Cary, G.J., Smith, A.L., Davies, I.D., Driscoll, D.A., Gill, A.M., *et al.* (2013). How does ecological disturbance influence genetic diversity? *Trends Ecol. Evol.*, **28**, 670–679.

Bazin, E., Glémin, S. & Galtier, N. (2006). Population size does not influence mitochondrial genetic diversity in animals. *Science*, **312**, 570.

Benson, D.A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., *et al.* (2013). GenBank. *Nucleic Acids Res.*, **41**, D36–D42.

Bickham, J.W., Sandhu, S., Hebert, P.D.N., Chikhi, L. & Athwal, R. (2000). Effects of chemical contaminants on genetic diversity in natural populations: implications for biomonitoring and ecotoxicology. *Mutat. Res. Rev. Mutat. Res.*, **463**, 33–51.

Bijlsma, R. & Loeschcke, V. (2012). Genetic erosion impedes adaptive responses to stressful environments. *Evol. Appl.*, **5**, 117–129.

Boettiger, C., Lang, D.T. & Wainwright, P.C. (2012). rfishbase: exploring, manipulating and visualizing FishBase data from R. *J. Fish Biol.*, **81**, 2030–2039.

Ceballos, G., Ehrlich, P.R. & Dirzo, R. (2017). Biological annihilation via the ongoing sixth mass extinction signaled by vertebrate population losses and declines. *PNAS*, **114**, E6089–E6096.

Chamberlain, S.A. & Szöcs, E. (2013). taxize: taxonomic search and retrieval in R. *F1000Res.*, **2**, 191.

Chave, J. (2013). The problem of pattern and scale in ecology: what have we learned in 20 years? *Ecol. Lett.*, **16**, 4–16.

Crispo, E., Moore, J.-S., Lee-Yaw, J.A., Gray, S.M. & Haller, B.C. (2011). Broken barriers: human-induced changes to gene flow and introgression in animals. *BioEssays*, **33**, 508–518.

DiBattista, J.D. (2008). Patterns of genetic variation in anthropogenically impacted populations. *Conserv. Genet.*, **9**, 141–156.

Dubrova, Y.E., Nesterov, V.N., Krouchinsky, N.G., Ostapenko, V.A., Neumann, R., Neil, D.L., *et al.* (1996). Human minisatellite mutation rate after the Chernobyl accident. *Nature*, **380**, 683–686.

Dudgeon, D., Arthington, A.H., Gessner, M.O., Kawabata, Z.-I., Knowler, D.J., Lévêque, C., *et al.* (2006). Freshwater biodiversity: importance, threats, status and conservation challenges. *Biol. Rev.*, **81**, 163–182.

Ellegren, H., Lindgren, G., Primmer, C.R. & Møller, A.P. (1997). Fitness loss and germline mutations in barn swallows breeding in Chernobyl. *Nature*, **389**, 593.

Ellis, E.C. & Ramankutty, N. (2008). Putting people in the map: anthropogenic biomes of the world. *Front. Ecol. Environ.*, **6**, 439–447.

Estes, L., Elsen, P.R., Treuer, T., Ahmed, L., Caylor, K., Chang, J., *et al.* (2018). The spatial and temporal domains of modern ecology. *Nat. Ecol. Evol.*, **2**, 819–826.

Foley, J.A., DeFries, R., Asner, G.P., Barford, C., Bonan, G., Carpenter, S.R., *et al.* (2005). Global consequences of land use. *Science*, **309**, 570–574.

Frankham, R. (2005). Genetics and extinction. *Biol. Conserv.*, **126**, 131–140.

Froese, R. & Pauly, D. (2018). *FishBase*. World Wide Web electronic publication. Available at: http://www.fishbase.org.

Fugère, V. & Hendry, A.P. (2018). Human influences on the strength of phenotypic selection. *PNAS*, **115**, 10070–10075.

Gonzalez, A., Cardinale, B.J., Allington, G.R.H., Byrnes, J., Endsley, K.A., Brown, D.G., *et al.* (2016). Estimating local biodiversity change: a critique of papers claiming no net loss of local diversity. *Ecology*, **97**, 1949–1960.

Gratton, P., Marta, S., Bocksberger, G., Winter, M., Keil, P., Trucchi, E., *et al.* (2017). Which latitudinal gradients for genetic diversity? *Trends Ecol. Evol.*, **32**, 724–726.

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

Hendry, A.P., Farrugia, T.J. & Kinnison, M.T. (2008). Human influences on rates of phenotypic change in wild animal populations. *Mol. Ecol.*, **17**, 20–29.

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

Hijmans, R.J. (2015). Raster: geographic data analysis and modeling. R package raster version 2.8-19. Available at: https://CRAN.R-project.org/package=raster.

Hughes, A.R., Inouye, B.D., Johnson, M.T.J., Underwood, N. & Vellend, M. (2008). Ecological consequences of genetic diversity. *Ecol. Lett.*, **11**, 609–623.

Hughes, J.B., Daily, G.C. & Ehrlich, P.R. (1997). Population diversity: its extent and extinction. *Science*, **278**, 689–692.

Jarzyna, M.A. & Jetz, W. (2018). Taxonomic and functional diversity change is scale dependent. *Nat. Commun.*, **9**, 2565.

Katoh, K. & Standley, D.M. (2013). MAFFT Multiple Sequence Alignment Software version 7: improvements in performance and usability. *Mol. Biol. Evol.*, **30**, 772–780.

Kimura, M. (1968). Evolutionary rate at the molecular level. *Nature*, **217**, 624–626.

Kimura, M. (1983). *The Neutral Theory of Molecular Evolution*. New York, Cambridge University Press.

King, J. & Jukes, T. (1969). Non-Darwinian evolution. *Science*, **164**, 788–798.

Klein Goldewijk, K., Beusen, A., Doelman, J. & Stehfest, E. (2017). Anthropogenic land use estimates for the Holocene – HYDE 3.2. *Earth Syst. Sci. Data*, **9**, 927–953.

Lawrence, E.R., Benavente, J.N., Matte, J.-M., Marin, K., Wells, Z.R.R., Bernos, T.A., *et al.* (2019). Geo-referenced population-specific microsatellite data across American continents, the MacroPopGen Database. *Sci. Data*, **6**, 14.

Levin, S.A. (1992). The problem of pattern and scale in ecology. *Ecology*, **73**, 1943–1967.

Lister, B.C. & Garcia, A. (2018). Climate-driven declines in arthropod abundance restructure a rainforest food web. *PNAS*, **115**, E10397–E10406.

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

McGill, B.J., Dornelas, M., Gotelli, N.J. & Magurran, A.E. (2015). Fifteen forms of biodiversity trend in the Anthropocene. *Trends Ecol. Evol.*, **30**, 104–113.

Mimura, M., Yahara, T., Faith, D.P., Vázquez-Domínguez, E., Colautti, R.I., Araki, H., *et al.* (2017). Understanding and monitoring the consequences of human impacts on intraspecific variation. *Evol. Appl.*, **10**, 121–139.

Miraldo, A., Li, S., Borregaard, M.K., Flórez-Rodríguez, A., Gopalakrishnan, S., Rizvanovic, M., *et al.* (2016). An Anthropocene map of genetic diversity. *Science*, **353**, 1532–1535.

Moritz, C., Dowling, T.E. & Brown, W.M. (1987). Evolution of animal mitochondrial DNA: relevance for population biology and systematics. *Ann. Rev. Ecol. Syst.*, **18**, 269–292.

Nei, M. & Li, W.H. (1979). Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proc. Natl. Acad. Sci., USA*, **76**, 5269–5273.

Palumbi, S.R. (2001). Humans as the world's greatest evolutionary force. *Science*, **293**, 1786–1790.

Pautasso, M. (2007). Scale dependence of the correlation between human population presence and vertebrate and plant species richness. *Ecol. Lett.*, **10**, 16–24.

Paz-Vinas Ivan, Loot Géraldine, Hermoso Virgilio, Veyssière Charlotte, Poulet Nicolas, Grenouillet Gaël, *et al.* (2018). Systematic conservation planning for intraspecific genetic diversity. *Proc. R. Soc. B*, **285**, 20172746.

Pebesma, E., Bivand, R. S., Rowlingson, B., Gomez-Rubio, V., Hijmans, R., Sumner, M., *et al.* (2018). sp: classes and methods for spatial data in R. *R News 5*, **2**.

Pentinsaari, M., Salmela, H., Mutanen, M. & Roslin, T. (2016). Molecular evolution of a widely-adopted taxonomic marker (COI) across the animal tree of life. *Sci. Rep.*, **6**, 35275.

Pereira, H.M., Ferrier, S., Walters, M., Geller, G.N., Jongman, R.H.G., Scholes, R.J., *et al.* (2013). Essential biodiversity variables. *Science*, **339**, 277–278.

Pope, L.C., Liggins, L., Keyse, J., Carvalho, S.B. & Riginos, C. (2015). Not the time or the place: the missing spatio-temporal link in publicly available genetic data. *Mol. Ecol.*, **24**, 3802–3809.

Porter, T.M. & Hajibabaei, M. (2018). Over 2.5 million COI sequences in GenBank and growing. *PLoS One*, **13**, e0200177.

R Core Team. (2018). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. Available at: http://www.R-project.org.

Ratnasingham, S. & Hebert, P.D.N. (2007). BOLD: The Barcode of Life Data System (http://www.barcodinglife.org). *Mol. Ecol. Notes*, **7**, 355–364.

Reed, D.H. & Frankham, R. (2003). Correlation between fitness and genetic diversity. *Conserv. Biol.*, **17**, 230–237.

Roskov Y., Ower G., Orrell T., Nicolson D., Bailly N., Kirk P.M., et al. (2018). *Species 2000 & ITIS Catalogue of Life, 2018 Annual Checklist*. Available at: http://www.catalogueoflife.org/annual-checklist/2018.

Sánchez-Bayo, F. & Wyckhuys, K.A.G. (2019). Worldwide decline of the entomofauna: a review of its drivers. *Biol. Conserv.*, **232**, 8–27.

Sarkar, D. & Andrews, F. (2016). latticeExtra: extra graphical utilities based on lattice. R package version 0.6-28. Available at: https://CRAN.R-project.org/package=latticeExtra.

Schiebelhut, L.M. & Dawson, M.N. (2018). Correlates of population genetic differentiation in marine and terrestrial environments. *J. Biogeogr.*, **45**, 2427–2441.

Schlaepfer, D.R., Braschler, B., Rusterholz, H.-P. & Baur, B. (2018). Genetic effects of anthropogenic habitat fragmentation on remnant animal and plant populations: a meta-analysis. *Ecosphere*, **9**, e02488.

Sneath, P.H. (1957). The application of computers to taxonomy. *J. Gen. Microbiol.*, **17**, 201–226.

South, A. (2011). rworldmap: a new R package for mapping global data. *The R Journal*, **3**, 35–43.

Stoms, D.M., Davis, F.W., Andelman, S.J., Carr, M.H., Gaines, S.D., Halpern, B.S., *et al.* (2005). Integrated coastal reserve planning: making the land–sea connection. *Front. Ecol. Environ.*, **3**, 429–436.

Tajima, F. (1993). Statistical analysis of DNA polymorphism. *Jpn. J. Genet.*, **68**, 567–595.

Thomas, C.D. (2015). Rapid acceleration of plant speciation during the Anthropocene. *Trends Ecol. Evol.*, **30**, 448–455.

Tucker, M.A., Böhning-Gaese, K., Fagan, W.F., Fryxell, J.M., Moorter, B.V., Alberts, S.C., *et al.* (2018). Moving in the Anthropocene: global reductions in terrestrial mammalian movements. *Science*, **359**, 466–469.

White, E.R. (2019). Minimum time required to detect population trends: the need for long-term monitoring programs. *BioScience*, **69**, 40–46.

Wickham, H. & RStudio. (2017). tidyverse: easily install and load the “tidyverse.”. R package version 1.2.1. Available at: https://CRAN.R-project.org/package=tidyverse.

Wiens, J.A. (1989). Spatial scaling in ecology. *Funct. Ecol.*, **3**, 385–397.

Wood, S.N. (2017). *Generalized Additive Models: An Introduction with R*. 2nd edn. Chapman and Hall/CRC.

Table 1. Number of COI sequences, populations, species, genera, families, and orders included in the dataset at each spatial scale.

scale	class	sequences	populations	species	genera	families	orders
10	Aves	6155	2026	1294 (12.50%)	677 (30.30%)	120 (52.86%)	31 (77.50%)
10	Actinopterygii	18057	4492	2297 (7.06%)	1022 (20.82%)	240 (49.18%)	43 (93.48%)
10	Insecta	105736	18904	9714 (1.05%)	4434 (5.53%)	430 (38.05%)	24 (85.71%)
10	Mammalia	15632	2166	531 (9.07%)	219 (17.03%)	42 (26.58%)	10 (34.48%)
100	Aves	6834	2174	1413 (13.64%)	711 (31.83%)	120 (52.86%)	31 (77.50%)
100	Actinopterygii	19194	3973	2366 (7.28%)	1044 (21.27%)	245 (50.20%)	44 (95.65%)
100	Insecta	114674	17677	10740 (1.16%)	4730 (5.90%)	442 (39.12%)	24 (85.71%)
100	Mammalia	16248	1517	547 (9.35%)	224 (17.42%)	43 (27.22%)	11 (37.93%)
1000	Aves	7931	1977	1561 (15.07%)	746 (33.39%)	121 (53.30%)	31 (77.50%)
1000	Actinopterygii	19919	2884	2401 (7.38%)	1054 (21.48%)	245 (50.20%)	44 (95.65%)
1000	Insecta	125169	14591	12006 (1.29%)	5125 (6.40%)	452 (40.00%)	25 (89.29%)
1000	Mammalia	16673	743	564 (9.64%)	228 (17.73%)	43 (27.22%)	11 (37.93%)
10,000	Aves	8597	1614	1607 (15.52%)	760 (34.02%)	123 (54.19%)	31 (77.50%)
10,000	Actinopterygii	20126	2427	2412 (7.42%)	1056 (21.52%)	246 (50.41%)	44 (95.65%)
10,000	Insecta	129705	12515	12496 (1.35%)	5261 (6.57%)	456 (40.35%)	25 (89.29%)
10,000	Mammalia	16819	568	567 (9.69%)	228 (17.73%)	43 (27.22%)	11 (37.93%)

Scale indicates maximum geographic distance (km) allowed between any two sequences when defining populations. Percentages in parenthesis for species, genera, families and orders are estimates of the proportion of taxa present in the database relative of the total number of taxa globally, as described in the Methods. Note that the percentage for Actinopterygii orders is an over-estimate caused by large discrepancies in order number in different taxonomic databases.

Figure captions

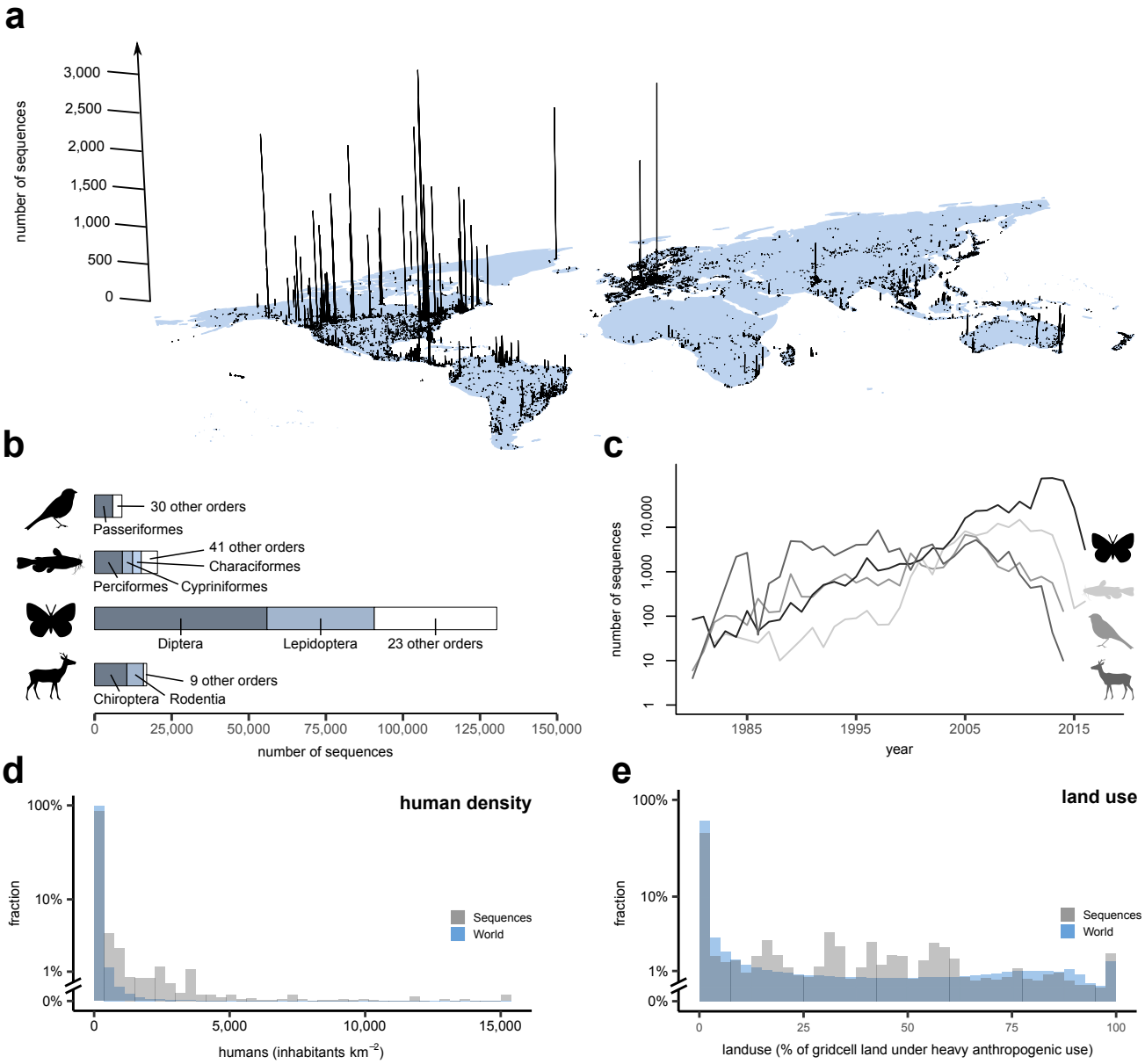
Figure 1. Sequence distribution across geographical space (a), taxonomic classes (b), time (c), and anthropogenic parameters (d, e). COI sequences depicted are those available in NCBI or BOLD with geographic coordinates and year of collection documented in the databases after data filtering steps. (a) Global distribution sequences. Bar heights represent the total number of sequences per 5' grid cell (min=1 sequence, max=3,258) from the four animal classes. (b) Distribution of sequences across the four taxonomic classes: birds (N=8,866 sequences), fish (N=20,388), insects (N=130,433), and mammals (N=16,890). Orders contributing ≥10% of a taxon's sequences are indicated. (c) Number of sequences available in the dataset for each year and animal class. (d, e) Distribution of sequences (grey) according to human density (d) and land use intensity (e) relative to the frequency of these parameter values worldwide (blue).

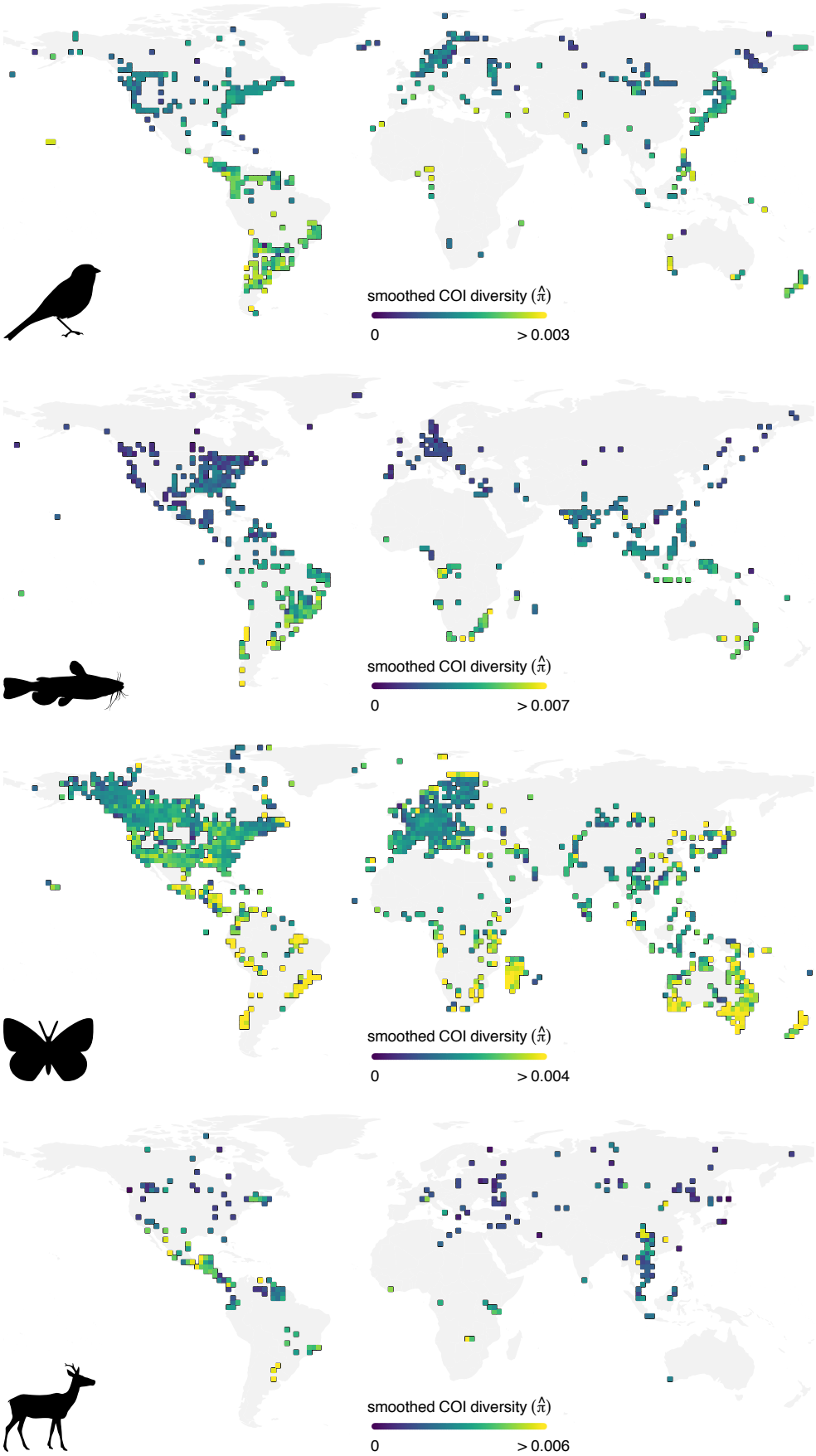
Figure 2. Spatial variation in COI nucleotide diversity of birds, inland and coastal bony fishes, insects, and mammals. Diversity values are GAMM predictions when assuming that all sequences from a given population were collected at the same location, thus removing the confounding effect of population spatial extent on nucleotide diversity. Note that diversity scales differ among animal classes.

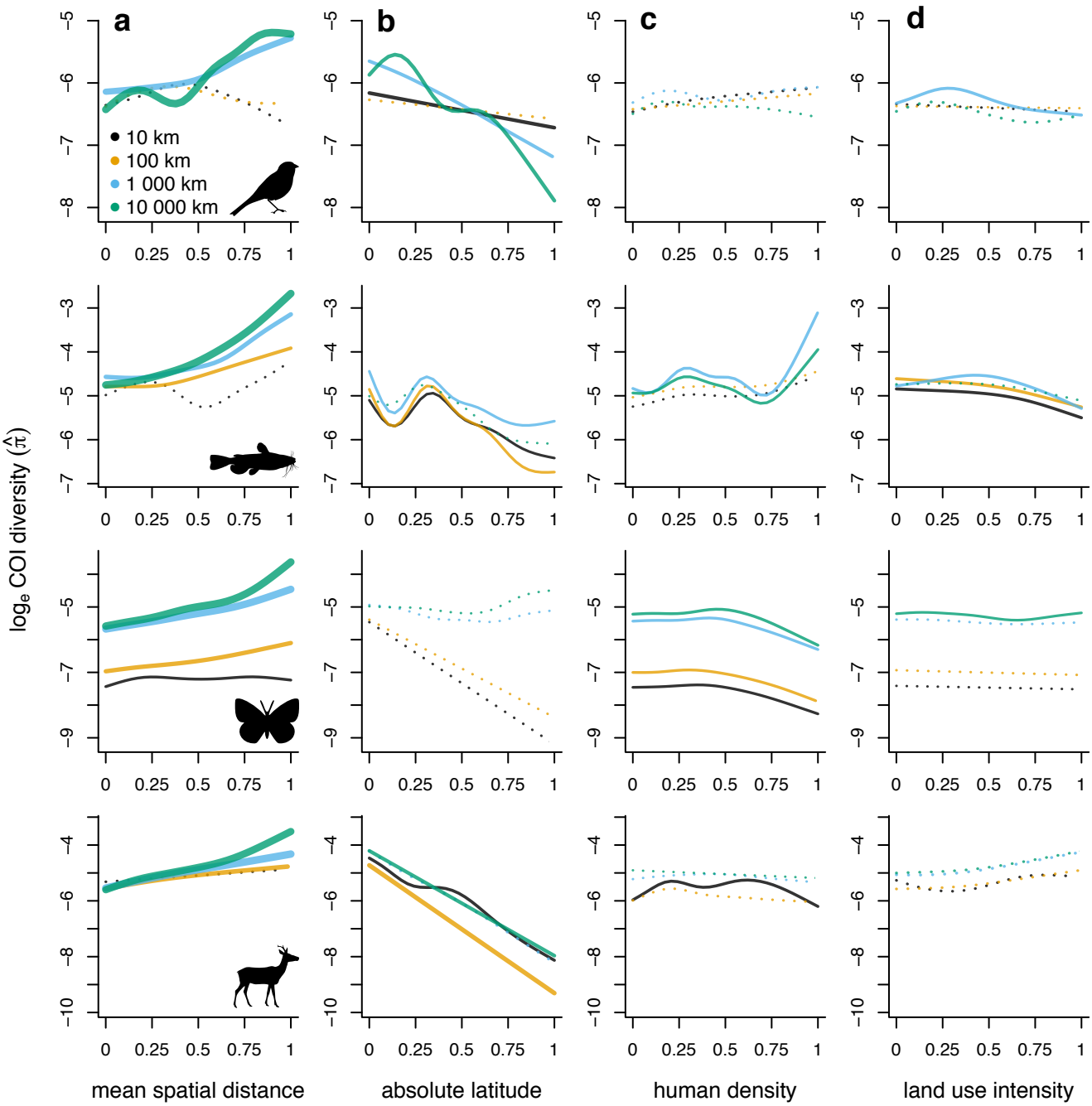
Figure 3. Relationship between COI nucleotide diversity and the mean distance among sequences in a population (a), the absolute latitude of the population centroid (b), and the mean human population density (c) and land use intensity (d) of HYDE 3.2 grid cells from which sequences originated. Panels on the same row belong to the same taxon. Lines represent fitted values from 16 scale and taxon-specific GAMMs (4 scales × 4 classes). Line colour indicates spatial scale; for each taxon, lines with the same colour belong to the same model. Line width indicates effect size (*F* value of predictor) standardized within taxa; thicker lines are stronger effects. Line type indicates whether the relationship is statistically-significant, i.e. not flat (solid line = $p < 0.01$; dotted line = $p > 0.01$). Predictor variables were scaled from 0 to 1. Predictions for one variable were made while setting the value of all other variables to their median value in the taxon and scale-specific dataset used to fit the GAMM. For visualization purposes, fitted

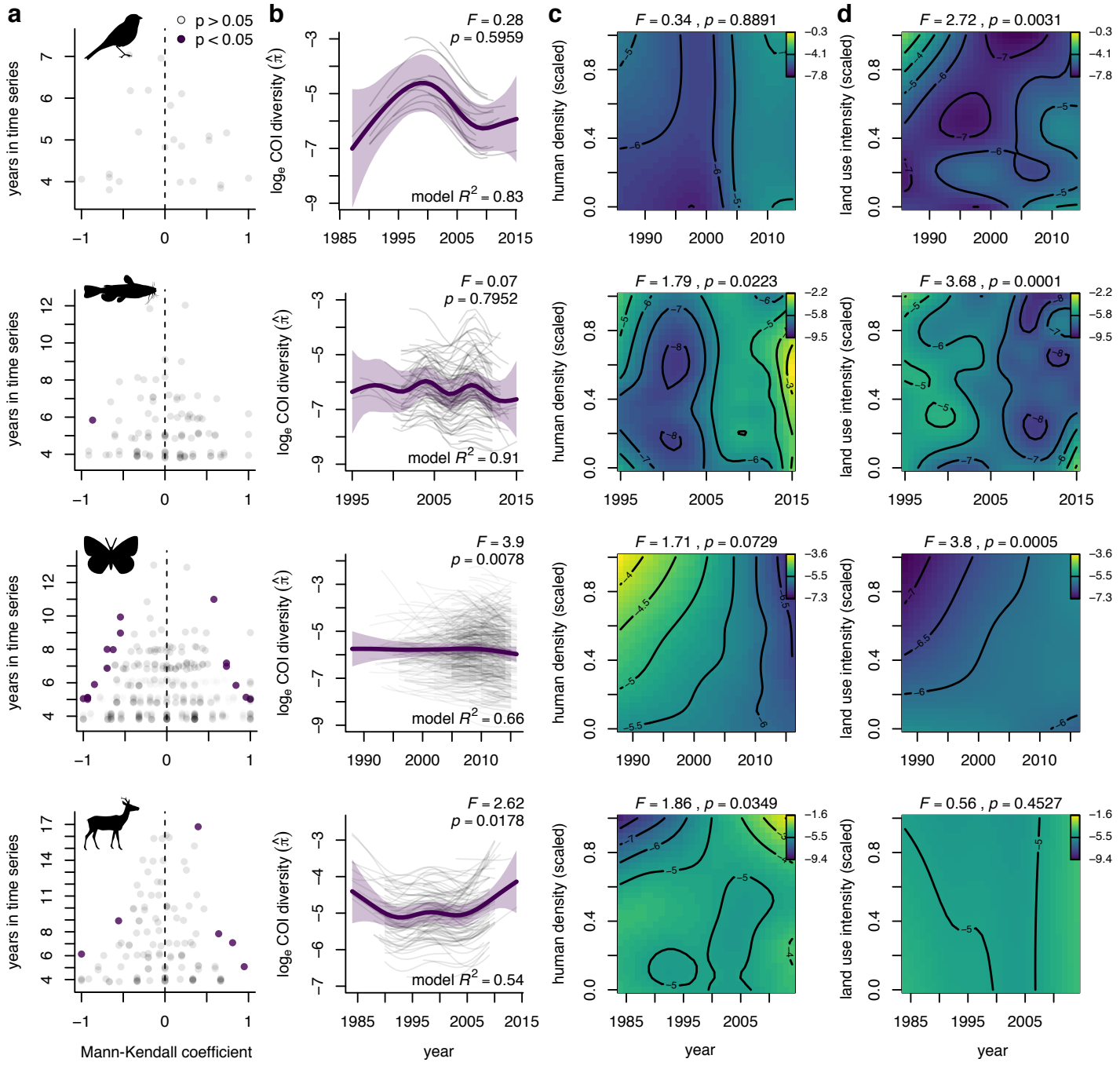
values are shown on the scale of the linear predictor. Additional model details are provided in Figs. S2-S5.

Figure 4. Time series analysis of COI nucleotide diversity in 909 populations of birds, inland and coastal bony fishes, insects, and mammals. Populations were defined at the 1000 km scale and included a minimum of 4 years of data. Panels on the same row belong to the same taxon. (a) Mann-Kendall trend coefficients (τ) of time series vs. the number of years of data that they contain. Purple symbols are time series with statistically-significant negative ($\tau < 0$) or positive ($\tau > 0$) trends in diversity. A small offset was added on the y axis to aid visualization. (b-d) GAMM results showing predicted values of diversity over time. Values are on the scale of the linear predictor. (b) Effect of year on diversity. Thick purple lines indicate overall trends (predicted value ± 95% confidence intervals) while thin grey lines show fitted values for individual populations. *F* and *p* values for the 'time' smooth and the overall model adjusted R^2 are shown. (c, d) Predicted temporal trends in diversity as a function of mean human population density (c) or land use intensity (d) in the HYDE 3.2 grid cells from which population-specific sequences originated. A colour transition from yellow to blue indicates a reduction in nucleotide diversity; for example, in insects (third row, column c), nucleotide diversity in areas of high human population density transitions from high diversity (yellow) to low diversity (blue) over time. The *F* and *p* values of both tensor-product interactions are provided; low *p* values indicate that a given two-dimensional surface is not flat. Note that the scales for diversity and year differ among animal classes.


a
number of sequences
3,000
2,500
2,000
1,500
1,000
500
0
b
30 other orders
Passeriformes
41 other orders
Characiformes
Perciformes
Cypriniformes
Diptera
Lepidoptera
23 other orders
9 other orders
Chiroptera
Rodentia
0
25,000
50,000
75,000
100,000
125,000
150,000
number of sequences
c
number of sequences
10,000
1,000
100
10
1
1985
1995
2005
2015
year
d
human density
100%
10%
1%
0%
fraction
Sequences
World
0
5,000
10,000
15,000
humans (inhabitants km^{-2})
e
land use
100%
10%
1%
0%
fraction
Sequences
World
0
25
50
75
100
landuse (% of gridcell land under heavy anthropogenic use)


smoothed COI diversity ($\hat{\pi}$)
0
> 0.003
smoothed COI diversity ($\hat{\pi}$)
0
> 0.007
smoothed COI diversity ($\hat{\pi}$)
0
> 0.004
smoothed COI diversity ($\hat{\pi}$)
0
> 0.006


a
b
c
d
10 km
100 km
1 000 km
10 000 km
log$_e$ COI diversity ($\hat{\pi}$)
mean spatial distance
absolute latitude
human density
land use intensity


a
b
c
d
p > 0.05
p < 0.05
years in time series
Mann-Kendall coefficient
log$_e$ COI diversity ($\hat{\pi}$)
year
human density (scaled)
land use intensity (scaled)
F = 0.28
p = 0.5959
model R^2 = 0.83
F = 0.34 , p = 0.8891
F = 2.72 , p = 0.0031
F = 0.07
p = 0.7952
model R^2 = 0.91
F = 1.79 , p = 0.0223
F = 3.68 , p = 0.0001
F = 3.9
p = 0.0078
model R^2 = 0.66
F = 1.71 , p = 0.0729
F = 3.8 , p = 0.0005
F = 2.62
p = 0.0178
model R^2 = 0.54
F = 1.86 , p = 0.0349
F = 0.56 , p = 0.4527