Co-limitation towards lower latitudes shapes global forest diversity gradients

The latitudinal diversity gradient (LDG) is one of the most recognized global patterns of species richness exhibited across a wide range of taxa. Numerous hypotheses have been proposed in the past two centuries to explain LDG, but rigorous tests of the drivers of LDGs have been limited by a lack of high-quality global species richness data. Here we produce a high-resolution (0.025° × 0.025°), map of local tree species richness using a global forest inventory database with individual tree information and local biophysical characteristics from ~1.3 million sample plots. We then quantify drivers of local tree species richness patterns across latitudes. Generally, annual mean temperature was a dominant predictor of tree species richness, which is most consistent with the metabolic theory of biodiversity (MTB). However, MTB underestimated LDG in the tropics, where high species richness was also moderated by topographic, soil and anthropogenic factors operating at local scales. Given that local landscape variables operate synergistically with bioclimatic factors in shaping the global LDG pattern, we suggest that MTB be extended to account for co-limitation by subordinate drivers.

Results and discussion
Global patterns of local tree species richness and LDG. Our analyses confirmed, with a high level of accuracy, one general spatial trend in local tree species richness worldwide that has led us to three conclusions regarding the mechanisms underlying patterns of tree species richness. We found that LDG for tree species richness was consistent with that of most other groups of organisms, with a decline from the tropics to the poles (Figs. 2 and 4). In the Northern Hemisphere, tree species richness dropped sharply from the equator (98 species per ha) to 10°N with an average rate of decline of 6 species per ha per 1° increase in latitude, after which the decline diminished and stabilized at 4 species per ha at 50°N. In the Southern Hemisphere, tree species richness declined from the equator to 25°S on average by 3 species per ha per 1° increase in latitude, after which tree species richness fluctuated before another steep drop from 25 species per ha at 43°S to 4 species per ha at 50°S. We were able to detect and map regional patterns and global peaks of tree species diversity, with a high spatial resolution (0.025°×0.025°). The Amazonian, Southeast Asian and Melanesian rainforests are the regions with the greatest local tree species richness worldwide, containing >200 tree species per ha above the 5 cm diameter-at-breast-height (DBH) threshold, confirming previous findings[15,16]. Tropical African rainforests generally contain 50% fewer tree species per hectare than Amazonian rainforests. In the temperate forests of the Northern Hemisphere, the Changbai Mountains in Northeast Asia (up to ~28 species per ha) and the
Central Appalachian forests in the Eastern United States (up to ~20 species per ha) display high local species richness. In the Southern Hemisphere, the sclerophyllous and Nothofagus-dominated forests in south-central Chile are among the most species-rich temperate communities (up to 50 species per ha). Boreal forest communities are consistently low in local tree species richness, with typically five or fewer tree species per hectare.

The above LDG pattern of tree species richness was generally consistent with the metabolic theory of biodiversity (MTB)\(^1\)\(^9\)\(^,\)\(^2\)\(^2\), except at low latitudes (Fig. 5). According to MTB, environmental temperature is largely responsible for the generation and maintenance of biodiversity\(^2\)\(^1\)\(^,\)\(^2\)\(^1\)\(^,\)\(^2\)\(^2\), and the natural logarithm of species richness is linearly associated with \(1,000T^{-1}\), where \(T\) is the absolute environmental temperature in Kelvin (mean annual temperature +273.15K), with a slope ranging from \(-7.5\) K to \(-9.0\) K. Our global tree species richness gradient was largely consistent with MTB, with a slope of \(-8.0\) K (\(P<0.001\)) and a coefficient of determination of 0.82 (Metabolic Theory of Biodiversity in Methods), indicating that environmental temperature is generally a good predictor of LDG. However, at low latitudes, MTB substantially underestimated LDG.

In fact, near the equator where the actual LDG peaked (98 species per ha), observed tree species richness was almost twice as high as predicted by MTB (56 species per ha) (Fig. 4a). Our results suggest that within this low latitudinal range, other factors are also important to the maintenance of biodiversity.

The under-estimation of local tree species richness by MTB at low latitudes is attributable, in part, to the lack of a definite dominant environmental factor, suggesting a co-limitation of multiple subordinate drivers at low latitudes (Fig. 5). In general, bioclimatic factors predominantly determined species richness in 82.6% of the forested areas, while co-limitation (that is, absence of any dominating factor) occurred in 11.7% of forested areas globally. However, in the low-latitude range between 5° N and 15° S, the percentage area of co-limitation increased to 37.1%, more than three times the global average. Furthermore, forested areas under co-limitation contained on average 81.1 ± 0.1 species per ha, much higher than the average local tree species richness of forested areas predominantly determined by topographic (43.9 ± 0.1), anthropogenic (35.6 ± 0.2), soil (33.9 ± 0.2) and bioclimatic (19.4 ± 0.02) factors (Fig. 5b). This suggests that the pattern of co-limitation is pervasive...
in species-rich tropical forests. In South America, transitional areas between Amazonia and savanna formations nearby are subject to co-limitation that is partly attributable to a dynamic equilibrium between closed forest and savanna, edaphic conditions and natural fire regimes. In Africa, anthropogenic influences such as selective timber extraction and fuelwood collection, together with large-scale degradation affect local tree species richness. In Central Africa, the evolution of anthropogenic influences from prehistoric to present times has imposed a substantial effect on species diversity and resulted in the development of a complex system of mixes with light-demanding and old-growth tree species.

Bioclimatic dominance and co-limitation. In addition to an overall positive response of local tree species richness to the rise of annual mean temperature (partial dependence plot of C1 in Fig. 3 and Extended Data Fig. 3), the importance of environmental temperature (2.7%) was topped by the total annual precipitation (C12, 7.6%) (Fig. 3). Our findings are consistent with previous discoveries of a joint role of water and temperature/energy—as a proxy for net primary productivity—on plant species richness, with water dominating particularly at warmer, lower latitudes. Predicted tree species richness accelerated exponentially with temperature and rainfall, although independently, as shown in the cold–dry quadrant and the convex contours of the 2D partial dependence plot (Extended Data Fig. 3), until each has reached its respective threshold (1,500 mm for total annual precipitation and 10°C for annual mean temperature). Beyond one of these thresholds, species richness is limited only by the predictor below its threshold (that is, by annual mean temperature in the cold–wet quadrant or by annual precipitation in the hot–dry quadrant). When both predictors have reached their thresholds, that is, in the hot–wet quadrant, co-limitation predominates in most tropical forests. Net primary productivity in the tropics, thus, requires co-limitation of other factors besides only temperature and rainfall. As the response of carbon flux mirror the low-latitude co-limitation pattern for tree species diversity, the matching determinants for both diversity and productivity may explain the similar latitudinal gradient in productivity and the positive diversity–productivity relationship.

Our findings also indicate that under climate change, intensified droughts coupled with increased annual mean temperature can potentially trigger declines of tree species richness, although possible increases in water-use efficiency from elevated CO2 and the dominance of highly contingent co-limiting factors may partially buffer this effect in the tropics.

Here we articulate evidence for co-limitation in LDG. Resource co-limitation is a common concept in ecology (for example, refs. 44–48), often used to describe how the synergistic interactions of two or more factors limit ecological productivity. Our use of the term co-limitation emphasizes the reduced effect of bioclimate on tree species richness at low latitudes, although bioclimate is the globally predominant driver of species richness, recognizing that several local subordinate factors synergistically contribute to increased tree species richness in this latitudinal range. We thus argue that the inclusion of co-limitation could substantially improve the explanatory power of biodiversity models in estimating alpha diversity by considering multiple subordinate factors where single-factor dominance is lacking, especially in the tropics. At high latitudes, bioclimatic conditions, particularly environmental temperature, are the major limiting factors and thus the dominant drivers of tree species diversity. As the latitude declines, the influence of bioclimatic conditions dwindles and the maintenance of tree species richness is moderated by many interacting drivers without a clear dominance, which is especially well expressed between 5°N and 15°S (Fig. 5). This prevalence of co-limiting factors is thus not a mere coincidence as to why the observed LDG at low latitudes is almost double that predicted by MTB (Fig. 2). While each of the existing hypotheses underpinning LDG addresses a certain process (for example, selection, drift, dispersal or speciation), the evidence of co-limitation highlights synergistic interactions of local processes across the latitudinal gradient.

Concluding remarks. More research is needed to fully elucidate patterns of LDG driven by climatic and other influences, especially those outlined in competing hypotheses. First, our analyses lack explicit consideration of some evolutionary, ecological and historical factors. These include mid-domain stochastic effects, the legacies of the poleward expansion of tree species after the Last Glacial Maximum and recent human land use/management. Alternative hypotheses, such as niche conservatism or climatic history, are more difficult to test due to data limitations. In addition, long-term effects at geological and millennial time scales also play a role, but it is difficult to disentangle these effects due to collinearity. A major source of uncertainty in our results (Fig. 4b) came from an uneven sample coverage between developed and developing countries (Extended Data Fig. 1). To address this gap, we argue that there needs to be a shared responsibility among forestry agencies at various levels of government, scientists, indigenous communities and other biodiversity monitoring groups to improve sample coverage of forest inventories in developing countries. Innovative biodiversity funding mechanisms, for example, forest inventories funded by carbon
initiatives such as REDD+, should be incorporated into a comprehensive global forest biodiversity database. Meanwhile, the severe shortage of experts and database management infrastructures, especially in developing countries, poses another major challenge to address this gap. The education and training of new generations of forest scientists, taxonomists and foresters can bring tangible benefits to biodiversity monitoring while improving local economies as well.

Considering co-limitation in addition to MTB enables a refined description of the biogeographic distribution of biodiversity and mechanisms underlying LDG. Our analysis has resulted in the production of a high-resolution map of tree species richness across the global forest range, along with visuals of those factors responsible for the moderation of local tree species richness. Such tools are necessary for conservation management which requires assessments of factors responsible for biodiversity patterns at multiple scales that matter—from local and regional to global scales. Patterns of local tree species richness and associated drivers may provide insights into how and why the diversity of other forest flora, fauna and microbes vary across space and time. Furthermore, the high-resolution map of local tree species richness presented here provides a benchmark for evaluating the impact of biodiversity loss on the productivity and functioning of forest ecosystems. Finally, aligned with current international calls for spatially explicit monitoring of ecosystem attributes, this study delivers detailed biogeographic information to support international endeavours focused on valuing natural capital and advancing global conservation.

Methods

As illustrated in Fig. 1, we conducted data analyses and modelling in three stages.

Stage 1. For this study, we compiled individual in situ tree data from all the regional and national GFBi forest inventory datasets (Supplementary Table 2) into a standardized GFBi dataframe, that is, the GFBi tree list. In this standardized GFBi dataframe, each row represents an individual tree, and columns represent nine key tree- and plot-level attributes. These attributes are tree ID (FID), a unique number assigned to each individual tree; plot ID (PLT), a unique string assigned to each plot; plot coordinates (LAT and LON); tree species name (SPCD); DBH or above buttress; year of measurement; and dataset name (DSN), a unique number to each plot; plot coordinates (LAT and LON); tree species name (SPCD); DBH or above buttress; year of measurement; and dataset name (DSN), a unique number assigned to each forest inventory dataset (Supplementary Table 2). With a total of 56 million trees surveyed, GFBi individual-based dataframe represents 1/50,000 of the approximately 2.7 trillion trees worldwide. Because all trees in each sample plot were identified and measured, GFBi data make it possible to quantify forest community structure, composition and species distribution.

To ensure consistency and maximum accuracy in species names, we standardized observations from different forest inventory datasets with the following protocol. First, all multi-stem trees were divided so that each stem represents an individual tree. The scientific names were extracted from original datasets, keeping only the genus and species (authority names were removed). Next, tree species names were compiled into five general species lists, one for each continent. We matched individual species names against 23 online taxonomic resources—nine tree and 14 species lists, each representing one tree genus, and columns represent nine key tree- and plot-level attributes. These attributes are tree ID (FID), a unique number assigned to each individual tree; plot ID (PLT), a unique string assigned to each plot; plot coordinates (LAT and LON); tree species name (SPCD); DBH or above buttress; year of measurement; and dataset name (DSN), a unique number assigned to each forest inventory dataset (Supplementary Table 2). With a total of 56 million trees surveyed, GFBi individual-based dataframe represents 1/50,000 of the approximately 2.7 trillion trees worldwide. Because all trees in each sample plot were identified and measured, GFBi data make it possible to quantify forest community structure, composition and species distribution.

To ensure consistency and maximum accuracy in species names, we standardized observations from different forest inventory datasets with the following protocol. First, all multi-stem trees were divided so that each stem represents an individual tree. The scientific names were extracted from original datasets, keeping only the genus and species (authority names were removed). Next, tree species names were compiled into five general species lists, one for each continent. We matched individual species names against 23 online taxonomic resources—nine tree and 14 species lists, each representing one tree genus, and columns represent nine key tree- and plot-level attributes. These attributes are tree ID (FID), a unique number assigned to each individual tree; plot ID (PLT), a unique string assigned to each plot; plot coordinates (LAT and LON); tree species name (SPCD); DBH or above buttress; year of measurement; and dataset name (DSN), a unique number assigned to each forest inventory dataset (Supplementary Table 2). With a total of 56 million trees surveyed, GFBi individual-based dataframe represents 1/50,000 of the approximately 2.7 trillion trees worldwide. Because all trees in each sample plot were identified and measured, GFBi data make it possible to quantify forest community structure, composition and species distribution.

To ensure consistency and maximum accuracy in species names, we standardized observations from different forest inventory datasets with the following protocol. First, all multi-stem trees were divided so that each stem represents an individual tree. The scientific names were extracted from original datasets, keeping only the genus and species (authority names were removed). Next, tree species names were compiled into five general species lists, one for each continent. We matched individual species names against 23 online taxonomic resources—nine tree and 14 species lists, each representing one tree genus, and columns represent nine key tree- and plot-level attributes. These attributes are tree ID (FID), a unique number assigned to each individual tree; plot ID (PLT), a unique string assigned to each plot; plot coordinates (LAT and LON); tree species name (SPCD); DBH or above buttress; year of measurement; and dataset name (DSN), a unique number assigned to each forest inventory dataset (Supplementary Table 2). With a total of 56 million trees surveyed, GFBi individual-based dataframe represents 1/50,000 of the approximately 2.7 trillion trees worldwide. Because all trees in each sample plot were identified and measured, GFBi data make it possible to quantify forest community structure, composition and species distribution.

To ensure consistency and maximum accuracy in species names, we standardized observations from different forest inventory datasets with the following protocol. First, all multi-stem trees were divided so that each stem represents an individual tree. The scientific names were extracted from original datasets, keeping only the genus and species (authority names were removed). Next, tree species names were compiled into five general species lists, one for each continent. We matched individual species names against 23 online taxonomic resources—nine tree and 14 species lists, each representing one tree genus, and columns represent nine key tree- and plot-level attributes. These attributes are tree ID (FID), a unique number assigned to each individual tree; plot ID (PLT), a unique string assigned to each plot; plot coordinates (LAT and LON); tree species name (SPCD); DBH or above buttress; year of measurement; and dataset name (DSN), a unique number assigned to each forest inventory dataset (Supplementary Table 2). With a total of 56 million trees surveyed, GFBi individual-based dataframe represents 1/50,000 of the approximately 2.7 trillion trees worldwide. Because all trees in each sample plot were identified and measured, GFBi data make it possible to quantify forest community structure, composition and species distribution.

To ensure consistency and maximum accuracy in species names, we standardized observations from different forest inventory datasets with the following protocol. First, all multi-stem trees were divided so that each stem represents an individual tree. The scientific names were extracted from original datasets, keeping only the genus and species (authority names were removed). Next, tree species names were compiled into five general species lists, one for each continent. We matched individual species names against 23 online taxonomic resources—nine tree and 14 species lists, each representing one tree genus, and columns represent nine key tree- and plot-level attributes. These attributes are tree ID (FID), a unique number assigned to each individual tree; plot ID (PLT), a unique string assigned to each plot; plot coordinates (LAT and LON); tree species name (SPCD); DBH or above buttress; year of measurement; and dataset name (DSN), a unique number assigned to each forest inventory dataset (Supplementary Table 2). With a total of 56 million trees surveyed, GFBi individual-based dataframe represents 1/50,000 of the approximately 2.7 trillion trees worldwide. Because all trees in each sample plot were identified and measured, GFBi data make it possible to quantify forest community structure, composition and species distribution.

To ensure consistency and maximum accuracy in species names, we standardized observations from different forest inventory datasets with the following protocol. First, all multi-stem trees were divided so that each stem represents an individual tree. The scientific names were extracted from original datasets, keeping only the genus and species (authority names were removed). Next, tree species names were compiled into five general species lists, one for each continent. We matched individual species names against 23 online taxonomic resources—nine tree and 14 species lists, each representing one tree genus, and columns represent nine key tree- and plot-level attributes. These attributes are tree ID (FID), a unique number assigned to each individual tree; plot ID (PLT), a unique string assigned to each plot; plot coordinates (LAT and LON); tree species name (SPCD); DBH or above buttress; year of measurement; and dataset name (DSN), a unique number assigned to each forest inventory dataset (Supplementary Table 2). With a total of 56 million trees surveyed, GFBi individual-based dataframe represents 1/50,000 of the approximately 2.7 trillion trees worldwide. Because all trees in each sample plot were identified and measured, GFBi data make it possible to quantify forest community structure, composition and species distribution.
the Congo named ‘CDi’. To maximize our species coverage, a tree was defined in this study as a perennial plant with an elongated woody stem that supports branches and leaves, including woody angiosperms, gymnosperms and taller palms (Areaceae). Tree ferns (Cyatheales) and bamboos (Bambusoideae) were excluded from our analysis.

From the GFBi individual tree-level dataframe, we derived a global species abundance matrix. The global species abundance matrix consisted of the number of individuals by species (column vectors) within individual sample plots (row vectors). The global species abundance matrix consisted of two complementary datasets: Phase I dataset contained 1,255,444 sample plots and Phase II dataset contained 22,131 sample plots, most of which are located in unsampled and under-sampled regions of Phase I dataset. Phase I sample plots cover 394 ecoregions across the world, and Phase II sample plots cover an additional 30 ecoregions in Africa, South America, Southeast Asia, Mexico, India and Japan. Together, our ground-based forest sample plots cover 424 of 435 (97.5%) forested ecoregions across the world. The global species abundance matrix contains ~1.3 million rows (plots) by 32,608 columns (species). Key plot-level information was added to the matrix, including PLT, DSN, plot coordinates, basal area (B), the total cross-sectional areas (m²) of living trees per ha calculated from DBH and TPH (expansion factor) and the year of measurement. TPH denotes the number of trees per hectare represented by each sampled individual. It ranged from 1 to 5,244 across the GFBi data, with a mean of 48 trees per ha.

We quantified for each sample plot tree species richness ($S$), which is the total number of tree species in a community. Due to the difference in plot size (standard deviation = 0.09 ha) and threshold DBH values (standard deviation = 2.52 cm) across GFBi sample plots, we developed machine learning models to standardize tree species richness for a common basis of 1 ha in area and 5 cm in threshold DBH. The models incorporated both plot area ($A$) and threshold DBH ($D$) as predictors to account for the underlying species-area relationship\textsuperscript{50–52} and species-individual size distribution\textsuperscript{53} in a rarefaction-based approach\textsuperscript{54}. This standardization approach justifies compiling direct tree species diversity estimates from GFBi in situ data of different sources and sampling protocols\textsuperscript{55–57}, an issue highlighted in earlier large-scale—although less extensive—forest biodiversity studies\textsuperscript{57,58}. To evaluate the accuracy of this standardization approach, we tested the machine learning models using cross-sample validation and compared our global maps of estimated tree species diversity against other standardization approaches based on sample completeness (Model Evaluation below).

The machine learning models employed 47 environmental covariates to predict tree species richness. These covariates, derived from satellite-based remote sensing and ground-based survey data, can be summarized into five general categories: bioclimatic (for example, annual mean temperature, total annual precipitation, potential evapotranspiration and indexed annual aridity); soil (bulk density, pH, electrical conductivity, C/N ratio and total nitrogen); topographic, including elevation, slope, aspect and terrain features; vegetation and survey attributes.
For rigorous model evaluation, we employed three very different cross validation approaches: randomized cross validation (RCV), spatial cross validation (SCV) and post-sample validation (PSV). In RCV, a model was trained for each continent with a random subsample that accounted for 90% of the training data from that continent, and the remaining 10% of the training data were used as the testing set. This process was repeated 20 times with sample replacement to examine the accuracy of estimated tree species diversity values. In SCV, all sample data from an ecoregion were reserved for testing the model that was trained with the remaining samples from the larger continent within which the ecoregion is situated. We decided to use ecoregions as spatial blocks because (1) unlike political units such as countries and provinces, ecoregions are delineated based on ecological and bioclimatic conditions; and (2) with a total of ~700 terrestrial ecoregions across the world, each ecoregion encompasses 1,800 sample plots on average, which is a large enough sample size for training RF models. This process was repeated until all the forested ecoregions across the world had been tested. SCV was more rigorous than RCV because samples from an entire ecoregion rather than random samples were withheld for validation. PSV was the most rigorous among the three validation processes. For PSV, we have collated an independent sample dataset from 22,131 forest sample plots, which we named Phase II sample plots to highlight their independence from the original GFBi dataset (that is, Phase I sample plots). In PSV, we used Phase II data as the testing set to evaluate the accuracy of the predictive models that were trained for each continent with the Phase I data.

Using these three cross validation processes, we also evaluated the performance of the RF model against two other predictive models, including multiple regression with ordinary least squares (OLS) and Extreme Gradient Boosting (XGBoost). For each model, we derived predicted values of tree species richness of the testing sets and compared these predicted values against observed data using mean absolute error, root-mean-squared error (RMSE), and coefficient of determination (R²) (ref. 15). The process was repeated 20 times to select the best model for each continent.

The OLS model estimated values of standardized point diversity for non-sampled point location \(s\), based on spatially explicit values of covariates:

\[ Y(s) = \alpha \cdot X(s) + e(s), \]

where \(Y(s)\) is tree species richness at location \(s\); \(X\) a design matrix for the predictor variables at location \(s\); \(\alpha\) is a vector of coefficients; and \(e\) is a random vector.
Fig. 6 | Patterns and variance of local tree species richness per ha by continent. The collage of maps shows the zoomed-in view of the distribution of predicted local tree species richness per ha (Fig. 4a) by continent. Circular Venn diagrams (with the legend in the centre) show, for each continent, the spatial variance in observed tree species richness partitioned as follows: a (mean = 14.3%) represents the fraction of variance uniquely explained by environmental factors (that is, bioclimatic, topographic, anthropogenic and soil variables) after latitudinal effects had been accounted for. b (mean = 68.2%) stands for the fraction of variance jointly explained by environmental factors and latitudinal effects. c (mean = 0.3%) represents the fraction of variance explained by latitudinal effects after removing environmental effects. d (mean = 17.2%) represents the fraction of unexplained variance in tree species richness. The fractions were based on contrasting the amount of local richness variations in sample data from ~1.3 million plots explained by the $R^2$ statistics from the continental-scale random forest models with the full set of factors versus those with targeted factors removed.

Following a Gaussian probability density function, with an expected value of zero and variance of $\sigma^2$. Spatial autocorrelation was not accounted for here due to computational limitations. GFBi data collected from sample plots of various sizes were harmonized to represent local forest community populations per ha using computational limitations. GFBi data collected from sample plots of various sizes using step-wise regression and Akaike information criterion.

As a result, we obtained a preliminary RF model. Because the RF model emerged as the most accurate model from all three cross validation processes (Extended Data Fig. 2), we selected the RF as the final model, and re-calibrated the final RF model using all the sample data (Phase I and Phase II data).

Stage 3. Global map of local tree species richness. To map community-level tree species richness over the global forest range, we first derived the global forest range map from version 1.3 of the Global Forest Change database (years 2000–2015). To ensure consistency with the definition of forest by the Food and Agriculture Organization of the United Nations (FAO), the global forest range in this study was defined as forested areas with $\geq$10% tree crown coverage per unit area. The titled ‘treecover2000’, ‘loss’ and ‘gain’ datasets were integrated to obtain current forest cover estimates for the year 2015. To minimize processing artefacts, the ~1 arcsecond spatial resolution tiles were spatially aggregated to an even multiple of their native resolution that approximated the resolution of our covariates. The datasets were then converted to vector point files before being reconverted to raster format with the exact resolution and origin of our covariates. After mosaicking each set of tiles, we computed ‘treecover’ (scaled) – ‘loss’ + ‘gain’ to obtain the 2015 global forest cover, represented as
percent forest cover per ~30 arcsecond pixel. Artefacts in the original data led to 0.08% of all terrestrial pixels having forest cover estimates greater than 100% and 1.9% of terrestrial pixels having estimates less than 0%. These values were truncated to 100% and inflated to 0%, respectively. Finally, the global forest cover range consisted of those pixels with a percent forest cover ≥10% in 2015. In total, each map consisted of 9,944,908 pixels of 0.025° × 0.025° (hereafter, the pixel) of forested areas. This range is rather conservative and potentially underestimates many remnant forests in drylands and grasslands75.

We then estimated tree species richness at a 1 ha scale for all pixels within a continent based on the final RF model trained for that continent, using both Phase I and Phase II data. Spatially explicit local environmental covariate data across the global forest range were used for the imputation, except that plot size and threshold DBH were set as 1 ha and 3 cm, respectively. For ecoregions with extremely low sample coverage, we further fine tuned the RF model using samples of similar environment characteristics from other continents. More specifically, we first identified two ecoregions of extremely low sample coverage, that is, the temperate forests in South America and the tropical forests in Oceania, as there were fewer than 1,000 sample plots for the entire biome on those continents. We then trained a new RF model for each ecoregion, using all the sample data from the same biome across the world and fine tuned the mapping data for that ecoregion using the biome-specific RF model.

We computed and mapped the width of the 95% confidence interval for our local estimates of tree species richness per ha across the global forest range. To this end, we employed a rigorous spatial-block approach, analogous to the spatial cross validation, to derive the 95% confidence interval. More specifically, we computed the width of the 95% confidence interval for each 0.025° × 0.025° mapping pixel by ecoregion. For a pixel in ecoregion e, we trained 20 RF models using random subsamples that accounted for 90% of the training data from the same continent, which included all samples except those from ecoregion e. We then derived the standard error and the width of the 95% confidence interval for this pixel p in ecoregion e from the predictions of the 20 RF models trained for this ecoregion. This process was repeated until all the forested ecoregions across the world had been assessed and mapped.

Uncertainty in our global diversity estimates was caused by two types of error. The first was measurement error from in situ forest inventories. We mitigated this type of error by implementing stringent species name check and data standardization protocols (Stage 1 Data Standardization). The second arose from the imputation process to map tree species diversity. We minimized this type of error using the three cross validation approaches introduced in Stage 2.

**Metabolic theory of biodiversity.** Using the global standardized tree species richness values predicted from the final RF models, we quantified the global LDG of tree species richness and tested the effect of environmental temperature based on the MTB11:

\[
\ln(S) = \frac{1000}{T_{an}} + \beta, 
\]

where 5 represents species richness and \(T_{an}\), here represents absolute environmental temperature (mean annual temperature +273.15 K); \(\alpha\) and \(\beta\) represent coefficients to be estimated by ordinary least squares. According to both the original and extended MTB11, the slope \(\alpha\) is expected to range between ~7.5 K and ~9.0 K, under the assumption that tree community abundance per area does not vary with latitude.

**Variance partitioning.** We used variance partitioning3 based on the sample data from ~1.3 million plots to quantify the unique and joint fractions of spatial variance in tree species richness explained by environmental factors and latitude. Due to the correlation between species and environment and between the spatially explicit environmental factors, the variance partitioning approach mitigates type I error inflated by spatial autocorrelation12. With variance partitioning, we tested the significance of environmental effects on tree species richness in a series of nested RF models. (A) The full model (Extended Data Fig. 4a) consisted of latitude and 47 environmental variables (including 21 bioclimatic ones). (B) The reduced model I (Extended Data Fig. 4b) consisted of all but the 21 bioclimatic variables. (C) The reduced model II (Extended Data Fig. 4c) consisted of only a zero constant. The overall significance of all environmental factors plus latitude was tested in an one-tailed F-test by comparing the residual sum of squares of error (RSS) of model (A) and model (C):

\[
F = \frac{RSS_A - RSS_S}{RSS_S}, 
\]

where \(n - n_c\) and \(n_c - n_{sc}\) stand for the degree of freedom for the full model and the difference in the degrees of freedom between the full model and the reduced model II, respectively.

The significance of bioclimatic factors, with the effect of latitude being controlled, was tested in an one-tailed F-test by comparing RSS of model (A) and model (B):

\[
F = \frac{RSS_A - RSS_S}{RSS_S}. 
\]

where \(n_c - n_{sc}\) = 21 stands for the difference in the degrees of freedom between the full model and the reduced model I.

We partitioned the spatial variance in observed species richness into four components: \(a\) represents the fraction of variance uniquely explained by environmental factors (that is, bioclimatic, topographic, anthropogenic and soil variables), after latitudinal effects have been taken into account; \(b\) represents the fraction of variance jointly explained by environmental factors and latitudinal effects; \(c\) represents the fraction of variance explained by latitudinal effects after removal of environmental effects; and \(d\) represents the fraction of variance not explained by the full RF model. Then, the total fraction of variance explained by both environmental factors and latitude was \(a + b + c\), the fraction of variance explained by environmental factors was \(a + b\), and the fraction of variance explained by latitude was \(b + c\). Components \(a + b + c\) and \(b + c\) were estimated by the \(R^2\) statistics from the RF models trained for each ecoregion using all factors, environmental factors and latitude, respectively (Stage 2 Model Training and Evaluation). Components \(a\), \(b\) and \(c\) were computed from the previous components using arithmetic relationships that ensure that \(a + b + c + d = 100\%\).

**Model sensitivity.** Based on the final RF models and sample data from ~1.3 million plots, we mapped the dominant drivers of tree species richness with a 0.025° × 0.025° resolution (that is, global map of co-limitation), following a standard procedure for model sensitivity analysis9:

Step 1: using the full RF model, and the values of environmental factors \(X(s)\) specific to a 0.025° pixel \(s\), we had already estimated local tree species richness \(S_{null}(s)\):

\[
S_{null}(s) = f_{null}(X(s)). 
\]

where \(f_{null}\) represents the RF model, and \(X(s)\) environmental factors in four categories, namely E1: bioclimatic, E2: topographic, E3: anthropogenic and E4: soil.

Step 2: for the above-mentioned pixel, we estimated a new local tree species richness value \(S_{E1} - S_{null}(s)\), using a reduced RF model in which all E1 (bioclimatic) variables were removed:

\[
S_{E1} - S_{null}(s) = f_{E1}(X(s)) + s, 
\]

where \(f_{E1}(\cdot)\) represents the RF model trained with all but 21 bioclimatic variables, and \(X(s)\) encompasses environmental factors in three categories, namely E2: topographic, E3: anthropogenic and E4: soil.

Step 3: for a given pixel, we calculated the relative sensitivity of predicted species richness to E1:

\[
R(E1) = \frac{S_{null}(s) - S_{E1} - S_{null}(s)}{S_{null}(s)} \times 100\%. 
\]

Step 4: we repeated steps 2 and 3 to calculate, for a given pixel, the relative sensitivity of each of the following categories (that is, E2: topographic, E3: anthropogenic and E4: soil), respectively. The dominant driver (that is, limiting factor) for this pixel was then the category with the highest relative sensitivity, provided that this relative sensitivity was greater than or equal to 0.2.

Step 5: if the relative sensitivities were less than 0.2 for all categories, we considered that this was a scenario of joint effects of multiple categories of factors (that is, co-limitation), rather than dominance of a single category. Where clear dominance of a single category was lacking, we denoted the dominant driver of this pixel as ‘E5: co-limitation.’

Step 6: we repeated the steps above to calculate, for all the remaining pixels of the global grid, the relative sensitivity of each of the five categories of environmental factors, namely E1: bioclimatic, E2: topographic, E3: anthropogenic, E4: soil and E5: co-limitation. On the basis of these values, we created a wall-to-wall map of dominant drivers of tree species richness across the global forest range by labelling the category with the highest relative sensitivity for each pixel (Fig. 5b).

Step 7: on the basis of the relative sensitivity obtained from the steps 1–6, we computed percent prevalence (0–100%) of bioclimatic, topographic, anthropogenic and soil factors and a lack of dominance (co-limitation) in all the forested pixels along each latitudinal band.

**Inclusion and ethics statement.** The international research collaboration leading to this research paper was conducted via Science-i.org, a transparent and FAIR (Findable, Accessible, Interoperable and Reusable) web platform for international research collaboration. Through this platform and our partner initiatives including the Global Forest Biodiversity Initiative (GFBI), we pursue excellence in high standards of performance, professionalism and ethical conduct. Science-i strictly prohibits any form of discrimination against individuals on the basis of gender, race, age, religion, sexual orientation, veteran status or disability status. Science-i continuously seeks and encourages underrepresented and underprivileged people and groups and the unique voices in global scientific research collaboration.

**Reporting summary.** Further information on research design is available in the Nature Research Reporting Summary linked to this article.
Data availability
The global map of tree species richness is available under license CC BY 4.0, with the identifier: 10.6084/m9.figshare.1723439. This map can be downloaded in two formats. One is a geotiff file (S_mean.raster.tif) containing the fully geo-referenced map of tree species richness worldwide at a 0.025°×0.025° resolution. The other is a comma-separated file (S_mean.grid.csv) with the following attributes: S is local average tree species richness per ha and x, y are centroid coordinates of all the species

1. References
Acknowledgements
The work presented in this thesis and manuscript development are supported by the web-based team science platform: science-iorg, with the project number 2022SFGFB2. We thank the following initiatives, agencies, teams and individuals for data collection and other technical support: the Global Forest Biodiversity Initiative (GBFI) for establishing the data standards and collaborative framework; United States Department of Agriculture, Forest Service, Forest Inventory and Analysis (FIA) Program; University of Alaska Fairbanks; the SODFOR, Ivory Coast; University Felix Hoenekoop (UFB); the Queensland Herbarium and past Queensland Government Forestry and Natural Resource Management departments and staff for data collection for over seven decades; and the National Forestry Commission of Mexico (CONAFOR). We thank M. Baker (Carbon Tanzania), together with a team of field assistants (Valentine and Lawrence); all persons who made the Third Spanish Forest Inventory possible, especially the main coordinators of the project: AFITRON and the Third Spanish Forest Inventory (NFI campaigns (raw data 2005 and following annual surveys), were downloaded by GBFI at https://inventario-forestier.ign.fr/es/php/brique195; site accessed on 1 January 2015); the Italian Forest Inventory (NFI campaigns raw data 2005 and following surveys were downloaded by GBFI at https://inventarioforestale.org/; site accessed on 27 April 2019); Swiss National Forest Inventory, Swiss Federal Institute for Forest, Snow and Landscape Research (WSL) and Federal Office for the Environment (FOEN), Switzerland; the Swedish NFI, Department of Forest Resource Management, Swedish University of Agricultural Sciences SLU; the National Research Foundation (NRF) of South Africa (89967 and 109244) and the South African Research Chair Initiative (DST); the Danish National Forestry, Department of Geosciences and Natural Resource Management, UCPH; the Federal Support for Indigenous Education Personnel of Brazil (CAPES, grant number 88881.064976/2014-01); R. Ávila and S. van Tuylen, Instituto Nacional de Bosques (INAB), Guatemala, for facilitating Guatemalan data; the National Focal Center for Forest condition monitoring of Serbia (NFC), Institute of Forestry, Belgrade, Serbia; the Thünen Institute of Forest Ecosystems (Germany) for providing National Forest Inventory data; the FAO and the United Nations High Commissioner for Refugees (UNHCR) for undertaking the SAFE (Safe Access to Fuel and Energy) and CBIF-Project forests; and the Amazon Forest Inventory Network (RAINFOR), the African Tropical Rainforest Observation Network (AfTRINFono) and the ForestPlots.net initiative for their contributions from Amazonian and African forests. The Natural Forest plot data collected between January 2009 and May 2011 for the LUCAS Programme for the New Zealand Landscape; the New Zealand Environment are provided by the New Zealand National Vegetation Survey Databank https://nvs.landcareresearch.co.nz/. We thank the International Boreal Forest Research Association (IBRA); the Forestry Corporation of New South Wales, Australia; the National Forest Directory of the Ministry of Environment and Sustainable Development of the Argentine Republic (MAyDS) for the plot data of the Second National Forest Inventory (DNBN2); the National Forestry Authority and Ministry of Water and Environment of Uganda for their National Biomass Survey (NBS) dataset; and the Sabah Biodiversity Council and the staff from Sabah Forest Research Centre. All TEAM data are provided by the Tropical Ecology Assessment and Monitoring (TEAM) Network, a collaboration between Conservation International, the Missouri Botanical Garden, the Smithsonian Institution and the Wildlife Conservation Society, and partially funded by these institutions, the Gordon and Betty Moore Foundation and other donors. We are thankful to all current and previous TEAM site manager and other collaborators that helped collect data. We thank the people of the Rediroti, Pierrekonde and Cassipora village who were instrumental in assisting with the collection of data and sharing local knowledge of their forest and the dedicated members of the field crew of Kabo 2012 census. We are also thankful to FAPEG, SFB, FAO and IMAFSC for supporting the IFFSC. This research was supported in part through computational resources provided by Information Technology at Purdue, West Lafayette, Indiana. This work is supported in part by the NASA grant number 12000401 ‘Multi-sensor biodiversity framework developed from bioacoustic and space based sensor platforms’ (J. Liang, B.P.); the USDA National Institute of Food and Agriculture McIntire Stennis projects 101771 (J. Liang) and 101776 (M.Z.); the US National Science Foundation Biological Dynamics of Forests (DBI) Grants NSF-DBI-2021889 (P.R.B.); the funding by H2020 VERIFY (contract 776810) and H2020 Resonate (contract 101000574) (G.-J.N.); the TEAM project in Uganda supported by the Moore foundation and Buffett Foundation through Conservation International (CI) and Wildlife Conservation Society (WCS); the Danish Council for Independent Research (HANAL); the Natural Sciences (THEF) (grant number 00078B) and VILLUM FONDEN grant number 16549 (J.-C.S.); the Natural Environment Research Council of the UK (NERC) project NE/T011084/1 awarded to J.A.-G. and NE/ S011811/1; ERC Advanced Grant 291585 (T-FORECES) and a Royal Society-Wolson Research Merit Award (O.L.P.); RAINFOR plots supported by the Gordon and Betty Moore Foundation and the UK Natural Environment Research Council, notably NERC Consortium Grants AMAZONICA (NE/F005806/1), ‘TROBIF’ (NE/D005590/1) and ‘BIO-RED’ (NE/NI012542/1); CIFORS Global Comparative Study on REDD+ funded by the Norwegian Agency for Development Cooperation, the Australian Department of Foreign Affairs and Trade, the European Union, the International Climate Initiative (IKI) of the German Federal Ministry for the Environment, Nature Conservation, Building and Nuclear Safety and the CGIAR Research Program on Forests, Trees and Agroforestry (CRP-FTF); the donor fund to the CGIAR-Research Program (Africa); the funding from the local communities and NERC, ERC, European Union, Royal Society and Leverhume Trust; a grant from the Royal Society and the Natural Environment Research Council, UK (S.L.L.); National Science Foundation CIF21 DDBE: EZ number 1724732 (A.C.C.); National Natural Science Foundation of China (31870374) and Shandong Provincial Natural Science Foundation (ZR2019BC083); the USDA National Institute of Food and Agriculture (USDA- NIFA) (6108-00076); the Amazon Forest Inventory Network (RAINFOR); the CGIAR Research Program on Forests, Trees and Agroforestry (CRP-FTF); the donor fund to the CGIAR-Research Program (AFRICA); the US National Science Foundation project number 19-77-300-12 (R.V.M.); the Takenaka Scholarship Foundation; the Research Council of the UK (NERC) project NE/P016365/1 and NE/P014196/4 (A.C.); the Romania National Council for Higher Education Funding, CNISF, project number CNISFSFDI-2022-0259 (O.B.); Natural Sciences and Engineering
Investigación y Postgrado, Universidad de La Frontera, Temuco, Chile. 189Departamento de Silvicultura y Conservación de la Naturaleza, Universidad de Chile, Santiago, Chile. 190Peoples Friendship University of Russia (RUDN University), Moscow, Russia. 191University of Freiburg, Faculty of Biology, Freiburg, Germany. 192Institution with City, Department of Geography, University of Zurich, Zurich, Switzerland. 193National Forest Centre, Zvolen, Slovak Republic. 194CNRS-UMR LEEISA, Campus Agronomique, Kourou, French Guiana. 195Université de Lorraine, AgroParTech, INRA, Nancy, France. 196Center for International Forestry Research (CIFOR), Situ Gede, Bogor Barat, Indonesia. 197Cirad, University of Montpellier, Montpellier, France. 198Universidade Federal do Rio Grande do Norte, Departamento de Ecologia, Natal, Brazil. 199School of Biological Sciences, University of Aberdeen, Aberdeen, UK. 200Herbarium Kew, Royal Botanic Gardens Kew, London, UK. 201Faculté des Sciences Appliquées, Université de Mbujimai, Mbujimai, Democratic Republic of Congo. 202Yale School of Forestry and Environmental Studies, New Haven, CT, USA. 203URAL State Forest Engineering University, Botanical Garden, Ural Branch of the Russian Academy of Sciences, Yekaterinburg, Russia. 204DIBAF Department, Tuscia University, Viterbo, Italy. 205LINCGlobal, MNCN, CSIC, Madrid, Spain. 206Plant Ecology and Nature Conservation Group, Wageningen University, Wageningen, Netherlands. 207Agricultural High School, ESAV, Polytechnic Institute of Viseu, IPV, Viseu, Portugal. 208Centre for the Research and Technology of Agro-Environmental and Biological Sciences, CITAB, UTAD, Quinta de Prados, Vila Real, Portugal. 209Department of Forest Engineering, Universidade Regional de Blumenau, Blumenau, Brazil. 210Nucleo de Estudos e Pesquisas Ambientais, Universidade Estadual de Campinas, Campinas (UNICAMP), SP, Campinas, Brazil. 211International Center for Tropical Botany, Department of Biological Sciences, Florida International University, Miami, FL, USA. 212Forest Research Institute, University of the Sunshine Coast, Sippy Downs, Queensland, Australia. 213Sanya Nanfan Research Institute, Hainan Yazhou Bay Seed Laboratory, Hainan University, Sanya, China. 214Kenya Forestry Research Institute, Taita Taveta Research Centre, Wundanyi, Kenya. 215Department of Forest Management, Centre for Agricultural Research in Suriname, Paramaribo, Suriname. 216Polish State Forests-Coordination Centre for Environmental Projects, Warsaw, Poland. 217Research Center of Forest Management Engineering of State Forestry and Grassland Administration, Beijing Forestry University, Beijing, China. 218Department of Statistics, University of Wisconsin–Madison, Madison, WI, USA. 219Institut National Polytechnique Félix Houphouët-Boigny, DFR Eaux, Forêts et Environnement, BP, Yamoussoukro, Ivory Coast. 220Centre for Invasion Biology, Department of Mathematical Sciences, Stellenbosch University, Matieland, South Africa. 221African Institute for Mathematical Sciences, Muizenberg, South Africa. 222E-mail: albeca.liang@gmail.com; chui@sun.ac.za
Extended Data Fig. 1 | Distribution of sample plots in the global forest biodiversity individual-based database (GFBi). (A) GFBi data consists of in situ measurements of approximately 1,255,444 forest sample plots (Phase I sample plots, blue). Together with an independent and complementary dataset of 22,131 forest sample plots (Phase II sample plots, yellow). (B) GFBi sample plots cover 424 ecoregions across the world, with the proportion of sample plots by ecoregion ranging from 0.001% to 19.333%. Phase I sample plots cover 394 ecoregions across the world, and Phase II sample plots cover an additional 30 ecoregions in Africa, South America, Southeast Asia, Mexico, India, and Japan. Together, our ground-based forest sample plots cover 424 of 435 forested ecoregions across the world.
Extended Data Fig. 2 | Examination of estimated tree species richness using three cross-validation approaches. (Left column) In randomized cross-validation (RCV), three imputation models—random forests (RF, top), multiple regression with ordinary least squares (OLS, centre), and XGBoost (XGB, bottom)—were trained for each continent with a random subsample comprising 90% of the training data from that continent; the remaining 10% of the training data were used as the testing set. This process was repeated 20 times with sample replacement to examine the accuracy of estimated tree species richness values for each sample plot. (Centre column) In spatial cross-validation (SCV), all sample data from an ecoregion were reserved for testing the three imputation models, trained with the remaining samples from the same continent. This process was repeated until all the forested ecoregions across the world had been tested. (Right column) For post-sample validation (PSV), we collated a new sample dataset from 22,131 forest sample plots (Phase II data), and used Phase II data as the testing set to evaluate the accuracy of the predictive models that were trained for each continent with the Phase I data. Scatter plots show observed (vertical axis) vs. predicted (horizontal axis) values of tree species richness per hectare, from which we calculated mean absolute error (MAE), root-mean-squared error (RMSE), and the coefficient of determination ($R^2$) of the trendlines (red) between the predicted and observed values.
Extended Data Fig. 3 | Partial dependence of tree species richness on each predictor variable. We drew a partial dependence plot (with 95% confidential interval in shaded bands) for each of the 47 explanatory variables in five categories, that is, bioclimatic (green), vegetation and survey (orange), topographic (blue), anthropogenic (pink), and soil (lime). The final panel in the lower right corner is a 2D partial dependence plot showing how tree species richness changes by annual mean temperature and total annual precipitation, while all the other predictors remained constant at their sample mean. Curves on top of all partial dependence plots depict the density of sample data for each explanatory variable. See Supplementary Table 1 for the units and a detailed description of the explanatory variables.
Extended Data Fig. 4 | See next page for caption.
Extended Data Fig. 4 | Analyses of the latitudinal gradients in the residuals from three nested models. For each model, we show a scatter plot of absolute residual values (blue circles) and a column plot of root-mean-squared error (RMSE) calculated from the training dataset, with 2° bins and vertical axis in reverse order (black columns). Model (A) consisted of all 47 explanatory variables plus latitude. Model (B) contained all the explanatory variables in (A) except for the 21 bioclimatic variables. Model (C) as the base model only contained a zero constant. RMSE values in the upper-right corner of each panel represent the total RMSE, with $p$-values calculated from the one-sided F-test (see Variance Partitioning for details). $p(A)$ stands for the $p$-value of all 47 explanatory variables plus latitude, and $p(B)$ stands for the $p$-value of 21 bioclimatic variables.
Extended Data Fig. 5 | Hyper-parameter selection for the random forests model. Using 10 bootstrapping iterations on random training sets consisting of 1000 random sample for each continent, we calculated the sensitivity of the root-mean-squared error (RMSE) to (left) the number of trees to grow and (right) the number of variables randomly sampled as candidates at each split in the random forests model. As RMSE converged at 50 trees and 10 variables, we selected them as optimal hyper-parameters for the random forests model. Black lines represent the mean of all the iterations (with red bands showing the standard error of the mean), and blues lines represent each iteration.
Extended Data Fig. 6 | Hyper-parameter selection for the XGBoost model. Using 20 bootstrapping iterations on random training sets consisting of 90% of sample for each continent, we calculated the sensitivity of the root-mean-squared error (RMSE) of the testing sets (consisting of the remaining 10% of sample) to (left) the maximum number of boosting iterations (that is number of rounds), and (right) the maximum depth of a tree for the XGBoost model. As RMSE converged at 50 rounds and 20 depth, we selected them as optimal hyper-parameters for the XGBoost model. The box plot represents the 25th and 75th percentiles (bounds of box), median (centre line), and the maximum and minimum (upper and lower whiskers) of the RMSE values for each level of tree depth.
Extended Data Fig. 7 | See next page for caption.
Extended Data Fig. 7 | Co-limiting environmental drivers of tree species richness per hectare in forested areas worldwide. Driver dominance was derived for each pixel from four driver categories (that is bioclimatic, topographic, anthropogenic, and soil), as well as ‘co-limitation’ which represents a lack of clear dominance among the four categories above. All map layers are displayed at a 0.025° pixel with an equirectangular projection (Plate-Carrée). See Fig. 5 for a map of all categories overlaid on top of each other.
Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

<table>
<thead>
<tr>
<th>Item</th>
<th>Confirmed</th>
</tr>
</thead>
<tbody>
<tr>
<td>n/a</td>
<td></td>
</tr>
<tr>
<td>The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement</td>
<td>☑</td>
</tr>
<tr>
<td>A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly</td>
<td>☑</td>
</tr>
<tr>
<td>The statistical test(s) used AND whether they are one- or two-sided</td>
<td>☑</td>
</tr>
<tr>
<td>Only common tests should be described solely by name; describe more complex techniques in the Methods section.</td>
<td>☑</td>
</tr>
<tr>
<td>A description of all covariates tested</td>
<td>☑</td>
</tr>
<tr>
<td>A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons</td>
<td>☑</td>
</tr>
<tr>
<td>A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)</td>
<td>☑</td>
</tr>
<tr>
<td>For null hypothesis testing, the test statistic (e.g. F, t, r) with confidence intervals, effect sizes, degrees of freedom and P value noted</td>
<td>☑</td>
</tr>
<tr>
<td>Give P values as exact values whenever suitable.</td>
<td>□</td>
</tr>
<tr>
<td>For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings</td>
<td>☑</td>
</tr>
<tr>
<td>For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes</td>
<td>☑</td>
</tr>
<tr>
<td>Estimates of effect sizes (e.g. Cohen’s d, Pearson’s r), indicating how they were calculated</td>
<td>☑</td>
</tr>
</tbody>
</table>

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

**Data collection**

No software was used

**Data analysis**

The development of the GEPI database, tabular data cleaning, creation of species abundance matrices, evaluation of diversity determinants, and geostatistical imputation were conducted in R49 (v.3.4.2) through the use of several Linux-based high-performance computing (HPC) resources at Purdue University, and a custom HPC interface developed using Amazon Web Services, each designed for batch processing, scalable resource distribution, embarrassingly parallel computations, and/or large RAM jobs. Compute nodes with up to 1TB of RAM and clusters of up to 64 nodes were employed in this study. Portions of the covariate preparation, mapping, and quality control assessment were conducted on Windows-based operating systems with up to 128 GB of RAM.

Final continental-level RF models and the R codes we developed to train the models are available under license MIT, with the identifier: 10.6084/m9.figshare.17234729.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.
Data

Policy information about availability of data
All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:
- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

Data availability

- The global map of tree species richness is available under license CC BY 4.0, with the identifier: 10.6084/m9.figshare.17232491. This map can be downloaded in two formats. One is a geoTIFF file (S_mean_raster.tif) containing the fully geo-referenced map of tree species richness worldwide at a 0.025°×0.025° resolution. The other is a comma-separated file (S_mean_grid.csv) with the following attributes:
  - S is local average tree species richness per hectare
  - x, y are centroid coordinates of all 0.025°×0.025° pixels;
- The global map of co-limitation is available under license CC BY 4.0, with the identifier: 10.6084/m9.figshare.17234339.
- The metadata of the entire training dataframe – including the characteristics and references of all the in situ Phase-I and Phase-II datasets, as well as the definitions, units, and summary statistics of the environmental covariates – is available under license CC BY 4.0, with the identifier: 10.6084/m9.figshare.19733449.v1
- The public version of the training dataframe including the plot-level species richness and all the covariates, which is needed to reproduce the models and results presented here, will be deposited to a public data repository Figshare upon the final acceptance of this paper. Meanwhile, we will also publish this dataframe on two international web research platforms: science-i.org, and gfbinitiative.org.
- Raw forest inventory data are commonly subject to a wide array of confidentiality clauses in regards to open access policies. Despite recent efforts to make some of these data fully open76,77, some governments and private data owners, especially those from the developing countries generally have decided to keep their data confidential. This decision is based on well-founded arguments to protect certain trees or forests (because of their large size or protected taxonomic status) from illegal logging or trespassing, and to protect landowners’ privacy, against the misuse of plot information such as the geographic coordinates. The sensitive information in the training dataframe, including the plot coordinates and tree-level information, will be available from the corresponding author (albeca.liang@gmail.com) upon a request via Science-I or GFBI, and an approval from data contributors.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences  ☐ Behavioural & social sciences  ☑ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

<table>
<thead>
<tr>
<th>Study description</th>
<th>We compiled a ground-sourced forest inventory database containing ca. 55 million trees in ~1.3 million forest sample plots, based on which we developed the first global maps of tree species diversity and global map of co-limitation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Research sample</td>
<td>In situ measurement of ~1.3 million forest sample plots spanning across 110 countries and territories</td>
</tr>
<tr>
<td>Sampling strategy</td>
<td>Sampling strategy is reported separately by the 96 national, regional, and local forest inventories underlying our database. See Table S2 for a complete list of all the forest inventories and references to the associated reports.</td>
</tr>
<tr>
<td>Data collection</td>
<td>Data collection is reported separately by the 96 national, regional, and local forest inventories underlying our database. See Table S2 for a complete list of all the forest inventories and references to the associated reports.</td>
</tr>
<tr>
<td>Timing and spatial scale</td>
<td>The overall GFBi database has a global coverage. The metadata of the entire training dataframe – including the characteristics and references of all the in situ Phase-I and Phase-II datasets, as well as the definitions, units, and summary statistics of the environmental covariates – is available under license CC BY 4.0, with the identifier: 10.6084/m9.figshare.19733449.v1</td>
</tr>
<tr>
<td>Data exclusions</td>
<td>Trees with incorrect or unrecognisable species names may have been excluded, according to the section Data collection and standardization.</td>
</tr>
<tr>
<td>Reproducibility</td>
<td>This study uses observational data from natural forest ecosystems.</td>
</tr>
<tr>
<td>Randomization</td>
<td>Random forests model was trained to estimate global forest tree species richness and co-limitation</td>
</tr>
<tr>
<td>Blinding</td>
<td>Blinding was not relevant to this study, because this study is based primarily on observational tree data from natural forest ecosystems.</td>
</tr>
</tbody>
</table>

Did the study involve field work? ☐ Yes  ☑ No
## Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

<table>
<thead>
<tr>
<th>Materials &amp; experimental systems</th>
<th>Methods</th>
</tr>
</thead>
<tbody>
<tr>
<td>[ ] n/a Involved in the study</td>
<td>[ ] n/a Involved in the study</td>
</tr>
<tr>
<td>☒ Antibodies</td>
<td>☒ ChIP-seq</td>
</tr>
<tr>
<td>☒ Eukaryotic cell lines</td>
<td>☒ Flow cytometry</td>
</tr>
<tr>
<td>☒ Palaeontology and archaeology</td>
<td>☒ MRI-based neuroimaging</td>
</tr>
<tr>
<td>☒ Animals and other organisms</td>
<td></td>
</tr>
<tr>
<td>☒ Human research participants</td>
<td></td>
</tr>
<tr>
<td>☒ Clinical data</td>
<td></td>
</tr>
<tr>
<td>☒ Dual use research of concern</td>
<td></td>
</tr>
</tbody>
</table>