PNAS
www.pnas.org

Supplementary Information for

Phylogenomic and ecological analyses reveal the spatiotemporal evolution of global pines

Wei-Tao Jin, David S. Gernandt, Christian Wehenkel, Xiao-Mei Xia, Xiao-Xin Wei, Xiao-Quan Wang

Corresponding authors:
Xiao-Quan Wang, Email: xiaoq_wang@ibcas.ac.cn
Xiao-Xin Wei, Email: weixx@ibcas.ac.cn

This PDF file includes:

Additional Methods Information
Figs. S1 to S22
Tables S1 to S10
SI References

Additional Methods

Sampling, transcriptome sequencing and data processing. A total of 117 species of *Pinus* were studied, including 109 recognized by ref. 1 and eight (*Pinus chiapensis*, *P. dabeshanensis*, *P. discolor*, *P. georginae*, *P. johannis*, *P. kwangtungensis*, *P. mastersiana*, and *P. washoensis*) recognized as distinct species by some previous studies (2-5). Most species were represented by more than one individual. For transcriptomic analysis, with the exception of five species (*P. amamiana*, *P. cubensis*, *P. johannis*, *P. occidentalis*, and *P. tropicalis*) that are unavailable to us, a total of 255 accessions representing 112 species were included (Dataset S1). Among them, 234 transcriptomes from 107 species were newly generated, and the remaining data were retrieved from NCBI. The raw sequencing reads are deposited in the Sequence Read Archive of GenBank (Dataset S1). *Cathaya argyrophylla* and three spruces, *Picea abies*, *P. breweriana and P. smithiana*, were used as outgroups based on the close relationships of the two genera with *Pinus* (6).

Total RNA was extracted from haploid megametophytes or young leaves with RNAprep Pure Plant Kit (Tiangen, Beijing). The cDNA libraries were prepared using NEBNext® Ultra™ Directional RNA Library Prep Kit for Illumina® (New England BioLabs Inc. Ipswich, MA, USA) and used for paired-end sequencing of a 2 x 100 bp run with an Illumina HiSeq2500 device in the State Key Laboratory of Systematic and Evolutionary Botany, Institute of Botany, the Chinese Academy of Sciences (LSEB, IB-CAS). Raw reads were first checked with FastQC (v.0.11.5) (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and trimmed using Trimmomatic v.0.36 (ILLUMINACLIP:TruSeq3-PE-2.fa:2:30:10 LEADING:20 TRAILING:20 HEADCROP:3 SLIDINGWINDOW:8:20 MINLEN:36) (7). The clean reads were then de novo assembled using Trinity (v.20140717) (8) with min_kmer_cov as 2, SS_lib_type as RF and other parameters with default settings. Redundant transcripts were removed by CD-HIT v.4.6.5 (-c 1.0 -n 10) (9, 10). Putative coding sequences (CDSs) were predicted using TransDecoder_r20140704 (https://github.com/TransDecoder/TransDecoder/releases) with default settings (11), and the isoform encoding the longest peptide was selected as unigene. The CDSs and corresponding amino acid (AA) sequences of the unigenes were used for further analyses.

Ortholog identification. HaMStR (v.13.2.6) (12) was used to search for single-copy orthogroups (OGs) for phylogenetic analyses. Firstly, we identified 2,996 core single-copy OGs using OrthoFinder (v.2.2.7) (13) based on a set of 17 primer taxa with sequenced transcriptomes which covered nine of eleven subsections of *Pinus* recognized by (4) except for subsections *Krempfianae* and *Nelsoniae*. The primer taxa were *P. aristata*, *P. armandii*, *P. canariensis*, *P. clausa*, *P. gerardiana*, *P. heldreichii*, *P. latteri*, *P. muricata*, *P. pinceana*, *P. pinea*, *P. resinosa*, *P. sabiniana*, *P. serotina*, *P. squamata*, *P. strobus*, *P. torreyana*, and *P. virginiana*, with *P. pinea* as the HaMStR reference taxon. Then the inferred AA sequences of core single-copy OGs were aligned using MAFFT (v.7.407) (14), and were used to build a profile HMM with hmmbuild using HMMER (v.3.2.1) (15). Finally, to accurately determine OGs for each species, HaMStR (v.13.2.6) (12) was performed for CDSs of all of the 255 pine samples and four outgroup species with strict parameters (-est -central -intron = remove -representative -strict -eval_blast =1e-20 -eval_hmmer =1e-20). Subsequently, each OG with CDS and AA sequences was extracted with Perl scripts.

To optimize the final OG dataset for phylogenetic analysis, we tested whether conspecific individuals form monophyletic groups. A total of 965 OGs were extracted and concatenated by FASconCAT-G (v.1.04) (16) for phylogenetic reconstruction, with each OG covering at least 252 individuals. IQ-TREE v.2.0-rc1 (17) was used for the generation of Maximum likelihood (ML) tree with 1,000 ultrafast bootstrap replicates (UFBoot) (18) using GTR+I+G model. The preliminary result showed that all conspecific individuals but those from *P. luzmariae* and *P. pringlei* formed monophyletic groups with strong support (*SI Appendix*, Fig. S1). To improve computing efficiency and obtain more OGs to study the phylogenetic relationships of pines, only one individual with the longest N50 value and the maximum number of core single-copy OGs (Dataset S1) was retained for each monophyletic species. A total of 113 individuals of 112 pine species and 4 outgroups were further used to identify OGs for subsequent analyses. 1,768 candidate OGs were obtained, and each of them covered all samples or only missed in a few samples from the five species (*P. herrerae*, *P. jaliscana*, *P. luzmariae*, *P. pringlei*, and *P. remota*) because of their low sequencing quality.

For the 1,768 candidate OGs, the AA sequences were aligned with MAFFT (v.7.407) (14), and then the corresponding CDS alignments were obtained using PAL2NAL (v.14) (19). To reduce the impact of low-quality OGs and sequences on phylogenetic analysis, we manually inspected each OG. The OG with over five ambiguous sequences that were almost impossible to be aligned with the rest sequences was removed. In addition, the ambiguously aligned positions in AA alignments were identified with ZORRO (20) and the positions with a confidence score less than 4 were removed. Finally, we obtained 1,662 OGs for subsequent analyses, each of which had a length longer than 100 aligned amino acids and covered at least 110 species.

Phylogenetic analyses. Phylogenomic analyses were performed based on two datasets, i.e., CDS and CDS (1st+2nd), using both concatenation and coalescence methods. For the concatenation analysis, the OGs were concatenated by FASconCAT-G (v.1.04) (16) and used for the ML analysis by IQ-TREE v.2.0-rc1 (17). The best substitution models for CDS and CDS (1st+2nd) were GTR+F+R11 and GTR+F+R15, respectively, which were determined by ModelFinder (21) using two options (-mset GTR -cmax 15), according to the Bayesian information criterion (BIC). Branch supports were generated with 1,000 ultrafast bootstrap replicates (18). For the coalescent analyses, to improve the accuracy of individual gene trees, the sequences with gap characters more than 60% were removed from each OG (22), and then individual gene trees were generated by IQ-TREE v.2.0-rc1 (16) with 1,000 ultrafast bootstrap replicates (18) under the best substitution model (21). We also contracted branches with very low support (below 10% UFBoot support) from each gene tree to improve the accuracy (23, 24). Based on these individual gene trees, we estimated the species tree in ASTRAL (v.5.7.3) (24) with local posterior probability (LPP) (25).

Divergence time estimation. The concatenated CDS dataset and corresponding ML tree were used for dating analysis. Despite that *Pinus* has rich and diverse fossils (26), a large discrepancy has surrounded the fossils' ages and phylogenetic placements. Recent studies (27, 28) assessed most pine fossils and gave some guidelines for molecular dating. Following their suggestions, we used four fossils for the calibration. *Pinus yorkshirensis*, the earliest cone fossil of *Pinus* from the early Cretaceous Wealden Formation in Yorkshire, UK (131-129 Ma) (29), was used to calibrate the stem age of *Pinus* with minage (129 Ma). *Pinuxylon* sp., a silicified fossil wood attributed to the subg. *Strobus* from the late Cretaceous (Santonian, 85.8-83.5 Ma) Aachen Formation of northeast Belgium (30), was used to calibrate the crown age of *Pinus* with a minage of 84 Ma (31). *Pinus florissantii*, a fossil of reproductive organs from the early Oligocene Florissant Flora (34 Ma) (32), was used to calibrate the divergence time between subsections *Strobus* and *Krempfianae*, and *Pinus baileyi*, a cone fossil from the early Eocene Thunder Mountain Flora in Idaho (45 Ma) (33), was used to calibrate the divergence time between subsections *Pinus* and *Pinaster*. To test the influence of *Pinuxylon* sp. on divergence time estimates, we also conducted the analysis by excluding this fossil. Considering that Saladin*,* et al. (27) used 12-21 fossils (different between the ND and FBD methods) as calibration points to estimate the divergence times of *Pinus*, we used the 16 fossils following their study with ND method to time calibration of our phylogeny (MCMCTree can not perform FBD analysis) (*SI Appendix*, Table S1).

We estimated divergence times in the MCMCTree program in PAML 4.9j (34). The approximate likelihood method was conducted to speed up the likelihood calculation for large datasets (35). The time unit was set to 100 Ma and a maximum bound for the root was set to 153 Ma based on the oldest fossil of Pinaceae (36). For the minimum and maximum bounds of all calibration points, the default 2.5 percent tail probability was used. The ML estimates of branch lengths, the gradient vector, and Hessian matrix were calculated using MCMCTree and BASEML (in PAML) programs under the GTR+G substitution models (model = 7). The overall substitution rate (rgene_gamma) and rate-drift parameter (sigma2_gamma) were set as G (1 33.8) and G (1 10), respectively. The independent-rates model (37) was used to specify the prior of rates, while the posterior times were estimated by using a Markov chain Monte Carlo (MCMC) algorithm in the MCMCTree program. Two independent MCMC runs were conducted. Each run discarded the first 10 million iterations as burn-in, and then sampled every 500 iterations until it gathered 20,000 samples. The stationary state and convergence of each run were checked in Tracer (v.1.7.1) (38) to ensure that all parameters had effective sample sizes (ESS) above 200.

The impact of environmental variables on the global pattern of pine species diversity. To explore the underlying driving forces of the distribution pattern of pine species diversity, all 31 environmental variables were extracted as mean values within each 100 km x 100 km grid cell (Dataset S4). To reduce collinearity among variables, the initial set of 31 environmental variables was finally reduced to 13 variables with weak pairwise correlation (|r|<0.7), including five bioclimatic variables (BIO1–annual mean temperature; BIO2–mean diurnal range; BIO8–mean temperature of wettest quarter; BIO15–precipitation seasonality; BIO16–precipitation of wettest quarter), one topographical variable (elevation range), five soil variables (phh2o–soil pH; cfvo–volumetric fraction of coarse fragments; clay–proportion of clay particles; silt–proportion of silt particles; cec–cation exchange capacity), and two landcover variables (consensus1–evergreen/deciduous needleleaf trees; consensus4–mixed/other trees) (*SI Appendix*, Table S10).

Multiple regression models (ordinary least squares, OLS) were built for the global pines, and for the two subgenera (*Pinus* and *Strobus*) and six biogeographic regions of *Pinus*, with pine species richness as the response variable and 13 environmental variables as predictors (*SI Appendix*, Table S10). The response variable was log-transformed in all statistical analyses. Multi-predictor models were evaluated for multi-collinearity using Variance Inflation Factors (VIF) in “car” package (39) and the predictors with VIFs more than 5 were removed before model selection. All predictors were scaled to a mean of zero and variance of one before the analysis to make the direct comparison of regression coefficients, and the OLS model residuals approximated a normal distribution. We used a stepwise regression based on the Akaike Information Criterion (AIC) (40) to derive a minimum adequate model which has the smallest number of predictors and retains the highest explanatory power. For the minimum adequate model, we obtained the standardized coefficient of each predictor to compare the relative importance of predictors in explaining pine species richness.

We also used Moran's *I* values to assess spatial autocorrelation in the residuals of OLS models, and the results indicated significant spatial autocorrelations (*SI Appendix*, Table S9), which could potentially affect significance tests (41). For comparison, we further built spatial simultaneous autoregressive (SAR) models using "spdep" package (42) to account for spatial autocorrelation (43). Finally, the results from the OLS models were used because it has been argued that spatial autocorrelation does not seriously affect OLS estimation of regression coefficients (41), and the results from the SAR models are also shown in *SI Appendix*, Table S9. All statistical analyses were done in R (v.3.6.2) (44).

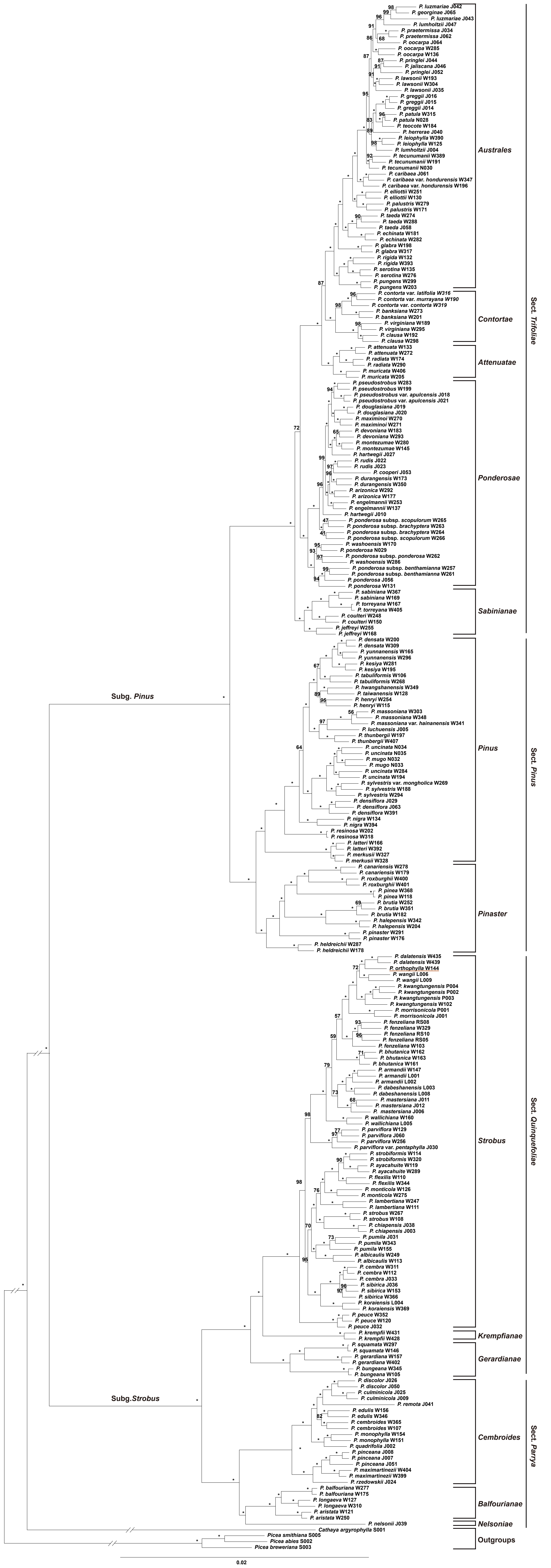

Fig. S1. ML tree of *Pinus* inferred from the concatenated CDS alignment of 965 OGs from all 259 samples. Numbers attached to the branches show the UFBoot supports (* for 100%).

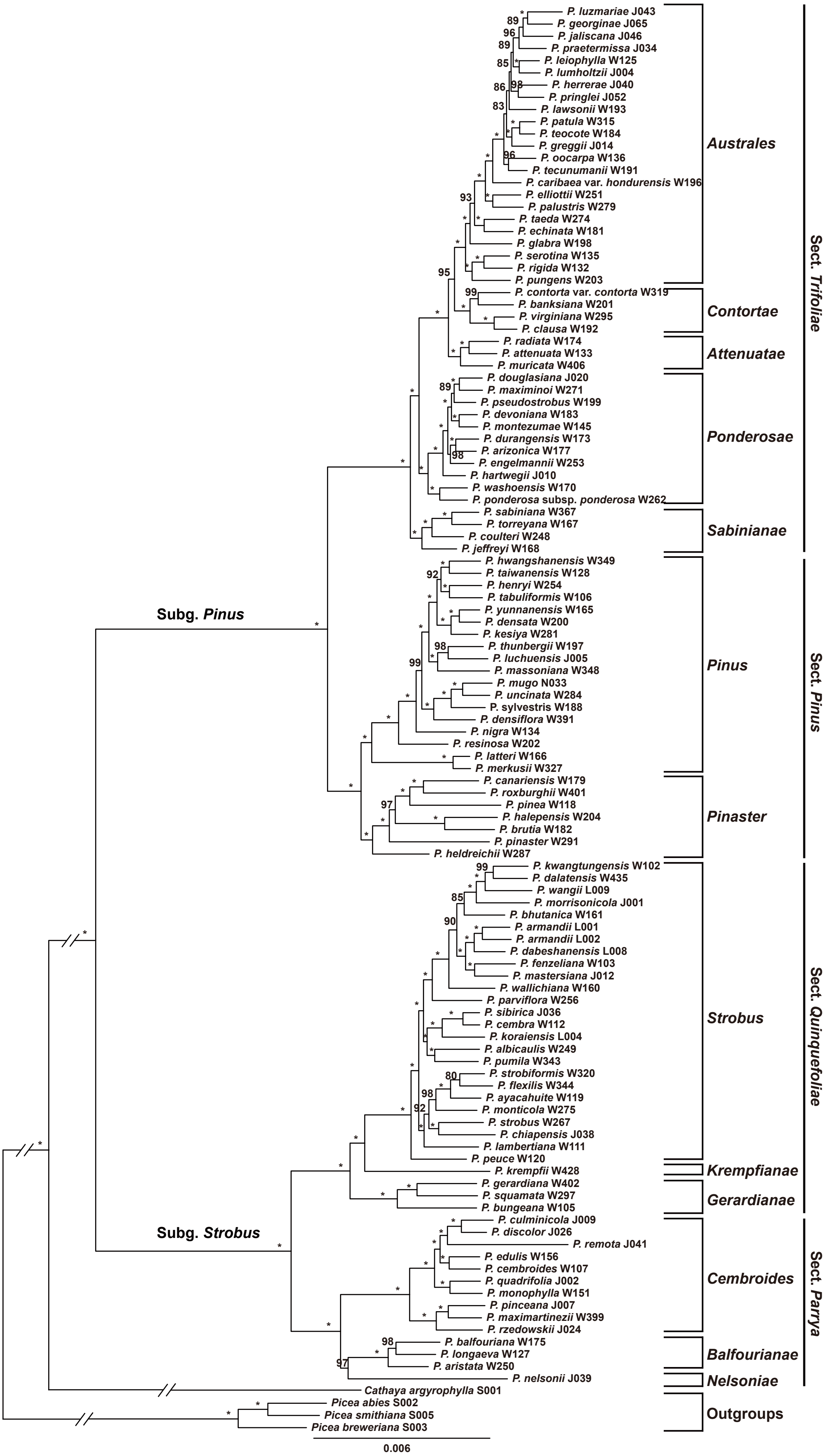

Fig. S2. ML tree of *Pinus* inferred from the concatenated CDS (1st+2nd) alignment of 1662 OGs from 117 samples. Numbers attached to the branches show the UFBoot supports (* for 100%).

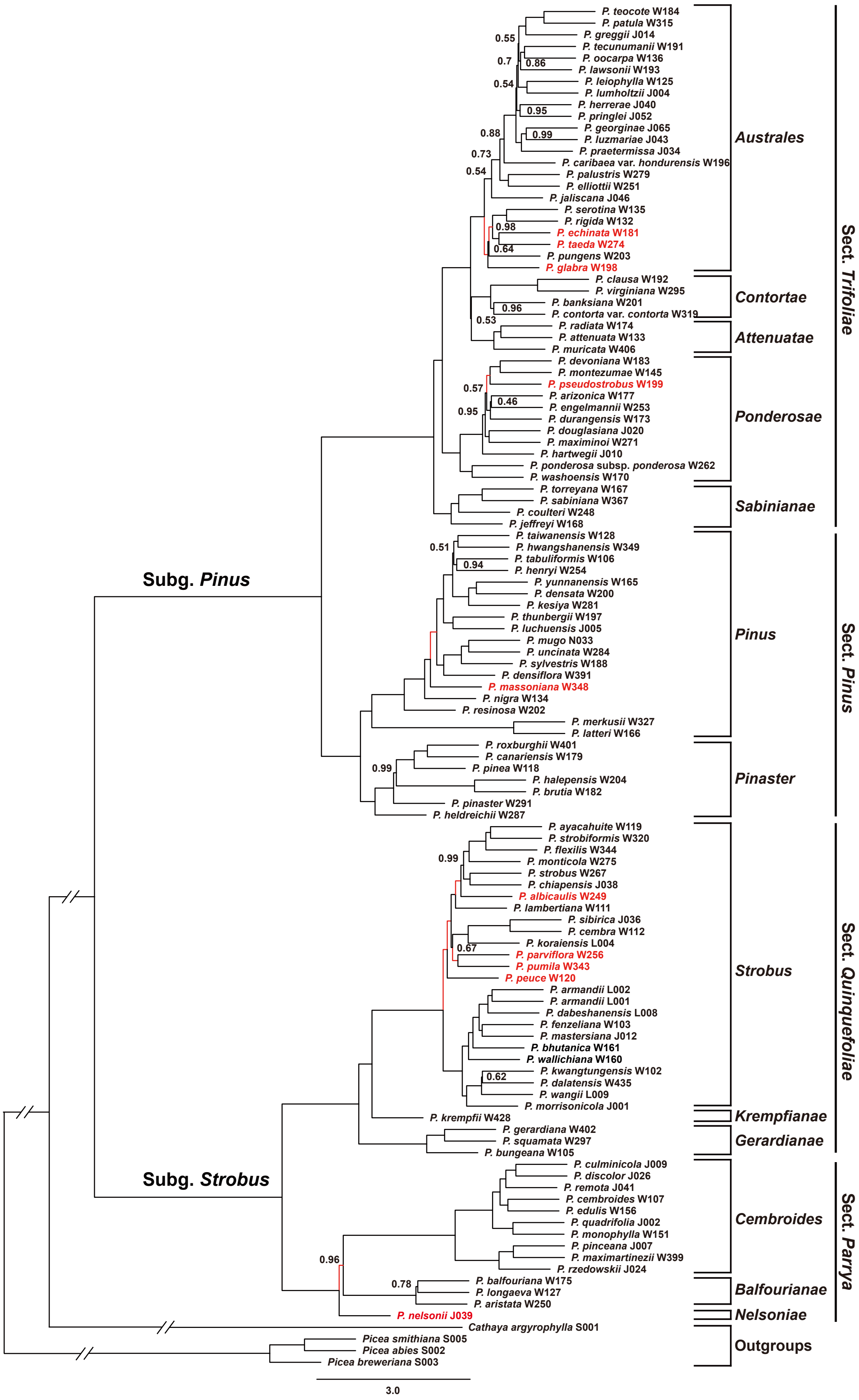

Fig. S3. A *Pinus* phylogeny generated using ASTRAL based on the CDS alignment of 1662 OGs. Except those indicated at the nodes, all other local posterior probability (LPP) values are 100%. Taxa and branches indicated in red refer to the highly inconsistent placements (LPP > 95%) with the ML tree in Fig. S1

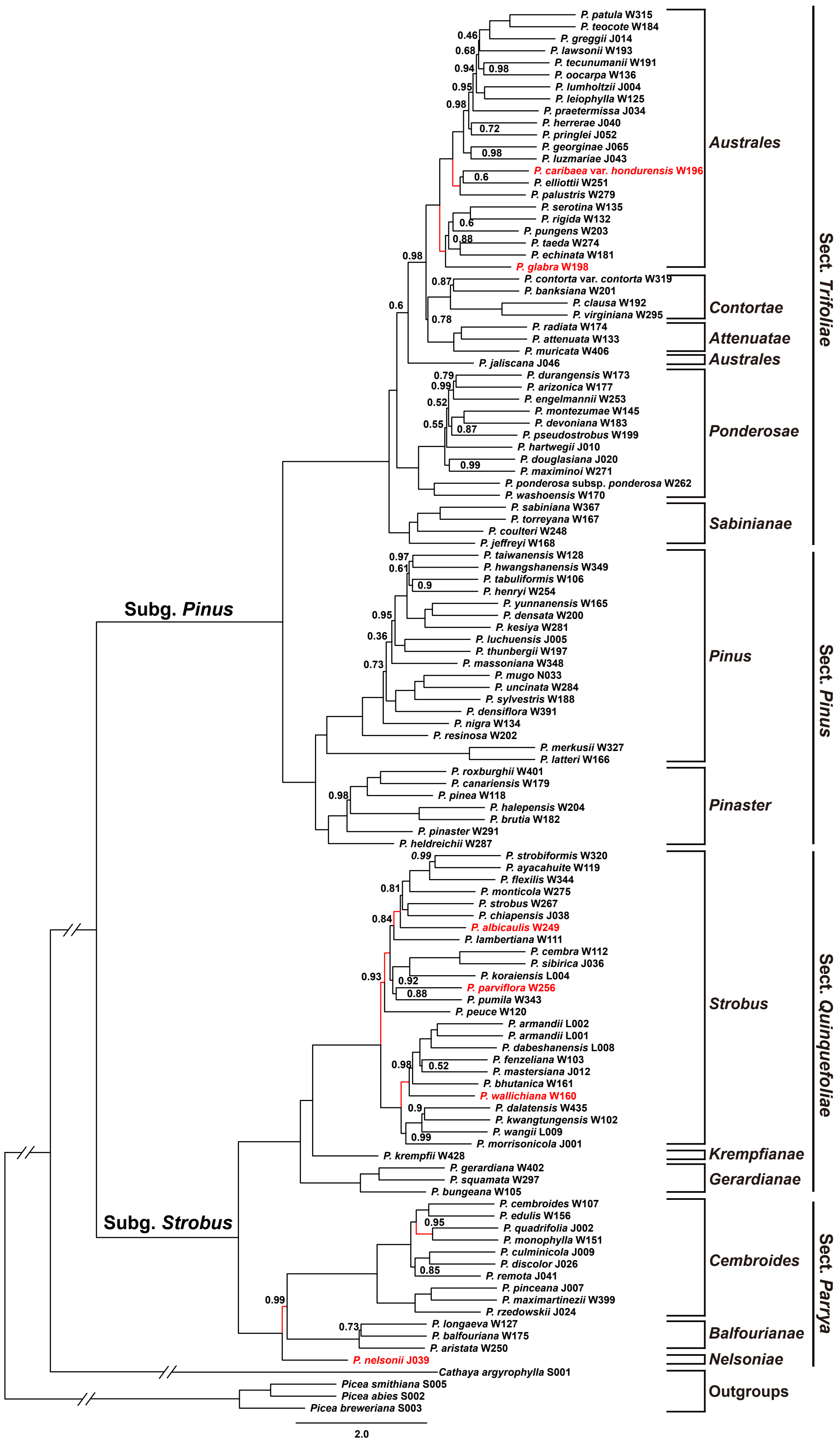

Fig. S4. A *Pinus* phylogeny generated using ASTRAL based on the CDS (1st+2nd) alignment of 1662 OGs. Except those indicated at the nodes, all other local posterior probability (LPP) values are 100%. Taxa and branches indicated in red refer to the highly inconsistent placements (LPP > 95%) with the ML tree in Supplementary Fig. S2.

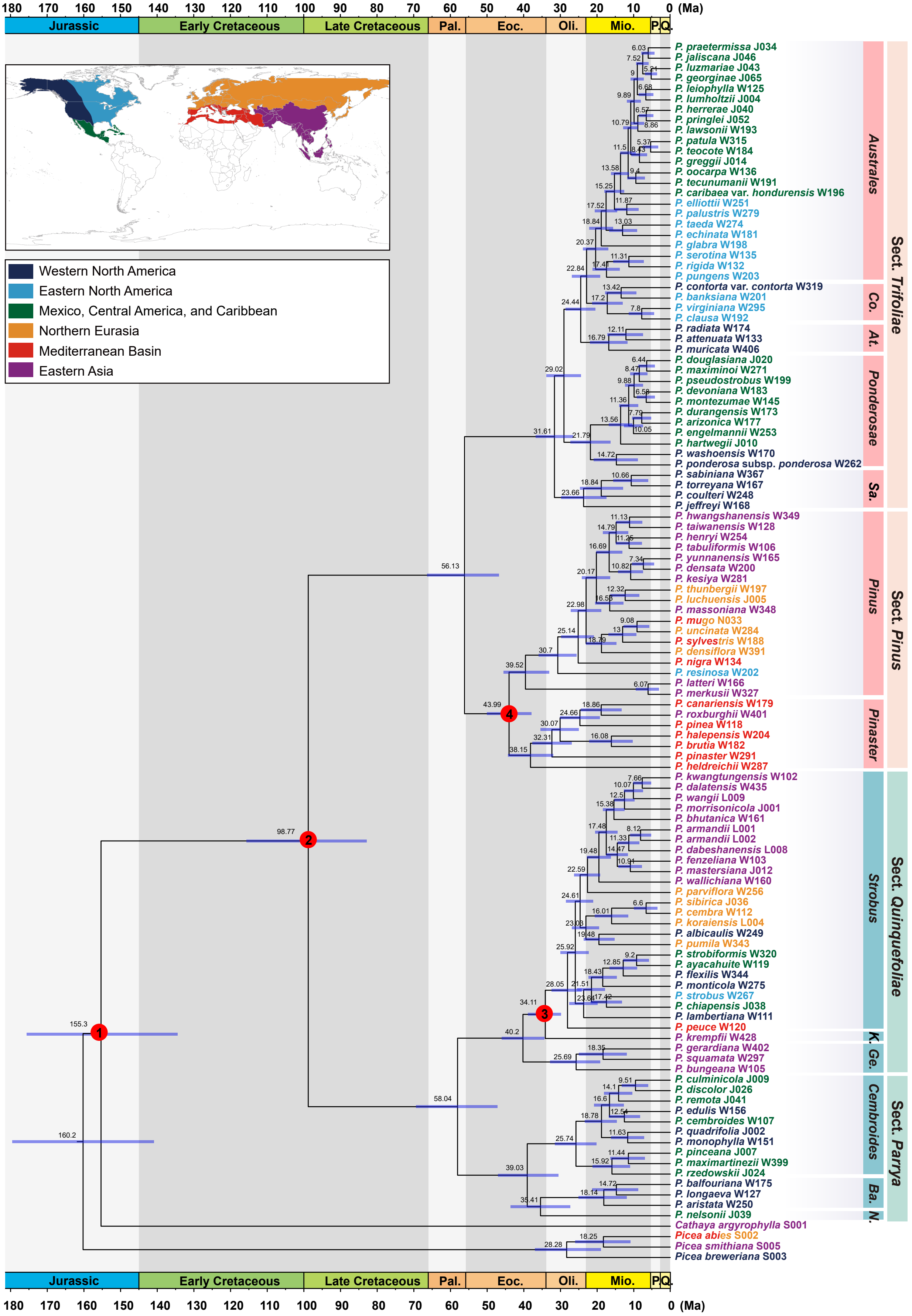

Fig. S5. Divergence times of *Pinus* estimated from the concatenated CDS dataset using MCMCTree. Red circles with numbers indicate the four fossil calibration points. Horizontal bars indicate 95% credible intervals of the divergence time estimates. *Co.*, *Contortae*; *At.*, *Attenuatae*; *Sa.*, *Sabinianae*; *K.*, *Krempfianae*; *Ge.*, *Gerardianae*; *Ba.*, *Balfourianae*; *N.*, *Nelsoniae*; Pal., Paleocene; Eoc., Eocene; Oli., Oligocene; Mio., Miocene; P., Pliocene; Q., Quaternary.

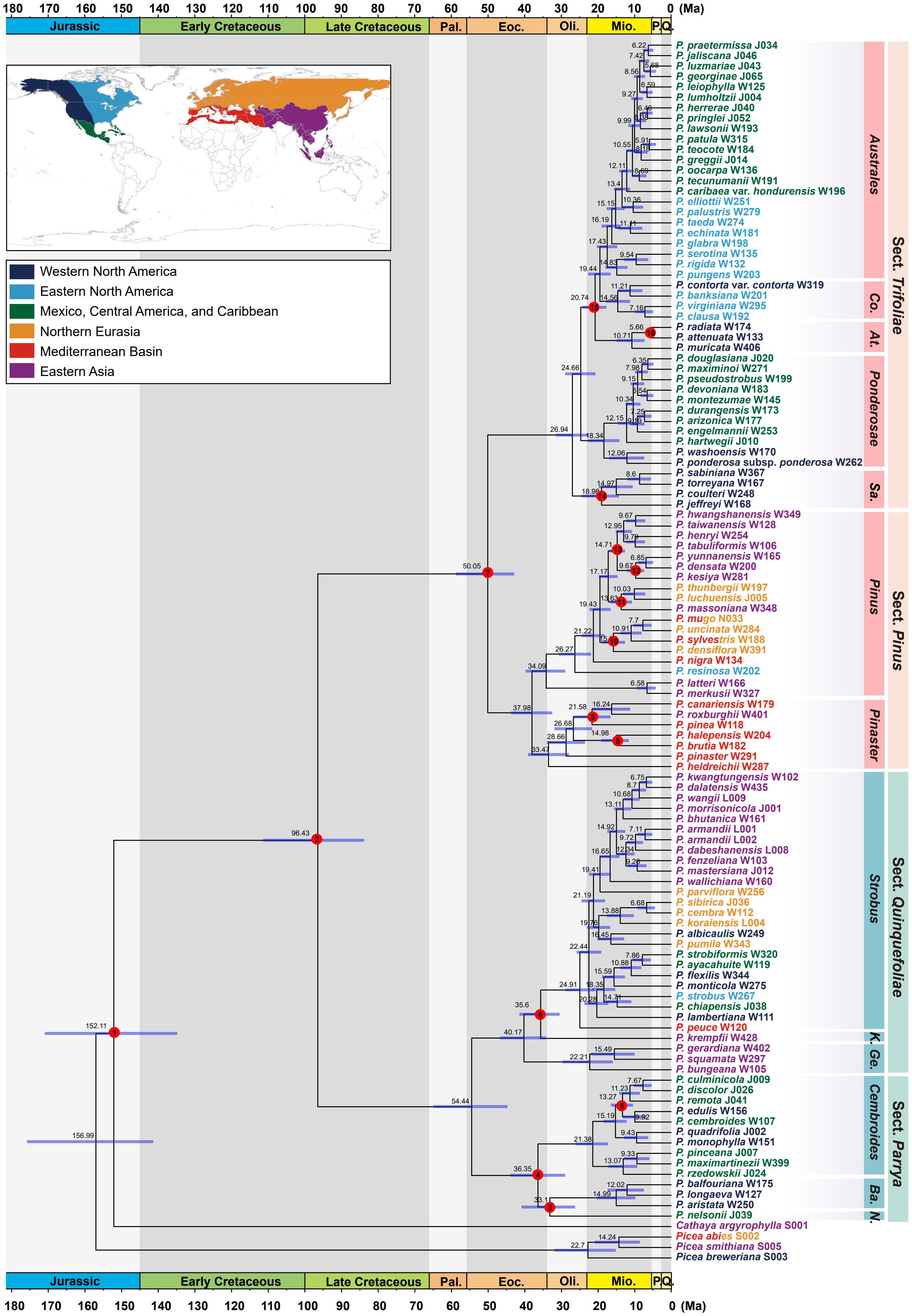

Fig. S6. Divergence times of *Pinus* estimated from the concatenated CDS dataset using MCMCTree. Red circles with numbers indicate the 16 fossil calibration points following ref. 2. Horizontal bars indicate 95% credible intervals of the divergence time estimates. *Co.*, *Contortae*; *At.*, *Attenuatae*; *Sa.*, *Sabinianae*; *K.*, *Krempfianae*; *Ge.*, *Gerardianae*; *Ba.*, *Balfourianae*; *N.*, *Nelsoniae*; Pal., Paleocene;Eoc., Eocene; Oli., Oligocene; Mio., Miocene; P., Pliocene; Q., Quaternary.

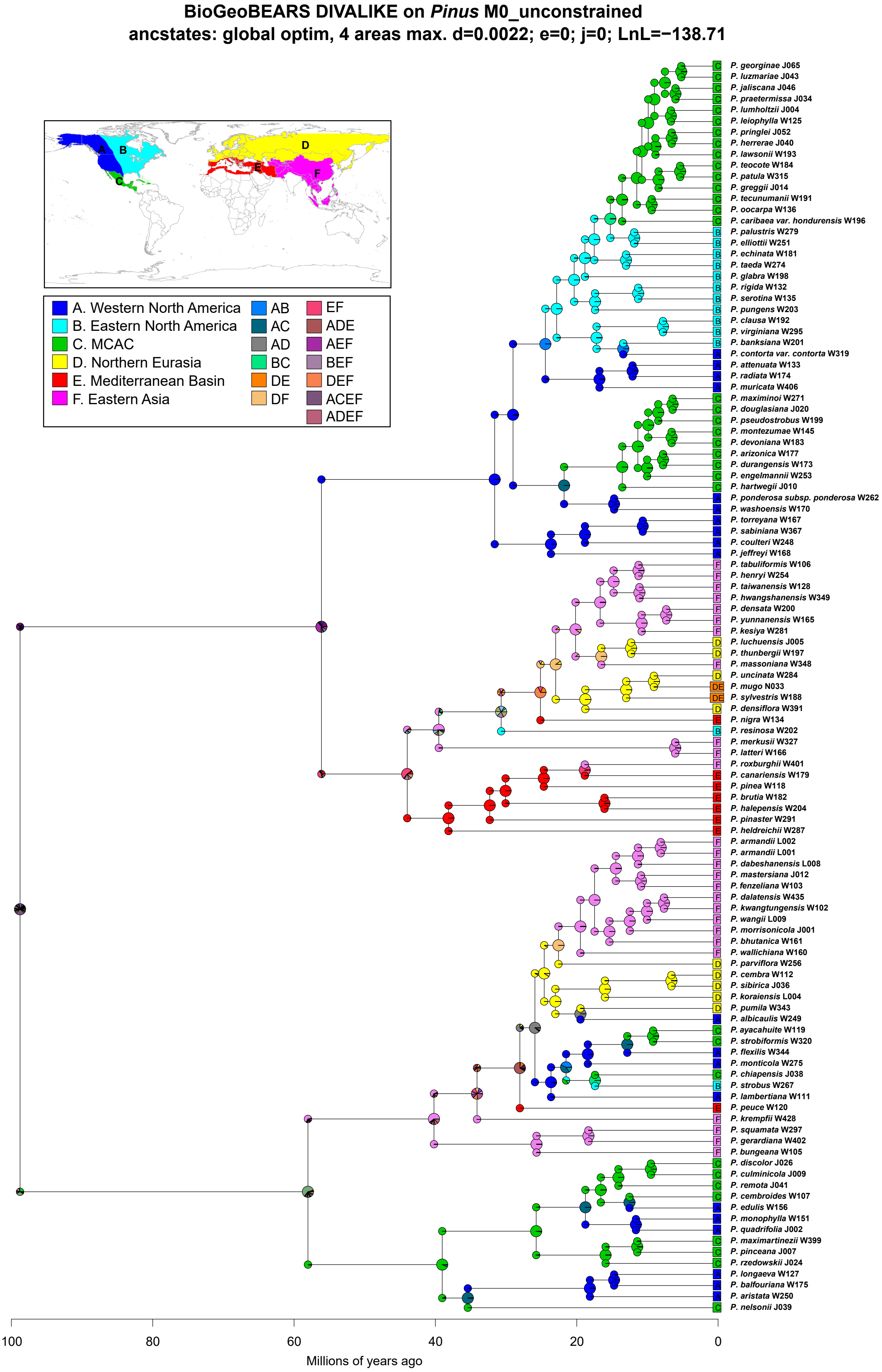

Fig. S7. Ancestral areas of *Pinus* estimated under the DIVALIKE model using BioGeoBEARS. Pie charts represent the probabilities of different possible geographical ranges just before and after a cladogenesis event. MCAC represents Mexico, Central America, and Caribbean.

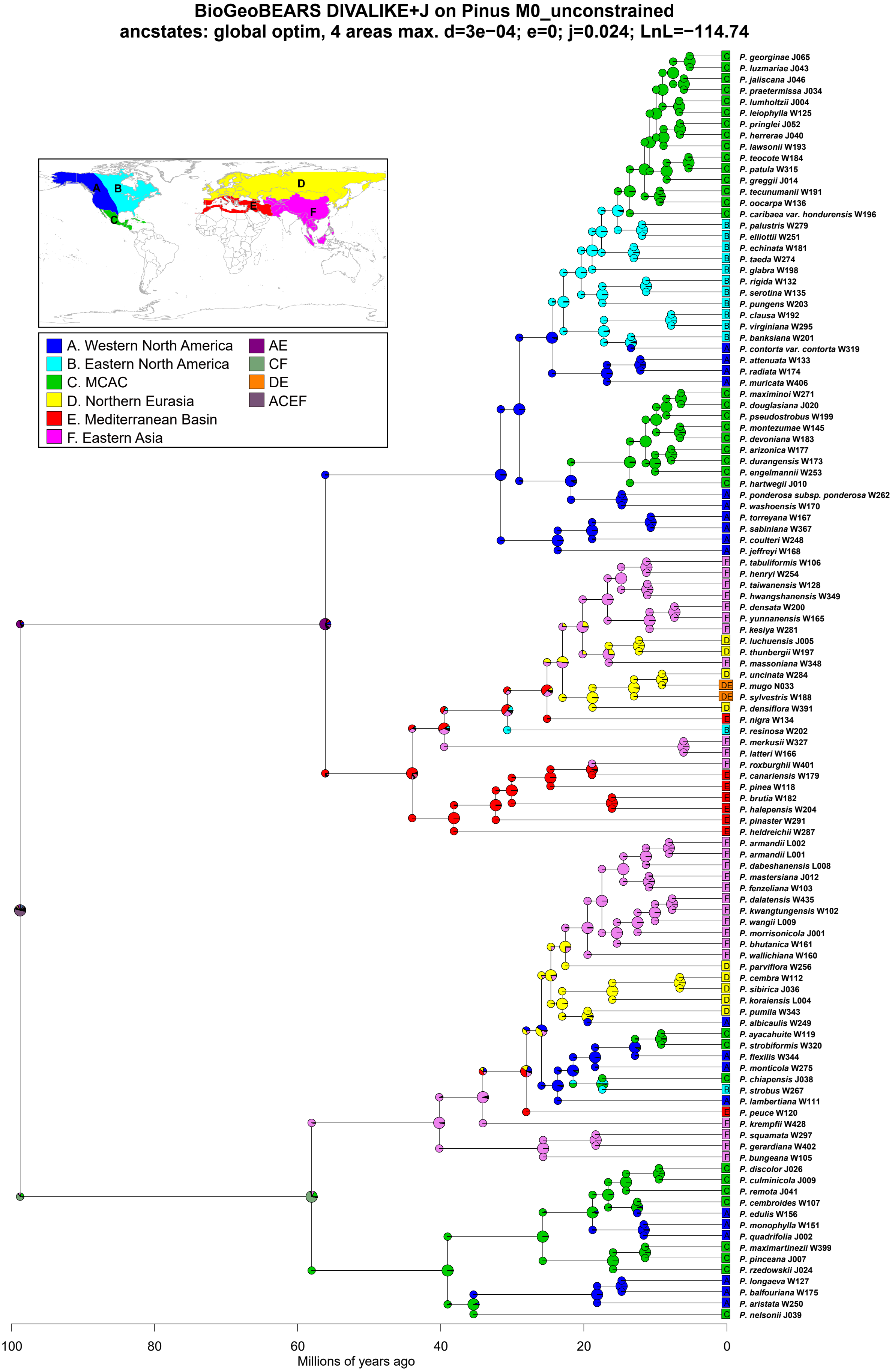

Fig. S8. Ancestral areas of *Pinus* estimated under the DIVALIKE + J model using BioGeoBEARS. Pie charts represent the probabilities of different possible geographical range just before and after a cladogenesis event. MCAC represents Mexico, Central America, and Caribbean.

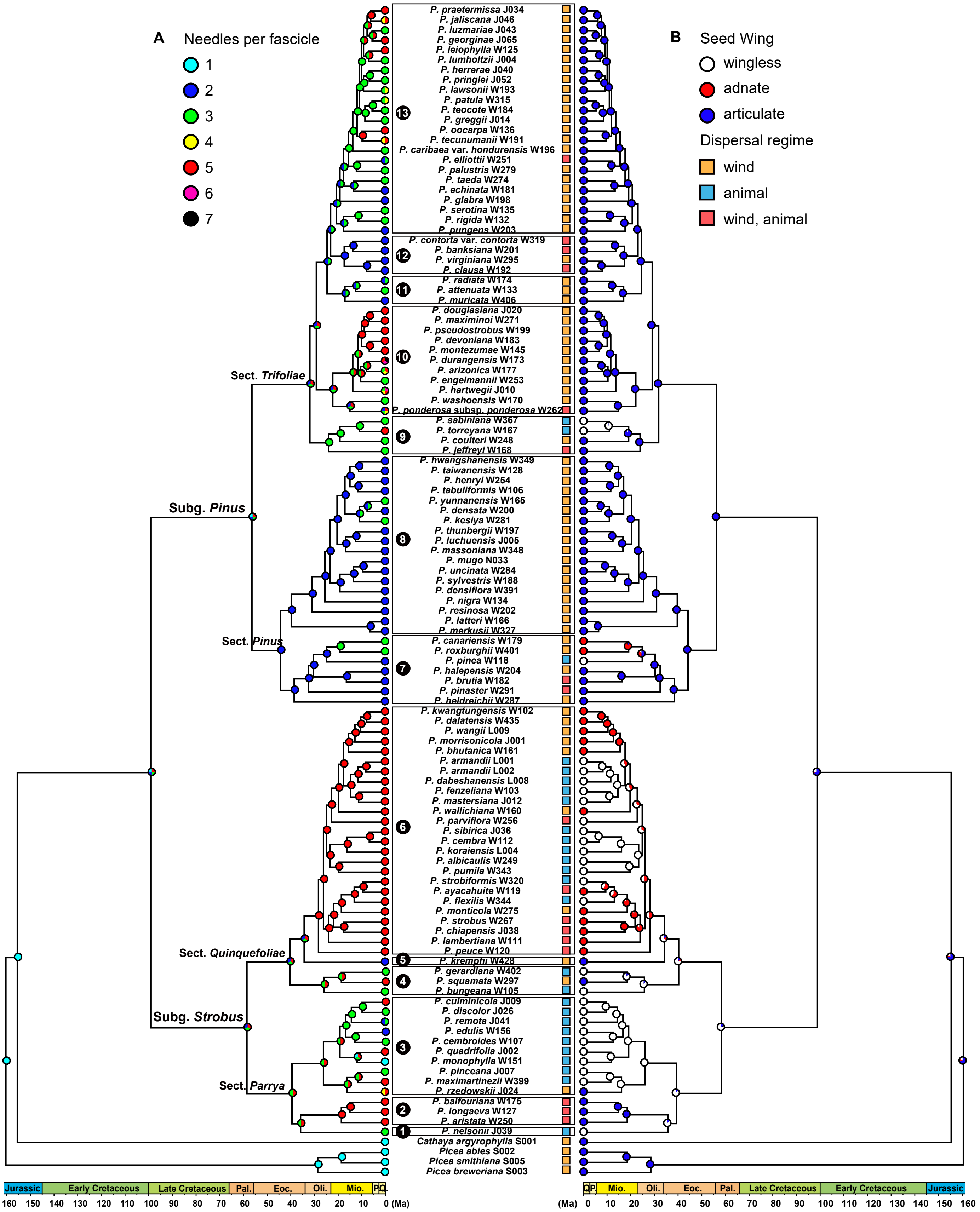

Fig. S9. Ancestral state reconstruction of two morphological characters. A, Needles per fascicle; B, Seed wing. Squares indicate dispersal regimes. Black circles with numbers indicate the thirteen subsections from bottom to top, including *Nelsoniae*, *Balfourianae*, *Cembroides*, *Gerardianae*, *Krempfianae*, *Strobus*, *Pinaster*, *Pinus*, *Sabinianae*, *Ponderosae*, *Attenuatae*, *Contortae*, and *Australes*.

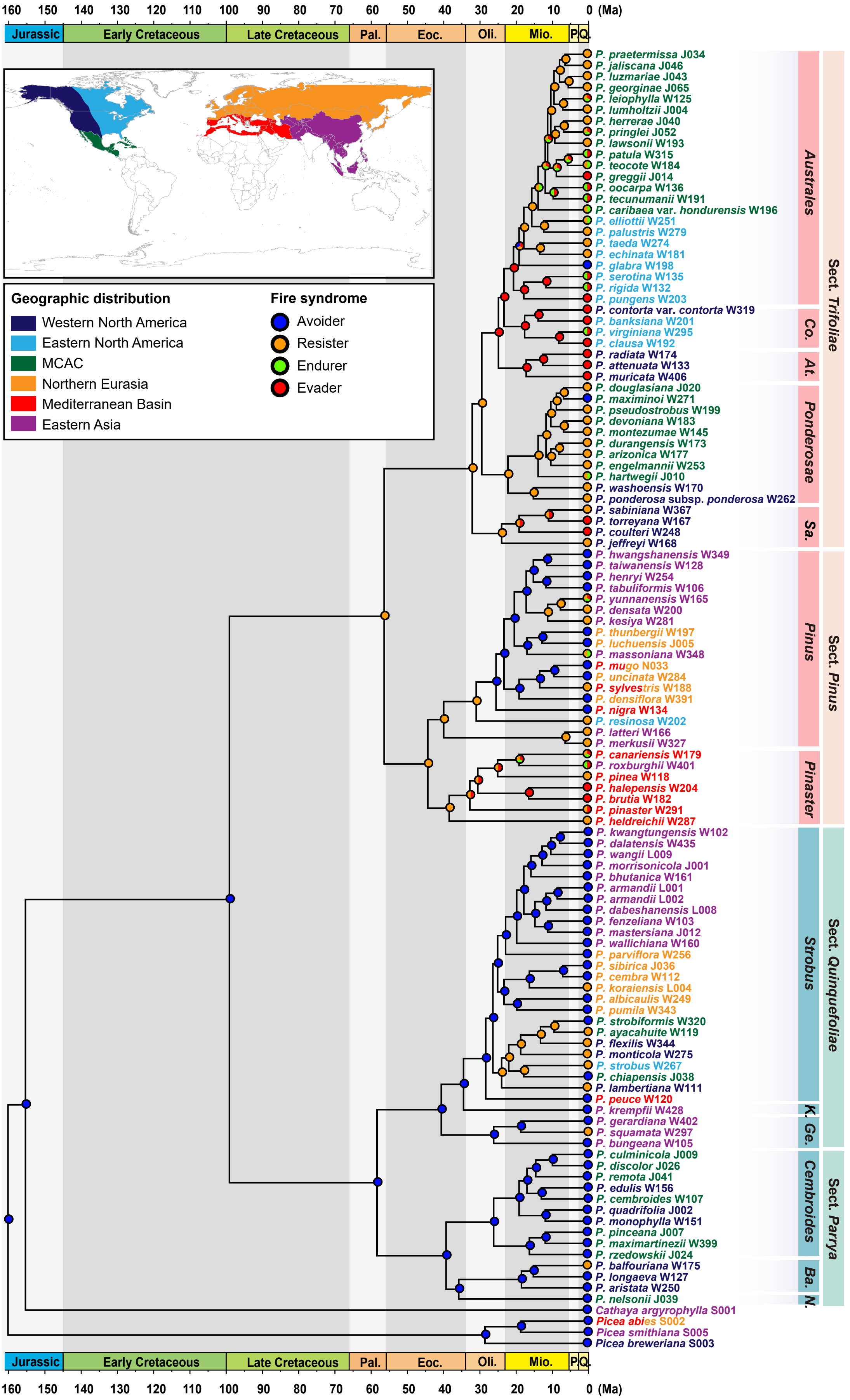

Fig. S10. Ancestral state reconstruction of fire syndromes. *Co.*, *Contortae*; *At.*, *Attenuatae*; *Sa.*, *Sabinianae*; *K.*, *Krempfianae*; *Ge.*, *Gerardianae*; *Ba.*, *Balfourianae*; *N.*, *Nelsoniae*; Pal., Paleocene; Eoc., Eocene; Oli., Oligocene; Mio., Miocene; P., Pliocene; Q., Quaternary; MCAC represents Mexico, Central America, and Caribbean.

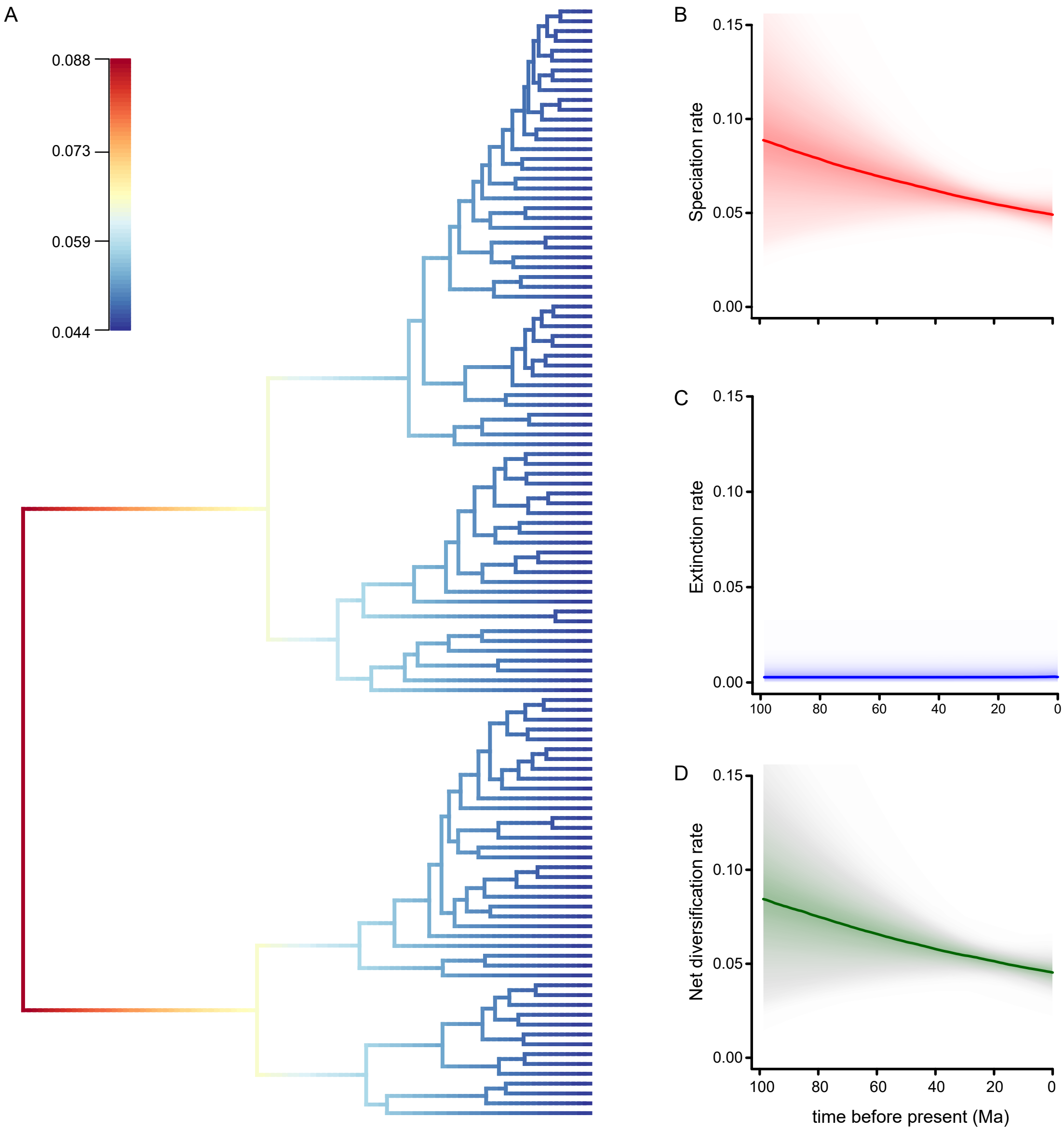

Fig. S11. Diversification rates of *Pinus* estimated by BAMM. A, Phylorate plot showing net diversification rate on a scale from low to high values (blue to red) along each branch of the pine phylogeny. B, Speciation rate through time, with color density shading to denote confidence on evolutionary rate. C, Extinction rate through time, with color density shading to denote confidence on evolutionary rate. D, Net diversification rate through time, with color density shading to denote confidence on evolutionary rate.

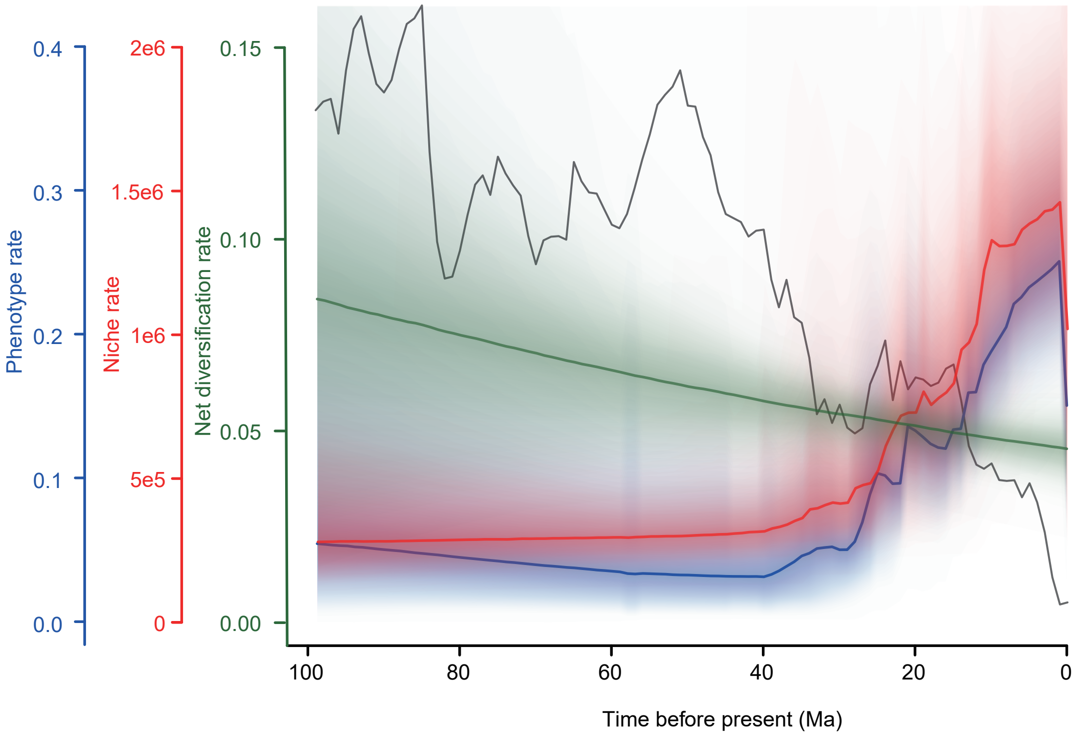

Fig. S12. Macroevolutionary rates of net diversification (green), niche (red), and phenotypic traits (blue) estimated by BAMM, with color density shading to denote confidence on evolutionary rates thought time. The gray curve in the background shows the fluctuation of global temperature, using dataset from ref. 45.

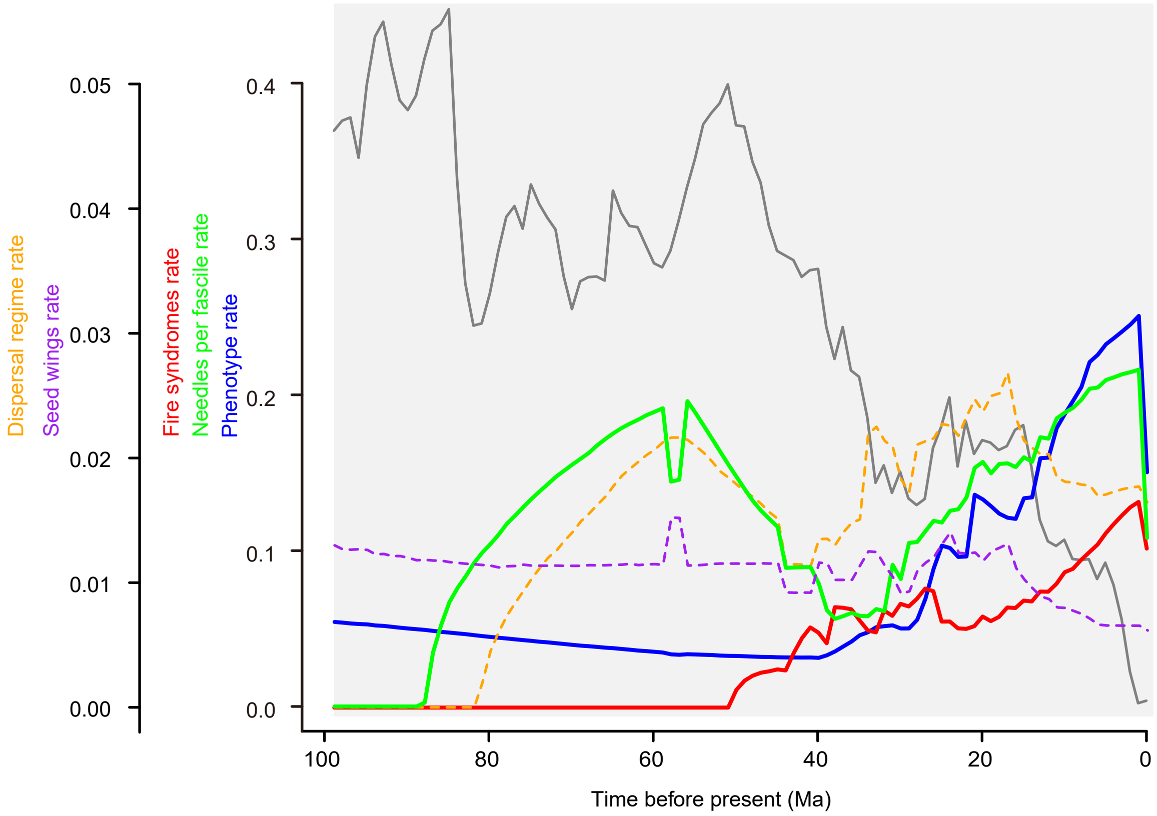

Fig. S13. Rates for phenotype (blue), needles per fascile (green), fire syndromes (red), seed wings (purple), and dispersal regime (orange) estimated by BAMM. The gray curve in the background shows the fluctuation of global temperature, using dataset from ref. 45.

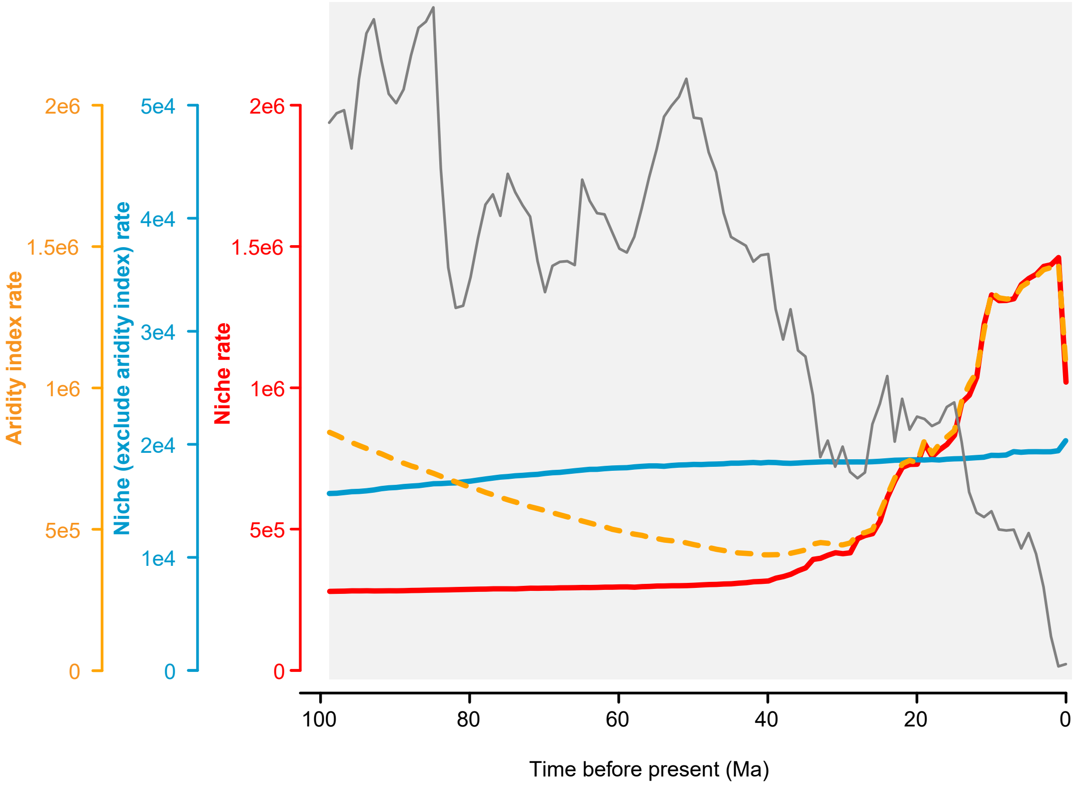

Fig. S14. Rates for niche (red), niche (exclude aridity index) (deepskyblue3), and aridity index (orange) estimated by BAMM. The gray curve in the background shows the fluctuation of global temperature, using dataset from ref. 45.

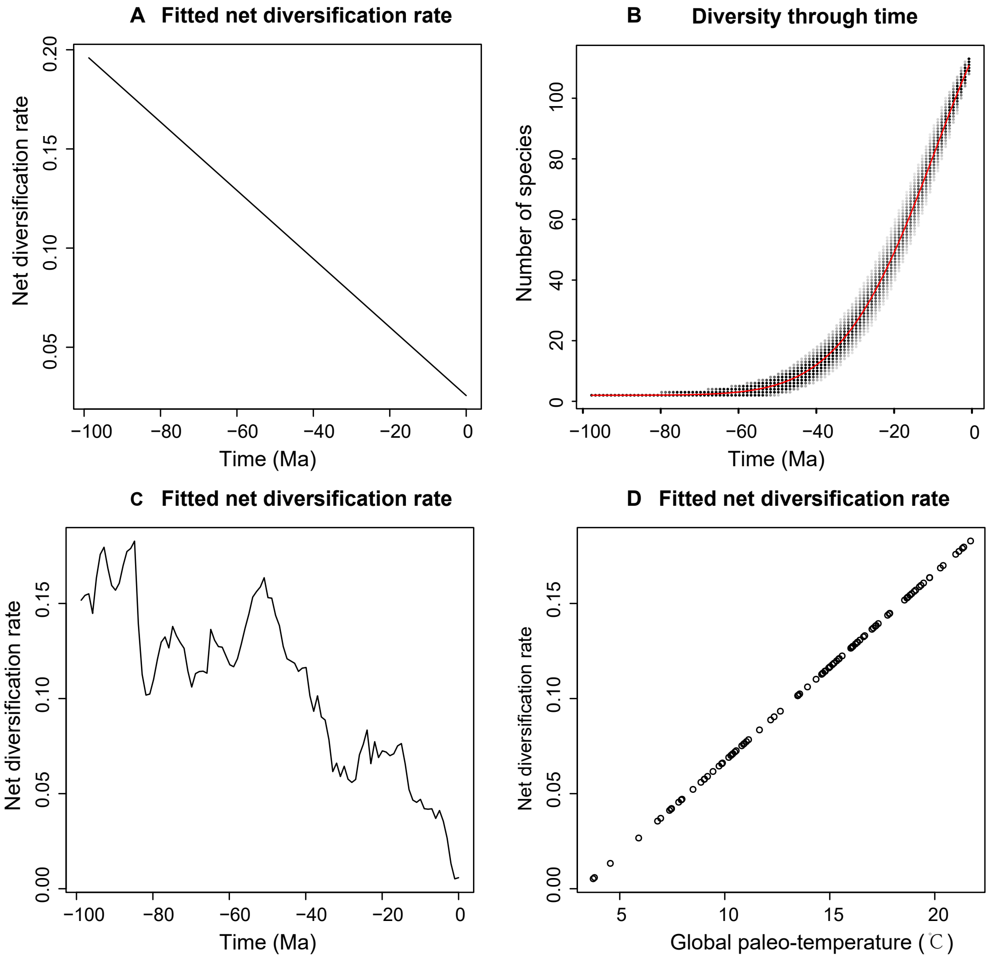

Fig. S15. Plots showing the estimated net diversification rates through time (A, B) and temperature (C, D) estimated by RPANDA. Solid red curve in B is the diversity-through-time (DDT) curve using the deterministic approach by RPANDA. Probability distribution (in gray scale) of the number of extant species is at each 1 Myr interval for the pine phylogeny.

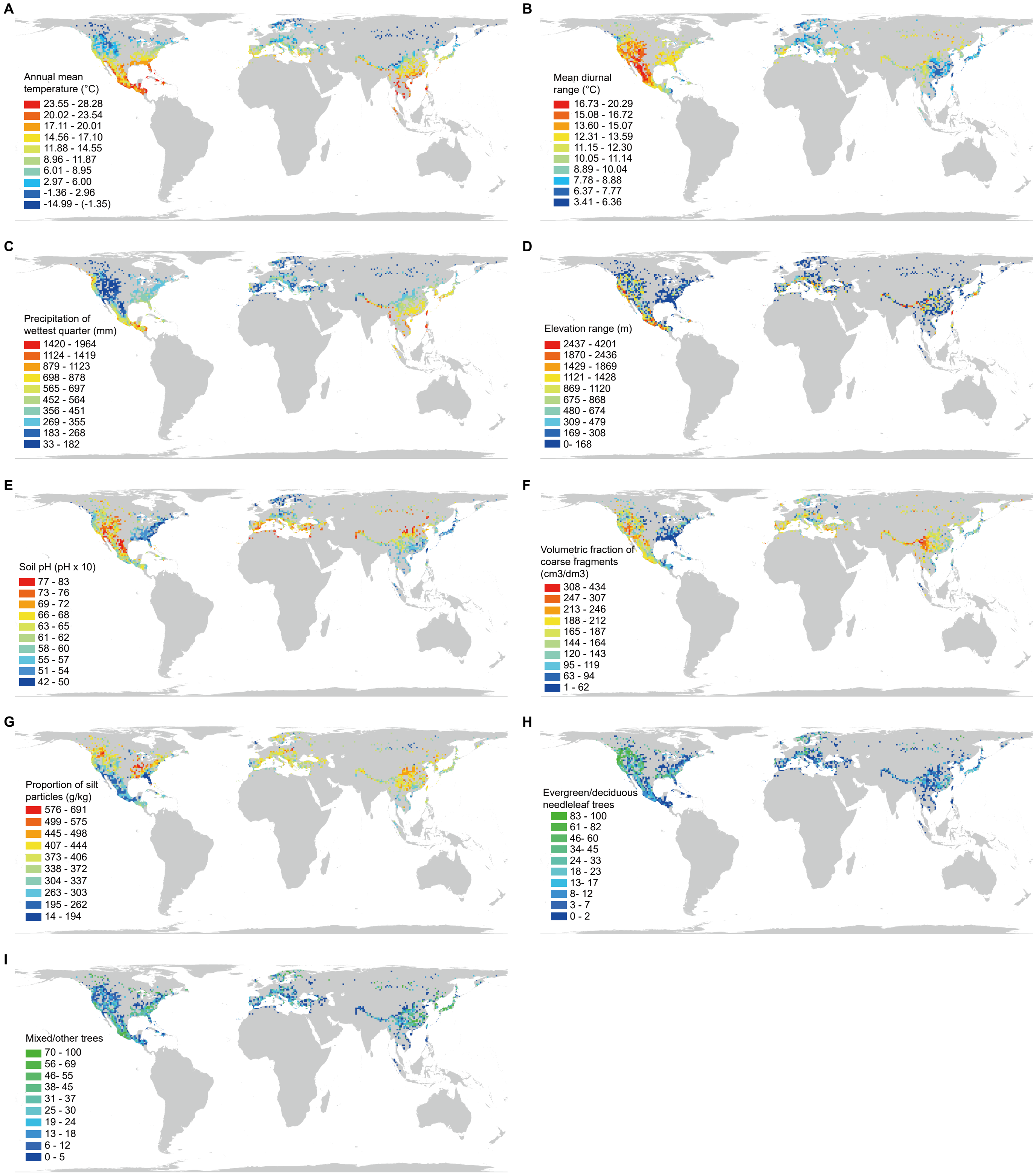

Fig. S16. Environmental variables and their global variation across grid cells of 100 km x 100 km. A - C, Climatic variables. D, Topographical variable. E - G, Soil variables. H, I, Landcover variables.

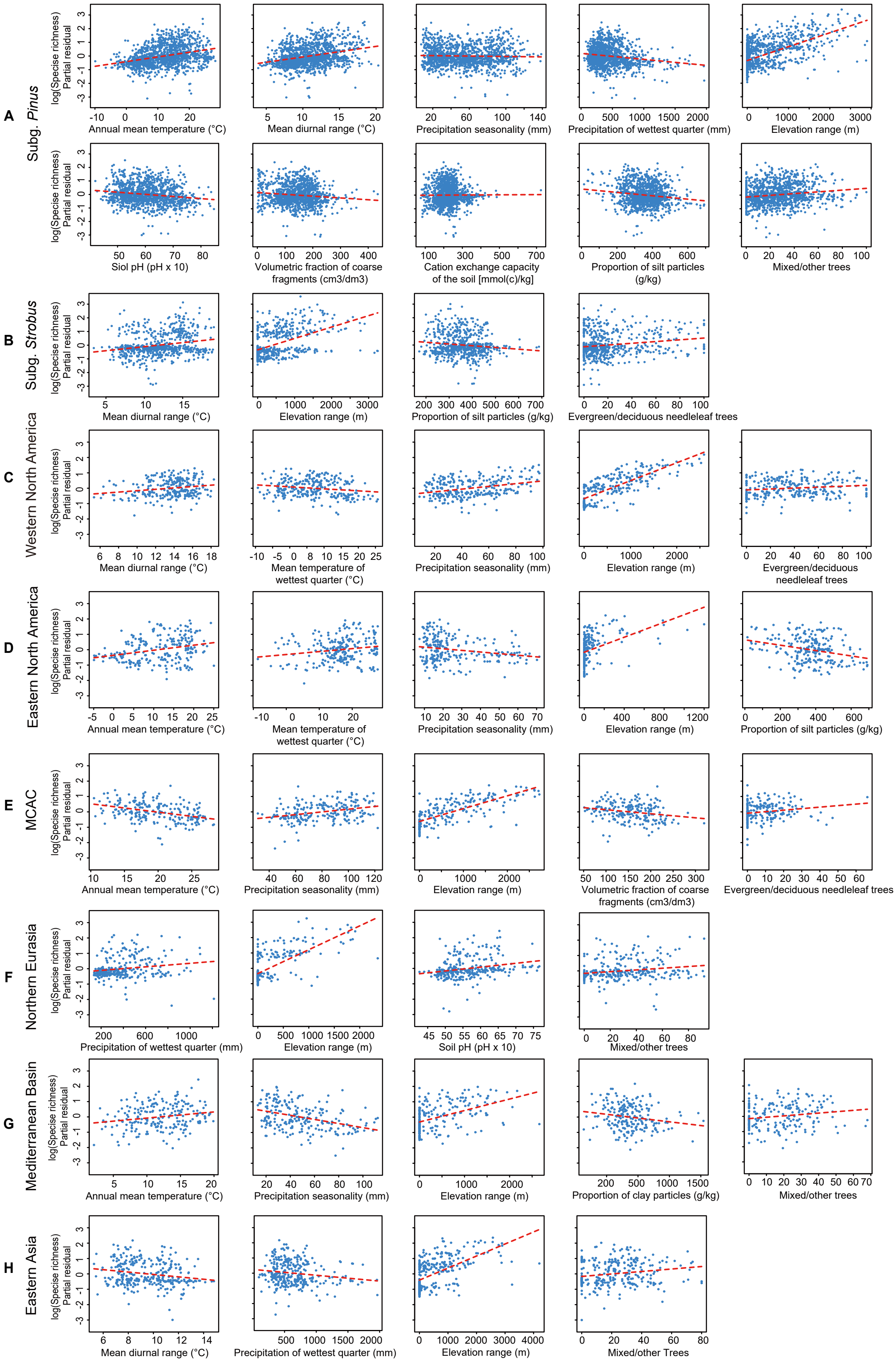

Fig. S17. Determinants of species diversity for two subgenera and six biogeographic regions. A - H, Relationships between predictor variables and species richness based on the multi-predictor model across grid cells of 100 km x 100 km. MCAC represents Mexico, Central America, and Caribbean

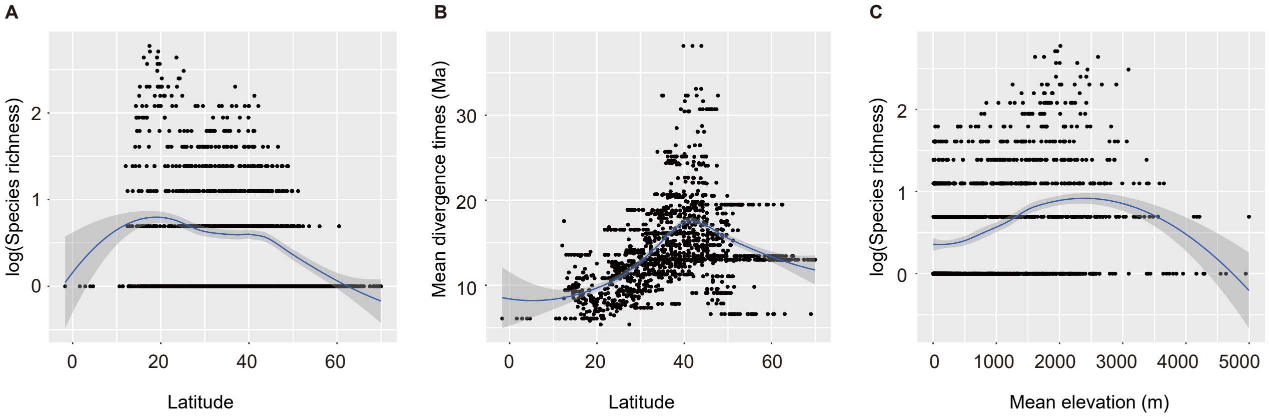

Fig. S18. Latitudinal patterns of species richness (A) and mean divergence times (B), and the elevational pattern of species richness of pines (C). The blue line indicates the general tendency fitting with the local polynomial regression (LOESS) implemented in the "ggplot2" package (46) using R (v.3.6.2) (44). The grey shadow shows an estimated standard error for each predicted value.

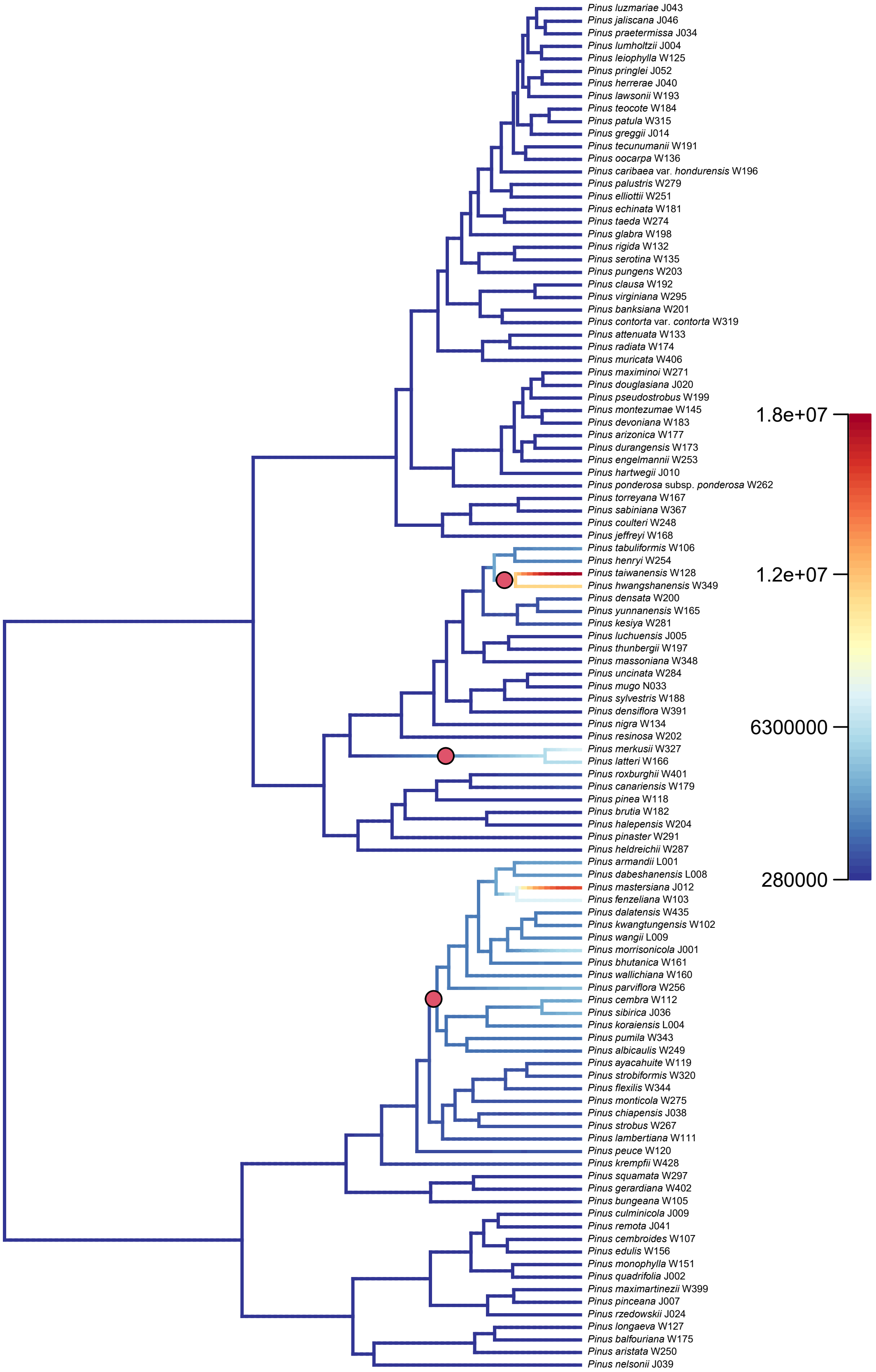

Fig. S19. Phylorate plot showing niche rate on a scale from low to high values (blue to red) along each branch of the pine phylogeny. Red circles along branches represent major ecological shifts.

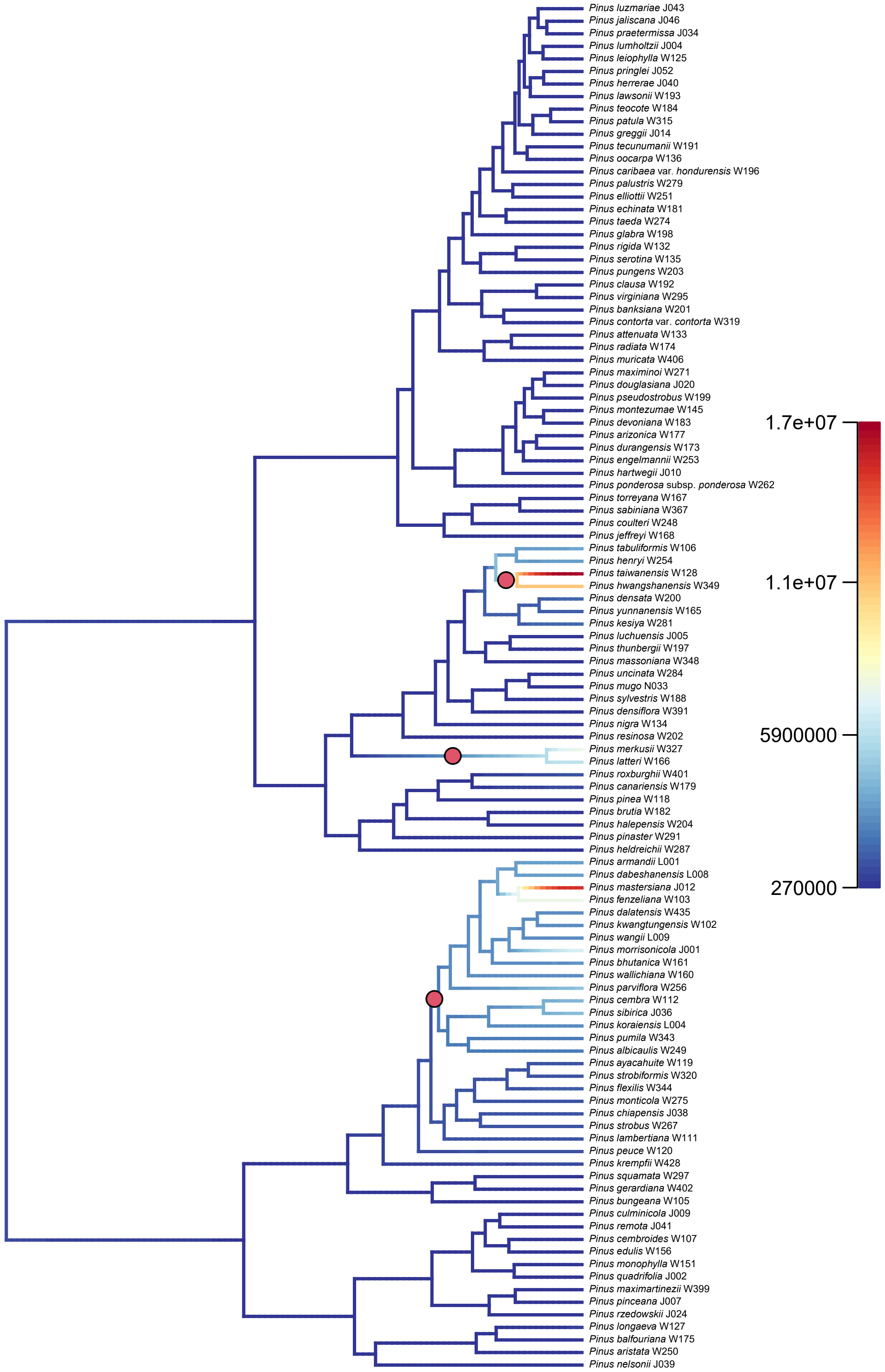

Fig. S20. Phylorate plot showing aridity index rate on a scale from low (blue, arid) to high values (red, semi-arid) along each branch of the pine phylogeny. Red circles along branches represent major aridity index shifts.

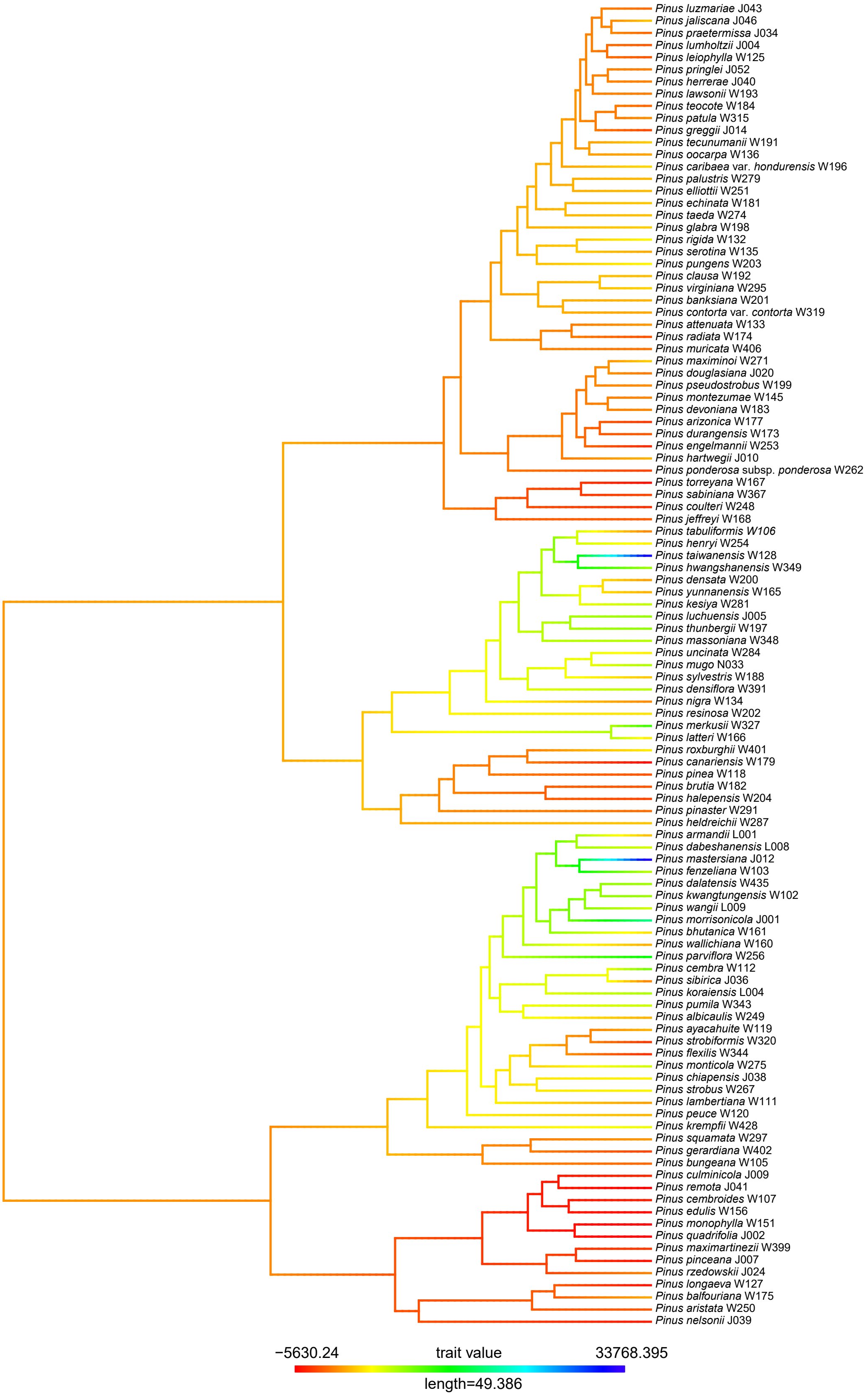

Fig. S21. Ancestral reconstruction across *Pinus* for PC1 (Principle Component 1) of the niche dataset. Branches are colored in a rainbow scale from low ordinated values (red and yellow; arid habitats) to high ordinated values (green and blue; semi-arid habitats).

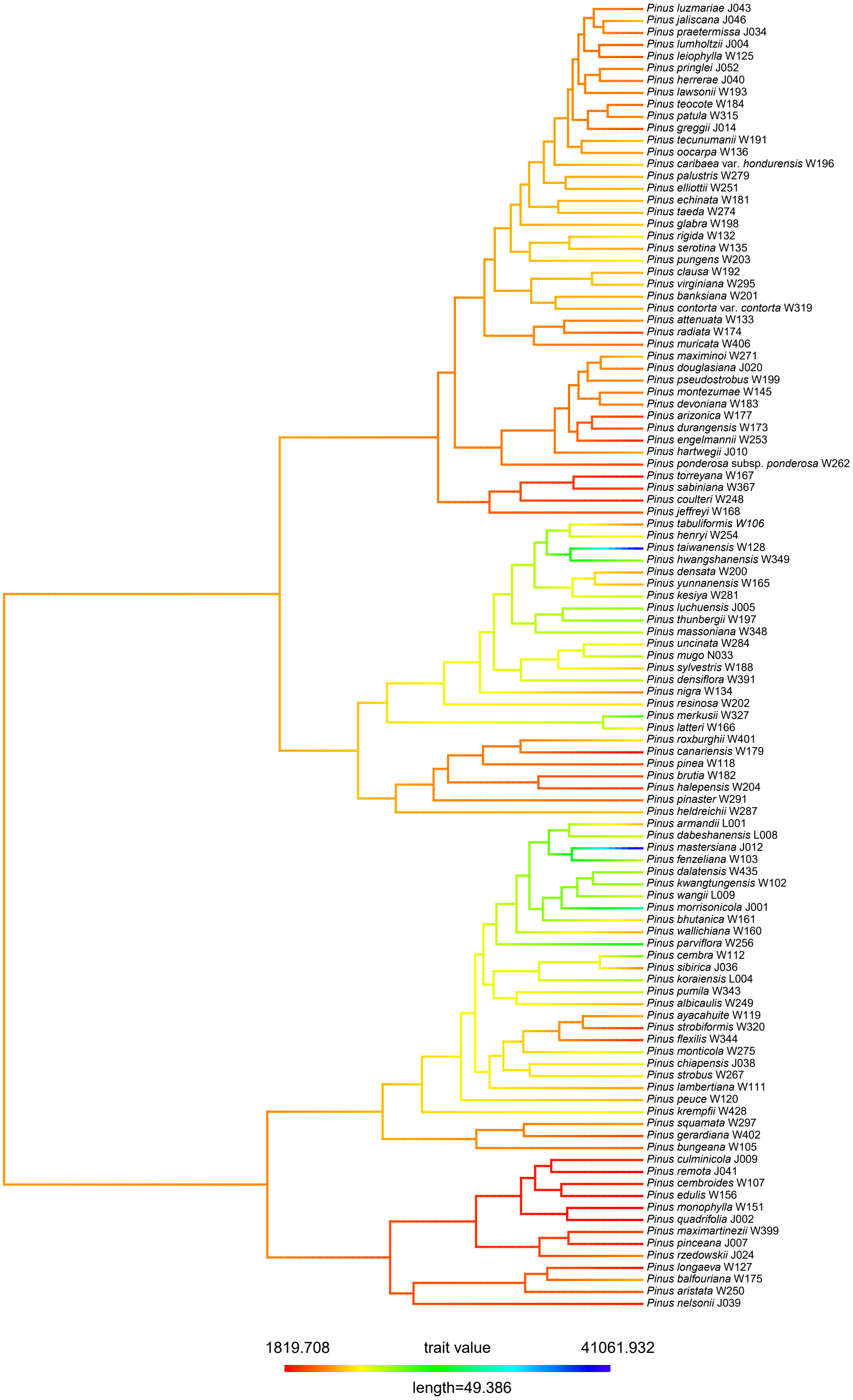

Fig. S22. Ancestral reconstruction across *Pinus* for PC1 (Principle Component 1) of the aridity index dataset. Branches are colored in a rainbow scale from low ordinated values (red and yellow; arid habitats) to high ordinated values (green and blue; semi-arid habitats).

Supplementary Table:

Table S1. The 16 fossils used in the calibration following ref. 27.

No.	Fossils	Divergence dated	Geological timescale
1	*P. yorkshirensis*	129 Ma; stem *Pinus*	Wealden Formation: 131 - 129 Ma
2	*P. triphylla*	90 Ma; stem subgenus *Pinus*	Turonian: 93.9 - 89.7 Ma
3	*P. crossii*	27 Ma; stem subsection *Balfourianae*	Creed Flora, Colorado: 27.2 Ma
4	*P. sanjuanenesis*	27 Ma; stem subsection *Cembroides*	Creed Flora, Colorado: 27.2 Ma
5	*P. lindgrenii*	6 Ma; MRCA of *P. edulis* - *P. johannis* clade	Chalk Hills Formation, Idaho: 7 - 6 Ma
6	*P. florissantii*	34 Ma; stem subsection *Strobus*	Early Oligocene: 33.9 - 28.1 Ma
7	*P. baileyi*	45 Ma; stem section Pinus	Thunder Mountain, Florida: 46 - 45 Ma
8	*P. halepensis*	12.8 Ma; stem *P. halepensis*	Badenian: 16.3 - 12.8 Ma
9	*P. canariensis*	12.8 Ma; stem (*P. canariensis* + *P. roxburghii)* clade	Styrian Basin: 16.3 - 12.8 Ma
10	*P. densiflora*	1.1 Ma; stem *P. densiflora*	Lower part of Osaka Group in Pleistocene: 0.86 - 1.75 Ma
11	*P. premassoniana*	5.3 Ma; stem *P. massoniana*	Shengxian Formation: 11.6 - 5.3 Ma
12	*P. prekesiya*	5.3 Ma; *P. yunnanensis* – *P. kesiya* divergence	Xiaolongtain Formation: 11.6 - 5.3 Ma
13	*P. fujiii*	15 Ma; stem MRCA of *P. kesiya* + *P. tabuliformis*	Lower to Upper Miocene: 23 - 5.3 Ma
14	*P. pieperi*	5 Ma; (*P. coulteri* + *P. sabiniana*) clade divergence	Mount Eden Flora: 6 - 5 Ma
15	*P. storeyana*	12 Ma; member *Attenuatae* clade	Coal Valley Formation: 12.5 Ma
16	*P. radiata*	0.4 Ma; stem *P. radiata*	Santa Barbara Formation: 0.8 - 0.4 Ma

Table S2. Biogeographic models tested in this study.

Model	**LnL**	**numparams**	**d**	**e**	**j**	**AICc**	**AICc_wt**
DEC	-136	2	0.0015	1.00E-12	0	276.1	1.20E-09
DEC+J	-115.7	3	0.0002	1.00E-12	0.024	237.6	0.28
DIVALIKE	-138.7	2	0.0022	1.00E-12	0	281.5	8.10E-11
DIVALIKE+J	-114.7	3	0.0003	1.00E-12	0.024	235.7	0.72
BAYAREALIKE	-180.2	2	0.0012	0.023	0	364.5	7.90E-29
BAYAREALIKE+J	-120	3	0.0001	0.0003	0.027	246.3	0.0036

Table S3. Stochastic mapping statistics and data from BioGeoBears results.

Summary of dispersal events counts (mean)

	A	B	C	D	E	F
A	-	1.58	2.72	0.42	0.32	0.04
B	1.06	-	1.5	0.06	0.12	0.14
C	2.92	0.5	-	0	0	0.2
D	1.32	0.1	0	-	2.34	1.44
E	0.32	0.52	0	0.46	-	2.04
F	0.48	0.4	0.24	2.1	1.04	-

Range expansion event counts (mean)

	A	B	C	D	E	F
A	-	0.04	0.06	0	0.04	0.02
B	0.02	-	0.02	0	0.06	0
C	0.04	0	-	0	0	0.04
D	0.04	0.02	0	-	2	0
E	0.02	0.06	0	0.02	-	0.08
F	0	0.02	0.08	0.1	0.06	-

Founder event counts (mean)

	A	B	C	D	E	F
A	-	1.54	2.66	0.42	0.28	0.02
B	1.04	-	1.48	0.06	0.06	0.14
C	2.88	0.5	-	0	0	0.16
D	1.28	0.08	0	-	0.34	1.44
E	0.3	0.46	0	0.44	-	1.96
F	0.48	0.38	0.16	2	0.98	-

Summary counts of BSMs

	founder	a	d	e	subset	vicariance	sympatry	ALL_disp	ana_disp	all_ana	all_clado	total_events
means	21.54 (19%)	0	2.84 (2%)	0	0	3.14 (3%)	87.32 (76%)	24.38 (21%)	2.84 (2%)	2.84 (2%)	112 (96%)	114.8
stdevs	1.45	0	0.91	0	0	1.26	0.91	1.18	0.91	0.91	0	0.91
sums	1077	0	142	0	0	157	4366	1219	142	142	5600	5742

A, Western North America; B, Eastern North America; C, Mexico, Central America, and Caribbean; D, Northern Eurasia; E, Mediterranean Basin; F, Eastern Asia. ALL_disp, all dispersal (mean of all observed anagenetic 'a', 'd' dispersals, plus cladogenetic founder/jump dispersal); ana_disp, Anagenetic dispersal (mean of all observed anagenetic 'a' or 'd' dispersals); all_ana, all Anagenetic (mean of all observed 'a', 'd' and 'e'); all_clado, all Cladogenetic (mean of all sympatry, plus founder, and vicariance).

Table S4. Niche PCA loadings.

variables	PC1 loading
phh2o	-0.613391889
bio2	-0.501968112
bio7	-0.405529848
pet	-0.337269134
cfvo	-0.33472032
bio4	-0.310469116
bio5	-0.298713285
sand	-0.222467155
bio15	-0.14895368
consensus1	-0.127323009
bio10	-0.126309514
bio8	0.008842674
elevation	0.02778372
consensus4	0.028489711
bio3	0.091107185
bio1	0.123633668
bio9	0.14722527
silt	0.155088217
clay	0.197336237
bio11	0.219586536
cec	0.222182933
bio6	0.278521715
bio19	0.451583192
bio14	0.572191209
bio17	0.581013812
soc	0.660461024
bio13	0.767956855
bio16	0.778427482
bio18	0.781159968
bio12	0.886943091
ai	0.999990858

Abbreviations and explanations of environmental variables refer to Dataset S3.

Table S5. Phenotype PCA loadings

Traits	PC1 loadings
Needles per fascile	0.989035885
Seed wing	-0.002285033
Dispersal regime	0.126078035
Fire syndrome	0.319801275

Table S6. RPANDA diversification models fit of time-dependence.

Model	λ parameters	μ parameters	AICc
bcst_d0.t	0.0516		881.3831
bcst_dcst.t	0.0517	-1.075e-9	883.4561
bcst_dexp.t	0.0518	-1.192e-8, 5.377e-3	885.5672
bcst_dlin.t	0.0558	0.0103, -0.0008	883.9389
bexp_d0.t	0.0460, 0.0072		881.4507
bexp_dcst.t	0.0460, 0.0072	1.337e-8	883.5618
bexp_dexp.t	0.0461, 0.0074	1.182e-8, 8.707e-3	885.7175
blin_d0.t	0.0256, 0.0017		875.557
blin_dcst.t	0.0256 0.0017	7.837e-10	877.6681

Table S7. RPANDA diversification models fit of temperature-dependence.

Model	λ parameters	μ parameters	AICc
bcst_d0.x	0.0062		861.1247
bcst_dcst.x	0.0062	-6.6383e-8	863.1978
bcst_dexp.x	0.0532	-5.0509e-9, 4.6616e-3	884.412
bcst_dlin.x	0.0569	0.0200, -0.0033	883.7449
bexp_d0.x	0.0272, 0.0745		873.9308
bexp_dcst.x	0.0272, 0.0744	1.6731e-8	876.0419
bexp_dexp.x	0.0272, 0.0744	-5.6800e-09, -6.1604e-02	878.1921
blin_d0.x	-0.0313, 0.0099		856.891
blin_dcst.x	-0.0312, 0.0099	-3.6707e-08	859.0023

bcst: birth with a constant speciation rate.

bexp: birth with an exponential speciation rate.

blin: birth with a linear speciation rate.

d0: without extinction.

dcst: death with a constant extinction rate.

dexp: death with an exponential extinction rate.

dlin: death with a linear extinction rate.

Table S8. Phylogenetic signals of niche and phenotypic traits evaluated by Blomberg's *K* and Pagel's λ values.

	Blomberg's *K*	P	Pagel's λ	P
Niche	0.358	0.004	0.619	1.17E-06
Phenotypic traits	0.354	0.001	0.669	1.41E-07

Table S9. Summary results from multi-predictor regression models explaining the pine species richness within 100 km × 100 km grid cells at global, subgeneric, and regional scales.

	Globe		Subgenus *Pinus*		Subgenus *Strobus*	
	OLS	SAR	OLS	SAR	OLS	SAR
Intercept	0	-0.004	0	-0.005	0	0
BIO1	0.182***	0.123***	0.240***	0.165***	-	-
BIO2	0.247***	0.225***	0.224***	0.202***	0.184***	0.171***
BIO8	-	-	-	-	-	-
BIO15	-	-	-0.022	-0.023	-	-
BIO16	-0.060*	-0.035	-0.121***	-0.061	-	-
Elevation range	0.603***	0.547***	0.560***	0.480***	0.496***	0.485***
phh2o	-0.123***	-0.107***	-0.123***	-0.089*	-	-
cfvo	-0.075***	-0.053*	-0.082**	-0.061*	-	-
clay	-0.033	-0.007	-	-	-	-
silt	-0.074***	-0.070**	-0.106***	-0.076**	-0.100**	-0.082*
cec	-	-	0.004	-0.004	-	-
consensus1	0.078***	0.0613**	-	-	0.146***	0.126***
consensus4	0.098***	0.0605**	0.124***	0.072**	-	-
R^2	0.53	0.53	0.49	0.49	0.36	0.36
R^2 *full*	-	0.59	-	0.58	-	0.40
Moran’s *I*	0.202***	-0.014	0.249***	-0.021	0.143***	-0.005

Continued Table S9

	Western North America		Eastern North America		MCAC	
	OLS	SAR	OLS	SAR	OLS	SAR
Intercept	0	-0.001	0	-0.001	0	-0.008
BIO1	-	-	0.229**	0.239**	-0.233***	0.242***
BIO2	0.110**	0.102*	-0.092	-0.092	-	-
BIO8	-0.100*	-0.094*	0.139*	0.140*	-	-
BIO15	0.193***	0.190***	-0.169**	-0.164*	0.193***	0.165**
BIO16	-	-	-	-	-	-
Elevation range	0.706***	0.698***	0.402***	0.397***	0.619***	0.600***
phh2o	-	-	-	-	-	-
cfvo	-	-	-0.105	-0.098	-0.131*	-0.126*
clay	-	-	-	-	-	-
silt	-	-	-0.234***	-0.227***	-0.080	-0.074
cec	-	-	-	-	-	-
consensus1	0.083*	0.075	-	-	0.104*	0.100*
consensus4	-	-	-	-	-	-
R^2	0.70	0.70	0.41	0.41	0.68	0.68
R^2 *full*	-	0.70	-	0.41	-	0.69
Moran's *I*	0.082*	-0.006	0.038	-0.002	0.108*	0.003

Continued Table S9

	Northern Eurasia		Mediterranean Basin		Eastern Asia	
	OLS	SAR	OLS	SAR	OLS	SAR
Intercept	0	-0.002	0	-0.004	0	0.004
BIO1	-	-	0.162*	0.0831	-	-
BIO2	-	-	-	-	-0.157**	-0.123
BIO8	0.087	0.075	-	-	-	-
BIO15	-0.101	-0.094	-0.319***	-0.219*	-	-
BIO16	0.119*	0.122*	-	-	-0.128*	-0.083
Elevation range	0.672***	0.656***	0.416***	0.403***	0.581***	0.544***
phh2o	0.165**	0.156**	-	-	-	-
cfvo	-	-	-0.116	-0.085	-	-
clay	-	-	-0.151*	-0.061	-	-
silt	-	-	-	-	-	-
cec	-	-	-	-	-	-
consensus1	-	-	0.111	0.120*	-	-
consensus4	0.129**	0.118**	0.142*	0.119*	0.149***	0.125*
R^2	0.56	0.56	0.40	0.39	0.41	0.40
R^2 *full*	-	0.57	-	0.47	-	0.45
Moran's *I*	0.059*	-0.002	0.171***	0.006	0.137**	0.002

The non-spatial ordinary least square (OLS) regression and the spatial simultaneous autoregressive (SAR) model are compared. All predictors were scaled to a mean of zero and variance of one before the analysis. Standardised coefficients for the minimum adequate mode with the lowest Akaike Information Criterion (AIC), the explained variance of the environmental variables (R^2), the explained variance of the full SAR model (R^2*full*) including both environment and space, and the Moran's *I* are given. Significance levels: ****p* < 0.001; ***p* < 0.01; **p* < 0.05. MCAC represents Mexico, Central America, and Caribbean. Abbreviations and explanations of predictor variables refer to *SI Appendix*, Table S10.

Table S10. The final predictor variables used for pine species richness analysis in this study.

Abbreviation	Variable description	Unit
Climate		
BIO1	Annual mean temperature	°C
BIO2	Mean diurnal range	°C
BIO8	Mean temperature of wettest quarter	°C
BIO15	Precipitation seasonality	mm
BIO16	Precipitation of wettest quarter	mm
Topography		
Elevation range	Max elevation – min elevation	m
Soil		
phh2o	Soil pH	pH x 10
cfvo	Volumetric fraction of coarse fragments	cm3/dm3
clay	Proportion of clay particles in the fine earth fraction (0–2 µm)	g/kg
silt	Proportion of silt particles in the fine earth fraction (2–50 µm)	g/kg
cec	Cation exchange capacity of the soil	mmol(c)/kg
Landcover		
consensus1	Evergreen/deciduous needleleaf trees	%
consensus4	Mixed/other trees	%

SI References

1. A. Farjon, *A Handbook of the World's Conifers* (Brill, Leiden, 2017).
2. R. Price, A. Liston, S. Strauss, "Phylogeny and systematics of *Pinus*" in Ecology and Biogeography of *Pinus,* D. M. Richardson, Ed. (Cambridge University Press, Cambridge, 1998), pp. 49-68.
3. L. G. Fu, N. Li, R. R. Mill, "Pinaceae" in Flora of China, vol. 4*,* Z. Y. Wu, P. H. Raven, D. Y. Hong, Eds. (Science Press and Missouri Botanical Garden Press, Beijing and St. Louis, 1999), pp. 11-52.
4. D. S. Gernandt, G. Geada López, S. Ortiz García, A. Liston, Phylogeny and classification of *Pinus*. *Taxon* **54**, 29-42 (2005).
5. J. A. Pérez-de la Rosa, *Pinus georginae* (Pinaceae), a new species from western Jalisco, Mexico. *Brittonia* **61**, 56-61 (2009).
6. J. H. Ran, T. T. Shen, H. Wu, X. Gong, X. Q. Wang, Phylogeny and evolutionary history of Pinaceae updated by transcriptomic analysis. *Mol. Phylogenet. Evol.* **129**, 106-116 (2018).
7. A. M. Bolger, M. Lohse, B. Usadel, Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114-2120 (2014).
8. M. G. Grabherr *et al.*, Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29**, 644-652 (2011).
9. W. Li, A. Godzik, Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658-1659 (2006).
10. L. Fu, B. Niu, Z. Zhu, S. Wu, W. Li, CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150-3152 (2012).
11. B. J. Haas *et al.*, De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat. Protoc.* **8**, 1494-1512 (2013).
12. I. Ebersberger, S. Strauss, A. von Haeseler, HaMStR: profile hidden markov model based search for orthologs in ESTs. *BMC Evol. Biol.* **9**, 157 (2009).
13. D. M. Emms, S. Kelly, OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol.* **16**, 157 (2015).
14. K. Katoh, D. M. Standley, MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772-780 (2013).
15. S. R. Eddy, Accelerated Profile HMM Searches. *PLoS Comput. Biol.* **7**, e1002195 (2011).
16. P. Kück, G. C. Longo, FASconCAT-G: extensive functions for multiple sequence alignment preparations concerning phylogenetic studies. *Front. Zool.* **11**, 81 (2014).
17. L. T. Nguyen, H. A. Schmidt, A. von Haeseler, B. Q. Minh, IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268-274 (2015).
18. D. T. Hoang, O. Chernomor, A. von Haeseler, B. Q. Minh, L. S. Vinh, UFBoot2: improving the ultrafast bootstrap approximation. *Mol. Biol. Evol.* **35**, 518-522 (2018).
19. M. Suyama, D. Torrents, P. Bork, PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res.* **34**, W609-612 (2006).
20. M. Wu, S. Chatterji, J. A. Eisen, Accounting for alignment uncertainty in phylogenomics. *PLoS ONE* **7**, e30288 (2012).
21. S. Kalyaanamoorthy, B. Q. Minh, T. K. F. Wong, A. von Haeseler, L. S. Jermiin, ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* **14**, 587-589 (2017).
22. E. Sayyari, J. B. Whitfield, S. Mirarab, Fragmentary gene sequences negatively impact gene tree and

species tree reconstruction. *Mol. Biol. Evol.* **34**, 3279-3291 (2017).

23. C. Zhang, E. Sayyari, S. Mirarab, "ASTRAL-III: increased scalability and impacts of contracting low support branches" in Comparative Genomics, J. Meidanis, L. Nakhleh, Eds. (Springer International Publishing, Cham, Switzerland, 2017), pp. 53-75.
24. C. Zhang, M. Rabiee, E. Sayyari, S. Mirarab, ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinformatics* **19**, 153 (2018).
25. E. Sayyari, S. Mirarab, Fast coalescent-based computation of local branch support from quartet frequencies. *Mol. Biol. Evol.* **33**, 1654-1668 (2016).
26. C. I. Millar, "Early evolution of Pines" in Ecology and Biogeography of *Pinus,* D. M. Richardson, Ed. (Cambridge University Press, Cambridge (UK), 1998), pp. 69-91.
27. B. Saladin *et al.*, Fossils matter: improved estimates of divergence times in *Pinus* reveal older diversification. *BMC Evol. Biol.* **17**, 95 (2017).
28. A. B. Leslie *et al.*, An overview of extant conifer evolution from the perspective of the fossil record. *Am. J. Bot.* **105**, 1531-1544 (2018).
29. P. E. Ryberg *et al.*, Reconsidering relationships among stem and crown group Pinaceae: oldest record of the genus *Pinus* from the early Cretaceous of Yorkshire, United Kingdom. *Int. J. Plant Sci.* **173**, 917-932 (2012).
30. J. J. F. Meijer, Fossil woods from the late Cretaceous Aachen Formation. *Rev. Palaeobot. Palynol.* **112**, 297-336 (2000).
31. A. Willyard, J. Syring, D. S. Gernandt, A. Liston, R. Cronn, Fossil calibration of molecular divergence infers a moderate mutation rate and recent radiations for *Pinus*. *Mol. Biol. Evol.* **24**, 90-101 (2007).
32. D. I. Axelrod, Cenozoic history of some eestern American pines. *Ann. Mo. Bot. Gard.* **73**, 565-641 (1986).
33. D. M. Erwin, H. E. Schorn, *Pinus baileyi* (section *Pinus*, Pinaceae) from the Paleogene of Idaho, USA. *Am. J. Bot.* **93**, 197-205 (2006).
34. Z. Yang, PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586-1591 (2007).
35. M. dos Reis, Z. Yang, Approximate likelihood calculation on a phylogeny for Bayesian estimation of divergence times. *Mol. Biol. Evol.* **28**, 2161-2172 (2011).
36. G. W. Rothwell, G. Mapes, R. A. Stockey, J. Hilton, The seed cone *Eathiestrobus* gen. nov.: fossil evidence for a Jurassic origin of Pinaceae. *Am. J. Bot.* **99**, 708-720 (2012).
37. B. Rannala, Z. Yang, Inferring speciation times under an episodic molecular clock. *Syst. Biol.* **56**, 453-466 (2007).
38. A. Rambaut, A. J. Drummond, D. Xie, G. Baele, M. A. Suchard, Posterior summarization in bayesian phylogenetics using Tracer 1.7. *Syst. Biol.* **67**, 901-904 (2018).
39. J. Fox, S. Weisberg, *An R Companion to Applied Regression, Third Edition* (Sage, Thousand Oaks, California, 2019).
40. K. P. Burnham, D. R. Anderson, *Model Selection and Multimodel Inference-A Practical Information-Theoretic Approach* (Springer-Verlag, New York, 2002).
41. B. A. Hawkins, J. A. F. Diniz-Filho, L. Mauricio Bini, P. De Marco, T. M. Blackburn, Red herrings revisited: spatial autocorrelation and parameter estimation in geographical ecology. *Ecography* **30**, 375-384 (2007).
42. R. S. Bivand, D. W. S. Wong, Comparing implementations of global and local indicators of spatial

association. *Test* **27**, 716-748 (2018).

43. W. D. Kissling, G. Carl, Spatial autocorrelation and the selection of simultaneous autoregressive models. *Global Ecol. Biogeogr.* **17**, 59-71 (2008).
44. R Core Team, *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna, Austria, 2019).
45. M. Sun *et al.*, Recent accelerated diversification in rosids occurred outside the tropics. *Nat. Commun.* **11**, 3333 (2020).
46. H. Wickham, *ggplot2: Elegant Graphics for Data Analysis* (Springer-Verlag New York, 2016).