Supplementary Material & Methods

Trait data collection and selection

We searched for trait combinations that promote species with the potential to become invasive. To investigate this question, we extracted ant species traits from the *AntProfiler* collaborative database (https://antprofiler.ese.u-psud.fr). This database includes information on 2,176 ant species (134 genera) distributed across ecozones for the entire world (Fig. S1), an ant phylogeny (Fig. S2), and ecological traits related to their occurrence, morphology, behavior, and invasive status (1).

From this database, all traits were used for the missing data imputation step, but we reduced the number of traits for the predictive model. The full description of the trait selection process is detailed and illustrated in Figure S3. Specifically, we excluded traits that were too correlated with other traits (i.e., Pearson's Chi-squared Test for Count Data $p<0.05$). We also removed two traits that are not available prior to species introduction to avoid restricting our profiling tool to species that already moved from their native range: exotic status and distribution by ecozone. We then removed traits that were not significantly different between invasive and non-invasive species. Finally, we removed species that had too many missing values in order to have a maximum of ~60 % missing values per trait (2).This resulted in a final subset of four traits: Colony foundation, Colony structure, Nesting type and Association with disturbed areas and 992 species (out of 2,176). Table S1details the coding rules for the four traits.

Missing data handling

Despite our data collection efforts, the database contained many missing values. The two possible approaches to deal with missing data are deletion and imputation. Deleting missing data is the simplest and most widely chosen option, but can be highly problematic when data are not missing at random, as this can lead to biased estimates of model parameters and incorrect conclusions (3–7). Removing data would also have drastically reduced the sample size and predictive power of the screening tool, since only 184 species had no missing values for all 4 traits. In this regard, we explored how missing data could be imputed in our database in order to minimize bias and maximize prediction accuracy.

Missing data imputation consists of replacing a missing value by a prediction of its value. It has been proven that repeating missing data imputation multiple times (called multiple imputation) is more accurate than replacing missing values only once in order to accurately recover missing observations and estimate model parameters (3, 4, 7, 8). It was also shown that random forest methods provide highly accurate imputations and that this algorithm performs better when phylogenetic information is included during the process (2, 9). In light of this, we implemented a multiple imputation of our database using the R package missForest (10, 11), while including phylogenetic information following the method of Penone et al. (2). The details of the multiple imputation parametrization are described below. First, we downloaded the most recently published ant genus phylogenies (12, 13), which cover all of the genera present in our database. Second, we measured the strength of the phylogenetic signal in each of our traits in order to decide whether it was informative to include genus-level phylogenetic information for the multiple imputation (14). Figure S4 and Table S4 present the results of this analysis, which enabled us to conclude that a phylogenetic signal was discernible in our four pre-selected traits. This meant that using phylogenetic information to replace the missing values for these traits was likely to improve the quality of the imputation. Third, we ran a principal coordinates analysis on the phylogenetic tree. This type of analysis is routinely used to represent the phylogenetic distances among clades as well as the phylogenetic structure of a tree in the form of a series of eigenvectors (following (15, 16)). These eigenvectors can then be easily incorporated in a model as additional variables to improve the quality of the missing values replacement (2). We calibrated the multiple imputation by choosing both the best parameters and the best trade-off between a low imputation error and as many species as possible (see the procedure summary in Fig. 1, Fig S6 and S7). The details of the multiple imputation calibration are given below. We used out-of-bag error estimations to calibrate the models as these were shown to accurately estimate imputation error (e.g., (11)). We selected the optimal number of eigenvectors to impute each trait (Fig. 1-Band Fig. S5). For the final multiple imputation, each trait was then imputed using its own optimal number of eigenvectors, varying between 1 and 26 according to the trait considered. Because the first eigenvectors tend to split the species relative to the most basal nodes, traits that need less eigenvectors for imputations might potentially be the ones that appeared earlier in the evolutionary history of ants. However, this also depends on the structure of the phylogeny. Note that by imputing the traits using eigenvectors, we might have inflated the phylogenetic signal of the traits and thus the phylogenetic structure of our models. Finally, we performed 100 imputations using the best parameters found in the preliminary analyses. These were 100 for the number of trees to grow in the forest (ntree=100), 6 for the number of variables randomly selected at each node to set-up the split of the random forest (mtry=6), and 15 for the maximum set of iterations (maxiter=15; see TableS1). We set a maximum of 15 iterations to limit the running time, while leaving enough iterations to meet the stopping criterion. In most cases, the stopping criterion was met before the maximum number of imputations. In the remaining cases, the imputation estimation might be slightly reduced, however, because we repeated the imputation 100 times, this should not affect our results. The percentage of missing values and the imputation error for each trait are provided in Table S1.

Predictive model building

Potential future invasive species

We modeled the invasive profile using generalized linear models with our four selected traits as predictors and a binomial distribution. We excluded interactions between traits to avoid model overfitting. The subsequent analyses were run on this list of 992 species, including all 19 known invasive species (Fig. S8).

We ran 100 models for each of the previously imputed datasets and used model predictions to determine future invasive species. For each model, we identified potentially new invasive ant species as those with a predicted invasiveness probability above the lower 5thcentile of the 19 known invasive species probability distribution. Each identified species was considered to be potentially invasive if it was selected in at least 90 of the 100 models. In addition, we verified whether our models were able to correctly classify known invaders by recoding them as non-invasive and predict them from the models with the remaining 18 invasive species (i.e., 19-1; leave-one-out models). Model-averaged coefficients were very similar to those of the global models; we thus present the results of the global models alone.

Areas at risk

Species Distribution Models (SDMs) are based on correlations between environmental variables and geolocalized species records and can be used to delineate potential species distributions (17). We built SDMs for the ant species found to have similar profiles to those of known invasive species in order to identify the areas that present suitable environmental conditions for these species and thus areas at risk of invasion. SDMs, even if based on climatic variables alone, have been widely recognized as powerful tools to predict the potential distribution of invasive ants (18).

We used as presence points for the models geo-referenced presence records from the Global Biodiversity Information Facility (GBIF, https://www.gbif.org) and Global Ant Biodiversity Informatics project (19), supplemented with data from the literature (20–22) (Fig. S9). Presence data were deeply explored before running the SDMs, removing duplicate observations and those taken from buildings and greenhouses. Pseudo-absences were selected in the study area taking into account the dispersal rate of the modeled species (10 km) to avoid to (1) potentially include presences as absence locations, (same niche), (2) pseudo-replication and (3) the possible issue that absences occur in locations that are suitable for the species but the species has not yet had the time to reach that environment. In addition, pseudo-absences were generated in large numbers: 1000 if we had fewer than 1000 occurrences for that species, 10,000 otherwise. We also made sure that they were equally weighted to the presences, as recommended to obtain the greatest accuracy of the predictions (23).We sourced the 19 bioclimatic variables (averaged from 1970 to 2000) available from the Worldclim 2.0 database (http://www.worldclim.org) at 10 arc min resolution (24). These variables represent a combination of means, extremes, variability, and seasonality of temperature and precipitation data that are known to influence species distribution (25). Then, to ensure that we did not use correlated variables in the SDMs, we measured the correlation between these 19 bioclimatic variables: we used a hierarchical classification method based on a distance metric (Pearson's correlation coefficient at a threshold of 0.8, similarly to (26)). This resulted in eight variable correlation groups (Fig. S10). We then retained only one variable from each correlation group by selecting those that are known to limit the distribution of terrestrial invertebrates and that have already been applied to a range of insect species (27–29). We used seven different SDM algorithms within the 'biomod2' package (30) using the R platform (10): generalized linear model, flexible discriminant analysis, artificial neural network, random forest, generalized boosting model, maximum entropy, and multiple adaptive regression splines. These seven algorithms were used to build an ensemble model (i.e. TSS-weighted average of all models used) that encompassed the variability between them and provided the central tendency (31). This final consensus distribution comprised the weighted mean, proportional to the True Skills Statistics (TSS), of the seven modeling techniques. Two metrics were used to evaluate the accuracy of each SDM: the True Skill Statistics (TSS) (32) and the Area Under the receiver operating characteristic Curve (AUC) (33).

Ensemble models were run for each of the predicted invaders with sufficient occurrence points in order to produce individual climatic suitability distribution maps (Fig. S11). It is noteworthy that two species (*Formica yessensis* and *Aphaenogaster spinosa*) had very few occurrence points (24 and 42, respectively), meaning that the resulting distribution maps should be taken with additional caution. Finally, we also combined these individual predictions by summing each species potential distribution as binary maps in order to obtain a cumulative invasion risk map from these future invaders. Binary transformation was based on the threshold that maximized the TSS for each species. For all steps, the R code is available at https://github.com/caterinap/Antprofiler. The cleaned-up dataset is available upon simple request to the corresponding author.

References

1. Bertelsmeier C, Luque GM, Confais A, Courchamp F (2013) Ant Profiler - A database of ecological characteristics of ants (Hymenoptera: Formicidae). *Myrmecological News* 18(March):73–76.
2. Penone C, et al. (2014) Imputation of missing data in life-history trait datasets: Which approach performs the best? *Methods Ecol Evol* 5(9):1–10.

3. Nakagawa S, Freckleton RP (2008) Missing inaction: the dangers of ignoring missing data. *Trends Ecol Evol* 23(11):592–596.

4. Nakagawa S, Freckleton RP (2011) Model averaging, missing data and multiple imputation: A case study for behavioural ecology. *Behav Ecol Sociobiol* 65(1):103–116.

5. Veron S, et al. (2016) Integrating data-deficient species in analyses of evolutionary history loss. *Ecol Evol* 6(23):8502–8514.

6. Hadfield JD, Nakagawa S (2010) General quantitative genetic methods for comparative biology: Phylogenies, taxonomies and multi-trait models for continuous and categorical characters. *J Evol Biol* 23(3):494–508.

7. Ellington EH, et al. (2015) Using multiple imputation to estimate missing data in meta-regression. *Methods Ecol Evol* 6(2):153–163.

8. Horton NJ, Kleinman KP (2007) Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models. *Am Stat* 61(1):79–90.

9. Shah AD, Bartlett JW, Carpenter J, Nicholas O, Hemingway H (2017) Practice of Epidemiology Comparison of Random Forest and Parametric Imputation Models for Imputing Missing Data Using MICE : A CALIBER Study. 179(October):764–774.

10. R Development Core Team (2013) R: A language and environment for statistical computing. *R Found Stat Comput Vienna, Austria* 1. doi:10.1017/CBO9781107415324.004.

11. Stekhoven DJ, Bühlmann P (2012) Missforest-Non-parametric missing value imputation for mixed-type data. *Bioinformatics* 28(1):112–118.

12. Blanchard BD, Moreau CS (2016) Defensive traits exhibit an evolutionary trade-off and drive diversification in ants. *Evolution (N Y)*:1–14.

13. Branstetter MG, Longino JT, Ward PS, Faircloth BC (2017) Enriching the ant tree of life: enhanced UCE bait set for genome-scale phylogenetics of ants and other Hymenoptera. *Methods Ecol Evol* 8(6):768–776.

14. Münkemüller T, et al. (2012) How to measure and test phylogenetic signal. *Methods Ecol Evol* 3(4):743–756.

15. Diniz-Filho JAF, et al. (2012) On the selection of phylogenetic eigenvectors for ecological analyses. *Ecography (Cop)* 35(3):239–249.

16. Guénard G, Legendre P, Peres-Neto P (2013) Phylogenetic eigenvector maps: A framework to model and predict species traits. *Methods Ecol Evol* 4(12):1120–1131.

17. Araújo MB, Peterson a. T (2012) Uses and misuses of bioclimatic envelope modeling. *Ecology* 93(7):1527–1539.

18. Roura-Pascual N, Suarez A (2008) The utility of species distribution models to predict the spread of invasive ants (Hymenoptera: Formicidae) and to anticipate changes in their ranges in the face of global climate change. *Myrmecol News* 11(August):67–77.

19. Guénard B, Weiser MD, Gómez K, Narula N, Economo EP (2017) The Global Ant Biodiversity Informatics (GABI) database: Synthesizing data on the geographic distribution of ant species (Hymenoptera: Formicidae). *Myrmecological News* 24:83–89.

20. Wetterer JK (2013) Worldwide spread of the difficult white-footed ant , Technomyrmex difficilis (Hymenoptera : Formicidae). *Myrmecological News* 18(April):93–97.

21. Wetterer JK (2012) Worldwide spread of Emery's sneaking ant, Cardiocondyla emeryi (Hymenoptera: Formicidae). *Myrmecological News* 17(August):13–20.

22. Wetterer JK (2014) Worldwide Spread of the Lesser Sneaking Ant, Cardiocondyla minutior (Hymenoptera: Formicidae). *Florida Entomol* 97(2):567–574.

23. Barbet-Massin M, Jiguet F, Albert CH, Thuiller W, Alpine E (2012) Selecting pseudo-absences for species distribution models : how , where and how many ? *Methods Ecol Evol* 3(2):327–338.

24. Hijmans RJ, Cameron SE, Parra JL, Jones PG, Jarvis A (2005) Very high resolution interpolated climate surfaces for global land areas. *Int J Climatol* 25(15):1965–1978.

25. Root TL, Price JT, Hall KR, Schneider SH (2003) Fingerprints of global warming on wild animals and plants. *Nature* (tier 2):57–60.

26. Bellard C, Bertelsmeier C, Leadley P, Thuiller W, Courchamp F (2012) Impacts of climate change on the future of biodiversity. *Ecol Lett* 15(4):no-no.

27. Hill MP, Terblanche JS (2014) Niche overlap of congeneric invaders supports a single-species hypothesis and provides insight into future invasion risk: Implications for global management of the Bactrocera dorsalis complex. *PLoS One* 9(2). doi:10.1371/journal.pone.0090121.

28. Hill MP, Gallardo B, Terblanche JS (2017) A global assessment of climatic niche shifts and human influence in insect invasions. *Glob Ecol Biogeogr* 26(6):679–689.

29. De Meyer M, et al. (2010) Ecological niche and potential geographic distribution of the invasive fruit fly Bactrocera invadens (Diptera, Tephritidae). *Bull Entomol Res* 100(1):35–48.

30. Thuiller W, Georges D, Engler R (2014) Package 'biomod2.' (1):89.

31. Araújo MB, New M, Arau MB (2007) Ensemble forecasting of species distributions. *Trends Ecol Evol* 22(1):42–47.

32. Allouche O, Tsoar A, Kadmon R (2006) Assessing the accuracy of species distribution models : prevalence , kappa and the true skill statistic (TSS). *J Appl Ecol* 46:1223–1232.

33. Fielding AH, Bell JF (1997) A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environ Conserv* 24(1):38–49.

Supplementary materials

Figure S1: **Geographic distribution of the species present in our database and included in our model**. Information in black represents invasive species (known invasive species according to the IUCN list), and in other colors non invasive species (all other species in our database). We had information about the ecozone for 961 out of the 992 species in our database.

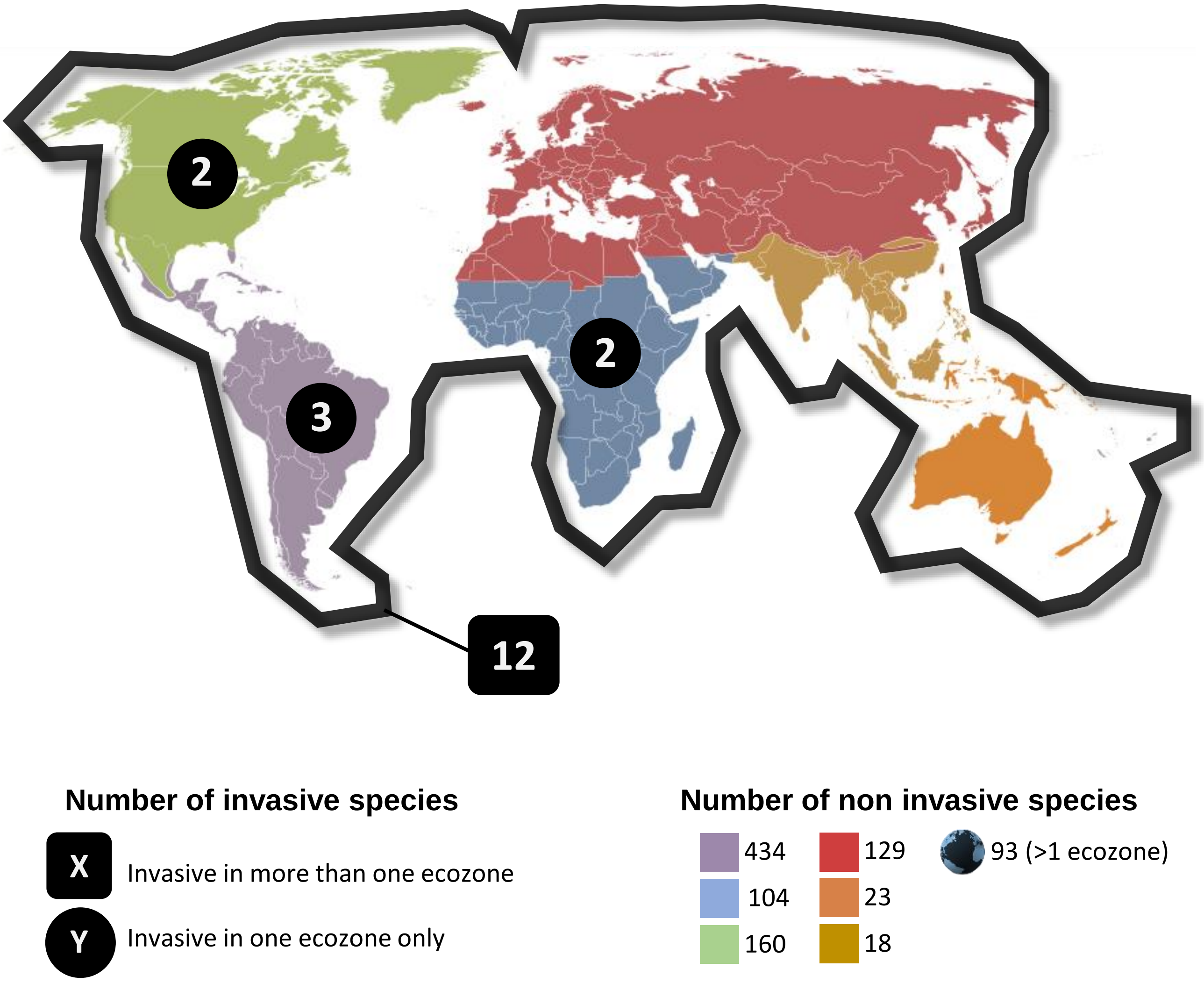

2
2
3
12
Number of invasive species
X Invasive in more than one ecozone
Y Invasive in one ecozone only
Number of non invasive species
🟪 434
🟥 129
🌍 93 (>1 ecozone)
🟦 104
🟧 23
🟩 160
🟨 18

Figure S2: **Taxonomic coverage of our database**. Grey tips show ant genera not present in the *antprofiler* database. Dark blue tips show genus present in the *antprofiler* database, but having too many NA (>70 %) to be included in the model (n = 22 genus). Red tips show genus comprising invasive species (n = 15 genus, included in the model), and light blue tips represent all other genus included in the model (n = 113 genus). This tree was adapted from Blanchard *et al.* 2016. The outgroup is the genus *Apis,* in black on this representation.

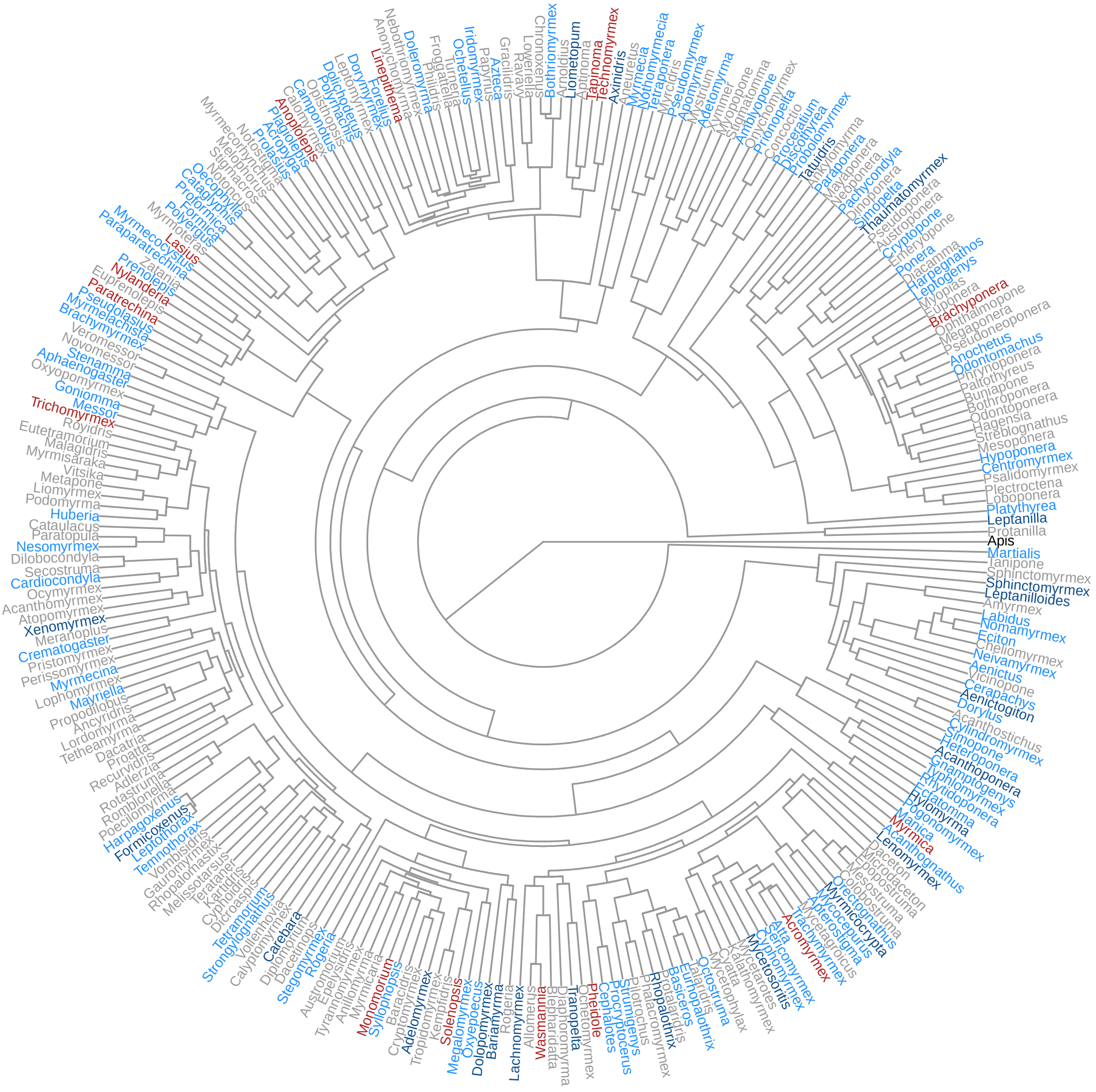
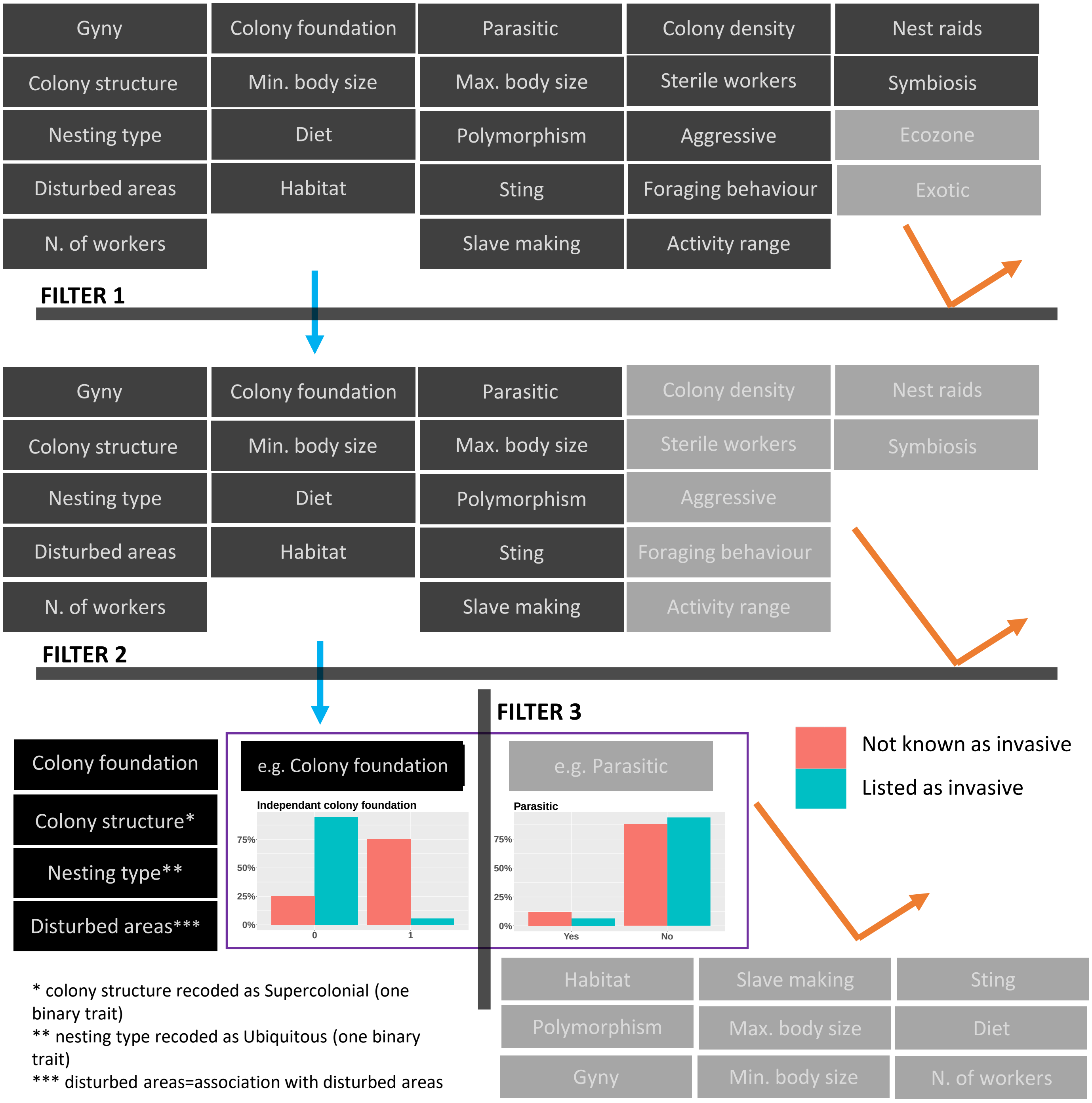


Figure S3: **Successive decisions made to select the variables used in our predictive model. FILTER 1:** we removed traits based on information that is not available prior to species introduction to avoid restricting our profiling tool to species that have already had the chance to be moved outside their native zone. **FILTER 2:** we removed traits that had too many missing values (> ~60%). **FILTER 3:** we computed pairwise correlations between traits and selected only non correlated traits and traits that were correlated with invasive status (based on p>0.05).

The light grey shading and the orange arrow show traits removed at each step. Dark shading (and blue arrow) show the selected traits at this step. The final traits retained are coloured in black.



* colony structure recoded as Supercolonial (one binary trait)
** nesting type recoded as Ubiquitous (one binary trait)
*** disturbed areas=association with disturbed areas

Fig S4: Measure of eigenvectors phylogenetic signal strength. This analysis was run before using the eigenvectors for missing data imputation. It is a distinct method

Various metrics exist for measuring phylogenetic signal in trait data. Here, we measured the amount of phylogenetic information in the categorical and binary traits using phylogenetic signal-representation curves for each traits (61–63). To obtain these curves, we computed phylogenetic eigenvector regressions, by successively adding eigenvectors to model trait variation, and plotted the corresponding R2 against the accumulated eigenvalues (Fig. S4).

The phylogenetic signal is measured following Guénard et al (2012) method. The principle of this method is to use an increasing number of eigenvectors as explanatory variables to model the variation in each trait and measure the R^2 of the model calibrated with each cumulated number of eigenvectors. A high R^2 (vertical axis) means that the eigenvectors are powerful in explaining the variation in that particular trait, and should therefore be used to impute the missing data in that trait.

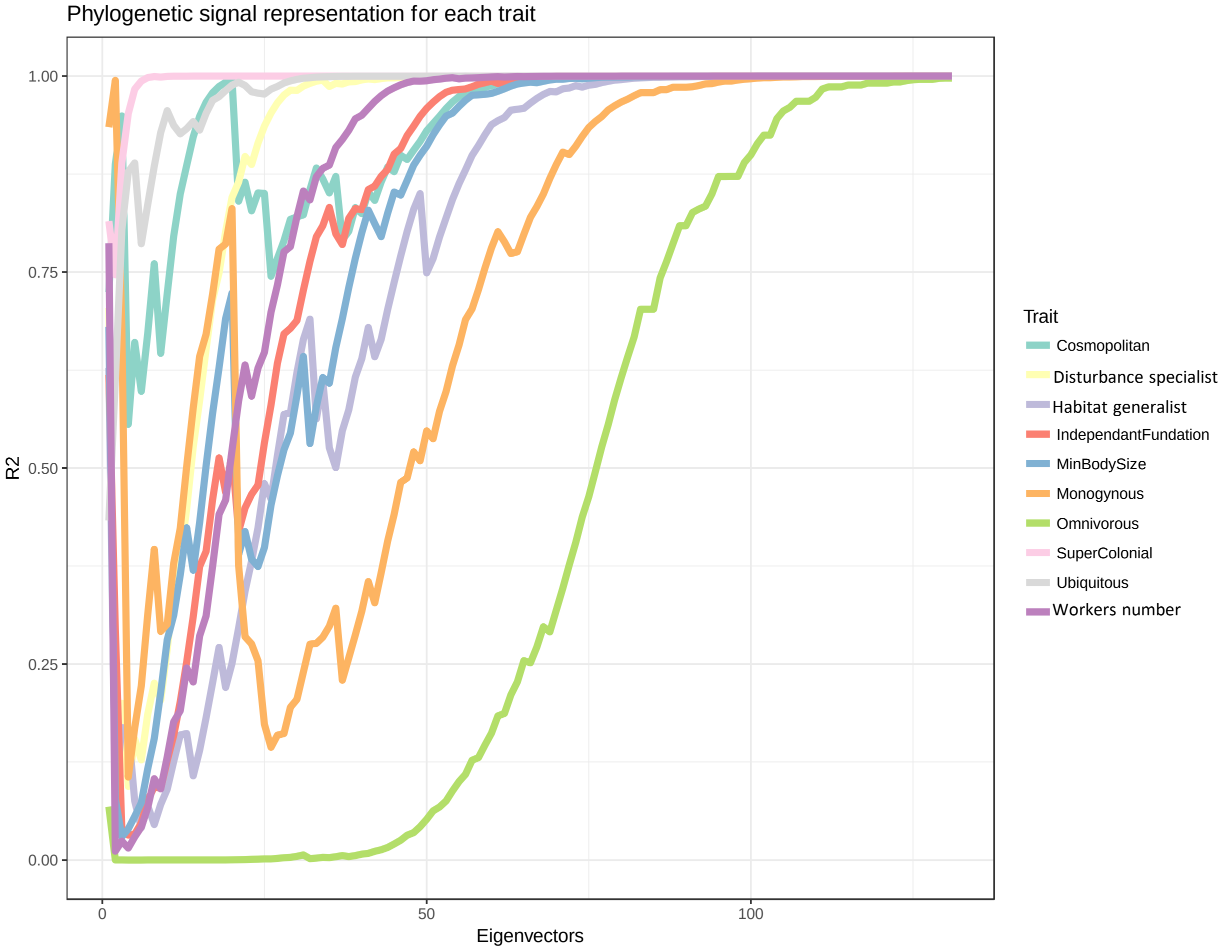


Fig S5: Imputation error (OOB error) according to number of eigenvectors (X axis) used for imputation for the 4 traits used in our models. The light blue dots show the number of eigenvectors retained for the imputation of each trait, i.e. the smallest number of eigenvector giving the minimum OOB error we could get. Each point on the graph is the OOB error averaged over 10 repeated imputations.

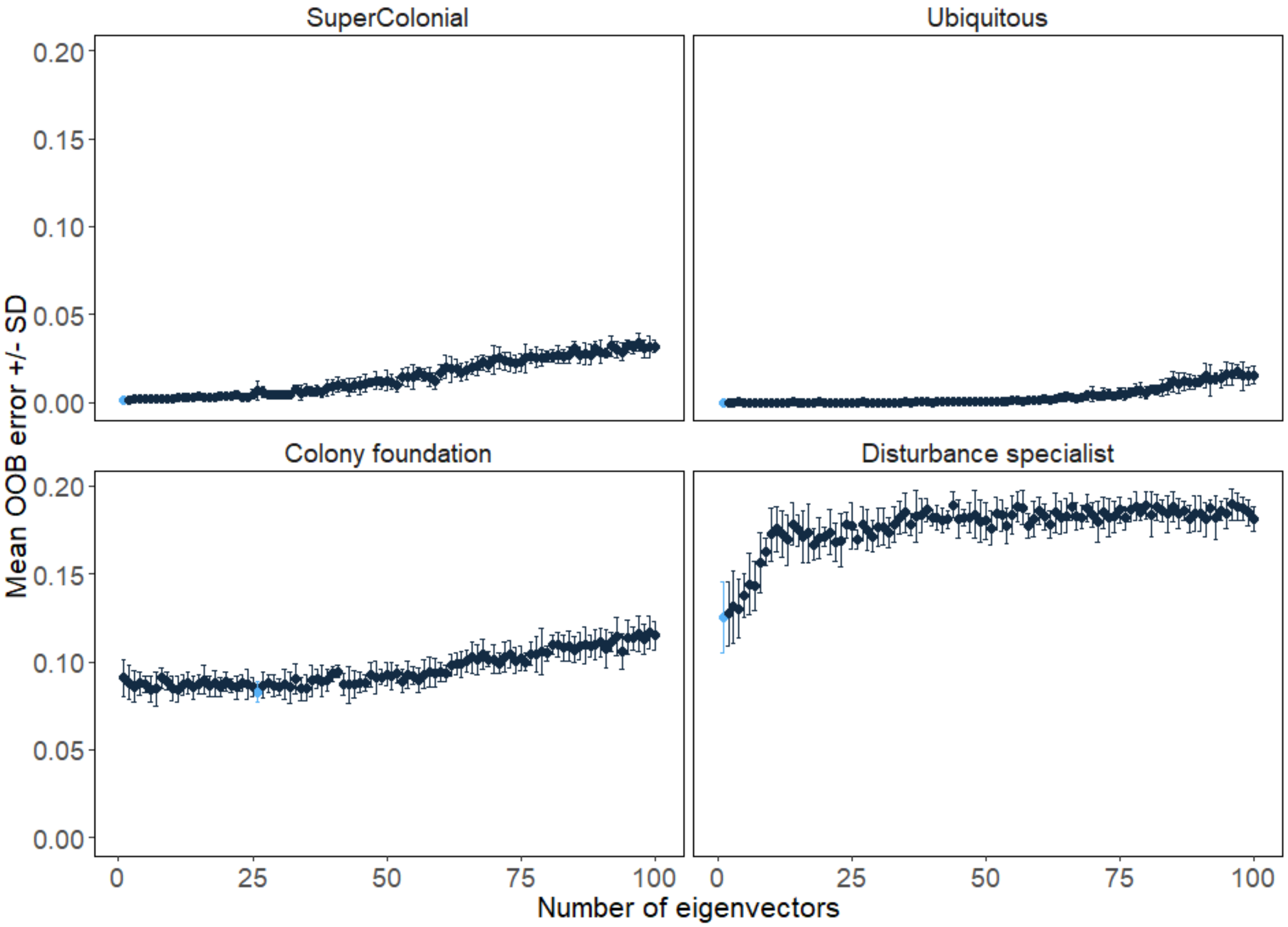


Fig. S6: Steps performed to calibrate MI.

A. Principle of missing value imputation using random forest (missForest)

Steps repeated for **each trait**;

1) Replacement of the missing data by the majority value in the column.
2) A random forest is fitted using the rest of the variables of the dataset (with n = 100 trees and 6 variables tested at each node). Replace NA with the most frequent imputed value in the forest, weighted by proximity.
3) Repeat this process iteratively until the stopping criterion is attained, i.e. when the difference between the nth imputed matrix and the n-1th imputed matrix increases for the first time. The details of the algorithm can be found in Stekhoven & Bühlmann, 2012 (*Missforest-Non-parametric missing value imputation for mixed-type data. Bioinformatics 28:112–118*).

B. Tuning of parameters (1) number of trees (***ntree***) to grow in each forest and **(2)** number of variables randomly selected at each node to set up the split (***mtry***). The final parameters chosen for the multiple imputation were 100 trees (ntree) and 6 variables (mtry). OOB error is the Out Of the Bag error.

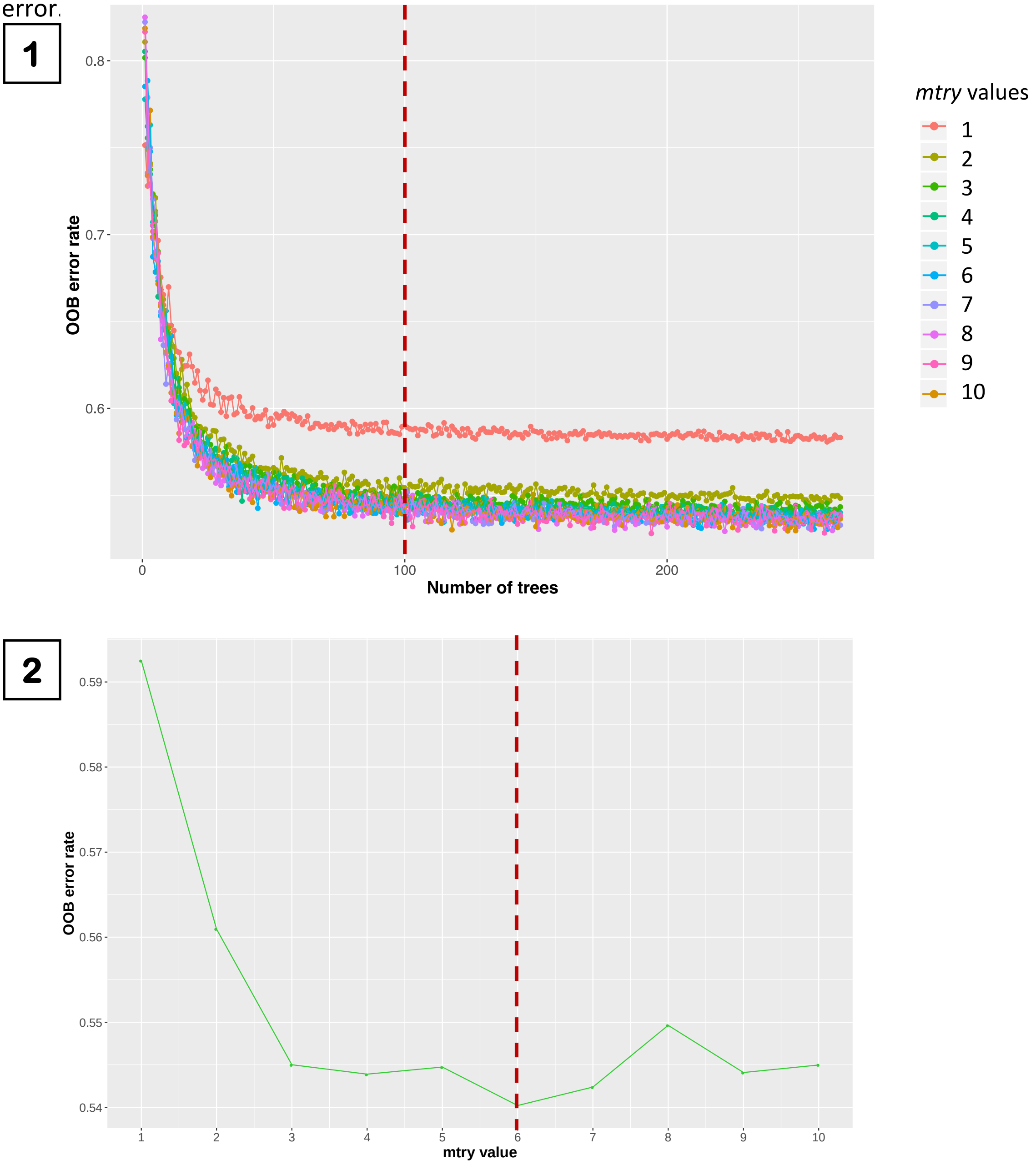


Fig. S7: Comparison of imputation error with categorical traits (left) and their corresponding binary traits (right). Imputation error is an order of magnitude greater when dealing with categorical traits, compared to binary ones. To keep imputation error as low as possible, we transformed the categorical trait “colony structure” and “nesting type” into binary traits. OOB: out-of-bag error.

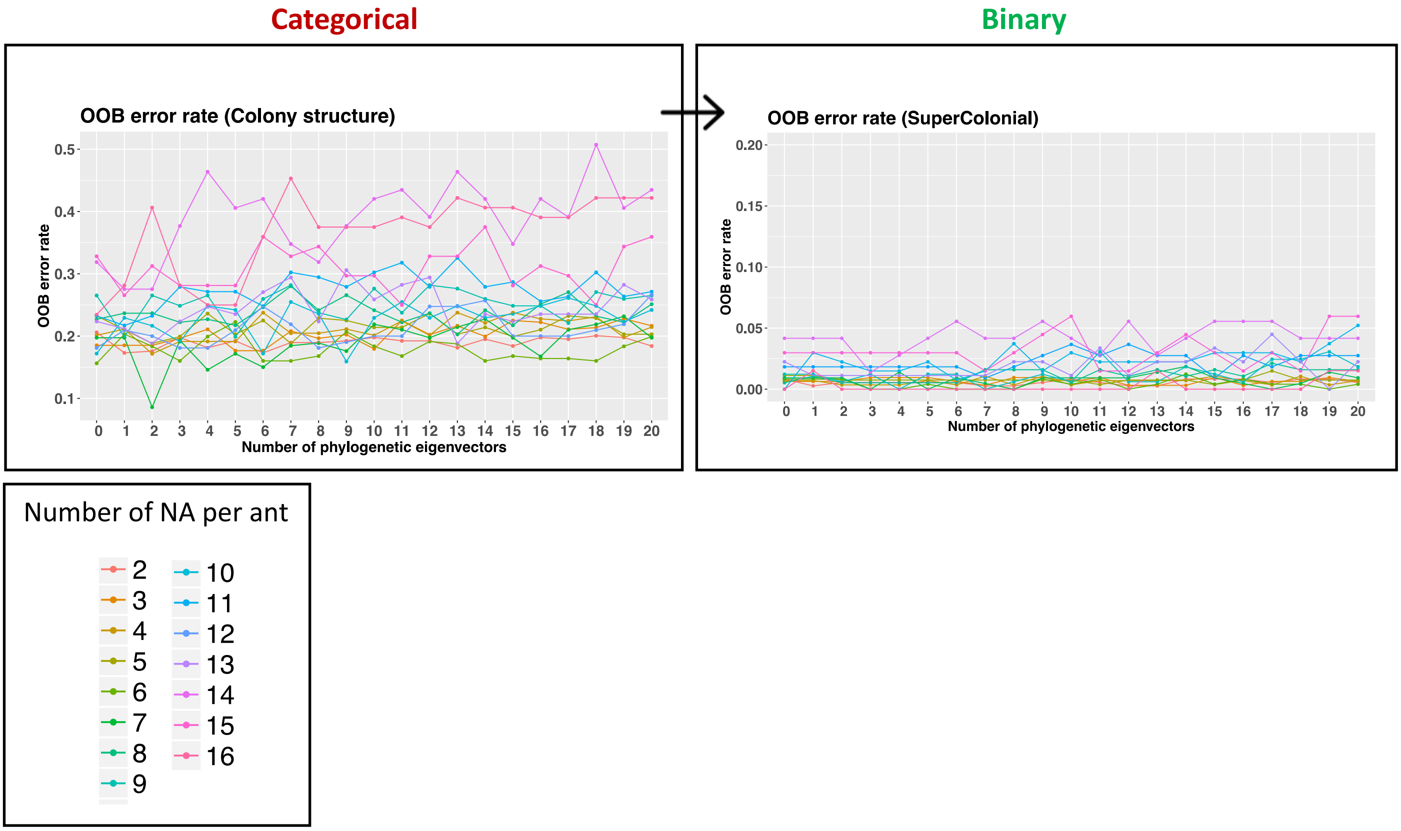


Fig. S8: (A) Visual analysis of missing values repartition among traits (columns) and species (lines) for the 992 ant species used for the model, and (B) evolution of the number of ants retained according to the number of NA tolerated per ant.

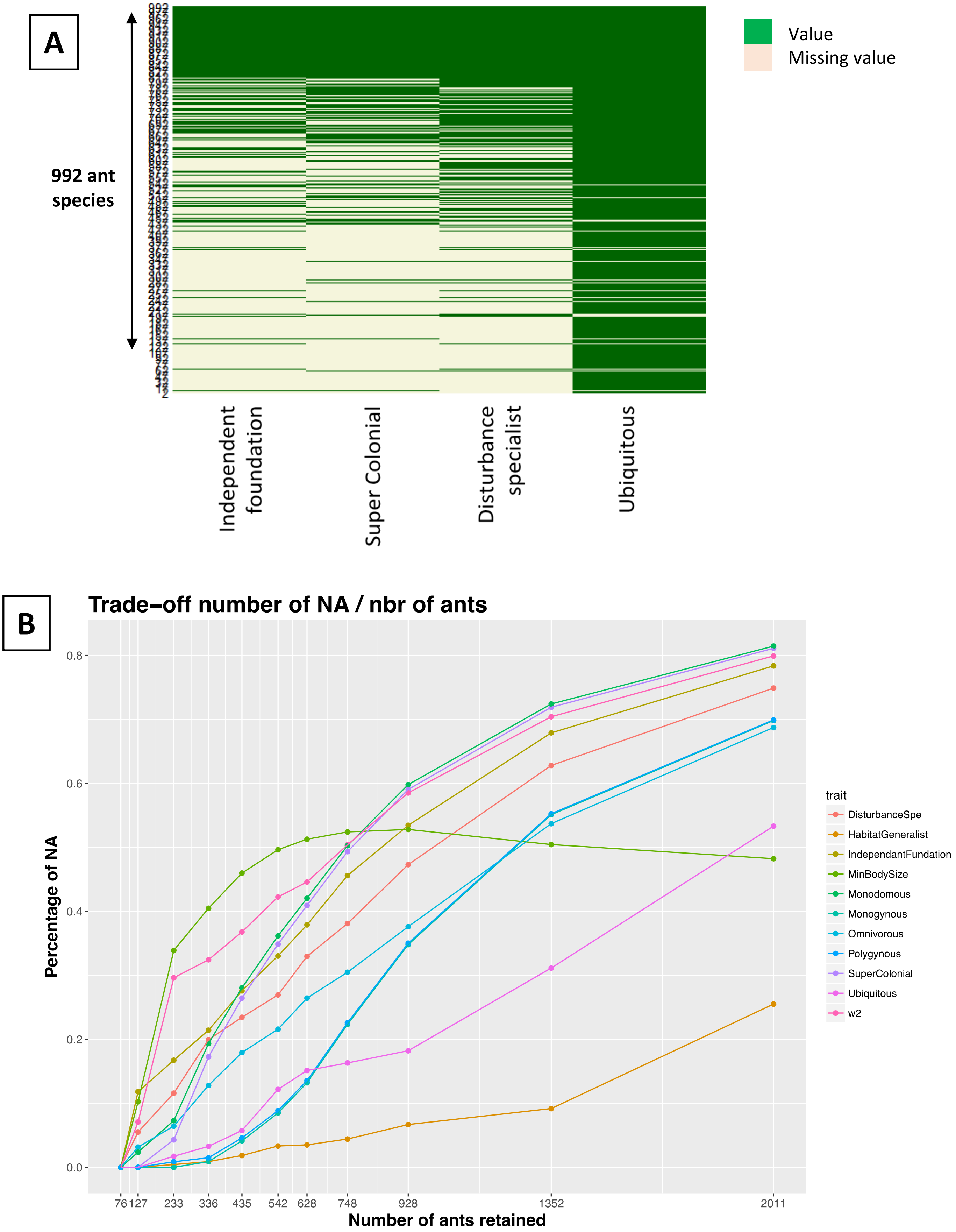
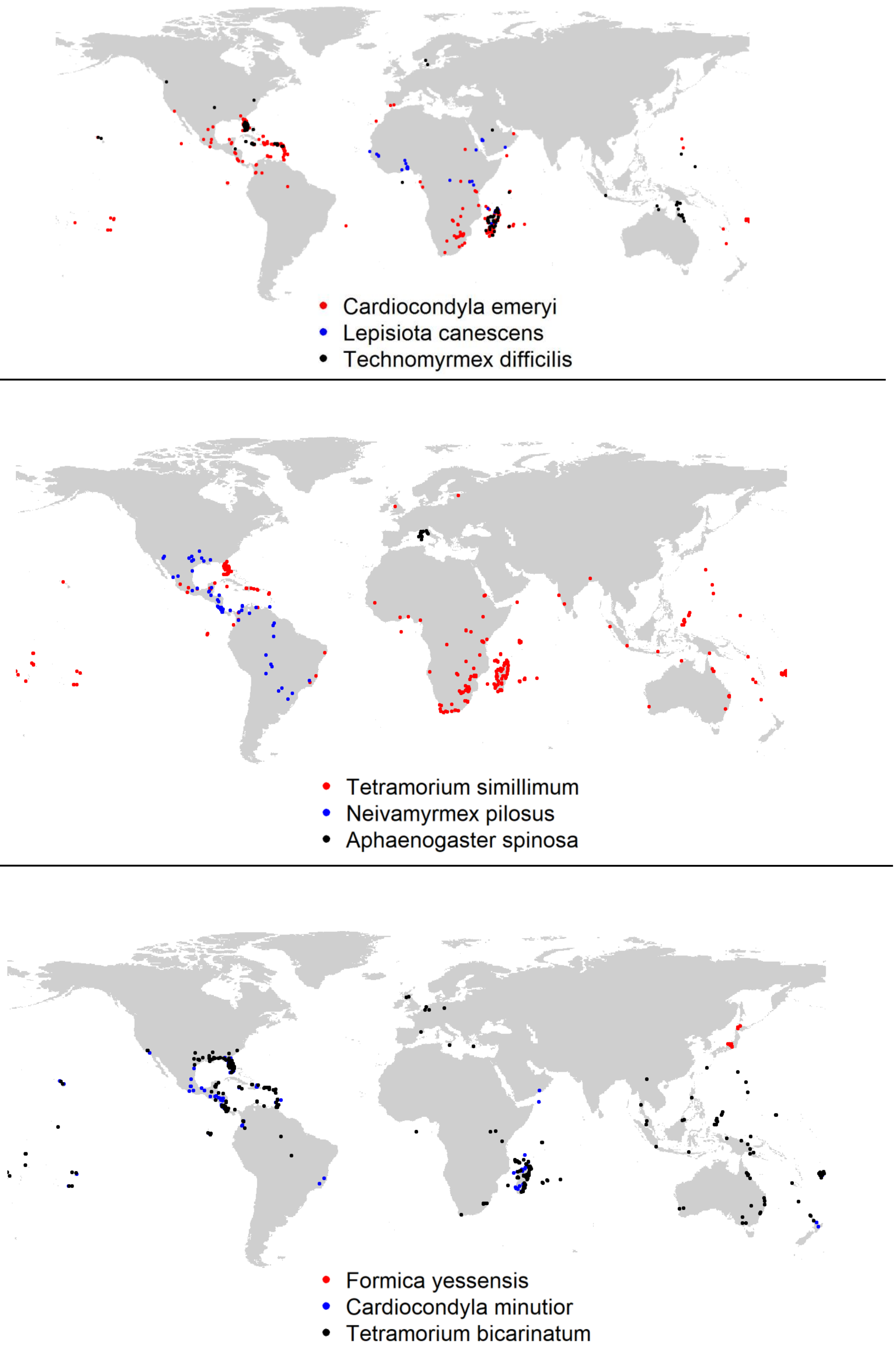
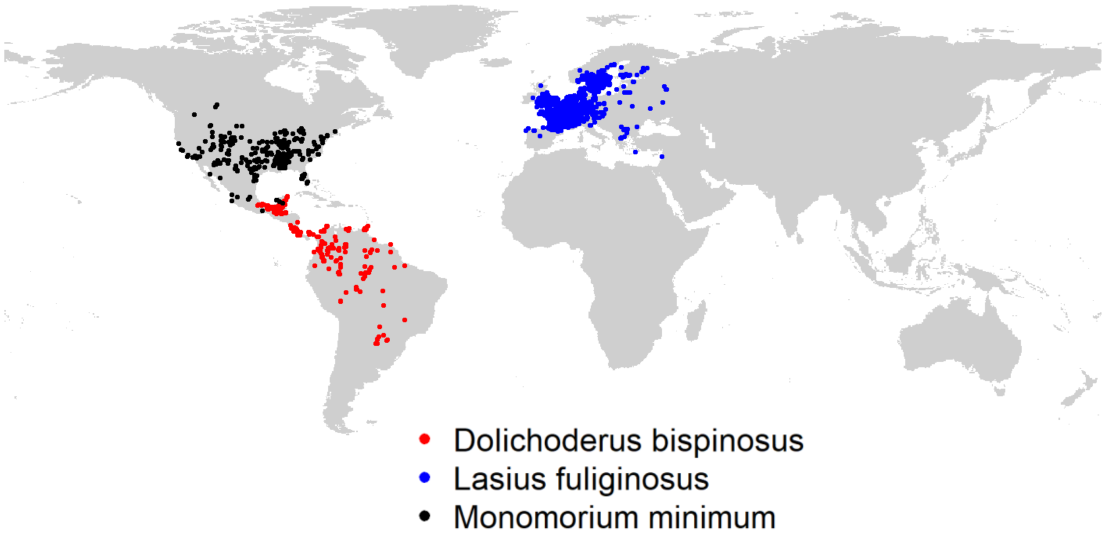
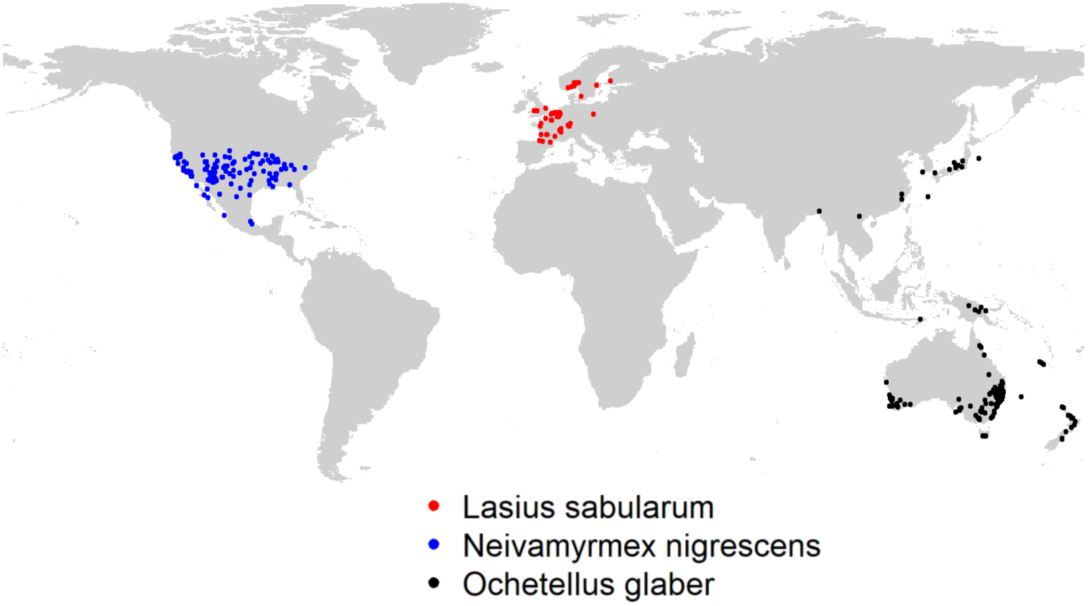


Fig. S9: Occurrences of 18 of the potential invasive and super-invasive ants used to build each species's SDM, gathered from GBIF & GABI databases, and from Wetterer's work.



Supplementary Figure S9 (cont.)





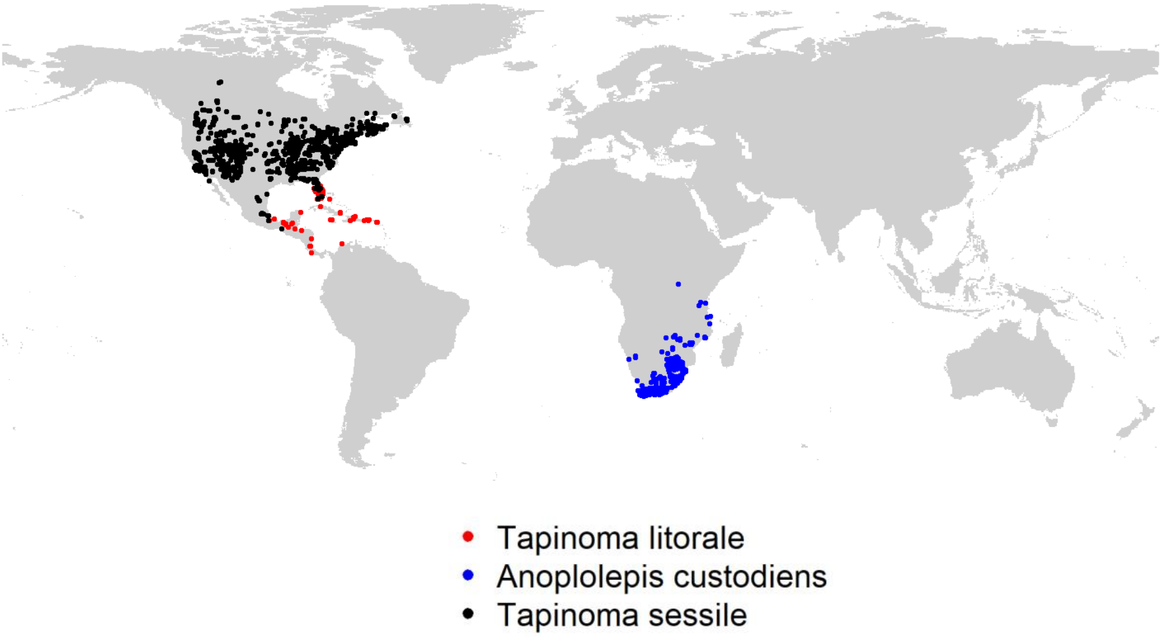
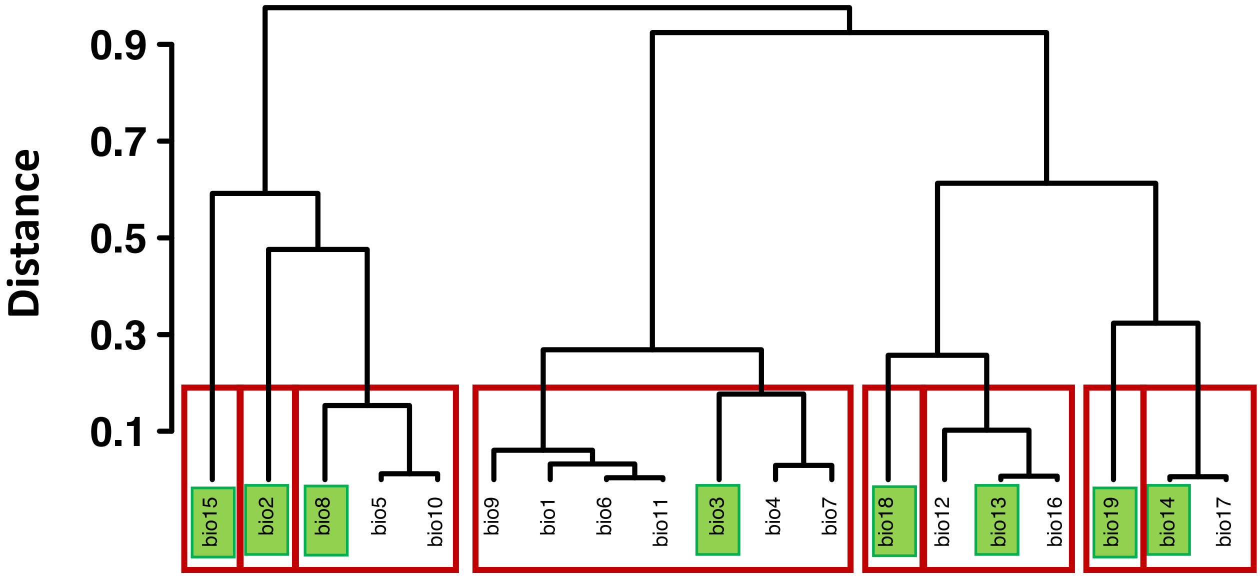
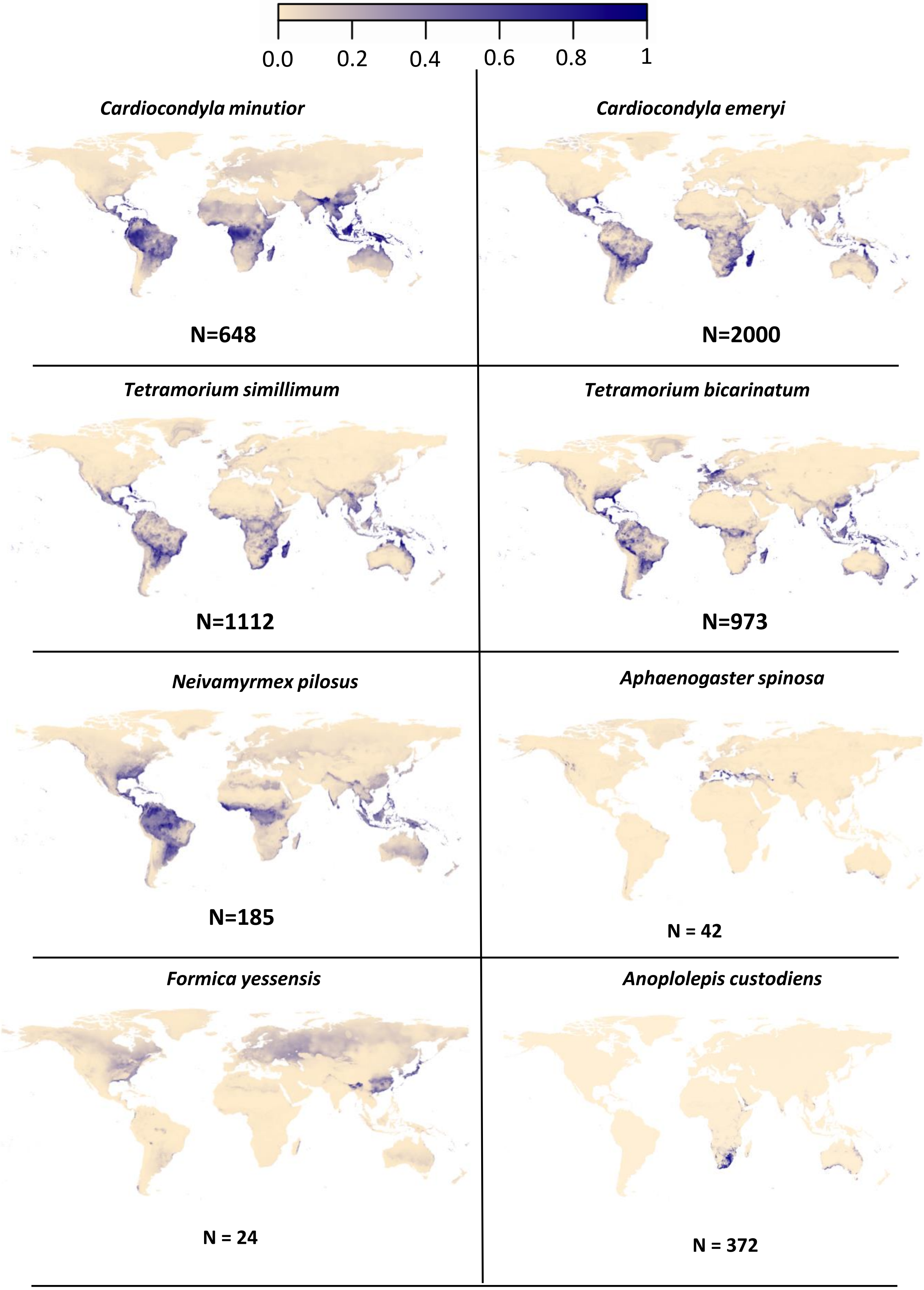


Fig. S10: Correlation groups between climatic variables and signification of climatic variables. Hierarchical ascendant classification based on distance = (1- Pearson's r), where red rectangles represent correlated variables at a threshold of 0.8 ((i.e. distance < 0.2) and green shaded variables are the ones we used for the SDMs.



bio1 = annual mean temperature
bio2 = mean diurnal range (mean of monthly (max temp - min temp))
bio3 = isothermality (bio2/bio7) (* 100)
bio4 = temperature seasonality (standard deviation *100)
bio5 = max temperature of warmest month
bio6 = min temperature of coldest month
bio7 = temperature annual range (bio5-bio6)
bio8 = mean temperature of wettest quarter
bio9 = mean temperature of driest quarter
bio10 = mean temperature of warmest quarter
bio11 = mean temperature of coldest quarter
bio12 = annual precipitation
bio13 = precipitation of wettest month
bio14 = precipitation of driest month
bio15 = precipitation seasonality (coefficient of variation)
bio16 = precipitation of wettest quarter
bio17 = precipitation of driest quarter
bio18 = precipitation of warmest quarter
bio19 = precipitation of coldest quarter

Fig. S11: Predicted climatic suitability distribution map for each of the 18 potential invasive ants, under current climatic conditions (the maps of the two super-invasive species are in Fig3). The number below each map corresponds to the number of occurrence points available to run the SDM. The suitability probability is increasing from pale to dark blue. Since the exact native range is often unknown, the projected climatic suitability includes the native range of species.



Cardiocondyla minutior

N=648

Cardiocondyla emeryi

N=2000

Tetramorium simillimum

N=1112

Tetramorium bicarinatum

N=973

Neivamyrmex pilosus

N=185

Aphaenogaster spinosa

N = 42

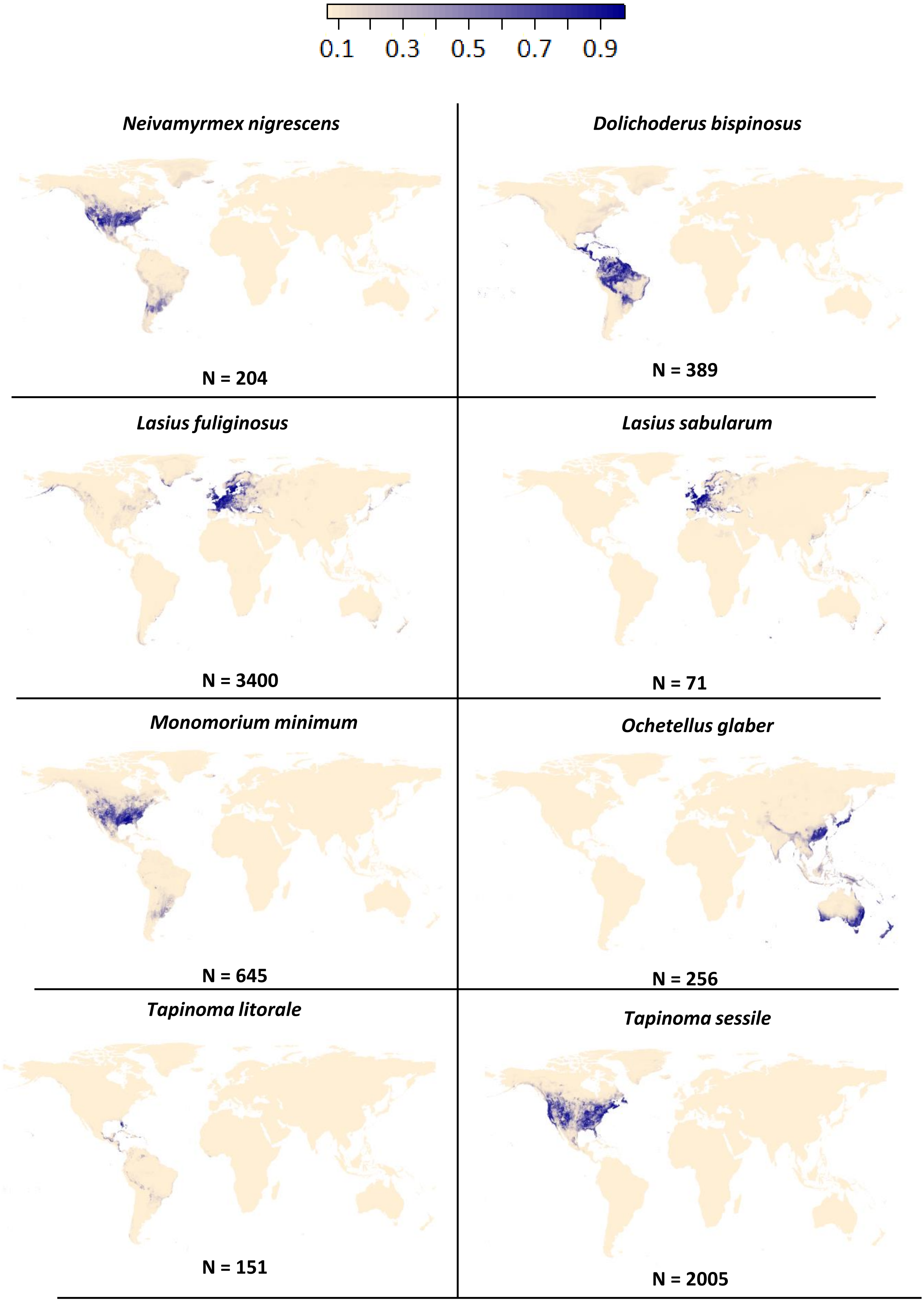
Formica yessensis

N = 24

Anoplolepis custodiens

N = 372

Supplementary Figure S11 (cont.)



Neivamyrmex nigrescens

N = 204

Dolichoderus bispinosus

N = 389

Lasius fuliginosus

N = 3400

Lasius sabularum

N = 71

Monomorium minimum

N = 645

Ochetellus glaber

N = 256

Tapinoma litorale

N = 151

Tapinoma sessile

N = 2005

Table S1: Traits used for the model, description of their coding rules, percentage of their missing values in the dataset with 992 species, number of eigenvectors and imputation error (out-of-bag error).

Trait name	**Colony Foundation**	**Colony structure**	**Association with disturbed areas**	**Nesting type**
Coding rules for binary traits	1: independent **colony foundation**, 0: otherwise	1: forms super-colonies (**supercolonialist**), 0: mono or polydomous	1: preferably found associated in disturbed areas (**disturbance specialist**), 2: otherwise	1: species found in > 1 nesting type (**ubiquitous**), 0: otherwise
Percentage NA in data	62.6	62.6	52.5	5.7
Number eigenvectors for imputations	26	1	1	1
Mean Imputation errror ± SD in 100 datasets (OOB)	0.010 ± 0.011	0.006 ± 0.005	0.153 ± 0.030	0.0005 ± 0.0009

Table S2: Phylogenetic signal calculated for the ten traits with lower percentage of missing values, based on Blanchard and Moreau (2016) genus level phylogenetic tree. This signal corresponds to the mean R^2 of the phylogenetic eigenvector regressions for categorical and binary traits, and to the Pagel's λ for the only continuous trait (min. body size). Traits used in the predictive models are in bold.

Trait	Phylogenetic signal
Habitat Generalist	0.766
Disturbance Specialist	**0.905**
Independent colony fundation	**0.828**
Polygyny	0.935
Super coloniality	**0.995**
Monodomous	0.816
Nesting generalist	**0.980**
Monogyny	0.725
Omnivorous (diet generalist)	0.420
Number of workers	0.848
Body Size	0.412

Table S3: Results of the linear model to predict invasiveness probability. Chisquare Wald values are averaged over 100 models fitted with 992 ant species. Model averaging provided very similar results as the ones presented here.

Trait	Df	Chisq mean (100 models)	Chisq sd (100 models)	Mean P-value
Super colonial (**colony structure**)	1	18.86	0.81	< 0.001
Disturbance Specialist (**association with disturbed areas**)	1	5.94	0.37	0.01
Ubiquitous (**nesting type**)	1	4.98	0.59	0.03
Independent foundation (**colony foundation**)	1	5.31	0.35	0.07

Table S4: Predicted invasive species using A) 100 models and 992 species (results presented in the main text); and for B) 1900 models (i.e. 100 models*19 invasive) and 991 species, where one of the 19 known invasives were excluded from the analyses. The percentage of models where the species was selected as invasive and the average invasiveness probability are given.

	A. Models with all 992 species		**B. Models with 991 species (leave-one-out invasive)**	
species	**Percentage models**	**Invasiveness probability (mean 100 models)**	**Percentage models**	**Invasiveness probability (mean 1900 models)**
Technomyrmex difficilis	100	0.87	100	0.86
Lepisiota canescens	100	0.83	100	0.82
Anoplolepis custodiens	98	0.38	98	0.37
Formica yessensis	100	0.23	95	0.23
Tapinoma litorale	100	0.17	100	0.16
Tapinoma sessile	100	0.17	100	0.16
Ochetellus glaber	100	0.17	100	0.16
Lasius fuliginosus	100	0.14	100	0.13
Aphaenogaster spinosa	100	0.13	100	0.12
Cardiocondyla emeryi	100	0.13	100	0.12
Cardiocondyla minutior	100	0.13	100	0.12
Dolichoderus_bispinosus	100	0.13	100	0.12
Lasius_sabularum	100	0.13	100	0.12
Monomorium_minimum	100	0.13	100	0.12
Neivamyrmex_nigrescens	100	0.13	100	0.12
Neivamyrmex pilosus	100	0.16	100	0.12
Tetramorium bicarinatum	100	0.23	100	0.12
Tetramorium simillimum	100	0.16	100	0.12

Table S5: Predicted invasiveness probabilities for already known invasive species when they were coded as 0 and then predicted by the model (leave-one-out invasives). These results show that our models were accurate in predicting invasiveness. Note that *Acromyrmex octospinosus* was not identified as invasive by our models due to its very low invasiveness probability (see manuscript)

Invasive species	**Percentage models**	**Invasiveness probability (mean 1900 models)**
Anoplolepis_gracilipes	100	0.83
Brachyponera_chinensis	100	0.17
Lasius_neglectus	100	0.87
Linepithema_humile	100	0.83
Monomorium_floricola	100	0.13
Monomorium_pharaonis	100	0.83
Myrmica_rubra	100	0.83
Nylanderia_pubens	100	0.83
Paratrechina_longicornis	100	0.83
Pheidole_megacephala	100	0.39
Solenopsis_geminata	100	0.87
Solenopsis_invicta	100	0.87
Solenopsis_papuana	100	0.13
Solenopsis_richteri	100	0.17
Tapinoma_melanocephalum	100	0.83
Technomyrmex_albipes	100	0.87
Trichomyrmex_destructor	100	0.87
Wasmannia_auropunctata	100	0.83