Common to rare transfer learning (CORAL) enables inference and prediction for a quarter million rare Malagasy arthropods

Otso Ovaskainen[1,2*], Steven Winter[3*], Gleb Tikhonov[2], Nerea Abrego[1], Sten Anslan[1], Jeremy R. deWaard[4], Stephanie L. deWaard[4], Brian L. Fisher[5,6], Brendan Furneaux[1], Bess Hardwick[2], Deirdre Kerdraon[7], Mikko Pentinsaari[8], Dimby Raharinjanahary[6], Eric Tsiriniaina Rajoelison[6], Sujeevan Ratnasingham[4], Panu Somervuo[2], Jayme E. Sones[4], Evgeny V. Zakharov[4], Paul D. N. Hebert[4,9], Tomas Roslin[7,2] and David Dunson[3].

* shared first authorship

[1]Department of Biological and Environmental Science, University of Jyväskylä, P.O. Box 35 (Survontie 9C), FI-40014 Jyväskylä, Finland

[2]Organismal and Evolutionary Biology Research Programme, Faculty of Biological and Environmental Sciences, University of Helsinki, P.O. Box 65, Helsinki 00014, Finland

[3]Department of Statistical Science, Duke University, Durham, NC, USA

[4]Centre for Biodiversity Genomics, University of Guelph, Guelph, ON, CANADA N1G 2W1

[5]Entomology, 55 Music Concourse Drive, California Academy of Sciences, San Francisco, CA 94118, USA

[6]Madagascar Biodiversity Center, Parc Botanique et Zoologique de Tsimbazaza, Antananarivo 101, Madagascar

[7]Department of Ecology, Swedish University of Agricultural Sciences (SLU), Uppsala, Sweden

[8]Canadian National Collection of Insects, Arachnids and Nematodes, Agriculture and Agri-Food Canada, 960 Carling Avenue, Ottawa, Ontario, K1A 0C6, Canada

[9]Canadian High Arctic Research Station, Polar Knowledge Canada, PO Box 2150, 1 Uvajuq Road, Cambridge Bay, Nunavut X0B 0C0, Canada

Abstract

Modern DNA-based biodiversity surveys result in massive-scale data, including up to millions of species – of which most are rare. Making the most of such data for inference and prediction requires modelling approaches that can relate species occurrences to environmental and spatial predictors, while incorporating information about taxonomic or phylogenetic placement of the species. Even if the scalability of joint species distribution models to large communities has greatly advanced, incorporating hundreds of thousands of species has been infeasible to date, leading to compromised analyses. Here we present a novel "common to rare transfer learning" approach (CORAL), based on borrowing information from the common species and thereby enabling statistically and computationally efficient modelling of both the common and the rare species. We illustrate that CORAL leads to much improved prediction and inference in the context of DNA metabarcoding data from Madagascar, comprising 255,188 arthropod species detected in 2874 samples.

MAIN TEXT

Earth is home to some ten million species[1]. Among these, the typical species is unknown[2] and rare[3]. Recent innovations in sensor technologies are now providing unprecedented means for recording patterns and changes in the abundance and distribution of all kinds of taxa, from the named to the previously unnamed and from the rare to the common. These technologies include DNA-based monitoring, passive acoustic monitoring, and visual sensors[4,5]. By allowing the efficient recording of thousands to hundreds of thousands of species in time and space, the accumulation of high-dimensional "novel community data" is transforming our access to information on species distributions and abundances[4]. As a particularly exciting development, the emergence of novel community data allows us to target the speciose groups accounting for the main part of global biodiversity[1,2]. Where species records to date have been massively biased towards vertebrates, one of the least species-rich taxa[3], recent methods are now making hyperdiverse taxa such as arthropods and fungi arguably *easier* to sample than vertebrates and plants. As these speciose taxa can be mass-sampled and mass-identified, we can derive automated characterizations of what taxa occur where[5–7]. Nonetheless, the recent revolution in the generation of data is awaiting a matching insurgence of novel methods to analyze the data.

While most species on Earth are rare, these are the species that we know least about, partially because rare species are the most challenging to model[8]. Paradoxically, the species in the highest need of protection and thus for which information on their distributions and ecological requirements is most needed, are typically rare ('the rare species paradox'[9]). Understanding biodiversity change necessitates models which can relate species occurrences to environmental, biotic and spatial predictors, and which can predict changes in species communities with changes in the state of these drivers[10]. Hence, the need for predictive tools for rare species has been repeatedly highlighted[11–14]. However, typical approaches to species-level modelling impose a cut-off on species occurrences or abundances[15,16] – arguing that for the rarest species, the data are simply insufficient for any quantitative inference regarding the drivers of their distribution. The inherent rarity of most species results in highly skewed species-abundance distributions, where few species are common whereas most species are present at few sites in low numbers. In a world where rare is common[3] this can and will typically amount to rejecting the majority of data, and all the information there then remains hidden. To make the most of increasingly available data, we need modelling approaches which can draw on the full depth of them.

With species distribution models (SDMs), rare species may be modelled through ensemble predictions from multiple small models, each of which contains just a few predictors to avoid overfitting[9,17,18]. Because closely related species are generally ecologically more similar than distantly related species[19,20], phylogenetic information may be used to make inference on the distributions of rare species[21–25]. Joint species distribution models (JSDM)[26] allow levelling up by modelling large numbers of species simultaneously. This enables efficient borrowing of information across species through their shared responses to environmental variation[27]. Furthermore, when data on species phylogenies and/or traits are available, information can be borrowed especially across similar species[28,29]. This can lead to improved predictions especially for rare species[30].

The high-dimensional, and often extremely sparse nature of the species occurrence data, compounded with spatiotemporal and phylogenetic dependencies, presents major challenges for statistical analyses and computation. High performance computing can scale some existing JSDMs to thousands of species[31,32]. Two-stage methods, which make small concessions by cutting dependence between species via approximate likelihoods[33,34], can scale to tens of thousands of species while still retaining reasonable uncertainty estimates. Unfortunately, these approaches do not yet scale to millions of

species as needed to comprehensively model Earth's biodiversity[1]. What is more, they may perform poorly for extremely sparse rare species, by compromising model structures in the interest of gaining computational advantage.

Results

The Hierarchical Modelling of Species Communities framework

In this paper, we apply Bayesian transfer learning[35] to develop the "common to rare transfer learning approach" (CORAL) (Fig. 1). CORAL builds on the Hierarchical Modelling of Species Communities (HMSC)[10,29,36] approach to joint species distribution modelling (JSDM)[26]. HMSC is a multivariate generalized linear model fitted in a Bayesian framework. As a response it considers a matrix of species occurrences or abundances. We exemplify our approach with presence-absence data, denoting by $y_{ij} = 1$ if species j (with $j = 1, \dots, n_s$) is present in sample i (with $i = 1, \dots, n_y$) and $y_{ij} = 0$ if this is not the case. Presence-absence data are modelled in HMSC with probit regression: $\Pr(y_{ij} = 1) = \Phi(L_{ij})$, where $\Phi(.)$ is the standard normal cumulative distribution function and L_{ij} is the linear predictor modelled as:

$$L_{ij} = \sum_k^{n_c} x_{ik}\beta_{kj} + \sum_k^{n_f} \eta_{ik}\lambda_{kj}, \qquad (1)$$

where x_{ik} are measured predictors, η_{ik} are latent predictors, and β_{kj} and λ_{kj} are regression coefficients quantifying responses of the species to the measured and the latent predictors. The latent features induce within-sample dependence across species; these features may encode characteristics of the habitat, the environment and the spatio-temporal setting not captured by the x_{ik}s. HMSC uses a Bayesian hierarchical model to (a) infer how many latent features n_f are needed and (b) to borrow information across species in inferring the β_{kj}s. For (b) HMSC estimates to what degree taxonomically or phylogenetically related species, or species with similar traits, show similar responses β_{kj} to environmental variation through the multivariate normal distribution[10,28]

$$\mathrm{vec}(B) \sim N(\mu, P \otimes V). \qquad (2)$$

Here, B is the matrix of the regression parameters β_{kj} of the p species, μ is the average response, $P = \rho C + (1-\rho) I_{n_s}$ is a weighted average between the phylogenetic or taxonomic correlation matrix C and the identity matrix I corresponding to unrelated species, $0 \le \rho \le 1$ is the strength of the phylogenetic signal, and V is the variance-covariance matrix of species-specific deviations from the average μ. The average response μ is further modelled as $\mu = \mathrm{vec}(\Gamma T^T)$, where T is a matrix of species traits and Γ are the estimated responses of the traits to environmental variation. With Eq. 2, HMSC learns if and to what extent related species, or species with similar traits, show similar environmental responses. This allows for effective borrowing of information among species. In particular, it allows improved parameter estimation for rare species, for which it would be difficult to obtain accurate estimates if considering the data in isolation from the community context[29]. As a result, HMSC shows higher predictive performance compared to approaches that do not enable such borrowing of information[30].

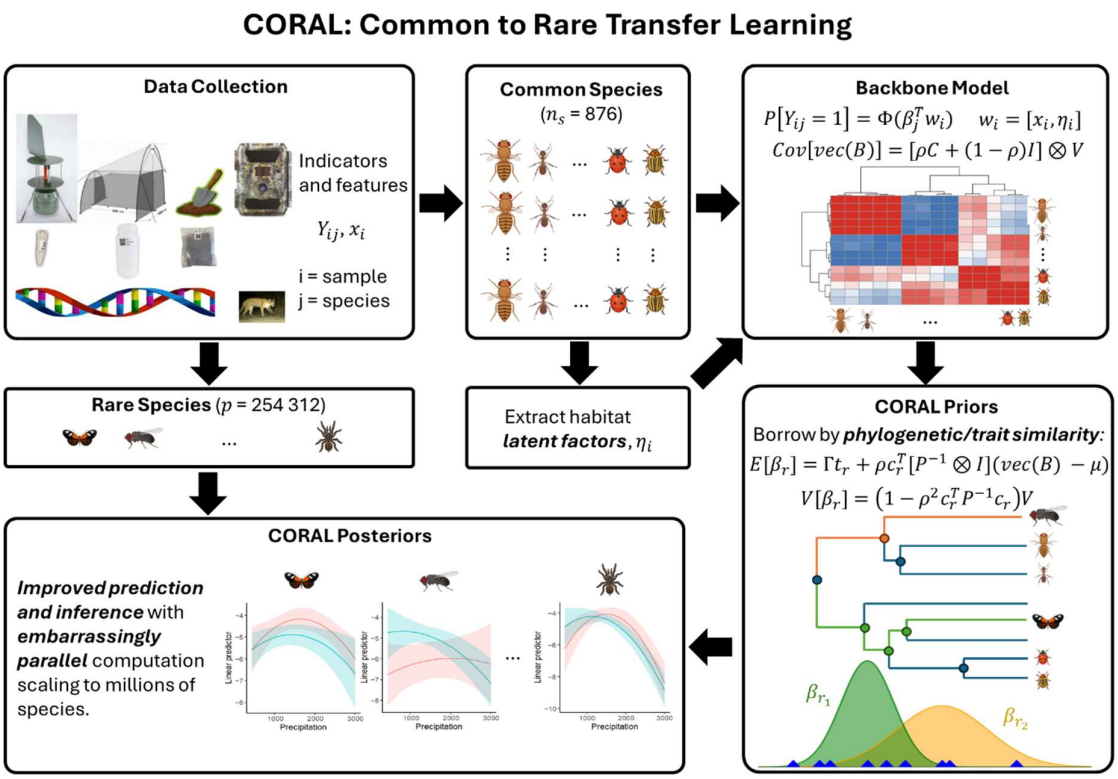

Figure 1. CORAL is based on fitting a backbone JSDM model to a subset of the most common species in the data and then borrowing information from this backbone to model the rare species. The backbone model learns about latent factors representing relevant missing environmental predictors, as well as about the species responses to both the measured and the latent predictors. The backbone model provides an informative prior distribution for each rare species. This is particularly efficient when we have access to phylogenetic or trait information, allowing us to borrow information especially from common species closely related to the rare species, or species sharing similar traits with the rare species. As we show with a case study targeting a quarter million rare Malagasy arthropods, such informative prior distributions greatly improves the modelling of the rare species, both in terms of inference and prediction. To enable easy application of the CORAL approach to high-dimensional biodiversity data, we provide an R pipeline for implementing all the steps outlined in this paper.

Deriving conditional priors for rare species by borrowing information from common species

Our key idea is that even if it is not feasible to include 100,000+ species in a JSDM model such as HMSC, one can still borrow information from common species. We propose to first fit an HMSC ***backbone model to the common species*** (Fig. 1). From this analysis we obtain a point estimate of the latent features $\eta_i = (\eta_{i1}, \dots, \eta_{in_f})$, which provide key information not captured in $x_i = (x_{i1}, \dots, x_{in_c})$ about the environmental/habitat conditions and overall biological community represented in sample i. We define a new covariate vector $\tilde{x}_i = \left(\tilde{x}_{i1}, \dots, \tilde{x}_{in_r}\right)^T$ by concatenating x_i and η_i to be used as a fixed predictor in separate probit models for each rare species: $\Pr(y_{ir} = 1) = \Phi(\tilde{x}_i^T \beta_r)$, for $r \in \mathcal{J}_r$ with $\mathcal{J}_r \subset \{1, \dots, p\}$ the set of rare species. To reduce mean square error in inferring β_r for $j \in \mathcal{J}_r$, we propose to adaptively shrink towards the common species coefficients while accounting for taxonomic/phylogenetic similarity. This is accomplished by first re-fitting the HMSC backbone model using the new covariate vector $\tilde{x}_i$ to learn relevant hyper-parameters and then using the following prior in fitting the rare species probit models:

$$\beta_r \sim N(m_r, S_r). \quad (3)$$

Here, the mean is given by

$$m_r = \Gamma t_r + \left[(\rho c_r^T)\left(P^{-1} \otimes I_{n_c}\right)\right](\mathrm{vec}(B) - \mu) \quad (4)$$

where $P = \rho C + (1-\rho) I_p$. The variance-covariance matrix by $S = k_r V$, where the variance scaling factor k_r is given by

$$k_r = 1 - \rho c_r^T P^{-1} c_r. \quad (5)$$

The vector c_r encodes relatedness between a rare species $r \in \mathcal{J}_r$ and all the n_s species in the backbone analysis and t_r is the trait vector for this rare species. The prior (3)-(5) is derived using conditional distributions from (2), as detailed in the Methods. As HMSC is fitted to data with Bayesian inference, parameter uncertainty can be accounted for by defining the prior as a mixture of multivariate normal distributions (each defined by Eqs. 3-5) over the posterior samples. To achieve a simple functional form for the prior, we approximated the mixture by a single multivariate normal distribution, the mean and variance-covariance matrix of which we set equal to those of the mixture (see Methods).

We refer to the above approach as **Co**mmon-to-**ra**re transfer **l**earning (**CORAL**). To evaluate the benefits of CORAL, we compare its performance to that of a **baseline model** that does not benefit from the backbone analysis. In other words, for the baseline model we fit $\Pr(y_{ir} = 1) = \Phi(x_i^T \beta_r)$, separately for $r \in \mathcal{J}_r$ using a simple Gaussian prior $\beta_r \sim N(m_r, S_r)$. We expect CORAL to have substantial advantages over the baseline model due to two considerations: $\tilde{x}_i$ contains important latent factor information on top of x_i, and CORAL allows the borrowing of information from the β_js for common species to rare species. As both CORAL and the baseline models can be fitted independently for $r \in \mathcal{J}_r$, computational time scales linearly with the number of species; these computations can be trivially parallelized allowing for inference and prediction for hundreds of thousands or even millions of species.

Case study on Malagasy arthropods

We tested the approach in the context of metabarcoding data on Malagasy arthropods. We applied Malaise trap sampling in 53 locations across Madagascar, each of which was relatively undisturbed and where the vegetation represented the conditions of the local environment. We then applied high-throughput COI metabarcoding[37] and the OptimOTU pipeline[38] to score the occurrences of 255,188 species-level OTUs (henceforth, species) in 2874 samples (see *Methods*). To create a backbone model of common species, we included those 876 species that occurred in at least 50 samples. This left those 254,312 species that occurred less than 50 times in the data as rare species, which we model by the CORAL approach. We note that the threshold of 50 occurrences is relatively high and that hence, some of the rare species are actually not so rare. This choice was made to test the hypothesis that borrowing information from the backbone model changes predictions and inference especially for the very rare species – but less so for more common species.

As predictors of species presence, we included covariates related to seasonality, climatic conditions, and sequencing effort. We modelled seasonality through periodic functions $\sin(2\pi d/365)$, $\cos(2\pi d/365)$, $\sin(4\pi d/365)$, and $\cos(4\pi d/365)$, where d is the day of sampling. Climatic conditions were modelled through the second order polynomials of mean annual temperature and mean annual precipitation[39], while including the interaction between these two climatic predictors. Variation in sequencing effort was modelled by including log-transformed sequencing depth as a predictor. To capture site- and sample-level variation not captured by the measured predictors, we

included the site ($n = 53$) and the sample ($n = 2874$) as random effects. As a proxy of phylogeny, we used taxonomic assignments at the levels of kingdom, phylum, class, order, family, subfamily, tribe, genus and species, including assignments to pseudotaxa for those cases that could not be reliably classified to previously known taxa (see *Methods*).

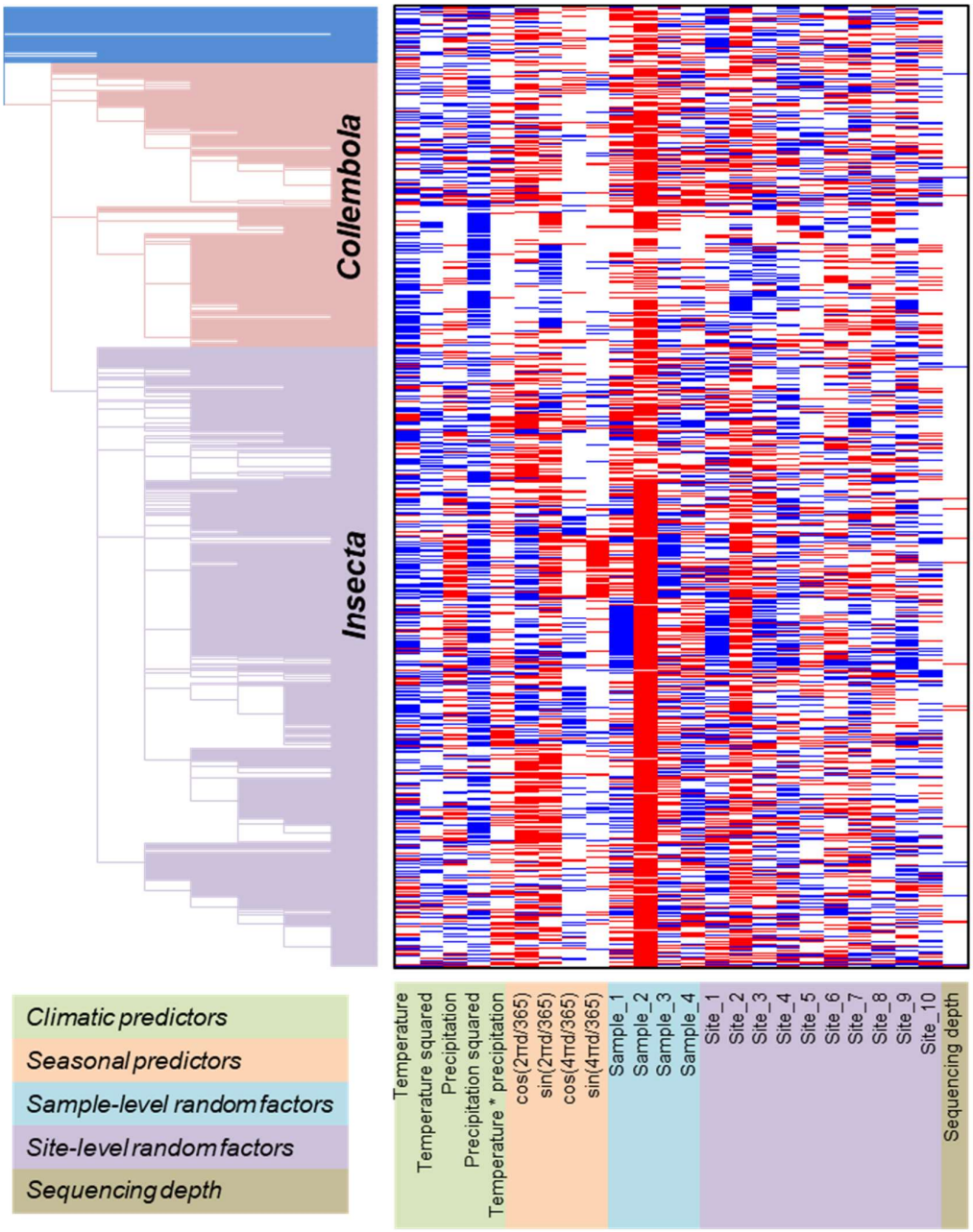

Figure 2. Responses of the common species to measured and latent predictors. Responses that were estimated to be positive (red) or negative (blue) with at least 95% posterior probability in the backbone model are highlighted. OTUs not assigned to classes *Insecta* or *Collembola* were placed in *Arachnida* (n=14), *Malacostraca* (n=1), *Mammalia* (n=2), or could not reliably be placed in any specific class at all (n=165).

The common species responded especially to site-level variation. This was shown both by responses to climatic variables, which contributed 48% of the explained variation, and by responses to the site-level random factors, which contributed 42% of the explained variation. The effects of the remaining predictors were much less pronounced, with seasonality contributing 3% of the explained variation, sample-level latent factors 7%, and sequencing depth 0.1%. The responses of the species to the predictors (Fig. 2) were strongly phylogenetically structured (posterior mean $\rho = 0.65$, posterior probability $\Pr(\rho > 0\ ) = 1.00$), thus providing potential for borrowing information especially from related species.

As we did not include any traits in the model, we based the CORAL models on borrowing of information on taxonomic relatedness only. The variance scaling factor k varied between 0.13 and 0.70, with a mean value of 0.34, thus showing a substantial reduction in variance. As expected, it was lowest for species with close relatives in the backbone model (Fig. 3A). The conditional prior models predicted variation in the occurrences of the rare species better than random (Fig. 3B). This result is non-trivial, as the predictions are made by a completely independent model that has not seen any data for the focal species. The accuracy of the prior predictions increased with the level of relatedness between the focal species and the species in the backbone model and the predictions were more accurate for species occurring at least 10 times in the data than for the very rare species (Fig. 3B).

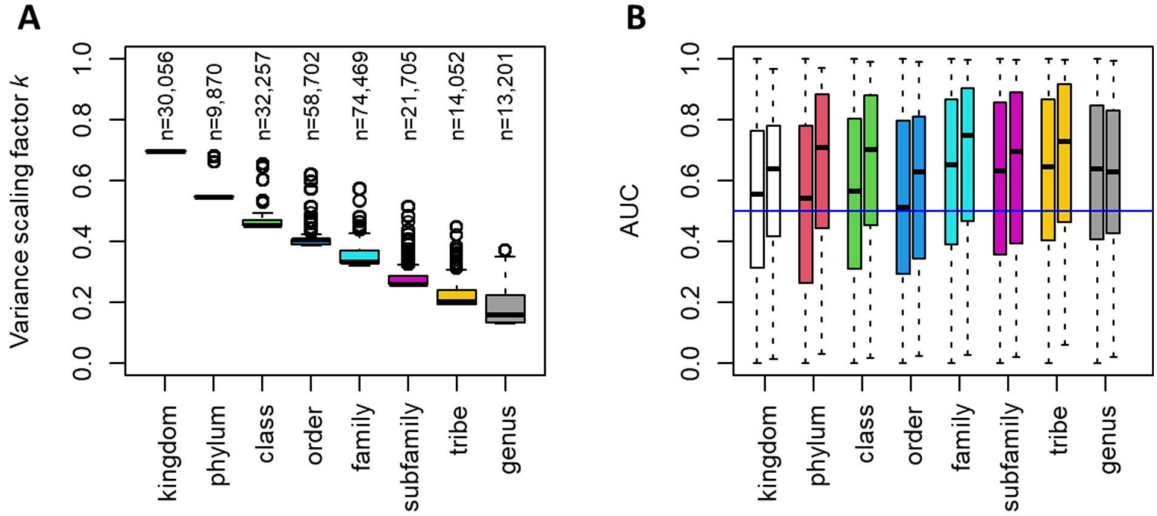

Figure 3. Conditional prior models for rare species constructed by borrowing information from the backbone model of common species. (A) Prior model precision measured by the variance scaling factor k (Eq. 5), as shown in relation to the taxonomic level shared with the closest relative in the backbone model. The numbers on top of the bars indicate the number of rare species in each category. (B) Discrimination powers of the conditional prior models, shown separately for each rank of the closest relative in the backbone model (different colors of bars) and for two prevalence classes (at most ten occurrences, left bars; more than ten occurrences, right bars). The blue line shows the null expectation AUC=0.

To compare the baseline and CORAL models in terms of predictive power, we considered those 22,140 species that were not included in the backbone model but occurred at least five times in the data. We applied two-fold cross validation, where we randomized the folds separately for each species, re-sampling until both the folds included at least 40% of the occurrences. We compared the models using AUC, Tjur's R^2, PRAUC, Brier score, negative log-likelihood and log determinant posterior covariance.

Together, these metrics provide a comprehensive overview of model performance covering predictive power, well calibrated probabilities, and useful inference (see *Methods* for their interpretations). All metrics of predictive performance improved considerably when moving from the baseline model to the CORAL model: AUC from 0.86 to 0.94, Tjur's R^2 from 0.03 to 0.08, PRAUC from 0.07 to 0.16, Brier score from 0.004 to 0.003, negative log-likelihood from 0.023 to 0.016, and log determinant from -28.2 to -36.4. All of these improvements were significant with $p<10^{-16}$ as measured by one-sided t-tests (see *Methods*). The improvement in the predictions was essentially independent of relatedness between the focal species and the species in the backbone model (Fig. 4), suggesting that most of the improvement derived from the inclusion of the latent factors estimated through the joint response of all common species, with less contribution from the direct borrowing of information from the related species. We validated this inference by fitting another set of models which included common species latent factors but not our novel prior; this approach retained about 75% of the gains in AUC over the baseline model. Additionally, the mean improvement in AUC did not essentially depend on the prevalence of the species (Fig. 4A), whereas for Tjur's R^2 the improvement was higher for the more common species (Fig. 4B).

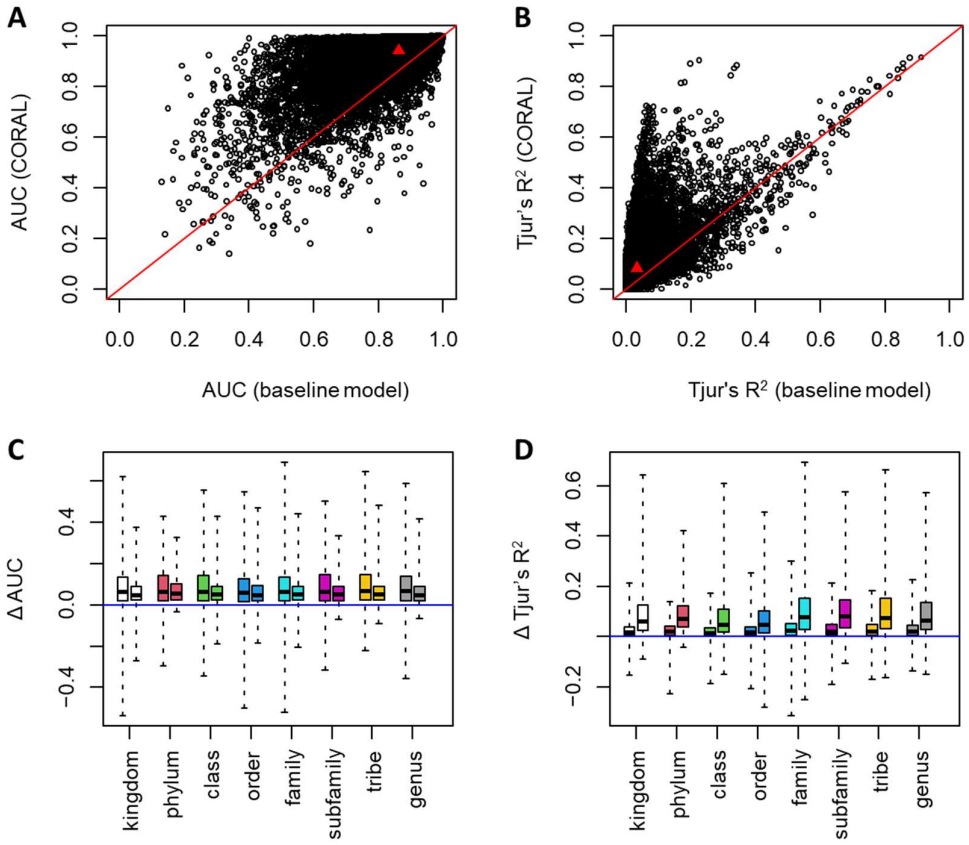

Figure 4. Comparison of predictive power for baseline and CORAL models based on two-fold cross validation. Predictive comparison is compared in terms of AUC (AC) and Tjur's R^2 (BD). The upper row of panels (AB) shows the raw values of the metrics for each species included in the analysis, with the red line showing the identity line and the red triangle showing mean values over the species. The lower row of panels (CD) shows the difference between the CORAL and baseline models. For values above 0 (the blue line), predictions by the CORAL model were more accurate. The results are shown separately for each rank of the closest relative in the backbone model (different colors of bars) and two prevalence classes (at most ten occurrences, left bars; more than ten occurrences, right bars).

To compare the baseline and CORAL models in terms of inference, we fitted them to all rare species, out of which 182,402 (71% of all species) were detected only once. By updating the conditional prior from the backbone model of common species with data from the focal rare species, we achieved improved predictions. The CORAL models showed higher precision than the baseline models (Fig. 5). This was especially the case for the very rare species (such as the one exemplified in Fig. 5B), for which the baseline models led to very large credible intervals, as may be expected for the low information contained in few occurrences. For more common species (such as the one exemplified in Fig. 5A), the increase in precision was smaller (Fig. 5C). The increase in precision increased with relatedness between the focal species and the species included in the backbone model (Fig. 5C), thus mirroring the relation seen between relatedness and the variance scaling factor (Fig. 3A).

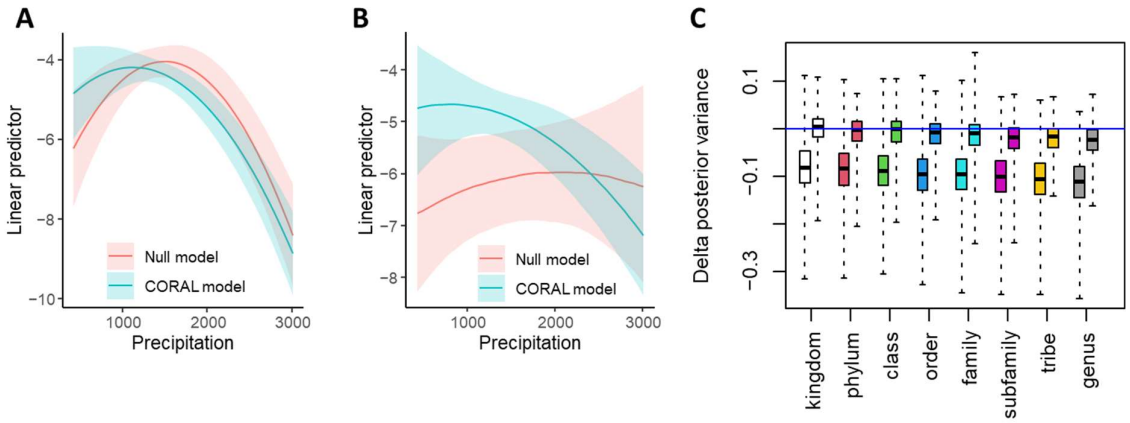

Figure 5. Comparison of inference between CORAL and baseline models. Panels A and B illustrate a specific prediction for two example species, out of which one is relatively common (A; the wall spider *Garcorops madagascar*, 10 occurrences) and the other very rare (B; the deer fly *Chrysops madagascarensis*, 1 occurrence). The panels show the posterior mean (line) and the interquartile posterior range (shaded area) of the linear predictor under changing precipitation, keeping temperature at its mean value over the data. Panel C compares the posterior variance between baseline and CORAL models systematically for all species. We averaged the posterior variance over the environmental predictors (excluding intercept and latent factors). The panel shows the difference between posterior variance in the CORAL and baseline models. Thus, for values below 0 (the blue line) the CORAL model shows smaller variance.

Discussion

The CORAL approach overcomes previous limitations on joint models of species communities with very large numbers of rare species. By borrowing information from a backbone model of common species, CORAL makes it possible to model even the rarest species in a statistically effective manner by combining an informative prior model with the limited data available for each rare species. As the rare species models can be parameterized independently, CORAL has an embarrassingly parallel implementation, making it feasible to fit it for even tens of millions of species, and hence to all species on Earth[1]. Rather than omitting rare species from all quantitative inference[15,16], the approach developed here enables one to draw on the full information inherent in novel community data[4]. This allows one to generate informed predictions about changes in communities and overall biodiversity with changes in the state of environmental drivers. In essence, this amounts to putting the "-diversity" back in "biodiversity".

We found species' responses to climatic, seasonal, and latent predictors to be phylogenetically structured (posterior mean $\rho = 0.65$, posterior probability $\Pr(\rho > 0) = 1.00$), forming the basis for borrowing information especially across related species. However, we note that even without phylogenetic signal in the data, or alternatively by fitting a model without phylogeny, CORAL would make it possible to borrow information from the backbone model of common species by identifying sample- and site-level latent factors, as well as by basing the conditional mean on the average response of all species.

To illustrate the scale of the gain, we reiterate the proportion of rare species in our samples: had we imposed a cutoff of species occurrence in 50 samples, we would have omitted 254,312 out of 255,188 species (99.7%), retaining 876 species (0.3%) of the species pool. Leaving the rare species unmodelled would hardly be an efficient use of the massive data painstakingly acquired. For the 22,140 species (8.7%) which occurred at least five times in the data, but which were not included in the backbone model, we scored a substantial improvement in predictive power by borrowing information from the more common species. This is a major achievement, as it shows how the limited information inherent in the distribution of rare species may be leveraged by gleaning information from more common species.

While this study focused on methodological development, our findings are also of major interest for understanding the eco-evolutionary community assembly processes of the Malagasy fauna. We found seasonality and climatic responses of arthropods to vary with their phylogenetic relatedness, suggesting that their distributions across Madagascar are partially constrained by their ancestral niche. This region is characterized by extreme levels of endemism at both a regional and a very small scale[40–42]. Nonetheless, in adapting to local conditions, the species appear to maintain a strong signal of their ancestral niche.

Acknowledgements

The work was funded by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 856506: ERC-synergy project LIFEPLAN to OO, DD and TR). OO's group was also funded by Academy of Finland (grant no. 336212 and 345110) and the European Union: HORIZON-CL6-2021-BIODIV-01 project 101059492 (Biodiversity Genomics Europe). JRd, SLd, MP, SR, JES, EVZ and PDNH were funded by support awarded to PDNH from the Government of Canada through its New Frontiers in Research Fund (NFRFT-2020-00073), the Large Scale Applied Research Program administered by Genome Canada and Ontario Genomics (OGI-208), and the Major Science Initiatives Fund administered by Canada Foundation for Innovation (project number 42450). The authors wish to acknowledge CSC – IT Center for Science, Finland, for computational resources.

References

1. Mora, C., Tittensor, D. P., Adl, S., Simpson, A. G. B. & Worm, B. How many species are there on Earth and in the ocean? *PLoS Biol* **9**, e1001127 (2011).

2. Stork, N. E. How many species of insects and other terrestrial arthropods are there on Earth? *Annu Rev Entomol* **63**, 31–45 (2018).

3. Callaghan, C. T., Borda-de-Água, L., van Klink, R., Rozzi, R. & Pereira, H. M. Unveiling global species abundance distributions. *Nat Ecol Evol* **7**, 1600–1609 (2023).

4. Hartig, F. *et al.* Novel community data in ecology-properties and prospects. *Trends Ecol Evol* **39**, 280–293 (2024).

5. van Klink, R. *et al.* Towards a global toolkit for insect biodiversity monitoring. *Philosophical Transactions of the Royal Society B: Biological Sciences* (2024).

6. Větrovský, T. *et al.* A meta-analysis of global fungal distribution reveals climate-driven patterns. *Nat Commun* **10**, 5142 (2019).

7. Tedersoo, L. *et al.* Global patterns in endemicity and vulnerability of soil fungi. *Glob Chang Biol* **28**, 6696–6710 (2022).

8. Jeliazkov, A. *et al.* Sampling and modelling rare species: Conceptual guidelines for the neglected majority. *Glob Chang Biol* **28**, 3754–3777 (2022).

9. Lomba, A. *et al.* Overcoming the rare species modelling paradox: A novel hierarchical framework applied to an Iberian endemic plant. *Biol Conserv* **143**, 2647–2657 (2010).

10. Ovaskainen, O. *et al.* How to make more out of community data? A conceptual framework and its implementation as models and software. *Ecol Lett* **20**, 561–576 (2017).

11. Helmstetter, N. A., Conway, C. J., Stevens, B. S. & Goldberg, A. R. Balancing transferability and complexity of species distribution models for rare species conservation. *Divers Distrib* **27**, 95–108 (2021).

12. Didham, R. K. *et al.* Interpreting insect declines: seven challenges and a way forward. *Insect Conserv Divers* **13**, 103–114 (2020).

13. Galante, P. J. *et al.* The challenge of modeling niches and distributions for data-poor species: a comprehensive approach to model complexity. *Ecography* **41**, 726–736 (2018).

14. Zhang, C., Chen, Y., Xu, B., Xue, Y. & Ren, Y. Improving prediction of rare species' distribution from community data. *Sci Rep* **10**, 12230 (2020).

15. Novotny, V. *et al.* Low beta diversity of herbivorous insects in tropical forests. *Nature* **448**, 692–695 (2007).

16. van Proosdij, A. S. J., Sosef, M. S. M., Wieringa, J. J. & Raes, N. Minimum required number of specimen records to develop accurate species distribution models. *Ecography* **39**, 542–552 (2016).

17. Le Lay, G., Engler, R., Franc, E. & Guisan, A. Prospective sampling based on model ensembles improves the detection of rare species. *Ecography* **33**, 1015–1027 (2010).

18. Breiner, F. T., Guisan, A., Bergamini, A. & Nobis, M. P. Overcoming limitations of modelling rare species by using ensembles of small models. *Methods Ecol Evol* **6**, 1210–1218 (2015).

19. Cadotte, M. W., Cardinale, B. J. & Oakley, T. H. Evolutionary history and the effect of biodiversity on plant productivity. *Proceedings of the National Academy of Sciences* **105**, 17012–17017 (2008).

20. Wiens, J. J. *et al.* Niche conservatism as an emerging principle in ecology and conservation biology. *Ecol Lett* **13**, 1310–1324 (2010).

21. Guénard, G., Ohe, P. C. von der, de Zwart, D., Legendre, P. & Lek, S. Using phylogenetic information to predict species tolerances to toxic chemicals. *Ecological Applications* **21**, 3178–3190 (2011).

22. Mondanaro, A. *et al.* ENphylo: A new method to model the distribution of extremely rare species. *Methods Ecol Evol* **14**, 911–922 (2023).

23. Qiao, H., Peterson, A. T., Ji, L. & Hu, J. Using data from related species to overcome spatial sampling bias and associated limitations in ecological niche modelling. *Methods Ecol Evol* **8**, 1804–1812 (2017).

24. Morales-Castilla, I., Davies, T. J., Pearse, W. D. & Peres-Neto, P. Combining phylogeny and co-occurrence to improve single species distribution models. *Global Ecology and Biogeography* **26**, 740–752 (2017).

25. Smith, A. B., Godsoe, W., Rodríguez-Sánchez, F., Wang, H.-H. & Warren, D. Niche estimation above and below the species Level. *Trends Ecol Evol* **34**, 260–273 (2019).

26. Warton, D. I. *et al.* So many variables: joint modeling in community ecology. *Trends Ecol Evol* **30**, 766–779 (2015).

27. Ovaskainen, O. & Soininen, J. Making more out of sparse data: hierarchical modeling of species communities. *Ecology* **92**, 289–295 (2011).

28. Abrego, N., Norberg, A. & Ovaskainen, O. Measuring and predicting the influence of traits on the assembly processes of wood-inhabiting fungi. *Journal of Ecology* **105**, 1070–1081 (2017).

29. Ovaskainen, O. & Abrego, N. *Joint Species Distribution Modelling*. (Cambridge University Press, 2020). doi:10.1017/9781108591720.

30. Norberg, A. *et al.* A comprehensive evaluation of predictive performance of 33 species distribution models at species and community levels. *Ecol Monogr* **89**, (2019).

31. Pichler, M. & Hartig, F. A new joint species distribution model for faster and more accurate inference of species associations from big community data. *Methods Ecol Evol* **12**, 2159–2173 (2021).

32. Rahman, A. U., Tikhonov, G., Oksanen, J., Rossi, T. & Ovaskainen, O. Accelerating joint species distribution modeling with Hmsc-HPC: A 1000x faster GPU deployment. *BioRxiv* (2024).

33. Ting, B., Wright, F. & Zhou, Y.-H. Fast multivariate probit estimation via a two-stage composite likelihood. *Stat Biosci* **14**, 533–549 (2022).

34. Chakraborty, A., Ou, R. & Dunson, D. B. Bayesian inference on high-dimensional multivariate binary responses. *J Am Stat Assoc* 1–12 (2023) doi:10.1080/01621459.2023.2260053.

35. Suder, P. M., Xu, J. & Dunson, D. B. Bayesian transfer learning. *ArXiv* 2312.13484, (2023).

36. Tikhonov, G. *et al.* Joint species distribution modelling with the R-package Hmsc. *Methods Ecol Evol* **11**, 442–447 (2020).

37. Elbrecht, V. *et al.* Validation of COI metabarcoding primers for terrestrial arthropods. *PeerJ* **7**, e7745 (2019).

38. Song, H., Buhay, J. E., Whiting, M. F. & Crandall, K. A. Many species in one: DNA barcoding overestimates the number of species when nuclear mitochondrial pseudogenes are coamplified. *Proceedings of the National Academy of Sciences* **105**, 13486–13491 (2008).

39. Wouters, H. Global bioclimatic indicators from 1979 to 2018 derived from reanalysis. Copernicus Climate Change Service (C3S) Climate Data Store (CDS). (Accessed on 14-04-2024). (2021).

40. Antonelli, A. *et al.* Madagascar's extraordinary biodiversity: Evolution, distribution, and use. *Science (1979)* **378**, (2022).

41. Vences, M., Wollenberg, K. C., Vieites, D. R. & Lees, D. C. Madagascar as a model region of species diversification. *Trends Ecol Evol* **24**, 456–465 (2009).

42. Wilmé, L., Goodman, S. M. & Ganzhorn, J. U. Biogeographic evolution of Madagascar's microendemic biota. *Science (1979)* **312**, 1063–1065 (2006).

METHODS

Deriving the CORAL Prior

CORAL is motivated by the default prior for coefficients in HMSC. Under this prior, the prior for a species r that is not part of the backbone model (i.e., a rare species) is given by $\beta_r \mid B \sim N(m_r, S_r)$, with the mean (4) and variance (5) given by the conditional multivariate normal formulas. The moments of this distribution are functions of HMSC parameters including Γ, ρ, V, and B and do not include information from the common species data, Y_c, a-priori. Fitting the backbone model produces a posterior distribution, π, over these parameters, which in turn implies a posterior marginal distribution for the rare species coefficients,

$$p(\beta_r \mid Y_c) = \int N(\beta_r; m_r, S_r)\pi(m_r, S_r|Y_c)dm_r\, dS_r$$

This updated distribution is our desired rare species prior. As this distribution is analytically intractable due to the integral over the posterior; we approximate it with a Gaussian:

$$p(\beta_r|Y_c) \approx N(\beta_r; m_r', S_r').$$

The mean m_r' and variance S_r' of this Gaussian are chosen to be the mean and variance of $p(\beta_r|Y_c)$, respectively. These can be calculated using the laws of total expectation/variance, resulting in simple expressions in terms of posterior means/variances: $m_r' = E_\pi[m_r]$ and $S_r' = E_\pi[S_r] + V_\pi[m_r]$. In practice, we approximate posterior means/variances using Monte-Carlo with posterior samples returned by HMSC. This completes specification of the CORAL prior.

Computational details

We fitted the backbone model with a high-performance computing accelerated version[32] of the R-package Hmsc[36], sampling each of the four chains for 37500 iterations. Of these chains, we omitted the first 12500 iterations as transient and then thinned the remaining chains by 100 to obtain 250 samples per chain and thus 1000 posterior samples in total.

For each rare species, we fitted a single-species model where we either did not (the baseline model) or did (the CORAL model) utilize information from the backbone models of common species. The baseline models were simple probit models with a Gaussian prior on the regression coefficients. The baseline models did not include the latent factors as predictors, and they assumed a default prior distribution for the species responses (N(0, 10) for the intercept and N(0, 1) for fixed effect coefficients). In the CORAL models, we included the latent factors as predictors, and assumed the

conditional prior distribution based on Eqs. 3-5. We obtained 5000 samples after 2500 transient iterations for each species for both the baseline and CORAL models using MCMCpack[43].

Metrics used to evaluate model performance

AUC is the probability a randomly chosen positive sample has a higher predicted probability than a randomly chosen negative sample. Tjur's R^2 is a pseudo-R^2 value which can be read like any other R^2 value, but typically reaches lower values[44]. PRAUC is the area under the precision recall curve – this metric quantifies true positives and is useful for analyzing highly imbalanced data where the minority class is of primary interest. The Brier score is the average squared error between predicted probabilities and labels – this metric penalizes overconfidence. Negative log-likelihood directly measures goodness of fit under the proposed Bernoulli model. The determinant of the posterior covariance determines the volume of a 95% credible interval for fixed effect coefficients under a Gaussian approximation, smaller intervals meaning more confident inference.

Sampling of Malagasy arthropods

We selected 53 locations across Madagascar that were relatively undisturbed and where the vegetation represents the conditions of the local environment. 28 of the sites were sampled in a spatially nested sampling design with decreasing distances between them (50 km, 5 km and 500 m apart). The other 25 sites were spread across different forested habitats in Madagascar (dry, lowland and montane forests), at elevations ranging from 8 to 1592 MASL. We continuously collected one-week samples of flying arthropods in 95% ethanol using Malaise traps (ez-Malaise Trap, MegaView Science Co.) A detailed sampling protocol is published at https://dx.doi.org/10.17504/protocols.io.kqdg3xkdqg25/v2.

DNA extraction and sequencing

Ethanol filtration and tissue lysis. Ethanol was filtered from five bulk samples of insect specimens at a time using five Microfunnel 0.45 uM Supor membrane filters (Pall Laboratory), an analytic and precision weighing balance (Sartorius), and a 6-Funnel manifold (Pall Laboratory). First the weight of each of the five membrane filters was tared with source weights recorded in CBG's Collection Information Management System and then placed on five of the six funnels on the manifold. The bulk sample of insect specimens were filtered through the membrane on the manifold until all specimens were transferred from the bulk bottle to the membrane and all ethanol was removed. The membrane and bulk sample contents were weighed using the balance with the initial tared weight subtracted, to measure the resulting wet arthropod biomass of the bulk sample. Lastly, the insect specimen contents and their membrane were transferred back to the bulk sample bottle for lysis. The wet arthropod biomass was recorded for the remaining four bulk samples of insect specimens in the subset by repeating these steps. The amount of insect lysis buffer was determined based on the wet arthropod biomass and this volume of buffer was added to the bulk bottle containing the filtered bulk sample of insect specimens, ensuring that all specimens were submerged in the insect lysis buffer. Bulk samples were incubated, rotating gently on a shaker, at 56°C overnight. After lysis, 3 technical replicates of 300 µl aliquots of lysate were taken from each sample in a set of 30 samples and transferred to a 500 µl 96-well deep well plate (Eppendorf) along with wells designated for 3 positive AMPtk synthetic[45] controls and 3 negative (template-free) controls per block. Subsequently, 50 µl of lysate for each well was transferred from the block to a fully skirted 96-well microplate for DNA extraction and the remaining 250 µl of lysates in the block were archived for future analysis. Lysate remaining in the bottles was filtered and discarded, whereas the specimens were archived in 96% ethanol for future use and stored at -20°C.

DNA extraction and PCR amplification DNA extractions followed a membrane-based protocol[46]. First, 100 µl of binding mix was added to the microplate containing 50 µl of lysate and the total volume of 150 µl was mixed and transferred to a 3.0 µm Pall Supor Membrane glass fiber plate (Pall Laboratory) and centrifuged for 5 min at 5000 g. Then, 180 µl of protein wash buffer was added and centrifuged for 2 min at 5000 g followed by two additional washes with 600 µl of wash buffer with centrifugation for 5 min at 5000 g for both washes. The 3.0 µm Pall Supor Membrane glass fiber plate was placed onto a new microplate used for DNA storage and incubated at 56°C for 30 min. Finally, 60 µl of elution buffer was added, incubated at room temperature for 1 min, and centrifuged for 5 min at 5000 g to elute the DNA. The components of each buffer used in tissue lysis and DNA extraction are the following[46]: Insect lysis buffer – 700 mm GuSCN, 30 mM EDTA pH 8.0, 30 mM Tris-HCl pH 8.0, 0.5% Triton X-100, and 5% Tween20; Binding buffer (BB) – 6 M GuSCN, 20 mM EDTA pH 8.0, 10 mM Tris-HCl pH 6.4, and 4% Triton X-100, prewarmed at 56 °C to dissolve; Binding mix – 50 mL of ethanol (96%) thoroughly mixed with 50mL of BB; Protein wash buffer – 70 mL of ethanol (96%) thoroughly mixed with 26 mL of BB; and Wash buffer – ethanol (60%), 50 mM NaCl, 10 mm, Tris-HCl pH 7.4 and 0.5 mM EDTA pH 8.0. A 418 bp region of cytochrome c oxidase subunit I (COI) was amplified from two rounds of PCR amplification using the forward primer BF3 and the reverse primer BR2[37,47]. The PCR1 reaction amplifies the BF3/BR2 fragment and the PCR2 reaction adds the P5/P7 Illumina adapters and inline indices. PCR1 reactions were carried out with the following thermocycler settings: initial denaturation at 95°C for 5 minutes, 30 cycles of denaturation at 94°C for 30 seconds, annealing at 46°C for 30 seconds, extension at 72°C for 50 seconds, and lastly extension for 5 minutes at 72°C. A 2 µl volume of DNA template and AMPtk positive control were added to corresponding wells. First-round PCR products were diluted with equal volume of ddH20 before the second round of PCR. PCR2 reactions employed fusion versions of BF3 and BR2 primers tailed with one of 24 6 to 10-base long in-line tags and P5 or P7 sequencing adaptors (Illumina) for a total of 1152 unique combinations of tag pairs. The total number of unique combinations factors in half of the samples being sequenced in reverse direction by swapping P5 and P7 tails on the indexed primers. PCR2 reactions were carried out with the following thermocycler settings: initial denaturation at 94°C for 2 minutes, 15 cycles of denaturation at 94°C for 40 seconds, annealing at 51°C for 1 minute, extension at 72°C for 1 minute, and lastly extension for 5 minutes at 72°C. A 2 µl volume of PCR1 template, including positive and negative controls, was added.

Library construction. The PCR2 products labeled with P5/P7 Illumina indexing adapters were checked by running 4 µL of the product on the Invitrogen e-Gel 2% Agarose gel for 4 min. An equal volume of each sample (6 µL) was pooled into one 96-well microplate using an Biomek i7 Automated Liquid Handling Workstation (Beckman Coulter LIfe Sciences, Indianapolis) robot. The pooled PCR products were then transferred into a 5 – 15 mL tube. The pooled DNA library was concentrated with a 0.8 – 1.2 ratio of magnetic beads (Cytiva Sera-Mag™ Carboxylate-Modified Magnetic SpeedBeads) to pooled PCR product. The DNA library was vortexed and spun. Four 1.5 mL Eppendorf LoBind tubes were prepared with pre-aliquoted 160 uL of magnetic beads. The pooled DNA library was vortexed and 200 µL of it was added to each tube. The standard recommendations for the library concentration were applied: 8 min on the bench at room temperature, two min on the magnet, three washes with 1 mL of 80% freshly prepared ethanol, elution with 45 µL of water, and collecting 40 µL from each tube. The total volume of the concentrated DNA library was 160 µL. To target the desired length of amplicons, and thereby eliminate non-specific, especially long fragments of DNA from the DNA library, we used an Invitrogen Size-Select 2% agarose gel, according to the protocol recommended by the manufacturer. The final volume of the target DNA was approximately 200 µL. The volume of the target DNA library was evaluated by using a Rainin pipette (Mettler Toledo, Mississauga, ON). The 0.8 ratio of Cytiva magnetic beads:pooled library was applied. For library

concentration, we used the standard recommendations of the manufacturer, including 8 min on the bench at room temperature, 2 min on the magnet, four washes with 1 mL of 80% freshly prepared ethanol, elution with 23 µL of water, and collecting 20 µL. To evaluate the concentration of the final DNA library, we used the Invitrogen Qubit High Sensitivity kit. The normalized 1 ng/uL concentration of the final library was checked on the Bioanalyzser (Agilent Technology, Santa Clara) using the High Sensitivity kit in order to visualize the length of the target DNA.

Sequencing. Sequencing was completed on an Illumina NovaSeq 6000 platform with the S Prime (SP) reagent kit, SP flow cell, and 500-cycle configuration. Each lane received its own library with 1152 dual-indexed amplicons. Because multiplexing employed inline tags and did not employ any native Illumina indices, indexing flows were redirected to sequencing of the insert for both reads, effectively making it 2x258bp paired-end sequencing. Up to 15% PhiX spike-in was used to ensure base diversity. Each sequencing run produced two sets of files, one for each lane, which included a single read 1 file and a single read 2 file. Reads from the three replicates for each sample were demultiplexed with Cutadapt v3.5[48] with no indels and no mismatches allowed in the index sequence search (--no-indels, -e0), requesting a minimum overlap of 9 bp with the 10-bp index sequence (--overlap 9) and minimum length of the demultiplexed output sequence was 99 bp (--minimum-length 99).

Bioinformatics

We used a development version of the OptimOTU pipeline[49] to process the raw sequencing reads into a species x sample occurrence matrix and taxonomy for each species. In brief, the workflow consisted of primer removal[48], quality filtering[48,50], denoising[50], de novo[50] and reference-based[51,52] chimera removal, flagging likely non-animal sequences[53], removal of putative nuclear-mitochondrial pseudogenes[38], probabilistic taxonomic assignment[54,55], and finally taxonomically-guided hierarchical clustering[49,56]. This yielded a full taxonomic tree with approximate placeholder taxa to group those sequences which could not be reliably identified.

Primer removal and quality filtering. Paired-end reads were trimmed and quality filtered using Cutadapt version 4.2[48]. Both ends of each read were first quality-trimmed with a threshold of 2 (--q 2,2 --Q 2,2), effectively removing completely ambiguous bases, i.e., N's. Then the BF3 and BR2 primer sequences were searched for and removed, taking into account the amplicon orientation for each sample (see above), with a maximum mismatch rate of 0.2 (-e 0.2) and a minimum overlap of 1 base (-O 10). Read pairs where the appropriate primer was not found at the 5' end of both R1 and R2 were discarded, but primers at the 3' end were not required. After trimming, read pairs where either read still included N's were discarded (--max-n 0), as were pairs where either read was less than 100bp long (-m 100).

Denoising. Trimmed read pairs were then further filtered, denoised, and merged using DADA2 version 1.26.0. First, read pairs where either read had an expected number of errors greater than 1 (maxEE=1), or matched the PhiX genome (rm.phix=TRUE), were discarded using the filterAndTrim() function. DADA2 error profiles were then fitted separately for each sequencing run, read, and orientation using default parameters with the learnErrors() function. Control samples were not included in the profile fitting process. Sequences were then denoised using the dada() function, using the relevant error profile and default settings. Finally, denoised read pairs were merged using the mergePairs() function with default settings. The denoised R1 and R2 reads from reverse-oriented samples were swapped at this stage, so that the merged, denoised amplicon sequence variants (ASVs) were all in the same orientation.

ASV table construction. Due to the very large number of samples and high diversity of ASVs, the standard DADA2 function to combine denoised sequences from multiple samples to form a matrix had prohibitively high memory requirements. We instead formed a list of all unique ASVs by first finding unique ASVs in each sample, then in each sequencing run, and then finally across the entire data set. These sequences were then stored in a single compressed fasta file, which we indexed with FastqIndEx version 0.9.0b (https://github.com/DKFZ-ODCF/FastqIndEx/releases/tag/0.9.0b) for fast extraction of contiguous subsets. The matrix of ASV read counts in each sample was then constructed and stored in coordinate list (COO) sparse matrix format, as a three column R data frame where the columns represent a sample identifier, index of the sequence in the master ASV file, and number of reads. Potential cases of tag-switching were detected and removed using an R implementation of the UNCROSS2 algorithm[57], using a threshold uncross score of f=0.01. Because different sequencing runs were processed and sequenced independently, UNCROSS2 was applied only within sequencing runs.

A number of operations were then performed on the ASVs, including de novo and reference-based chimera detection, scoring and aligning to a hidden markov model, removal of potential nuclear-mitochondrial pseudogenes, flagging potential non-animal sequences, and taxonomic assignment. These steps, most of which flag some ASVs for removal, could in principle be performed in stages, thereby reducing the computational load of later stages – because only sequences which passed all previous stages would then need to be considered. We instead performed them in parallel, so that it is possible to compare the results of different stages. As such comparisons are not relevant to CORAL, we do not present them here.

Reference sequences and taxonomy. Reference sequences and taxonomy were retrieved from BOLD[53]. The taxonomic hierarchy on BOLD is a consolidation of numerous online databases and taxonomic publications. The hierarchy is constantly updated through both user submissions of new taxon names and the incorporation of new taxon-specific checklists. Some of the key data sources are listed on the Taxonomic Resources page on BOLD (https://boldsystems.org/index.php/resources/taxresource). The version of the hierarchy used in this study was retrieved from BOLD on April 21, 2022. Two BOLD data packages were utilized in this study. A 7M reference barcode set, used for training the Protax taxonomic classifier (see below), consisted of all COI data publicly available on BOLD as of June 1, 2022. Any sequences flagged due to data quality issues, such as reading frame shifts, or suspected cross-contamination or misidentification, as well as non-animal sequences, were excluded from this reference dataset. An additional 8.7M barcode data set, in which non-animal sequences were retained for use in chimera detection and flagging non-animal sequences (see below), was downloaded Feb 9, 2024 (https://www.boldsystems.org/index.php/datapackage?id=BOLD_Public.09-Feb-2024). The majority of sequences in both datasets originate from research programs led by the International Barcode of Life (iBOL) Consortium and the Centre for Biodiversity Genomics, University of Guelph, such as the Global Malaise Trap Program (https://biodiversitygenomics.net/projects/gmp/) and BIOSCAN (https://biodiversitygenomics.net/research/bioscan/).

Chimera detection. De novo chimera detection was performed using a partial re-implementation of the consensus method of DADA2 in order to operate on the sparse matrix representation. The internal DADA2 function C_table_bimera2 was used to flag potential bimeras in small subsets of sequences, and the results were combined into a master list. Each ASV which was flagged as a potential bimera in at least 90% - 1 of the samples it occurred in was removed.

Reference-based chimera detection was performed using the UCHIME algorithm[51] as implemented in VSEARCH (--uchime_ref), using as reference the 8.7M barcode dataset described above.

HMM alignment and NuMt detection: Both the ASV sequences and the 7M and 8.7M barcode reference sets were aligned to the nucleotide hidden Markov model (HMM) of the COI barcode region from[58] using the hmmalign command of HMMER version 3.3.2[59]. Extra nucleotides at the beginning and end of the aligned region were removed (--trim). Insertions, represented by HMMER as lower-case characters in the aligned sequences, were removed from the reference data sets, but retained at this stage in the ASV sequences. ASV sequences with frame shift mutations (insertions or deletions with length not divisible by 3) or in-frame TAG/TAA stop codons were discarded as potential nuclear mitochondrial pseudogenes (NuMts)[38]. Additionally, the ASV sequences were scored against the same HMM with the nhmmer command, also from HMMER 3.3.2. ASVs which did not match the model from at least position 245 to 652, corresponding to the BF3-BR2 amplicon with a few bases truncated at each end, or for which the calculated match score was less than 200, were regarded as likely sequencing artifacts and discarded.

Taxonomic placement. Taxonomic placement of animal sequences was conducted with the Protax classifier[59] using parameterization similar to the one used in[60] and trained on the 7M barcode reference set. During classification, the nucleotides were represented as 64-bit integers in order to gain speedup when computing pairwise distances. Each nucleotide was converted into a 4-bit vector (A: '1000', C: '0100', G: '0010', T: '0001', N: '0000') so that 16 consecutive nucleotides could be represented in one 64-bit integer. The number of matching nucleotides along the alignment could then be calculated efficiently by using a fast built-in function 'popcountl' in C. The number of taxa after adding unknown branches (one outside-of-known-taxonomy node under each parent node) was 37, 157, 882, 7963, 18676, 25832, 102736, and 572097 in phylum, class, order, family, subfamily, tribe, genus, and species level, respectively. There were 7279191 reference sequences whose annotations reached 99% family level, 69% subfamily level, 64% genus level, and 42% species level. The number of taxonomy nodes with at least one reference sequence was 29 at the phylum level, 103 at the class level, 605 at the order level, 4909 at the family level, 7580 at the subfamily level, 9161 at the tribe level, 50659 at the genus level, and 199841 at the species level.

The same pairwise distance algorithm was also used to find the single closest match for each ASV in the 8.7M barcode reference set. Sequences whose closest match was to a non-animal sequence were flagged but retained until after the clustering step.

Taxonomically guided clustering: Due to frequent intra-species variation in the COI gene[60], ASVs are not a reliable proxy for unique species. We therefore performed a multi-stage hierarchical clustering of the ASV sequences, taking into account the results of taxonomic classifications from Protax. Clustering was performed first at the phylum level, then proceeded through the taxonomic ranks until the final round resulted in approximately species-level OTUs, which were used as the main unit of analysis for CORAL. At each rank, clustering proceeded in three stages: core formation, closed-reference clustering, and de-novo clustering. In the core formation stage, ASVs which were identified at the current taxonomic rank with at least 90% confidence by Protax were grouped according to their taxonomic identifications. In the closed-reference clustering stage, ASVs which could not be identified with 90% confidence were joined to cluster cores if the pairwise sequence dissimilarity, calculated by the Protax algorithm, between the ASV and a member of the cluster core was less than a rank-specific threshold of 0.12 for phylum, 0.10 for class, 0.07 for order, 0.05 for family, 0.04 for subfamily, 0.035 for tribe, 0.03 for genus, and 0.02 for species. If an ASV matched more than one cluster core at below the given threshold, it was joined only to the one which it matched most closely, so that cluster cores were never merged. This stage was repeated, checking for matches

between unclustered ASVs and newly clustered sequences, until no new matches were found. Finally, in the de novo clustering stage, remaining unclustered OTUs were single-linkage clustered, with the same rank-specific dissimilarity thresholds. Placeholder "pseudotaxon" names at each rank were given to the de novo clusters, such that the final OTUs were placed in a full taxonomic hierarchy consisting of both known taxa and potentially novel taxa. Finally, all ASVs in pseudophyla (i.e., de novo clusters at phylum rank) which contained a large number of ASVs which were flagged as non-animal (see above) were removed.

Implementation and execution: The OptimOTU pipeline is implemented using the targets 1.5.1 workflow management package[61], here executed using the crew 0.9.0[62] and crew.cluster 0.3.0[63] backends in R 4.2.3[64] on the Puhti cluster at CSC – IT Center For Science, Finland.

Additional references

43. Martin, A. D., Quinn, K. M. & Park, J. H. MCMCpack: Markov chain Monte Carlo in R. *J Stat Softw* **42**, (2011).

44. Abrego, N. & Ovaskainen, O. Evaluating the predictive performance of presence–absence models: Why can the same model appear excellent or poor? *Ecol Evol* **13**, (2023).

45. Palmer, J. M., Jusino, M. A., Banik, M. T. & Lindner, D. L. Non-biological synthetic spike-in controls and the AMPtk software pipeline improve mycobiome data. *PeerJ* **6**, e4925 (2018).

46. Ivanova, N. V., DeWaard, J. R. & Hebert, P. D. N. An inexpensive, automation-friendly protocol for recovering high-quality DNA. *Mol Ecol Notes* **6**, 998–1002 (2006).

47. Elbrecht, V. & Steinke, D. Scaling up DNA metabarcoding for freshwater macrozoobenthos monitoring. *Freshw Biol* **64**, 380–387 (2019).

48. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J* **17**, 10 (2011).

49. Saine, S. *et al.* Natural deadwood hosts more diverse pioneering wood-inhabiting fungal communities than restored deadwood. *Restor Ecol* **32**, (2024).

50. Callahan, B. J. *et al.* DADA2: High-resolution sample inference from Illumina amplicon data. *Nat Methods* **13**, 581–583 (2016).

51. Edgar, R. C., Haas, B. J., Clemente, J. C., Quince, C. & Knight, R. UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics* **27**, 2194–2200 (2011).

52. Rognes, T., Flouri, T., Nichols, B., Quince, C. & Mahé, F. VSEARCH: a versatile open source tool for metagenomics. *PeerJ* **4**, e2584 (2016).

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

54. Somervuo, P., Koskela, S., Pennanen, J., Henrik Nilsson, R. & Ovaskainen, O. Unbiased probabilistic taxonomic classification for DNA barcoding. *Bioinformatics* **32**, 2920–2927 (2016).

55. Roslin, T. *et al.* A molecular-based identification resource for the arthropods of Finland. *Mol Ecol Resour* **22**, 803–822 (2022).

56. Vu, D., Nilsson, R. H. & Verkley, G. J. M. Dnabarcoder: An open-source software package for analysing and predicting DNA sequence similarity cutoffs for fungal sequence identification. *Mol Ecol Resour* **22**, 2793–2809 (2022).

57. Edgar, R. C. UNCROSS2: identification of cross-talk in 16S rRNA OTU tables. https://doi.org/10.1101/400762. *BioRxiv* (2018).

58. Roslin, T. *et al.* A molecular-based identification resource for the arthropods of Finland. *Mol Ecol Resour* **22**, 803–822 (2022).

59. Eddy, S. R. Accelerated profile HMM searches. *PLoS Comput Biol* **7**, e1002195 (2011).

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

61. Landau, W. The targets R package: a dynamic Make-like function-oriented pipeline toolkit for reproducibility and high-performance computing. *J Open Source Softw* **6**, 2959 (2021).

62. Landau, W. M. crew: A Distributed Worker Launcher Framework. https://wlandau.github.io/crew/, https://github.com/wlandau/crew. Preprint at (2023).

63. Landay, W. M. & Levin, M. G. crew.cluster: Crew Launcher Plugins for Traditional High-Performance Computing Clusters. R package version 0.3.0, https://CRAN.R-project.org/package=crew.cluster. Preprint at (2024).

64. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/. Preprint at (2023).