PNAS RESEARCH ARTICLE | ECOLOGY OPEN ACCESS

Soil eDNA reflects regionally dominant species rather than local composition of tropical tree communities

Francesc Borràs Sayas[a], Ottavia Iacovino[b], María Uriarte[c], Jess K. Zimmerman[d], Christopher J. Nytch[d], Glenn Dunshea[b,1,2], and Robert Muscarella[a,1,2]

Affiliations are included on p. 9.

Edited by Janet Franklin, San Diego State University, San Diego, CA; received April 7, 2025; accepted July 23, 2025

Environmental DNA (eDNA) is increasingly used for biodiversity monitoring, but validation of the spatial scale(s) at which eDNA reflects extant communities is scarce, particularly in tropical forests: the terrestrial biome with the most concentrated diversity on earth. We leveraged spatially explicit tree inventory data from the 16-ha Luquillo Forest Dynamics Plot (LFDP) in Puerto Rico to validate soil eDNA as a spatially explicit indicator of tree diversity/composition. Using a comprehensive local chloroplast *trnL*-P6 reference library, we analyzed soil samples at multiple scales through eDNA *trnL*-P6 metabarcoding. We compared eDNA taxonomic diversity/composition, considering several bioinformatic thresholds, with inventory data across a range of spatial scales, as well as random points to compare observed correlations with random expectations. Despite considerable fine-scale heterogeneity in soil plant eDNA composition, we detected 53 tree Operational Taxonomic Units (OTUs) across the LFDP, corresponding to 68% of tree OTUs from the census data. Encouragingly, this equated to 98% of the total basal area (and 98% of the total stems). An initial confusion matrix evaluation suggested a highly localized eDNA signal (within 5 m of the sampled locations). However, comparison with random expectations revealed a lack of support for a fine-scale spatial signal due to misclassification (i.e., eDNA false presence or/and false absence) of relatively common taxa. Our study shows that "universal" PCR primer metabarcoding of tree eDNA in tropical soils may be useful for assessing dominant taxa at landscape scales, but not for spatially explicit characterization of rare species and community composition at local scales.

environmental DNA | Luquillo Forest Dynamics Plot | Puerto Rico | trnL | biodiversity monitoring

Environmental DNA (eDNA) has emerged as a powerful and popular tool for characterizing and monitoring spatiotemporal patterns of biodiversity (1–7). Despite exciting potential, significant challenges limit inferences about the multidimensional concept of biodiversity based on DNA molecules in the environment (3, 8–10). Especially in terrestrial ecosystems, a largely unaddressed challenge of eDNA metabarcoding studies is understanding the spatial scale(s) of biodiversity reflected in eDNA samples (8, 9).

Few studies in terrestrial systems have explicitly investigated the spatial scales that relate to eDNA data samples (11–13). Results from some studies have supported the intuitive interpretation that plant eDNA in soil represents a highly localized community, as nearby plants are assumed to be the most likely source for local eDNA (13–15). However, cellular material of individual organisms (e.g., pollen, seeds, leaves) can disperse long distances via stochastic processes (e.g., wind, animal movement) (16). Additionally, extracellular DNA itself can also be mobile in soils (17, 18). Thus, the distribution of cellular material and DNA from taxa detected from eDNA samples may not accurately reflect the local distribution of living organisms. Accordingly, other studies have suggested that soil eDNA samples may better reflect more regional patterns of biodiversity (11, 17). Existing comparisons of soil eDNA to botanical inventories across relatively uniform habitats have typically been restricted to either fairly large (e.g., >20 ha) or small (<16 m^2) spatial scales (e.g., (14, 15). A systematic, spatially explicit approach encompassing a range of scales in a single habitat may better elucidate the spatial scale(s) of plant community composition captured in eDNA samples.

Studies aiming to evaluate eDNA approaches for biodiversity assessment by comparison to other data sources require selection of validation criteria. While some studies have emphasized alpha diversity metrics (e.g., taxonomic richness), for eDNA data, these can depend on laboratory considerations such as PCR replication and sequencing depth (18, 19). Beta diversity metrics (e.g., Bray–Curtis dissimilarity) may show congruence between eDNA and other community composition approaches (e.g., 20), but are challenging to

Significance

Spatially explicit observations of organisms are fundamental to ecological research. Environmental DNA (eDNA) is increasingly used for biomonitoring, but comprehensive tests of the spatial scale(s) at which eDNA reflects biodiversity are rare. By comparing spatially explicit tree inventory data from the 16-hectare Luquillo Forest Dynamics Plot (LFDP) in Puerto Rico with soil eDNA metabarcoding data, we found that while most common taxa were detected, many rare taxa were undetected, and eDNA was no better an indicator of the standing tree community in a specific location than at other random locations in the plot. Our study provides an explicit assessment of spatial scales for interpreting tropical forest tree eDNA in soil, offering critical insights for applying this technique in forest biodiversity research.

Author contributions: G.D. and R.M. designed research with input from F.B.S.; F.B.S., O.I., G.D., and R.M. performed research; all authors contributed data; G.D. and R.M. analyzed data with input from F.B.S.; F.B.S., R.M., and G.D wrote the paper with input from all authors.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1] G.D. and R.M. contributed equally to this work.

[2] To whom correspondence may be addressed. Email: glenn.dunshea@ntnu.no or robert.muscarella@ebc.uu.se.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2505772122/-/DCSupplemental.

Published August 19, 2025.

interpret with respect to taxon-specific information about how eDNA corresponds to the actual taxa present, even where critical experimental choices such as marker/primer selection, sequence variability, and reference library completeness are optimal. For direct comparison of eDNA with community data, focusing on only the proportion of correctly classified presences/absences can yield a biased view of performance (20), particularly since such datasets are inevitably unbalanced (i.e., many more absences than presences). A confusion matrix approach that incorporates both true and false presences and absences can provide a more comprehensive view of how biodiversity is reflected in eDNA samples, and more clear conclusions regarding the detection or nondetection of individual taxa (21, 22).

An emergent pattern from soil metabarcoding studies for plants is that they tend to detect taxa that are highly abundant across the landscape (11, 17). By default, more abundant taxa are expected to occur near any random sampling location relatively more often than low abundance (i.e. rare) taxa, by chance alone. Biased detection of regionally abundant (and thus also nearby) taxa in eDNA samples could limit the utility of eDNA in describing local community composition. Comparing observed composition of eDNA with random expectations based on the abundance and distribution of taxa across a study area can help disentangle the effects of abundant taxa on our ability to infer a truly localized spatial signature of eDNA.

The 16-ha Luquillo Forest Dynamics Plot (LFDP) in Puerto Rico (Fig. 1*A*) offers an exceptional opportunity to examine the spatial correspondence between soil eDNA and tree composition of a diverse tropical forest. In the LFDP, the size, identity, and location of all stems ≥1 cm diameter at 1.3 m above the ground have been recorded approximately every 5 years since 1990, and most recently in 2023, thus providing a spatially explicit reference map of observed tree diversity. To assess the congruence between tree biodiversity based on soil eDNA metabarcoding and the LFDP tree census, we collected leaf samples from 104 species to generate a local reference library for the trnL (UAA) intron P6 loop, a commonly used marker for plant eDNA studies because of relatively high performance with degraded material (23). We then performed metabarcoding on soil samples from 40 locations positioned systematically across the entire LFDP ("big grid" locations; Fig. 1*A*). For each location, we homogenized three subsamples haphazardly collected from within 4 m^2 to get a representative sample at each location for comparison with the LFDP census data. We used two supplemental approaches to examine fine-scale heterogeneity in soil plant eDNA. First, we performed metabarcoding on each of the three subsamples from three of the 40 LFDP locations ("unpooled" samples). Second, we intensively sampled one randomly located 4 m^2 area with a set of 12 soil samples ("intensive plot" samples). By analyzing the "unpooled subsamples" and the "intensive plot" samples, we aimed to contextualize results from the 40 LFDP samples by 1) assessing the effectiveness of subsample pooling to capture OTUs present in subsamples and 2) assessing the extent of very fine scale (10 - 100 cm) soil plant eDNA heterogeneity to estimate the amount of sampling required to capture all OTUs present within the area of a "big grid" sampling point and to examine whether there was a trade-off between sampling intensity and misclassification rates comparing eDNA with census data. Several bioinformatic iterations of the soil metabarcoding data, increasing in filtering stringency, were used for all subsequent analyses to examine potential effects on results.

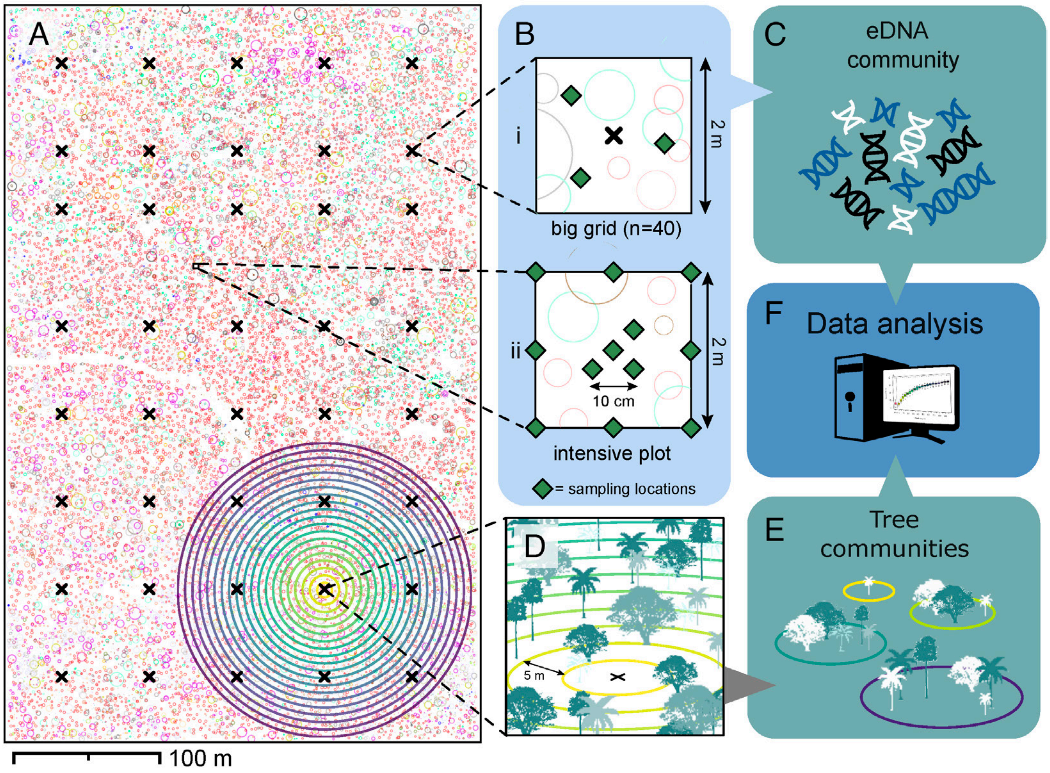

Fig. 1. A schematic diagram of the sampling regime and workflow for this study. (*A*) A map of the 16-ha Luquillo Forest Dynamic Plot (LFDP) with measured, identified, and individually mapped trees. Crosses show 40 locations for "big grid" soil eDNA sampling. (*B*) Big grid sampling: At each of the 40 points, three soil samples were collected haphazardly from within 2 m and homogenized prior to sequencing. At 3 of the 40 locations, we also sequenced the 3 "unpooled" subsamples. "Intensive plot" sampling: To examine eDNA heterogeneity at fine spatial scales, we collected 12 individual samples from a single 4 m^2 area. (*C*) We generated a community matrix of OTUs based on the soil eDNA samples. (*D* and *E*) We used LFDP tree census data (completed in 2023) to characterize tree neighborhoods in concentric circles around each sampling location at radii of 5 to 100 m, in 5 m increments. (*F*) We used a suite of analyses to compare the soil eDNA community with the tree community (in different sized neighborhoods) at each sample location.

Overall, we examined two main questions: First, to what extent does community composition based on eDNA correspond with results from census data? And second, at which spatial scale(s) does eDNA most accurately reflect census data?

Results

Reference Library. Of the 104 tree species collected for the reference library, all amplified and sequenced successfully. Seventy-one (68%) had a unique haplotype while 33 (32%) did not have a unique haplotype (and shared a total of 12 haplotypes), which is expected for this marker (24). We pooled species with shared haplotypes into Operational Taxonomic Units (OTUs) and also merged—based on taxonomy—an additional 15 LFDP species for which we did not have sequence data (see *Methods* and Dataset S1). In four cases, we used a conservative approach to pool unique haplotypes because BLAST results indicated potential field errors or ambiguity given the taxonomic composition of the tree census data. This process resulted in a final number of 79 OTUs representing 123 species [note that our downstream results were nearly identical if we excluded the 15 species without sequence data from the analysis (*SI Appendix*, Figs. S27–S38)]. Together, these OTUs comprised ~88% of all species (and >99% of stems and basal area) in the 2023 LFDP census. Full details related to bioinformatic procedures are provided in *SI Appendix* and merging of OTUs are provided in Dataset S1.

eDNA Sequencing and Small-Scale Heterogeneity. Full details related to bioinformatic procedures are provided in the *SI Appendix*. Briefly, following initial bioinformatics and quality control of sequencing libraries, there were ≈14.2 M reads across samples/controls in three PCR technical replicates, with ≈170 K ± 24 K (mean ± SE) reads per soil sample (*SI Appendix*, Table S3 and Figs. S2 and S3).

Different bioinformatic approaches gave qualitatively similar downstream results in terms of comparing eDNA data with the LFDP tree census (*SI Appendix*, Table S4 and Figs. S12–S36), so we present results from the least stringent approach that retained the most samples and LFDP OTUs (all samples with >10,000 reference library reads and no further filtering). This dataset consisted of 38 "big grid" samples (as well as 7 unpooled subsamples from three "big grid" sampling points, and 12 "intensive plot" samples). There were 61 Amplicon Sequence Variants (ASVs) with 100% matches with unique LFDP reference library sequences [equating to 53 LFDP tree Operational Taxonomic Units (OTUs)], with nearly all plant reads (median 99%; range 46 to 100%) and the majority of plant ASVs (median 71%; range 50 to 100%) amplified from soil samples matching LFDP reference library sequences (*SI Appendix*, Fig. S4). Thus, 53 of 79 (67%) LFDP tree OTUs were detected in the soil eDNA samples, together comprising 98% of the total LFDP tree basal area (and 98% of total stems). On average, 11.8 (±4.9 SD, range 1 to 22) LFDP tree OTUs were identified in each 'big grid' sample (Fig. 2*A*).

We found considerable fine-scale heterogeneity in the occurrence of LFDP tree OTUs in soil eDNA. Comparing the unpooled subsamples with their corresponding 'big grid' (pooled) sample showed 1-4 OTUs unique to each subsample at each station, including a single OTU that was detected exclusively in a single unpooled subsample but undetected in any of the "big grid" samples (*SI Appendix*, Figs. S5–S7). Among the 12 "intensive plot" samples, we detected 22 LFDP tree OTUs with (mean ± SD) 12 ± 4.4 per sample. Depending on the bioinformatic approach, 26-39% of the OTUs detected in the "intensive plot" samples occurred in only ≤2 individual samples despite being separated by only 10 s of cm (*SI Appendix*, Fig. S8). Rarefaction of both the "big grid" and "intensive plot" samples indicated sufficient intrasample sequencing depth to capture LFDP reference library ASV diversity (*SI Appendix*, Fig. S9*A*). Hill numbers of reference library ASV incidence richness (q = 0) among "intensive plot" samples revealed that theoretically, >45 individual samples (depending on bioinformatic filtering threshold) would be required to sample the estimated asymptotic LFDP tree ASV richness in the 4 m^2 area (*SI Appendix*, Fig. S9*B*).

Comparison of eDNA and Tree Census Data. Taxonomic richness of the tree census data increased with area around the sample locations, with an average of 11.3 OTUs in a 5 m radius (± 3.1 SD, range 5 to 17) up to an average of 64.9 OTUs (± 2.2 SD, range 59 to 69) in a 100 m radius (Fig. 2*A*). Notably, the magnitude of taxonomic richness recorded in the eDNA samples (mean 11.8 ± 4.9 SD) corresponded closely to that of the tree census data at the 5 m scale (mean 11.3). Considering all 'big grid' samples together, rank abundance of eDNA reads per OTU was positively associated with stem rank abundance in the 2023 census data (Fig. 2*B*; Pearson correlation = 0.70, $P < 0.0001$). Additionally, there was a positive correlation between OTUs that were more abundant in terms of total basal area in the 2023 census and detection rate in soil eDNA (Fig. 2*C*; Pearson correlation = 0.65, $P < 0.0001$). Across all OTUs, detection in eDNA samples was negatively related to the distance from the sample to the nearest known individual of that OTU (Fig. 2*D*; $P < 0.0001$). Correlations of OTU richness were positive and marginally significant when based on the tree community in neighborhoods of 5 to 15 m and values approached zero at larger scales (Fig. 2*E*; note that different bioinformatic approaches gave somewhat different results, *SI Appendix*, Fig. S12).

Based on the confusion matrix analysis of the "big grid" samples, the median value of sensitivity (i.e., true positive rate) declined from 0.59 at the 5 m scale to 0.17 at the 100 m scale (Fig. 3*A*). In contrast, the median value of specificity (i.e., true negative rate) increased from 0.93 at the 5 m scale to 1.00 at the 100 m scale (Fig. 3*B*). The median value of the Matthews correlation coefficient (MCC) declined from 0.50 at the 5 m scale to 0.19 at the 100 m scale (Fig. 3*C*). Overall patterns of these metrics for the "intensive plot" samples followed the same trend albeit with higher and lower raw values of sensitivity and specificity, respectively. Taken together, these patterns suggest that the strongest correspondence between eDNA and stem data is at the smallest (i.e. 5 m) spatial scale and declines to essentially random predictions for larger (>50 m) scale neighborhoods. However, we found little support for this interpretation when considering the standardized effect size of the confusion matrix metrics. In particular, median values of sensitivity, specificity, and MCC fell within the range of random expectation across spatial scales, indicating that eDNA samples did not describe the corresponding local tree composition significantly better than they did for random locations throughout the LFDP (Fig. 3 *D*–*F*).

Further investigation of which OTUs were driving confusion matrix results revealed that, at 5 m scales, abundant OTUs (>10,000 individuals) were correctly detected as present in a majority of sites while very rare OTUs (<100 individuals) tended to be correctly identified as absent in most sites (Fig. 4 *A* and *B*). Irrespective of total abundance, OTUs had low rates of false detection in eDNA (i.e., detected when absent from the census data) except for a handful of taxa with intermediate abundance (Fig. 4*C*). Relatively abundant OTUs were recorded as falsely absent (i.e., undetected in eDNA when present in the census) in a higher proportion of sites than rare taxa (Fig. 4*D*).

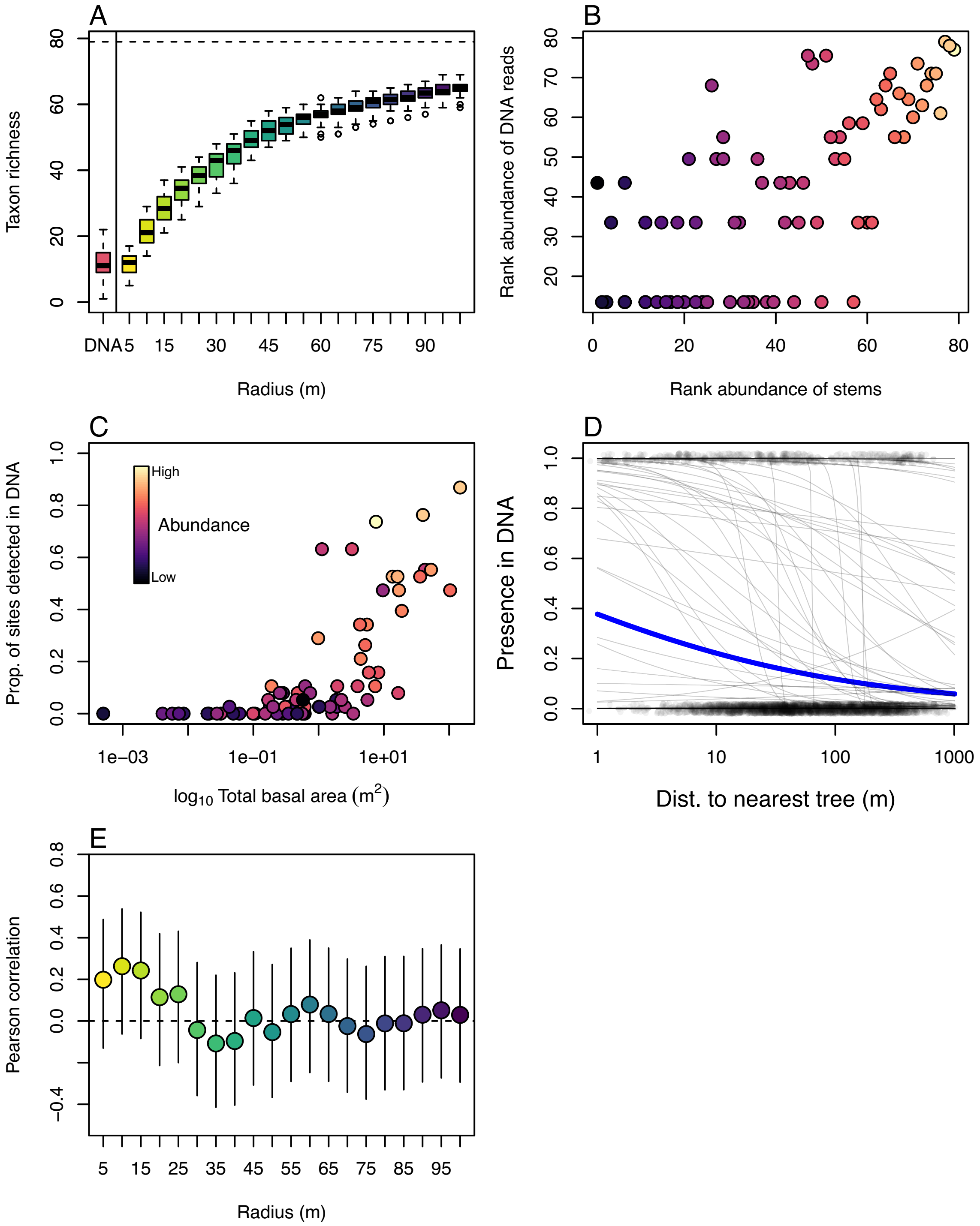

Fig. 2. (*A*) Boxplot of taxon richness of eDNA samples and LFDP census data based on increasing neighborhood radii. Box plot midlines show medians, box edges show first and third quartiles, whiskers show minima and maxima, and points are outliers. (*B*) Rank abundance of eDNA reads versus rank abundance of stems for 79 OTUs in the 2023 LFDP census data. Point color corresponds to total abundance in the 2023 LFDP census where darker colors are less abundant taxa. (*C*) The proportion of sample sites where a given OTU was detected as a function of the total basal area (m^2) in the 16-ha LFDP; point color is as in (*B*). (*D*) Presence–absence of OTUs in soil eDNA samples as a function of distance to the nearest known individual in the LFDP. Light gray curves show fitted logistic relationships for individual OTUs; the thicker blue line shows the fitted logistic relationship across all OTUs. (*E*) Pearson correlation (with 95% CI) between stem OTU and DNA OTU richness across spatial scales using data rarefied based on the number of reads for eDNA and number of individual trees for census data using the 'rarefy' function in the vegan R package (25).

Discussion

As eDNA becomes widely adopted for biodiversity assessments (1, 2, 6, 7), robust evaluations of the spatial scale at which eDNA samples reflect the spatial distribution of taxa in ecological communities are crucial. This is particularly important for any approach (including biodiversity monitoring and conservation applications) that relates species presence/abundance with covariates, as the scales of response and predictor variables must align for robust inference (26). In this study, we leveraged spatially explicit data from a large (16-ha) tropical forest plot to evaluate the ability of soil eDNA metabarcoding to capture patterns of tree diversity based on comprehensive and contemporaneous field sampling of woody stems ≥1 cm DBH. Our main findings include

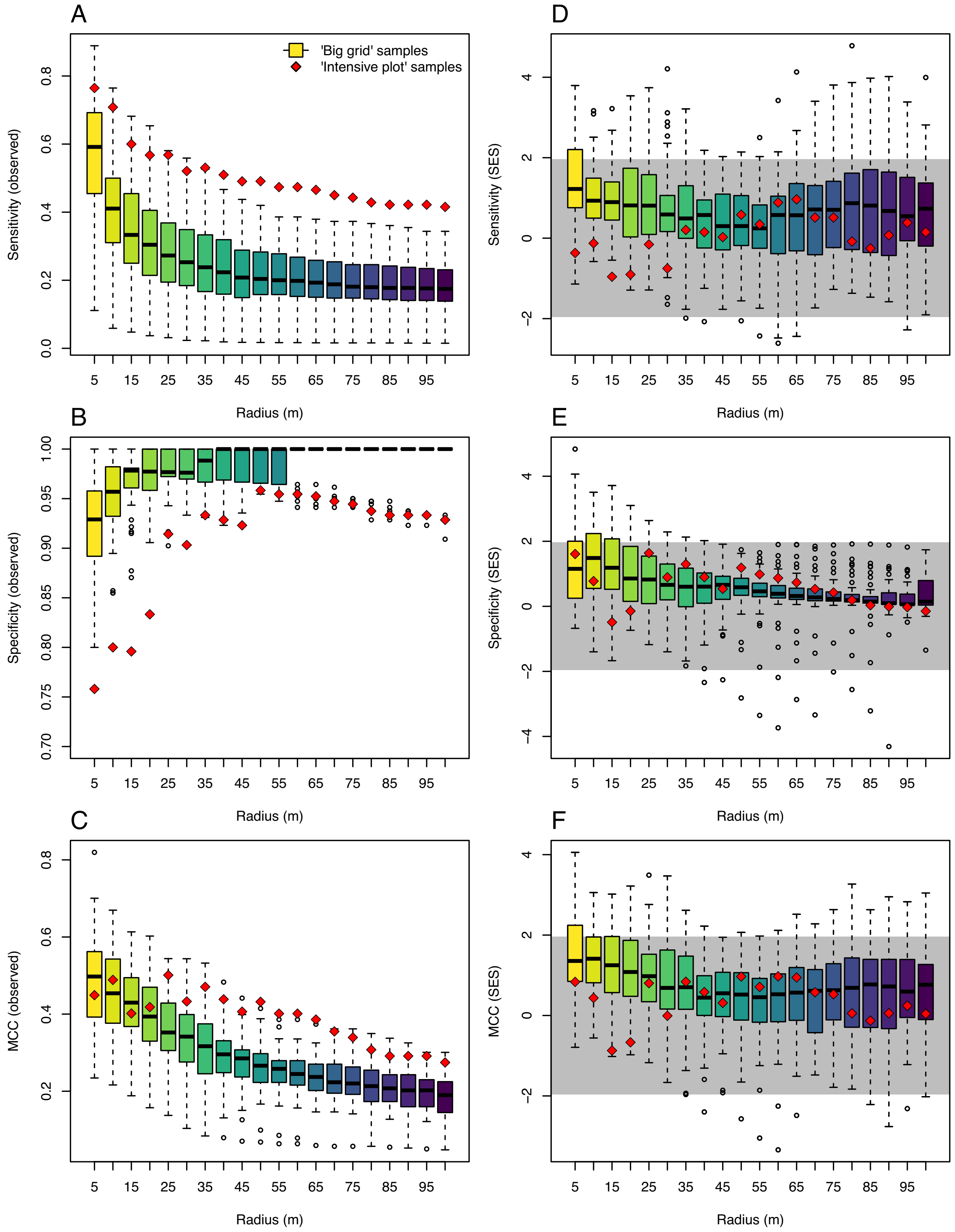

Fig. 3. Confusion matrix statistics comparing eDNA communities with tree communities across spatial scales. Boxes show values from 40 "big grid" sample points, red diamonds represent values from the "intensive plot" samples. Boxes represent interquartile range (IQR), dark lines indicate medians, whiskers extend to the smallest and largest values within 1.5 × IQR, and points denote outliers. (*A*) Sensitivity (true positive rate; the proportion of all OTUs present in the census data at a given spatial neighborhood that were detected in the eDNA data), (*B*) specificity (true negative rate; the proportion of all OTUs absent in the census data at a given spatial neighborhood that were also absent in the eDNA data), and (*C*) Matthews correlation coefficient (MCC; quantifies the overall classification performance considering true and false positives and negatives). (*D*–*F*) Standardized effect size of sensitivity, specificity, and MCC based on comparison of the observed values with values based on 1,000 random locations in the LFDP. Area outside of gray shading indicates values <−1.96 or >1.96, corresponding to $P < 0.05$ statistical significance

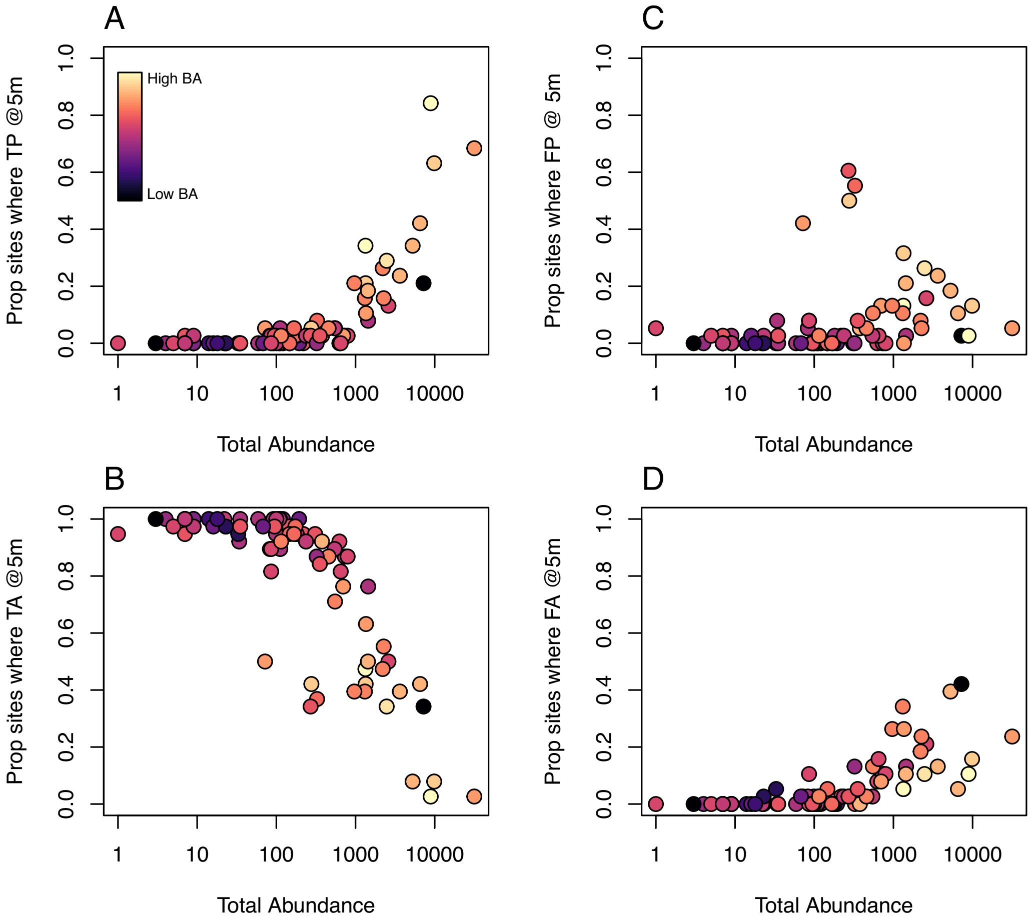

Fig. 4. (*A–D*) Confusion matrix components for 5 m radius at each big grid sampling point for each OTU as a function of total abundance in the 2023 LFDP census. TP/TA: True Presence/Absence; FP/FP: False Presence/Absence. Point colors are scaled to log total basal area (BA in m^2) in the LFDP.

1) mixed evidence for the ability to characterize tropical tree diversity and composition with eDNA and 2) weak evidence that soil eDNA metabarcoding data corresponds to a localized spatial signature of the tree community.

Correspondence of eDNA and Tree Census Data. The primary aim of our study was to determine the ability of soil eDNA to capture observed spatial patterns of tree diversity and composition. We note that our validation dataset (the LFDP tree census) is limited to woody stems ≥1 cm DBH, so for our analyses "true and false presence" are defined as when DNA was detected for a taxa present or absent, respectively, from the LFDP census data in any given area. These definitions are pragmatically necessary to compare datasets, but it is possible that some DNA detections defined as "false presence" reflect DNA originating from individuals below the census size threshold (e.g., seedlings). However, we consider it more likely that the origin of most DNA detected is from censused individuals rather than unsurveyed seedlings, given the orders of magnitude difference in biomass between large trees and seedlings, and that tree species are often spatially clumped so seedlings are typically found nearby conspecific trees that would be captured by the LFDP census data (27). Additionally, the LFDP census size threshold aligns with standardized ForestGEO sampling protocols (28), reflecting an ecological focus on individuals recruited into the community that contribute meaningfully to forest biomass and community dynamics. Thus, despite inherent limitations, our approach provides a practical and ecologically grounded framework for assessing the spatial correspondence between soil eDNA and tree communities.

Similar to some previous studies (15, 24), rare taxa tended to be missing from our eDNA data. In fact, the combined basal area of the 26 OTUs that were completely undetected in the eDNA samples was only ~2% of the total basal area (and ~2% total stems) recorded in the 2023 census, despite accounting for ~33% of the OTU richness. This is expected as, on average, OTUs with higher basal area should shed relatively more DNA into the environment. Fine-scale variation of biological diversity (i.e., variation within a single habitat type) can, however, be strongly affected by relatively rare taxa, whereas floristic distinctiveness among different habitats and regions is typically defined more by turnover of common taxa (29–31). Thus, the lack of detection of rare OTUs is problematic for characterizing some aspects of spatial variation of community composition, both in terms of taxonomic as well as functional diversity, given for example, the disproportionate contribution of rare taxa to functional distinctiveness (24, 29, 32, 33).

Overall, our results suggest that relatively rare taxa are unlikely to be reliably detected in soil metabarcoding studies that employ general (i.e. "universal") PCR primers without a very high level of sampling/sequencing effort, combined with a robust local reference library. Tropical forest metabarcoding-based biodiversity assessments that prioritize detection of relatively rare taxa should therefore either employ a more targeted molecular approach or incorporate a much higher sampling intensity—either using different approaches for larger soil volumes (e.g., (34)) or homogenizing many more individual subsamples/samples (e.g., 11)—than in our study. The level of sampling intensity required to address this issue remains unclear, although asymptotic diversity estimates from our "intensive plot" samples suggest an exceedingly large

(>45) number of samples if using typical soil eDNA sampling approaches [e.g., this study, (11)]. However, it is important to note that increased sampling density will also lead to more false presences (see below).

Current knowledge about the spatial signal of eDNA in diverse terrestrial ecosystems is extremely limited (11, 12, 34); our study explicitly evaluates the spatial resolution of soil eDNA to infer tropical forest standing tree community composition across a continuous range of spatial scales from 78 m^2 to 160,000 m^2. Despite detecting the most common taxa across the entire plot (in terms of abundance and basal area), several of our results suggest that it is unlikely soil eDNA can yield spatially explicit measures of individual taxa occurrence and diversity at fine spatial scales. This is somewhat surprising, as the majority of leaf fall—a major source of biomass from living trees (35)—falls within 10 m of individuals in the LFDP (36) although pollen and other cellular material can certainly be transported further. Despite this, we found that tree OTU composition in tropical forest soil eDNA is highly heterogeneous at fine spatial scales. The "intensive plot" samples detected many different OTUs in adjacent subsamples (i.e., separated by 10's of cm), and our "unpooled subsamples" had taxa that were not found in the associated (pooled) "big grid" sample, and vice versa. Thus, considerably more intensive sampling would be required to capture all tree OTUs in soil within a 4 m^2 plot. Yet the "intensive plot" confusion matrix analyses also demonstrated that more sampling did not improve fine-scale spatial signal because while discovering more taxa increased sensitivity, it also increased the discovery of false presences and thus reduced specificity.

The "big grid" sample confusion matrix analyses provided further insight to the performance of eDNA at different spatial scales. Median sensitivity (i.e., true positive rate) across sites dropped from ca. 0.59 to 0.17 with increasing neighborhood size, indicating that at the smallest (5 m) scale examined here, about half of the OTUs present in the census data were detected with eDNA, whereas at broader scales this proportion was much lower. The decline in sensitivity with spatial scale is intuitive because more taxa are truly present in larger census data neighborhoods by default, whereas the number of detections in the eDNA does not change. Similarly, specificity (i.e., true negative rate) increases with spatial scale because as neighborhood size increases, there are fewer true absences in the census data by default, and these absences are also likely to be undetected in the eDNA data. Specificity was already high at the smallest (5 m) spatial scale (median = 0.92), indicating that most census data OTUs absent at this scale (as trees ≥1 cm DBH) were indeed undetected in the corresponding eDNA samples. Additionally, this result also indicates few false presences (1-specificity, i.e., taxa recorded in the eDNA but absent from the tree census) in "big grid samples" at fine spatial scales, in relative contrast to the "intensive plot" results. Considering these metrics together, the MCC was highest at the small spatial scales and decreased with neighborhood size. We also note that alpha diversity at each sample location was congruent with true taxonomic richness at the smallest (5 m radius) spatial scale. A first interpretation of our results would therefore suggest that eDNA best captured tree diversity and composition at the smallest spatial scales.

However, the spatially explicit tree map of the LFDP enabled us to disentangle the raw confusion matrix results from random expectations by accounting for the spatial autocorrelation and plot-level abundance of OTUs in the LFDP. Standardized effect sizes of sensitivity, specificity, and MCC did not differ from random expectations when we compared taxa recorded in eDNA samples with the tree census data from random points throughout the LFDP. Notably, our simulation analyses demonstrated the potential to recover significant spatial signatures under certain conditions (*SI Appendix*, Figs. S10–S13). In the context of tropical forest tree monitoring, this implies that metabarcoding approaches (with a similar sampling intensity as ours, and perhaps even higher) are useful to characterize beta-diversity patterns and turnover of dominant taxa at landscape scales (e.g., tens of hectares and larger), but less so for monitoring rare taxa or for tasks that require characterizing taxonomic composition at finer scales. Given the aforementioned trade-offs between sensitivity and specificity with increased sampling effort, we believe this result is representative of metabarcoding sampling of soil eDNA pools, and thus robust, despite that greater sampling intensity, sequencing depth, and/or more loci could have been employed. Critically, the results of our randomization approach were robust across a broad range of bioinformatic filtering options (*SI Appendix*, Figs. S14–S37), which can influence the numbers of OTUs detected and samples retained, and consequently, derived composition and diversity metrics (37, 38).

Several processes related to overall abundance (or basal area) of taxa in the study site and laboratory considerations could contribute to variation of OTU detection among sites and, ultimately, the lack of spatial signature between eDNA and tree census data. First, we note that the absence of many rare OTUs from the eDNA data appears to have weak effects on confusion matrix metrics since rare OTUs are also absent from most sampling locations and so predominantly classified as true absences. As a corollary, misclassification (i.e. eDNA false presence and false absence) of intermediately and highly abundant taxa appears to drive confusion matrix metrics as the majority of OTUs have misclassification rates <20%, except for a handful of intermediately abundant (100 to 1,000 individuals) and highly abundant (1,000 to 10,000) OTUs. Second, PCR primer bias and mismatch are known to affect metabarcoding datasets, which may explain some misclassification of certain OTUs. While use of taxa-specific correction factors for read counts can account for PCR primer bias in read count data (39), these would not ameliorate false negatives (i.e. OTUs present but not detected) without increased sampling, sequencing depth, or the number of loci amplified. These approaches would, again, lead to increased false presence discovery, and consequently require further empirically derived thresholds to correct false presences in the already corrected read count data. Furthermore, it is unlikely such an approach is feasible without a priori knowledge on taxon-specific PCR primer biases and the spatial distribution and abundance of target taxa which, if available, may limit the novelty of information gained by eDNA studies. Thus, the complexities of interpreting eDNA data lie in both methodological uncertainties and uncertainties in inferring individual-level processes in populations and communities from molecules detected in environmental media (37).

Conclusions. Our study provides insight into the utility of eDNA as a biomonitoring tool for capturing tropical forest tree communities. If the objective of a study is to characterize the composition and diversity of dominant taxa across landscape scales, then our study suggests a sampling regime and laboratory approach similar to ours (~40 samples systematically collected) may provide a fairly robust picture of dominant (and some less abundant) taxa in the landscape. However, if the research objective requires reliable detection of rare taxa over landscape scales, then different molecular approaches and/or a much more intensive sampling regime should be adopted. We have demonstrated that typical soil eDNA metabarcoding is inaccurate for characterizing standing tree diversity or composition at smaller spatial scales (e.g., <1 ha) and particularly local scales (<100 m). We have also shown that increased sampling/sequencing effort leads to a trade-off between detections and false presences. Both of these latter results illustrate important boundaries for the use of eDNA for tropical forest biomonitoring, which has relevance not only for ecological research but also in the context of policy initiatives that involve biodiversity monitoring/reporting (e.g., the EU Corporate Sustainability Reporting Directive). Arguably, the default assumption regarding macro-organism detection via eDNA metabarcoding is that variation between samples is due to each sample representing the local community at the sampling point better than a random expectation. We have shown that this is not the case over 5-100 m distances in a diverse tropical forest. The lack of spatial resolution is counterintuitive but very important: understanding the relevant spatial scale to interpret eDNA results—i.e. linking molecular to actual presence—is critical for facilitating their effective use in ecological research and biodiversity monitoring. As the field advances, our approach contributes to clarifying which questions eDNA metabarcoding can be used to answer and how it can be integrated with other approaches to gain a richer understanding of biodiversity.

Methods

Study Area. The LFDP encompasses 16-ha (320 m × 500 m) of subtropical wet forest in northeastern Puerto Rico. The LFDP lies at 333-428 m a.s.l.; mean annual precipitation is ca. 3,500 mm, and the average monthly temperature ranges from 23.5 °C to 27 °C (38). In the most recent LFDP census (conducted 2021–2023), >102,000 stems belonging to 139 species were identified, tagged, and mapped (40).

Soil Sampling, eDNA Analysis, and Reference Library. Detailed field, laboratory, and bioinformatic methods are provided in the *SI Appendix*.

Soil sampling: Surface soil was sampled in a systematic grid of forty points across the LFDP, where three ≈10 to 30 g subsamples were haphazardly collected within a 2 × 2 m area around each of the 40 points ("big grid" samples; Fig. 1 *A* and *B*), combined, and homogenized. To supplement our analyses and examine fine-scale heterogeneity of plant soil eDNA, we also extracted DNA from 1) 0.5 g of each subsample before homogenization at 3 big grid locations ("unpooled" samples), and 2) 0.5 g of an intensively sampled (n = 12) 2 × 2 m area at one location ("intensive plot" samples) (Fig. 1*B*). Extraction of eDNA from soil followed a two-step approach based on Ariza et al. (11).

LFDP Soil metabarcoding & bioinformatics: We amplified a small fragment of the chloroplast trnL P6-loop from soil eDNA using the "universal" trnL-g (GGGCAATCCTGAGCCAA) & trnL-h (CCATTGAGTCTCTGCACCTATC) primer pair for vascular plants (23) with fusion primer adapters, facilitating a two-step PCR approach to Illumina library builds with quadruple indexing of amplicons (41). All PCRs used Amplitaq Gold 360 hotstart chemistry, and all PCR batches included appropriate experimental controls (42). Four uniquely indexed replicate PCRs were performed per sample, combined, and Illumina adaptors and indexes incorporated via the second PCR. Amplicons were combined for NOVOSEQ 6000 partial lane sequencing (via NOVOGENE) using PE150 chemistry, aiming for 25-50 K sequences per PCR replicate, per sample. Sequence data were demultiplexed to intrasample PCR replicates and adaptors/primers trimmed (43, 44), denoised with DADA2 (41), ASV tables curated with LULU (45) and soil sequences were mapped to LFDP reference library sequences at 100% match in DADA2, then finally a BLASTn search and the MEGAN lowest common ancestor algorithm (46) used to taxonomically annotate the remaining sequences. For comparison to LFDP census data, ASVs not matching reference library sequences were removed, and the "pseudopooled" DADA2 core inference algorithm was used for all analyses. Potential bioinformatic choice/filtering artifacts on results were examined by processing data using twelve combinations of bioinformatic iterations increasing in stringency that vary common quality control, ASV filtering and intersample normalization thresholds (minimum sequencing depth, ASV co-occurrence between PCR replicates, rarefaction, and removing low read count ASVs). Downstream results were all qualitatively similar and did not alter conclusions, so results from lowest stringency bioinformatic filtering (all samples with >10,000 reference library reads and no further filtering, "big grid" n = 38, "unpooled" subsamples n = 7, "intensive plot" n = 12) are presented here. See *SI Appendix* for results from all other bioinformatic iterations.

LFDP tree reference library and bioinformatics: Leaves were collected from 104 preidentified, tagged species (out of 139 total tree species recorded in the 2023 LFDP census) to build a local sequence reference library. DNA extracted from ≈8 mg of tissue using the DNEasy® Plant Mini Kit (Qiagen, Netherlands) and was PCR amplified with unique index combinations of the *trnL-g* & *trnL-h* fusion primers and Illumina sequence libraries prepared as above. This library was sequenced as above, aiming for 50,000 reads per species. Bioinformatic processing was similar to eDNA sample processing, except demultiplexing was to the individual sample, the "unpooled" DADA2 core inference algorithm was used to denoise, and the OTU table did not require curation. For quality control, BLASTn (47) indicated the taxonomic affinity of putative haplotype (the most abundant ASV). Reference library species with shared ASVs were pooled into OTUs. To reconcile species haplotype data with LFDP census data, we merged an additional 15 LFDP species for which we did not have sequence data based on taxonomy (see *Methods* and Dataset S1). In four cases, we used a conservative approach to pool unique ASVs because BLAST results indicated potential field errors or ambiguity given the taxonomic composition of the tree census data. We excluded 16 unsequenced LFDP species from the analysis because there was no sequence data for closely related (confamilial) species, and they were exceedingly rare in the census data. There were 79 final OTUs (Dataset S1). We pooled species from the LFDP census data into the same OTUs as the sequence reference library for further analyses.

Comparing Soil eDNA with Tree Census Data. To evaluate the spatial signal in the correspondence between eDNA and tree census data, we compared the taxonomic richness of eDNA communities from each sample with the tree census data based on increasing neighborhood sizes (from 5 to 100 m radius, in 5 m increments) (Fig. 1 *A* and *D*). As the focus was on richness and LFDP taxa presence/absence in predefined neighborhoods, metabarcoding data were converted to presence/absence. To gain further insight to the ability of eDNA to capture known taxonomic composition, we used a confusion matrix approach (22). For each sample location and each spatial scale, we used the caret R package (48) to compute the number of OTUs present/absent in the LFDP census data and present/absent in the eDNA data. Based on these values, we quantified sensitivity (true positive rate), specificity (true negative rate), and MCC (also known as the Yule phi coefficient). Sensitivity ranges from 0 to 1; higher values indicate that a higher proportion of OTUs present in the census data in a given neighborhood were also detected in the eDNA data (i.e., eDNA samples correctly capture OTUs that were present from the census data). Specificity also ranges from 0 to 1; higher values indicate that a higher proportion of OTUs absent from the census data in a given neighborhood were also undetected in the eDNA data (i.e., eDNA samples correctly exclude OTUs that were absent from the census data). MCC quantifies the overall classification performance by balancing true and false positives and negatives.

To understand how the observed values of these metrics compare with random expectations based on the abundance and spatial distribution of trees in the census, we also computed sensitivity, specificity, and MCC for the eDNA community at each sample site using the census data around 1,000 random locations in the LFDP as the reference (i.e., spatially decoupling the eDNA and census data). We computed the standardized effect size (SES) of each metric as the difference between the observed value and the median of the values based on random points, divided by the SD of the randomized values (49). Positive (or negative) values of SES sensitivity, specificity, and balance accuracy indicate higher (or lower) values compared to a random expectation after controlling for the overall abundance and spatial distribution of tree taxa in the study area; absolute values greater than 1.96 indicate statistically significant effects (alpha = 0.05). We used a series of simulations of artificial eDNA data to benchmark this approach, and we confirmed the possibility to recover a significant spatial signature of eDNA given certain conditions (*SI Appendix*). All analyses were conducted in R version 4.4.1 (50). Raw DNA sequence data for the reference library and the soil eDNA samples are posted to Zenodo (https://zenodo.org/records/15649282). Scripts and data to reproduce our analyses are posted to a Github repository (https://github.com/bobmuscarella/eDNA-LFDP), including instructions on how to incorporate the raw data posted on Zenodo.

Data, Materials, and Software Availability. 1. Tree census data; 2. Sequence data for reference library; 3. Soil metabarcoding data have been deposited in Zenodo (https://zenodo.org/records/15649282) (51). Scripts and data to reproduce our analyses are posted to a Github repository (https://github.com/bobmuscarella/eDNA-LFDP) (52), including instructions on how to incorporate the raw data posted on Zenodo.

ACKNOWLEDGMENTS. Research was funded by a grant from the Swedish Research Council, Formas (2020-00921 to R.M.), a grant from the Swedish Research Council, Vetenskapsrådet (2019-03758 to R.M.), and support from the Swedish Phytogeographical Society to F.B.S. and R.M. Logistical support for G.D. was provided by NTNU University Museum and support for developing eDNA data analysis methods by the Horizon Europe project DiverSea (Project 101082004). Tree census data from the LFDP were collected with support from NSF DEB 2028688 to M.U. and J.K.Z., and additional support was provided by the Luquillo LTER (US NSF grant DEB-1831952), the University of Puerto Rico-Río Piedras, and the USDA Forest Service International Institute of Tropical Forestry. The manuscript was improved by comments from two anonymous reviewers.

Author affiliations: [a]Department of Ecology and Genetics, Uppsala University, Uppsala 75236, Sweden; [b]Department of Natural History, Norwegian University of Science and Technology, Trondheim 7491, Norway; [c]Department of Ecology, Evolution and Environmental Biology, Columbia University, New York, NY 10027; and [d]Department of Environmental Sciences, University of Puerto Rico, Rio Piedras, PR 00936-8377

1. P. F. Thomsen, E. Willerslev, Environmental DNA – An emerging tool in conservation for monitoring past and present biodiversity. *Biol. Conserv.* **183**, 4–18 (2015).
2. K. Deiner *et al.*, Environmental DNA metabarcoding: Transforming how we survey animal and plant communities. *Mol. Ecol.* **26**, 5872–5895 (2017).
3. A. Burian *et al.*, Improving the reliability of eDNA data interpretation. *Mol. Ecol. Resour.* **21**, 1422–1433 (2021).
4. K. C. Beng, R. T. Corlett, Applications of environmental DNA (eDNA) in ecology and conservation: Opportunities, challenges and prospects. *Biodivers. Conserv.* **29**, 2089–2121 (2020).
5. J. Pawlowski, A. Bonin, F. Boyer, T. Cordier, P. Taberlet, Environmental DNA for biomonitoring. *Mol. Ecol.* **30**, 2931–2936 (2021).
6. L. Jiang, Y. Yang, Visualization of international environmental DNA research. *Curr. Sci.* **112**, 1659–1664 (2017).
7. M. Seymour, Rapid progression and future of environmental DNA research. *Commun. Biol.* **2**, 80 (2019).
8. B. K. Hansen, D. Bekkevold, L. W. Clausen, E. E. Nielsen, The sceptical optimist: Challenges and perspectives for the application of environmental DNA in marine fisheries. *Fish Fish (Oxf.)* **19**, 751–768 (2018).
9. M. D. Johnson *et al.*, Environmental DNA as an emerging tool in botanical research. *Am. J. Bot.* **110**, e16120 (2023).
10. F. Keck, M. Couton, F. Altermatt, Navigating the seven challenges of taxonomic reference databases in metabarcoding analyses. *Mol. Ecol. Resour.* **23**, 742–755 (2023).
11. M. Ariza *et al.*, Plant biodiversity assessment through soil edna reflects temporal and local diversity. *Methods Ecol. Evol.* **14**, 415–430 (2022), 10.1111/2041-210x.13865.
12. I. Hiiesalu *et al.*, Plant species richness belowground: Higher richness and new patterns revealed by next-generation sequencing. *Mol. Ecol.* **21**, 2004–2016 (2012).
13. E. Duley, A. Iribar, C. Bisson, J. Chave, J. Donald, Soil environmental DNA metabarcoding can quantify local plant diversity for biomonitoring across varied environments. *Restor. Ecol.* **31**, e13831 (2022).
14. N. G. Yoccoz *et al.*, DNA from soil mirrors plant taxonomic and growth form diversity. *Mol. Ecol.* **21**, 3647–3655 (2012).
15. M. E. Edwards *et al.*, Metabarcoding of modern soil DNA gives a highly local vegetation signal in Svalbard tundra. *Holocene* **28**, 2006–2016 (2018).
16. R. Nathan, Long-distance dispersal of plants. *Science* **313**, 786–788 (2006).
17. M. Vasar *et al.*, Metabarcoding of soil environmental DNA to estimate plant diversity globally. *Front. Plant Sci.* **14**, 1106617 (2023).
18. A. J. Drummond *et al.*, Evaluating a multigene environmental DNA approach for biodiversity assessment. *Gigascience* **4**, 46 (2015).
19. S. Shirazi, R. S. Meyer, B. Shapiro, Revisiting the effect of PCR replication and sequencing depth on biodiversity metrics in environmental DNA metabarcoding. *Ecol. Evol.* **11**, 15766–15779 (2021).
20. V. López, A. Fernández, S. García, V. Palade, F. Herrera, An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. *Inf. Sci.* **250**, 113–141 (2013).
21. J. S. Hleap, J. E. Littlefair, D. Steinke, P. D. N. Hebert, M. E. Cristescu, Assessment of current taxonomic assignment strategies for metabarcoding eukaryotes. *Mol. Ecol. Resour.* **21**, 2190–2203 (2021).
22. C. M. Nugent, S. J. Adamowicz, Alignment-free classification of COI DNA barcode data with the Python package Alfie. *Metabarcoding Metagenomics* **4**, e55815 (2020).
23. P. Taberlet *et al.*, Power and limitations of the chloroplast *trnL* (UAA) intron for plant DNA barcoding. *Nucleic Acids Res.* **35**, e14 (2007).
24. C. J. Barnes *et al.*, Metabarcoding of soil environmental DNA replicates plant community variation but not specificity. *Environ. DNA* **4**, 732–746 (2022).
25. J. Oksanen *et al.*, vegan: Community Ecology Package (Version 2.6-10, R package, 2025). http://cran.r-project.org/package=vegan. Accessed 29 January 2025.
26. S. A. Levin, The problem of pattern and scale in ecology. *Ecology* **73**, 1943–1967 (1992).
27. R. Condit *et al.*, Spatial patterns in the distribution of tropical tree species. *Science* **288**, 1414–1418 (2000).
28. S. J. Davies *et al.*, Forestgeo: Understanding forest diversity and dynamics through a global observatory network. *Biol. Conserv.* **253**, 108907 (2021).
29. R. P. Leitão *et al.*, Rare species contribute disproportionately to the functional structure of species assemblages. *Proc. Biol. Sci.* **283**, 20160084 (2016).
30. F. C. Draper *et al.*, Dominant tree species drive beta diversity patterns in western Amazonia. *Ecology* **100**, e02636 (2019).
31. H. Morlon *et al.*, A general framework for the distance-decay of similarity in ecological communities. *Ecol. Lett.* **11**, 904–917 (2008).
32. D. Mouillot *et al.*, Rare species support vulnerable functions in high-diversity ecosystems. *PLoS Biol.* **11**, e1001569 (2013).
33. R. Tang, S. Li, X. Lang, X. Huang, J. Su, Rare species contribute greater to ecosystem multifunctionality in a subtropical forest than common species due to their functional diversity. *For. Ecol. Manage.* **538**, 120981 (2023).
34. N. Rota *et al.*, Evaluation of soil biodiversity in alpine habitats through eDNA metabarcoding and relationships with environmental features. *For. Trees Livelihoods* **11**, 738 (2020).
35. W. Lonsdale, Predicting the amount of litterfall in forests of the world. *Ann. Bot.* **61**, 319–324 (1988).
36. M. Uriarte, B. L. Turner, J. Thompson, J. K. Zimmerman, Linking spatial patterns of leaf litterfall and soil nutrients in a tropical forest: A neighborhood approach. *Ecol. Appl.* **25**, 2022–2034 (2015).
37. N. Rodríguez-Ezpeleta *et al.*, Biodiversity monitoring using environmental DNA. *Mol. Ecol. Resour.* **21**, 1405–1409 (2021).
38. A. Ramírez, Meteorological data from El Verde Field Station: NADP Tower. *Environ. Data Initiative* (2022), 10.6073/PASTA/2193C7DD5F79A28BE7A0D467646D63AD. Deposited 26 January.
39. A. O. Shelton *et al.*, Toward quantitative metabarcoding. *Ecology* **104**, e3906 (2023).
40. J. K. Zimmerman, Census of species, diameter and location at the Luquillo Forest Dynamics Plot (LFDP). *Puerto Rico. EDI.* (2023), 10.6073/pasta/7d937e27dfd99308362049d6c4495deb. Deposited 14 November.
41. B. J. Callahan *et al.*, DADA2: High-resolution sample inference from Illumina amplicon data. *Nat. Methods* **13**, 581–583 (2016).
42. L. Zinger *et al.*, DNA metabarcoding-Need for robust experimental designs to draw sound ecological conclusions. *Mol. Ecol.* **28**, 1857–1862 (2019).
43. M. Martin, Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.J* **17**, 10–12 (2011).
44. Sabre. https://github.com/najoshi/sabre.sabre [Accessed 1 November 2023].
45. T. G. Frøslev *et al.*, Algorithm for post-clustering curation of DNA amplicon data yields reliable biodiversity estimates. *Nat. Commun.* **8**, 1188 (2017).
46. D. H. Huson, A. F. Auch, J. Qi, S. C. Schuster, MEGAN analysis of metagenomic data. *Genome Res.* **17**, 377–386 (2007).
47. National Center for Biotechnology Information. https://www.ncbi.nlm.nih.gov/ [Accessed 26 April 2024].
48. M. Kuhn, Building predictive models in R using the caret package. *J. Stat. Softw.* **48**, 1–26 (2008), https://www.jstatsoft.org/index.php/jss/article/view/v028i05.
49. J. Cohen, *Statistical Power Analysis for the Behavioral Sciences* (Routledge, ed. 2, 2013).
50. R Development Core Team, *R: A language and environment for statistical computing* (R Foundation for Statistical Computing, 2021).
51. G. Dunshea *et al.*, "Soil Environmental DNA metabarcoding data from the Manuscript: Soil eDNA reflects regionally dominant species rather than local composition of tropical tree communities." Zenodo. https://doi.org/10.5281/ZENODO.15649282. Deposited 12 June 2025.
52. R. Muscarella *et al.*, eDNA-LFDP. GitHub. https://github.com/bobmuscarella/eDNA-LFDP. Deposited 12 June 2025.