Within-Community-Sampling Power Analysis to Detect Richness Change

Eden Tekwa[1], Jake Lawlor[1], Matt Lemay[2], Ryan R.E. Stanley[3], Nicholas W. Jeffery[3], Emily Rubidge[4], Jennifer Sunday[1*]

Affiliations:

1. Department of Biology, McGill University, 1205 Dr. Penfield, Montreal, PQ,Canada.
2. Hakai Institute, street address, Hariott Bay, British Columbia
3. Fisheries and Oceans Canada, Bedford Institute of Oceanography, 1 Challenger Dr., Dartmouth, NS, Canada
4. Fisheries and Oceans Canada, Institute of Ocean Sciences, 9860 West Saanich Rd., Sidney, BC, Canada

Corresponding Author: Jennifer.Sunday@mcgill.ca

Abstract

Reliable biodiversity monitoring requires understanding how sampling effort influences detectability of meaningful changes in species richness. Increased sampling within independent units has been shown to reduce measurement error, while sampled richness estimates are often subject to bias. However, robust methods for quantifying the relationship between sampling effort and the power to detect biodiversity change remain limited. Here we present a simulation-based power analysis framework that explicitly links sampling effort within communities to the ability to detect richness change between them. The approach uses empirical pilot data to simulate sampling and applies a dimensionless coverage metric to translate simulated effort into real sample sizes within communities. This produces direct relationships among power, effect size, and coverage, enabling estimation of the sampling required to achieve a specified probability of detecting richness change in the correct direction. We also provide a variant designed for before–after monitoring scenarios in which pilot data are available from a single community and the post-impact state is unknown. We demonstrate the application of this framework using a field-based environmental DNA biodiversity dataset. Additionally, we show that sample size recommendations rapidly converge when we apply the initial recommendation and reanalyze the next survey. The method is implemented in the R package *BioDivPoweR*, enabling broad application for designing efficient biodiversity monitoring programs and evaluating emerging sampling technologies.

Keywords: biodiversity, species richness, monitoring, sampling effort, quality control, environmental DNA

Introduction

Biodiversity underpins the functioning of ecosystems [1] that ultimately support and sustain human well-being [2]. As such, monitoring biodiversity change is essential for maintaining ecosystem functions and the well-being they provide. However, reliably detecting and attributing causes to these changes have been one of the most difficult technical problems in biodiversity science [3–5]. Enhancing the effectiveness of biodiversity monitoring programs will be vital for achieving international goals and targets[6], such as the commitment among nations to effectively protect 30% of areas under the Global Biodiversity Framework [7]. At more local scales, the ability to reliably detect changes in species richness or abundance-weighted diversity indicators is fundamental to evaluate ecological conditions, inform management decisions, and consider attaching financial incentives to biodiversity outcomes [8].

Despite the importance of biodiversity change assessments, we lack tools for understanding how much monitoring effort is required to robustly assess policy-relevant biodiversity changes within a region of interest. In a recent meta-analysis of species richness change associated with implementation of Marine Protected Areas (MPAs), about one half of studies showed no increase in diversity compared to reference sites [9] (from Figure S3 therein). Not observing an increase could have two very different explanations: (1) there could be no biologically-meaningful biodiversity differences, or (2) studies could lack the sampling effort needed to detect differences that were present. An estimate of power to detect a known global effect with increasing independent replicates advised caution in interpreting stable local richness as evidence that biodiversity loss is not occurring [3]. When we lack independent replicates, such as in a monitoring program at a local scale, we may still be interested in evaluating our power to detect an effect with different within-community sampling designs. Yet no analysis exists to our knowledge on how to advise local managers on how much sampling effort is needed to robustly detect change of a given effect size. Here we review why this is the case and propose a user-friendly method to achieve this.

What is a power analysis?

Statistical power is the rate of positive identification of an effect (such as differences in body size between two populations) when in fact there is an effect (i.e., the true positive rate, or 1-false negative rate). In prospective power analysis [10], we ask how many independent replicates are needed to detect a difference between 2 or more normally-distributed random variables, with an acceptable power level (1-B, often set to 80%), a prespecified significance level α (usually 0.05), and the population effect

size, to be detected with probability 1-B [10]. For example, statistical power for detecting a difference in body size between the two treatment groups can be estimated as a function of the true difference in body sizes (effect), user-set α (probably of type I error), and number of independent treatments (sample size)[11]. In ecology, examples of detection objectives amenable to standard power analyses include detecting differences or trends in ecological values when effect means are estimated using multiple independent samples [12].

Why is power in biodiversity change difficult to assess?

Biodiversity monitoring poses a special challenge to assessing statistical power for two reasons: biased measurement error, and a tendency to not replicate independent samples. Natural communities are typically composed of species that cannot be found in every location because of spatial heterogeneity and rarity (reviewed in [4]). As such, biodiversity is typically assessed by drawing 'samples' within a unit of study, and as more samples are drawn, the total number of species encountered is pooled (note, they are not averaged), bringing the measurement increasingly close to a maximum and correct (or *census*) species richness. Sampled richness is therefore not typically a mean across independent samples, but a cumulative metric of a single independent unit (i.e. a place or single community), and is necessarily downward-biased relative to true richness (Fig. 1b), at a scale that diminishes with increased observation effort (such as area, volume, or numbers of samples). In addition, in richness estimates, each individual's species identity is only used toward a community's richness if it belongs to a previously unrecorded species. In other words, the weight of the individual's datapoint is dependent on other individuals. When comparing two or more communities, a central challenge is determining how many samples within each community are needed to robustly detect a difference between them in the correct direction, but because samples contribute to a tally and not a mean, this question is not addressed with a traditional power analysis.

Sometimes in biodiversity studies, independent treatments are replicated (e.g., multiple communities representing each treatment and controls) [13], and we may then apply a traditional power analysis to assess differences between mean richness of two groups. However, even in these cases, measurement error within each replicate can affect outcomes. For example, a survey may repeatedly observe 5 species in community type A and 3 species in community type B. However, if the survey always misses 3 rare species in community type A and 10 rare species in community type B (sampling bias, [4]), the survey misidentifies community type A as richer even though the species estimates appear to be precise. Spatial heterogeneity and rarity of species occurrences calls for attention to the role of sampling effort to assure that values from each replicate are meaningful to test the effect at the scale of interest [14], yet the role of measurement error has been little discussed in ecological power analysis [12].

Furthermore, in biodiversity monitoring applications, we are often only interested in detecting temporal change between two communities without replication, such as comparing before vs. after an environmental impact, or an affected vs. reference area [15]. In these cases, it is even more important to consider the role of observation effort on the rate of detecting true differences between biodiversity metrics. To focus efforts on biologically meaningful change, we may consider a threshold of biodiversity change that calls for management action, such as a 10% difference [12,16]. In either case, the question of how many samples are needed in each community to detect a biodiversity difference of a set size with a desired rate of correct detections (power) is a question of quality assurance against measurement error, i.e. we are discussing the quality of information contributing to essentially two data points (richness) that we then rank.

Biodiversity measurement quality assurance

We propose that the question of how to draw samples from communities to assess species richness is analogous to how to draw blood from patients in a standardized way - a quality assurance step. The objective of quality assurance in this case is to obtain sufficient within-subject sampling to increase the probability of detecting a change in the correct direction. For the case of drawing blood in a single human subject (analogous to a community), we can test how the needle, volume of blood drawn, and blood processing affect test results in controlled experiments where the conditions and expected outcomes can be precisely defined. This can be formally justified by a power analysis on the number of blood particles one needs to draw in order to compare two blood samples. Note, in this case the power analysis is framed on two subjects, not on two treatment groups with multiple replicates. It still makes sense to talk about power for two data points if the binary empirical decision on the observed state (person one has higher iron concentration than person two) is simply whether the point readings are higher (without a t-test). It is the same concept when we envision a power analysis for how to sample two communities to detect a biodiversity difference; we are still able to ask about power (e.g., probability of detecting a difference in the correct direction) across different sampling qualities, even without replication needed for a t-test across groups. The quality assurance power analysis is more difficult for biodiversity science, because we generally do not know, or have control over, the true states of ecological communities.

Both kinds of power analysis are needed in biodiversity science

A quality assurance power analysis can be useful when we only have two communities, or even a single community of interest in which we aim to detect change over time. A quality assurance power analysis can also be useful to improve inference when we have true replication of communities, by assessing how well each replicate captures true richness. In the latter case, analyses can additionally leverage the broader

suite of traditional statistical tools, including hypothesis testing and power analyses. For simplicity, we will use “power analysis” at times in this text to mean quality assurance power analysis for detecting differences in two communities, not two groups of communities. “Sample size” will be used to refer to the number of samples drawn from a single community, rather than the number of communities sampled.

Proposed Solution

We develop a method for conducting a quality-assurance power analysis by using simulation to evaluate the effects of progressively reduced sampling efforts starting from a pilot biodiversity dataset. In most biological communities, no practical method approaches a true census. Therefore, we use simulated communities where true biodiversity states are assumed and then subsampled, starting with an artificial census.

The approach we propose is to computationally resample empirical community composition datasets. While down-sampling these datasets to evaluate the relationship between sampling effort and power is straightforward and has been done in previous studies (e.g. [17]), the key challenge lies in translating the resulting power-effort relationships to the real system, which exists at a broader spatial scale. Specifically, we need to convert the abscissa variable (unit of effort) from the number of samples in each community, which is a unit-specific measure, to a dimensionless value that represents a percentage of the total sampling effort. This can be achieved using empirical relationships between effort and percent coverage [18,19]. Using coverage as a universal measure of sampling effort allows the simulated relationships between effect size, power and sampling effort from pilot data can be rescaled to the real communities, including effort levels beyond the pilot dataset, through the principle of scale-invariance (analogous to modeling turbulence in fluid mechanics [20]). We explore this framework to make full use of natural and detection-based variation in species occurrences across pilot samples, in a manner analogous to how traditional power analyses use pilot data to estimate sample variance around a mean.

We focus on two simple but common scenarios encountered by biodiversity practitioners. The first is estimating differences between two “treatments” (e.g., affected vs. control), where pilot data from two communities are used to estimate power-effort-effect size relationships. The second is a biodiversity monitoring scenario where the goal is to detect compositional changes across time between present and future states. In the latter case, a range of possible future states and effect sizes are inferred based on the present state and assuming future states to be within the range of random reassembly of the observed community. To demonstrate the application of the power analysis, we use pilot eDNA metabarcoding data on marine biodiversity collected from nearshore coastal regions of the northeast Pacific Ocean. We also provide and describe

the implementation of the analysis in an openly accessible R package whereby users can enter their own pilot data for power analysis in any system following our method.

Methods

Overview of Quality Assurance Power Analysis for Species Richness

Our methods are organized into four subsections. First, we derive a relationship between coverage and the number of samples so that coverage becomes the universal measure of sampling effort. Second, we develop the computational method that uses a pilot study of species occupancies from two pre-defined communities (a two-community comparison scenario) to link effect size, coverage, and power. We repeat this for the scenario when we have pilot samples from only one defined community (scenario for a present-future comparison). Power in these cases was defined as the probability of detecting a difference in the same direction as the true simulated difference, so the statistic is the sign (+/-) of the difference in richness. Third, we develop analyses for the same two scenarios, but this time assess the probability of detecting a difference that exceeds a predefined minimum effect size (akin to *minimum effect power analysis* [12,16]). Finally, we demonstrate the application with an eDNA metabarcoding dataset, and run all four power analyses to assess how estimates of sample coverage needed to monitor for changes between two habitats at a given level of power is predicted by pilot study sample size.

All methods are implemented in the R-package *BioDivPoweR* [21], enabling general application to user-supplied datasets, allowing users to evaluate their own pilot data.

Developing a relationship between sample size and sample coverage

To design a quality-assurance power analysis for detecting biodiversity change, we first transform sampling effort to be comparable in both communities and transferable from pilot studies to the hypothetical real study. If sampling coverage of communities being compared is not the same, the comparison of richness (including direction of richness differences between them) may be biased and not directly comparable [22].

Here, we derive a relationship between coverage and sample number for each community, where a sample is a population of individuals from any species collected as a single unit of effort (e.g., one eDNA sample filter from the ocean or one quadrat from a terrestrial survey). Coverage (Eq. 1) is defined as the portion of species captured (*C*, [18]), and a sample is an independent population draw (*k*) within each community. *S* is the observed number of species and *S'* is an asymptotic richness estimator. If the same percent of species is captured in two communities (i.e., coverage is equal), we can determine which community has higher diversity.

(1) $C(k)=S(k)/S'$

We can use an asymptotic estimator of S' to estimate the relationship between coverage C and species richness observed S. Here we use the Ω asymptotic estimator that bases estimates on the entire observed pattern of species occupancies [19], that is more accurate than the *Chao2* estimator of S' based on the number of species occupying 1 or 2 samples [19].

Let's define $C_i(k)$ as the observation probability of species i after k samples. We first make the assumption that each sample is a random independent draw of community composition. $C(k)$ can be understood as the average observation probability across species after k samples. For each species i, C_i is a function of P_i (observed occupancy, or portion of samples where species i is observed). P_i is an observation, whereas C_i is the estimated occupancy we would observe if k were to increase beyond the original study's sample size. The across-species average probability C ranges from 0 to 1 and is an increasing function of sample size k. $< >_i$ is the mathematical operator for averaging over i. As sample size increases, the average probability approaches one.

(2) $C(k) = < C_i(k) >_i = < 1 - (1 - P_i)^k >_i$

In general, C in Eq. 2 is the average of a series of exponential functions belonging to i, where the exponent is i and the base is P_i. C approaches one as k approaches infinity. Unlike non-mechanistic statistical extrapolation on the rarefaction curve that does not asymptote, Eq. 2 implies that the number of species captured (CS' where S' is the true number of species - a constant) necessarily approaches a finite number as k approaches infinity. The inverse of Eq. 2 allows us to rescale any function of coverage (C) below 100% to the actual number of samples (k) by solving k given C. Generally, every marginal increase in coverage (C) corresponds to an increasing increment in sample size (k), meaning that a large increase in samples is required to gain an increase in coverage when the coverage is already high.

Coverage (C) serves us to rarify to equal coverage when we directly compare communities, and also to provide a dimensionless measure of sampling effort. We therefore apply simulated relationships between power and coverage to the whole community, by assuming that the pattern of occupancies of unobserved species is the same as that of observed species. Since coverage is unitless, applying simulated *in silico* relationships between power and coverage to real ecosystems is dimensionally consistent.

Inferring power from data in two communities (Scenario 1)

Our first power analysis scenario assesses relationships between power, sample size, and effect size from a pilot dataset in which multiple samples are taken from two

‘treatment’ communities we wish to compare. The two treatments could represent different habitat types, sites experiencing different environmental stressors or regulatory regimes, two time points from the same community, etc. Importantly, replicate samples in this design are not independent replicates of the treatment but instead multiple draws from two community types (i.e., two treatments, no replication) that we wish to compare. Random samples from within each community provide estimates of occupancy across samples that inform power. These are represented as multiple samples in red and blue communities in Fig. 1b and Fig. 2.

From the dataset (first row of boxes in Fig. 2), we first use Eq. 2 to compute the relationship between sample size and coverage for each community. We can also determine rarified richness differences from the pilot communities for later comparison.

Estimating power as a function of sample coverage using simulation

From each community, we create multiple bootstraps (resample with replacement) through a Monte Carlo procedure, and consider these as an array of simulated communities (second row of boxes in Fig. 1). We randomly pair bootstraps of each of two communities, and for each pair calculate the rarefied sample sizes using Eq. 2. These numbers will be different for each bootstrap pair and different from the rarefaction in the original communities.

For each bootstrapped pair of communities, individuals are subsampled (without replacement) according to the number of samples that correspond to coverage values ranging from zero to the maximum rarefied coverage (unique for each pair). We then make the assumption that the maximum rarefied coverage is 100% for the subsequent steps. For example, the rarefied coverage of 55% achieved by one bootstrapped pair is relabelled as 100%, and all other rarefied coverage values below 55% are multiplied by 100/55. The richness difference is measured (in terms of $\log_2$ of the $\text{richness}_2/\text{richness}_1$ ratio, or fold difference for short) at each subsampled step. Using the assumption that the observed difference at maximum coverage is the true difference (i.e., not from samples but from a census, third row of boxes in Fig. 1), the coverage is rescaled to pool results from different bootstrap pairs. We record these richness difference values for the next step.

Compiling results

The bootstrap pairs of communities are next binned according to their richness differences at maximum coverage as in a histogram (Fig. 2a), grouping bootstraps with assumed-to-be similar ‘true’ effect sizes, and the continuous range of bins that contain a minimum number of replicates desired for power analyses are extracted from each (e.g., 40; coloured portion of bars in Fig. 2a). This range of effect size is proportional to the heterogeneity in species compositions and represents a likely range of expected richness for communities similar to the original data source. Smaller bin sizes increase

the effect size resolution of the final power analysis, while holding computation load constant, thereby decreasing the range of effect sizes with sufficient replicates.

Then, for each effect size bin, we compare the measured richness difference at different coverage values including full coverage with the again assumed to be 'true' richness difference. The proportion of simulated differences in the correct direction is taken as the statistical power or the probability of detecting the correct direction of differences. The results taken together provide a 3-dimensional relationship between coverage, effect size, and power to detect change in the correct direction (Fig. 1 bottom plots, example in Fig. 2b).

We then rescale sample coverage to sample size from each community using our dimensionless scalar (Eq. 2) that relates sample size to coverage for each community. This allows for a specific relationship between power, effect size, and sample size (Fig. 1 bottom plots, upper x-axis).

One particularly useful view of this relationship comes if we hold power constant (e.g. typically 0.8 or 80% correct detections), and express the relationship between coverage (or sample size in each community) and the minimum detectable effect size (or amount of biodiversity change that is detectable in the correct direction at a given sample size; bottom right of Fig. 1). This relationship provides potentially critical information for planning monitoring efforts: *How much sampling is needed to detect a particular effect size?* For every possible level of coverage, we can extract the minimum effect size that achieves a given power (e.g., 0.8), and fit a linear regression to rank of coverage to describe the general relationship between coverage and log2 of effect size. We transformed coverage to exaggerate small increments at higher coverage values, according to the non-linear sequence of coverage values: ([250:250:10000].^0.5)/100. This sequence tends to generate a linear relationship between detectable effect size and coverage increments.

Inferring power from data in one community (Scenario 2)

When empirical effect sizes between treatments are unavailable, the relationships between power, coverage, and effect size may still be of interest. If no treatment is defined in the initial dataset (i.e., sample data are from a single community in one time point) we may still be interested in detecting changes over time. For this, we can use a similar procedure described in the previous section, except that samples are drawn from the original single dataset into two simulated communities leading to a richness difference between them. The logic of this application is illustrated in Figure S1. Here, the simulated community pairs will on average have an effect size of zero. However, the variation generated from the between-sample compositional heterogeneities can be understood as an estimate of how biodiversity could change. This approach assumes that the community is a closed system within a particular time frame of interest for tracking environmental changes, and that species retain their

occupancy distributions. Here, biodiversity changes must be selected from existing variations in the original community, which is compatible with short term stability in species occupancy patterns [23,24]. Therefore, our proposal to feed the same community data to two simulated communities is not only constrained by data limitation but has plausible biological justification.

Detecting effect sizes above a conservation-relevant minimum effect

Because large changes in species richness could be important for triggering conservation action, we also applied a "minimum effect" power analysis to estimate how many samples are needed to detect a change greater than a predefined benchmark to trigger conservation intervention [12,16]. We used a threshold of 0.1 log2 fold change in richness as our action-oriented benchmark (~7% increase in richness). Note, this is not the same as asking how many samples are needed to detect a difference greater than 0 and in the correct direction if they were truly different by 0.1 log2 fold change (that would be attainable in our above analysis), instead, this shows how many samples would be needed to detect a difference greater than 0.1 log2 fold change in the correct direction if they were truly different by at least that amount. We applied this to both scenarios above, and expected a greater number of samples to be required for this more stringent detection.

Applying our power analysis to biodiversity data derived from eDNA

We apply our simulation under both scenarios to a dataset of marine community assemblages, sampled using environmental DNA (eDNA) metabarcoding in the central coast of British Columbia, Canada (methods and data are described in [25]). Briefly, triplicate 1L-seawater samples were collected at 208 locations across a temperate marine region roughly 100km^2 in scale, focussed on nearshore habitats of kelp, seagrass, and rocky reef, and nearby deep-water sites. Samples were collected at depth using a Niskin bottle and filtered using 0.22uM Sterivex filters. DNA metabarcoding was implemented to target fish communities by amplifying a short region of the 12S rRNA gene (MiFish-U; [26]). The resulting libraries were sequenced using Illumina MiSeq V3 chemistry. Raw sequences were clustered into ASVs using the DADA2 pipeline [27] and annotated using BLAST searches of the complete NCBI nucleotide database (complete methods are available with the data package [25]). Vegetation cover of each sample was estimated from video footage for every sample and scored; we used a coarse category of high vs. low vegetation cover for our two-treatment analysis, and used all of the data as pilot for our one-treatment analysis. We refer to this dataset as the *Coastal Highly Density eDNA* dataset. Although this dataset has three replicated samples within each of 208 sites, for the sake of our simulation we pooled samples within sites, and treated every individual site as an independent "sample" of the whole community.

Sensitivity Analyses

We assessed, through simulation, the sensitivity of our power analysis to the size of the initial pilot dataset and explored the possibility of convergence if the recommendations of the power analysis were iteratively used. We assessed sensitivity to pilot size by randomly drawing (without replacement) smaller subsets of the full dataset (208 samples, from 5% to 100% of the total) and using it to estimate number of samples needed to detect a 0.1 fold change at 0.8 power in the 1-treatment power analysis. We assessed 'convergence' by drawing, for each initial pilot data simulation, a new dataset that matched the newly recommended sample size, and using that for a subsequent power analysis, to estimate the number of samples needed to detect a 0.1 fold change at 0.8 power. This simulates 'adaptive monitoring', or iteratively using new samples to run a new power analysis, assuming stationarity in mean species occupancy across samples and richness. We repeated the adaptive process 10 times to visualize trajectories and possible convergence on a stable recommendation.

Results

Application

Our *Coastal Highly Density eDNA* fish dataset shows different compositional characteristics. In 208 samples, the total richness of fish was 86 species. When the dataset was divided into two treatments of high vs. low-vegetation, rarified fish richness difference between treatments was 11.4 species (81 in the high-vegetation sites, and an average of 69.6 species in the rarefied low-vegetation sites - a fold difference of 0.22, Fig. 3A). The ranges of effect sizes that contained enough replications to conduct our quality-assurance power analysis by simulation were between -0.38 to 0.07 fold difference in fish (coloured bins in Fig. 3A).

Using the datasets with 2 treatments (high vs. low vegetation), the minimum detectable fish richness difference at a power of 0.8 decreased with coverage values (Fig. 3C, high values on the y-axis correspond to smaller effect sizes). The power analysis suggests that, assuming the richness difference between communities is 0.1 or greater, a study of these communities would require sampling sufficient to reach 89% coverage (or 98 samples in the high-vegetation community and 116 samples in the low-vegetation community) to detect a true difference 80% of the time. For scale, this assumed difference (0.1 log fold) is equivalent to a difference of 6 or 7 species between the communities. Note, we consider only the application of these simulated relationships to future samples or with assumed true effect sizes, to avoid the pitfalls of retroactive power analysis [28]. Our analysis also suggests that with smaller sample sizes (e.g., 6-10 samples), we would only be able to detect differences with 80% power when true differences are large (e.g., log fold difference of 0.3, equivalent to a difference of ~19 species).

The 1-treatment analyses estimated relationships between coverage and detectable effect size, providing estimates of sample sizes needed to detect differences assuming different true (future) effect sizes. For fish, the full pilot dataset achieved a coverage of 93% (Fig. 4C) with a detectable effect size of 0.09 fold difference (±6 species from the total of 86, at a power of 0.8). In other words, the pilot study size can detect changes of ±6 starting from 86 species.

The minimum effect quality assurance power analysis provides recommendations for sampling effort needed to detect changes above a user-defined biologically-relevant threshold. When we set this to 0.1 log2 fold-difference, or a 7.18% increase, we found that, as expected, the effect size always needs to be greater than this minimum level, in order to detect change of that scale or greater in 80% of the runs when it was indeed that large (Fig. 5). Also as expected, the extent to which the true effect needed to be greater than 0.1 decreased with sample size. As an example, for the sample size of the pilot to have a power of 0.8 to detect a change greater than 0.1 log2 change, the true effect size would need to be 0.19 log2 fold change, or a 14% difference (Fig. 5).

Sensitivity Analysis

We tested to see if our method converges on a stable coverage recommendation from repeated application to subsequent samples as 'pilot datasets' for the next set of samples. We used the fish dataset without treatments and arbitrarily chose 0.1 (~7% increase) as the target effect size we want to detect, which is close to one standard deviation from bootstrapped communities. Repeating the power analysis starting from different pilot study sizes and successively adopting new recommendations showed convergent coverage recommendations. Most pilot studies recommended sample sizes between 95 and 145 in the test case (Fig. 6) after the second iteration. At very small sample sizes (e.g., k=6), our analysis could not proceed due to cases in which the same sample is drawn by chance 100% of the time within a bootstrap, which fixes occupancies of all observed species to 100% (coverage $C(k)=1$ for all sample sizes k in Eq.2), resulting in no possible control over coverage values. This illustrates that a minimal number of samples is required for our power analysis, but it does not imply these small studies have no statistical power. For example, Fig. 2 shows that with 6 samples, a fold difference of 0.35 can be detected, but such an analysis was only possible from a larger pilot study. Our R package *BioDivPoweR* provides a warning when the simulations could not generate coverage variations.

Discussion

We present a novel, non-parametric quality-assurance power analysis for assessing sampling intensity needed within communities to robustly detect biodiversity change between them. This addresses a unique problem in biodiversity detection, and

contrasts from traditional power analyses that rely on independent replicates of treatment groups (e.g. Valdez et al). Our work supports previous findings of power relationships increasing with sampling effort within communities[17], but provides a method to apply the relationships from pilot data to the unsampled community to guide future sampling. Our main analysis estimates the rate of detecting biodiversity differences in the correct direction, but our modified *minimum-effect* power analysis estimates sampling effort needed to detect a change greater than an ecologically-relevant benchmark with sufficient power. This provides a practical approach to designing field surveys to detect change relevant for conservation action [12]. Through application to a large eDNA metabarcoding dataset, we found that the method yields convergent results on the recommended sample coverage required to detect a fixed change.

A key innovation that makes the power analysis applicable to real communities is the use of coverage, which was used to both rarefy within measurements and scale simulated findings for application. Our derivation shares the same mechanics as the best available asymptotic richness estimator [19] which uses information from observed species occupancies from the pilot study, and provides an analytical rarefaction curve that is used to obtain measurable effect sizes at different levels of effort. We consider coverage to be a dimensionless universal measure of sampling effort under the assumption that the pattern of occupancies of unobserved species is the same as that of observed species. We recognize that sampling biases stemming from rarity and spatial heterogeneity will likely violate the above assumption, but we do not yet know how to correct these second-order biases [19]. Therefore, we suggest an adaptive monitoring approach where, after adopting the sample coverage recommended for the next field survey, one repeats the power analysis and adopts the new recommendation until a convergent pattern becomes clear (Fig. 5). In addition, our single-sample method assumes stationarity in occupancy distributions and other features by which a treatment can affect a species accumulation curve [29], and we therefore suggest conservatism by erring on the side of more samples for planning before-after surveys where a pilot dataset is not available for both treatments.

We caution that like all power analyses, our approach should not be used to retroactively assess the power of a study to detect its own effect size [28,30]. This has been shown to be invalid because the probability to detect an observed effect size is not relevant after the effect has been observed (reviewed in [28]). Instead, our quality-assurance power analysis makes it possible to estimate the sampling effort required to detect a difference with a given probability under an assumed true effect in future monitoring (i.e. a prospective power analysis [10,12]), or to estimate how large of an effect would be necessary to detect a difference in the correct direction with an acceptable error rate if the sample size is fixed (minimum detectable effect size [31]). We also made it possible to assess sampling effort needed to detect a difference at

least as large as a given biologically-meaningful effect size (minimum effects power analysis [12,16]). We expect our approach to be useful to guide monitoring planning, especially with new technologies, when an effect size is in mind or a conservation action is tied to an ecologically-relevant amount of change. With caution, our analysis could be also used to ask if previous studies had sufficient power to detect a change of a given *externally-expected* size*,* but refer to the previous literature on cautions against incorrect retrospective power analysis [28,30].

We built our quality-assurance power analysis with relatively simple scenarios in mind, but expect that our general approach could be extended further. For example, we focused on detected changes in species richness, i.e. Hill number q=0 [32]. We believe it could be extended to abundance-weighted biodiversity estimators (q>0) in the future, possibly by implementing coverage estimators designed for q>0 [18]. Similarly, our method is focused on differences in richness across two treatments, but we believe the concept can be extended for detecting trends across multiple levels of treatments, or detecting trends across multiple timepoints. For example, a previous study found relatively lower sampling effort is needed to detect trends across multiple time points compared to detecting direction from two timepoints [17]. Our method is not currently designed to detect differences in community compositions (beta diversity); to do this, we would need a measure of effort analogous to coverage but specific to beta-diversity estimators, where future work is needed.

We made our method accessible through the R-package *BioDivPoweR* [21], which generates quality-assurance power analyses for any monitoring program that have collected pilot data with species and presence/absence occupancies across samples. The package provides relationships between sampling effort, true effect sizes, and rate of detecting a change in the correct direction or above a use-set difference. Additionally, the package can be used to compare the costs of sampling associated with statistical power that can be used to compare methods or different monitoring technologies; we believe this will help decisions on the relative value of changing sampling technologies.

Our method addresses an urgent need for designing and assessing biodiversity monitoring programs. Efficient and scalable monitoring systems are paramount, particularly as investments are made in new sampling technologies and survey programs to improve both the detection and attribution of biodiversity change. So far, such investments have often proceeded under the assumption that increased sampling effort and improved technology will enhance detectability; our framework provides a new quantitative basis for evaluating how much improvement is actually achieved.

Acknowledgements

We thank Evan Morien, Carolyn Prentice, Rute Carvalho, Kristin Meagher Robinson, Rosie Savage, Kyle Hall, Ben Millard-Martin, and Kate Sheridan for their contributions to field collections and processing of the pilot eDNA dataset. This work was funded by a

Genomics Applications Partnership Program grant from Genome Canada, and a Convergent Research Themes grant from the Computational and Data Systems Institute at McGill University

Author Contributions

ET designed the quality-control power analysis and wrote the original paper draft; JL contributed substantially to methods development and created the R-package; JS contributed to methods development, writing, and organized the team; ER, ML, RS, and NJ contributed substantially to project conceptualization, multiple rounds of feedback on methods, and contributed to writing.

Funding Statement

Eden Tekwa - Genomics Applications Partnership Program grant from Genome Canada
Jennifer Sunday - supported in part by the NSERC Canada Research Chairs Program
Jake Lawlor - supported by a Convergent Research Themes grant from the Computational and Data Systems Institute at McGill University, and by a Réal Decoste-Ouranos Partnership Training Award from the Province of Quebec.

Data availability

Pilot species occupancy dataset from eDNA survey are available here: https://doi.org/10.21966/vdyq-r660.
R-package and associated GitHub repository will be made available upon project publication
Code to replicate analyses for this manuscript will be made available upon project publication

Competing Interests

We have no competing interests.

References

1. Gonzalez A *et al.* 2020 Scaling-up biodiversity-ecosystem functioning research. *Ecol. Lett.* **23**, 757–776. (doi:10.1111/ele.13456)

2. Dasgupta P, Levin S. 2023 Economic factors underlying biodiversity loss. *Philos. Trans. R. Soc. B Biol. Sci.* **378**, 20220197. (doi:10.1098/rstb.2022.0197)

3. Valdez JW, Callaghan CT, Junker J, Purvis A, Hill SLL, Pereira HM. 2023 The undetectability of global biodiversity trends using local species richness. *Ecography* **2023**, e06604. (doi:10.1111/ecog.06604)

4. Magurran, McGill. 2011 *Biological Diversity: Frontiers in Measurement and Assessment*. Oxford University Press.

5. Cardinale BJ, Gonzalez A, Allington GRH, Loreau M. 2018 Is local biodiversity declining or not? A summary of the debate over analysis of species richness time trends. *Biol. Conserv.* **219**, 175–183. (doi:10.1016/j.biocon.2017.12.021)

6. Gonzalez A *et al.* 2023 A global biodiversity observing system to unite monitoring and guide action. *Nat. Ecol. Evol.* **7**, 1947–1952. (doi:10.1038/s41559-023-02171-0)

7. Convention on Biological Diversity (CBD). 2022 Convention on Biological Diversity. Kunming-Montreal Global Biodiversity Framework (Decision CBD/COP/DEC/15/4).

8. Wunder S *et al.* 2025 Biodiversity Credits: An Overview of the Current State, Future Opportunities, and Potential Pitfalls. *Bus. Strategy Environ.* **34**, 8470–8499. (doi:10.1002/bse.70018)

9. Jacquemont J, Blasiak R, Le Cam C, Le Gouellec M, Claudet J. 2022 Ocean conservation boosts climate change mitigation and adaptation. *One Earth* **5**, 1126–1138. (doi:10.1016/j.oneear.2022.09.002)

10. Cohen J. 1988 *Statistical Power Analysis for the Behavioral Sciences*. 0 edn. Routledge. (doi:10.4324/9780203771587)

11. Steidl RJ, Hayes JP, Schauber E. 1997 Statistical Power Analysis in Wildlife Research. *J. Wildl. Manag.* **61**, 270. (doi:10.2307/3802582)

12. Morrison LW. 2007 Assessing the Reliability of Ecological Monitoring Data: Power Analysis and Alternative Approaches. *Nat. Areas J.* **27**, 83–91. (doi:10.3375/0885-8608(2007)27[83:ATROEM]2.0.CO;2)

13. Nichols JD, Kendall WL, Boomer GS. 2019 Accumulating evidence in ecology: Once is not enough. *Ecol. Evol.* **9**, 13991–14004. (doi:10.1002/ece3.5836)

14. Chase JM, McGill BJ, McGlinn DJ, May F, Blowes SA, Xiao X, Knight TM, Purschke O, Gotelli NJ. 2018 Embracing scale-dependence to achieve a deeper understanding of biodiversity and its change across communities. *Ecol. Lett.* **21**, 1737–1751. (doi:10.1111/ele.13151)

15. Davies GM, Gray A. 2015 Don't let spurious accusations of pseudoreplication limit our ability to learn from natural experiments (and other messy kinds of ecological monitoring). *Ecol. Evol.* **5**, 5295–5304. (doi:10.1002/ece3.1782)

16. Myors B, Murphy KR. 2023 *Statistical Power Analysis: A Simple and General Model for Traditional and Modern Hypothesis Tests, Fifth Edition*. 5th edn. New York: Routledge. (doi:10.4324/9781003296225)

17. Gwinn DC, Allen MS, Bonvechio KI, V. Hoyer M, Beesley LS. 2016 Evaluating estimators of species richness: the importance of considering statistical error rates. *Methods Ecol. Evol.* **7**, 294–302. (doi:10.1111/2041-210X.12462)

18. Chao A *et al.* 2020 Quantifying sample completeness and comparing diversities among assemblages. *Ecol. Res.* **35**, 292–314. (doi:10.1111/1440-1703.12102)

19. Tekwa EW, Whalen MA, Martone PT, O'Connor MI. 2023 Theory and application of an improved species richness estimator. *Philos. Trans. R. Soc. B Biol. Sci.* **378**, 20220187. (doi:10.1098/rstb.2022.0187)

20. Moin P, Chan WHR. 2024 *Fundamentals of Turbulent Flows*. 1st edn. Cambridge University Press. (doi:10.1017/9781009431385)

21. BioDivPoweR: Analyze Power to Detect Change in Biodiversity Sampling Schemes. R package version 0.1.0, *Package DOI withheld for anonymity. Instead, a local, anonymized version of the package is available as a compressed file within the project repository uploaded as a supplementary file for reviewers (BioDivPoweR-87C7_June02).*

22. Gotelli NJ, Colwell RK. 2001 Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness. *Ecol. Lett.* **4**, 379–391. (doi:10.1046/j.1461-0248.2001.00230.x)

23. Webb TJ, Noble D, Freckleton RP. 2007 Abundance–occupancy dynamics in a human dominated environment: linking interspecific and intraspecific trends in British farmland and woodland birds. *J. Anim. Ecol.* **76**, 123–134. (doi:10.1111/j.1365-2656.2006.01192.x)

24. Zuckerberg B, Porter WF, Corwin K. 2009 The consistency and stability of abundance–occupancy relationships in large-scale population dynamics. *J. Anim. Ecol.* **78**, 172–181. (doi:10.1111/j.1365-2656.2008.01463.x)

25. *DOI of data redacted for anonymity. Data are instead available within the anonymized project R repository, uploaded as a supplementary file for review. Details of data collection methods are also available with the online dataset; an anonymized version of these methods are provided with the data within the anonymized project repository for review (Anonymized eDNA data methods.doc,* within *BioDivPoweR_manuscript_figs-master).*

26. Miya M *et al.* 2015 MiFish, a set of universal PCR primers for metabarcoding environmental DNA from fishes: Detection of more than 230 subtropical marine species. *R. Soc. Open Sci.* (doi:10.1098/rsos.150088)

27. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP. 2016 DADA2: High-resolution sample inference from Illumina amplicon data. *Nat. Methods* **13**, 581–583. (doi:10.1038/nmeth.3869)

28. Dormann CF, Ellison AM. 2025 *Statistics by Simulation: A Synthetic Data Approach*. Princeton and Oxford: Princeton University Press.

29. Chase JM, Knight TM. 2013 Scale-dependent effect sizes of ecological drivers on biodiversity: why standardised sampling is not enough. *Ecol. Lett.* **16**, 17–26. (doi:10.1111/ele.12112)

30. O'Keefe DJ. 2007 Brief Report: Post Hoc Power, Observed Power, A Priori Power, Retrospective Power, Prospective Power, Achieved Power: Sorting Out Appropriate Uses of Statistical Power Analyses. *Commun. Methods Meas.* **1**, 291–299. (doi:10.1080/19312450701641375)

31. Rotenberry JT, Wiens JA. 1985 Statistical Power Analysis and Community-Wide Patterns. *Am. Nat.* **125**, 164–168. (doi:10.1086/284335)

32. Jost L. 2006 *Entropy and diversity*. *Oikos* **113**, 363–375. (doi:10.1111/j.2006.0030-1299.14714.x)

Figures

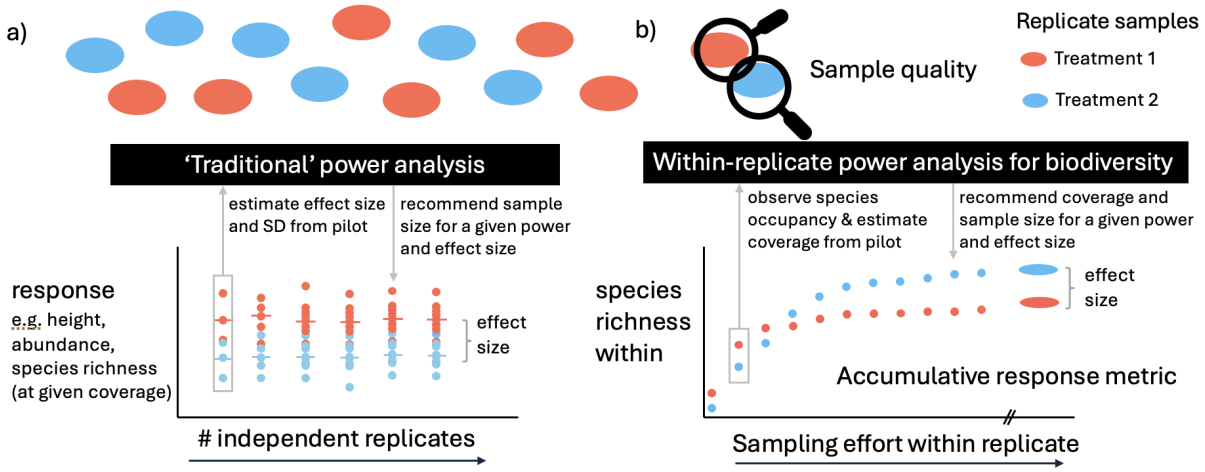

Fig 1. Traditional power analysis vs. within-replicate power analysis for biodiversity assessment. (a) In traditional power analysis, a small number of independent replicates, across two treatments (red vs. blue) are used to estimate the mean effect and second moment of effect (SD, standard error), to assess recommended sample size at a given effect size and power. Points represent estimates within each replicate, and horizontal lines represent mean across replicates (sample mean); mean and standard error are invariant with increasing numbers of replicates. Response values are any metric assessed within a replicate, and include species richness at a given level of sampling, or sample coverage. (b) In the quality-assurance power analysis for biodiversity, we focus on sampling effort *within* replicates, i.e. sample quality. Here, we are interested in how sampling effort (such as volume, area, or number of sites visited) improves the estimated richness difference between communities. An accumulated response metric like species richness estimates the true (census) richness with bias that is reduced with increased coverage. In this power analysis, we assess occupancy and estimate coverage from a pilot study and use simulations from the data to recommend a sample size at a given power and effect size.

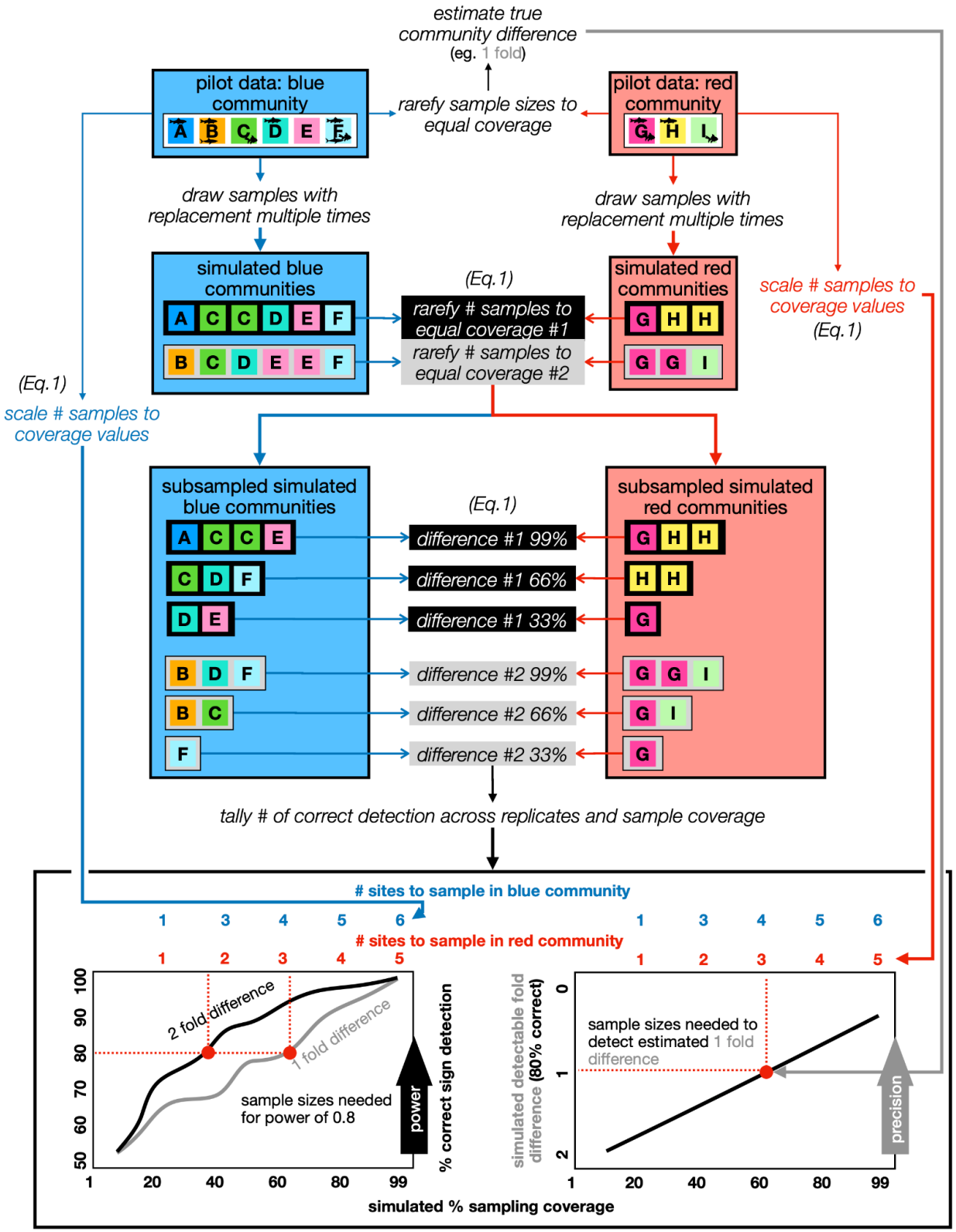

Figure 2. Power analysis from two-treatment data. Coloured boxes labelled A to F represent independent replicated samples from pilot data, and blue vs. red communities represent two communities with different treatments that are being compared.

Workflows and calculations are described in methods. Bottom plots summarize hypothetical simulation results for estimates of power/accuracy (LHS) and detectable effect size/precision (RHS), each as a function of sampling intensity in the pilot area given the same sampling method. LHS shows the percent of times that an observed effect size (differences in bottom red and blue boxes) over a range of coverage (from subsampling) is in the correct direction as the simulated "truth" (eg. 1 or 2 fold difference, measured between the rarefied simulated communities). RHS shows the minimum effect size that can be correctly detected at a given level of power (eg. 0.8). Note here that the y-axis is reversed so that increases in precision (smaller detectable effect sizes) are in the positive direction. All sample-to-coverage calculations, including rarefactions, subsampling, and scaling to final applications, use Eq. 2 (samples vs. coverage relationship).

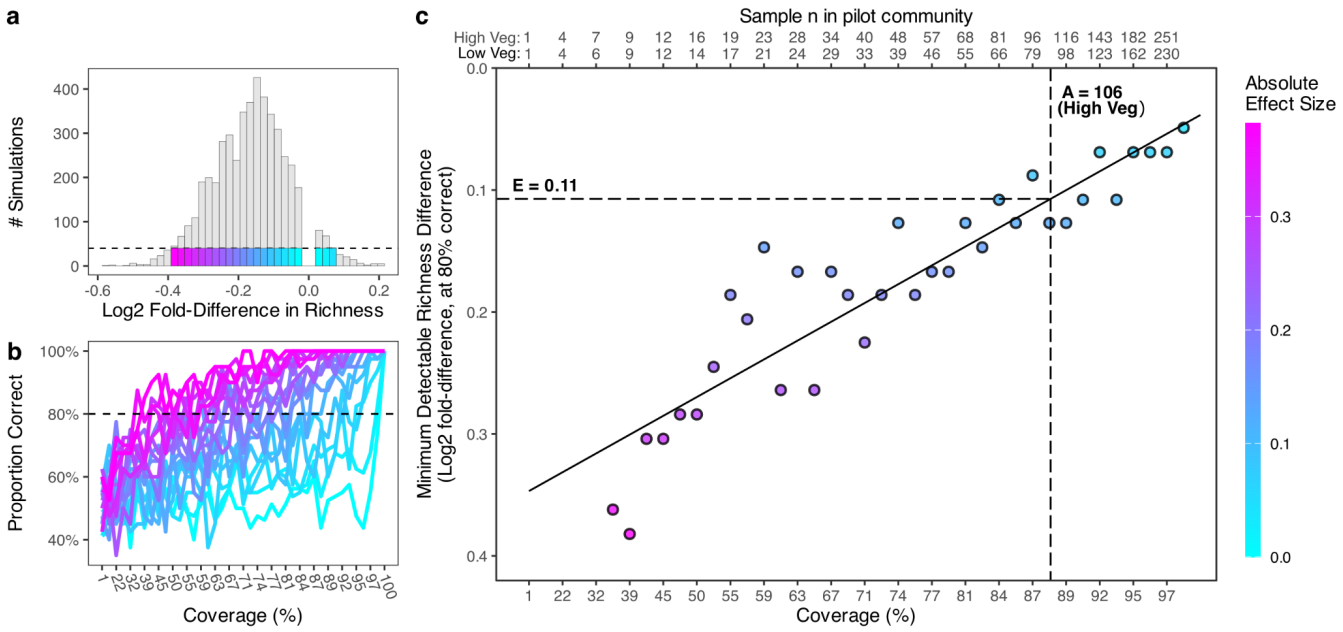

Figure 3. Power analysis results for two-treatment fish data. Simulations are based on data from two marine communities that we wish to compare: fish species at high-vegetation sites vs. fish species at low-vegetation sites (n = 106 & 102, rarefied richness = 81 & 69.6. **A.** Simulated richness log-ratio from random draws of each community (log_2(high-veg richness / low-veg richness), with zero-change simulations removed, provides a range of effect sizes with an equal number of simulations in each, colour coded from blue to magenta in the order of increasing absolute effect size; dashed line shows number of simulations retained (user-set, default is 40 replicates). Retained simulations are downsampled once to provide an equal number of replicates for each simulated effect size for subsequent power analysis. **B.** Relationship between power to detect change and richness coverage from simulations. Percent of simulations with correctly detected difference sign is shown as a function of down-sampled % coverage, assuming observed community is 100% coverage. Lines show results for each effect size bin; dashed line indicates a target power (user-set, default 80%). **C.** Minimum detectable effect size as a function of sample size and % coverage in each community, at 80% power. The upper x-axis shows sample size recommended from each community, while lower x-axis shows the corresponding equal sample coverage. The y-axis is reversed in scale to show smaller differences towards the top, indicating greater precision. The solid line is an ordinary least-squares regression fit to the data; vertical dashed line is achieved equal coverage from the original dataset; the horizontal dashed line is the minimum detectable effect size at achieved equal coverage. The regression estimates that the current sample size (A) can detect an effect size (E) higher than 0.11, 80% of the time in the correct direction.

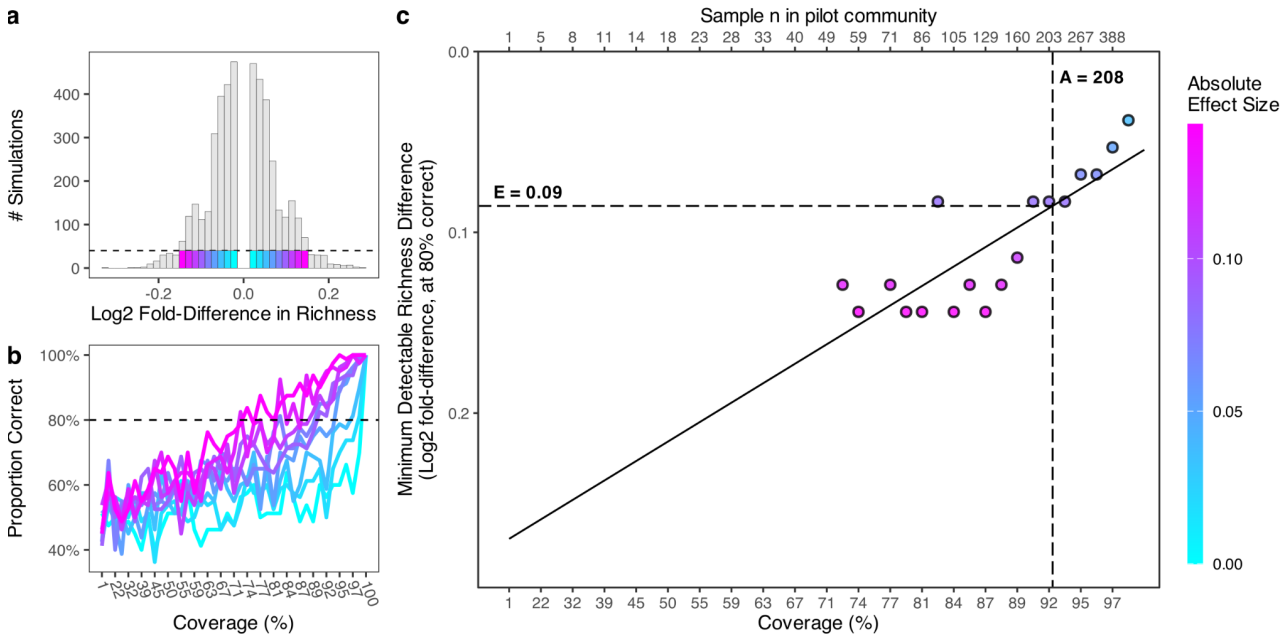

Figure 4. Power analysis results for single-treatment fish data. Simulations are based on data from a single community in time. Species richness is 86. **A.** Simulated rarefied richness fold difference (log2(richness at time 2 / richness at time 1) colour coded from blue to magenta in the order of increasing effect size, with data from time points being two bootstraps from the original dataset. The range of differences (binned effect sizes) with 40 or more replicates (user-set, dashed line) is used for subsequent power analysis. **B.** Percent of simulations with correct difference sign across true effect size magnitude bins and coverage. **C.** Detectable effect size as a function of sample coverage at a power of 0.8. The fitted line is from an ordinary least-squares regression. The regression estimates that the current sample size (A) can detect an effect size (E) higher than 0.09, 80% of the time in the correct direction.

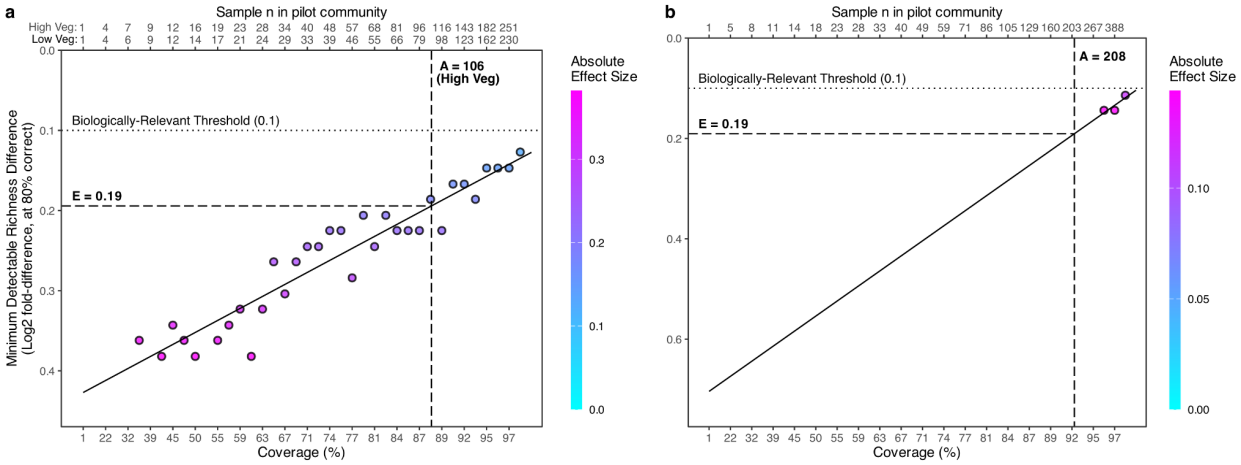

Figure 5. Minimum Effects Power Analysis for detecting relevant thresholds. Recommendations for sampling effort to detect changes above a user-defined biologically-relevant threshold (here, 0.1 log2 fold-difference, a 7.18% increase) for the two-treatment method (a) and the single-treatment method (b) using the example fish eDNA dataset. Plots mirror Figures 3c and 4c, respectively, but here power is set as a 0.8 probability of detecting change at least as great and in the correct direction as the user-defined minimum effect size for biological or conservation relevance. Because of measurement error, the effect size always needs to be greater than this minimum (threshold) level, and needs to be increasingly greater as sample size decreases (higher values are lower on y-axis as samples size is decreased on x-axis).

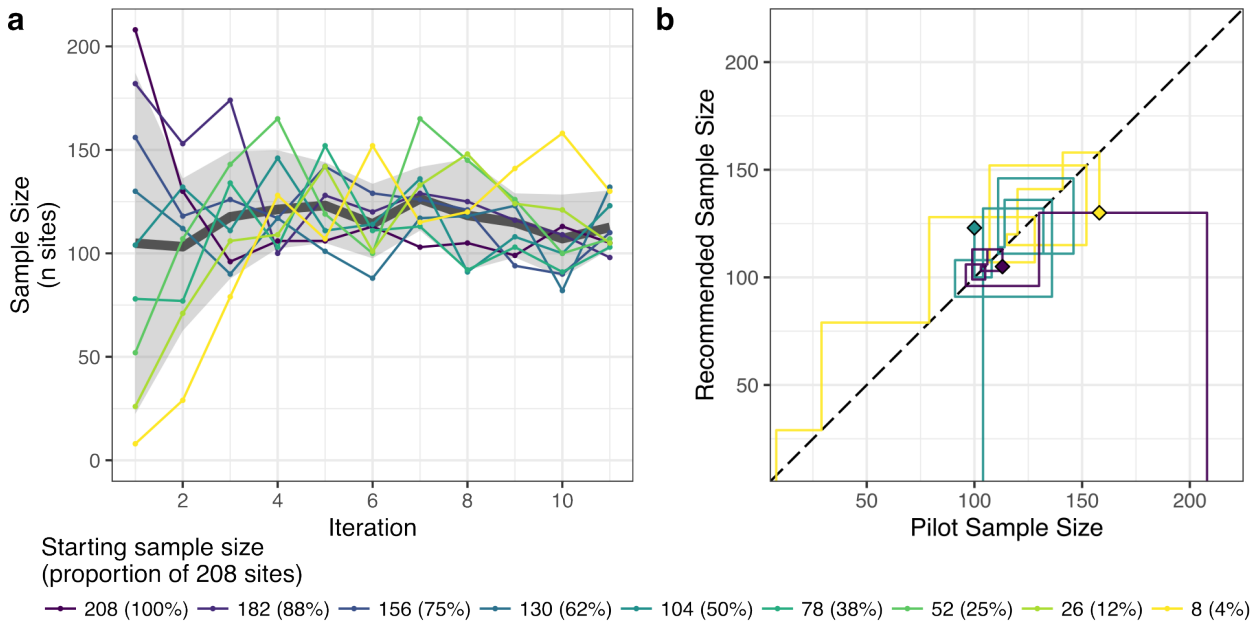

Figure 6. Convergence of sample size recommendations. Recommendations from repeating simulated power analysis and adapting through iterations starting from 9 different pilot study sizes. Simulations are from the 1-treament dataset with an objective of detecting a log-fold difference of 0.1 with a power of 0.8. **A.** Recommended sample size through 11 adaptive iterations. Black trend shows the mean recommendation over iterations, and shade shows the 10th and 90th quantiles. **B.** Phase diagram relating pilot sample size (iteration *t*) to recommended sample size (iteration *t*+1) for three trajectories (origins o; pilot sample sizes 8, 104, 208 sites) over 11 iterations.