57

A hierarchical modeling approach for environmental DNA metabarcoding: inference of species detection process, site occupancy, and study design

Environmental DNA (eDNA) metabarcoding is an emerging technology for gauging the distribution and diversity of species. Compared to the traditional survey methods such as capturing or sighting individuals, species detection by eDNA metabarcoding can be more sensitive. Nevertheless, the detection of species using eDNA metabarcoding is yet imperfect, owing to various factors that can cause false negatives in the inherent multistage workflow. False negatives in the multistage workflow of eDNA metabarcoding also complicate the identification of the ideal allocation of resources among the different stages in optimizing research efficiency. To address these issues, we propose a variant of the multispecies site occupancy model for eDNA metabarcoding studies, where samples are collected at multiple sites within a region of interest. In contrast to traditional site occupancy models with a binomial (or Bernoulli) observation model, this model employs a multinomial observation model to describe the variation in sequence read counts, the output of the high-throughput sequencers. It explicitly accounts for the hierarchical workflow of eDNA metabarcoding and interspecific heterogeneity and allows the analysis of the sources of variation in the detectability of species throughout the different stages of the workflow. The model can also be used to identify the study design that optimizes the effectiveness of species detection using a Bayesian decision analysis approach. An application of the model to freshwater fish communities in the Lake Kasumigaura watershed in Japan highlighted a remarkable inhomogeneity in the detectability of species, indicating a potential risk of the biased detection of specific species. Species with lower site occupancy probabilities tended to be difficult to detect as they had lower capture probabilities and fewer sequence reads. A study design analysis suggested that ensuring multiple within-site replications of the environmental samples is preferred to achieve higher species detection effectiveness, provided that tens of thousands of sequence reads were secured per replicate. These results suggest that the use of hierarchical models that explicitly account for the inherent multistage process of species detection in eDNA metabarcoding makes the application of eDNA metabarcoding more error-tolerant and allows ecologists to monitor ecological communities more efficiently.