How to quantify immigration from community abundance data using the Neutral Community Model

Ramis Rafay[a], Eric W. Jones[b,c,d], David A. Sivak[d], S. Jane Fowler[a]
[a]*Department of Biological Sciences, Simon Fraser University, Burnaby, BC, Canada V5A1S6*
[b]*Department of Physics and Energy Science, University of Colorado Colorado Springs, Colorado Springs, CO, USA 80918*
[c]*BioFrontiers Center, University of Colorado Colorado Springs, Colorado Springs, CO, USA 80918*
[d]*Department of Physics, Simon Fraser University, Burnaby, BC, Canada V5A1S6*

Corresponding author: S. Jane Fowler

Email: sjfowler@sfu.ca

Competing Interest Statement: N/A

Classification: Ecology

Keywords: dispersal, neutral theory, immigration, community assembly, sampling, sequencing

Abstract

Biological communities are connected through dispersal, which regulates diversity across local and regional scales. However, dispersal is difficult to measure directly, limiting what is known about dispersal's impact on species composition in complex communities. One method to measure dispersal employs the Neutral Community Model (NCM) to quantify how a local community is influenced by the immigration of individuals from a larger source community. Conveniently, the immigration rate $N_T m$ of the NCM can be fit from biological sequence abundance datasets, which are plentiful. Yet it is neither known if these estimated values reflect the ground truth, nor what sampling effort is required to yield accurate estimates. In this study we introduce two inference methods, a variance-based and a Dirichlet-Multinomial Log-Likelihood (DM-LL) method, to complement the established occupancy-based inference method. In simulations of communities that resemble activated sludge microbiomes, all inference methods were capable of estimating $N_T m$ within 10% of ground-truth, with the variance-based and DM-LL methods requiring less sampling effort. Accurate inferences require read depths greater than $N_T m$ in each sample. The three methods agree in their inferred $N_T m$ in simulations of communities experiencing weak non-neutral effects (e.g., selection), and in applications to an empirical dataset from wastewater activated sludge. Based on these findings, we propose practical sampling and methodological guidelines for quantifying immigration between highly diverse, complex communities using the NCM.

Introduction

A central goal in ecology is to describe how biological communities assemble and vary over time and space. The four fundamental community assembly processes—selection, speciation, dispersal, and drift—determine taxa composition and diversity (*i.e.*, community structure) in any habitat (1). While selection, speciation, and drift operate at local scales, dispersal is a strictly regional process that connects communities across larger spatial scales through immigration and emigration. Moreover, the rate of dispersal mediates the relative influence of stochastic and deterministic assembly processes: high dispersal homogenizes communities and reduces the influence of selection and drift, while low dispersal increases the impact of local differentiation and taxa turnover on community structure (2). Advances in high-throughput sequencing methods enable researchers to census micro- and macro-organismal communities with unprecedented resolution (3, 4), facilitating the development of models that describe community assembly and community structure at finer spatial and temporal scales. The Neutral Community Model (NCM) has emerged as a powerful and analytically tractable framework for quantifying how dispersal affects community structure in ideal 'neutral' communities in which all taxa are assumed to have equal growth and death rates (5). Though biological communities are obviously not made of species with identical fitness, NCMs nonetheless recapitulate observed ecological patterns like taxa-area, and distance-decay relationships (6–9).

In the NCM, the abundance of each taxon in a 'local' community is determined by stochastic birth-death processes and immigration from an external 'source' community. At each timestep a random individual in the local community is lost to death or emigration, then replaced by (i) an immigrant from a larger source community with a probability m, or (ii) by the reproduction of a random individual in the local community with a probability $1-m$ (Fig. 1). The long-term taxa abundance distributions predicted by the NCM are approximated by a Dirichlet distribution that depends on the immigration rate $N_T m$, the expected number of immigrants colonizing the local community per N_T replacement events (10, 11). The larger the immigration rate, the more local-community composition is driven by immigration from external sources, which can be an important process for maintaining the presence of rare (low abundance) taxa that would otherwise go extinct due to drift.

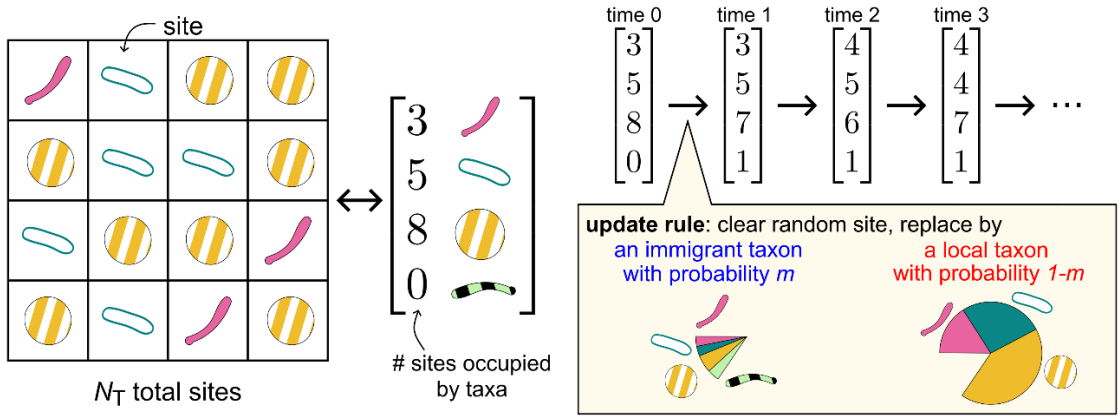

Figure 1. Schematic of the Neutral Community Model, describing the temporal dynamics of a 'local' community with N_T sites that are each occupied by some taxon i. At each timestep, one site is cleared and replaced by an individual from a 'source' community with probability m or by an individual from the local community with probability $1-m$. In the source community, the relative abundance p_i of each taxon is fixed.

In this paper, we set out to examine how sampling affects the reliability of $N_T m$ inference. In any ecological sampling campaign, both the number of samples taken from a community (e.g., aliquots of biomass or count surveys) and the total number of counts in each sample (i.e., read depth in sequencing parlance) influence the accuracy and reliability of measured taxon abundances. A limited number of samples leads to an imprecise understanding of how community structure varies over space and/or time, while a limited read depth hampers the detection of rare taxa leading to an underestimation of biodiversity (12, 13). However, increasing sampling effort requires additional material and labor investment, highlighting the importance of rational sampling campaigns that optimize the allocation of resources.

In addition to sampling challenges, there are methodological limitations in how $N_T m$ is inferred. Though the standard 'occupancy-based' $N_T m$ inference method has been widely applied to viral (14), microbial (15–20) and macro-organismal (21, 22) datasets, it relies only on the presence/absence of each taxon, a small fraction of the total amount of information contained in abundance datasets. To complement the established occupancy-based method, here we introduce two $N_T m$ inference methods: a variance-based method that matches the variance in measured taxon relative abundances to the variance predicted by the NCM, and a likelihood-maximization method grounded in the Dirichlet-multinomial distribution (DM-LL). With neutral community assembly simulations, we validate these inference methods and measure their accuracy as a function of number of samples and read depth. Last, we apply these inference methods to simulated non-neutral communities and an empirical dataset to evaluate the generality of each method. Our results show that the range of immigration rates that can be inferred is limited by the read depth, and that the variance-based and DM-LL methods yield more accurate estimates than the occupancy-based method. Our findings lay out practical guidelines for quantifying immigration in complex communities using the Neutral Community Model.

Results

Methods to infer the immigration rate $N_T m$ from sampled abundance data

Following the NCM update rule (Fig. 1), the relative abundance X_i (of each taxon i out of S taxa) will fluctuate over time and eventually attain a stationary steady state (i.e., a statistical steady state). In the limit that the number of sites in the local community is very large, the long-term relative abundances $X_1, \ldots, X_S$ of the local community are Dirichlet distributed (23):

$$X_1, \ldots, X_S \sim \mathrm{Dirichlet}(N_\mathrm{T} m p_1, \ldots, N_\mathrm{T} m p_S). \tag{1}$$

Empirical relative abundance distributions can be compared against this theoretical distribution. More pointedly, the Dirichlet distribution's dependence on the immigration rate $N_T m$ means that the relative abundance distribution of each taxon is shaped by immigration, giving rise to the following variance, occupancy, and Dirichlet-Multinomial $N_T m$ inference methods.

Variance-based inference

Marginalizing over the Dirichlet distribution, the relative abundance of each taxon in the local community is Beta distributed with $X_i \sim \mathrm{Beta}\big(N_\mathrm{T} m p_i, N_\mathrm{T} m(1 - p_i)\big)$. Therefore, the mean relative abundance of a taxon in the local community is equal to its relative abundance in the source community $\mathrm{E}[X_i] = p_i$, and its variance is inversely proportional to $N_\mathrm{T}m$, $\mathrm{Var}[X_i] = p_i(1-p_i)/(N_\mathrm{T}m+1)$. Accordingly, $N_\mathrm{T}m$ can be inferred from the observed variances and average relative abundances of each taxon. To account for sampling noise, observed variance $\widehat{\mathrm{Var}[X_i]}$ in each taxon's relative abundance also includes a contribution from the binomial sampling variance $p_i(1-p_i)/R$, where R is the read depth. The best fit $N_\mathrm{T}m$ is obtained by least-squares regression between observed and predicted variance:

$$\widehat{\mathrm{Var}[X_i]} = \frac{p_i(1-p_i)}{N_\mathrm{T}m+1} + \frac{p_i(1-p_i)}{R}. \tag{2}$$

The larger the parameter $N_\mathrm{T}m$, the lower the variance of relative abundances of local-community taxa (i.e., the 'tighter' the relative abundance distribution across samples).

Occupancy-based inference

A taxon's sample occupancy O_i is the proportion of samples in which it has nonzero abundance (it is 'present'). The occupancy O_i of neutral communities produced by the NCM can be calculated (24) by integrating the probability density function of the beta distribution from the sample detection limit D to 1. The detection limit is the smallest measurable value of the relative abundance, $1/R$. The best fit $N_\mathrm{T}m$ is obtained by least-squares regression between observed and predicted sample occupancies:

$$O_i = \int_D^1 \mathrm{P}[X_i = x_i]\mathrm{d}x_i = \int_D^1 x_i^{N_\mathrm{T}mp_i-1}(1-x_i)^{N_\mathrm{T}m(1-p_i)-1}\mathrm{d}x_i. \tag{3}$$

As $N_\mathrm{T}m$ increases, the occupancy of all observed taxa approaches 1.

Dirichlet-Multinomial Log-Likelihood inference (DM-LL)

Last, the Dirichlet-Multinomial Log-Likelihood (DM-LL) method identifies the $N_\mathrm{T}m$ value that maximizes the likelihood of having observed the measured distribution of reads in local-community samples. Formally, this likelihood is the probability of drawing J samples from a multinomial distribution whose probabilities are themselves drawn from a Dirichlet distribution (Eq. 1),

$$\mathrm{P}(\mathbf{y}) = \prod_{j=1}^{J} \int_{\hat{\mathbf{p}}} \mathrm{Mult}\big(\boldsymbol{y}^{(j)} \,\big|\, R,\, \hat{\mathbf{p}}\big)\, \mathrm{Dir}(\hat{\mathbf{p}} \,|\, N_\mathrm{T}m\,\mathbf{p})\, \mathrm{d}\hat{\mathbf{p}} = \prod_{j=1}^{J} \frac{R\,\mathrm{B}(N_\mathrm{T}m, R)}{\prod_{k:y^{(j)}>0} y_k^{(j)}\,\mathrm{B}\big(N_\mathrm{T}m\,p_i, y_k^{(j)}\big)}, \tag{4}$$

where the Dirichlet distribution captures the dependence on the relative abundance of each taxon in the source community p_i and $N_\mathrm{T}m$, and the multinomial distribution captures the effects of sampling. Each multinomial draw consists of R reads, $\mathbf{y}^{(j)}$ is the vector of

read abundances for sample *j*, $\boldsymbol{y}$ is the set of all samples, $\mathrm{B}(z_1, z_2) = \frac{\Gamma(z_1)\Gamma(z_2)}{\Gamma(z_1+z_2)}$ is the beta function, and $\Gamma(z) = \int_0^\infty t^{z-1}e^{-t}\mathrm{d}t$ is the gamma function. The resulting Dirichlet-multinomial distribution is well-studied (25) and analytically tractable, as demonstrated by the striking simplification in Eq. 4. The maximum-likelihood estimate of $N_\mathrm{T}m$, $\mathcal{L}(N_\mathrm{T}m|\mathbf{y})$, maximizes $\mathrm{P}(\mathbf{y})$; equivalently, it maximizes the log-likelihood:

$$\log[\mathcal{L}(N_\mathrm{T}m\,|\boldsymbol{y})] \sim \sum_{j=1}^{S}\Big[\log R + \log\Gamma\,(N_\mathrm{T}m) + \log\Gamma\,(R) - \log\Gamma\,(N_\mathrm{T}m + R) - \sum_{k:y_k^{(j)}>0}\Big[\log y_k^{(j)} + \log\Gamma\,(N_\mathrm{T}mp_k) + \log\Gamma\left(y_k^{(j)}\right) - \log\Gamma\left(N_\mathrm{T}mp_k + y_k^{(j)}\right)\Big]\Big]. \quad (5)$$

Inference improves with increased read depth and number of samples

The Neutral Community Model posits an inverse relationship between the immigration rate and the variance of the relative abundance of each taxon Var[X_i] in the local community according to $\mathrm{Var}[X_i] = \frac{p_i(1-p_i)}{N_\mathrm{T}m+1}$. However, when communities are sampled, the estimate of a taxon's relative abundance X_i itself has a variance $\mathrm{Var}[\widehat{X}_i] = \frac{p_i(1-p_i)}{R}$ due to sampling noise, which scales inversely with the read depth R. When $R < N_\mathrm{T}m$, sampling noise dominates the ecological variance, obscuring the underlying community dynamics. Put simply, sample read depths must be larger than $N_\mathrm{T}m$ for accurate inference. Since $N_\mathrm{T}m$ cannot be known prior to sampling, researchers can avoid situations where inferences are limited by read depth by sequencing samples as deeply as possible (Figs. S2-S3). Increasing the number of samples improves the resolution for taxon sample occupancies, sampled abundances, and their variances.

Variance-based and Dirichlet-Multinomial inference methods require less sampling

In this idealized sampling scenario, we explore how estimated immigration rates depend only on local-community sampling intensity and the inference method used. To this end, inferences were performed on simulated abundance datasets of neutrally assembled communities, with source-community relative abundances (Fig. S1) chosen to resemble a wastewater microbiome (Methods). Increasing the read depth and the number of samples led to more accurate and consistent $N_\mathrm{T}m$ estimates (Fig. 2). Increasing the number of local-community samples increases the accuracy of all inference methods, but little improvement is achieved beyond taking 10 samples (Figs. S2 and S3). The DM-LL inference method provided the most accurate estimates of $N_\mathrm{T}m$ across all combinations of read depth and number of samples, whereas the occupancy method required the largest read depth and number of samples for inference within 10% of the simulated value. Provided that a community is well described by neutral dynamics and read depths are larger than $N_\mathrm{T}m$, our findings suggest that all methods will estimate immigration rates within 10% of the true value with at least 10 local-community samples.

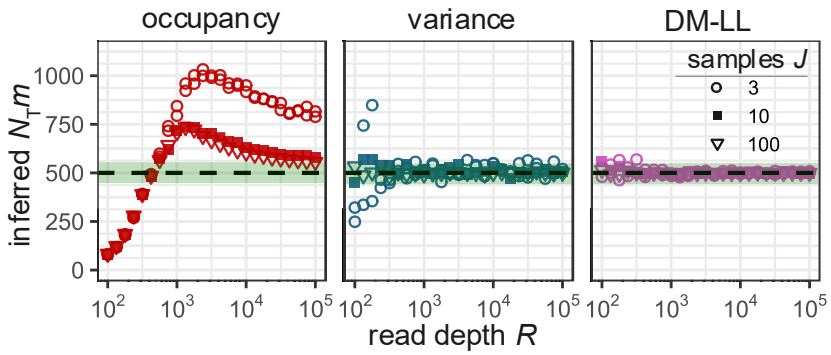

Figure 2. Inference accuracy depends on the sampling effort of the local community. Inferred N_Tm from 3, 10, and 100 local-community samples at read depths of 10^2-10^5 using the occupancy-based (red), variance-based (blue), and DM-LL (pink) methods. Source-community abundances were chosen to resemble a wastewater microbiome (Methods). The shaded area indicates N_Tm within 10% of the ground-truth value of 500 (dashed line).

Sample a separate source community at least as much as the local community

Important ecosystems like wastewater bioreactors and gut microbiomes consist of microbes that at one point immigrated and colonized from an 'upstream' community (18). In the NCM, the flow of immigrants 'downstream' is proportional to N_Tm, and accurately estimating this quantity requires knowledge of taxa relative abundances in the upstream source community p_i as well as taxa relative abundances in the downstream local community. Given that the source and local communities must be sampled separately, we next investigate how sampling of the source community (which determines p_i in the NCM) affects N_Tm inference (Fig. 3). In the limit that sampling of the source community achieves exact estimates of p_i, inferences achieve the accuracy of the idealized sampling scenario discussed in the previous section. Importantly, N_Tm is accurately inferred only when samples have read depths greater than N_Tm, regardless of the number of samples taken (Figs. S3). When at least 10 source-community samples are taken, typical read depths of 10^3–10^4 (26, 27) are sufficient to infer the immigration rate within 10% of simulated ground-truth when $N_Tm \leq 5000$ (Figs. S4 and S5).

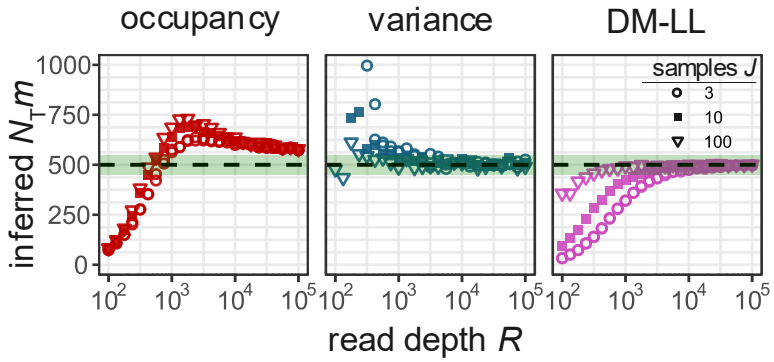

Figure 3. Increasing source-community sampling effort improves inference accuracy, approaching the ideal case that taxon p_i values are exactly known. Inferred N_Tm from 3, 10, and 100 source-community samples for the occupancy-based (red), variance-based (blue), and Dirichlet-multinomial log-likelihood (DM-LL, pink) methods. In all simulations, 10 samples were taken from the local community, and read depths for both source

and local communities are varied from 10^2-10^5. The shaded area indicates N_Tm within 10% of the ground-truth value of 500 (dashed line).

Quantifying immigration in non-neutral communities

In nature, communities are not strictly neutral. Selection, for example, can favor certain taxa over others, causing local-community abundances to deviate from neutral expectations. To investigate how deviations from neutrality influence N_Tm inferences, we randomly biased the relative abundance of local-community taxa by an amount proportional to a 'non-neutrality' parameter σ. Specifically, the bias α_i in the relative abundance of taxon i was drawn from a normal distribution $N(0, \sigma^2)$ where larger σ increases the likelihood of stronger deviations from neutrality (Methods). These biases cause some taxa to increase in abundance while others decline or disappear, perturbing the otherwise Dirichlet-distributed local community expected under purely neutral dynamics. Four values of the non-neutrality parameter σ were tested: 0 (neutral), 5×10^{-5} (weak), 5×10^{-4} (moderate), and 5×10^{-3} (strong). Deviations from neutrality affect the accuracy of N_Tm inference because even weak selection can drive rare taxa to extinction—eliminating any possibility of measurement—or increase their abundances far beyond neutral expectations.

The variance-based and DM-LL methods converged to accurate estimates with increasing read depth for weak and moderate levels of non-neutrality (Fig. 4 and S6). The occupancy-based method was the most sensitive to deviations from neutrality and did not converge to a stable N_Tm estimate with increasing read depth, unlike the neutral case, because previously rare taxa regularly achieved occupancies of 1 (Fig. S7). To correct this, we modified the occupancy method to only consider taxa with sample occupancies less than 1, giving better performance at weak non-neutrality (Fig. S10). Indeed, none of the inference methods could accurately infer N_Tm from local communities where strong non-neutrality impacts community structure (Figs S7-S9). At increasing levels of non-neutrality, the variance method tends to overestimate N_Tm, whereas the DM-LL and occupancy-based methods tend to underestimate N_Tm.

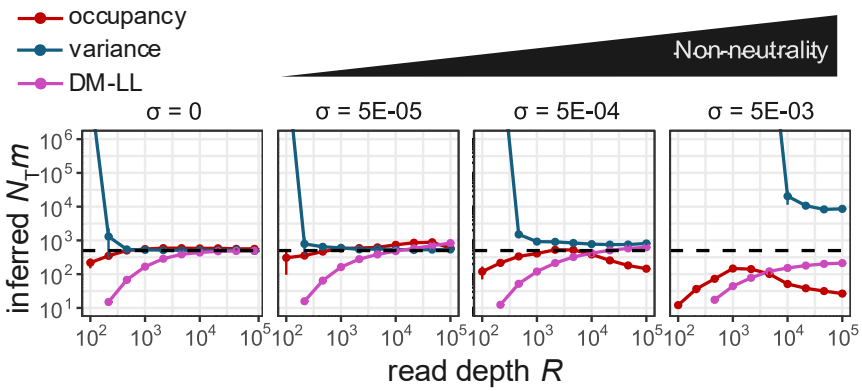

Figure 4. Non-neutrality reduces agreement between inference methods. Inferred N_Tm in neutral (σ = 0) and non-neutral (σ = 5×10^{-5}, 5×10^{-4}, and 5×10^{-3}) local communities using the occupancy (red), variance (blue), and DM-LL (pink) methods at a simulated N_Tm of 500 (dashed line). Each data point consists of 10 simulations in which selection advantages are randomly drawn, local-community abundances are simulated, the source and local communities are sampled, and inference is performed. Error bars show one standard deviation.

To infer N_Tm from real data: rarefy, infer, and repeat

Finally, we apply these methods to a real 16S rRNA gene amplicon dataset consisting of samples from a full-scale activated sludge system and influent wastewater (15). This dataset provides an opportunity to validate our recommendations for sampling and N_Tm inference under natural conditions, where the true N_Tm value is unknown and additional non-neutral ecological complexities may influence community assembly. To assess inference stability and detect potential deviations from neutrality in each method, abundance tables with progressively smaller read depths were used for N_Tm inference (Methods). Inference profiles—inferred N_Tm values at varying rarefied read depths— show agreement between the estimates at read depths of 10^5 (N_Tm_{occ} = 905, N_Tm_{var} = 1252, N_Tm_{DM} = 626) within one order of magnitude (Fig. 5); moreover, estimates from all three inference methods were stable across high read depths, confirming that inferences are not read depth limited. This approach demonstrates that rarefaction, repeated inference, and comparison across methods can provide confidence towards estimates of the immigration rate while revealing when selection or other non-neutral processes impact community assembly.

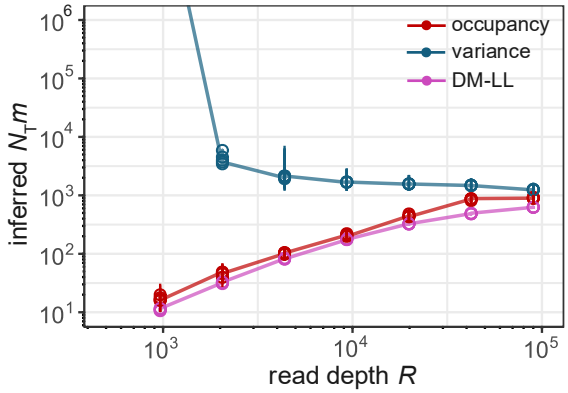

Figure 5. Inferring the immigration rate from wastewater to an activated sludge microbiome in a full-scale wastewater treatment plant. Inference used abundances rarefied to $5x10^2$–$1x10^5$ reads per sample (Methods). Error bars show one standard deviation.

Discussion

Pitfalls abound when attempting to quantify immigration rates using the Neutral Community Model. Motivated by our findings, we summarize key considerations here.

What is your source community?

Originally, the NCM was formulated within the context of island biogeography (28), considering a large mainland community that serves as the source of immigrants to a smaller island community. This scale asymmetry creates a directional influence of immigration in which the island (local) community is shaped by immigration from the mainland (source) but not vice versa. On the other hand,, many ecological systems are better represented as metacommunities—networks of interacting local communities connected by dispersal without a single external source community (29). If an experimental system possesses upstream (mainland) and downstream (island) communities, both communities must be sampled separately, and the mainland-

community abundances serve as the inputs for estimating the immigration rate $N_T m$. By contrast, under a metacommunity framework the composition of the source community is not an independent entity but is instead the average of the local-community samples. Identifying the appropriate spatial model—mainland-island or metacommunity—impacts sampling requirements, NCM estimation, and the interpretation of estimated parameters (30, 31).

Take at least 10 samples from the local and source communities and maximize sequence read depths

Resource constraints often limit the scale of sampling efforts, making optimized sampling strategies essential for studying community assembly. Our results show that read depth upper-bounds the $N_T m$ that can be inferred. Increasing the number of samples improves the precision and resolution of observed sample occupancies and improves inferences regardless of the inference method applied, although diminishing returns were observed beyond 10 local-community samples. Though more is always better, based on our simulations, we recommend taking at least 10 samples and prioritizing deep sequencing for robust $N_T m$ estimation. Similar recommendations apply when a source community must also be independently sampled.

Apply all three inference methods to build confidence in estimates

The three inference methods each leverage different aspects of abundance data—occupancy, variance in relative abundances, and abundance distributions—to estimate $N_T m$. Under neutral conditions, all three methods can infer $N_T m$ within 10% of its ground-truth value. However, non-neutrality impacts their estimates in distinct ways (Fig 4, S6). For empirical datasets, researchers should apply all three methods and evaluate the consensus between their estimates: provided that read depth is not the limiting factor, non-neutral effects likely dominate if the three methods yield inconsistent estimates. In this case, community assembly is not well described by the NCM.

Make the most of your abundance data

In practice, read depths can vary up to 100-fold between individual samples, making normalization techniques such as rarefaction/subsampling a common preprocessing step before constructing relative abundance tables (32). When incorporated as part of a bootstrapping method, rarefaction provides a systematic way to evaluate confidence in $N_T m$ estimates. With evenly subsampled abundance tables, researchers can evaluate how $N_T m$ estimates change across read depths and determine if inferences are limited by sampling intensity or unreliable due to non-neutrality. Applying the 'rarefy, infer, repeat' approach in the activated sludge dataset revealed distinct patterns among the three methods across read depths. The occupancy-based, variance-based, and DM-LL methods converged to stable $N_T m$ estimates at higher read depths, confirming that read depths were sufficient for reliable inference. Crucially, the similarity between $N_T m$

estimates lends confidence to the inferred rate of microbial immigration from the influent wastewater to the activated sludge community.

Estimating immigration rates in complex communities using the Neutral Community Model

Our results demonstrate that the NCM can address not only the qualitative question of whether community assembly is 'neutral' but also provide quantitative estimates of immigration rates shaping local communities. Through simulations, we codify practical recommendations for sampling and inference methodologies to guide researchers interested in applying the NCM to their study ecosystems; code that implements these inference methods is freely available. Furthermore, our 'rarefy, infer, repeat' approach enables researchers to extract robust immigration estimates from high-resolution sequence datasets. Although we found that $N_T m$ inference methods are resilient to some implementations of non-neutrality, future work should evaluate their robustness to real and simulated datasets exhibiting other forms of non-neutrality. By broadly applying these methods across diverse ecosystems and contexts, researchers can deepen ecological understanding of how dispersal influences community biodiversity.

Materials and Methods

Simulating neutral and non-neutral community assembly

The source-community taxa abundances were simulated by a log-series abundance distribution ($\Theta = 250$) with a total of 10^{10} individuals and 3453 taxa (Fig. S1). Parameters for defining the source community were motivated by taxa richness in activated sludge (33), expected cell counts per litre of wastewater (34), and considerations of computational feasibility.

The neutral local-community taxa abundances were generated using a Dirichlet distribution (Eq. 1), parameterized by the source-community abundances and one of three simulated $N_T m$ values: 5×10^2, 5×10^3, and 5×10^4. These values were chosen since pprevious studies have estimated $N_T m$ in this range (35–37). The NCM assumes that there are no fitness differences among taxa, so the expected relative abundances of local-community taxa mirror those in the source community.

To simulate violations of the neutrality assumption, an advantage term α_i is sampled from a normal distribution $N(0, \sigma^2)$ and randomly assigned for each local-community taxon i (20, 24, 38). The randomly drawn advantage terms have a 99.7% probability of falling within the range $(-3\sigma, 3\sigma)$. Briefly, the α terms are added to Dirichlet-distributed local-community relative abundances—advantaged taxa ($\alpha_i > 0$) and disadvantaged taxa ($\alpha_i < 0$) have their local abundances increased and decreased relative to their source-community abundances, respectively. Taxa with relative abundances less than zero after this addition have their abundance set to zero (simulating extinction). Finally, the now 'non-neutral' local-community relative abundances are re-normalized to sum to one. We simulated three levels of neutrality violation ($\sigma = 5\times10^{-5}$, 5×10^{-4}, and 5×10^{-3}) alongside the neutral case ($\sigma = 0$).

All simulations, analyses, and visualizations were created using *R 4.2.2* (39).

Sampling source- and local-community abundances

To simulate sampling and sequencing, J samples (each containing R reads) are drawn from simulated community abundances. Formally, each of these J samples is a draw from a multinomial distribution parameterized by the relative abundances of each taxon in the source- or local-community and by the read depth R.

Simulated Scenarios

Three scenarios were used to evaluate the performance of the three $N_T m$ inference methods under varying ecological and sampling conditions:

- Scenario 1: Inferring $N_T m$ assuming known source-community abundances. Local communities were simulated as neutral. Simulated source-community abundances were used as input p_i for the inference methods. The number of samples taken from the local community were varied from 2 to 100. Read depths were varied from 10^2-10^5.
- Scenario 2: Inferring $N_T m$ in cases where source-community abundances need to be sampled. Local communities were simulated as neutral. The number of samples taken from the source community were varied from 2 to 100 to determine input p_i for the inference methods. Ten samples of the local community were taken. Read depths were varied from 10^2-10^5.
- Scenario 3: Inferring $N_T m$ from non-neutral local communities. Local communities were simulated to deviate from neutrality, testing four levels of non-neutrality (σ = 0, 5×10^{-5}, 5×10^{-4}, and 5×10^{-3}). Ten samples were taken from both source and local communities. Read depths were varied from 10^2-10^5.

For each combination of (i) $N_T m$ values, (ii) number of samples taken, and (iii) read depths, ten simulations were run; each simulation generated relative abundance datasets. The reported $N_T m$ is the average of the $N_T m$ values inferred from each simulation.

Inferring immigration from a real wastewater bioreactor dataset

The three $N_T m$ inference methods were applied to 16S rRNA amplicon sequence dataset (V4 region) from a long-term sampling campaign of a full-scale activated sludge bioreactor (15). The activated sludge system and its influent wastewater were sampled twice every week over a year. The resulting dataset consists of 28 samples from each of the influent wastewater and activated sludge communities, and the read depth ranges from 31,309 to 163,733 reads per sample. Raw sequences were downloaded from the NCBI Sequence Read Archive (SRA) using the *SRA Toolkit* (https://www.ncbi.nlm.nih.gov/sra). Quality control, sequence denoising, pair-end read merging and chimera removal were performed using the *dada2* pipeline (40). Abundance tables were packaged into a *phyloseq* (41) object along with sample metadata, and used for analysis. The influent wastewater was considered the 'source' community, and the activated sludge system was considered the 'local' community. To build inference depth profiles, the abundance tables used as input for the inference methods were first rarified to different read depths before being converted

to relative abundance. To ensure at least 10 samples from both source and local communities are used for inference, the maximum read depth considered was chosen to be 10^5.

Data, material and software availability

R code that replicates Figs. 2-5 and that can be applied to any sequencing dataset is available at https://github.com/ramisrafay/NCM-Inference-Methods

Acknowledgments

This work was supported by an SFU Graduate Deans Entrance Scholarship awarded to R. R., Banting and Pacific Institute for the Mathematical Sciences Postdoctoral Fellowships awarded to E.W.J., a Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant and Discovery Accelerator Supplement RGPIN-2020-04950 and a Tier-II Canada Research Chair CRC-2020-00098 awarded to D.A.S., and an NSERC Discovery Grant RGPIN-2020-05086 awarded to S.J.F.

References

1. D. R. Nemergut, *et al.*, Patterns and Processes of Microbial Community Assembly. *Microbiol. Mol. Biol. Rev.* **77**, 342–356 (2013).

2. S. Evans, J. B. H. Martiny, S. D. Allison, Effects of dispersal and selection on stochastic assembly in microbial communities. *ISME J.* **11**, 176–185 (2017).

3. K. Deiner, *et al.*, Environmental DNA metabarcoding : Transforming how we survey animal and plant communities. 5872–5895 (2017). https://doi.org/10.1111/mec.14350.

4. D. P. Agustinho, *et al.*, Unveiling microbial diversity: harnessing long-read sequencing technology. *Nat. Methods* **21**, 954–966 (2024).

5. S. P. Hubbell, *The Unified Neutral Theory of Biodiversity and Biogeography* (2001).

6. P. L. Wennekes, J. Rosindell, R. S. Etienne, The Neutral-Niche Debate: A Philosophical Perspective. *Acta Biotheor.* **60**, 257–271 (2012).

7. M. Holyoak, M. Loreau, Reconciling empirical ecology with neutral community models. *Ecol. Soc. Am.* **87**, 1370–1377 (2006).

8. J. Rosindell, S. P. Hubbell, F. He, L. J. Harmon, R. S. Etienne, The case for ecological neutral theory. *Trends Ecol. Evol.* **27**, 203–208 (2012).

9. S. R. Zhou, D. Y. Zhang, A nearly neutral model of biodiversity. *Ecology* **89**, 248–258 (2008).

10. S. Woodcock, *et al.*, Neutral assembly of bacterial communities. *FEMS Microbiol. Ecol.* **62**, 171–180 (2007).

11. B. K. Harris, *et al.*, Linking Statistical and Ecological Theory : Hubbell ' s Unified Neutral Theory of Biodiversity as a Hierarchical Dirichlet Process. *Proc. IEEE* **105** (2017).

12. C. Sanchez-cid, *et al.*, Sequencing Depth Has a Stronger Effect than DNA Extraction on Soil Bacterial Richness Discovery. 1–15 (2022).

13. T. P. Curtis, W. T. Sloan, J. W. Scannell, Estimating prokaryotic diversity and its limits. *Proc. Natl. Acad. Sci. U. S. A.* **99**, 10494–10499 (2002).

14. Z. Sam, J. Mei, Stochastic neutral drifts seem prevalent in driving human virome assembly : Neutral , near-neutral and non-neutral theoretic analyses. *Comput. Struct. Biotechnol. J.* **20**, 2029–2041 (2022).

15. W. Zheng, X. Wen, How exogenous influent communities and environmental conditions affect activated sludge communities in the membrane bioreactor of a wastewater treatment plant. *Sci. Total Environ.* **692**, 622–630 (2019).

16. R. Mei, W. T. Liu, Quantifying the contribution of microbial immigration in engineered water systems. *Microbiome* **7**, 1–8 (2019).

17. B. Guo, Z. Sheng, Y. Liu, Evaluation of influent microbial immigration to activated sludge is affected by different-sized community segregation. *npj Clean Water* **4**, 1–5 (2021).

18. A. R. Burns, *et al.*, Contribution of neutral processes to the assembly of gut microbial communities in the zebrafish over host development. *ISME J.* **10**, 655–664 (2016).

19. K. L. Adair, M. Wilson, A. Bost, A. E. Douglas, Microbial community assembly in wild populations of the fruit fly Drosophila melanogaster. *ISME J.* **12**, 959–972 (2018).

20. I. D. Ofiţeru, *et al.*, Combined niche and neutral effects in a microbial wastewater treatment community - supplementary information. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 15345–15350 (2010).

21. Y. Gu, *et al.*, Body size as key trait determining aquatic metacommunity assemblies in benthonic and planktonic habitats of Dongting Lake , China. *Ecol. Indic.* **143**, 109355 (2022).

22. Q. Lu, *et al.*, Multi-group biodiversity distributions and drivers of metacommunity organization along a glacial – fluvial – limnic pathway on the Tibetan plateau. **220** (2023).

23. W. T. Sloan, S. Woodcock, M. Lunn, I. M. Head, T. P. Curtis, Modeling taxa-abundance distributions in microbial communities using environmental sequence data in *Microbial Ecology*, (Springer-Verlag, 2007), pp. 443–455.

24. W. T. Sloan, *et al.*, Quantifying the roles of immigration and chance in shaping prokaryote community structure. *Environ. Microbiol.* **8**, 732–740 (2006).

25. K. Chaloner, G. T. and Duncan, Some properties of the dirichlet-multinomial distribution and its use in prior elicitation. *Commun. Stat. - Theory Methods* **16**, 511–523 (1987).

26. S. Katiraei, *et al.*, Evaluation of Full-Length Versus V4-Region 16S rRNA Sequencing for Phylogenetic Analysis of Mouse Intestinal Microbiota After a Dietary Intervention. *Curr. Microbiol.* **79**, 1–9 (2022).

27. G. Dottorini, *et al.*, Mass-immigration determines the assembly of activated sludge microbial communities. *Proc. Natl. Acad. Sci. U. S. A.* **118** (2021).

28. S. P. Hubbell, *The unified neutral theory of biodiversity and biogeography* (Princeton University Press, 2001).

29. M. A. Leibold, *et al.*, The metacommunity concept: A framework for multi-scale community ecology. *Ecol. Lett.* **7**, 601–613 (2004).

30. C. R. Beeravolu, P. Couteron, R. Pélissier, F. Munoz, Studying ecological communities from a neutral standpoint: A review of models' structure and parameter estimation. *Ecol. Modell.* **220**, 2603–2610 (2009).

31. R. A. Chisholm, S. W. Pacala, Niche and neutral models predict asymptotically equivalent species abundance distributions in high-diversity ecological communities. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 15821–15825 (2010).

32. P. D. Schloss, Rarefaction is currently the best approach to control for uneven sequencing effort in amplicon sequence analyses. *mSphere* **9** (2024).

33. L. Wu, *et al.*, Global diversity and biogeography of bacterial communities in wastewater treatment plants. *Nat. Microbiol.* **4**, 1183–1195 (2019).

34. P. Foladori, L. Bruni, S. Tamburini, G. Ziglio, Direct quantification of bacterial biomass in influent, effluent and activated sludge of wastewater treatment plants by using flow cytometry. *Water Res.* **44**, 3807–3818 (2010).

35. S. J. Fowler, E. Torresi, A. Dechesne, B. F. Smets, Biofilm thickness controls the relative importance of stochastic and deterministic processes in microbial community assembly in moving bed biofilm reactors. *Interface Focus* (2023). https://doi.org/10.1098/rsfs.2012.0069.

36. Y. Mo, *et al.*, Biogeographic patterns of abundant and rare bacterioplankton in three subtropical bays resulting from selective and neutral processes. *ISME J.* **12**, 2198–2210 (2018).

37. I. D. Ofiţeru, *et al.*, Combined niche and neutral effects in a microbial wastewater treatment community. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 15345–15350 (2010).

38. Z. Peng, S. Zhou, D. Zhang, Dispersal and recruitment limitation contribute differently to community assembly. **5**, 89–96 (2012).

39. R Core Team, R: A language and environment for statistical computing. (2022). Available at: https://www.r-project.org/.

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

41. P. J. McMurdie, S. Holmes, Phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. *PLoS One* **8** (2013).