

Linking statistical and ecological theory: Hubbell's unified neutral theory of biodiversity as a hierarchical Dirichlet process 1

Keith Harris¹, Todd L Parsons², Umer Z Ijaz³, Leo Lahti⁴, Ian Holmes⁵, Christopher Quince^{6,*}

1 School of Mathematics and Statistics, University of Sheffield, Sheffield, UK

2 Laboratoire de Probabilités et Modèles Aléatoires, CNRS UMR 7599, UPMC Univ Paris 06, Paris, France

3 Infrastructure and Environment Research Division, School of Engineering, University of Glasgow, Glasgow, G12 8LT, UK

4 Department of Veterinary Biosciences, University of Helsinki, Helsinki, Finland & Laboratory of Microbiology, Wageningen University, Wageningen, Netherlands

5 Department of Bioengineering, University of California, Berkeley, California, USA

6 Warwick Medical School, University of Warwick, Coventry, CV4 7AL, UK

* E-mail: c.quince@warwick.ac.uk

Abstract

Neutral models which assume ecological equivalence between species provide null models for community assembly. In Hubbell's Unified Neutral Theory of Biodiversity (UNTB), many local communities are connected to a single metacommunity through differing immigration rates. Our ability to fit the full multi-site UNTB has hitherto been limited by the lack of a computationally tractable and accurate algorithm. We show that a large class of neutral models with this mainland-island structure but differing local community dynamics converge in the large population limit to the hierarchical Dirichlet process. Using this approximation we developed an efficient Bayesian fitting strategy for the multi-site UNTB. We can also use this approach to distinguish between neutral local community assembly given a non-neutral metacommunity distribution and the full UNTB where the metacommunity too assembles neutrally. We applied this fitting strategy to both tropical trees and a data set comprising 570,851 sequences from 278 human gut microbiomes. The tropical tree data set was consistent with the UNTB but for the human gut neutrality was rejected at the whole community level. However, when we applied the algorithm to gut microbial species within the same taxon at different levels of taxonomic resolution, we found that species abundances within some genera were almost consistent with local community assembly. This was not true at higher taxonomic ranks. This suggests that the gut microbiota is more strongly niche constrained than macroscopic organisms, with different groups adopting different functional roles, but within those groups diversity may at least partially be maintained by neutrality. We also observed a negative correlation between body mass index and immigration rates within the family Ruminococcaceae. This provides a novel interpretation of the impact of obesity on the human microbiome as a relative increase in the importance of local growth versus external immigration within this key group of carbohydrate degrading organisms.

Introduction

A key question in ecology is what maintains species diversity in communities. The classical view is that every species occupies a distinct niche and the species observed in a community are then determined by the niches present. The niche itself is viewed as an n -dimensional hyper-volume in a space of abiotic and biotic environmental variables [1]. If two species occupy the same niche then one will outcompete the other [2]. This viewpoint has been challenged by neutral theory. Neutral models of species abundance combine stochastic population dynamics with the assumption of ecological equivalence between species, formally defined as equivalent forms for all *per capita* demographic rates, e.g., birth and death. Ecological equivalence is assumed to operate between species with a similar functional role deriving from the same

broad functional group or guild of species [3]. The result of the neutrality assumption is that rather than one species always outcompeting another the abundances within the neutral guild fluctuate. The diversity at a single site is then generated as a balance between the immigration of new species and local extinction [4]. In Hubbell’s Unified Neutral Theory of Biodiversity (UNTB) these ideas were extended to multiple sites [5] using a mainland-island structure [6]. The local communities experiencing neutral dynamics are coupled through migration to a metacommunity where neutral dynamics are again assumed but diversity is generated through speciation on a longer time-scale.

The relative importance of niche versus neutral processes in macroscopic organisms is controversial. The first attempts to address this question fitted the UNTB to species abundance distributions (SADs) from a single site and compared model fit to non-neutral alternatives, e.g., log-normal or log-series [7]. The development of Etienne’s genealogical approach, which allowed the calculation of an exact sampling formula or likelihood for a single-site UNTB model [8], was key in allowing the UNTB to be fit efficiently to abundance data [8, 9]. Maximising this likelihood with respect to the model parameters generates a model fit. However, single samples do not provide enough information to reliably fit the UNTB [10] and it has been demonstrated that niche models can generate identical SADs to a single-site neutral model [11]. A more powerful test of the UNTB is to fit a data set from multiple sites simultaneously assuming the same metacommunity but different immigration rates. The genealogical approach has been generalised to multiple sites with identical migration rates [12] but for the fully general case of multiple sites with different immigration rates the resulting sampling formula is computationally intractable for more than a few sites [13]. Instead, an approximate two-stage method has to be used [14–16].

If the importance of neutrality is still an open question for macroscopic organisms then it is even more pertinent for microbes. It is only the recent coupling of molecular methods for characterising species identity with next generation sequencing that has allowed the efficient determination of microbial community structure *in situ* [17]. However, we are now regularly generating data sets comprising hundreds of sites and tens of thousands of sampled individuals per site [18]. In order to accurately fit the multi-site UNTB to these data we developed an alternative to the likelihood based genealogical approach. We are able to show that the UNTB is, in the limit of large population sizes, equivalent to a model from machine learning, the hierarchical Dirichlet process (HDP) [19]. Moreover, our result is more general than the UNTB, as this limit applies irrespective of the exact local community dynamics, provided species are neutral and the total community size is fixed. We can use this result to adapt the existing Bayesian fitting strategy for the HDP to the problem of fitting the UNTB [15].

Using this strategy it is possible to efficiently fit even the largest data sets in a reasonable amount of time with the added advantage of generating full posterior distributions over the parameters rather than just a maximum likelihood prediction. This method also reconstructs the metacommunity distribution enabling us to separate the key question of whether a community appears neutral into two parts. We can generate samples from the full neutral model with our fitted parameters and, as in [12], compare their likelihood with that of the observed samples to test for neutrality, but we can also generate samples given the observed metacommunity and, hence, test for neutral local community assembly alone.

We will validate this method by applying it to twenty-nine tropical tree plots from Panama [20]. We will then use it to determine the extent to which gut microbial communities are neutrally assembled [18]. The human gut is not a closed system, being constantly subjected to immigration events mainly through the diet, hence a metacommunity description is appropriate. However, it is not obvious for microbes at what level we would expect neutrality to operate, as different types of microorganisms perform very different roles. Indeed, there is evidence of clustering of gut microbiota into different enterotypes [21–23], which implies non-neutral structuring at the whole community level. We will address this issue by subdividing the species according to their taxa at multiple taxonomic levels. There is increasing evidence of ecological coherence at higher taxonomic levels for bacteria, with particular taxonomic groupings correlating with broad traits and metabolic functions [24–26]. Thus, even though within a species there may be variability in gene content and the precise niche occupied by strains, e.g. commensal and pathogenic

Escherichia coli [27], at higher levels an ecological signal is preserved [24]. We will test whether this signal leads to species within taxa being distributed neutrally in the human gut.

This is the first time that the full multi-site neutral model has been fit to microbial community data. Earlier studies fitted the proportion of sites that a given species was observed in as a function of its abundance in the metacommunity [28]. However, this approach models local neutral community assembly only, cannot allow for different immigration rates between sites and does not utilise the actual abundances of species, only their presence or absence. Similarly, although [29] showed that the bacterial taxa-abundance distributions in tree-holes scaled across sites in a way that was consistent with the neutral model, they were not fitting to the actual species abundances directly, but rather the shapes of those distributions in individual sites. Recently, an attempt was made to determine the degree of neutrality in human gut microbiota but again by fitting the single-site distribution only [30]. By testing for neutrality at both the local and metacommunity level, and by resolving to different taxonomic groups, we will address the question of what is structuring the newly revealed microbial diversity of the human gut.

Methods

Hubbell’s Unified Neutral Theory of Biodiversity (UNTB)

The UNTB separates the dynamics in the metacommunity from that in the local communities but both are neutral. Assume that there are M local communities indexed $i = 1, \dots, M$ each with a fixed number of N_i individuals. Each iteration of the local community dynamics for site i comprises two steps: choose an individual at random and remove it; with probability m_i migration occurs and this individual is replaced by a randomly chosen member of the metacommunity or with probability $1 - m_i$ it is replaced by a randomly chosen member of local community i . A generation in the model consists of replacing each individual on average once which will require N_i iterations of these two steps. These dynamics will generate a stochastic Markov chain for the abundance of each species [31], which given a sufficiently long time will converge to a stationary, or time-invariant, distribution. In the UNTB it is assumed that the local communities are at this stationary state which we will denote as a vector for each site $\bar{\pi}_i$, with elements $(\pi_{i,1}, \dots, \pi_{i,S})$ giving the probability of observing a particular species at site i . The two parameters m_i and N_i can be conveniently replaced by a single immigration rate $I_i = \frac{m_i}{1-m_i}(N_i - 1)$ [9]. The parameter I_i controls the coupling of the local community to the metacommunity. As $I_i \rightarrow \infty$, the local community stationary distribution will approach the metacommunity distribution and the number of species at that site will increase, while as $I_i \rightarrow 0$, the local community will become dominated by a single species.

In the metacommunity equivalent neutral dynamics operate but with new species generated through speciation with a probability ν . This occurs on a longer time-scale than the local community dynamics so that the metacommunity can be assumed fixed relative to the local communities. Just as in the local communities where I_i is preferred to m_i , it is more convenient to use the speciation rate (or fundamental biodiversity number) to parameterise the metacommunity distribution, $\theta = \frac{\nu}{1-\nu}(N - 1)$ [9], where N is the fixed number of individuals in the metacommunity. The parameter θ can be viewed as the rate at which new individuals are appearing in the metacommunity as a result of speciation. As it increases, the total number of species in the metacommunity also increases and the species abundance distribution becomes increasingly skewed to rare individuals. The final component of the UNTB is to realise that the observed data, the $M \times S$ frequency matrix \mathbf{X} with elements x_{ij} giving the number of times species j is observed at site i , is a sample from the local community [9]. The simplest approach is to assume sampling with replacement so that the multinomial distribution describes the vector of observations at a given site:

$$\bar{X}_i \sim MN(J_i, \bar{\pi}_i), \quad (1)$$

where $J_i = \sum_{j=1}^S x_{ij}$ is the sample size.

In the SI Appendix we show that a wide class of neutral models including the UNTB converge in the large population limit to the same hierarchical Dirichlet process (HDP) approximation. This approximation captures the essential hypothesis of the UNTB – namely neutrality, finite populations, and multiple panmictic geographically isolated populations linked by rare migration – whilst being robust to the specific details of the local community dynamics. Analogous to the relationship between Kingman’s coalescent, Kimura’s diffusion, and the Wright-Fisher model and its many generalisations (*e.g.*, Cannings’ models), we find that under suitable conditions on the higher moments of the individual reproductive output (namely, that when one considers the corresponding genealogical process, the coalescent, mergers of three or more ancestral lines happen with vanishingly small probability as the population size tends to infinity), it is sufficient to introduce local effective population sizes for each deme to accurately approximate many disparate models.

For example, just as Hubbell’s UNTB has population dynamics analogous to the Moran model of population genetics, we could equally well consider a “Wright-Fisher” neutral model, in which all individuals perish at the end of each time step, but each leaves behind a Poisson distributed number of offspring (conditioned on the total population size). Whilst qualitatively different, this model retains the notion of neutrality: each individual is equally likely to be the parent of a randomly chosen individual in the next generation. With an appropriate choice of time rescaling (see Example 2 in the SI), this model also gives rise to the HDP in the large population limit, much as both the Moran and Wright-Fisher models give rise to the same diffusive limits for appropriate choices of effective population size. By contrast, if we consider the highly-skewed reproduction model in which the offspring of one randomly chosen individual replaces all other individuals, we do not obtain the HDP, even though we preserve the neutral hypothesis - as we discuss in the SI (Section 1.2), we require that the offspring distribution is not so fat-tailed that one individual is reasonably likely to be parent to a significant portion of the next generation. In this latter case, there is still a well-defined limit, but it is poorly understood; in particular, there is no known analogue to the Antoniak equation (Equation 6) upon which our approach rests.

It has been shown previously that for large local population sizes, and assuming a fixed finite-dimensional metacommunity distribution with S species present then the local community distribution, $\bar{\pi}_i$, can be approximated by a Dirichlet distribution [28,32]. The parameters of this Dirichlet distribution are proportional to the immigration rate multiplied by the metacommunity distribution:

$$\bar{\pi}_i | I_i, \bar{\beta} \sim Dir(I_i \bar{\beta}), \tag{2}$$

where $\bar{\beta} = (\beta_1, \dots, \beta_S)$ is the relative frequency of each species in the metacommunity. In the SI Appendix (see Section 1.4: Corollary 1), we generalise this to the case where as for the UNTB, there is a potentially infinite number of species that can be observed in the local community. Then the stationary distribution is a Dirichlet process (DP) [33]:

$$\bar{\pi}_i | I_i, \bar{\beta} \sim DP(I_i, \bar{\beta}). \tag{3}$$

The DP can be viewed as an infinite dimensional generalisation of the Dirichlet. It generates an infinite set of samples from the base distribution, which in this case is the metacommunity $\bar{\beta}$, while the concentration parameter, which is I_i here, controls the distribution of weights of those samples. Indeed, these weights are generated by a stick-breaking process (see below) with parameter I_i .

In the metacommunity, a Dirichlet process also applies (SI Appendix: Section 1.5), but now the base distribution is simply a uniform distribution over arbitrary species labels, and the concentration parameter is the biodiversity parameter, θ . This is not a new observation, as it is implicit in the use of Ewens’s sampling formula [34] for the metacommunity in Etienne’s approach [9]. In this case the metacommunity distribution is purely the stick-breaking process. Define an infinite set of random variables drawn from a beta distribution $\{\beta'_k\}_{k=1}^\infty$:

$$\beta'_k \sim Beta(1, \theta). \tag{4}$$

Then we can define the k^{th} element of the metacommunity vector as:

5

$$\beta_k = \beta_k' \cdot \prod_{l=1}^{k-1} (1 - \beta_l'). \quad (5)$$

We will denote this process $\bar{\beta} \sim \text{Stick}(\theta)$. Since the local communities are also DPs the model becomes a hierarchical Dirichlet process (HDP) in the parlance of machine learning [19]. The stick-breaking process is one way to view the DP but an alternative perspective can be obtained by considering successive draws from a DP, which yields the Chinese restaurant process, where each new draw has a probability proportional to the number of individuals already assigned to an existing type (which in our case would be species) of deriving from that type and a probability proportional to θ of deriving from a previously unseen type (or species). From this process the Antoniak equation for the number of types or species S observed following N draws from a DP with concentration parameter θ can be derived:

$$P(S|\theta, N) = s(N, S)\theta^S \frac{\Gamma(\theta)}{\Gamma(\theta + N)} \quad (6)$$

where $s(N, S)$ is the unsigned Stirling number of the first kind [35] and $\Gamma(x)$ denotes the gamma function.

Gibbs sampler for the Neutral-HDP model

Combining the model elements described above, we obtain the complete Neutral-HDP model as:

$$\begin{aligned} \bar{\beta}|\theta &\sim \text{Stick}(\theta), \\ \bar{\pi}_i|I_i, \bar{\beta} &\sim \text{DP}(I_i, \bar{\beta}), \\ \bar{X}_i|\bar{\pi}_i, J_i &\sim \text{MN}(J_i, \bar{\pi}_i). \end{aligned}$$

To this we add gamma hyper-priors for the biodiversity parameter, θ , and the immigration rates, I_i :

$$\theta|\alpha, \zeta \sim \text{Gamma}(\alpha, \zeta), \quad (7)$$

$$I_i|\eta, \kappa \sim \text{Gamma}(\eta, \kappa), \quad (8)$$

where α, ζ, η and κ are all constants.

In any given sample although the potential number of species is infinite we only observe S different types. It is convenient therefore to represent the model in terms of these finite dimensional number of types and one further class corresponding to all unobserved species. We will represent the proportions of the S observed species explicitly as β_k with $k = 1, \dots, S$ and the unrepresented component as $\beta_u = \sum_{k=S+1}^L \beta_k$, in the limit as $L \rightarrow \infty$. In this finite dimensional representation we can determine the species distributions in the local communities:

$$\bar{\pi}_i \sim \text{Dir}(I_i\beta_1, \dots, I_i\beta_S, I_i\beta_u). \quad (9)$$

We can then marginalise the local community distributions and derive the probability of the observed frequencies given the metacommunity distribution $\bar{\beta}$ and the immigration rates $I_i, i = 1, \dots, M$:

$$P(\mathbf{X}|\bar{\beta}, I_1, \dots, I_M) = \prod_{i=1}^M \frac{J_i!}{X_{i1}! \cdots X_{iS}!} \frac{\Gamma(I_i)}{\Gamma(J_i + I_i)} \prod_{j=1}^S \frac{\Gamma(x_{ij} + I_i\beta_j)}{\Gamma(I_i\beta_j)}. \quad (10)$$

The observation that the UNTB is actually a hierarchical Dirichlet process allows us to utilise an efficient Gibbs sampling method to fit it. A Gibbs sampler is a type of Bayesian Markov chain Monte Carlo (MCMC) algorithm. An MCMC algorithm generates samples from the posterior distribution of the parameters given the data [36], which in this case is $P(\theta, I_1, \dots, I_M|\mathbf{X})$. In general, the posterior

is too complex to sample from directly and, in Gibbs sampling, samples are instead generated from the conditional distribution of one parameter given all the others. These full conditionals are often much simpler than the joint posterior distribution, and, crucially, if repeated samples are taken in this way, then they will converge onto the posterior after sufficient iterations. By introducing extra auxiliary variables, it is possible to devise an efficient Gibbs sampler for the UNTB-HDP approximation. One of these auxiliary variables is the metacommunity distribution itself $\bar{\beta}$ and the other is the number of ancestors in site i that gave rise to species j , denoted T_{ij} , i.e., the number of independent immigration events from the metacommunity. Using these variables a Gibbs sampling iteration proceeds as follows:

1. Sample the biodiversity parameter θ from the conditional:

$$P(\theta|S, T) \propto s(T, S)\theta^S \frac{\Gamma(\theta)}{\Gamma(\theta + T)} \text{Gamma}(\theta|\alpha, \zeta), \quad (11)$$

where $T = \sum_{i=1}^M \sum_{j=1}^S T_{ij}$. The first part of the above expression derives from the Antoniak equation (Equation 6) for the number of unique species observed, S , when we sample T ancestors from the metacommunity Dirichlet process with concentration parameter, θ , the second part is simply the prior on θ [35]. To sample from this we use the auxiliary variable approach of [37].

2. Sample the metacommunity distribution:

$$\bar{\beta} = (\beta_1, \beta_2, \dots, \beta_S, \beta_u) \sim \text{Dir}(T_{.1}, T_{.2}, \dots, T_{.S}, \theta), \quad (12)$$

where $T_{.j} = \sum_{i=1}^M T_{ij}$. This exploits the conjugacy between the stick breaking prior for the metacommunity, $\bar{\beta}$, and the likelihood of the ancestor numbers T_{ij} [19].

3. Sample the immigration rates:

$$P(I_i|T_{ij}) \propto \frac{\Gamma(I_i)}{\Gamma(J_i + I_i)} I_i^{T_{i.}} \text{Gamma}(I_i|\eta, \nu). \quad (13)$$

This is again just Antoniak's equation multiplied by the prior but here the number of unique types observed, are the ancestors from the metacommunity, $T_{i.} = \sum_{j=1}^S T_{ij}$, in J_i samples from the local community DP with concentration parameter, I_i .

4. Sample the ancestral states:

$$P(T_{ij}|x_{ij}, I_i, \beta_j) = \frac{\Gamma(I_i\beta_j)}{\Gamma(x_{ij} + I_i\beta_j)} s(x_{ij}, T_{ij})(I_i\beta_j)^{T_{ij}}, \quad (14)$$

where again we recognise the Antoniak equation. This summarises the Gibbs sampling but in SI Appendix 2 we rigorously derive the above conditional distributions.

In general we found that this MCMC procedure quickly converges but to ensure that we were sampling from the stationary distribution we generated either 50,000 Gibbs samples for each fitted data set and discarded the first 25,000 iterations as burn-in or for the human gut microbiota when testing multiple taxa we used 10,000 Gibbs sample and discarded 5,000 iterations as burn-in. The results below are quoted as the median values over these last 25,000 or 5,000 samples with upper and lower credible (Bayesian confidence) limits given by the 2.5% and 97.5% quantiles of these samples.

An MCMC approach was used in an early method to fit the single-site model [8], but it required the use of the more complicated Metropolis-Hastings algorithm, not Gibbs sampling, which is central to the efficiency of our method. In SI Appendix Section 2 we present detailed results demonstrating that on samples generated from the UNTB with known parameters that our method outperforms the

two-stage approximate method of [16], providing accurate and reliable estimates of both θ and I_i except when $I_i \gg \theta$. In this case there is a consistent bias towards under-estimating I_i , which, as we explain in SI Appendix Section 2, is preferable to the large variation in the parameter estimates exhibited by the two-stage approximation. The HDP method also has two further advantages: it generates a full posterior distribution of the model parameters, which provides a realistic estimate of the uncertainty around their point estimates, and it also recovers the metacommunity distribution.

To determine whether an observed data set appears neutral we used a similar Monte Carlo significance test to that in [12]. Given the k^{th} posterior sample of fitted UNTB parameters, $\theta^k, I_1^k, \dots, I_M^k$, an artificial data matrix with the same number of samples M and the same sample sizes J_i as the original data matrix is generated by sampling from the full neutral-HDP, which we will denote by \mathbf{X}_0^k . Given this sample we can also generate a neutral metacommunity distribution, $\bar{\beta}_0^k$, using Equation 12, since the ancestral frequencies $T_{.j} = \sum_{i=1}^M T_{ij}$ are known. This will be a true neutral metacommunity since the distribution will correspond to stick-breaking with parameter θ . Note that the number of species observed can differ from S . We then calculate the likelihood $P(\mathbf{X}_0^k | \bar{\beta}_0^k, I_1^k, \dots, I_M^k)$ using Equation 10. These likelihoods were then compared to the actual likelihood of the observed sample, $P(\mathbf{X} | \bar{\beta}^k, I_1^k, \dots, I_M^k)$, and the proportion that exceeded that value calculated to give a pseudo p-value, denoted p_N , that the data is consistent with the neutral model. In addition, we generated data sets, \mathbf{X}_1^k , with the metacommunity fixed at the model fitted values, $\bar{\beta}^k$. Due to the hierarchical nature of the model, the metacommunity DP only gives a prior on the metacommunity distributions, the observed meta-community can deviate from the neutral expectation. This enables us to test for local neutral community assembly but with a fitted potentially non-neutral metacommunity. We do this in the same way calculating the likelihood for each of the samples, $P(\mathbf{X}_1^k | \bar{\beta}^k, I_1^k, \dots, I_M^k)$, and comparing to $P(\mathbf{X} | \bar{\beta}^k, I_1^k, \dots, I_M^k)$, the proportion of samples with likelihood greater than this forms our pseudo p-value for local neutral community assembly, which we denote by p_L . For both tests, samples were generated either from 2,500 sets of fitted parameters taken from every tenth iteration of the last 25,000 Gibbs samples or from 500 sets of fitted parameters taken from every tenth iteration of the last 5,000 Gibbs samples for the human gut microbiota when testing multiple taxa.

There are many ways in which a distribution could appear non-neutral. A clear example is provided by the situation where communities fall into a finite number of distinct types such that community configurations cluster together. It has been suggested that the human gut microbiome can be clustered into three distinct enterotypes [21–23]. This will appear non-neutral since a single metapopulation distribution will be unable to describe all the community configurations observed. In addition, communities can also appear non-neutral at the level of the observed taxa abundances, if the abundances within individual samples are more or less skewed to rare species than expected for a Dirichlet process then this will appear non-neutral at the local community level. If this occurs for the metacommunity then neutrality will be rejected there too.

Identifying neutral subsets of species

For the microbial community data, we will separate species by their taxa and fit the model to taxa separately in an attempt to identify neutral subsets. The validity of this approach rests on two observations. Firstly, that if there are multiple neutral guilds of species in a community, where the abundance of a guild varies from site to site in a non-neutral fashion, then the community as a whole will appear non-neutral but if we just sample species from one guild then the neutral patterns will be recovered [38]. This is self-evident. The second observation is that if only a subset of the species in a neutral guild are sampled, then that subset will still fluctuate neutrally but with renormalised probabilities. This derives from the following property of the Dirichlet distribution, that if only a subset of the S dimensions are observed, say U , then that subset is still distributed as a Dirichlet on the reduced space with the same parameters. For the neutral model the result is that the biodiversity parameter is unchanged but that the immigration rate at each site is reduced, $I_i^U = I_i(1 - \sum_{i \notin U} \beta_i)$, according to the weight of the missing species in the

metacommunity. The result is that if at some level of taxonomic resolution all species are from the same neutral guild, if not necessarily representing all that guild, then they will still be identified as neutral.

The key ideas used in the above derivations are summarised in Table 1.

Data

Neutral simulation

In SI Appendix Section 2 we show that the UNTB-HDP fitting method accurately determines the parameters of data sets generated from the UNTB. To provide a further test of the model fitting from a sample that relaxes the mainland-island structure of the UNTB but maintains the assumption of neutrality we performed a neutral model simulation. This comprised 50 sites indexed $i = 1, \dots, 50$, with a fixed population number of $N_i = 20,000$ individuals per site. Discrete dynamics were used with a probability that an individual was removed at each iteration of 5%. Deleted individuals were then replaced, with speciation probability $\nu = 10^{-5}$ by an entirely new species, by an individual chosen at random from the local community in the previous iteration with probability $(1 - \nu)(1 - m_i)$, or by an individual chosen at random from all the other sites with probability $(1 - \nu)m_i$. The migration probability was varied across sites according to the rule $m_i = i \times 10^{-4}$, so that the immigration rate, $I_i = m_i N_i = 2i$, varied from 2 to 100. The model was run for 2,000 generations, i.e., 40,000 iterations, at which point the species number appeared stationary, then 1,000 individuals were sampled with replacement from each site. The UNTB-HDP model was fit by Gibbs sampling to this data set as was the two-stage approximate method of [16]. This simulation although it has strictly neutral dynamics does not correspond exactly to Hubbell's UNTB because rather than an explicit mainland-island structure with diversity only generated in the metapopulation, it has speciation occurring in the local populations themselves, with a metapopulation which is an implicit aggregate of the local populations rather than an explicit distribution.

Tropical trees from Panama

To provide a well-distributed sample of tropical trees at a regional level we took twenty-nine of the one hectare forest plots considered in [20]. These comprised all the one hectare samples from the Panama region with an elevation of less than 200 metres. This restriction ensured that all samples were from the same environment of lowland tropical forest. We also did not use data from the three larger Panama plots in order to maintain an even sampling at the regional level. Within each plot all trees ≥ 10 cm in diameter were censused and their morpho-species recorded. The network of sample sites was spread across a 15×50 km region along the Panama canal, see [39] for details. A total of 13,263 trees were sampled from 367 species. The number of individuals observed in each plot ranged from a minimum of 302 to 647 with a median of 450. The UNTB-HDP model was fit to this data as described above.

Human gut microbiota

To compare with the tropical tree analysis we also fitted the UNTB-HDP model to a study of the gut microbiomes of twins and their mothers [18]. These comprised fecal samples from 154 different individuals characterised by family and body mass index (BMI). Each individual was sampled at two time points approximately two months apart. The V2 hypervariable region of the 16S rRNA gene was amplified by PCR and then sequenced using 454. We reprocessed this data set filtering the reads, denoising and removing chimeras using the AmpliconNoise pipeline [40,41]. This gave a total of 570,851 reads split over 278 samples, since out of the 308 collected samples thirty failed to possess any reads following filtering. The size of individual samples varied from just 53 to 10,580 with a median of 1,598. The number of unique sequences remaining following noise removal was 19,647. These were then taxonomically classified using the RDP stand-alone classifier of [42]. We constructed Operational Taxonomic Units (OTUs) at

3% sequence difference using average linkage clustering to approximate species [43]. This was done for the entire data set generating 7,238 OTUs. We fitted the UNTB-HDP model to this data set.

To explore the impact of sample size and number on the ability of our pseudo p-values to correctly identify a community as non-neutral at the local and metacommunity levels we generated a series of subsampled data sets from this study. First, we selected at random without replacement either 20, 50, 100 or 200 samples from all those that had 1,000 reads or greater (247 in total). Then we generated a series of data sets where we sampled increasing numbers of individuals or reads from these selected samples, from 20 individuals per sample to 400 inclusive in increments of 20. We used sampling with replacement i.e. multinomial sampling so that expected OTU proportions were equal to those in the observed communities. For each number of samples and number of reads we generated ten replicate communities. We then fitted the UNTB-HDP model to these communities and tested for neutrality at the local and metapopulation level.

Starting with the full data set, we split the unique sequences according to the phylum to which they were classified, using a cut-off of 70% bootstrap confidence. OTUs were then reconstructed at 3% for each phylum and the UNTB-HDP fit to each phylum separately. We repeated this process for family and genus too. Only samples that had more than 150 representatives from a taxa were included in the analysis and the model was only fit to taxa that had at least 50 samples satisfying this criterion. This ensured a sufficiently large data set for parameters to be inferred and if a taxa dominates a neutral guild occupying a particular role we would expect it to appear in a large proportion of samples. We also generated ten replicate data sets from the full data set with the same number of samples and same number of reads per sample as the data sets split by taxa at each level. Applying the UNTB-HDP to these then gives us an equivalent bench-mark for the effect of subsampling on our ability to detect non-neutrality. We also did this for the tropical tree data.

Results

Neutral simulation

In Figure 1, we give the immigration rates estimated by the UNTB-HDP fitting algorithm for the neutral simulation. From this single sample we are able to accurately predict the immigration rates across all the sites. The uncertainty in our predictions increases for higher I_i but there is no consistent bias. In contrast, the two-stage approximation substantially underestimates the immigration rate as I_i increases. This is most likely because although the simulation appears locally neutral ($p_L = 0.57$) as we would expect, the hypothesis that the neutral model applies at the metacommunity level too is rejected, $p_N = 0.0096$. The deviation from the mainland-island structure and the occurrence of speciation within the islands themselves results in a metacommunity distribution that deviates from the neutral stick-breaking process. This illustrates that in contrast to the two-stage approximation the UNTB-HDP model can still correctly predict immigration rates when neutral community assembly operates only at the local community level.

Tropical Trees from Panama

By fitting the UNTB-HDP model to the twenty-nine tropical tree communities we found that they have a distribution of abundances across sites that is consistent with the neutral model at both the metacommunity and local community levels, $p_N = 0.81$ and $p_L = 0.23$. The median fitted θ obtained was 109.3. The median fitted immigration rates varied across sites from 20.69 to 76.93 with a median of 41.7. In Figure 2, we use non-metric multi-dimensional scaling (NMDS) to position each community in two-dimensions in such a way as to preserve Bray-Curtis distances between communities. This was done using the metaMDS function of the vegan package in R [44]. The fitted metacommunity distribution is also shown in this plot. The sites are represented as bubbles with size proportional to their fitted

immigration rates and contours calculated using the `ordisurf` function. From this it is apparent that the communities with higher I_i are in general more similar to the metacommunity. The fitted immigration rates are also related to the spatial location of the sites. Although there is no spatial location associated with the metacommunity, if we assign it to the location of the site with the highest I_i , site 14, and calculate the distance from this site to each of the others, then we find a significant negative correlation ($p = 0.03$) between distance and immigration rate.

Human gut microbiota

In contrast to the tropical trees, the human gut samples do not appear neutral at the whole community level, $p_N = 0$ and $p_L = 0$. This was not purely an effect of the tropical trees comprising a data set of fewer samples and fewer individuals. Reducing the gut data set to an equivalent number of samples (29) with the same sizes we would still always reject neutrality at the metacommunity level, at the local level we observed a median p_L of 0.062 across the ten replicates. We would falsely fail to reject neutrality therefore but not as strongly as for the real tree data ($p_L = 0.23$). Therefore, we can conclude that the human gut is convincingly less neutral than tropical trees even accounting for the different sample numbers and sizes.

In Figure 3 we show the impact of sample number and sample size on the pseudo p-values for the test of neutrality for whole community and local community assembly. With sufficient samples (i.e. at least 200) we have power to reject neutrality at both levels provided the sample size exceeds 150 but as sample number decreases our power to correctly reject neutrality particularly for local community assembly decreases.

The results of subdividing the OTUs at different taxonomic levels and fitting the UNTB-HDP model are given in a nested format in Table 2. The families associated with each phylum are indented below as are the genera in each family. We see some evidence that as we move down the taxonomic hierarchy from phyla, through families to genera, the subdivided communities appear more consistent with neutral local community assembly. We would reject local neutrality for both major phyla found in the human gut, the Bacteroides and Firmicutes, but there are two families out of four for which we cannot confidently reject neutral local community assembly at the 1% level, the Bacteroidaceae and Incertae Sedis XIV, with $p_L = 0.03$ and 0.05, respectively. At the level of genera, two out of three appear close to neutral at the local level, the exception being the *Faecalibacterium*. This is not the case when we do not use the fitted metacommunities and instead test for both neutral local community assembly and a neutral metacommunity. Then for all data sets we would completely reject neutrality. The figures in parentheses give pseudo p-values for the equivalent complete data set randomly sampled down to the same size as the taxa. This gives us a benchmark to verify that these effects are not purely due to small sample sizes. From these we see that in all cases the probability of incorrectly concluding that the subsampled data set is neutral is less than 1%.

To quantify how the metacommunity deviates from the neutral assumption for those data sets that appear locally neutral we compared the fitted metacommunities averaged over 500 Gibbs samples with the metacommunity observed in samples from the full neutral model with the equivalent parameters. These two distributions are shown in Figure 4 for the three genera, Bacteroides, Blautia and *Faecalibacterium*. These distributions are shown as rank-abundance plots with the OTUs ordered in terms of the relative frequency with that frequency given on the y-axis, which is log-scaled. It is clear that the fitted metacommunities from the three genera all have a small number of highly abundant OTUs and then a long tail of rare OTUs. The neutral model cannot fit a metacommunity of this shape.

We also looked for correlations between the fitted immigration rates for the different taxa and the body mass index of subjects. No significant relationships were found at the genus level but for the family Ruminococcaceae a significant negative relationship was observed (p -value = 0.014 see Figure 5). The same negative correlation was also observed for their parent phylum the Firmicutes but it was slightly stronger (p -value = 0.007).

The results clearly demonstrate the usefulness of the UNTB-HDP Gibbs sampler, its ability to fit large multi-sample data sets, and its robustness to deviations of the metacommunity from neutrality and the ability to detect those deviations whilst still correctly inferring immigration rates. The resulting significance tests and fitted parameters reveal a great deal about the ecology of the human gut microbiota in comparison to macroscopic organisms such as the tropical trees. The human gut is clearly much more strongly structured by functional niches. Only at the genus level do we see some evidence of neutral local community assembly in the gut, whilst tropical trees were well described by the neutral model without any subdivision of species. In some ways, this is to be expected, given the multiplicity of metabolic roles performed by the human microbiota we would not expect ecological equivalence at the whole community level. However, the borderline neutral patterns we did observe suggest the possibility that neutral local community assembly may be operating within the species occupying those roles, and that neutral processes may be responsible for maintaining some of the vast diversity that is observed in the human gut. This has to be a tentative conclusion as pattern does not imply process [10], but, regardless, the fact the observed abundances are consistent with the neutral model means that its importance for explaining fine-scale gut microbial diversity cannot be ruled out.

It is important to address the question of whether the tests have the power necessary to detect non-neutrality. It is clear from Figure 3 that as the number of samples in particular decreases it becomes hard to detect non-neutral distributions — this is actually a strong motivation for the use of the UNTB-HDP which can be efficiently fit in the multi-site case. However, our benchmarking against the full gut data set allows us to conclude that some genera and the tropical trees appear more neutral than the equivalent sized complete gut microbiome. It is also important to note that the model was unable to detect the spatial signature in the tropical tree data as a deviation from neutrality. In the absence of that spatial information we would have included that a spatially inhomogenous metapopulation was sufficient to explain these patterns. That certainly motivates inference strategies for spatially explicit neutral models [45].

It is highly significant that the metacommunity distributions could not be explained by the neutral process for any taxa. Instead, the metacommunity was dominated by a small number of very abundant OTUs, with in all cases the most abundant OTU possessing a relative abundance exceeding 10% of the metacommunity. This may be a signature of non-neutral processes. The dominant OTUs may have a competitive advantage, or interactions with bacteriophages [46] or the host immune system may be structuring these distributions [47], and that is skewing their apparent metacommunity abundance, or it may genuinely reflect the abundance of these organisms in the metacommunity perhaps coupled with an improved dispersal ability over their competitors.

The parameters of the fitted models, in particular, the immigration rates, are also highly informative. For the Panamanian tree data set we showed that these correlated with spatial location of the sites. A strong effect of distance on community similarity was found in the original study and a spatially explicit version of the neutral model was fit to the data [20], but we have shown that even in the UNTB where space is only implicit, this signal can be recovered from the fitted immigration rates. For the gut microbiota samples, we have no spatial position, but here, remarkably, the immigration rates for the family Ruminococcaceae and phylum Firmicutes correlated negatively with body mass index. This provides an unique interpretation of the impact of obesity on the human gut microbiota: an increase in the rate of input of nutrients to the gut effectively results in an increase in microbial growth rates in the key carbohydrate metabolising group the Ruminococcaceae [48] and these equate to a decrease in immigration rate relative to local birth.

It is also instructive to compare immigration rates between fitted models. There has been debate as to the importance of dispersal on microbial community structure, the theory that “everything is everywhere, but the environment selects” [49]. However, comparing the tropical tree fits with the gut microbiota at the phylum level we find that the predicted immigration rates are comparable, implying that dispersal

limitation may be just as important between human guts as it is between tropical forests. Interesting¹² patterns also appear comparing immigration rates between gut taxa. They are much lower, for example, for the Bacteroides than the Firmicutes, probably reflecting the much higher tendency for the latter to be spore-forming.

Finally, whilst these results are of great interest in themselves, perhaps our most significant achievement is formally linking a model from ecology, the Unified Neutral Theory of Biodiversity, with a model from machine learning, the hierarchical Dirichlet process. In addition, by showing that the details of the local community dynamics are irrelevant for the HDP approximation to hold, provided the neutrality assumption is met, we may explain why we were able to fit communities as different as tropical trees and the gut microbiota. This strongly motivates the HDP as an ecological null model. What is more the mathematical structure of the HDP is easily extendable to for example, niche-neutral models or further hierarchical levels. Therefore, we believe that the connection we have made here will lead to an explosion of hierarchical Bayesian modelling in community ecology.

Software for fitting the UNTB-HDP can be downloaded from:
<https://github.com/microbiome/NMGS>.

Acknowledgments

CQ is funded through an EPSRC Career Acceleration Fellowship EP/H003851/1. KH was funded through a Unilever research grant whilst conducting this research. LL is funded by the Academy of Finland (decision 256950). TLP is funded by the Fondation Sciences Mathématiques de Paris. We thank three anonymous reviewers for constructive comments.

Figures

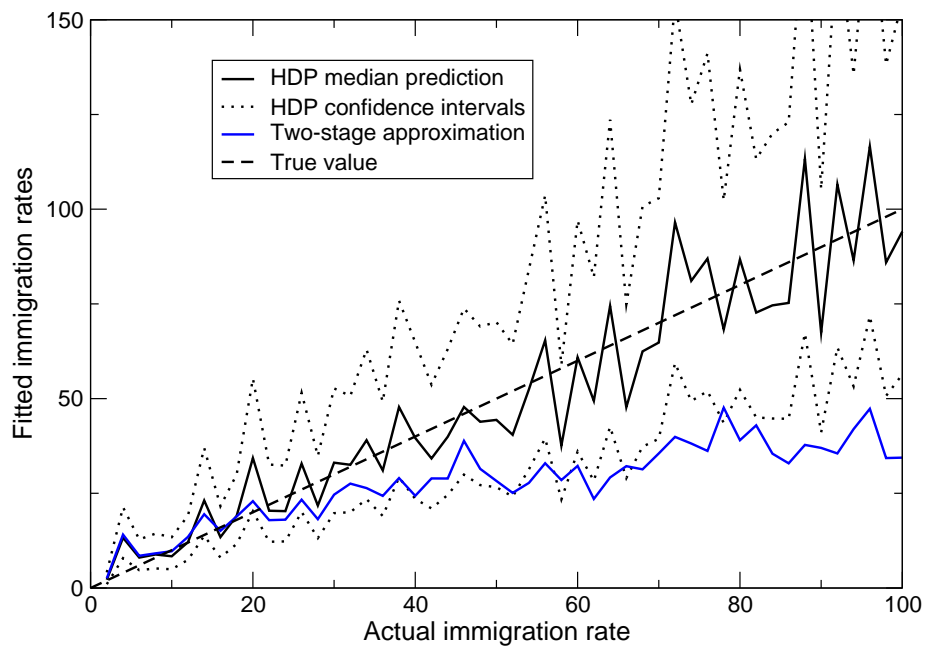


Figure 1. Estimated immigration rates vs. true values for the UNTB-HDP model fit to a neutral model simulation. Predictions are medians (solid line) from 25,000 posterior samples together with lower (2.5%) and upper (97.5%) Bayesian confidence intervals (dotted lines). The predictions from the two-stage approximation are also given (blue line).

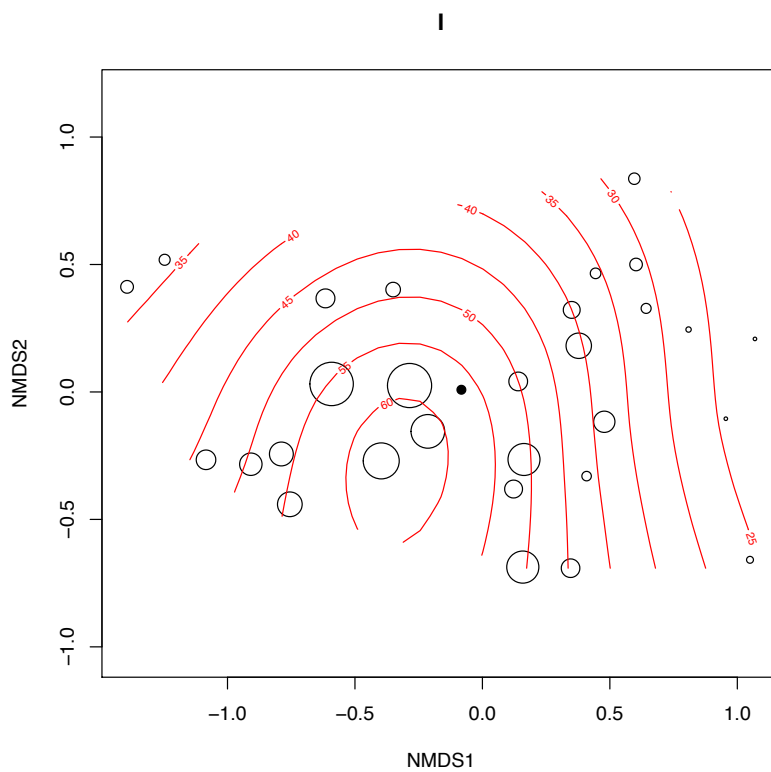


Figure 2. An NMDS plot of the twenty-nine Panama tropical tree communities. Communities are visualised as bubbles with size proportional to the median I_i values obtained from the UNTB-HDP Gibbs sampler. Contours calculated using the `ordisurf` function of the R `vegan` package are also shown. The metacommunity distribution is denoted by a solid black point.

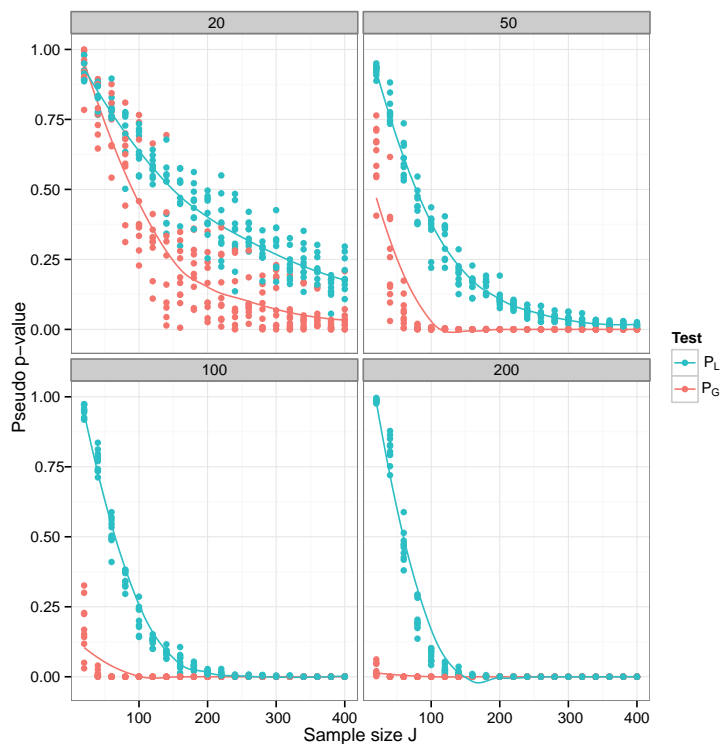


Figure 3. Impact of sample number and size on detection of non-neutrality in the human gut data. The figures show the pseudo p-values for neutrality for both the complete neutral model (P_G) and local community assembly (P_L). We generated ten replicate communities by sampling without replacement either 20, 50, 100 or 200 samples from those that had 1,000 reads or greater (247 in total) and from the selected samples we generated a fixed number of reads sampling with replacement. We increased read numbers from 20 individuals per sample to 400 inclusive in increments of 20. We then tested the subsampled communities for neutrality.

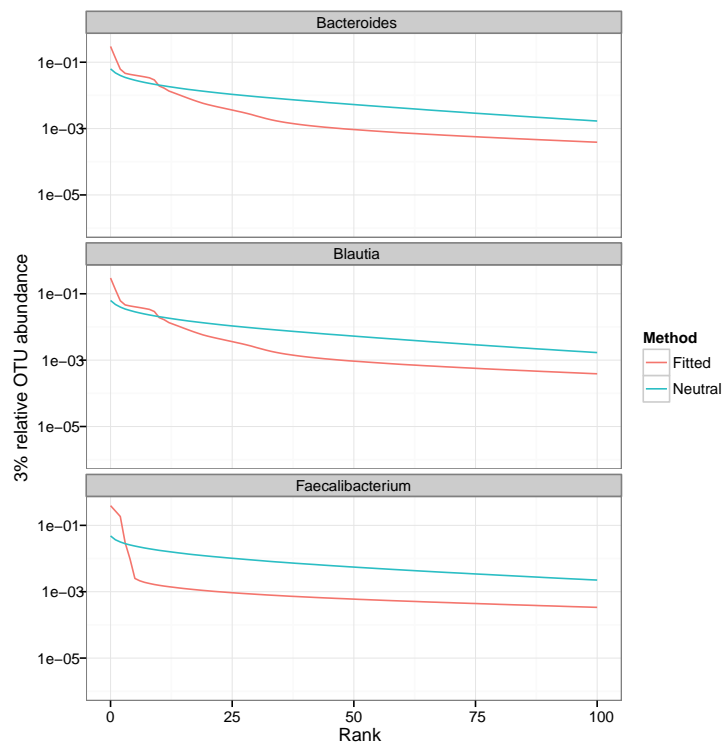


Figure 4. Human gut metacommunity distributions. The fitted metacommunity distributions (red line) and neutral metacommunity predictions (blue line) as rank-abundance curves for three genera: Bacteroides, Blautia, and Faecalibacterium.

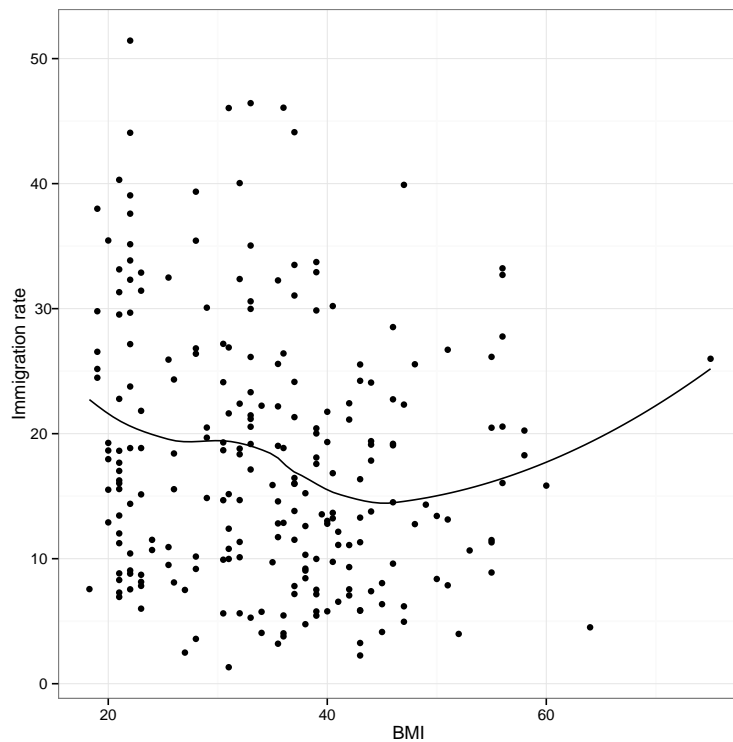


Figure 5. Immigration rate vs. BMI. Median immigration rate for the family Ruminococcaceae determined by the UNTB-HDP model plotted against body mass index. A significant negative correlation is observed (p-value = 0.014 - Pearson's correlation).

Table 1. Key ideas used in this paper.

Neutral model	A population model in which all types are <i>functionally</i> equivalent
Unified Neutral Theory of Biodiversity (UNTB)	A discrete time stochastic model of an island-mainland metacommunity proposed by Stephen Hubbell [5]. At each time step, one individual on the island dies, and is either replaced by the offspring of a randomly chosen individual on the island, or, with fixed probability, by the offspring of an individual chosen at random from the mainland.
Chinese Restaurant Process (CRP)	A discrete time stochastic model proposed by Davis Aldous [50] in which he imagines a Chinese restaurant with an unlimited number of tables. At each time step, a new customer arrives, who will either choose a new table with a fixed probability θ , or sit at an already occupied table with probability proportional to the number of individuals already seated at that table. It is mathematically equivalent to Hoppe's urn [51], which generates samples from a Kingman coalescent with neutral mutations that occur at a fixed rate, and which always give rise to a new allelic type.
Dirichlet Process (DP)	A random variable taking value in the set of discrete probability distributions on a set \mathcal{X} , obtained by drawing random points in \mathcal{X} according to a given probability measure μ , and assigning these to the tables in a stationary Chinese Restaurant Process (thus, there are infinitely many customers seated at infinitely many tables), so that the probability of drawing a given point is equal to the proportion of customers seated at the corresponding table.
Hierarchical Dirichlet Process (HDP)	A Dirichlet Process for which the underlying measure μ is itself an instance of a Dirichlet Process.

Tables

Table 2. Fitting the UNTB-HDP model to human gut microbiota.

Taxa	N	S	\tilde{J}	θ	I_i			p_N	p_L
					l	m	u		
Bacteroidetes	231	569	596	148.6	1.5	5.5	13.7	0.0 (0.0)	0.0 (0.0)
Bacteroidaceae	208	224	506	51.4	0.7	3.3	7.6	0.0 (0.0)	0.03 (0.0)
Bacteroides	208	224	506	51.4	0.7	3.3	7.6	0.0 (0.0)	0.03 (0.0)
Firmicutes	277	4770	1009	1382.3	21.4	44.8	81.0	0.0 (0.0)	0.0 (0.0)
Incertae Sedis XIV	87	176	264	39.2	1.7	9.8	27.5	0.0 (0.0)	0.05 (0.004)
Blautia	87	175	264	38.9	1.6	10.1	27.1	0.0 (0.0)	0.06 (0.003)
Lachnospiraceae	164	873	248	262.9	6.5	13.0	21.2	0.0 (0.0)	0.0 (0.0)
Ruminococcaceae	239	1471	409	411.0	4.5	16.1	38.1	0.0 (0.0)	0.0 (0.0)
Faecalibacterium	141	301	297	71.7	1.0	7.5	21.4	0.0 (0.0)	0.004 (0.0)

Results are given for 3% OTUs at different levels, quantities given in the table are: N - the no. of samples with > 150 reads; S - the number of 3% OTUs; \tilde{J} - the median sample size; θ - the fitted biodiversity parameter; I_i - the fitted immigration rates where l, m and u are the lower 2.5%, median and upper 97.5% quantiles respectively; p_N - the proportion of simulated neutral samples exceeding the observed data likelihood; and p_L - the proportion of simulated locally neutral samples exceeding the observed data likelihood. The figures in parantheses give pseudo p-values for the equivalent complete gut microbiome data set randomly sampled down to the same size as the individual taxa.

1 Large Population Limits for a Neutral Metacommunity

1.1 Summary and Outline

Given the length and technical nature of this supplement, we will begin with a summary that outlines the results herein. Our intent is to formulate a class of models that generalize Hubbell’s formulation of the Unified Neutral Theory of Biodiversity and Biogeography (UNTB) and a number of variants that have appeared in the community ecology literature, whilst retaining the essential feature of neutrality. Our inspiration in this are Cannings’ models [52], which have become the standard in theoretical population genetics. We discuss coalescent theory and these models in detail below, but in brief, a Cannings’ model allows any reproduction law with discrete generations that keeps the total population size fixed, provided that relabeling the parents leaves the distribution of offspring unchanged. More generally, we could consider models replacing fixed population sizes with density dependent population dynamics, as in [53], [54, 55] and [56], but this would have further lengthened and complicated this supplement.

We formulate a mainland-island Cannings’ model, in which the mainland has size $N_0 = N$ and the islands have size N_i that grow with N , but are approximately equal. We allow migration between any pair of island and mainland, and further allows mutations to give rise to new types on both island and mainland. After collecting a few results regarding the reproduction law for a Cannings’ model, we show in Section 1.4, provided that:

- the islands are asymptotically smaller than the mainland (in both census and effective population size; see the discussion below),
- migration between demes is rare (we assume that the probability that a migrant arrives in a local community is inversely proportional to the size of that community), and
- the probability of multiple mergers is asymptotically smaller in N than the rate of pairwise coalescence,

then Proposition 1 shows that if we rescale time proportionally to the effective population size of the islands (*i.e.*, we measure time so that one time step corresponds to N_e generations) for large values of N , the population dynamics on the islands converge to the dynamics of Moran’s infinitely many alleles model, with the migration rate from the mainland taking the place of the mutation rate in the population genetic model, and such that the type of all new mutants/migrants is drawn from the initial type distribution for the mainland (*i.e.*, the probability of migration between islands or novel mutations appearing on an island becomes vanishingly small as N grows large, and can be completely ignored in the limit), and moreover, the composition of the mainland remains constant on this timescale - the dynamics are sufficiently slow that one cannot see changes when time is scaled according to the effective population size of the islands. Moreover, this limit is independent of the specific reproduction law for the islands, provided it satisfies Cannings’ conditions - indeed, we don’t even need to assume the same law between islands. As a consequence of the identification of the islands’ dynamics as a variation of the infinitely many alleles model, we can use previous results from theoretical population genetics to conclude that the stationary distribution for the islands is a Dirichlet Process, and that the composition of a sample is distributed according to Ewens’ sampling formula.

In Section 1.5 we turn our attention to the mainland. We first observe that for large values of time, the species distribution on the islands converge onto stationary processes governed by the Dirichlet process above. We can then apply this with results from [57] to obtain Proposition 3, which tells us that we need to rescale time according to the effective population size of the mainland (again, so that one time step

corresponds to N_e generations, but now N_e for the mainland, which is substantially larger). On this slow²¹ scale, the islands will essentially instantaneously arrive at their stationary state (an instant in this “slow” time scale is in fact an extremely long time in the natural “intermediate” time scale for the islands), whilst now the population on mainland follows the “real” infinitely many alleles model (with the actual mutation rate), and again, migrations from an island to the mainland become vanishingly rare as N becomes large, and, as before, the stationary distribution is again a Dirichlet process, where each newly appearing genotype is assigned a label chosen uniformly at random from $[0, 1]$ (thus the probability of two distinct mutations giving rise to the same type is 0). In particular, the islands have the Hierarchical Dirichlet Process for their stationary distribution: they are Dirichlet Processes in which the types are drawn from the underlying Dirichlet Process that describes the mainland.

1.2 A Mainland-Island “Cannings’ Model”

We begin by formulating a broad class of haploid models that includes Hubbell’s Unified Neutral Theory of Biodiversity and Biogeography (UNTB) [5]. Our inspiration are Cannings’ population genetic models [52], which use exchangeability as a general mathematical formulation of neutrality: random variables ν_1, \dots, ν_N are *exchangeable* if the random vectors $(\nu_{\pi(1)}, \dots, \nu_{\pi(N)})$ are equal in distribution for all permutations π of $\{1, \dots, N\}$. Informally, the labels $1, \dots, N$ are arbitrary, and can be changed without essentially changing the process. In a Cannings’ model, one assumes a fixed population of size N and discrete generations; $\nu_i(n)$ is the number of offspring in the $n + 1^{\text{st}}$ generation of the i^{th} individual of the n^{th} generation. (ν_1, \dots, ν_N) is assumed to be exchangeable and must satisfy

$$\sum_{i=1}^N \nu_i = N.$$

Under suitable conditions on the higher moments (*n.b.*, as a consequence of exchangeability, we must have $\mathbb{E}[\nu_i] = 1$ for all i), one can show [58] that as $N \rightarrow \infty$ the frequency of types (here, the type of an individual is inherited from its ancestor in the initial population) and the genealogical process converge to the Wright-Fisher diffusion and Kingman’s coalescent, respectively (relaxing the moment conditions leads to a Λ -coalescent limit for the genealogical process). In particular, if $X_i^{(N)}(n)$ is the number of descendants alive in the n^{th} generation of the i^{th} ancestral individual in the 0^{th} generation, and c_{N_i} is the coalescence probability, *i.e.*, the probability two individuals sampled without replacement from deme i have the same parent,

$$c_N := \frac{\mathbb{E}[(\nu_1)_2]}{N-1},$$

where

$$(x)_k := x(x-1)\cdots(x-k+1)$$

is the *falling factorial* or *Pochhammer symbol*. Then, [58] shows that

$$\lim_{N \rightarrow \infty} \frac{\mathbb{E}[(\nu_1)_3]}{N\mathbb{E}[(\nu_1)_2]} = 0$$

is a necessary and sufficient condition for $X_i^{(N)}(\lfloor c_N^{-1}t \rfloor)$ to converge weakly¹ as $N \rightarrow \infty$ to a Wright-Fisher²² diffusion, *i.e.*, to a diffusion process with probability density

$$p(\mathbf{y}, t | \mathbf{x}) := \mathbb{P} \{ \mathbf{X}(t) \in \mathbf{y} + d\mathbf{y} | \mathbf{X}(0) = \mathbf{x} \}$$

satisfying the Kolmogorov backward equation

$$\frac{\partial p}{\partial t} = \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N x_i (\delta_{ij} - x_j) \frac{\partial^2 p}{\partial x_i \partial x_j}.$$

The quantity c_N^{-1} has been referred to as the coalescent effective population size, and can be shown to generalize previously defined notions of an effective population size [59].

Here, we take our cues from the discussion of infinite-alleles models in [57], which we will closely follow, in formulating a ‘‘Cannings’ UNTB’’ with migration and mutation. As in previous models, we will assume a mainland, which supports a population of size $N_0 = N$, together with a collection of islands labelled $i = 1, \dots, M$ which support populations of size N_i . We will assume that the islands are all approximately the same size, and substantially smaller than the mainland; for Section 1.4, we will require $N_i \ll N_0^2$, whereas we will need to impose sharper estimates of the relative sizes in Section 1.5. In what follows, we will refer to the mainland and each of the islands as having N_0 or N_i niches respectively, we will use the term deme when we are referring to a local community that can be either an island or the mainland, and will refer to *e.g.*, the individual in the j^{th} niche in the i^{th} deme.

We will assume discrete generations, and that at each time step the current residents reproduce and are replaced by their offspring. The j^{th} individual has $\nu_{ij}^{(N)}$ offspring so that

$$\sum_{j=1}^{N_i} \nu_{ij} = N_i,$$

and model neutrality by assuming that each random vector $(\nu_{i1}^{(N)}(n), \dots, \nu_{iN_i}^{(N)}(n))$ is exchangeable. We further assume that $(\nu_{i1}^{(N)}(n), \dots, \nu_{iN_i}^{(N)}(n))$ is independent of $(\nu_{j1}^{(N)}(m), \dots, \nu_{jN_j}^{(N)}(m))$ unless $i = j$ and all $m = n$. Following [58], we define

$$c_{N_i} := \frac{\mathbb{E}[(\nu_{i1})_2]}{N_i - 1},$$

¹A family of random variables $\{X^{(N)}\}$ taking values in a space S is said to *converge weakly* to X if

$$\lim_{N \rightarrow \infty} \mathbb{E}[f(X^{(N)})] = \mathbb{E}[f(X)]$$

for all $f \in C(S)$; the values $\mathbb{E}[f(X)]$ completely characterize the distribution of X . Weak convergence is denoted by

$$X^{(N)} \Rightarrow X.$$

²We will write $a_N = o(b_N)$, or $a_N \ll b_N$, if

$$\lim_{N \rightarrow \infty} \frac{a_N}{b_N} = 0,$$

and use $a_N \asymp_N b_N$ to indicate that

$$\lim_{N \rightarrow \infty} \frac{a_N}{b_N} = 1.$$

We will also write $a_N = \mathcal{O}(b_N)$ if there exists a constant C such that

$$a_N \leq C b_N,$$

for all N .

for $i = 0, \dots, M$, and assume the analogue of Möhle's condition:

$$\lim_{N_i \rightarrow \infty} \frac{\mathbb{E}[(\nu_{i1})_3]}{N_i \mathbb{E}[(\nu_{i1})_2]} = 0, \quad (15)$$

which has the following consequence [60]:

Lemma 1. *Assume (15). Then,*

$$\lim_{N_i \rightarrow \infty} c_{N_i} = 0,$$

and

$$\lim_{N_i \rightarrow \infty} \frac{\mathbb{E}[(\nu_1)_2(\nu_2)_2]}{c_{N_i}} = 0.$$

We will further assume that there exists a_N such that

$$\lim_{N \rightarrow \infty} \frac{c_{N_i}}{a_N} = \begin{cases} \gamma_i & \text{if } i > 0, \text{ and} \\ 0 & \text{otherwise.} \end{cases} \quad (16)$$

which formalises the notion that the populations on the islands are all of the same order of magnitude (their effective population sizes are asymptotically proportional $c_{N_i} \sim \gamma_i a_N$) and asymptotically smaller than the mainland ($a_N \ll c_N$).

We will further assume that each individual has a type, which is a label in $[0, 1]$, which we think of as a probability space with the uniform (Lebesgue) measure λ . The labels are more of a mathematical convenience for tracking ancestries, and have no effect on fitness, so we could equally well take labels in any compact Polish space \mathfrak{X} that is equipped with a probability measure $\gamma(dx)$. We write $X_{ij}(n) \in [0, 1]$ for the type of the individual in the j^{th} niche of the i^{th} deme in generation n – the labels are inherited from the parent, except when an individual is subject to mutation at birth. We discuss the processes of reproduction and mutation below. The state of the i^{th} deme in the n^{th} generation is conveniently represented by an atomic probability measure on $[0, 1]$,

$$G_i^{(N)}(n) = \frac{1}{N_i} \sum_{j=1}^{N_i} \delta_{X_{ij}(n)},$$

where $\delta_{X_{ij}(t)}$ is the Dirac point mass at $X_{ij}(t)$, and the superscript (N) emphasizes the dependence on the “system size” N , *i.e.*, for any subset $A \subseteq [0, 1]$, $G_i^{(N)}(n)(A)$ is the number of individuals in the i^{th} deme with a type in the set A . We write $\mathbf{G}^{(N)}(n) = G_0^{(N)}(n) \otimes \dots \otimes G_M^{(N)}(n)$ for the product measure,

$$\mathbf{G}^{(N)}(n)(A) = G_0^{(N)}(n)(A) \dots G_M^{(N)}(n)(A).$$

Given a measure μ and a continuous function f on $[0, 1]$, we will use the shorthand

$$\langle f, \mu \rangle := \int f(x) \mu(dx)$$

for the integral. More generally, if $f \in C([0, 1]^{M+1})$, then

$$\langle f, \mu_0 \otimes \dots \otimes \mu_M \rangle := \int f(x_0, \dots, x_M) \mu_0(dx_0) \dots \mu_M(dx_M).$$

By definition, we have

$$\langle f, G_i^{(N)}(n) \rangle = \frac{1}{N_i} \sum_{j=1}^{N_i} f(X_{ij}(n)).$$

We model migration by assuming that with probability $c_{N_i} \frac{\varpi_i}{2}$ (the factor of $\frac{1}{2}$ is to maintain consistency of notation with the cited population genetics literature), a given individual in the $n + 1^{\text{st}}$ generation is replaced by the migrant offspring of a parent chosen uniformly at random from the entire metapopulation, *i.e.*, we assume a parent of type $X_{pq}(n)$, where the p and q are drawn uniformly from $\{0, \dots, M\}$ and $\{1, \dots, N_p\}$, respectively. Thus, the average number of migrants to a given island is asymptotically independent of N ; this is a weak migration limit. Equivalently, the parent is drawn from the *metapopulation* measure,

$$G^{(N)}(n) := \frac{1}{\sum_{k=0}^M N_k} \sum_{i=0}^M N_i G_i^{(N)}(n). \quad (17)$$

Finally, we allow for the possibility that individuals mutate after birth; we assume that there is a probability measure $P^{(N)}$ such that the offspring of a parent with type $x \in [0, 1]$ mutates to a type in $A \subseteq [0, 1]$ with probability $P^{(N)}(x, A)$. Define an operator $Q^{(N)}$ on $C([0, 1])$ by

$$(Q^{(N)}f)(x) = \int_0^1 f(y) P^{(N)}(x, dy).$$

Then, for all $f \in C([0, 1])$, we define

$$\begin{aligned} (Q_i^{(N)}f)(x) &:= \mathbb{E} \left[f(X_{ij}) \mid \mathbf{G}^{(N)}(n), \text{parent of type } x \right] \\ &= (1 - c_{N_i} \frac{\varpi_i}{2})(Q^{(N)}f)(x) + c_{N_i} \frac{\varpi_i}{2} \int (Q^{(N)}f)(y) G^{(N)}(n)(dy) \end{aligned} \quad (18)$$

and

$$(B_i^{(N)}f)(x) := \frac{\varpi_i}{2} \left(\int f(y) G^{(N)}(n)(dy) - f(x) \right). \quad (19)$$

While it may at first appear unusual, this notation will greatly simplify subsequent calculations.

We will assume mutation is weak:

$$B := \lim_{N \rightarrow \infty} c_N^{-1} (I - Q^{(N)})$$

exists and B is a bounded operator. Thus, for any set $A \subseteq [0, 1]$, the probability that the offspring has a type in A approaches 1 as $N \rightarrow \infty$, if the parent has a type in A , and approaches 0 otherwise. Here, c_N is the coalescent effective population size for the mainland, and we are making the standard assumption that mutation rates scale like the reciprocal of the effective population size. For the sake of clarity in the arguments that follow, we emphasize that our assumptions entail that

$$Q_i^{(N)} = I + c_{N_i} B_i^{(N)} + c_N B + o(c_N).$$

One can consider many forms for the operator B ; the operator

$$(B^{(L)}f) \left(\frac{i}{L} \right) = \frac{\theta}{L-1} \sum_{j=1}^L \left(f \left(\frac{j}{L} \right) - f \left(\frac{i}{L} \right) \right)$$

corresponds to the classical population genetic models, in which the number of possible types is discrete and finite (here, there are L) and mutation is symmetric (*i.e.*, the offspring of an individual have the same type as their parent with probability $1 - \frac{\theta}{N}$, and mutate to any other type with probability $\frac{\theta}{N(L-1)}$). Since the labels are arbitrary, they can be assumed to be chosen from the set $\{\frac{1}{L}, \frac{2}{L}, \dots, 1\}$. Now, as $L \rightarrow \infty$, $B^{(L)}$ converges to the operator

$$(Bf)(x) = \frac{\theta}{2} \int_0^1 f(y) dy - f(x) = \theta(\langle f, \lambda \rangle - f(x)),$$

which corresponds to the infinitely many alleles model; the probability that two mutations give rise to the same type is 0. We will henceforth assume B is of this form.

Remark 1. Although we have formulated the community dynamics in discrete time, we could equally well consider a continuous time Markov process $\tilde{G}_i^{(N)}(t)$ in which disturbances happen at some rate D ; in the latter case, we consider the embedded Markov chain: if disturbances happen at random times τ_1, τ_2, \dots , then the embedded chain is the process $G_i^{(N)}(n) := \tilde{G}_i^{(N)}(\tau_n)$. The limiting (continuous time) process as $N \rightarrow \infty$ is the same for both $G^{(N)}$ and $\tilde{G}_i^{(N)}$.

In the next section, we will consider the limiting behaviour as first N and then L are taken to infinity. We will see that under moment assumptions corresponding to those in [58], all of these models converge to the same limiting process. First, however, we illustrate how Hubbell's original UNTB is an example of our class of models.

Example 1 (Hubbell's UNTB). In Hubbell's original model [5], only a single individual is replaced in each deme at each time step. We then have ν_{0i} takes values in $\{0, 1\}$, with

$$\mathbb{P}\{\nu_{0i} = 1\} = m.$$

We then have that the remaining offspring numbers are either

$$(\nu_{i1}, \dots, \nu_{iN_i}) = (1, \dots, 1, 0, 1, \dots, 1)$$

(the vector with i^{th} entry 0, and all others 1), if $\nu_{0i} = 1$, and is

$$(\nu_{i1}, \dots, \nu_{iN_i}) = (1, \dots, 1, 0, 1, \dots, 1, 2, 1, \dots, 1)$$

(the vector with i^{th} entry 0 and j^{th} entry 2 for some $i \neq j$), if $\nu_{0i} = 0$, with conditional probabilities equal to $\frac{1}{N_i}$ and $\frac{1}{N_i(N_i-1)}$, respectively (and thus the ν_{ij} are exchangeable, given ν_{i0}).

For this model, we have

$$c_{N_i} = \frac{2}{N_i(N_i - 1)},$$

whereas by definition, $(\nu_{i1})_3 = 0$, so (15) holds.

In Hubbell's model, immigrants are always from the mainland, which is assumed to have a fixed, stationary distribution (usually taken so that samples from the mainland are distributed according to Ewens' sampling formula [34]), and no mutations are assumed to occur on the islands. We will not need to make these assumptions, but will instead derive them (in the limit as $N \rightarrow \infty$) as a consequence of the relative size of the mainland and the islands.

Example 2 ("Wright-Fisher" UNTB). We can regard Hubbell's UNTB as a community analogue of the discrete Moran model. We could similarly define a community analogue to the Wright-Fisher model by assuming that the vector $(\nu_{i1}, \dots, \nu_{iN_i})$ follows a multinomial distribution with parameters N_i and $(\frac{1}{N_i}, \dots, \frac{1}{N_i})$, *i.e.*, for each i :

$$\mathbb{P}\{(\nu_{i1}, \dots, \nu_{iN_i}) = (k_1, \dots, k_{N_i})\} = \frac{N_i!}{k_1! \dots k_{N_i}!} \left(\frac{1}{N_i}\right)^{k_1} \dots \left(\frac{1}{N_i}\right)^{k_{N_i}}.$$

Here, $c_{N_i} = \frac{1}{N_i}$, whereas $\mathbb{E}[(\nu_{i1})_3] = \frac{1}{N_i}^2$.

Example 3. We briefly note that it is possible to have $c_{N_i} \equiv 0$, by assuming that $(\nu_{i1}, \dots, \nu_{iN_i}) = (1, \dots, 1)$ with probability 1 (a trivial case that we will ignore), whereas it need not be the case that

$$\lim_{N_i \rightarrow \infty} c_{N_i} = 0$$

if (15) is violated: if we assume that with probability $\frac{1}{N_i}$, $\nu_{ij} = N_i$ and $\nu_{ik} = 0$ for all $k \neq j$, then $c_{N_i} \equiv 1$.

It is well known [61] that

$$\frac{\binom{N_i}{j}}{\binom{N_i}{k}} \mathbb{E} [(\nu_{i1})_{k_1} \cdots (\nu_{ij})_{k_j}],$$

where $j, k_1, \dots, k_j \in \mathbb{N}$ and $k := k_1 + \cdots + k_j$, is the probability that k individuals, sampled uniformly at random without replacement from the i^{th} deme have exactly j parents in the previous generation, *n.b.*, exchangeability implies that

$$\frac{\binom{N_i}{j}}{\binom{N_i}{k}} \mathbb{E} [(\nu_{i1})_{k_1} \cdots (\nu_{ij})_{k_j}] = \frac{\binom{N_i}{j}}{\binom{N_i}{k}} \mathbb{E} [(\nu_{i\pi(1)})_{k_1} \cdots (\nu_{i\pi(j)})_{k_j}]$$

for any permutation π of $\{1, \dots, N_i\}$, so that these probabilities only depend on j, k , and the unordered list of values k_1, \dots, k_j . In [58], we find the following monotonicity result for these probabilities:

Lemma 2. *Let $j \geq l, k_1 \geq m_1, \dots, k_l \geq m_l$, and $m := m_1 + \cdots + m_l$. Then,*

$$\frac{\binom{N_i}{j}}{\binom{N_i}{k}} \mathbb{E} [(\nu_{i1})_{k_1} \cdots (\nu_{ij})_{k_j}] \leq \frac{\binom{N_i}{l}}{\binom{N_i}{m}} \mathbb{E} [(\nu_{i1})_{m_1} \cdots (\nu_{il})_{m_l}].$$

Remark 2. In particular, in conjunction with Lemma 2, we have

$$\frac{\binom{N_i}{j-1}}{\binom{N_i}{j}} \mathbb{E} [(\nu_{i1})_2 \nu_{i2} \cdots \nu_{ij-1}] \leq c_{N_i}, \tag{20}$$

(and, by exchangeability, whenever at least one $k_i \geq 2$) and

$$\frac{\binom{N_i}{j}}{\binom{N_i}{k}} \mathbb{E} [(\nu_{i1})_{k_1} \cdots (\nu_{ij})_{k_j}] = o(c_{N_i}) \tag{21}$$

whenever $k_q, k_r \geq 2$ for at least two distinct indices q, r or $k_q \geq 3$ for some index q .

Remark 3. In particular, in (21), $\frac{\binom{N_i}{j}}{\binom{N_i}{k}} \mathbb{E} [(\nu_{i1})_{k_1} \cdots (\nu_{ij})_{k_j}]$ is always smaller than one of

$$\frac{\binom{N_i}{1}}{\binom{N_i}{3}} \mathbb{E} [(\nu_{i1})_3]$$

or

$$\frac{\binom{N_i}{2}}{\binom{N_i}{4}} \mathbb{E} [(\nu_{i1})_2 (\nu_{i2})_2].$$

In what follows, all terms $o(c_{N_i})$ will be of order at most equal to one of these two quantities (which are the probability of three individuals sampled at random having the same parent in the previous generation, or a sample of four individuals consisting of two pairs of descendants of two distinct parents, respectively) or will be of order less than or equal to $\frac{c_{N_i}}{N_i}$. This will be very important when we consider the long timescale.

We will have use of some general relations between exchangeable random variables in the sequel:

Lemma 3. *For all $j > 1$*

$$\mathbb{E} [\nu_{i1} \cdots \nu_{ij-1}] - \mathbb{E} [\nu_{i1} \cdots \nu_{ij}] = (j-1) \frac{\binom{N_i}{j-1}}{\binom{N_i}{j}} \mathbb{E} [(\nu_{i1})_2 \cdots \nu_{ij-1}].$$

Proof. We begin by observing that

$$\begin{aligned}
N_i \mathbb{E} [\nu_{i1} \cdots \nu_{ij-1}] &= \mathbb{E} [N_i \nu_{i1} \cdots \nu_{ij-1}] \\
&= \mathbb{E} [(\nu_{i1} + \cdots + \nu_{iN_i}) \nu_{i1} \cdots \nu_{ij-1}] \\
&= \mathbb{E} \left[\sum_{k=1}^{N_i} \nu_{i1} \cdots \nu_{ij-1} \nu_{ik} \right] \\
&= \mathbb{E} \left[\sum_{k=1}^{j-1} \nu_{i1} \cdots \nu_{ij-1} \nu_{ik} + \sum_{k=j}^{N_i} \nu_{i1} \cdots \nu_{ij-1} \nu_{ik} \right] \\
&= \sum_{k=1}^{j-1} \mathbb{E} [\nu_{i1} \cdots \nu_{ij-1} \nu_{ik}] + \sum_{k=j}^{N_i} \mathbb{E} [\nu_{i1} \cdots \nu_{ij-1} \nu_{ik}] \\
&= \sum_{k=1}^{j-1} \mathbb{E} \left[\nu_{ik}^2 \prod_{\substack{l=1 \\ l \neq k}}^{j-1} \nu_{il} \right] + \sum_{k=j}^{N_i} \mathbb{E} [\nu_{i1} \cdots \nu_{ij-1} \nu_{ik}]
\end{aligned}$$

and thus, exploiting the exchangeability of the ν_{ij} ,

$$= (j-1) \mathbb{E} [\nu_{i1}^2 \cdots \nu_{ij-1}] + (N_i - j + 1) \mathbb{E} [\nu_{i1} \cdots \nu_{ij}].$$

On the other hand,

$$N_i \mathbb{E} [\nu_{i1} \cdots \nu_{ij-1}] = (j-1) \mathbb{E} [\nu_{i1} \cdots \nu_{ij-1}] + (N_i - j + 1) \mathbb{E} [\nu_{i1} \cdots \nu_{ij-1}].$$

Equating the two sides and subtracting, we get

$$(N_i - j + 1) (\mathbb{E} [\nu_{i1} \cdots \nu_{ij-1}] - \mathbb{E} [\nu_{i1} \cdots \nu_{ij}]) = (j-1) (\mathbb{E} [\nu_{i1}^2 \cdots \nu_{ij-1}] - \mathbb{E} [\nu_{i1} \cdots \nu_{ij-1}]).$$

The result follows. □

Remark 4. In conjunction with (20), the lemma tells us that for all $j > 1$,

$$\mathbb{E} [\nu_{i1} \cdots \nu_{ij-1}] - \mathbb{E} [\nu_{i1} \cdots \nu_{ij}] = \mathcal{O}(c_{N_i}),$$

and thus,

$$\mathbb{E} [\nu_{i1} \cdots \nu_{iq}] - \mathbb{E} [\nu_{i1} \cdots \nu_{ir}] = \mathcal{O}(c_{N_i}),$$

for any $q < r$.

Next, we observe that

Lemma 4. *For all j ,*

$$\mathbb{E} [\nu_{i1} \cdots \nu_{ij}] = 1 - \binom{j}{2} \frac{(N_i)_{j-1}}{(N_i)_j} \mathbb{E} [(\nu_{i1})_2 \cdots \nu_{ij-1}] - o(c_{N_i}).$$

Proof. This is a consequence of the identity

$$(N_i)_j = (\nu_{i1} + \cdots + \nu_{iN_i})_j = \sum_{j_1 + \cdots + j_{N_i} = j} \frac{j!}{j_1! \cdots j_{N_i}!} (\nu_{i1})_{j_1} \cdots (\nu_{iN_i})_{j_{N_i}},$$

where we assume $0! = 1$ for ease of notation, and we assume that most of the j_i are equal to zero. Equivalently, if we only consider non-zero values,

$$(N_i)_j = \sum_{m=1}^j \sum_{\substack{n_1, \dots, n_m \\ \text{distinct}}} \sum_{k_1 + \dots + k_m = k} \frac{j!}{k_1! \dots k_m!} (\nu_{in_1})_{k_1} \dots (\nu_{in_m})_{k_m}.$$

Taking expectations on both sides, and using the exchangeability of $(\nu_{i1}, \dots, \nu_{iN_i})$, we have

$$(N_i)_j = \sum_{m=1}^j \sum_{\substack{n_1, \dots, n_m \\ \text{distinct}}} \sum_{k_1 + \dots + k_m = k} \frac{j!}{k_1! \dots k_m!} \mathbb{E}[(\nu_{i1})_{k_1} \dots (\nu_{im})_{k_m}].$$

Now, observe that the expected value of the summand is independent of the choice of the values n_1, \dots, n_m , that can be chosen in $\binom{N}{m}$ ways. Moreover, the expectation $\mathbb{E}[(\nu_{i1})_{k_1} \dots (\nu_{im})_{k_m}]$ remains unchanged under permutations, and thus are all equal to

$$\mathbb{E}[(\nu_{i1})_{\tilde{k}_1} \dots (\nu_{im})_{\tilde{k}_m}],$$

where $\tilde{k}_1 \geq \tilde{k}_2 \geq \dots \geq \tilde{k}_m$ are the values k_1, \dots, k_m listed in decreasing order. If we let a_p be the number of indices q such that $k_q = p$,

$$a_p = \#\{q : k_q = p\},$$

then

$$\begin{aligned} & \sum_{m=1}^j \sum_{\substack{n_1, \dots, n_m \\ \text{distinct}}} \sum_{k_1 + \dots + k_m = k} \frac{j!}{k_1! \dots k_m!} \mathbb{E}[(\nu_{i1})_{k_1} \dots (\nu_{im})_{k_m}] \\ &= \sum_{m=1}^j \sum_{\substack{\tilde{k}_1 + \dots + \tilde{k}_m = k \\ \tilde{k}_1 \geq \tilde{k}_2 \geq \dots \geq \tilde{k}_m}} \frac{j!}{\tilde{k}_1! \dots \tilde{k}_m!} \frac{m!}{a_1! \dots a_j!} \binom{N}{m} \mathbb{E}[(\nu_{i1})_{\tilde{k}_1} \dots (\nu_{im})_{\tilde{k}_m}], \end{aligned}$$

so that, simplifying and dividing through by $(N)_j$, we have

$$\begin{aligned} 1 &= \sum_{m=1}^j \sum_{\substack{\tilde{k}_1 + \dots + \tilde{k}_m = k \\ \tilde{k}_1 \geq \tilde{k}_2 \geq \dots \geq \tilde{k}_m}} \frac{j!}{\tilde{k}_1! \dots \tilde{k}_m!} \frac{1}{a_1! \dots a_j!} \frac{(N)_m}{(N)_j} \mathbb{E}[(\nu_{i1})_{\tilde{k}_1} \dots (\nu_{im})_{\tilde{k}_m}] \\ &= \mathbb{E}[\nu_{i1} \dots \nu_{ij}] + \binom{j}{2} \frac{(N_i)_{j-1}}{(N_i)_j} \mathbb{E}[(\nu_{i1})_2 \dots \nu_{ij-1}] + o(c_{N_i}), \end{aligned}$$

where, using (21), we have truncated after the two highest order terms in the sum. \square

We conclude this section with a final observation,

Lemma 5. *Let $j > 1$. Then,*

$$\frac{(N_i)_j}{(N_i)_{j+1}} \mathbb{E}[(\nu_{i1})_2 \nu_{i2} \dots \nu_{ij}] = \frac{(N_i)_{j-1}}{(N_i)_j} \mathbb{E}[(\nu_{i1})_2 \nu_{i2} \dots \nu_{ij-1}] + o(c_{N_i}).$$

Proof. Again exploiting exchangeability, we see that

$$\begin{aligned}
& (N_i - j + 1) \mathbb{E}[(\nu_{i1})_2 \nu_{i2} \cdots \nu_{ij}] \\
&= \mathbb{E}[(\nu_{i1})_2 \nu_{i2} \cdots \nu_{ij-1} \nu_{ij}] + \mathbb{E}[(\nu_{i1})_2 \nu_{i2} \cdots \nu_{ij-1} \nu_{ij+1}] + \cdots + \mathbb{E}[(\nu_{i1})_2 \nu_{i2} \cdots \nu_{ij-1} \nu_{iN_i}] \\
&= \mathbb{E}[(\nu_{i1})_2 \nu_{i2} \cdots \nu_{ij-1} (\nu_{ij} + \cdots + \nu_{iN_i})] \\
&= \mathbb{E}[(\nu_{i1})_2 \nu_{i2} \cdots \nu_{ij-1} (N_i - \nu_{i1} - \cdots - \nu_{ij-1})] \\
&= \mathbb{E}[(\nu_{i1})_2 \nu_{i2} \cdots \nu_{ij-1} (N_i - j + 2 - (\nu_{i1} - 2) - (\nu_{i2} - 1) - \cdots - (\nu_{ij-1} - 1))] \\
&= (N_i - j + 2) \mathbb{E}[(\nu_{i1})_2 \nu_{i2} \cdots \nu_{ij-1}] - \mathbb{E}[(\nu_{i1})_3 \nu_{i2} \cdots \nu_{ij-1}] \\
&\quad - \mathbb{E}[(\nu_{i1})_2 (\nu_{i2})_2 \nu_{i3} \cdots \nu_{ij-1}] - \cdots - \mathbb{E}[(\nu_{i1})_2 \nu_{i2} \cdots (\nu_{ij-1})_2].
\end{aligned}$$

In particular, dividing both sides by $(N_i - j + 1)(N_i - j + 2)$, we have

$$\begin{aligned}
\frac{\binom{N_i}{j}}{\binom{N_i}{j+1}} \mathbb{E}[(\nu_{i1})_2 \nu_{i2} \cdots \nu_{ij}] &= \frac{\binom{N_i}{j-1}}{\binom{N_i}{j}} \mathbb{E}[(\nu_{i1})_2 \nu_{i2} \cdots \nu_{ij-1}] - \frac{\binom{N_i}{j-1}}{\binom{N_i}{j+1}} \mathbb{E}[(\nu_{i1})_3 \nu_{i2} \cdots \nu_{ij-1}] \\
&\quad - \frac{\binom{N_i}{j-1}}{\binom{N_i}{j+1}} \mathbb{E}[(\nu_{i1})_2 (\nu_{i2})_2 \nu_{i3} \cdots \nu_{ij-1}] - \cdots - \frac{\binom{N_i}{j-1}}{\binom{N_i}{j+1}} \mathbb{E}[(\nu_{i1})_2 \nu_{i2} \cdots (\nu_{ij-1})_2]
\end{aligned}$$

and the result again follows by (21). \square

Remark 5. Iterating the previous lemma, we see that

$$\frac{\binom{N_i}{j}}{\binom{N_i}{j+1}} \mathbb{E}[(\nu_{i1})_2 \cdots \nu_{ij}] = \cdots = \frac{\mathbb{E}[(\nu_{i1})_2]}{N_i - 1} + o(c_{N_i}) = c_{N_i} + o(c_{N_i}).$$

1.4 Convergence to a Limit

We will be interested in weak limits of the random measures $\mathbf{G}^{(N)}(n)$ in two time-scales determined by N , a “slow-time” process, $\mathbf{G}^{(N)}(\lfloor c_N^{-1} t \rfloor)$, and an “intermediate-time” process $\mathbf{G}^{(N)}(\lfloor a_N^{-1} t \rfloor)$, where $t > 0$ is a continuous time variable, and we will consider the limits as $N \rightarrow \infty$.

Our principal tool in doing this is the generator of $\mathbf{G}^{(N)}(n)$, an operator on $C(\mathcal{P}([0, 1])^{M+1})$ defined by

$$(\mathcal{G}_N F)(\boldsymbol{\mu}) = \mathbb{E} \left[F(\mathbf{G}^{(N)}(n+1)) \middle| \mathbf{G}^{(N)}(n) = \boldsymbol{\mu} \right] - F(\boldsymbol{\mu}).$$

Knowing $(\mathcal{G}_N F)(\boldsymbol{\mu})$ for all $F \in C(\mathcal{P}([0, 1])^{M+1})$ and all $\boldsymbol{\mu} \in \mathcal{P}([0, 1])^{M+1}$ completely characterizes the transition probabilities of $\mathbf{G}^{(N)}$, and thus, together with the initial value $\mathbf{G}^{(N)}(0)$, allow us to characterize the process (although not necessarily the limit, see *e.g.*, [57]).

Our limiting processes are continuous, rather than discrete time random variables, but also have associated generators; in general, if $\mathbf{H}(t)$ is a continuous time process taking values in $\mathcal{P}([0, 1])^{M+1}$ and $F \in C(\mathcal{P}([0, 1])^{M+1})$, then $\mathbf{H}(t)$ has generator \mathcal{H} :

$$(\mathcal{H}F)(\boldsymbol{\mu}) = \lim_{h \rightarrow 0^+} \frac{\mathbb{E}[F(\mathbf{H}(t+h)) | \mathbf{H}(t) = \boldsymbol{\mu}] - F(\boldsymbol{\mu})}{h},$$

with domain $\mathcal{D}(\mathcal{H})$, consisting of all functions F for which the limit exists.

The notion of a generator simultaneously generalizes the transition matrix, master equation, and diffusion equations of classical probability. The typical proof of convergence proceeds by first showing that a limit exists, then characterizing the limit by first determining the limit of the generators, and finally showing that given the initial conditions (via a distribution from which they are drawn), there is a unique process with that generator (*e.g.*, [57] is a standard reference).

Remark 6. Note that $(\mathcal{H}F)(\boldsymbol{\mu})$ is the right-hand derivative of $\mathbb{E}[F(\mathbf{H}(t+h))|\mathbf{H}(t) = \boldsymbol{\mu}]$ at $t = 0$. In particular, if the generator vanishes, then $\mathbb{E}[F(\mathbf{H}(t))|\mathbf{H}(0) = \boldsymbol{\mu}] = F(\boldsymbol{\mu})$ for all $t > 0$, and all F , and the process $\mathbf{H}(t) \equiv \boldsymbol{\mu}$ is constant. This will be important when we come to consider the limit on the intermediate time scale.

We will make use of the fact that the set of functions

$$\mathcal{C} := \left\{ F(\boldsymbol{\mu}) = \prod_{i=0}^M \prod_{k=1}^{K_i} \langle f_{ik}, \mu_i \rangle \mid K_i \in \mathbb{N}_0, f_{ik} \in C([0, 1]) \right\}$$

is separating, and convergence determining [57], so that for the purpose of characterizing our process and its limits, we need only compute the value the generator takes on functions $F \in \mathcal{C}$ and its limits.

We will evaluate the generator on this class of functions, but we first begin with a pair of lemmas. We will use \coprod to indicate the disjoint union of sets and, for all integers $M > 0$, we use the shorthand $[M] = \{1, \dots, M\}$.

Lemma 6. *Let $\mu_i = \frac{1}{N_i} \sum_{j=1}^{N_i} \delta_{x_{ij}}$ for $x_{ij} \in [0, 1]$. Then,*

$$\begin{aligned} \prod_{k=1}^{K_i} \langle f_{ik}, \mu_i \rangle &= \frac{1}{N_i^{K_i}} \sum_{m=1}^{K_i} \sum_{\substack{j_1, \dots, j_m \\ \text{distinct}}} \sum_{A_1 \coprod \dots \coprod A_m = [K_i]} \prod_{q=1}^m \prod_{r \in A_q} f_{ir}(x_{ij_q}) \\ &= \frac{1}{N_i^{K_i}} \sum_{\substack{j_1, \dots, j_{K_i} \\ \text{distinct}}} \prod_{k=1}^{K_i} f_{ik}(x_{ij_k}) + \mathcal{O}(N^{-1}), \end{aligned}$$

where the sum is over all partitions of $[K_i]$ into m disjoint sets.

Proof. The first statement is simply a matter of collecting terms according to the number of distinct values j_k :

$$\begin{aligned} N_i^{K_i} \prod_{k=1}^{K_i} \langle f_{ik}, \mu_i \rangle &= N_i^{K_i} \prod_{k=1}^{K_i} \left(\frac{1}{N_i} \sum_{j=1}^{N_i} f_{ik}(x_{ij}) \right) = \sum_{j_1=1}^{N_i} \dots \sum_{j_{K_i}=1}^{N_i} \prod_{k=1}^{K_i} f_{ik}(x_{ij_k}) \\ &= \sum_{m=1}^{K_i} \sum_{\substack{j_1, \dots, j_m \\ \text{distinct}}} \sum_{A_1 \coprod \dots \coprod A_m = [K_i]} \prod_{q=1}^m \prod_{r \in A_q} f_{ir}(x_{ij_q}). \end{aligned}$$

Now, for the final term, we have $m = 1$, $A_1 = [K_i]$, so it takes the form:

$$\sum_{j_1=1}^{N_i} \prod_{k=1}^{K_i} f_{ik}(x_{ij_1}) = N_i \langle \prod_{k=1}^{K_i} f_{ik}, \mu_i \rangle,$$

whilst for $m = 2$, we have:

$$\begin{aligned} &= \sum_{j_1=1}^{N_i} \sum_{j_2 \neq j_1} \sum_{A_1 \coprod A_2 = [K_i]} \prod_{k \in A_1} f_{ik}(x_{ij_1}) \prod_{k \in A_2} f_{ik}(x_{ij_2}) \\ &= \sum_{A_1 \coprod A_2 = [K_i]} \sum_{j_1=1}^{N_i} \prod_{k \in A_1} f_{ik}(x_{ij_1}) \sum_{j_2=1}^{N_i} \prod_{k \in A_2} f_{ik}(x_{ij_2}) - \sum_{A_1 \coprod A_2 = [K_i]} \sum_{j_1=1}^{N_i} \prod_{k \in A_1} f_{ik}(x_{ij_1}) \prod_{k \in A_2} f_{ik}(x_{ij_1}) \\ &= N_i^2 \sum_{A_1 \coprod A_2 = [K_i]} \langle \prod_{k \in A_1} f_{ik}, \mu_i \rangle \langle \prod_{k \in A_2} f_{ik}, \mu_i \rangle - S(K_i, 2) N_i \langle \prod_{k \in A_1} f_{ik}, \mu_i \rangle, \end{aligned}$$

where $S(K_i, 2)$ is a Stirling number of the second kind [62] and gives the number of distinct partitions of K_i elements into 2 sets. 31

Proceeding inductively in this manner completes the proof of the lemma. \square

The previous lemma shows we will be interested in products over distinct indices j_1, \dots, j_m . In particular, we have the result of Lemma 7.

Lemma 7. *For distinct values j_1, \dots, j_{K_i} in $\{1, \dots, N_i\}$,*

$$\begin{aligned} \mathbb{E} \left[\prod_{k=1}^{K_i} f_{ik}(X_{ij_k}(n+1)) \middle| \{X_{ij}(n) = x_{ij}\} \right] &= \frac{\mathbb{E}[\nu_{i1} \cdots \nu_{iK_i}]}{(N_i)_{K_i}} \sum_{\substack{p_1, \dots, p_{K_i} \\ \text{distinct}}} \prod_{k=1}^{K_i} (Q_i^{(N)} f_{ik})(x_{ip_k}) \\ &+ \frac{\mathbb{E}[(\nu_{i1})_2 \nu_{i2} \cdots \nu_{iK_i-1}]}{(N_i)_{K_i}} \sum_{q < r} \sum_{\substack{p_1, \dots, p_{K_i} \\ p_q = p_r}} \prod_{\substack{k=1 \\ k \neq q, r}}^{K_i} (Q_i^{(N)} f_{ik})(x_{ip_k}) (Q_i^{(N)} f_{iq} Q_i^{(N)} f_{ir})(x_{ip_q}) + o(c_{N_i}). \end{aligned}$$

Proof. We begin by recalling that conditional on an individual's parent having type x , its type is independently distributed according to the probability measure $P(x, \cdot)$, *i.e.*,

$$\mathbb{E} \left[f(X_{ij}(n+1)) \middle| \mathbf{G}^{(N)}(n), \text{parent of type } x \right] = (Q_i^{(N)} f)(x).$$

We can thus, similar to the previous lemma, write:

$$\begin{aligned} \mathbb{E} \left[\prod_{k=1}^{K_i} f_{ik}(X_{ij_k}(n+1)) \middle| \{X_{ij}(n) = x_{ij}\} \right] \\ = \sum_{m=1}^{K_i} \sum_{\substack{p_1, \dots, p_m \\ \text{distinct}}} \sum_{A_1 \amalg \cdots \amalg A_m = [K_i]} \frac{\mathbb{E}[(\nu_{ip_1})_{|A_1|} \cdots (\nu_{ip_m})_{|A_m|}]}{(N_i)_{K_i}} \prod_{q=1}^m \prod_{r \in A_q} (Q_i^{(N)} f_{ir})(x_{ip_q}), \end{aligned}$$

where

$$\frac{\mathbb{E}[(\nu_{ip_1})_{|A_1|} \cdots (\nu_{ip_m})_{|A_m|}]}{(N_i)_{K_i}} = \frac{\mathbb{E}[(\nu_{i1})_{|A_1|} \cdots (\nu_{im})_{|A_m|}]}{(N_i)_{K_i}}$$

is the probability that the K_i distinct individuals have m ancestors p_1, \dots, p_m (with types $x_{ip_1}, \dots, x_{ip_m}$), and that the individuals in A_q had parent p_q .

Next, we observe that since $\|Q_i^{(N)}\| \leq 1$,

$$\begin{aligned} \left| \sum_{\substack{p_1, \dots, p_m \\ \text{distinct}}} \sum_{A_1 \amalg \cdots \amalg A_m = [K_i]} \frac{\mathbb{E}[(\nu_{ip_1})_{|A_1|} \cdots (\nu_{ip_m})_{|A_m|}]}{(N_i)_{K_i}} \prod_{q=1}^m \prod_{r \in A_q} (Q_i^{(N)} f_{ir})(x_{ip_q}) \right| \\ \leq \sum_{A_1 \amalg \cdots \amalg A_m = [K_i]} \frac{(N_i)_m}{(N_i)_{K_i}} \mathbb{E}[(\nu_{i1})_{|A_1|} \cdots (\nu_{im})_{|A_m|}] \prod_{k=1}^{K_i} \|f_{ik}\| \end{aligned}$$

and is thus $o(c_{N_i})$ whenever $|A_q| \geq 3$ for some q or $|A_q|$ and $|A_r|$ are both ≥ 2 for distinct indices q, r by (21). The result follows. \square

We now turn to the main result of this section:

Proposition 1. Let $\mu_i^{(N)} = \frac{1}{N_i} \sum_{j=1}^{N_i} \delta_{x_{ij}}$, for $x_{ij} \in [0, 1]$ and let $\boldsymbol{\mu}^{(N)} = \mu_1^{(N)} \otimes \cdots \otimes \mu_M^{(N)}$ converge weakly to a measure $\boldsymbol{\mu} \in \mathcal{P}([0, 1]^{M+1})$.

Let $F(\boldsymbol{\mu}) = \prod_{i=0}^M \prod_{k=1}^{K_i} \langle f_{ik}, \mu_i \rangle \in \mathcal{C}$ and, for $i = 1, \dots, M$, let

$$(\mathcal{G}_i F)(\boldsymbol{\mu}) = \prod_{\substack{j=0 \\ j \neq i}}^M \prod_{k=1}^{K_j} \langle f_{jk}, \mu_j \rangle \left(\sum_{q=1}^{K_i} \frac{\varpi_i}{2} \langle f_{iq}, \mu_i - \mu_0 \rangle \prod_{\substack{k=1 \\ k \neq q}}^{K_i} \langle f_{ik}, \mu_i \rangle + \frac{1}{2} \sum_{q \neq r} \prod_{\substack{k=1 \\ k \neq q, r}}^{K_i} \langle f_{ik}, \mu_i \rangle (\langle f_{iq} f_{ir}, \mu_i \rangle - \langle f_{iq}, \mu_i \rangle \langle f_{ir}, \mu_i \rangle) \right) \quad (22)$$

define an operator on $C(\mathcal{P}([0, 1]^{M+1}))$.

Then,

$$\lim_{N \rightarrow \infty} a_N^{-1} (\mathcal{G}_N F)(\boldsymbol{\mu}^{(N)}) = (\mathcal{G} F)(\boldsymbol{\mu}) := \sum_{i=1}^M \gamma_i (\mathcal{G}_i F)(\boldsymbol{\mu}).$$

Moreover, given $\tilde{\mu}_i \in \mathcal{P}(\mathcal{P}([0, 1]))$, there exist unique independent processes $G_i(t)$ with generators \mathcal{G}_i , such that $G_i(0)$ is distributed according to $\tilde{\mu}_i$ and such that

$$\mathbf{G}^{(N)}(\lfloor a_N^{-1} t \rfloor) \Rightarrow \mathbf{G}(t) := G_0(0) \otimes G_1(\gamma_1 t) \otimes \cdots \otimes G_M(\gamma_M t),$$

for all $t > 0$, where convergence is in the space of càdlàg functions endowed with the Skorokhod topology, $\mathbb{D}_{\mathcal{P}([0, 1]^{M+1})}[0, \infty)$ (see e.g., [57]).

Remark 7. Because

$$\lim_{N \rightarrow \infty} \frac{c_N}{a_N} = 0,$$

the component of the generator acting on the mainland vanishes in the limit; if

$$\mathcal{C}_0 := \left\{ F \in \mathcal{C} \mid F(\boldsymbol{\mu}) = \prod_{k=1}^{K_0} \langle f_{0k}, \mu_0 \rangle \right\},$$

then $\mathcal{G}_i F \equiv 0$ for all $F \in \mathcal{C}_0$ and thus the generator vanishes on this set. Equivalently, the process $G_0(t) \equiv \mu_0$.

Remark 8. Recall from Equation 16 that the effective population size of the i^{th} island is $c_{N_i} \sim \gamma_i a_N$; since we have rescaled time by a_N rather than the individual effective population sizes, the factors γ_i appear in the generator and in the components G_i . These reflect the fact that the different effective population sizes on the different islands result in their population dynamics having different rates (*i.e.*, different expected inter-event times), which are given by the γ_i .

Remark 9. This theorem tells us that on the intermediate time scale, the islands have essentially independent dynamics, coupled only by immigration from a mainland which remains unchanged on the intermediate timescale. The generator of the dynamics on the island is identical to that in the infinite population limit for the infinitely many alleles model, with the rescaled migration rate, $\frac{\varpi_i}{2}$ taking the place of the rescaled mutation rate θ , and the mainland density measure μ_0 taking the place of Lebesgue measure.

Proof. Applying Lemmas 6 and 7, we have

$$\begin{aligned}
\mathbb{E} \left[F(\mathbf{G}^{(N)}(n+1)) \middle| \mathbf{G}^{(N)}(n) = \boldsymbol{\mu} \right] &= \mathbb{E} \left[\prod_{i=0}^M \prod_{k=1}^{K_i} \langle f_{ik}, G_i^{(N)}(n+1) \rangle \middle| \{X_{ij}(n) = x_{ij}\} \right] \\
&= \prod_{i=0}^M \mathbb{E} \left[\prod_{k=1}^{K_i} \langle f_{ik}, G_i^{(N)}(n+1) \rangle \middle| \{X_{ij}(n) = x_{ij}\} \right] \\
&= \prod_{i=0}^M \frac{1}{N_i^{K_i}} \sum_{m=1}^{K_i} \sum_{\substack{j_1, \dots, j_m \\ \text{distinct}}} \sum_{A_1 \amalg \dots \amalg A_m = [K_i]} \mathbb{E} \left[\prod_{q=1}^m \prod_{r \in A_q} f_{ir}(x_{ij_q}) \middle| \{X_{ij}(n) = x_{ij}\} \right] \\
&= \prod_{i=0}^M \frac{1}{N_i^{K_i}} \sum_{m=1}^{K_i} \sum_{\substack{j_1, \dots, j_m \\ \text{distinct}}} \sum_{A_1 \amalg \dots \amalg A_m = [K_i]} \left(\frac{\mathbb{E} [\nu_{i1} \cdots \nu_{im}]}{(N_i)_m} \sum_{\substack{p_1, \dots, p_m \\ \text{distinct}}} \prod_{k=1}^m (Q_i^{(N)} \prod_{l \in A_k} f_{il})(x_{ip_k}) \right. \\
&\quad \left. + \frac{\mathbb{E} [(\nu_{i1})_2 \nu_{i2} \cdots \nu_{im-1}]}{(N_i)_m} \sum_{q < r} \sum_{\substack{p_1, \dots, p_m \\ p_q = p_r}} \prod_{\substack{k=1 \\ k \neq q, r}}^m (Q_i^{(N)} \prod_{l \in A_k} f_{il})(x_{ip_k}) ((Q_i^{(N)} \prod_{l \in A_q} f_{il})(Q_i^{(N)} \prod_{l \in A_r} f_{il}))(x_{ip_q}) + o(c_{N_i}) \right).
\end{aligned}$$

Now, observing that the term in brackets is independent of the values j_k , we note that j_1, \dots, j_m can be chosen in $(N_i)_m$ ways, and we are left with a product over sums of the form:

$$\begin{aligned}
&\frac{1}{N_i^{K_i}} \sum_{m=1}^{K_i} \mathbb{E} [\nu_{i1} \cdots \nu_{im}] \sum_{A_1 \amalg \dots \amalg A_m = [K_i]} \sum_{\substack{p_1, \dots, p_m \\ \text{distinct}}} \prod_{k=1}^m (Q_i^{(N)} \prod_{l \in A_k} f_{il})(x_{ip_k}) \\
&\quad + \frac{1}{N_i^{K_i}} \sum_{m=1}^{K_i} \mathbb{E} [(\nu_{i1})_2 \nu_{i2} \cdots \nu_{im-1}] \sum_{A_1 \amalg \dots \amalg A_m = [K_i]} \sum_{q < r} \sum_{\substack{p_1, \dots, p_m \\ p_q = p_r}} \prod_{\substack{k=1 \\ k \neq q, r}}^m (Q_i^{(N)} \prod_{l \in A_k} f_{il})(x_{ip_k}) ((Q_i^{(N)} \prod_{l \in A_q} f_{il})(Q_i^{(N)} \prod_{l \in A_r} f_{il}))(x_{ip_q}) + o(c_{N_i}).
\end{aligned}$$

We will focus our attention on the first sum in the first line. Using Lemma 6 in reverse, we have

$$\begin{aligned}
\frac{1}{N_i^{K_i}} \sum_{\substack{p_1, \dots, p_{K_i} \\ \text{distinct}}} \prod_{k=1}^{K_i} (Q_i^{(N)} f_{ik})(x_{ip_k}) &= \prod_{k=1}^{K_i} \langle Q_i^{(N)} f_{ik}, \mu_i \rangle \\
&\quad - \frac{1}{N_i^{K_i}} \sum_{m=1}^{K_i-1} \sum_{A_1 \amalg \dots \amalg A_m = [K_i]} \sum_{\substack{p_1, \dots, p_m \\ \text{distinct}}} \prod_{k=1}^m (Q_i^{(N)} \prod_{l \in A_k} f_{il})(x_{ip_k}),
\end{aligned}$$

where the terms on the second line are $\mathcal{O}\left(\frac{1}{N_i}\right)$. Thus,

$$\begin{aligned} & \frac{1}{N_i^{K_i}} \sum_{m=1}^{K_i} \mathbb{E} [\nu_{i1} \cdots \nu_{im}] \sum_{A_1 \amalg \cdots \amalg A_m = [K_i]} \sum_{\substack{p_1, \dots, p_m \\ \text{distinct}}} \prod_{k=1}^m (Q_i^{(N)} \prod_{l \in A_k} f_{il})(x_{ip_k}) \\ &= \mathbb{E} [\nu_{i1} \cdots \nu_{iK_i}] \prod_{k=1}^{K_i} \langle Q_i^{(N)} f_{ik}, \mu_i \rangle \\ &+ \frac{1}{N_i^{K_i}} \sum_{m=1}^{K_i-1} (\mathbb{E} [\nu_{i1} \cdots \nu_{im}] - \mathbb{E} [\nu_{i1} \cdots \nu_{iK_i}]) \sum_{A_1 \amalg \cdots \amalg A_m = [K_i]} \sum_{\substack{p_1, \dots, p_m \\ \text{distinct}}} \prod_{k=1}^m (Q_i^{(N)} \prod_{l \in A_k} f_{il})(x_{ip_k}). \end{aligned}$$

Further, we observed in Remark 4 that the differences $\mathbb{E} [\nu_{i1} \cdots \nu_{im}] - \mathbb{E} [\nu_{i1} \cdots \nu_{iK_i}]$ are $\mathcal{O}(c_{N_i})$, so that the first sum reduces to

$$\mathbb{E} [\nu_{i1} \cdots \nu_{iK_i}] \prod_{k=1}^{K_i} \langle Q_i^{(N)} f_{ik}, \mu_i \rangle + o(c_{N_i}).$$

Proceeding similarly, applying Lemma 6 with the set of $K_i - 1$ distinct functions $\{Q_i^{(N)} f_{ik}\}_{k \neq q, r} \cup \{(Q_i^{(N)} f_{iq})(Q_i^{(N)} f_{ir})\}$, we see that

$$\begin{aligned} & \frac{1}{N_i^{K_i}} \sum_{m=1}^{K_i} \mathbb{E} [(\nu_{i1})_2 \nu_{i2} \cdots \nu_{im-1}] \sum_{A_1 \amalg \cdots \amalg A_m = [K_i]} \sum_{q < r} \sum_{\substack{p_1, \dots, p_m \\ p_q = p_r}} \\ & \prod_{\substack{k=1 \\ k \neq q, r}}^m (Q_i^{(N)} \prod_{l \in A_k} f_{il})(x_{ip_k}) ((Q_i^{(N)} \prod_{l \in A_q} f_{il})(Q_i^{(N)} \prod_{l \in A_r} f_{il}))(x_{ip_q}) \\ &= \frac{1}{N_i} \mathbb{E} [(\nu_{i1})_2 \nu_{i2} \cdots \nu_{iK_i-1}] \sum_{q < r} \prod_{\substack{k=1 \\ k \neq q, r}}^{K_i} \langle Q_i^{(N)} f_{ik}, \mu_i \rangle \langle (Q_i^{(N)} f_{iq})(Q_i^{(N)} f_{ir}), \mu_i \rangle + o(c_{N_i}), \end{aligned}$$

where we have used the fact that $\frac{1}{N_i} \mathbb{E} [(\nu_{i1})_2 \nu_{i2} \cdots \nu_{im-1}] = \mathcal{O}(c_{N_i})$ in bounding the higher order terms. Thus,

$$\begin{aligned} \mathbb{E} \left[F(\mathbf{G}^{(N)}(n+1)) \middle| \mathbf{G}^{(N)}(n) = \boldsymbol{\mu} \right] &= \prod_{i=0}^M \left(\mathbb{E} [\nu_{i1} \cdots \nu_{iK_i}] \prod_{k=1}^{K_i} \langle Q_i^{(N)} f_{ik}, \mu_i \rangle \right. \\ &+ \left. \frac{1}{N_i} \mathbb{E} [(\nu_{i1})_2 \nu_{i2} \cdots \nu_{iK_i-1}] \sum_{q < r} \prod_{\substack{k=1 \\ k \neq q, r}}^{K_i} \langle Q_i^{(N)} f_{ik}, \mu_i \rangle \langle (Q_i^{(N)} f_{iq})(Q_i^{(N)} f_{ir}), \mu_i \rangle + o(c_{N_i}) \right) \\ &= \prod_{i=0}^M \mathbb{E} [\nu_{i1} \cdots \nu_{iK_i}] \prod_{k=1}^{K_i} \langle Q_i^{(N)} f_{ik}, \mu_i \rangle + \sum_{i=0}^M \prod_{\substack{j=0 \\ j \neq i}}^M \mathbb{E} [\nu_{j1} \cdots \nu_{jK_j}] \prod_{k=1}^{K_j} \langle Q_j^{(N)} f_{jk}, \mu_j \rangle \\ &\quad \times \frac{1}{N_i} \mathbb{E} [(\nu_{i1})_2 \nu_{i2} \cdots \nu_{iK_i-1}] \sum_{q < r} \prod_{\substack{k=1 \\ k \neq q, r}}^{K_i} \langle Q_i^{(N)} f_{ik}, \mu_i \rangle \langle (Q_i^{(N)} f_{iq})(Q_i^{(N)} f_{ir}), \mu_i \rangle + o(c_{N_i}). \end{aligned}$$

Further, recalling (18), by assumption

$$Q_i^{(N)} = I + c_{N_i} B_i^{(N)} + \mathcal{O}(c_{N_i}),$$

we have

$$\begin{aligned}
& \mathbb{E} \left[F(\mathbf{G}^{(N)}(n+1)) \middle| \mathbf{G}^{(N)}(n) = \boldsymbol{\mu} \right] \\
&= \prod_{i=0}^M \mathbb{E} [\nu_{i1} \cdots \nu_{iK_i}] \prod_{k=1}^{K_i} \langle f_{ik}, \mu_i \rangle \\
&+ \sum_{i=0}^M c_{N_i} \mathbb{E} [\nu_{i1} \cdots \nu_{iK_i}] \prod_{\substack{j=0 \\ j \neq i}}^M \mathbb{E} [\nu_{j1} \cdots \nu_{jK_j}] \prod_{k=1}^{K_j} \langle f_{jk}, \mu_j \rangle \sum_{q=1}^{K_i} \langle B_i^{(N)} f_{iq}, \mu_i \rangle \prod_{\substack{k=1 \\ k \neq q}}^{K_i} \langle f_{ik}, \mu_i \rangle \\
&+ \sum_{i=0}^M \prod_{\substack{j=0 \\ j \neq i}}^M \mathbb{E} [\nu_{j1} \cdots \nu_{jK_j}] \prod_{k=1}^{K_j} \langle f_{jk}, \mu_j \rangle \frac{1}{N_i} \mathbb{E} [(\nu_{i1})_2 \nu_{i2} \cdots \nu_{iK_i-1}] \sum_{q < r} \prod_{\substack{k=1 \\ k \neq q, r}}^{K_i} \langle f_{ik}, \mu_i \rangle \langle f_{iq} f_{ir}, \mu_i \rangle \\
&+ o(c_{N_i}). \quad (23)
\end{aligned}$$

Now recall,

$$(B_i^{(N)} f)(x) := \frac{\varpi_i}{2} \left(\int f(y) G^{(N)}(n)(dy) - f(x) \right),$$

and, from (17), we have

$$G^{(N)}(n) = \frac{1}{\sum_{k=0}^M N_k} \sum_{i=0}^M N_i G_i^{(N)}(n) = \frac{1}{\sum_{k=0}^M N_k} \sum_{i=0}^M N_i \mu_i = \mu_0 + \mathcal{O}\left(\frac{N_i}{N_0}\right),$$

so that

$$(B_i^{(N)} f)(x) = \frac{\varpi_i}{2} \left(\int f(y) \mu_0(dy) - f(x) \right) + o(1).$$

Now, recalling Lemma 4, we have

$$\begin{aligned}
\mathbb{E} [\nu_{i1} \cdots \nu_{iK_i}] &= 1 - \binom{K_i}{2} \frac{(N_i)_{K_i-1}}{(N_i)_{K_i}} \mathbb{E} [(\nu_{i1})_2 \nu_{i2} \cdots \nu_{iK_i-1}] - o(c_{N_i}) \\
&= 1 - \binom{K_i}{2} \frac{1}{N_i - K_i + 1} \mathbb{E} [(\nu_{i1})_2 \nu_{i2} \cdots \nu_{iK_i-1}] - o(c_{N_i}) \\
&= 1 - \binom{K_i}{2} \left(\frac{1}{N_i} + \frac{K_i - 1}{N_i(N_i - K_i + 1)} \right) \mathbb{E} [(\nu_{i1})_2 \nu_{i2} \cdots \nu_{iK_i-1}] - o(c_{N_i}) \\
&= 1 - \binom{K_i}{2} \frac{1}{N_i} \mathbb{E} [(\nu_{i1})_2 \nu_{i2} \cdots \nu_{iK_i-1}] - o(c_{N_i}),
\end{aligned}$$

so that

$$\begin{aligned}
F(\boldsymbol{\mu}) &= \prod_{i=0}^M \prod_{k=1}^{K_i} \langle f_{ik}, \mu_i \rangle \\
&= \prod_{i=0}^M \left(\mathbb{E} [\nu_{i1} \cdots \nu_{iK_i}] \prod_{k=1}^{K_i} \langle f_{ik}, \mu_i \rangle + \binom{K_i}{2} \frac{1}{N_i} \mathbb{E} [(\nu_{i1})_2 \nu_{i2} \cdots \nu_{iK_i-1}] \prod_{k=1}^{K_i} \langle f_{ik}, \mu_i \rangle + o(c_{N_i}) \right) \\
&= \prod_{i=0}^M \mathbb{E} [\nu_{i1} \cdots \nu_{iK_i}] \prod_{k=1}^{K_i} \langle f_{ik}, \mu_i \rangle \\
&+ \sum_{i=0}^M \prod_{\substack{j=0 \\ j \neq i}}^M \mathbb{E} [\nu_{j1} \cdots \nu_{jK_j}] \prod_{k=1}^{K_j} \langle f_{jk}, \mu_j \rangle \frac{1}{N_i} \mathbb{E} [(\nu_{i1})_2 \nu_{i2} \cdots \nu_{iK_i-1}] \sum_{q < r} \prod_{k=1}^{K_i} \langle f_{ik}, \mu_i \rangle + o(c_{N_i}). \quad (24)
\end{aligned}$$

Thus, taking the difference of (23) and (24) and using Lemmas 4 and 5 respectively to replace $\mathbb{E}[\nu_{i1} \cdots \nu_{iK_i}]$ and $\frac{1}{N_i} \mathbb{E}[(\nu_{i1})_2 \nu_{i2} \cdots \nu_{iK_i-1}]$ by $1 - \mathcal{O}(c_{N_i})$ and $c_{N_i} + o(c_{N_i})$, we see that

$$\mathbb{E} \left[F(\mathbf{G}^{(N)}(n+1)) \middle| \mathbf{G}^{(N)}(n) = \boldsymbol{\mu} \right] - F(\boldsymbol{\mu}) = c_{N_i} \sum_{i=0}^M (\mathcal{G}_i F)(\boldsymbol{\mu}) + o(c_{N_i}).$$

The first assertion follows directly.

We now observe that, restricted to the space of functions

$$C_i := \left\{ F(\boldsymbol{\mu}) = \prod_{k=1}^{K_i} \langle f_{ij}, \mu_i \rangle \middle| K_i \in \mathbb{N}_0, f_{ij} \in C([0, 1]) \right\} \subseteq C(\mathcal{P}([0, 1])),$$

\mathcal{G}_i is exactly the generator (4.4) of the infinitely many alleles model of Chapter 10 of [57]. In particular, Theorem 4.1 of the same chapter tells us that given a fixed initial measure $\tilde{\mu}_i \in \mathcal{P}(\mathcal{P}([0, 1]))$, the martingale problem for $(\mathcal{G}_i, \tilde{\mu}_i)$ is well posed, *i.e.*, there exists a unique in distribution process $G_i(t)$ with initial value $G_i(0)$ distributed according to $\tilde{\mu}_i$ with generator \mathcal{G}_i . Moreover, using Theorem 1.1 of Chapter 6 of [57], we see that $G_i(\gamma_i t)$ is the unique process with generator $\gamma_i \mathcal{G}_i$. We can thus appeal to Theorem 10.1 in [57] to conclude that given an initial measure $\tilde{\boldsymbol{\mu}} = \tilde{\mu}_0 \otimes \cdots \otimes \tilde{\mu}_M$, then the martingale problem for

$$\sum_{i=1}^M \gamma_i \mathcal{G}_i$$

is well posed and has solution $G_0(0) \otimes G_1(\gamma_1 t) \otimes \cdots \otimes G_M(\gamma_M t)$.

Given convergence of the generators, and well-posedness of the limiting generator, the second assertion then follows by Lemma 5.1 in Chapter 4 of [57]. \square

Finally, we conclude this section by observing that our characterization of the limiting generator in terms of the generator of the infinitely many alleles diffusion model also allows us to characterize the stationary distribution:

Corollary 1. *The stationary process for the islands is the joint law of M independent Dirichlet processes with scaling parameters ϖ_i and base probability measure μ_0 , $DP(\varpi_i, \mu_0)$.*

Proof. This is immediate from the result for a single copy of the infinitely many alleles model. See *e.g.*, Theorem 4.1, Chapter 9 in [57]. \square

1.5 Long-Term Behaviour

In the previous section, we simply assumed that the mainlands were asymptotically smaller in size (as measured by the coalescence probability of two randomly selected individuals) than the mainland, in order to show that the Cannings' UNTB converged to a sum of independent copies of the infinitely many alleles diffusion process, with migration from the mainland playing the role of mutation. In this section, we will show that in a slow timescale, the dynamics on the mainland converge to the standard infinitely many alleles model as well, from which we can conclude, as before, that the stationary distribution of the mainland is that of the Dirichlet process $DP(\theta, \lambda)$, where λ is the Lebesgue measure on $[0, 1]$. Thus, after a transient period, the mainland will approach a measure $\mu_0 \sim DP(\theta, \lambda)$, whereas the islands will converge on Hierarchical Dirichlet Processes $DP(\varpi_i, \mu_0)$ [19].

Let $\tilde{\nu}_i$, $i = 1, \dots, n$, be the law of the stationary process $DP(\varpi_i, \mu_0)$ from Corollary 1 above, and let $\tilde{\nu} = \tilde{\nu}_1 \otimes \cdots \otimes \tilde{\nu}_M$, *i.e.*, given a function $F \in C(\mathcal{P}([0, 1])^M)$,

$$\int F(\boldsymbol{\mu}) \tilde{\nu}(d\boldsymbol{\mu}) = \int \cdots \int F(\mu_1, \dots, \mu_M) \tilde{\nu}_1(d\mu_1) \cdots \tilde{\nu}_M(d\mu_M),$$

then $\tilde{\nu}$ is a stationary distribution for $\mathbf{G}(t)$: we have

$$\int (\mathcal{G}F)(\boldsymbol{\mu}) \tilde{\nu}(d\boldsymbol{\mu}) = 0$$

for all $F \in C(\mathcal{P}([0, 1])^{M+1})$, or equivalently, writing $\mathcal{T}(t)$ for the semi-group generated by \mathcal{G} , (*i.e.*,

$$(\mathcal{T}(t)F)(\boldsymbol{\mu}) = \mathbb{E}[F(\mathbf{G}(t)) | \mathbf{G}(0) = \boldsymbol{\mu}],$$

where $\mathbf{G}(t)$ is the process with generator \mathcal{G} of Proposition 1) we have

$$\int (\mathcal{T}(t)F)(\boldsymbol{\mu}) \tilde{\nu}(d\boldsymbol{\mu}) = \int F(\boldsymbol{\mu}) \tilde{\nu}(d\boldsymbol{\mu})$$

for all $F \in C(\mathcal{P}([0, 1])^M)$.

We start by showing that as $t \rightarrow \infty$, $\mathbf{G}(t)$ converges to a stationary process \mathbf{G}^* distributed according to $\tilde{\nu}$, (*i.e.*,

$$\mathbb{P}\{\mathbf{G}^*(t) \in A | \mathbf{G}^*(0) \sim \tilde{\nu}\} = \tilde{\nu}(A)$$

for all subsets $A \subseteq \mathcal{P}([0, 1])^{M+1}$). To this end, we begin with a series of lemmas, which are essentially the same as results appearing in [63]:

Lemma 8. *Let $\boldsymbol{\mu} = \mu_0 \otimes \cdots \otimes \mu_M \in \mathcal{P}([0, 1])^{M+1}$ and let $F(\boldsymbol{\mu}) = \prod_{i=0}^M \prod_{k=1}^{K_i} \langle f_{ik}, \mu_i \rangle \in \mathcal{C}$. Let $K = \sum_{i=0}^M K_i$ be the degree of F . If $K \geq 1$, there exists a scalar $\lambda > 0$ and a function ψ , which is a sum of functions of the same form as F , but of degree $K - 1$, such that*

$$\mathcal{G}F = -\lambda F + \psi.$$

Thus,

$$(\mathcal{T}(t)F)(\boldsymbol{\mu}) = e^{-\lambda t} F + \int_0^t e^{-\lambda(t-s)} \mathcal{T}(s)\psi ds.$$

Proof. Recalling Equation (22), we have

$$\begin{aligned} (\mathcal{G}_i F)(\boldsymbol{\mu}) &= \sum_{i=0}^M \sum_{1 \leq j \neq k \leq K_i} (\langle f_{ij} f_{ik}, \mu_i \rangle - \langle f_{ij}, \mu_i \rangle \langle f_{ik}, \mu_i \rangle) \prod_{l \neq j, k} \langle f_{il}, \mu_i \rangle \\ &\quad + \sum_{i=0}^M \sum_{j=1}^{K_i} \frac{\varpi_i}{2} \langle f_{ij}, x_0 - \mu_i \rangle \prod_{k \neq j} \langle f_{ik}, \mu_i \rangle \\ &= - \left(\sum_{i=0}^M \frac{K_i(K_i - 1)}{2} + \frac{\varpi_i}{2} \right) F + \sum_{i=0}^M \langle \sum_{1 \leq j \neq k \leq K_i} f_{ij} f_{ik}, \mu_i \rangle \prod_{l \neq j, k} \langle f_{il}, \mu_i \rangle \\ &\quad + \sum_{i=0}^M \langle \sum_{j=1}^{K_i} \frac{\varpi_i}{2} f_{ij}, x_0 \rangle \prod_{k \neq j} \langle f_{ik}, \mu_i \rangle, \end{aligned}$$

giving the first statement. In particular, if $K = 1$, say $K_i = 1$, we have

$$\mathcal{G}F = -\frac{\gamma_i \varpi_i}{2} F + \langle \frac{\gamma_i \varpi_i}{2} f_{i1}, x_0 \rangle.$$

For the second statement, we observe that

$$\frac{d}{dt} e^{\lambda t} \mathcal{T}(t)F = e^{\lambda t} (\lambda \mathcal{T}(t)F + \mathcal{T}(t)\mathcal{G}F) = e^{\lambda t} \mathcal{T}(t)\psi.$$

The result follows by integrating both sides over $(0, t)$. \square

With this lemma, we can show that the process $\mathbf{G}(t)$ is ergodic, *i.e.*, the distribution of $\mathbf{G}(t)$ converges on $\tilde{\nu}$, independently of the initial condition.

Proposition 2. *Let $F \in C(\mathcal{P}([0, 1]^M))$. As $t \rightarrow \infty$,*

$$\lim_{t \rightarrow \infty} \left\| \mathcal{T}(t)F - \int F(\boldsymbol{\mu}) \tilde{\nu}(d\boldsymbol{\mu}) \right\| = 0.$$

Proof. Since they are convergence-determining, it suffices to show the result for functions of the form $F \in \mathcal{C}$. We then have

$$(\mathcal{T}(t)F) = e^{-\lambda t}F + \int_0^t e^{-\lambda(t-s)}\mathcal{T}(s)\psi ds$$

for $\lambda > 0$ and ψ of degree $K - 1$. Integrating both sides, and recalling that

$$\int (\mathcal{T}(t)F)(\boldsymbol{\mu})\tilde{\nu}(d\boldsymbol{\mu}) = \int F(\boldsymbol{\mu})\tilde{\nu}(d\boldsymbol{\mu})$$

we have

$$\int F(\boldsymbol{\mu})\tilde{\nu}(d\boldsymbol{\mu}) = e^{-\lambda t} \int F(\boldsymbol{\mu})\tilde{\nu}(d\boldsymbol{\mu}) + \int_0^t e^{-\lambda(t-s)}\mathcal{T}(s) \int \psi(\mathbf{x})\tilde{\nu}(d\boldsymbol{\mu}) ds,$$

so that

$$\left\| \mathcal{T}(t)F - \int F(\boldsymbol{\mu})\tilde{\nu}(d\boldsymbol{\mu}) \right\| \leq e^{-\lambda t} \int F(\boldsymbol{\mu})\tilde{\nu}(d\boldsymbol{\mu}) + \int_0^t e^{-\lambda(t-s)} \left\| \mathcal{T}(s)\psi - \int \psi(\boldsymbol{\mu})\tilde{\nu}(d\boldsymbol{\mu}) \right\| ds.$$

The first term on the right hand side clearly vanishes as $t \rightarrow \infty$; for the latter, we can iterate the above inequality, relying on the fact that the process will eventually terminate when the degree reaches 1; when $K = 1$, say $\psi(\boldsymbol{\mu}) = \langle f_{i1}, \mu_i \rangle$, we have

$$(\mathcal{T}(t)\psi) = e^{-\frac{\omega_i}{2}t}\psi + \int_0^t e^{-\frac{\omega_i}{2}(t-s)}\mathcal{T}(s)\langle \frac{\omega_i}{2}f_{i1}, x_0 \rangle ds = e^{-\frac{\omega_i}{2}t}\psi + \int_0^t e^{-\lambda(t-s)}\langle \frac{\omega_i}{2}f_{i1}, x_0 \rangle ds$$

whereas

$$\int \psi(\boldsymbol{\mu})\tilde{\nu}(d\boldsymbol{\mu}) = e^{-\frac{\omega_i}{2}t} \int \psi(\boldsymbol{\mu})\tilde{\nu}(d\boldsymbol{\mu}) + \int_0^t e^{-\lambda(t-s)}\langle \frac{\omega_i}{2}f_{i1}, x_0 \rangle ds,$$

so that

$$\left\| \mathcal{T}(t)\psi - \int \psi(\boldsymbol{\mu})\tilde{\nu}(d\boldsymbol{\mu}) \right\| = e^{-\frac{\omega_i}{2}t} \left\| \psi - \int \psi(\boldsymbol{\mu})\tilde{\nu}(d\boldsymbol{\mu}) \right\| \rightarrow 0$$

as $t \rightarrow \infty$. □

Define a linear map \mathcal{P} on $C(\mathcal{P}([0, 1]^M))$ by

$$\mathcal{P}F = \int F(\boldsymbol{\mu})\tilde{\nu}(d\boldsymbol{\mu}),$$

i.e., \mathcal{P} sends $F \in C(\mathcal{P}([0, 1]^M))$ to a constant function; more generally, if $F \in C(\mathcal{P}([0, 1]^{M+1}))$, $\mathcal{P}F$ is a function of μ_0 alone. In particular,

$$(\mathcal{P}F)(\mu_0) = \mathbb{E}[F(\mu_0, G_1(t), \dots, G_M(t)) | G_i(0) \sim \tilde{\nu}_i],$$

so that applying the operator \mathcal{P} is equivalent to conditioning on the islands being at their stationary state.

Note that $\mathcal{P}^2 = \mathcal{P}$, so that \mathcal{P} is a projection. Moreover,

$$\mathcal{P}(\mathcal{G}F) = \int (\mathcal{G}F) \tilde{\nu}(d\boldsymbol{\mu}) = 0,$$

so the range of \mathcal{G} is contained in the null space of \mathcal{P} , $\mathcal{R}(\mathcal{G}) \subseteq \mathcal{N}(\mathcal{P})$, whereas $\mathcal{G}1 = 0$, so that $\mathcal{R}(\mathcal{P}) \subseteq \mathcal{N}(\mathcal{G})$. In fact, we have:

Lemma 9. \mathcal{P} is the spectral projection onto $\mathcal{N}(\mathcal{G})$.

Proof. By definition, the spectral projection onto $\mathcal{N}(\mathcal{G})$, \mathcal{Q} , is the residue of the resolvent of \mathcal{G} at $\lambda = 0$:

$$\mathcal{Q} = \lim_{\lambda \rightarrow 0^+} \lambda(\lambda - \mathcal{G})^{-1} = \lim_{\lambda \rightarrow 0^+} \lambda \int_0^\infty e^{-\lambda t} \mathcal{T}(t) dt.$$

Now, fix $\varepsilon > 0$ and choose $t_0 > 0$ so that $\|\mathcal{T}(t) - \mathcal{P}\| < \varepsilon$ for $t > t_0$. Then, for $\lambda > 0$,

$$\begin{aligned} \left\| \lambda \int_0^\infty e^{-\lambda t} \mathcal{T}(t) dt - \mathcal{P} \right\| &= \left\| \lambda \int_0^\infty e^{-\lambda t} (\mathcal{T}(t) - \mathcal{P}) dt \right\| \\ &\leq \lambda \int_0^\infty e^{-\lambda t} \|\mathcal{T}(t) - \mathcal{P}\| dt \\ &= \lambda \int_0^{t_0} e^{-\lambda t} \|\mathcal{T}(t) - \mathcal{P}\| dt + \lambda \int_{t_0}^\infty e^{-\lambda t} \|\mathcal{T}(t) - \mathcal{P}\| dt \\ &\leq \lambda t_0 \sup_{t \leq t_0} \|\mathcal{T}(t) - \mathcal{P}\| + \varepsilon. \end{aligned}$$

$\|\mathcal{T}(t) - \mathcal{P}\|$ is a continuous function, and thus bounded on $[0, t_0]$. Thus the first term vanishes as $\lambda \rightarrow 0^+$, whereas ε can be chosen arbitrarily small. We conclude $\mathcal{Q} = \mathcal{P}$. \square

With this, we are able to obtain our final result.

Proposition 3. Assume, as before, that

$$\lim_{N \rightarrow \infty} \frac{c_N}{a_N} = 0.$$

Let \mathcal{P} be the projection defined above. Define an operator \mathcal{G}_0 on \mathcal{C}_0 by

$$\begin{aligned} (\mathcal{G}_0 F)(\boldsymbol{\mu}) &= \left(\sum_{q=1}^{K_0} \frac{\theta}{2} \langle f_{0q}, \lambda - \mu_0 \rangle \prod_{\substack{k=1 \\ k \neq q}}^{K_0} \langle f_{0k}, \mu_0 \rangle \right. \\ &\quad \left. + \frac{1}{2} \sum_{q \neq r}^{K_0} \prod_{\substack{k=1 \\ k \neq q, r}}^{K_0} \langle f_{0k}, \mu_0 \rangle (\langle f_{0q} f_{0r}, \mu_0 \rangle - \langle f_{0q}, \mu_0 \rangle \langle f_{0r}, \mu_0 \rangle) \right), \quad (25) \end{aligned}$$

and let $\mathcal{T}_0(t)$ be the semigroup generated by $\mathcal{P}\mathcal{G}_0$. Then, for all $F \in C(\mathcal{P}([0, 1])^M)$, and all $\delta \in (0, 1)$ we have

$$(I + \mathcal{G}^{(N)})^{\lfloor c_N^{-1} t \rfloor} F \rightarrow \mathcal{T}_0(t) \mathcal{P}F$$

uniformly in $\delta \leq t \leq \delta^{-1}$. If in addition, we assume that $G_i(0) \sim \tilde{\nu}_i$ for all $i = 1, \dots, M$, and $G_0(t)$ is a stochastic process with generator \mathcal{G}_0 , then

$$\mathbf{G}^{(N)}(\lfloor c_N^{-1} t \rfloor) \Rightarrow \mathbf{G}(t) = G_0(t) \otimes G_1(t) \cdots \otimes G_M(t),$$

where the processes $G_i(t)$ are stationary for all $i = 1, \dots, M$.

$$\mathcal{G}^{(N)} = c_N^{-1} a_N \mathcal{G} + c_N^{-1} \mathcal{G}_0 + \text{lower order terms}$$

where $\mathcal{H}\mathcal{P} \equiv 0$. Now $c_N^{-1} a_N \rightarrow \infty$ as $N \rightarrow \infty$, so the first term dominates. $c_N^{-1} a_N$ is essentially the rate at which the first term shapes the dynamics of the process, and so as N grows large, the first term, which acts only on the islands, causes them to rapidly approach their equilibrium state (which, as we have already seen, corresponds to projection by \mathcal{P}). The first term, however, has no effect on the mainland. Moreover, the mainland only changes at the slower rate c_N^{-1} . Thus, the first term has already forced the faster terms to equilibrium, and we can assume that they are at equilibrium when we consider the mainland. Finally, the first two terms completely specify the limit, so what remains can only contribute a higher order correction. This is essentially the infinite dimensional analogue of the following simple dynamical system:

$$\begin{aligned} \dot{x} &= -Nax + f(x, y), \\ \dot{y} &= -\sqrt{N}by + g(x, y), \end{aligned}$$

for $a, b > 0$. Using variation of constants, we have

$$\begin{aligned} x(t) &= e^{-Nat} x(0) + \int_0^t e^{-Na(t-s)} f(x(s), y(s)) ds, \\ y(t) &= e^{-\sqrt{N}bt} y(0) + \int_0^t e^{-\sqrt{N}b(t-s)} g(x(s), y(s)) ds. \end{aligned}$$

Thus, provided f and g are bounded,

$$\int_0^t e^{-Na(t-s)} f(x(s), y(s)) ds \leq \frac{1}{Na} \|f\|,$$

and

$$\int_0^t e^{-\sqrt{N}b(t-s)} g(x(s), y(s)) ds \leq \frac{1}{\sqrt{N}b} \|g\|,$$

so that as $N \rightarrow \infty$, we have $x(t) = 0 + \mathcal{O}(\frac{1}{N})$. We can thus substitute this back into the equation for $y(t)$ to conclude that

$$y(t) = e^{-\sqrt{N}bt} y(0) + \int_0^t e^{-\sqrt{N}b(t-s)} g(0, y(s)) ds + \mathcal{O}(\frac{1}{N}),$$

(setting $x(t) \equiv 0$ is equivalent to the action of the projection \mathcal{P}). Thus, similarly, $y(t) = 0 + \mathcal{O}(\frac{1}{\sqrt{N}})$.

Remark 11. It is necessary to assume $G_i(0) \sim \tilde{\nu}_i$ to obtain continuity of $\mathcal{T}_0(t)\mathcal{P}$ at $t = 0$, which in turn is required to ensure weak convergence. More generally, Proposition 3 tell us that in the slow timescale, the island demes instantaneously jump to their stationary states, and henceforth evolve as stationary processes; see [64] and [56] for more detailed discussions of processes with this behaviour.

Proof. Calculations essentially identical to those in Proposition 1 show that, when restricted to \mathcal{C}_0 , $c_N^{-1} \mathcal{G}^{(N)} = \mathcal{G}_0 + o(c_N)$, with the primary difference being with the operator $Q_0^{(N)}$. Here,

$$Q_0^{(N)} = I + c_N B_0^{(N)} + c_N B + o(c_N),$$

where, as before

$$(B_0^{(N)} f)(x) = \frac{\varpi_i}{2} (\langle f, \mu_0 \rangle - f(x)) + o(1),$$

but now

$$(Bf)(x) = \frac{\theta}{2} \int_0^1 f(y) dy - f(x) = \theta(\langle f, \lambda \rangle - f(x))$$

(recall that λ is Lebesgue measure, $\lambda(dx) = dx$) is of the same asymptotic order. Moreover, we now only consider terms of the form $\langle Q_0^{(N)} f_{0k}, \mu_0 \rangle$, and

$$\langle B_0^{(N)} f_{0k}, \mu_0 \rangle = \frac{\varpi_i}{2} (\langle f, \mu_0 \rangle - \langle f, \mu_0 \rangle) + o(1),$$

which vanishes in the limit. Thus,

$$c_N \langle Q_0^{(N)} f_{0k}, \mu_0 \rangle - \langle f_{0k}, \mu_0 \rangle = \frac{\theta}{2} \int_0^1 f(y) dy - f(x) + o(1) = \theta(\langle f, \lambda \rangle - f(x)) + o(1),$$

giving the corresponding terms in the generator (25).

The first statement is then a restatement of Corollary 7.7, Chapter 1 of [57]; translating our notation into theirs, we have

$$\begin{aligned} \varepsilon_N &= c_N, \\ \alpha_N &= c_N^{-1} a_N, \\ A_N &= c_N^{-1} \mathcal{G}^N, \end{aligned}$$

$B = \mathcal{G}$, and $A = \mathcal{G}_0$. That \mathcal{G}_0 generates a strongly continuous semigroup is Theorem 4.1, Chapter 10 of [57], which we used previously.

The second statement is a consequence of Corollary 8.9, Chapter 4, [57], where our initial condition ensures continuity of the semigroup $\mathcal{T}_0(t)$ at $t = 0$. \square

2 Gibbs Sampling for the UNTB-HDP

2.1 Observed abundances

The observed data takes the form of an $N \times S$ matrix of counts \mathbf{X} whose elements x_{ij} are the observed frequency of species j in community sample i . Here, N denotes the total number of communities and S the total number of different species found in those communities. We will also denote the row vectors of \mathbf{X} , which give the observed frequency distribution of species in each individual sample, by \bar{X}_i , $i = 1, \dots, N$. The size of each sample is simply $J_i = \sum_{j=1}^S x_{ij}$.

2.2 Neutral-HDP model

$$\bar{\beta} | \theta \sim \text{Stick}(\theta), \tag{26}$$

$$\bar{\pi}_i | I_i, \bar{\beta} \sim \text{DP}(I_i, \bar{\beta}), \tag{27}$$

$$\bar{X}_i | \bar{\pi}_i, J_i \sim \text{MN}(J_i, \bar{\pi}_i). \tag{28}$$

This model for the observed frequencies can be interpreted as the generation of an infinite dimensional metacommunity distribution $\bar{\beta}$ which is obtained from a stick-breaking or GEM distribution with concentration parameter θ . From this, for each community i we sample using the Dirichlet process a vector of taxa probabilities $\bar{\pi}_i$ which has concentration I_i , the immigration rate for that site, and base distribution

$\bar{\beta}$. Finally, we sample the observed frequencies for each community \bar{X}_i from $\bar{\pi}_i$ using the multinomial distribution. We also include gamma hyper-priors for θ and the I_i :

$$\theta|\alpha, \zeta \sim \text{Gamma}(\alpha, \zeta), \quad (29)$$

$$I_i|\eta, \sim \text{Gamma}(\eta, \kappa), \quad (30)$$

where α, ζ, η and κ are all constants. This completes the definition of our model.

2.3 Finite dimensional representation

In any given sample although the potential number of species is infinite we only observe S different types. It is convenient therefore to represent the model in terms of these finite dimensional number of types and one further class corresponding to all unobserved species. We will derive this as the limit of L total types as $L \rightarrow \infty$. We will represent the proportions of the S observed species explicitly as β_k with $k = 1, \dots, S$ and the unrepresented component as $\beta_u = \sum_{k=S+1}^L \beta_k$. Let $\theta_r = \theta/L$ and $\theta_u = \theta(L-S)/L$, then we will have a Dirichlet prior on $\bar{\beta} \sim \text{Dir}(\theta_r, \dots, \theta_r, \theta_u)$. In this finite dimensional representation we can also determine the distributions in the local communities:

$$\bar{\pi}_i \sim \text{Dir}(I_i\beta_1, \dots, I_i\beta_S, I_i\beta_u). \quad (31)$$

We can then marginalise the local community distributions and derive the probability of the observed frequencies given $\bar{\beta}$:

$$P(\mathbf{X}|\bar{\beta}, I_1, \dots, I_N) = \prod_{i=1}^N \frac{J_i!}{X_{i1}! \dots X_{iS}!} \frac{\Gamma(I_i)}{\Gamma(J_i + I_i)} \prod_{j=1}^S \frac{\Gamma(x_{ij} + I_i\beta_j)}{\Gamma(I_i\beta_j)}. \quad (32)$$

2.4 Gibbs sampling

To devise a Gibbs sampling strategy we need to determine the full conditional distributions of the parameters we wish to sample, θ and I_i , for $i = 1, \dots, N$. Our starting point will be the joint distribution of these parameters and the data, that is, Equation 32 multiplied by the prior distributions for $\bar{\beta}$, θ and I_i , marginalised over $\bar{\beta}$:

$$P(\theta, I_1, \dots, I_N, \mathbf{X}) = \int_{\bar{\beta}} P(\mathbf{X}|\bar{\beta}, I_1, \dots, I_N) P(\bar{\beta}|\theta) d\bar{\beta} \text{Gamma}(\theta|\alpha, \zeta) \prod_{i=1}^N \text{Gamma}(I_i|\eta, \nu). \quad (33)$$

The key to simplifying this expression is to expand the terms $\Gamma(x_{ij} + I_i\beta_j)/\Gamma(I_i\beta_j)$ in Equation 10 as polynomials [19]:

$$\frac{\Gamma(x_{ij} + I_i\beta_j)}{\Gamma(I_i\beta_j)} = \sum_{T_{ij}=0}^{T_{ij}=x_{ij}} s(x_{ij}, T_{ij})(I_i\beta_j)^{T_{ij}}, \quad (34)$$

where the coefficients $s(x_{ij}, T_{ij})$ are unsigned Stirling numbers of the first kind. We substitute these sums into Equation 33 and then introduce the T_{ij} and $\bar{\beta}$ as auxilliary variables to give:

$$Q(\theta, \bar{\beta}, I_1, \dots, I_N, T_{ij}) \propto \left(\prod_{i=1}^N \frac{J_i!}{X_{i1}! \dots X_{iS}!} \frac{\Gamma(I_i)}{\Gamma(J_i + I_i)} \prod_{j=1}^S s(x_{ij}, T_{ij})(I_i\beta_j)^{T_{ij}} \right) P(\bar{\beta}|\theta) \text{Gamma}(\theta|\alpha, \zeta) \prod_{i=1}^N \text{Gamma}(I_i|\eta, \nu). \quad (35)$$

2.4.1 Full conditional for the ancestral states

43

From Equation 35, we see that the full conditional distribution for the number of ancestors (tables in the Chinese restaurant franchise analogy) of species j in sample i is given by:

$$P(T_{ij}|x_{ij}, I_i, \beta_j) \propto s(x_{ij}, T_{ij})(I_i\beta_j)^{T_{ij}}. \quad (36)$$

The reciprocal of Equation 34 is the normalising constant of this probability distribution and thus:

$$P(T_{ij}|x_{ij}, I_i, \beta_j) = \frac{\Gamma(I_i\beta_j)}{\Gamma(x_{ij} + I_i\beta_j)} s(x_{ij}, T_{ij})(I_i\beta_j)^{T_{ij}}. \quad (37)$$

2.4.2 Full conditional for the metapopulation

In their derivation of a posterior sampling scheme for the hierarchical Dirichlet process mixture model using an augmented Chinese restaurant franchise representation, [19] showed that the full conditional distribution for the metapopulation vector $\bar{\beta}$ was:

$$\bar{\beta} = (\beta_1, \beta_2, \dots, \beta_S, \beta_u) \sim \text{Dir}(T_1, T_2, \dots, T_S, \theta), \quad (38)$$

where $T_{\cdot j} = \sum_{i=1}^N T_{ij}$.

2.4.3 Full conditional for the immigration rates

To derive the full conditional distribution of each I_i given the other parameters we simply pull out all terms that depend on I_i from Equation 35. This gives:

$$P(I_i|T_{ij}) \propto \frac{\Gamma(I_i)}{\Gamma(J_i + I_i)} I_i^{T_i} \text{Gamma}(I_i|\eta, \nu), \quad (39)$$

where $T_i = \sum_{j=1}^S T_{ij}$. We can use the auxiliary variable approach of [37] to develop a Gibbs sampling update for I_i , $i = 1, \dots, N$. Here, for each i , we can write:

$$\frac{\Gamma(I_i)}{\Gamma(I_i + J_i)} = \frac{1}{\Gamma(J_i)} \int_0^1 w_i^{I_i} (1 - w_i)^{J_i - 1} \left(1 + \frac{J_i}{I_i}\right) dw_i \quad (40)$$

(cf. with equation (A.2) of [19]). We now define auxiliary variables $\bar{w} = (w_i)_{i=1}^N$ and $\bar{s} = (s_i)_{i=1}^N$, where each w_i is a variable taking on values in $[0, 1]$ and each s_i is a binary $\{0, 1\}$ variable, and define the following distribution:

$$q(I_i, \bar{w}, \bar{s}) \propto \prod_{i=1}^N I_i^{\eta - 1 + T_i} e^{-\nu I_i} w_i^{I_i} (1 - w_i)^{J_i - 1} \left(\frac{J_i}{I_i}\right)^{s_i} \quad (41)$$

(cf. with equation (A.3) of [19]). Now marginalising q to I_i gives the desired conditional distribution for I_i . Hence q defines an auxiliary variable sampling scheme for I_i . Given \bar{w} and \bar{s} , we have:

$$q(I_i|\bar{w}, \bar{s}) \propto I_i^{\eta - 1 + T_i - s_i} e^{-I_i(\nu - \log w_i)}, \quad (42)$$

which is a Gamma distribution with parameters $\eta + T_i - s_i$ and $\nu - \log w_i$ (cf. with equation (A.4) of [19]). Given I_i , the w_i and s_i are conditionally independent, with distributions:

$$q(w_i|I_i) \propto w_i^{I_i} (1 - w_i)^{J_i - 1} \quad (43)$$

and

$$q(s_i|I_i) \propto \left(\frac{J_i}{I_i}\right)^{s_i}, \quad (44)$$

which are Beta($I_i + 1, J_i$) and Bernoulli($\frac{J_i}{J_i + I_i}$), respectively (cf. with equations (A.5) and (A.6) of [19]).

A direct consequence of the stick-breaking prior for $\bar{\beta}$ is that the probability of observing S species from a total number of $T = \sum_{i=1}^N \sum_{j=1}^S T_{ij}$ ancestors is given by:

$$P(S|\theta, T) = s(T, S)\theta^S \frac{\Gamma(\theta)}{\Gamma(\theta + T)} \quad (45)$$

(cf. with equation (A.7) of [19]). The biodiversity parameter θ does not govern any other aspects of the joint distribution in Equation 35, hence Equation 11, along with the prior for θ in Equation 29, is all that is needed to derive a Gibbs sampling update for θ . The auxiliary variable approach of [37] can also be applied here, which leads to the following auxiliary variable sampling scheme for θ :

$$\theta|\rho, \phi, S \sim \text{Gamma}(\alpha + S - \rho, \zeta - \log \phi), \quad (46)$$

$$\rho|\theta, T \sim \text{Bernoulli}\left(\frac{T}{T + \theta}\right), \quad (47)$$

$$\phi|\theta, T \sim \text{Beta}(\theta + 1, T). \quad (48)$$

2.5 Results

In order to examine how well our HDP estimation approach performed in comparison with existing methods [12, 16, 65], we used a combination of simulated data and real data that had been analysed before. Firstly, we generated 1,000 simulated data sets of three local samples with 1,000 individuals each for the eight parameter combinations given in Table 3. Note that the migration probability is simply $m_i = I_i/(I_i + J_i - 1)$. These data sets were generated using the PARI/GP code provided in [12], which is an urn algorithm based on coalescence theory. We then estimated the parameters using the Gibbs sampling approach based on the HDP approximation and the approximate two stage approach of [16]. Tables 4 and 5 gives the means, coefficients of variation and mean absolute deviations from the true values of our approach and Etienne's two stage approximate method, respectively, across the 1,000 data sets for each parameter combination.

For all parameter combinations considered the HDP approximation outperforms Etienne's approximation as an estimator of θ , as in each case the overall means are closer to the true values and the coefficients of variations and mean absolute deviations from the true values are considerably smaller. The HDP approximation provides a less biased and more reliable estimator of θ than Etienne's approximation.

A similar pattern is observed with the estimates of the immigration probabilities m_i , as for the parameter combinations considered our approach gives lower coefficients of variation and mean absolute deviations from the true value than Etienne's approximate method. Both approximations break down when the immigration rate I is significantly larger than the fundamental biodiversity parameter θ (for example, see the estimates of m_3 for synthetic data sets 1-5 in Tables 4 and 5), but in different ways. Our method underestimates the immigration probability m in such cases, but the standard deviation around that estimate remains low, and thus our estimator for m is biased when $I > \theta$, but as this bias is consistent it would be possible to correct for it. On the other hand, Etienne's approximate approach gives an overall mean over the 1,000 simulated data sets that is much closer to the true value in such a case than our method does. However, the variability around Etienne's approximate estimate of m is much higher because the algorithm often converges to an immigration probability of 1, even when the true value is much lower. It is also worth noting that Etienne's approximate method also breaks down badly for data sets 7 and 8 where the immigration probabilities are very low, whereas the HDP approximation copes much better in such scenarios. Thus, we conclude that the HDP approximation is a better estimator of the neutral model's parameters than Etienne's approximation unless $I \gg \theta$ and the immigration probabilities are close to 1.

Data set	J_i	θ	I_1	I_2	I_3	m_1	m_2	m_3
1	1000	5	111	249.75	666	0.1	0.2	0.4
2	1000	50	111	249.75	666	0.1	0.2	0.4
3	1000	500	111	249.75	666	0.1	0.2	0.4
4	1000	5	10.0909	52.5789	333	0.01	0.05	0.25
5	1000	50	10.0909	52.5789	333	0.01	0.05	0.25
6	1000	500	10.0909	52.5789	333	0.01	0.05	0.25
7	1000	5	1	2.002	4.012	0.001	0.002	0.004
8	1000	50	1	2.002	4.012	0.001	0.002	0.004

Table 3. The parameter values chosen for the synthetic neutral model data sets that composed our simulation study.

Data set	$\hat{\theta}$	CV	MAD	\hat{m}_1	CV	MAD	\hat{m}_2	CV	MAD	\hat{m}_3	CV	MAD
1	5.4092	0.20	0.8950	0.0934	0.29	0.0232	0.1508	0.23	0.0522	0.2002	0.19	0.1998
2	51.5476	0.09	3.9993	0.0990	0.14	0.0114	0.1923	0.15	0.0242	0.3262	0.12	0.0749
3	498.8622	0.07	25.8993	0.0999	0.08	0.0067	0.1982	0.07	0.0119	0.3836	0.07	0.0252
4	5.4477	0.22	1.0088	0.0110	0.42	0.0032	0.0526	0.36	0.0144	0.1417	0.26	0.1083
5	51.7504	0.12	4.8836	0.0101	0.21	0.0017	0.0504	0.17	0.0065	0.2211	0.16	0.0387
6	488.8805	0.10	40.7537	0.0100	0.17	0.0014	0.0503	0.10	0.0040	0.2495	0.09	0.0171
7	5.3388	0.46	1.8189	0.0014	0.96	0.0007	0.0030	0.98	0.0015	0.0066	0.95	0.0035
8	55.0994	0.43	17.2483	0.0010	0.44	0.0004	0.0022	0.34	0.0006	0.0043	0.29	0.0009

Table 4. Estimates of θ and m_i from the various scenarios of simulated data sets of Table 3 using the hierarchical Dirichlet process approximation. The values reported are the means, coefficients of variation and mean absolute deviations from the true value of the parameter estimates over 1000 such data sets.

Data set	$\hat{\theta}$	CV	MAD	\hat{m}_1	CV	MAD	\hat{m}_2	CV	MAD	\hat{m}_3	CV	MAD
1	5.9130	0.40	1.9880	0.1899	1.45	0.1621	0.2763	1.14	0.2300	0.4057	1.10	0.3260
2	51.9033	0.20	8.2626	0.1071	0.44	0.0274	0.2239	0.56	0.0776	0.4231	0.48	0.1556
3	507.2382	0.12	50.4488	0.1006	0.09	0.0070	0.2010	0.09	0.0138	0.4032	0.12	0.0356
4	6.0710	0.45	2.1911	0.0410	3.62	0.0356	0.1177	1.88	0.1042	0.3086	1.11	0.2666
5	54.2026	0.29	12.6540	0.0102	0.55	0.0020	0.0580	0.83	0.0190	0.2897	0.72	0.1440
6	578.4131	0.36	166.5742	0.0100	0.18	0.0014	0.0503	0.13	0.0048	0.2601	0.34	0.0503
7	9.9517	1.41	6.5506	0.0164	7.03	0.0158	0.0348	4.69	0.0338	0.0473	3.88	0.0450
8	860.1590	7.00	824.9333	0.0011	1.61	0.0004	0.0022	0.73	0.0007	0.0075	6.32	0.0045

Table 5. Estimates of θ and m_i from the various scenarios of simulated data sets of Table 3 using Etienne's approximate method. The values reported are the means, coefficients of variation and mean absolute deviations from the true value of the parameter estimates over 1000 such data sets.

In Table 6, we present the average times in seconds of Etienne’s approximate method using the code⁴⁶ given in [16] and PARI/GP’s default settings, and our Gibbs sampling approach coded in C++ when it was run for 50,000 iterations with half of these being conservatively discarded as burn-in. Under these settings, for all but one of the simulated data set scenarios of Table 3, Etienne’s approximate method is two to three times faster than our approach. However, we are being very conservative with sample number and equivalent results could be achieved with as little as 10,000 iterations when the two methods would be of comparable speed.

We were unable to replicate these results using Etienne’s ‘exact’ maximum likelihood method, so instead we quote those that he gave in a similar simulation study [65] in Table 7. We see that Etienne’s ‘exact’ method slightly outperforms the HDP approximation as an estimator of θ , as although the coefficients of variation are broadly similar, the overall means are generally closer to their true values and thus Etienne’s ‘exact’ method is less biased for this parameter. Regarding the estimation of immigration probabilities, the results are comparable when $\theta \leq I$. When $\theta > I$, there is a tendency for Etienne’s ‘exact’ method to overestimate the immigration probability, but not as badly as the HDP approximation underestimates it. The advantage of the HDP approximation is that our code is easier to implement than Etienne’s ‘exact’ method’s PARI/GP algorithm, it is much faster, and our approach can handle the large data sets often encountered in microbiomics.

As an example of how the methods compare on real data, we reanalysed the tropical tree data set used as an example in [12,16,65]. The data consists of three forest plots in Panama called Barro Colorado Island (50 ha), Cocoli (4 ha) and Sherman (5.96 ha), which lie along a precipitation gradient [66]. Table 8 shows the results of the parameter estimation for Etienne’s three methods and our HDP approach. We see that in this case the results from the HDP approximation closely match Etienne’s ‘exact’ method, while his approximate method overestimates θ and underestimates the immigration rates. The matching results of our approach and Etienne’s ‘exact’ method is unsurprising as in this case $\theta \gg I_i$.

References

1. Hutchinson GE (1957) Concluding remarks. *Cold Spring Harbor Symposia on Quantitative Biology* 22: 415-427.
2. Hardin G (1960) The competitive exclusion principle. *Science* 131: 1292-1297.
3. Simberloff D, Dayan T (1991) The guild concept and the structure of ecological communities. *Ann Rev Ecol Syst* 22: 115-143.
4. Caswell H (1976) Community structure: A neutral model analysis. *Ecol Monograph* 46: 327-354.
5. Hubbell SP (2001) *The Unified Neutral Theory of Biodiversity and Biogeography*. Princeton University Press, Princeton, NJ.
6. MacArthur RH, Wilson EO (1967) *The theory of island biogeography*. Princeton University Press, Princeton, N.J.
7. McGill BJ (2003) A test of the unified neutral theory of biodiversity. *Nature* 422: 881-885.
8. Etienne RS, Olf H (2004) A novel genealogical approach to neutral biodiversity theory. *Ecol Lett* 7: 170-175.
9. Etienne RS (2005) A new sampling formula for neutral biodiversity. *Ecol Lett* 8: 253-260.
10. Rosindell J, Hubbell SP, He F, Harmon LJ, Etienne RS (2012) The case for ecological neutral theory. *Trends Ecol Evol* 27: 203-208.

Data set	Etienne's approximation	HDP approximation
1	13.8583	40.6223
2	21.5615	41.1254
3	208.6595	41.5881
4	14.9588	41.8532
5	14.9767	40.6765
6	27.3442	42.4084
7	20.0091	56.1613
8	17.8649	57.5658

Table 6. Average time in seconds that Etienne's approximate method and the HDP approximation took to run on the various scenarios of simulated data sets of Table 3. Note that the HDP approximation was run for 50,000 iterations and half of these were conservatively discarded as burn-in.

Data set	$\hat{\theta}$	CV	\hat{m}_1	CV	\hat{m}_2	CV	\hat{m}_3	CV
1	4.9689	0.21	0.1119	0.44	0.2353	0.49	0.4727	0.50
2	49.9838	0.10	0.1022	0.16	0.2041	0.16	0.4105	0.18
3	501.5142	0.07	0.1005	0.08	0.2009	0.08	0.4007	0.08
4	4.8982	0.25	0.0108	0.43	0.0572	0.46	0.3658	0.70
5	49.9892	0.12	0.0103	0.21	0.0513	0.16	0.2643	0.25
6	504.0792	0.11	0.0101	0.17	0.0504	0.11	0.2521	0.09
7	5.0388	0.45	0.0012	0.67	0.0027	1.27	0.0066	4.85
8	56.0378	0.55	0.0010	0.42	0.0020	0.35	0.0042	0.30

Table 7. Estimates of θ and m_i from the various scenarios of simulated data sets of Table 3 using Etienne's 'exact' maximum likelihood method. The values reported are the means and coefficients of variation over 1000 such data sets, and were obtained from [65].

Method	θ	I_{BCI}	I_C	I_S
Etienne fixed I	259	44.2	44.2	44.2
Etienne approx	342	53.7	30.8	33.9
Etienne 'exact'	235 \pm 23	65.3 \pm 5.9	31.5 \pm 3.9	35.7 \pm 3.9
HDP approx	231 \pm 22	65.5 \pm 5.9	31.6 \pm 3.8	35.8 \pm 3.9

Table 8. Neutral parameter estimates for samples from three local tree communities (Sherman, BCI and Cocoli) in the Panama Canal Zone using Etienne's approaches and the hierarchical Dirichlet process approximation. Standard errors are given for the methods where they are available.

11. Chisholm RA, Pacala SW (2010) Niche and neutral models predict asymptotically equivalent species abundance distributions in high-diversity communities. *Proc Nat Acad Sci USA* 107: 15821-15825.
12. Etienne RS (2007) A neutral sampling formula for multiple samples and an ‘exact’ test of neutrality. *Ecol Lett* 10: 608-618.
13. Etienne RS (2009) Maximum likelihood estimation of neutral model parameters for multiple samples with different degrees of dispersal limitation. *J Theor Biol* 257: 510-514.
14. Munoz F, Couteron P, Ramesh BR, Etienne RS (2007) Estimating parameters of neutral communities: from one single large to several small samples. *Ecology* 88: 2482-2488.
15. Jabot F, Etienne RS, Chave J (2008) Reconciling neutral community models and environmental filtering: theory and an empirical test. *Oikos* 117: 1308-1320.
16. Etienne RS (2009) Improved estimation of neutral model parameters for multiple samples with different degrees of dispersal limitation. *Ecology* 90: 847-852.
17. Hamady M, Walker JJ, Harris JK, Gold NJ, Knight R (2008) Error-correcting barcoded primers for pyrosequencing hundreds of samples in multiplex. *Nat Methods* 5: 235-237.
18. Turnbaugh PJ, Hamady M, Yatsunenko T, Cantarel BL, Duncan A, et al. (2009) A core gut microbiome in obese and lean twins. *Nature* 457: 480-484.
19. Teh YW, Jordan MI, Beal MJ, Blei DM (2006) Hierarchical Dirichlet processes. *J Amer Statist Assoc* 101: 1566-1581.
20. Condit R, Pitman N, Leigh EG, Chave J, Terborgh J, et al. (2002) Beta-diversity in tropical forest trees. *Science* 295: 666-669.
21. Arumugam M, Raes J, Pelletier E, Le Paslier D, Yamada T, et al. (2011) Enterotypes of the human gut microbiome. *Nature* 473: 174-180.
22. Holmes I, Harris K, Quince C (2012) Dirichlet multinomial mixtures: Generative models for microbial metagenomics. *PLoS ONE* 7: e30126.
23. Ding T, Schloss PD (2014) Dynamics and associations of microbial community types across the human body. *Nature* 509: 357-360.
24. Fierer N, Bradford MA, Jackson RB (2007) Toward an ecological classification of soil bacteria. *Ecology* 88: 1354-1364.
25. Philippot L, Bru D, Saby NPA, Cuhel J, Arrouays D, et al. (2009) Spatial patterns of bacterial taxa in nature reflect ecological traits of deep branches of the 16S rRNA bacterial tree. *Environ Microbiol* 11: 3096-3104.
26. Philippot L, Andersson SGE, Battin TJ, Prosser JI, Schimel JP, et al. (2010) The ecological coherence of high bacterial taxonomic ranks. *Nature Rev Microbiol* 8: 523-529.
27. Rasko DA, Rosovitz MJ, Myers GSA, Mongodin EF, Fricke WF, et al. (2008) The pangenome structure of *Escherichia coli*: Comparative genomic analysis of E-coli commensal and pathogenic isolates. *J Bacteriol* 190: 6881-6893.
28. Sloan W, Lunn M, Woodcock S, Head I, Nee S, et al. (2006) Quantifying the roles of immigration and chance in shaping prokaryote community structure. *Environ Microbiol* 8: 732-740.

29. Woodcock S, van der Gast CJ, Bell T, Lunn M, Curtis TP, et al. (2007) Neutral assembly of bacterial communities. *FEMS Microbiol Ecol* 62: 171-180.
30. Jeraldo P, Sipos M, Chia N, Brulc JM, Dhillon AS, et al. (2012) Quantification of the relative roles of niche and neutral processes in structuring gastrointestinal microbiomes. *Proc Nat Acad Sci USA* 109: 9692-9698.
31. McKane A, Alonso D, Sole R (2004) Analytic solution of Hubbell's model of local community dynamics. *Theor Pop Biol* 65: 67-73.
32. Sloan WT, Woodcock S, Lunn M, Head IM, Curtis TP (2007) Modeling taxa-abundance distributions in microbial communities using environmental sequence data. *Microb Ecol* 53: 443-455.
33. Ferguson TS (1973) A Bayesian analysis of some nonparametric problems. *Ann Stat* 1: 209-230.
34. Ewens WJ (1972) The sampling theory of selectively neutral mutations. *Theor Pop Biol* 3: 87-112.
35. Antoniak CE (1974) Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The annals of statistics* : 1152-1174.
36. Mackay DJ (1992) Bayesian interpolation. *Neural Comput* 4: 415-417.
37. Escobar MD, West M (1995) Bayesian density estimation and inference using mixtures. *J Amer Statist Assoc* 90: 577-588.
38. Walker SC (2007) When and why do non-neutral metacommunities appear neutral? *Theor Pop Biol* 71: 318-331.
39. Pyke C, Condit R, Aguilar S, Lao S (2001) Floristic composition across a climatic gradient in a neotropical lowland forest. *J Veg Sci* 12: 553-566.
40. Quince C, Lanzen A, Curtis TP, Davenport RJ, Hall N, et al. (2009) Accurate determination of microbial diversity from 454 pyrosequencing data. *Nat Methods* 6: 639-641.
41. Quince C, Lanzen A, Davenport RJ, Turnbaugh PJ (2011) Removing noise from pyrosequenced amplicons. *BMC Bioinf* 12.
42. Wang Q, Garrity GM, Tiedje JM, Cole JR (2007) Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl Environ Microb* 73: 5261-5267.
43. Youssef N, Sheik CS, Krumholz LR, Najjar FZ, Roe BA, et al. (2009) Comparison of species richness estimates obtained using nearly complete fragments and simulated pyrosequencing-generated fragments in 16S rRNA gene-based environmental surveys. *Appl Environ Microb* 75: 5227-5236.
44. R Development Core Team (2010) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. URL <http://www.R-project.org>. ISBN 3-900051-07-0.
45. Rosindell J, Wong Y, Etienne R (2008) A coalescence approach to spatial neutral ecology. *Ecological Informatics* : 259-271.
46. Minot S, Sinha R, Chen J, Li H, Keilbaugh SA, et al. (2011) The human gut virome: Inter-individual variation and dynamic response to diet. *Genome Res* 21: 1616-1625.
47. Quince C, Lundin E, Andreasson AN, Greco D, Rafter J, et al. (2013) The impact of Crohn's disease genes on healthy human gut microbiota: a pilot study. *Gut* 62: 952-954.

48. Ze X, Duncan SH, Louis P, Flint HJ (2012) *Ruminococcus bromii* is a keystone species for the degradation of resistant starch in the human colon. *ISME J* 6: 1535-1543.
49. Finlay BJ, Fenchel T (2004) Cosmopolitan metapopulations of free-living microbial eukaryotes. *Protist* 155: 237-244.
50. Aldous D (1985) Exchangeability and related topics. In: *École d'Été de Probabilités de Saint-Flour XIII-1983*, Berlin: Springer. pp. 1—198.
51. Hoppe FM (1984) Pólya-like urns and the Ewens' sampling formula. *J Math Biol* 20: 91-94.
52. Cannings C (1974) The latent roots of certain markov chains arising in genetics: A new approach, I. haploid models. *Adv Appl Prob* 6: 260-290.
53. Parsons TL, Quince C (2007) Fixation in haploid populations exhibiting density dependence II: The quasi-neutral case. *Theor Popul Biol* 72: 468-479.
54. Parsons TL, Quince C, Plotkin JB (2008) Expected times to absorption and fixation for quasi-neutral and neutral haploid populations exhibiting density dependence. *Theor Popul Biol* 74: 302-310.
55. Parsons TL, Quince C, Plotkin JB (2010) Some consequences of demographic stochasticity in population genetics. *Genetics* 185: 1345-1354.
56. Parsons TL (2012) *Asymptotic Analysis of Some Stochastic Models from Population Dynamics and Population Genetics*. Ph.D. thesis, University of Toronto.
57. Ethier SN, Kurtz TG (1986) *Markov Processes: Characterization and Convergence*. New York: John Wiley and Sons.
58. Möhle M (2001) Forward and backward diffusion approximations for haploid exchangeable population models. *Stochastic Processes Appl* 95: 133-149.
59. Sjödin P, Kaj I, Krone S, Lascoux M, Nordborg M (2005) On the meaning and existence of an effective population size. *Genetics* 169: 1061-70.
60. Möhle M, Sagitov S (2003) Coalescent patterns in diploid exchangeable population models. *J Math Biol* 47: 337-352.
61. Kingman JFC (1982) On the genealogy of large populations. *J Appl Prob* 19A: 27-43.
62. Abramowitz M, Stegun IA (1964) *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. New York: Dover.
63. Ethier SN, Kurtz TG (1981) The infinitely-many-neutral-alleles diffusion model. *Advances in Applied Probability* : 429-452.
64. Katzenberger GS (1991) Solutions of a stochastic differential equation forced onto a manifold by a large drift. *Ann Probab* 19: 1587-1628.
65. Etienne RS (2009) Maximum likelihood estimation of neutral model parameters for multiple samples with different degrees of dispersal limitation. *Journal of Theoretical Biology* 257: 510-514.
66. Condit R, Pitman N, Leigh Jr E, Chave J, Terborgh J, et al. (2002) Beta-diversity in tropical forest trees. *Science* 295: 666-669.