INVITED PAPER

Linking Statistical and Ecological Theory: Hubbell's Unified Neutral Theory of Biodiversity as a Hierarchical Dirichlet Process

This paper addresses the issue of a species occupying a specific ecological niche by introducing a new algorithmic model that overcomes shortcomings of the traditional neutral models.

By Keith Harris, Todd L. Parsons, Umer Z. Ijaz, Leo Lahti, Ian Holmes, and Christopher Quince

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 multisite 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 multisite UNTB. We can also use this approach to distinguish between neutral local community assembly given a nonneutral 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

Manuscript received May 29, 2014; revised October 14, 2014; accepted March 13, 2015. Date of publication August 13, 2015; date of current version February 16, 2017.
The work of K. Harris was supported by a Unilever Research Grant. The work of T. L. Parsons was supported by the CNRS, and was also supported in part by the Fondation Sciences Mathématiques de Paris. The work of U. Z. Ijaz was supported by NERC IRF NE/L011956/1. The work of L. Lahti was supported by the Academy of Finland (Decision 256950). The work of I. Holmes was supported by NIH R01 Grant HG004483. The work C. Quince was supported by an MRC Fellowship MR/M50161X/1 as part of the CLIMB consortium MR/L015080/1.
K. Harris is with the School of Mathematics and Statistics, University of Sheffield, Sheffield S10 2TN, U.K. (e-mail: k.j.harris@sheffield.ac.uk).
T. L. Parsons is with the Laboratoire de Probabilités et Modèles Aléatoires, Paris, France. He is also with the Collège de France, Center for Interdisciplinary Research in Biology, 75005 Paris, France (e-mail: tlparsons@gmail.com).
U. Z. Ijaz is with the Infrastructure and Environment Research Division, School of Engineering, University of Glasgow, Glasgow, G12 8LT, U.K. (e-mail: Umer.Ijaz@glasgow.ac.uk).
L. Lahti is with the Department of Veterinary Biosciences, University of Helsinki, Helsinki 0170, Finland. He is also with the Laboratory of Microbiology, Wageningen University, Wageningen, Netherlands (e-mail: leo.lahti@iki.fi).
I. Holmes is with the Department of Bioengineering, University of California, Berkeley, CA 94720 USA (e-mail: Iihholmes@gmail.com).
C. Quince is with Warwick Medical School, University of Warwick, Coventry, CV4 7AL, U.K. (e-mail: c.quince@warwick.ac.uk).
This paper has supplementary downloadable material available at http://ieeexplore.ieee.org, provided by the authors. This consists of a single 410-KB pdf containing two SI Appendices, entitled 1) "Large Population Limits for a Neutral Metacommunity" and 2) "Gibbs Sampling for the UNTB-HDP." The former contains a detailed derivation of the convergence of the UNTB onto the HDP and the latter both a derivation of the Gibbs sampling conditional posterior distributions and additional numerical results.

Digital Object Identifier: 10.1109/JPROC.2015.2428213

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.

KEYWORDS | Diversity; ecological modelling; hierarchical Dirichlet process; Hubbell's unified neutral theory of biodiversity; microbial communities

I. 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 hypervolume 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 nonneutral 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 multisite 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 nonneutral 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 multisite 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.

II. METHODS

A. 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 $i = 1, \ldots, M$ 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 $(\pi_{i,1}, \ldots, \pi_{i,S})$ 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 $I_i = (m_i/(1 - m_i))(N_i - 1)$ [9]. The parameter I_i controls the coupling of the local community to the metacommunity. As $I_i \to \infty$, the local community stationary distribution will approach the metacommunity distribution and the number of species at that site will increase, while as $I_i \to 0$, 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, $\theta = (\nu/(1 - \nu))(N - 1)$ [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, which we will denote S, 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

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

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

B. HDP Limit to Neutral Metacommunities

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). While 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 (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

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

where $\bar{\beta} = (\beta_1, \ldots, \beta_S)$ is the relative frequency of each species in the metacommunity. In the SI Appendix (see Section 1.4: Corollary 1), we generalize 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]

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

The DP can be viewed as an infinite dimensional generalization 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 $\{\beta'_k\}_{k=1}^{\infty}$

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

Then we can define the kth element of the metacommunity vector as

$$\beta_k = \beta'_k . \prod_{l=1}^{k-1} (1 - \beta'_l). \tag{5}$$

We will denote this process $\bar{\beta} \sim \mathrm{Stick}(\theta)$. 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

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

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

C. Gibbs Sampler for the Neutral-HDP Model

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

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

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

$$\theta|\alpha,\zeta \sim \text{Gamma}(\alpha,\zeta) \tag{7}$$
$$I_i|\eta,\kappa \sim \text{Gamma}(\eta,\kappa) \tag{8}$$

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 $k = 1,\ldots,S$ and the unrepresented component as $\beta_u = \sum_{k=S+1}^{L} \beta_k$, in the limit as $L \to \infty$. In this finite dimensional representation we can determine the species distributions in the local communities

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

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

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

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 $P(\theta,I_1,\ldots,I_M|\mathbf{X})$. 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

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

where $T = \sum_{i=1}^{M}\sum_{j=1}^{S} T_{ij}$. The first part of the above expression derives from the Antoniak (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

$$\bar{\beta} = (\beta_1,\beta_2,\ldots,\beta_S,\beta_u) \sim \text{Dir}(T_{\cdot 1},T_{\cdot 2},\ldots,T_{\cdot S},\theta) \tag{12}$$

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

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

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, $T_{i\cdot} = \sum_{j=1}^{S} T_{ij}$, in J_i samples from the local community DP with concentration parameter, I_i.

4) Sample the ancestral states

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

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 5000 iterations as burn-in. The results below are quoted as the median values over these last 25 000 or 5000 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 kth posterior sample of fitted UNTB parameters, $\theta^k, I_1^k, \ldots, I_M^k$, 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 (12), since the ancestral frequencies $T_{\cdot j} = \sum_{i=1}^{M} T_{ij}$ 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 $P(\mathbf{X}_0^k|\bar{\beta}_0^k, I_1^k, \ldots, I_M^k)$ using (10). These likelihoods were then compared to the actual likelihood of the observed sample, $P(\mathbf{X}|\bar{\beta}^k, I_1^k, \ldots, I_M^k)$, 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 nonneutral metacommunity. We do this in the same way calculating the likelihood for each of the samples, $P(\mathbf{X}_1^k|\bar{\beta}^k, I_1^k, \ldots, I_M^k)$, and comparing to $P(\mathbf{X}|\bar{\beta}^k, I_1^k, \ldots, I_M^k)$, the proportion of samples with likelihood greater than this forms our pseudo p-value for local neutral community assembly, which we denoteby p_L. For both tests, samples were generated either from 2500 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 5000 Gibbs samples for the human gut microbiota when testing multiple taxa.

There are many ways in which a distribution could appear nonneutral. 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 nonneutral since a single metapopulation distribution will be unable to desribe all the community configurations observed. In addition, communities can also appear nonneutral 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 nonneutral at the local community level. If this occurs for the metacommunity then neutrality will be rejected there too.

D. 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. First, 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 nonneutral fashion, then the community as a whole will appear nonneutral 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 renormalized 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 summarized in Table 1.

E. Data

1) Neutral Simulation: In SI Appendix Section 2 we show that the UNTB-HDP fitting method accurately determines

Table 1 Key Ideas Used in This Paper

Neutral model	A populaton model in which all types are *functionally* 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 [39] 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 [40], 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.

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 $i = 1, \ldots, 50$, 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 2000 generations, i.e., 40 000 iterations, at which point the species number appeared stationary, then 1000 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 occuring in the local populations themselves, with a metapopulation which is an implicit aggregate of the local populations rather than an explicit distribution.

2) 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 [41] 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.

3) 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 characterized 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 [42], [43]. 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 1598. The number of unique sequences remaining following noise removal was 19 647. These were then taxonomically classified using the RDP stand-alone classifier of [44]. We constructed operational taxonomic units (OTUs) at 3% sequence difference using average linkage clustering to approximate species [45]. This was done for the entire data set generating 7238 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 nonneutral 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 1000 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 nonneutrality. We also did this for the tropical tree data.

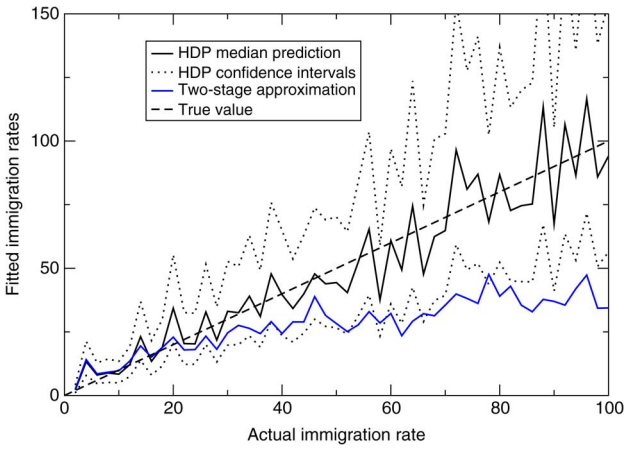


Fig. 1. ***Estimated immigration rates versus true values for the UNTB-HDP model fit to a neutral model simulation. Predictions are medians (solid line) from 25 000 posterior samples together with lower (2.5%) and upper (97.5%) Bayesian confidence intervals (dotted lines). The predictions from the two-stage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

III. RESULTS

A. Neutral Simulation

In Fig. 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

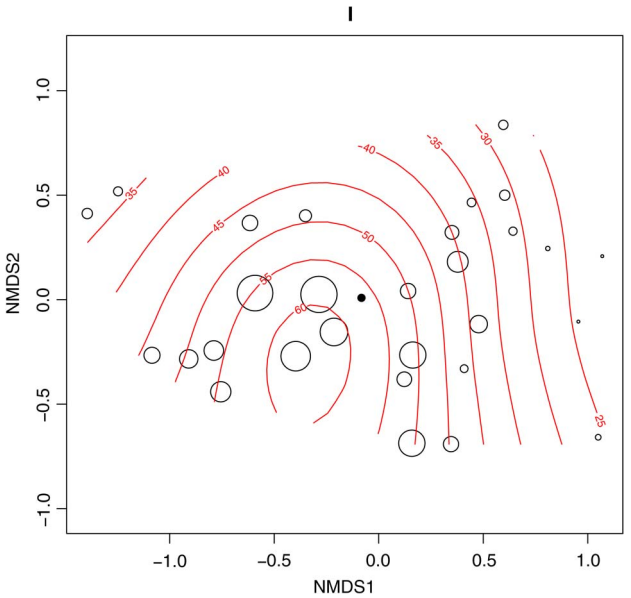


Fig. 2. ***An NMDS plot of the twenty-nine Panama tropical tree communities. Communities are visualised as bubbles with size proportional to the median I_i values obtained from the UNTB-HDP Gibbs sampler. Contours calculated using the ordisurf function of the R vegan package are also shown. The metacommunity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

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.

B. 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 Fig. 2, we use nonmetric multidimensional 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 [46]. 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

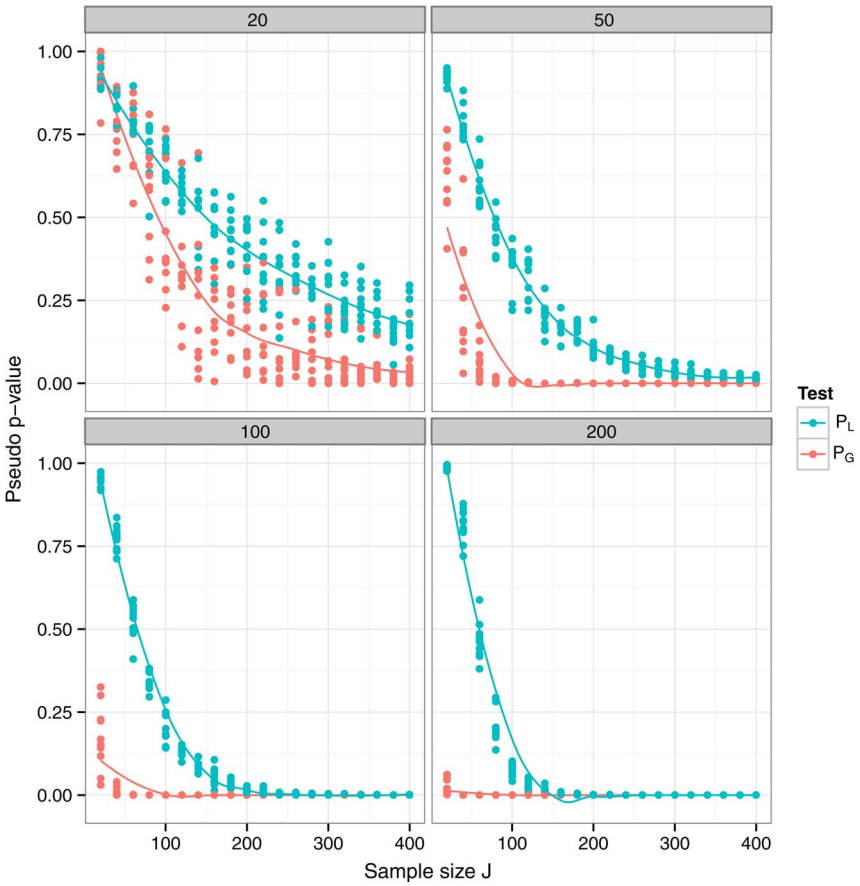


Fig. 3. *Impact of sample number and size on detection of nonneutrality in the human gut data. The figures show the pseudo p-values for neutrality for both the complete neutral model (P_G) and local community assembly (P_L). We generated ten replicate communities by sampling without replacement either 20, 50, 100, or 200 samples from those that had 1000 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.*

distance from this site to each of the others, then we find a significant negative correlation ($p = 0.03$) between distance and immigration rate.

C. 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 Fig. 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 parantheses 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 affects 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 Fig. 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.

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)
Bacteroidaceae	208	224	506	51.4	0.7	3.3	7.6	0 (0)	0.03 (0)
Bacteroides	208	224	506	51.4	0.7	3.3	7.6	0 (0)	0.03 (0)
Firmicutes	277	4770	1009	1382.3	21.4	44.8	81.0	0 (0)	0(0)
Incertae Sedis XIV	87	176	264	39.2	1.7	9.8	27.5	0 (0)	0.05 (0.004)
Blautia	87	175	264	38.9	1.6	10.1	27.1	0 (0)	0.06 (0.003)
Lachnospiraceae	164	873	248	262.9	6.5	13.0	21.2	0 (0)	0 (0)
Ruminococcaceae	239	1471	409	411.0	4.5	16.1	38.1	0 (0)	0 (0)
Faecalibacterium	141	301	297	71.7	1.0	7.5	21.4	0 (0)	0.004 (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; 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.

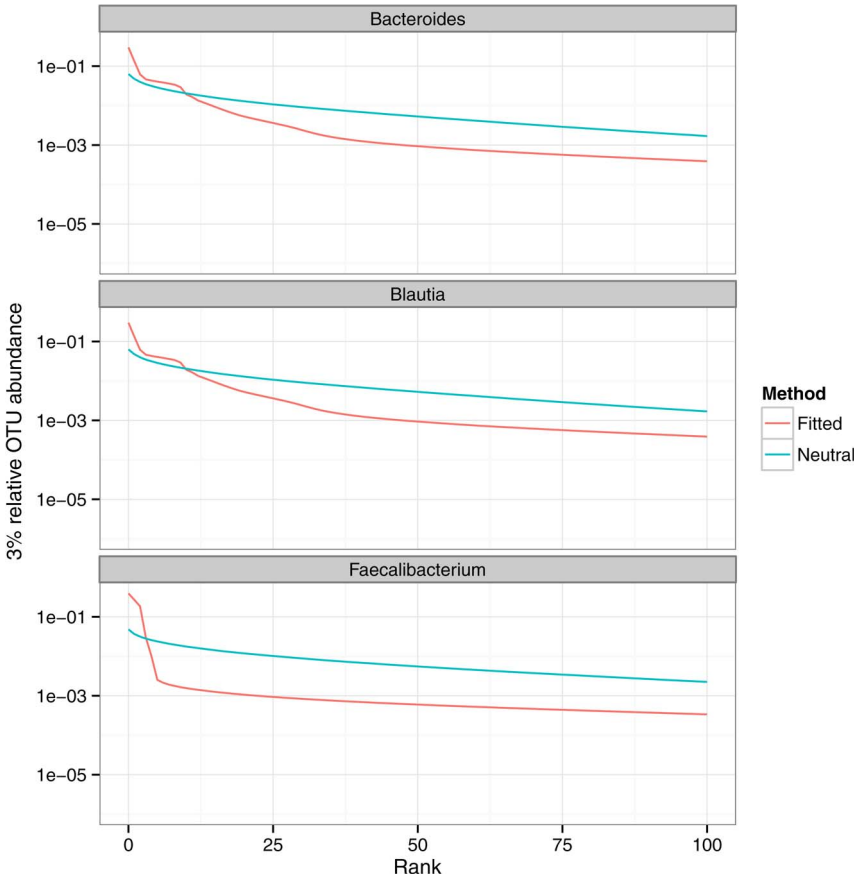


Fig. 4. ***Human gut metacommunity distributions. The fitted metacommunity distributions (red line) and neutral metacommunity predictions (blue line) as rank-abundance curves for three genera: Bacteroides, Blautia, and Faecalibacterium.***

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 Fig. 5). The same negative correlation was also observed for their parent phylum the Firmicutes but it was slightly stronger (p-value $= 0.007$).

IV. DISCUSSION

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 nonneutrality. It is clear from Fig. 3 that as the number of samples in particular decreases it becomes hard to detect nonneutral distributions—this is actually a strong motivation for the use of the UNTB-HDP which can be efficiently fit in the

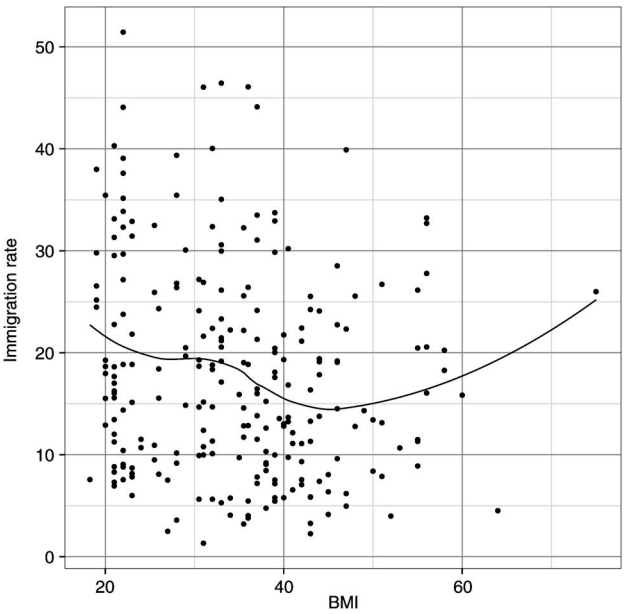


Fig. 5. ***Immigration rate versus BMI. Median immigration rate for the family Ruminococcaceae determined by the UNTB-HDP model plotted against body mass index. A significant negative correlation is observed (p-value = 0.014—Pearson's correlation).***

multisite 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 [47].

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 nonneutral processes. The dominant OTUs may have a competitive advantage, or interactions with bacteriophages [48] or the host immune system may be structuring these distributions [49], 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 [50] 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" [51]. 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. Interesting 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 hierachical 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. ■

Acknowledgment

The authors would like to thank three anonymous reviewers for constructive comments.

REFERENCES

[1] G. E. Hutchinson, "Concluding remarks," in *Proc. Cold Spring Harbor Symp. Quantitative Biol.*, 1957, vol. 22, pp. 415–427.

[2] G. Hardin, "The competitive exclusion principle," *Science*, vol. 131, pp. 1292–1297, 1960.

[3] D. Simberloff and T. Dayan, "The guild concept and the structure of ecological communities," *Ann. Rev. Ecol. Syst.*, vol. 22, pp. 115–143, 1991.

[4] H. Caswell, "Community structure: A neutral model analysis," *Ecol. Monograph.*, vol. 46, pp. 327–354, 1976.

[5] S. P. Hubbell, *The Unified Neutral Theory of Biodiversity and Biogeography*. Princeton, NJ, USA: Princeton Univ. Press, 2001.

[6] R. H. MacArthur and E. O. Wilson, *The Theory of Island Biogeography*. Princeton, NJ, USA: Princeton Univ. Press, 1967.

[7] B. J. McGill, "A test of the unified neutral theory of biodiversity," *Nature*, vol. 422, pp. 881–885, 2003.

[8] R. S. Etienne and H. Olff, "A novel genealogical approach to neutral biodiversity theory," *Ecol. Lett.*, vol. 7, pp. 170–175, 2004.

[9] R. S. Etienne, "A new sampling formula for neutral biodiversity," *Ecol. Lett.*, vol. 8, pp. 253–260, 2005.

[10] J. Rosindell, S. P. Hubbell, F. He, L. J. Harmon, and R. S. Etienne, "The case for ecological neutral theory," *Trends Ecol. Evol.*, vol. 27, pp. 203–208, 2012.

[11] R. A. Chisholm and S. W. Pacala, "Niche and neutral models predict asymptotically equivalent species abundance distributions in high-diversity communities," *Proc. Nat. Acad. Sci. USA*, vol. 107, pp. 15 821–15 825, 2010.

[12] R. S. Etienne, "A neutral sampling formula for multiple samples and an 'exact' test of neutrality," *Ecol. Lett.*, vol. 10, pp. 608–618, 2007.

[13] R. S. Etienne, "Maximum likelihood estimation of neutral model parameters for multiple samples with different degrees of dispersal limitation," *J. Theor. Biol.*, vol. 257, pp. 510–514, 2009.

[14] F. Munoz, P. Couteron, B. R. Ramesh, and R. S. Etienne, "Estimating parameters of neutral communities: From one single large to several small samples," *Ecology*, vol. 88, pp. 2482–2488, 2007.

[15] F. Jabot, R. S. Etienne, and J. Chave, "Reconciling neutral community models and environmental filtering: Theory and an empirical test," *Oikos*, vol. 117, pp. 1308–1320, 2008.

[16] R. S. Etienne, "Improved estimation of neutral model parameters for multiple samples with different degrees of dispersal limitation," *Ecology*, vol. 90, no. 3, pp. 847–852, 2009.

[17] M. Hamady, J. J. Walker, J. K. Harris, N. J. Gold, and R. Knight, "Error-correcting barcoded primers for pyrosequencing hundreds of samples in multiplex," *Nat. Methods*, vol. 5, no. 3, pp. 235–237, Mar. 2008.

[18] P. J. Turnbaugh *et al.*, "A core gut microbiome in obese and lean twins," *Nature*, vol. 457, no. 7228, pp. 480–484, Jan. 22, 2009.

[19] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei, "Hierarchical Dirichlet processes," *J. Amer. Statist. Assoc.*, vol. 101, no. 476, pp. 1566–1581, Dec. 2006.

[20] R. Condit *et al.*, "Beta-diversity in tropical forest trees," *Science*, vol. 295, no. 5555, pp. 666–669, Jan. 25, 2002.

[21] M. Arumugam *et al.*, "Enterotypes of the human gut microbiome," *Nature*, vol. 473, no. 7346, pp. 174–180, May 12, 2011.

[22] I. Holmes, K. Harris, and C. Quince, "Dirichlet multinomial mixtures: Generative models for microbial metagenomics," *PLoS ONE*, vol. 7, no. 2, Feb. 2012, Art. ID. e30126. [Online]. Available: http://dx.doi.org/10.1371%2Fjournal.pone.0030126

[23] T. Ding and P. D. Schloss, "Dynamics and associations of microbial community types across the human body," *Nature*, vol. 509, pp. 357–360, 2014.

[24] N. Fierer, M. A. Bradford, and R. B. Jackson, "Toward an ecological classification of soil bacteria," *Ecology*, vol. 88, no. 6, pp. 1354–1364, Jun. 2007.

[25] L. Philippot *et al.*, "Spatial patterns of bacterial taxa in nature reflect ecological traits of deep branches of the 16S rRNA bacterial tree," *Environ. Microbiol.*, vol. 11, no. 12, pp. 3096–3104, Dec. 2009.

[26] L. Philippot *et al.*, "The ecological coherence of high bacterial taxonomic ranks," *Nature Rev. Microbiol.*, vol. 8, no. 7, pp. 523–529, Jul. 2010.

[27] D. A. Rasko *et al.*, "The pangenome structure of Escherichia coli: Comparative genomic analysis of E-coli commensal and pathogenic isolates," *J. Bacteriol.*, vol. 190, no. 20, pp. 6881–6893, Oct. 2008.

[28] W. Sloan, M. Lunn, S. Woodcock, I. Head, S. Nee, and T. Curtis, "Quantifying the roles of immigration and chance in shaping prokaryote community structure," *Environ. Microbiol.*, vol. 8, no. 4, pp. 732–740, Apr. 2006.

[29] S. Woodcock *et al.*, "Neutral assembly of bacterial communities," *FEMS Microbiol. Ecol.*, vol. 62, no. 2, pp. 171–180, Nov. 2007, Joint Symposium of the Environmental-Microbiology-Group/British-Ecological-Society/Society-for- General-Microbiology, Univ. York, York, England, Sep. 13, 2006.

[30] P. Jeraldo *et al.*, "Quantification of the relative roles of niche and neutral processes in structuring gastrointestinal microbiomes," *Proc. Nat. Acad. Sci. USA*, vol. 109, no. 25, pp. 9692–9698, Jun. 19, 2012.

[31] A. McKane, D. Alonso, and R. Sole, "Analytic solution of Hubbell's model of local community dynamics," *Theor. Pop. Biol.*, vol. 65, no. 1, pp. 67–73, Feb. 2004.

[32] W. T. Sloan, S. Woodcock, M. Lunn, I. M. Head, and T. P. Curtis, "Modeling taxa-abundance distributions in microbial communities using environmental sequence data," *Microb. Ecol.*, vol. 53, no. 3, pp. 443–455, Apr. 2007, Workshop on Microbial Environmental Genomics, Shanghai Jiao Tong Univ., Manhang Campus, Shanghai, China, Jun. 12–15, 2005.

[33] T. S. Ferguson, "A Bayesian analysis of some nonparametric problems," *Ann. Stat.*, vol. 1, no. 2, pp. 209–230, 1973.

[34] W. J. Ewens, "The sampling theory of selectively neutral mutations," *Theor. Pop. Biol.*, vol. 3, pp. 87–112, 1972.

[35] C. E. Antoniak, "Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems," *Ann. Statist.*, vol. 2, pp. 1152–1174, 1974.

[36] D. J. Mackay, "Bayesian interpolation," *Neural Comput.*, vol. 4, pp. 415–417, 1992.

[37] M. D. Escobar and M. West, "Bayesian density estimation and inference using mixtures," *J. Amer. Statist. Assoc.*, vol. 90, no. 430, pp. 577–588, Jun. 1995.

[38] S. C. Walker, "When and why do non-neutral metacommunities appear neutral?" *Theor. Pop. Biol.*, vol. 71, pp. 318–331, 2007.

[39] D. Aldous, "Exchangeability and related topics *École d'Été de Probabilités de Saint-Flour XIII-1983*. Berlin: Springer-Verlag, 1985, pp. 1–198.

[40] F. M. Hoppe, "Pólya-like urns and the Ewens' sampling formula," *J. Math. Biol.*, vol. 20, no. 1, pp. 91–94, 1984.

[41] C. Pyke, R. Condit, S. Aguilar, and S. Lao, "Floristic composition across a climatic gradient in a neotropical lowland forest," *J. Veg. Sci.*, vol. 12, no. 4, pp. 553–566, Aug. 2001.

[42] C. Quince *et al.*, "Accurate determination of microbial diversity from 454 pyrosequencing data," *Nat. Methods*, vol. 6, pp. 639–641, 2009.

[43] C. Quince, A. Lanzen, R. J. Davenport, and P. J. Turnbaugh, "Removing noise from pyrosequenced amplicons," *BMC Bioinf.*, vol. 12, no. 38, 2011.

[44] Q. Wang, G. M. Garrity, J. M. Tiedje, and J. R. Cole, "Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy," *Appl. Environ. Microb.*, vol. 73, no. 16, pp. 5261–5267, Aug. 2007.

[45] N. Youssef *et al.*, "Comparison of species richness estimates obtained using nearly complete fragments and simulated pyrosequencing-generated fragments in 16S rRNA gene-based environmental surveys," *Appl. Environ. Microbiol.*, vol. 75, no. 16, pp. 5227–5236, Aug. 15, 2009.

[46] R Development Core Team, *R: A Language and Environment for Statistical Computing*, Vienna, Austria, 3-900051-07-0, 2010. [Online]. Available: http://www.R-project.org

[47] J. Rosindell, Y. Wong, and R. Etienne, "A coalescence approach to spatial neutral ecology," *Ecol. Informat.*, pp. 259–271, 2008.

[48] S. Minot *et al.*, "The human gut virome: Inter-individual variation and dynamic response to diet," *Genome Res.*, vol. 21, no. 10, pp. 1616–1625, Oct. 2011.

[49] C. Quince *et al.*, "The impact of Crohn's disease genes on healthy human gut microbiota: A pilot study," *Gut*, vol. 62, pp. 952–954, Jan. 2013.

[50] X. Ze, S. H. Duncan, P. Louis, and H. J. Flint, "Ruminococcus bromii is a keystone species for the degradation of resistant starch in the human colon," *ISME J.*, vol. 6, pp. 1535–1543, 2012.

[51] B. J. Finlay and T. Fenchel, "Cosmopolitan metapopulations of free-living microbial eukaryotes," *Protist*, vol. 155, pp. 237–244, 2004.

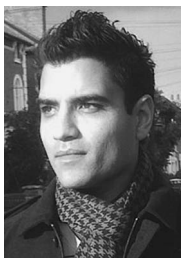
ABOUT THE AUTHORS

Keith Harris received the B.Sc. degree in mathematics and statistics from the University of York, York, U.K., in 2002, and the M.Sc. and Ph.D. degrees in statistics from the University of Sheffield, Sheffield, U.K., in 2003 and 2008, respectively.

From 2008 to 2013, he was a PDRA at the University of Glasgow, first working on a project entitled ''Classifiers in Medicine and Biology (Advancing Machine Learning Methodology for New Classes of Prediction Problems)'' in the Department of Computing Science, and later on a Unilever funded project in the School of Engineering that focused on developing new statistical methods for the analysis of genomics data from microbial communities. He is currently a PDRA in the School of Mathematics and Statistics at the University of Sheffield working on an EPSRC funded project called ''Simulation Tools for Automated and Robust Manufacturing.''

Todd L. Parsons received the B.Sc. degree in pure mathematics from the University of Waterloo, Waterloo, ON, Canada, and the M.Sc. and Ph.D. degee in mathematics from the University of Toronto, Toronto.

From 2007 to 2011, he was a Postdoctoral Researcher in biology at the University of Pennsylvania. He is now a permanent Research Scientist (CR2) with the Centre National de la Recherche Scientifique (CNRS) with a primary appointment in the Probability and Stochastic Models Laboratory at l'Université Pierre et Marie Curie (Paris 06), and a courtesy appointment in the Centre for Interdisciplinary Research in Biology at the Collège de France. His research interests include stochastic population models and combinatorial stochastic processes arising from ecology, epidemiology, and population genetics.



Umer Z. Ijaz received the B.S. and M.S. degrees in computer systems engineering from Ghulam Ishaq Khan Institute, Pakistan, in 2000 and 2003, respectively, and the Ph.D. degree in electrical and electonics engineering from Jeju National University, South Korea, in 2008.

Between 2008 to 2014, he worked as Postdoctoral Researcher at the Universities of Cambridge, Oxford, and Glasgow. He is currently a NERC Independent Research Fellow and Lord Kelvin Smith Fellow leading his own group on Environmental Omics at School of Engineering, Glasgow. The purpose of his current research is to integrate different sources of 'omics data (metagenomics, metatranscriptomics, metabolomics, and metaproteomics) in environmental science for microbial community analysis, by focusing on software development and numerical ecology.

Leo Lahti received the doctoral degree in machine learning and bioinformatics from the Department of Computer Science, Aalto University, Espoo, Finland, in 2010.

He is currently affiliated with the Laboratory of Microbiology, Wageningen University, The Netherlands, and the Department of Veterinary Biosciences, University of Helsinki, Finland, as a Postdoctoral Research Fellow of the Academy of Finland. His recent research focuses on understanding the population dynamics and health associations of the human microbiome.

Ian Holmes received the Ph.D. degree from the Sanger Centre in the University of Cambridge (now the Sanger Institute), Cambridge, U.K., in 1998.

Following a spell at Los Alamos National Laboratory as a Fulbright-Zeneca Research Fellow, he worked on the Berkeley Drosophila Genome Project and the EBI's Ensembl project before being appointed as a Lecturer in Bioinformatics at the Department of Statistics, University of Oxford (2002–2004). In 2004, he was hired by the Department of Bioengineering, University of California, Berkeley, CA, USA, and was promoted to tenure in 2010. His research involves building realistic stochastic models of various aspects of genome evolution, and making these models useful as practical tools for biological discovery.

Christopher Quince received the Ph.D. degree in food web modeling, in 2002, from the Theoretical Physics Group at the University of Manchester, Manchester, U.K.

He has pioneered the development of bioinformatics and statistics for the interpretation of next generation sequence data from microbial communities. He then worked on theoretical population genetics and fish growth models during postdoctoral positions at Arizona State University and the University of Toronto before obtaining a LKAS fellowship at the University of Glasgow in 2006 to study microbial communities. Since then he has held successive fellowships from the EPSRC and MRC enabling him to devote his time to developing algorithms and software for microbial community analysis. These include the widely used software, AmpliconNoise and uchime, for the analysis of 16S rRNA sequence data and CONCOCT for extracting genomes from shotgun metagenomics data.