Received: 25 March 2019 | Revised: 26 June 2019 | Accepted: 28 June 2019

DOI: 10.1111/1755-0998.13063

RESOURCE ARTICLE

MOLECULAR ECOLOGY RESOURCES WILEY

Characterizing symbiont inheritance during host–microbiota evolution: Application to the great apes gut microbiota

Benoît Perez-Lamarque[1,2] | Hélène Morlon[1]

[1]Institut de Biologie de l'ENS (IBENS), Département de biologie, École normale supérieure, CNRS, INSERM, PSL University, Paris, France

[2]Muséum national d'Histoire naturelle, UMR 7205 CNRS-MNHN-UPMC-EPHE "Institut de Systématique, Evolution, Biodiversité – ISYEB", Herbier National, 16 rue Buffon, Paris, France

Correspondence
Benoît Perez-Lamarque and Hélène Morlon, Institut de Biologie de l'ENS (IBENS), Département de biologie, École normale supérieure, CNRS, INSERM, PSL University, 75005 Paris, France.
Emails: benoit.perez@ens.fr; morlon@biologie.ens.fr

Funding information
Ecole Doctorale FIRE, Grant/Award Number: Programme Bettencourt; Ecole normale supérieure; Centre National de la Recherche Scientifique; European Research Council, Grant/Award Number: PANDA ; European Research Council; École Normale Supérieure

Abstract

Microbiota play a central role in the functioning of multicellular life, yet understanding their inheritance during host evolutionary history remains an important challenge. Symbiotic microorganisms are either acquired from the environment during the life of the host (i.e. environmental acquisition), transmitted across generations with a faithful association with their hosts (i.e. strict vertical transmission), or transmitted with occasional host switches (i.e. vertical transmission with horizontal switches). These different modes of inheritance affect microbes' diversification, which at the two extremes can be independent from that of their associated host or follow host diversification. The few existing quantitative tools for investigating the inheritance of symbiotic organisms rely on cophylogenetic approaches, which require knowledge of both host and symbiont phylogenies, and are therefore often not well adapted to DNA metabarcoding microbial data. Here, we develop a model-based framework for identifying vertically transmitted microbial taxa. We consider a model for the evolution of microbial sequences on a fixed host phylogeny that includes vertical transmission and horizontal host switches. This model allows estimating the number of host switches and testing for strict vertical transmission and independent evolution. We test our approach using simulations. Finally, we illustrate our framework on gut microbiota high-throughput sequencing data of the family Hominidae and identify several microbial taxonomic units, including fibrolytic bacteria involved in carbohydrate digestion, that tend to be vertically transmitted.

KEYWORDS

great apes, holobiont, likelihood-based framework, microbiota, molecular evolution, symbiont transmission

1 | INTRODUCTION

Microbiota – host-associated microbial communities – play a major role in the functioning of multicellular organisms (Hacquard et al., 2015). For example, the gut microbiota plays a significant nutritional role for animals by synthesizing essential nutrients and by helping digestion and detoxification (McFall-Ngai et al., 2013). It is also involved in a broad range of other mutualistic functions important for host protection, development, behaviour and reproduction (Zilber-Rosenberg & Rosenberg, 2008). Other less-studied microbiota, such as those found on animal skins or plant roots, also play major ecological roles (Philippot, Raaijmakers, Lemanceau, & van der Putten, 2013).

Host–microbiota associations have evolved for thousand million years with three major modes of inheritance across phylogenetic host lineages: (a) strict vertical transmission within a host lineage (Rosenberg & Zilber-Rosenberg, 2016), which can happen either by transmission from mother to child (e.g. directly through ovaries during reproduction or at birth), or by social contact while sharing life with related individuals (Bright & Bulgheresi, 2010), (b) vertical

transmission with occasional horizontal switches between host lineages (Henry et al., 2013), which can for example happen through direct interactions, via vectors or via shared habitats (Engel & Moran, 2013), and (c) environmental acquisition, with microbes coming from the environment independently from other related hosts (Bright & Bulgheresi, 2010). The vertical transmission of a given microbial lineage within host lineages can lead to cophylogenetic patterns, with the microbial phylogeny mirroring the host phylogeny (e.g. *Helicobacter pylori* in humans; Linz et al., 2007). Horizontal switches and environmental acquisitions can play key roles in adaptation, for example, by allowing host lineages to adapt to new feeding regimes (McKenney, Maslanka, Rodrigo, & Yoder, 2018; Muegge et al., 2011). They will tend to erase cophylogenetic patterns linked to vertical transmission. The relative importance of each of the three modes of inheritance depends on the type of host and the type of microbes. For example, vertical transmission is thought to be far more preponderant in the 'core' microbial species, which are shared across hosts regardless of environmental conditions, than in the 'flexible' microbial species, facultative and dependent on internal and external conditions (Shapira, 2016).

Quantifying the relative importance of different modes of inheritance during host–microbiota co-evolution remains a major challenge. Patterns of 'phylosymbiosis', that is a pattern of concordance between a given host phylogeny and the dendrogram reflecting the similarity of microbial communities across these hosts, are frequently observed (Bordenstein & Theis, 2015), for example, for great apes gut microbiota (Ochman et al., 2010). Although these phylosymbiotic patterns suggest that some microbial species within the microbiota are vertically transmitted, such community-wide comparisons of microbiota across hosts do not allow identifying which microbial species are vertically transmitted, nor quantifying the relative importance of the different modes of inheritance across distinct microbial species. More recently, approaches have been developed to apply cophylogenetic concepts to microbial taxa (Bailly-Bechet et al., 2017; Groussin et al., 2017). Cophylogenetic methods were originally developed to study the co-evolution between hosts and their symbionts, with the underlying idea that close and long-term associations lead to congruent phylogenies with similar topologies and divergence times (Page & Charleston, 1998; de Vienne et al., 2013), while processes such as host switches disrupt this congruence. Cophylogenetic tools either quantify the congruence between symbiont and host trees using distance-based methods – for example ParaFit (Legendre, Desdevises, & Bazin, 2002), generalizations of the Mantel test (Hommola, Smith, Qiu, & Gilks, 2009), or PACo (Balbuena, Míguez-Lozano, & Blasco-Costa, 2013) – or try to find the most parsimonious sets of events (e.g. host switches) that allow reconciling both trees (e.g. TreeMap or Jane; Conow, Fielder, Ovadia, & Libeskind-Hadas, 2010). In the context of microbiota, Groussin et al. (2017) and Bailly-Bechet et al. (2017) have used the ALE program (Szöllősi, Rosikiewicz, Boussau, Tannier, & Daubin, 2013; Szöllosi, Tannier, Lartillot, & Daubin, 2013), which was initially designed to solve the gene tree–species tree reconciliation problem. Importantly, all these methods require first a reconstruction of the microbial tree for each individual microbial taxon. However, microbiota data are typically generated using next-generation sequencing (NGS) metabarcoding techniques, providing short nucleotidic reads of a targeted slow-evolving universal gene (e.g. the 16S rRNA gene). Such data often contain limited variability within each microbial taxon, which can be problematic for reconstructing their tree.

Here, we develop a probabilistic model of host–symbiont evolution, which aims at studying modes of inheritance in the microbiota without building first microbial phylogenies. The main idea is to use the host phylogenetic tree to inform the microbial trees, which reduces the problem of low phylogenetic resolution of metabarcoding microbial markers. Huelsenbeck, Rannala, and Larget (2000) developed a model of cospeciation and host switches similar to ours, focused on host–parasite associations. However, the authors developed an inference framework reconstructing host and parasite phylogenetic trees jointly, which is not well adapted to the case when the host phylogenetic tree is robust and the symbionts are represented by a sequence alignment with limited phylogenetic information. Here, we fix the host phylogeny and follow the evolution of individual microbial taxa on the host tree. We compute likelihoods associated with microbial sequence alignments under a model including vertical inheritance and host switches. We find estimates of the number of host switches and develop tests for evaluating model support in comparison with scenarios of independent evolution and strict vertical transmission. We test our approach using simulations and apply it to gut microbiota high-throughput sequencing data of the family Hominidae.

2 | MATERIALS AND METHODS

2.1 | HOME: A general framework for studying host–microbiota evolution

2.1.1 | From metabarcoding microbiota data to separate alignments

Given a host species tree and metabarcoding microbiota data sampled from each host species (e.g. sequences from the 16S rRNA gene, ITS or any other DNA metabarcoding marker), our framework begins by clustering sequences into operational taxonomic units (OTUs) using bioinformatics pipelines. Each OTU is made of distinct microbial populations, each corresponding to a specific host species (Figure 1a). We assume as a starting point that there is no within-host genetic variability (we discuss later how we relaxed this assumption), such that each microbial population is represented by a unique sequence. In our analysis of these data, for each OTU and each host, we use the most abundant microbial sequence as the representative sequence. The data we consider thus consist of a series of microbial alignments A, each corresponding to a sequence alignment for a specific OTU; a given alignment is composed of N-nucleotidic sites long sequences (with potential gaps considered as missing data), each corresponding to a specific host. In each alignment, we

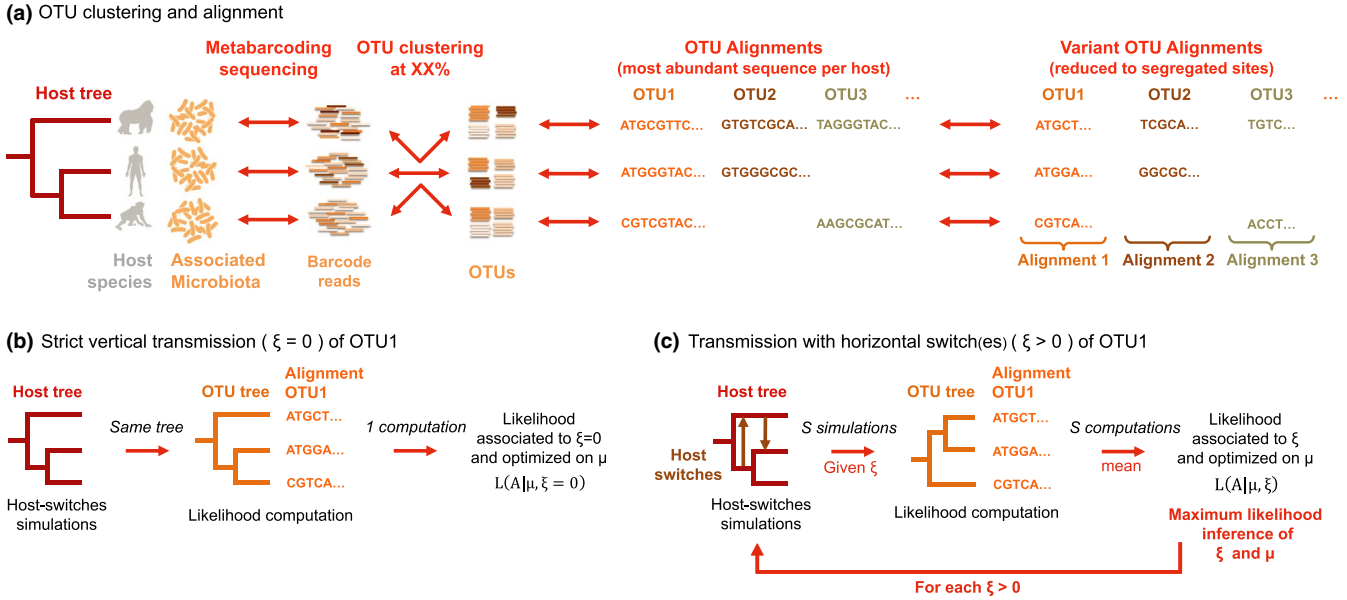


FIGURE 1 Illustration of the various steps for assessing microbial modes of inheritance in host–microbiota evolution from metabarcoding data. (a) The first step consists in clustering the microbial sequences into OTUs and building for each OTU the corresponding alignment of segregating sites (A_S). (b, c) The second step consists in fitting different models of inheritance to each microbial alignment. We compute the probability of the microbial alignment on hypothetical microbial trees. Under a model with strict vertical transmission ($\xi = 0$, b), the microbial is the same as the host tree; under a model with vertical transmission and host switches ($\xi > 0$, c), microbial trees are simulated from the host tree with various numbers of switches ξ. We find the substitution rate $\hat{\mu}$ and the number of switches $\hat{\xi}$ that maximize the probability of the alignment.

distinguish the segregating sites (i.e. those that vary in at least one sequence) to those that do not vary across sequences. Some microbial OTUs may not be represented in all host species (i.e. there might be missing sequences in the alignment), which can either be true absences (i.e. the corresponding host species do not host the OTU), or a lack of detection (i.e. the OTU is present but has not been sampled in these host species). Because we cannot distinguish these two possibilities, we simply treat missing sequences as missing data; we do not explicitly model the extinction of symbiotic populations in certain host species, nor the microbial sampling process. We apply our model separately to each alignment.

2.1.2 | Modelling the evolution of an OTU on a host phylogeny

We consider the evolution of a given microbial OTU on a host phylogeny T (Figure 1); T is assumed to be a known, ultrametric, rooted and binary n-tips tree. The model is defined as follows:

(1) Vertical transmission: From an ancestral microbial population at the root of the host phylogeny represented by a N-nucleotidic sites long sequence with N_v 'variable' sites (i.e. those that can experience substitutions), substitutions occur along host branches. Following classical models of molecular evolution (Strimmer & von Haeseler, 2009), we assume that each variable site evolves independently from the others according to a substitution model with a rate μ that is supposed to be the same for all variable sites and constant along the evolutionary branches (strict-clock model). The substitution model is represented by a continuous-time reversible Markov process, characterized by an invariant measure π (i.e. the vector of base frequencies at equilibrium) and an instantaneous transition rate matrix Q between different states (Strimmer & von Haeseler, 2009).

At a host speciation event, the two daughter host lineages inherit the microbial sequence from the ancestral host, after which microbial populations on distinct host lineages evolve independently.

(2) Host switches: A discrete number (ξ) of host switches happen during the evolution of the OTU on the host tree. The switches occur from a 'donor' branch, with a probability proportional to its branch length, and at a time uniformly distributed on the branch, to a 'receiving' branch, with equiprobability among the co-existing branches (we do not consider the phylogenetic proximity from the donor branch). When a host-switch happens, for convenience we assume that the microbial sequence from the donor host replaces that of the receiving host and the microbial sequence from the donor host remains unchanged.

Each series of host switches on T defines a tree of microbial populations T_B that summarizes which populations descended from which ones and when their divergences occurred (Figure 1). In the absence of host switches ($\xi = 0$), T_B and T are identical. When host switches occur, they break the congruence between T_B and T (e.g. Figure 1c). Hence, the model can be decomposed in two steps: first, host switches generate T_B from T; second, a sequence (representing a microbial population) evolves on T_B with a constant substitution rate.

2.1.3 | Likelihood computation and inference

We develop a likelihood-based framework in order to fit the above model to data comprising a given (fixed) tree T of hosts and an alignment A_S of microbial sequences characterizing populations of a given microbial OTU for these hosts (here the alignment A_S is reduced to the segregating sites). This will allow estimating the number of switches $\hat{\xi}$ on the host tree. The probability of the alignment assuming that the substitution rate is μ and that there are ξ switches is given by:

$$L\left(A_S|\mu,\xi\right)=\int_{T_B} L\left(A_S|\mu,T_B\right)dT_B \quad (1)$$

where $L\left(A_S|\mu,T_B\right)$ is the probability of the alignment assuming that the substitution rate is μ and the (dated) microbial tree is T_B, and the integral is taken over the space of dated trees obtained with ξ switches on T. In practice, we compute this integral using Monte Carlo simulations: we simulate a large number (S) of dated microbial trees obtained with ξ switches on T (see next section), compute for each T_B the probability of the alignment assuming that the substitution rate is μ, and sum these probabilities:

$$L\left(A_S|\mu,\xi\right)\sim\frac{1}{S}\sum_{T_B} L(A_S|\mu,T_B) \quad (2)$$

This approximate expression converges to the exact integral form when S is large.

We compute the probability $L\left(A_S|\mu,T_B\right)$ of the sequence alignment A_S on a given dated microbial tree T_B using the Felsenstein pruning algorithm (Felsenstein, 1981). We take into account the possibility of gaps in the microbial alignment, considering them as 'missing values' by pruning off the tips of the tree with a gap (Truszkowski & Goldman, 2016). First, we choose the model of DNA substitution between the K80, F81 and HKY matrices from the alignment reduced to segregating site (A_S) using the function modelTest (R package PHANGORN) and based on a BIC selection criterion: this function estimates Q and π directly from A_S, where Q, the reversible transition rate matrix, depends on the invariant measure π. We also obtain estimates of the transition/transversion rate ratio κ (K80 and HKY) and of the base frequencies at equilibrium π (F81 and HKY) from these models. Second, we compute the probability of the alignment at each nucleotidic site ν using the pruning algorithm. For a given segregating site among A_S, let $P(t)$ be the vector of probabilities of states A, C, G and T at time t. $P(t)$ is given by $P(t)=M(t)*P(0)$ where $P(0)=\left(1_A,1_C,1_G,1_T\right)$ with 1_A equals 1 if A is the initial nucleotide is A and 0 otherwise, and $M(t)=e^{t\mu Q}$. Let $P_\nu(s)$ be the probability of the alignment corresponding to the clade descending from node s in the phylogeny for site ν. We have:

$$P_\nu(\text{leaf})=\left(1_A,1_C,1_G,1_T\right) \text{ and } P_\nu(s)=\left(M(t_1)P_\nu(s_1)\right)\cdot\left(M(t_2)P_\nu(s_2)\right) \quad (3)$$

Where s_1 and s_2 are the two nodes descending from s and t_1 and t_2 are their respective times of divergence (t_1 and t_2 are fixed, given by the branch lengths of the simulated dated tree T_B). We iterate this pruning calculation from the leaves to the root of the tree, and obtain the probability of the alignment at site ν:

$$L_\nu=\pi P_\nu(\text{root}) \quad (4)$$

Because we consider only segregating sites, we condition this probability on the occurrence of at least one substitution. The probability of a substitution happening on a tree T_B of total branch length B is given by $\left(1-e^{-\mu B}\right)$. Finally, the probability of the alignment A_S is obtained by multiplying the probabilities corresponding to each site. Hence, the probability of the variable alignment A_S is given by:

$$L\left(A_S|\mu,T_B\right)=\left(1-e^{-\mu B}\right)^{-N_s}\prod_{\nu=1}^{N_s}L_\nu \quad (5)$$

where N_s is the number of segregating sites.

In practice, we used $S = 10^4$ and plotted the resulting value of $L\left(A_S|\mu,\xi\right)$ with an increasing number of trees T_B to ensure that S was large enough to obtain a reliable approximation of the likelihood. For each ξ, we find μ that maximizes $L\left(A_S|\mu,\xi\right)$. Finally, we repeat these analyses for a range of realistic ξ values (typically ξ = [0, 1, 2,..., 2n]) and deduce the couple of parameters $\hat{\xi}$ and $\hat{\mu}$ that maximizes the probability of the alignment. Likelihood landscapes typically have a well-defined peak (Figure S1), suggesting that ξ and μ are identifiable. We also show later that we can properly estimate them under a wide set of scenarios. Low $\hat{\xi}$ values are indicative of OTUs that are transmitted mostly vertically, while high $\hat{\xi}$ values are indicative of those that perform frequent host switches.

2.1.4 | *Simulations of host switches: from* T *to* T_B

By simulating ξ switches on T, we obtain a (dated) bacterial T_B characterized by its topology and its branch lengths. Each switch is characterized by its 'donor' branch, by its position on the branch, and by its 'receiving' branch. The donor branch is chosen with a probability proportional to its branch length, the time of the switch is drawn uniformly on the branch, and the receiving branch is chosen with equiprobability among the lineages alive at time t. A switch replaces the existing microbial sequence in the receiving host and creates a new branching event in the microbial tree T_B. Four types of switches can occur, and each of them results in different rules to obtain T_B from T (Figure 2):

1. The switch occurs just after the root on the host tree, before any other speciation event: T_B is obtained from T by re-dating the root of the tree to the time of the host-switch. This switch does not change the topology of the tree (i.e. it only affects the branch lengths).
2. The switch occurs from an internal branch to a branch directly related to the root, that is one of the sequences originating at root no longer has descendants in the current sequences: T_B is obtained from T by re-rooting the tree to the most recent common

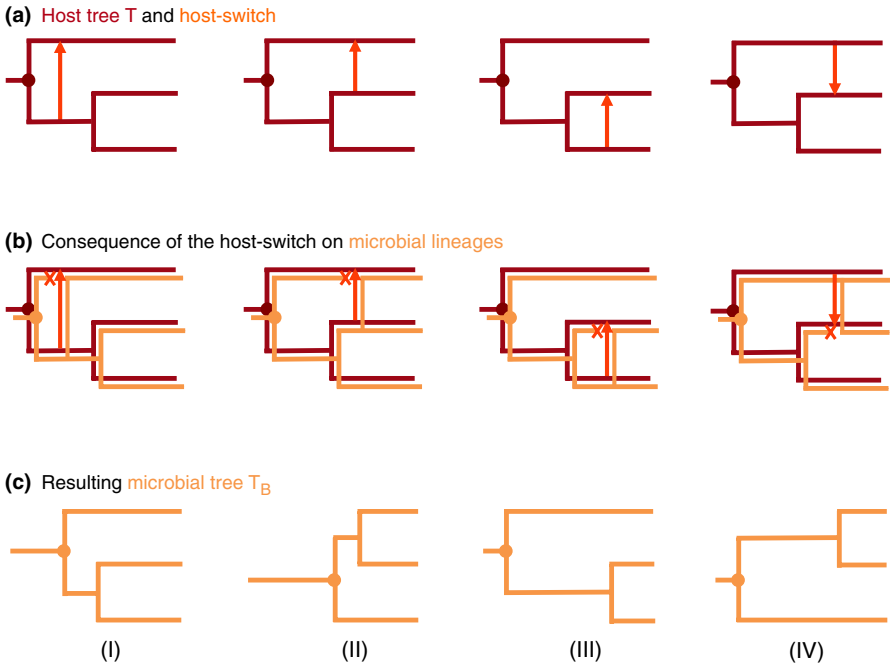


FIGURE 2 Host-switch simulations. (a) Four types of host-switch can occur on the host tree T, and (b, c) these host switches generate distinct microbial trees T_B. Orange arrows represent host switches. Orange crosses represent the extinction of the microbial lineage on the receiving branch

ancestor to all the current microbial sequences. This switch changes both the topology of the tree and the branch lengths.

3. The switch occurs between two sister lineages: T_B is obtained from T by re-dating the divergence between the two sister lineages to the time of the host-switch. This switch only affects the branch lengths of the tree.
4. The switch occurs between two distantly related lineages, and the receiving branch is not related to the root: T_B is obtained from T by an internal reorganization of the tree. This switch changes both the topology of the tree and the branch lengths.

Technically, in order to reduce computation time, we simulated a 'bank of trees' with ξ switches on the host tree and use these same trees in our different analyses. [Colour figure can be viewed at wileyonlinelibrary.com]

2.1.5 | Model selection

In addition to the general model fitting procedure described above, we designed two model selection procedures: the first aims at testing whether the presence of horizontal switches is statistically supported (versus a simpler model with only strict vertical transmission); the second aims at testing support for a model with a limited number of host switches versus environmental acquisition (OTUs that are environmentally acquired will provide high $\hat{\mu}$ and $\hat{\xi}$ estimates and could thus be interpreted as vertical transmission with frequent horizontal switches and high substitution rates instead of environmental acquisition).

In order to test support for a scenario with horizontal host switches versus strict vertical transmission, we compute $L_0 = L(A_S|\hat{\mu},T)$, the likelihood corresponding to the best scenario of evolution of the microbial sequences directly on the host tree (i.e. no switch), and compare it to the likelihood $L_1 = L\left(A_S|\hat{\mu},\hat{\xi}\right)$ corresponding to the best scenario with horizontal switches, using a likelihood ratio test. In order to test support for a scenario of vertical transmission with horizontal host switches versus environmental acquisition, we test its support when compared to a scenario where microbial populations are acquired at random by host species (thereafter referred to as a scenario of 'independent evolution'): we randomize R times the host–microbe association and run our model on each of these randomized data. Next, we analyse the rank of $\hat{\xi}$ and $\hat{\mu}$ estimated from the original alignment in the distribution of ξ_R and μ_R estimated from the randomized alignments. Ideally, we would perform a large number of randomizations (e.g. $R > 100$) and directly compute p-values from the ranks of $\hat{\xi}$ and $\hat{\mu}$. However, for computational reasons we used only $R = 10$ randomized alignments and chose to reject the hypothesis of independent evolution if $\hat{\xi} < \xi_R$ and $\hat{\mu} < \mu_R$ for all R. Conversely, if the estimated number of switches ξ or the substitution rate μ are ranked within the distribution of ξ_R and μ_R, we consider that a scenario of independent evolution cannot be rejected. There are thus two (indistinguishable) scenarios that will produce microbial alignments that won't reject our test of independent evolution: environmental acquisition and vertical transmission with highly frequent host switches.

2.1.6 | Detecting transmitted OTUs

Based on the analyses above and our definition of modes of inheritance, we sort the OTUs into two different categories: the *transmitted* OTUs (those that reject the hypothesis of independent evolution, either because they are strictly vertically transmitted, or because they are vertically transmitted with few host switches) and the *independent* OTUs (those that do not reject the hypothesis of independent evolution, either because they are environmentally acquired, or because they experienced enough host switches to be

indistinguishable from a scenario of environmental acquisition). In practice, there is no universal similarity threshold that will provide the 'right' biological unit delineation across all microbial groups (Sanders, Powell, Kronauer, et al., 2014; Figure S2). 'Over-splitting' a biological unit using a similarity threshold that is too high for that biological unit will reduce statistical signal (each subunit will be represented in fewer hosts) and will miss host switches between subunits (given that subunits will be analysed independently). 'Over-merging' OTUs using a similarity threshold that is too low will tend to blur a signal of transmission and will overestimate substitution rates, because alignments will mix sequences from distinct biological units. By using several clustering thresholds, we can hope to find one that properly delimitates biological units. Given that vertical transmission tends to be erased by improper delimitation, if it is detected for at least one threshold, then it suggests that it is the 'right' threshold and that vertical transmission does indeed occur.

2.1.7 | Implementation

All the scripts of our model are written in R (R Core Team 2019), using the packages APE, PHANGORN and PHYTOOLS for the manipulations of phylogenetic trees (Paradis, Claude, & Strimmer, 2004; Revell, 2012; Schliep, 2011), and are freely available on GitHub (https://github.com/hmorlon/PANDA) and in the R package RPANDA (Morlon et al., 2015). Some internal functions computing the likelihood are coded in C++. We also used the packages PARALLEL, EXPM, GGPLOT2, RESHAPE2, RCPP and R2HTML for the technical aspects of the scripts. All outputs of our model (e.g. parameter estimation and model selection) are concatenated in a user-friendly HTML file with different formats (e.g. tables, values, pdf plot and diagrams). We provide a tutorial in https://github.com/BPerezLamarque/HOME/blob/master/README.md.

The computational time depends both on the number of host (n) and on the number of trees (S) used in the likelihood inference; examples of computation time are provided in Figure S3.

2.2 | Testing our approach with simulations

We performed a series of simulations to test the ability of our approach to recover simulated parameter values and evolutionary scenarios. We calibrated our choices of tree size, alignment size and parameter values so as to obtain simulated data comparable to those of the great ape microbiota data (Figure S9 and Table S2). We considered 3 independent host trees of size n = 20 (T_1, T_2 and T_3) simulated under a Yule model (no extinction) using the function PBTREE from PHYTOOLS. We scaled these trees to a total branch length of 1. On each of these host trees, we considered a scenario of strict vertical transmission (ξ = 0), scenarios of vertical transmission with host switches ξ = [1, 2, 3, 5, 7, 10], and a scenario of environmental acquisition; each of these scenarios was obtained by simulating the corresponding microbial trees T_B. For the scenario of strict vertical transmission, T_B = T. For scenarios of host switches, 15 T_B per ξ value were derived from T. For the scenario of environmental acquisition, 20 T_B with n tips were simulated under a Yule model independently from T, using the same procedure as above. Finally, we simulated on each T_B the evolution of microbial sequences of a total length N = 300 using our own codes, with a probability 0.1 for each site to be variable. We simulated the K80 stochastic nucleotide substitution process with a ratio of transition/transversion rate κ = 0.66 and three different values of substitution rate (μ = 0.5, 1 or 1.5). The realized proportion of segregating sites was quite variable and comparable to empirical alignments (Fig. S9). We simulated 20 alignments A per substitution rate on T for the scenario of strict vertical transmission (180 alignments total), and 1 alignment per T_B per substitution rate for the scenarios of host-switch (135 alignments per ξ value) and environmental acquisition (180 alignments). Thereafter, we call 'ξ-switches alignment' an alignment simulated with ξ switches on T and 'independent alignment' an alignment simulated under the environmental acquisition scenario (i.e. independently from T).

We applied our inference approach to each simulated couple of T and A and compared the estimated parameters ($\hat{\xi}$, $\hat{\mu}$, and $\hat{\kappa}$) to the simulated values. We used mixed linear models with the host tree (T_1, T_2 and T_3) as a random effect (R package NLME). We tested homoscedasticity and normality of the model residuals and considered a p-value of 0.05 as significant. We also evaluated the type I and type II errors associated with our tests of strict vertical transmission and independent evolution.

2.3 | Empirical application: great apes microbiota

We illustrate our approach using data from Ochman et al. (2010); this paper is one of the first paper testing hypotheses about codiversification in the well-studied clade of great apes (using phylosymbiotic patterns), and the associated data have been used in other papers aimed at studying codiversification (Sanders, Powell, Kronauer, et al., 2014). The data set consists of faecal samples collected from 26 wild-living hominids, including eastern and western African gorillas (two individuals of *G. gorilla* and two individuals of *G. beringei*), bonobos (6 individuals of *P. paniscus*) and three subspecies of chimpanzees (five individuals of *P. t. schweinfurthii*, seven individuals of *P. t. troglodytes* and two individuals of *P. t. ellioti*), as well as two humans from Africa and America (*H. sapiens*).

Ochman et al. (2010) extracted DNA from the faecal samples, PCR-amplified the DNA for the 16S rRNA V6 gene region using universal primers and finally sequenced the PCR product using 454 (Life Sciences/Roche). They obtained 1,292,542 reads after sequence quality trimming and barcodes removal. Gut microbiota are now sequenced with more coverage than what was possible at the time of the Ochman paper, yet these data represent a good application of our approach.

We obtained the reads from Dryad (http://datadryad.org/resource/ https://doi.org/10.5061/dryad.023s6). We used python scripts from the Brazilian Microbiome Project (BMP, available on http://www.brmicrobiome.org/) (Pylro et al., 2014) which combines scripts from QIIME 1.8.0 (Caporaso et al., 2010) and USEARCH 7 (Edgar, 2013) as well as our own bash codes. We

merged raw reads from all the hosts and processed them step by step:

1. Dereplication: we discarded all the singletons and sorted the sequences by abundance using USEARCH commands derep_fulllength and sortbysize
2. Chimera filtering and OTU clustering: we removed chimers and clustered sequences into OTUs using the -cluster_otus command of the UPARSE pipeline (Edgar, 2013). We chose a 1.0, 3.0 or 5.0 OTU radius (the maximum difference between pairs of OTU member sequences), which corresponds to a minimum identity of 99%, 97% and 95%. We performed an additional chimera filtering step using uchime_ref with the RDP database as a reference (http://drive5.com/uchime/rdp_gold.fa). We obtained 1,074 OTUs at 95%, 1,793 at 97%, and 4,935 at 99% (Table S1).
3. Taxonomic assignation: we assigned taxonomy using a representative sequence for each OTU generated (with -cluster_otus), using assign_taxonomy.py from QIIME and the latest version of the Greengenes database (http://greengenes.secondgenome.com), or using BLAST when Greengenes did not assign taxonomy with enough resolution.
4. Mapping reads to OTUs and OTU table construction: we used the usearch_global command to map all the reads from the different samples to these taxonomy-assigned OTUs. Then, we used make_otu_table.py and BMP scripts to build the OTU table (a list of all the OTUs with their abundance by host individual).
5. Core-OTUs selection: we selected the 'core' OTUs as the ones that occurred in at least 75% of the host individuals, using the compute_core_microbiome.py script from QIIME. This resulted in 134 core OTUs at 95%, 120 at 97%, and 71 at 99% (there are more OTUs at 99% than at 97% and 95%, but a much smaller proportion that are core OTUs, Table S1).
6. Making intra-OTU alignments: discarding few OTUs that had unvaried alignments, we obtained 130 core OTUs at 95%, 110 core OTUs at 97%, and 66 core OTUs at 99% similarity thresholds (Table S1). Microbial genetic variability within each OTU and within each host individual (hereafter referred to as 'intra-individual variability') was quite high, sometimes higher than inter-individual variability (Figure S10a–c), suggesting that it was due to PCR and sequencing artefacts rather than true variability. Therefore, we built the bacterial alignment for a given OTU by selecting for each host individual the most abundant sequence among all the reads mapped to that OTU. This sequence is less likely to be subject to sequencing errors.

Finally, we applied our approach to each core OTU separately, and to the nexus tree of the 26 host individuals, constructed with mitochondrial markers provided in the supplementary data of the article, scaled to a total branch length of 1. We used this individual-level tree instead of the species- or subspecies-level tree in order to increase tree size (there are only seven subspecies in our great apes tree); this approach also provides a way to account for microbial genetic variability within host subspecies (hereafter referred to as 'intraspecific variability'). We arbitrarily resolved intra subspecies polytomies by assigning quasi-null branch lengths (10^{-4}) to the corresponding branches. We classified the OTUs into 'transmitted' and 'independent' OTUs; among the transmitted OTUs, we distinguished those where the transmission is strictly vertical, and for the others, we recorded the estimated number of host switches. In order to get an idea of the proportion of the microbiota that is transmitted, we also recorded the number of reads corresponding to the transmitted OTUs.

2.4 | Accounting for intra-host genetic variability

Our treatment of the great ape data illustrates an approach to account for intra-host microbial genetic variability: instead of running HOME on a species-level host tree (with a single representative microbial sequence per host species), it can be run on an individual-level host tree, with arbitrarily small intraspecific branch lengths. Because this usage of HOME is slightly different from the case envisioned in our description of the approach, we tested its behaviour. We simulated the evolution of microbial alignments on the great apes subspecies tree with a range of intraspecific variability similar to the range observed in the great apes alignments. For each OTU alignment, we defined intraspecific variability (V) as the mean nucleotidic diversity within host subspecies (computed using Nei's estimator; Ferretti, Raineri, & Ramos-Onsins, 2012) divided by the total nucleotidic diversity computed on the entire alignment. We simulated a total of 180 alignments according to three scenarios: strict vertical transmission ($\xi = 0$), transmission with five host switches ($\xi = 5$), and environmental acquisition. For every scenario, we simulated intraspecific variability by extending the stochastic process generating nucleotidic substitution on every sequence for a time range that allowed to obtain levels of intraspecific variability that corresponded to the empirical level of intraspecific variability (Figure S10d–i). We ran HOME on each of these simulated alignments and evaluated its performance, in terms of parameter estimation and model selection, when there was no intraspecific variability ($V = 0$), low and intermediate intraspecific variability ($0 < V < 0.5$), and high intraspecific variability ($V > 0.5$).

3 | RESULTS

3.1 | Performance of HOME

Likelihood landscapes typically display a single peak, illustrating that ξ and μ are in general identifiable (Figure S1). Rarefaction curves also indicate that using $S = 10^4$ trees to compute the likelihood provides a good approximation (Figure S4). Testing the performance of HOME using intensive simulations, we find a reasonable ability to recover simulated parameter values (Figure 3). Estimates of the number of switches $\hat{\xi}$ are highly correlated with simulated ξ values, although the approach tends to overestimate the true number of switches when there are very few (<2) and to underestimate this number when there are many (Figure 3a). The linear regression confirms these results $\hat{\xi} = 2.15$ ($F_{dl=606} = 1{,}015$, p-value <.0001) + $\xi * 0.58$

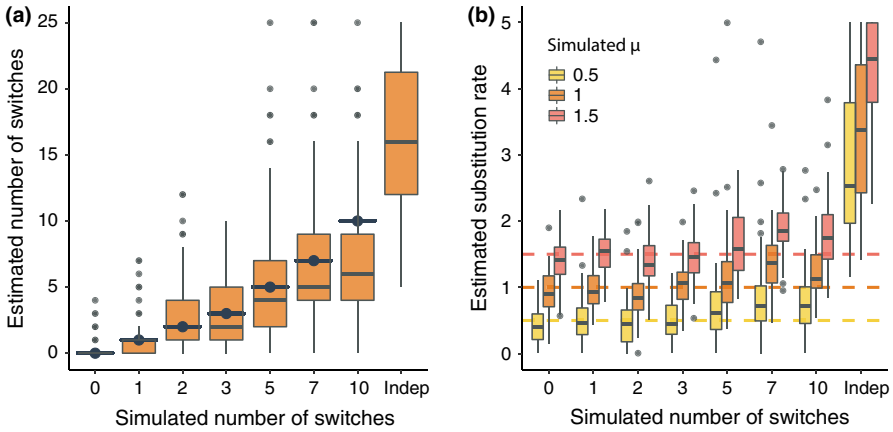


FIGURE 3 Parameter estimation. Estimated versus simulated number of switches ξ (a) and substitution rate μ (b) under various evolutionary scenarios (strict vertical transmission, vertical transmission with a given number of switches, and independent evolution, referred in the figure as 'indep'). Simulated values are represented by blue ticks in (a) and dashed lines in (b). Boxplots present the median surrounded by the first and third quartile, and whiskers extended to the extreme values but no further than 1.5 of the inter-quartile range.

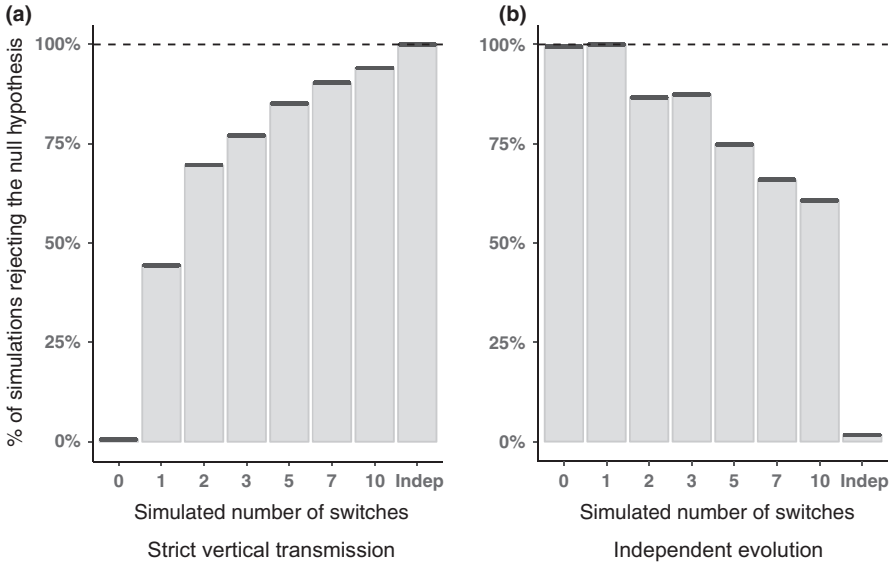


FIGURE 4 Model selection. Percentage of simulated alignments for which the null hypothesis of strict vertical transmission (a) or independent evolution (b) is rejected under various evolutionary scenarios (strict vertical transmission, vertical transmission with a given number of switches, and independent evolution, referred in the figure as 'indep')

($F_{dl=606}$ = 141, *p*-value <.0001). The ability to recover the true number of switches does not depend on the simulated substitution rate ($F_{dl=606}$ = 0.2601, *p*-value = .61; Figure S5). The substitution rate is rather well estimated (Figure 3b), although it tends to be slightly overestimated when the simulated number of switches exceeds 3 (slope 0.04; $F_{dl=606}$ = 45.9, *p*-value <.0001; Figure 3b). The simulated transition/transversion rate ratio κ is well estimated (median ± *SD* = 0.68 ± 0.17), although it is slightly underestimated when the substitution rate is high (slope of −0.015; $F_{dl=606}$ = 12, *p*-value = .0007). For alignments simulated independently from the host tree, the approach estimates a high number of switches (median ± *SD* = 16 ± 6.2, Figure 3a), and highly overestimates the substitution rate (Figure 3b). The type of host tree (T_1, T_2 or T_3) has little impact on the estimation of ξ (it explains <3% of the total variance, Figure S5), μ (around 10%, Figure S6) and κ (<0.01%).

Our model selection procedure has very low type I error rates, and type II error rates that depend on the situation (Figure 4): the hypothesis of strict vertical transmission was nearly never rejected when transmission was indeed strictly vertical (1/180, type I error = 0.0056%) and always rejected under environmental acquisition (Figure 4a); conversely, the hypothesis of independent evolution was almost always rejected when transmission was strictly vertical (1/180) and almost never rejected under environmental acquisition (3/180, type I error = 0.017%, Figure 4b). While the type I error rates of the two tests are low, their power to detect a scenario of strict vertical transmission with host switches is variable. In the case of the test of strict vertical transmission, the power ranges from 95% for ξ = 10 to 45% when ξ = 1 (Figure 4a). In the case of the test of independent evolution, the power ranges from 100% for ξ = 1 to 60% for ξ = 10, and it would decrease further with more switches (Figure 4b). In both cases, the power increases when the substitution rate μ is larger (Figure S7).

When HOME is applied to an individual-level host tree in order to account for intraspecific microbial genetic variability, type I error rates associated to the test of independent evolution remain very low

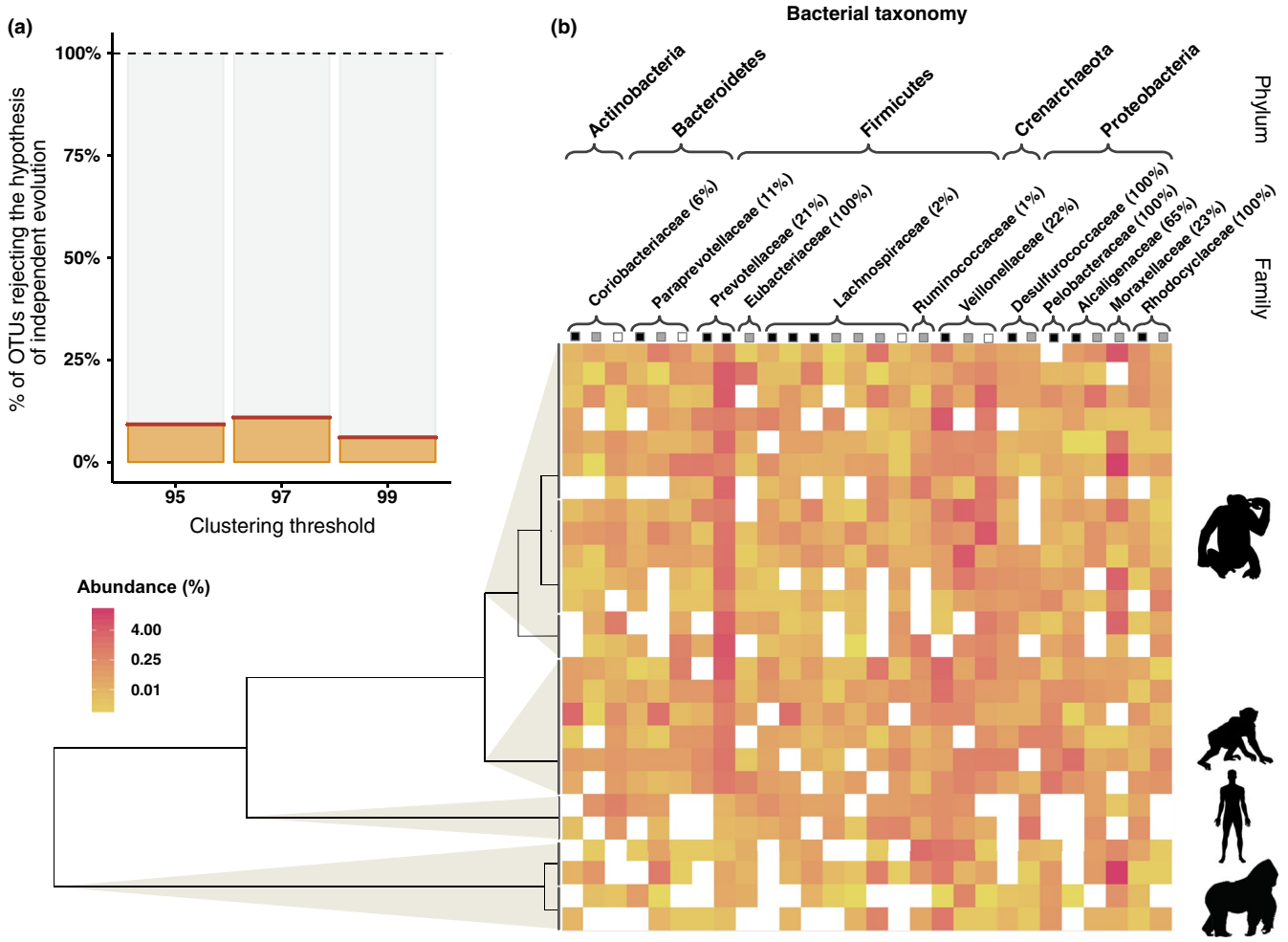


FIGURE 5 Transmitted OTUs in the great ape microbiota: (a) Percentage of OTUs rejecting the hypothesis of independent evolution at the three % similarity clustering thresholds. (b) Phylogenetic tree of greats apes and their associated transmitted OTUs (black: 95% similarity threshold, grey: 97%, white: 99%). The percentage indicated in parenthesis for each family is the estimated percentage of transmitted raw reads in the family. The colour of the heat map represents for OTU each host the percentage of raw reads of the OTU in the entire microbiota of the host. A white square means that the OTU is not found in the host.

regardless of the magnitude of the variability (Figure S8). The confidence in the estimation of the parameters (ξ and μ) remains good for low values of intraspecific variability ($V < 0.5$), but decreases with increasing variability ($V > 0.5$). The type I error rate associated to the test of strict vertical transmission increases with increasing variability, and the power of the two tests decreases with increasing variability.

3.2 | Modes of inheritance in the great apes microbiota

Applying HOME to great apes gut microbiota data, we found that among the core OTUs with at least one segregating site, ~1 in 10 OTUs is transmitted (i.e. rejects the test of independent evolution, Figure 5a); more specifically, the ratios of transmitted OTUs (and strictly vertically transmitted OTUs) were the following: 12(8)/130 at 95%, 12(10)/110 at 97%, and 4(4)/66 at 99%. In terms of relative abundance, 108,206 raw sequences in a total of 1,292,542 (8.4%) belonged to transmitted OTUs (Table S3). Almost half of the sequences from transmitted OTUs (49,508) were from an *Acinetobacter* bacterium (Moraxellaceae family); another important pool of these sequences was from the family Prevotellaceae (28,843 reads). In total, 12 bacterial families (in 27) contained OTUs that were transmitted, including Veillonellaceae, Lachnospiraceae, Ruminococcaceae and Paraprevotellaceae (Figure 5b, Table S4). Some of these families (e.g. Desulfurococcaceae, Pelobacteraceae, Rhodocyclaceae and Eubacteriaceae) were entirely made of a transmitted OTU, while others also had many OTUs and/or sequences that were independent (e.g. Ruminococcaceae, Lachnospiraceae and Coriobacteriaceae). Transmitted OTUs were in general not more abundant in a particular group of host species, except for the Prevotellaceae, that were overall more abundant in bonobos and chimpanzees than in gorillas and humans (Figure 5b).

The sequence length, proportion of segregating sites and intra-individual variability of the OTUs inferred as transmitted were similar to those of other OTUs (Figure S9 and Table S2), suggesting that HOME is not biased towards detecting vertical transmission in OTUs

with specific characteristics. However, the intraspecific variability of OTUs inferred as transmitted tends to be smaller than that of other OTUs (Table S5), which is consistent with our simulation results showing that the power to detect vertical transmission decreases with increasing intraspecific variability.

4 | DISCUSSION

We developed HOME, a likelihood-based approach for studying the inheritance of microbiota during the evolution of their hosts from metabarcoding data. We showed using simulations that even relatively short reads can help identify modes of inheritance, without the need to build a microbial phylogenetic tree. Applying HOME to great apes microbiota data, we identified a set of transmitted gut bacteria that account on average for 8.4% of the total reads of the gut microbiota.

Our combination of model fitting and hypothesis testing helps identify modes of inheritance. We see the estimate of the number of switches as a good indicator of modes of inheritance (from strict vertical transmission for low ξ values to transmission with high rates of horizontal switches or environmental acquisition for high ξ values) rather than as an accurate estimation of past host switches. We have indeed shown that ξ tends to be underestimated when quite many switches are simulated on a fixed host tree. In nature, this underestimation may be even more pronounced, as our model ignores host switches that happened in lineages not represented in the phylogeny, as a result of either extinction or undersampling (Szöllosi et al., 2013). In line with these results, we find that the hypothesis of strict vertical transmission is often not rejected when there are in fact host switches. On the other hand, we can also estimate a positive ξ from data simulated under strict vertical transmission; however, in this case, a model with host switches will in general not be selected when compared to a model of strict vertical transmission. Hence, if the hypothesis of strict vertical transmission is rejected, one can conclude with confidence that host switches occurred (or that the microbial unit was environmentally acquired). Similarly, the hypothesis of independent evolution is often not rejected when the transmission is actually vertical with rather frequent host switches, and rarely rejected in scenarios of environmental acquisition, such that when it is rejected, one can conclude with confidence that the microbial unit is transmitted. Said differently, our approach is conservative in its identification of transmitted OTUs; and when an OTU is identified as being transmitted, our approach is conservative in its identification of switches.

We assessed the performance of HOME in a limited set of conditions (e.g. host tree size, sequence length, substitution rates) calibrated on the great apes microbiota data. We can expect that the power of the model will increase with host tree size and the number of segregating sites in the microbial alignment. As the latter is a combination of sequence length, substitution rate and hosts divergence times, there is no universal guidelines on the applicability of the model to a particular marker, sequencing technology, and host clade age. Rather, the marker and sequencing technologies must be adapted to the study system. For example, the 200–300 bp-long 16S rRNA V6 gene region sequenced with 454 sequencing used on great apes in our empirical application was enough to identify some transmitted microbial OTUs, but it probably missed others that had too low substitution rates to leave a detectable signal. Similarly, it would probably have a too low resolution to detect variability between host species that diverged more recently than the great apes. In such cases, using longer sequences and/or markers that evolve more quickly can be necessary. Finally, we can expect that PCR and sequencing errors will blur the signal and reduce the power to detect transmitted OTUs, although this should be limited by selecting the most abundant sequence representative of each OTU for each host.

HOME is currently best suited to the study of microbiota transmission in recent, well-sampled host clades in which no or few extinctions occurred, since it does not account for unsampled host lineages, nor for host extinctions. For example, HOME would be well adapted to the study of microbiota transmission in some vertebrates and invertebrate clades, for which microbiota sequencing data are already available (Amato et al., 2019; Brooks, Kohl, Brucker, van Opstal, & Bordenstein, 2016; Ren, Kahrl, Wu, & Cox, 2016). Ignoring extinction is reasonable at the small evolutionary scales of such groups or the great apes (Ochman et al., 2010), but it would not be at larger evolutionary timescales such as across invertebrate or vertebrate species; in this case, accounting for host switches from now-extinct lineages is necessary (Szöllosi et al., 2013). Another reason why HOME is currently better adapted to studying recent rather than ancient host clades is that it does not account for extinction of symbiont lineages and therefore can only model the inheritance of OTUs shared across most species (i.e. core OTUs); the more divergent the host species, the less core OTUs there will be. Further developments of the model that would allow extending its relevance to a broader range of data include accounting for extinction and incomplete sampling in the host clade, as well as incorporating symbiont extinctions.

When it occurs, the support for vertical transmission of a given microbial unit arises from a phylogenetic signal in microbial sequences (i.e. a congruence between the phylogenetic similarity of host species and the molecular similarity of the microbes they host). However, such congruence can also arise from processes not accounted for in our model, such as geographic or environmental effects; for example, if there is a phylogenetic/molecular signal in the geographic or habitat distribution of hosts/ microbes, or if the host environment creates microbial selective filters, this could result in a phylogenetic signal in microbial sequences that could be misleadingly interpreted as vertical transmission. We have not evaluated the robustness of our approach to such effects. Future developments could involve reconstructing ancestral areas/habitats or host environments on the host phylogeny in order to distinguish a phylogenetic signal truly driven by vertical transmission versus other effects.

In the construction of the model, we have made the important assumption that there is no microbial genetic variability within host species, such that each microbial OTU is represented by at most one sequence in each host. This is quite unlikely in natural microbial

populations where multiple microbial strains can colonize a host species (Ellegaard & Engel, 2016). In our empirical application, we tackled this limitation by representing each host species by several individuals, using approximately zero-length branches to split conspecifics in the host phylogeny. Although our simulations show that the statistical power of our tests decreases strongly when intraspecific variability is high, they also show that the hypothesis of environmental acquisition is rarely rejected when the acquisition is indeed environmental. Hence, HOME is unlikely to misleadingly identify transmitted OTUs, especially in the presence of intraspecific variability. Another (more satisfying) approach would be to directly account for intraspecific variability in microbial sequences in the likelihood computation; this could for example be done by representing the data by – at each tip of the host phylogeny and for each nucleotidic site – a vector of probabilities of states A, C, G and T representing the intra-host relative abundance of the four bases at the given nucleotidic position. In this case, we would directly use the variation given at the level of amplicon sequence variants (ASVs) (Callahan et al., 2016). Alternatively, further developments of HOME incorporating horizontal host switches without replacement (i.e. the persistence of both ancestral and newly acquired symbionts in a lineage), as well as dynamics of duplication and recolonization, would allow better accounting for intra-host genetic variability. In addition, rather than considering each OTU as a separately evolving unit, it would be interesting to account for interactions between these units, that can for example lead to competitive exclusion (Koeppel & Wu, 2014) or interdependency (e.g. adaptive gene loss; Morris, Lenski, & Zinser, 2012), and are crucial aspects of microbial community assembly.

In the great apes gut microbiota, we found that the major part of the microbiota (91.6%) is constituted of bacteria which acquisition scenario is not distinguishable from one that is independent from the great apes phylogeny (Amato et al., 2019; Moeller et al., 2013). Still, we identified OTUs representing 8.4% of the total number of reads that are transmitted across generations during millions of years of evolution. Given the low phylogenetic signal in the geographic distribution of the hosts (see Ochman et al., 2010), these OTUs are likely truly transmitted vertically. And given that HOME is conservative in its identification of transmitted OTUs, 8.4% is a lower bound estimate of the relative abundance of the microbiota that is vertically transmitted. Thus, our results suggest that the phylosymbiosis pattern observed by Ochman et al. (Ochman et al., 2010) is partially driven by vertically transmitted bacteria, as suggested by Sanders, Powell, Kronauer, et al. (2014). Our approach offers the advantage of investigating the whole microbiota without an a priori on which families might be transmitted; it identified 12 microbial families with transmitted OTUs. This is a good complement to approaches that focus on few candidate families, such as in Moeller et al.'s study (Moeller et al., 2016). In the later study, the authors used 3 specific primer pairs to focus on 3 families (Bacteroidaceae, Bifidobacteriaceae and Lachnospiracea) and showed that phylogenies representing the Bifidobacteriaceae and Bacteroidaceae were congruent with the apes phylogeny, suggesting that codiversification occurred in these two families. Unfortunately, neither Bifidobacteriaceae nor Bacteroidaceae were represented in the core OTUs in Ochman et al.'s data, even with a 95% similarity threshold: those bacteria were either not sampled, badly processed during DNA extraction and PCR, poorly taxonomically annotated, or too divergent to be merged into a single core OTU defined at 95%. Conversely, while Moeller et al. did not find any signal of cophylogeny in the Lachnospiraceae family, we found 3 transmitted OTUs belonging to this family. The authors investigated the phylogenetic relationships between all the amplified strains of Lachnospiraceae and whether they match the phylogenetic tree of great apes. This illustrates the utility of our approach, which investigates transmission modes of separate OTUs within bacterial families, rather than considering in a single evolutionary framework all the sequences from the same family.

Among the families in which we found transmitted OTUs, some are well known for having mutualistic properties. For example, the Lachnospiraceae, Paraprevotellaceae and Rhodocyclales families are involved in breaking down complex carbohydrates in the gut; they have even evolved fibrolytic specialization in gut communities (Biddle, Stewart, Blanchard, & Leschine, 2013). These vertically transmitted fibrolytic bacteria, which have been codiversifying for millions of years with the great apes, would thus constitute for the great apes a conserved reservoir of gut symbionts able to digest carbohydrates and might have facilitated frequent and rapid dietary shifts during the evolutionary history of hominids (Hardy, Brand-Miller, Brown, Thomas, & Copeland, 2015; Head, Boesch, Makaga, & Robbins, 2011; Muegge et al., 2011). However, why these particular bacteria are faithfully vertically transmitted while other digesting gut bacteria seem largely environmentally acquired (or vertically transmitted with frequent host switches) remains unclear.

DNA metabarcoding data for microbiota are being collected across multiple hosts at an unprecedented scale. Our approach allows identifying, among numerous microbial units, those that are vertically transmitted and potentially co-evolving with their hosts. The current implementation of our model is entirely adapted to applications to other data sets using different sequencing techniques, clustering methods and de-noising algorithms. Being able to identify vertically transmitted microbial units is an important step towards a better understanding of the role of microbial communities on the long-term evolution of their hosts.

ACKNOWLEDGEMENTS

The authors thank Ana Alfonso Silva, Leandro Aristide, Julien Clavel, Carmelo Fruciano, Eric Lewitus, Sophia Lambert, Odile Maliet, Marc Manceau, Olivier Missa and Guilhem Sommeria-Klein for helpful comments on the article. They also thank Florian Hartig, Marc-André Selosse and Florent Martos for helpful discussions, as well as the Associate Editor and the anonymous reviewers for their constructive comments on the previous version of the manuscript. This work was supported by the Centre national de la recherche scientifique (CNRS), the grant PANDA from the European Research Council (ERC-CoG) attributed to H.M., the

Ecole Doctorale FIRE - Programme Bettencourt, and a doctoral fellowship from the École Normale Supérieure de Paris attributed to B.P.L.

CONFLICTS OF INTEREST

The authors declare no conflicts of interest.

AUTHOR CONTRIBUTIONS

B.P.L and H.M designed research, B.P.L performed research, and B.P.L and H.M analysed data and wrote the paper.

DATA AVAILABILITY STATEMENT

The implementation of HOME is available on github (https://github.com/hmorlon/PANDA) and in the R package RPANDA (Morlon et al., 2015). We provide a tutorial and scripts to prepare the data in https://github.com/BPerezLamarque/HOME/blob/master/README.md. The sequences used in our empirical applications are available in https://doi.org/10.5061/dryad.023s6/3 (Sanders, Powell, Kronaue, et al., 2014).

ORCID

Benoît Perez-Lamarque https://orcid.org/0000-0001-7112-7197
Hélène Morlon https://orcid.org/0000-0002-3195-7521

REFERENCES

Amato, K. R., G. Sanders, J., Song, S. J., Nute, M., Metcalf, J. L., Thompson, L. R., … R. Leigh, S. (2019). Evolutionary trends in host physiology outweigh dietary niche in structuring primate gut microbiomes. *The ISME Journal*, *13*(3), 576–587. https://doi.org/10.1038/s41396-018-0175-0

Bailly-Bechet, M., Martins-Simões, P., Szöllősi, G. J., Mialdea, G., Sagot, M.-F., & Charlat, S. (2017). How long does Wolbachia remain on board? *Molecular Biology and Evolution*, *34*(5), 1183–1193. https://doi.org/10.1093/molbev/msx073

Balbuena, J. A., Míguez-Lozano, R., & Blasco-Costa, I. (2013). PACo: A novel procrustes application to cophylogenetic analysis. *PLoS ONE*, *8*(4), e61048. https://doi.org/10.1371/journal.pone.0061048

Biddle, A., Stewart, L., Blanchard, J., & Leschine, S. (2013). Untangling the genetic basis of fibrolytic specialization by lachnospiraceae and ruminococcaceae in diverse gut communities. *Diversity*, *5*(3), 627–640. https://doi.org/10.3390/d5030627

Bordenstein, S. R., & Theis, K. R. (2015). Host biology in light of the microbiome: Ten principles of holobionts and hologenomes. *PLoS Biology*, *13*(8), 1–23. https://doi.org/10.1371/journal.pbio.1002226

Bright, M., & Bulgheresi, S. (2010). A complex journey : Transmission of microbial symbionts. *Nature Reviews Microbiology*, *8*(3), 218–230. https://doi.org/10.1038/nrmicro2262.A

Brooks, A. W., Kohl, K. D., Brucker, R. M., van Opstal, E. J., & Bordenstein, S. R. (2016). Phylosymbiosis: Relationships and functional effects of microbial communities across host evolutionary history. *PLOS Biology*, *14*(11), e2000225. https://doi.org/10.1371/journal.pbio.2000225

Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., & Holmes, S. P. (2016). DADA2: High-resolution sample inference from Illumina amplicon data. *Nature Methods*, *13*(7), 581–583. https://doi.org/10.1038/nmeth.3869

Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., … Knight, R. (2010). QIIME allows analysis of high- throughput community sequencing data. *Nature Methods*, *7*(5), 335–336. https://doi.org/10.1038/nmeth0510-335

Conow, C., Fielder, D., Ovadia, Y., & Libeskind-Hadas, R. (2010). Jane: A new tool for the cophylogeny reconstruction problem. *Algorithms for Molecular Biology*, *5*(1), 1–10. https://doi.org/10.1186/1748-7188-5-16

de Vienne, D. M., Refrégier, G., López-Villavicencio, M., Tellier, A., Hood, M. E., & Giraud, T. (2013). Coespeciation vs hos-shift speciation: Methods for testing, evidence from natural associations and relation to coevolution. *New Phytologist*, *198*(2), 347–385.

Edgar, R. C. (2013). UPARSE: Highly accurate OTU sequences from microbial amplicon reads. *Nature Methods*, *10*(10), 996–998. https://doi.org/10.1038/nmeth.2604

Ellegaard, K. M., & Engel, P. (2016). Beyond 16S rRNA community profiling: Intra-species diversity in the gut microbiota. *Frontiers in Microbiology*, *7*, 1475. https://doi.org/10.3389/fmicb.2016.01475

Engel, P., & Moran, N. A. (2013). The gut microbiota of insects - diversity in structure and function. *FEMS Microbiology Reviews*, *37*(5), 699–735. https://doi.org/10.1111/1574-6976.12025

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: A maximum likelihood approach. *Journal of Molecular Evolution*, *17*(6), 368–376. https://doi.org/10.1007/BF01734359

Ferretti, L., Raineri, E., & Ramos-Onsins, S. (2012). Neutrality tests for sequences with missing data. *Genetics*, *191*(4), 1397–1401. https://doi.org/10.1534/genetics.112.139949

Groussin, M., Mazel, F., Sanders, J. G., Smillie, C. S., Lavergne, S., Thuiller, W., & Alm, E. J. (2017). Unraveling the processes shaping mammalian gut microbiomes over evolutionary time. *Nature Communications*, *8*, 14319. https://doi.org/10.1038/ncomms14319

Hacquard, S., Garrido-Oter, R., González, A., Spaepen, S., Ackermann, G., Lebeis, S., … Schulze-Lefert, P. (2015). Microbiota and host nutrition across plant and animal kingdoms. *Cell Host and Microbe*, *17*(5), 603–616. https://doi.org/10.1016/j.chom.2015.04.009

Hardy, K., Brand-Miller, J., Brown, K. D., Thomas, M. G., & Copeland, L. (2015). The importance of dietary carbohydrate in Human evolution. *The Quarterly Review of Biology*, *90*(3), 251–268. https://doi.org/10.1086/682587

Head, J. S., Boesch, C., Makaga, L., & Robbins, M. M. (2011). Sympatric chimpanzees (Pan troglodytes troglodytes) and gorillas (Gorilla gorilla gorilla) in Loango National Park, Gabon: Dietary composition, seasonality, and intersite comparisons. *International Journal of Primatology*, *32*(3), 755–775. https://doi.org/10.1007/s10764-011-9499-6

Henry, L. M., Peccoud, J., Simon, J.-C., Hadfield, J. D., Maiden, M. J. C., Ferrari, J., & Godfray, H. C. J. (2013). Horizontally transmitted symbionts and host colonization of ecological niches. *Current Biology*, *23*(17), 1713–1717. https://doi.org/10.1016/J.CUB.2013.07.029

Hommola, K., Smith, J. E., Qiu, Y., & Gilks, W. R. (2009). A permutation test of host-parasite cospeciation. *Molecular Biology and Evolution*, *26*(7), 1457–1468. https://doi.org/10.1093/molbev/msp062

Huelsenbeck, J. P., Rannala, B., & Larget, B. (2000). A Bayesian framework for the analysis of cospeciation. *Evolution; International Journal of Organic Evolution*, *54*(2), 352–364. https://doi.org/10.1111/j.0014-3820.2000.tb00039.x

Koeppel, A. F., & Wu, M. (2014). Species matter: The role of competition in the assembly of congeneric bacteria. *The ISME Journal*, *8*, 531–540. https://doi.org/10.1038/ismej.2013.180

Legendre, P., Desdevises, Y., & Bazin, E. (2002). A statistical test for host–parasite coevolution. *Systematic Biology*, *51*(2), 217–234. https://doi.org/10.1080/10635150252899734

Linz, B., Balloux, F., Moodley, Y., Manica, A., Liu, H., Roumagnac, P., … Achtman, M. (2007). An African origin for the intimate association between humans and Helicobacter pylori. *Nature*, *445*(7130), 915–918. https://doi.org/10.1038/nature05562

McFall-Ngai, M., Hadfield, M. G., Bosch, T. C. G., Carey, H. V., Domazet-Lošo, T., Douglas, A. E., … Wernegreen, J. J. (2013). Animals in a bacterial world, a new imperative for the life sciences. *Proceedings of the National Academy of Sciences*, *110*(9), 3229–3236. https://doi.org/10.1073/pnas.1218525110

McKenney, E. A., Maslanka, M., Rodrigo, A., & Yoder, A. D. (2018). Bamboo specialists from two mammalian orders (Primates, Carnivora) share a high number of low-abundance gut microbes. *Microbial Ecology*, *76*(1), 272–284. https://doi.org/10.1007/s00248-017-1114-8

Moeller, A. H., Caro-Quintero, A., Mjungu, D., Georgiev, A. V., Lonsdorf, E. V., Muller, M. N., … Ochman, H. (2016). Cospeciation of gut microbiota with hominids. *Science*, *353*(6297), 380–382. https://doi.org/10.1126/science.aaf3951

Moeller, A. H., Peeters, M., Ndjango, J. B., Li, Y., Hahn, B. H., & Ochman, H. (2013). Sympatric chimpanzees and gorillas harbor convergent gut microbial communities. *Genome Research*, https://doi.org/10.1101/gr.154773.113

Morlon, H., Lewitus, E., Condamine, F. L., Manceau, M., Clavel, J., & Drury, J. (2015). RPANDA: An R package for macroevolutionary analyses on phylogenetic trees. *Methods in Ecology and Evolution*, https://doi.org/10.1111/2041-210X.12526

Morris, J. J., Lenski, R. E., & Zinser, E. R. (2012 The black queen hypothesis: Evolution of dependencies through adaptive gene loss. *MBio*, *3*(2), e00036-12. https://doi.org/10.1128/mBio.00036-12

Muegge, B. D., Kuczynski, J., Knights, D., Clemente, J. C., Fontana, L., Henrissat, B., … Gordon, J. I. (2011). Diet drives convergence in gut microbiome functions across mammalian phylogeny and within humans. *Science*, *332*(6032), 970–974. https://doi.org/10.1126/science.1198719.Diet

Ochman, H., Worobey, M., Kuo, C. H., Ndjango, J. B. N., Peeters, M., Hahn, B. H., & Hugenholtz, P. (2010). Evolutionary relationships of wild hominids recapitulated by gut microbial communities. *PLoS Biology*, *8*(11), 3–10. https://doi.org/10.1371/journal.pbio.1000546

Page, R. D. M., & Charleston, M. A. (1998). Trees within trees: Phylogeny and historical associations. *Trends in Ecology and Evolution*, *13*(9), 356–359. https://doi.org/10.1016/S0169-5347(98)01438-4

Paradis, E., Claude, J., & Strimmer, K. (2004). APE: Analyses of phylogenetics and evolution in R language. *Bioinformatics*, *20*(2), 289–290. https://doi.org/10.1093/bioinformatics/btg412

Philippot, L., Raaijmakers, J. M., Lemanceau, P., & van der Putten, W. H. (2013). Going back to the roots: The microbial ecology of the rhizosphere. *Nature Reviews Microbiology*, *11*(11), 789–799. https://doi.org/10.1038/nrmicro3109

Pylro, V. S., Roesch, L. F. W., Ortega, J. M., do Amaral, A. M., Tótola, M. R., Hirsch, P. R., … Brazilian Microbiome Project Organization Committee. (2014). Brazilian Microbiome Project: Revealing the unexplored microbial diversity - challenges and prospects. *Microbial Ecology*, *67*(2), 237–241. https://doi.org/10.1007/s00248-013-0302-4

R Core Team (2019). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.Rproject.org/

Ren, T., Kahrl, A. F., Wu, M., & Cox, R. M. (2016). Does adaptive radiation of a host lineage promote ecological diversity of its bacterial communities? A test using gut microbiota of *Anolis* lizards. *Molecular Ecology*, *25*(19), 4793–4804. https://doi.org/10.1111/mec.13796

Revell, L. J. (2012). PHYTOOLS: An R package for phylogenetic comparative biology (and other things). *Methods in Ecology and Evolution*, *3*(2), 217–223. https://doi.org/10.1111/j.2041-210X.2011.00169.x

Rosenberg, E., & Zilber-Rosenberg, I. (2016). Microbes drive evolution of animals and plants: The hologenome concept. *MBio*, *7*(2), e01395-15. https://doi.org/10.1128/mBio.01395-15

Sanders, J. G., Powell, S., Kronaue, D. J. C., Vasconcelos, H. L., Fredrickson, M. E., & Pierce, N. E. (2014). Data from: Stability and phylogenetic correlation in gut microbiota: lessons from ants and apes. *Dryad Digital Repository*, https://doi.org/10.5061/dryad.023s6.

Sanders, J. G., Powell, S., Kronauer, D. J. C., Vasconcelos, H. L., Frederickson, M. E., & Pierce, N. E. (2014). Stability and phylogenetic correlation in gut microbiota: Lessons from ants and apes. *Molecular Ecology*, *23*(6), 1268–1283. https://doi.org/10.1111/mec.12611

Schliep, K. P. (2011). PHANGORN: Phylogenetic analysis in R. *Bioinformatics*, *27*(4), 592–593. https://doi.org/10.1093/bioinformatics/btq706

Shapira, M. (2016). Gut microbiotas and host evolution: Scaling up symbiosis. *Trends in Ecology and Evolution*, *31*(7), 539–549. https://doi.org/10.1016/j.tree.2016.03.006

Strimmer, K., & von Haeseler, A. (2009). Genetic distances and nucleotide substitution models. In A. -M. Vandamme, M. Salemi,& P. Lemey (Eds.), *The phylogenetic handbook: A practical approach to phylogenetic analysis and hypothesis testing* (pp. 111–125). Cambridge, UK: Cambridge University Press.

Szöllősi, G. J., Rosikiewicz, W., Boussau, B., Tannier, E., & Daubin, V. (2013). Efficient exploration of the space of reconciled gene trees. *Systematic Biology*, *62*(6), 901–912. https://doi.org/10.1093/sysbio/syt054

Szöllosi, G. J., Tannier, E., Lartillot, N., & Daubin, V. (2013). Lateral gene transfer from the dead. *Systematic Biology*, *62*(3), 386–397. https://doi.org/10.1093/sysbio/syt003

Truszkowski, J., & Goldman, N. (2016). Maximum likelihood phylogenetic inference is consistent on multiple sequence alignments, with or without gaps. *Systematic Biology*, *65*(2), 328–333. https://doi.org/10.1093/sysbio/syv089

Zilber-Rosenberg, I., & Rosenberg, E. (2008). Role of microorganisms in the evolution of animals and plants: The hologenome theory of evolution. *FEMS Microbiology Reviews*, *32*(5), 723–735. https://doi.org/10.1111/j.1574-6976.2008.00123.x

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Perez-Lamarque B, Morlon H. Characterizing symbiont inheritance during host–microbiota evolution: Application to the great apes gut microbiota. *Mol Ecol Resour.* 2019;00:1–13. https://doi.org/10.1111/1755-0998.13063

MOLECULAR ECOLOGY RESOURCES

Supplemental Information for:

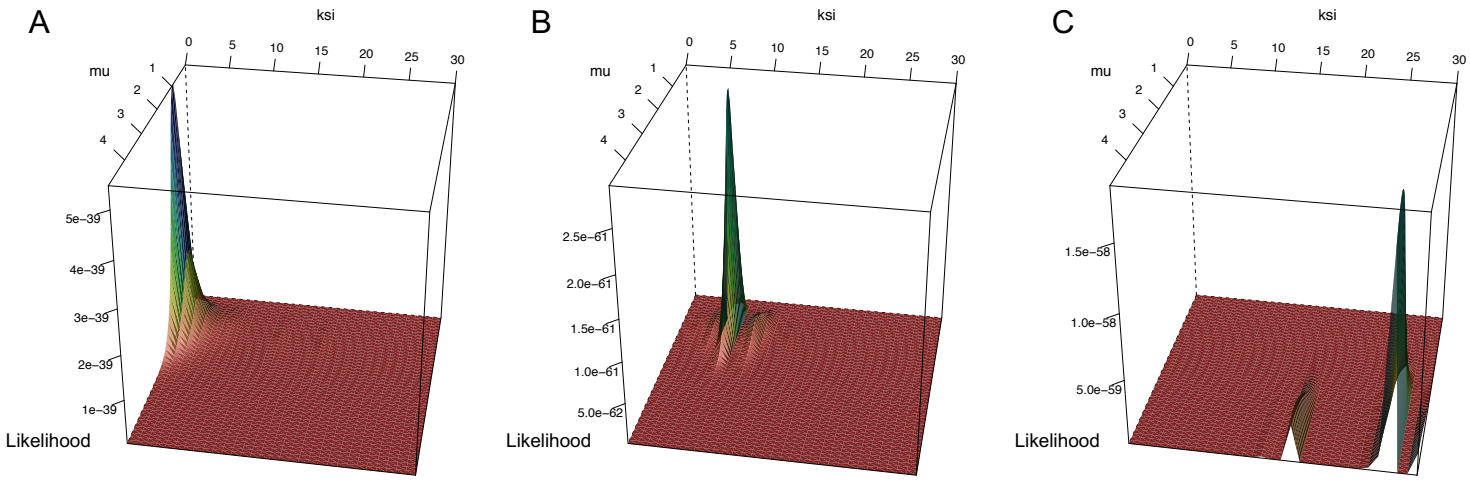
Characterizing symbiont inheritance during host-microbiota evolution: application to the great apes gut microbiota

Benoît Perez-Lamarque, Hélène Morlon

Supplemental Figures

Supp. Figure S1: Likelihood landscapes

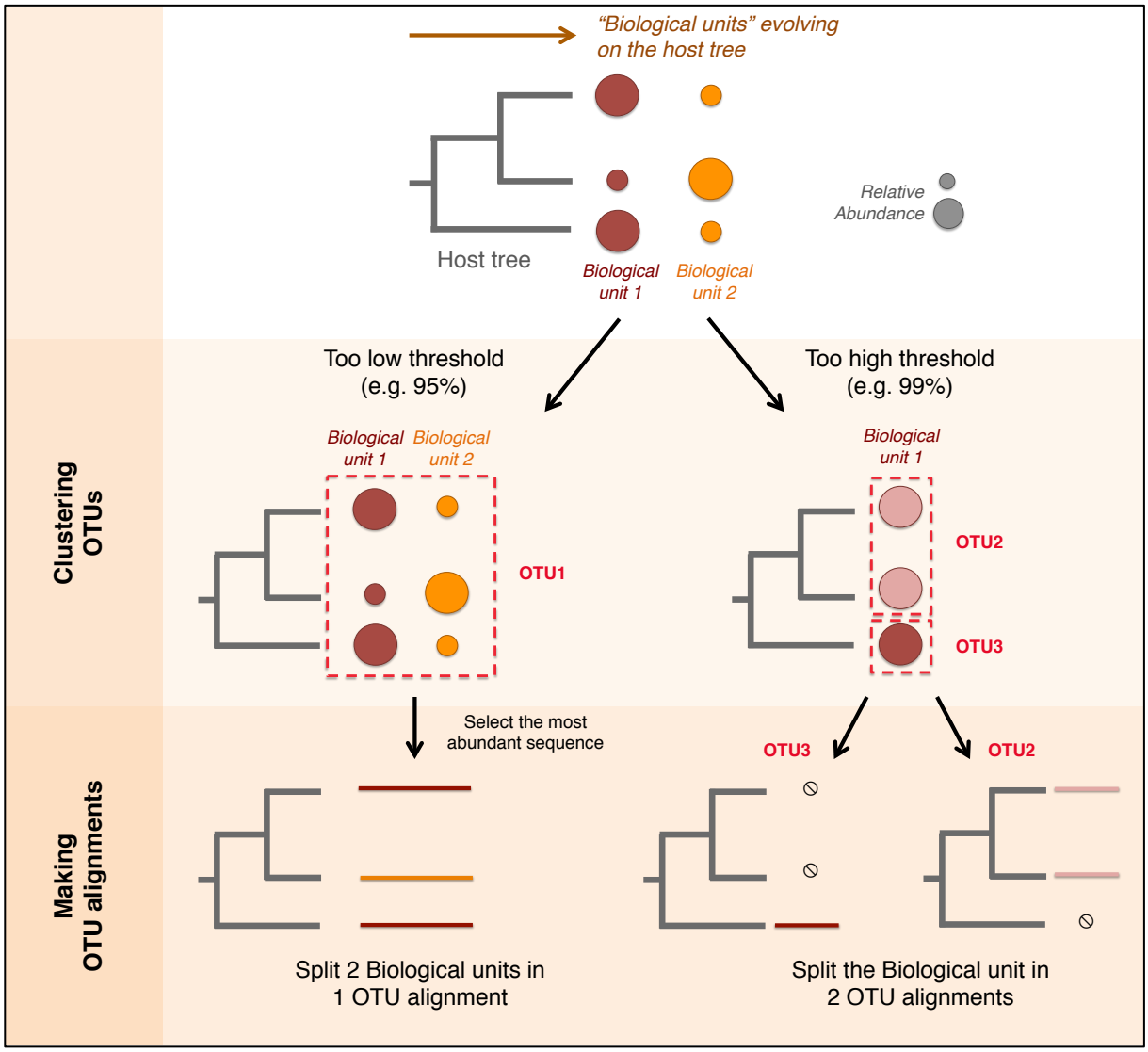
Examples of likelihood landscapes under scenarios of: A) strict vertical transmission, B) vertical transmission with 5 host-switches, and C) environmental acquisition. These landscapes were obtained from alignments simulated on a 20 tips host tree (simulated under the Yule process), with parameters similar to those used in the section "*Testing our approach with simulations*". These landscapes have a clearly identified peak that corresponds to the most likely parameter values, illustrating the identifiability of the model.



MOLECULAR ECOLOGY
RESOURCES

Supp. Figure S2: Detecting transmitted OTUs

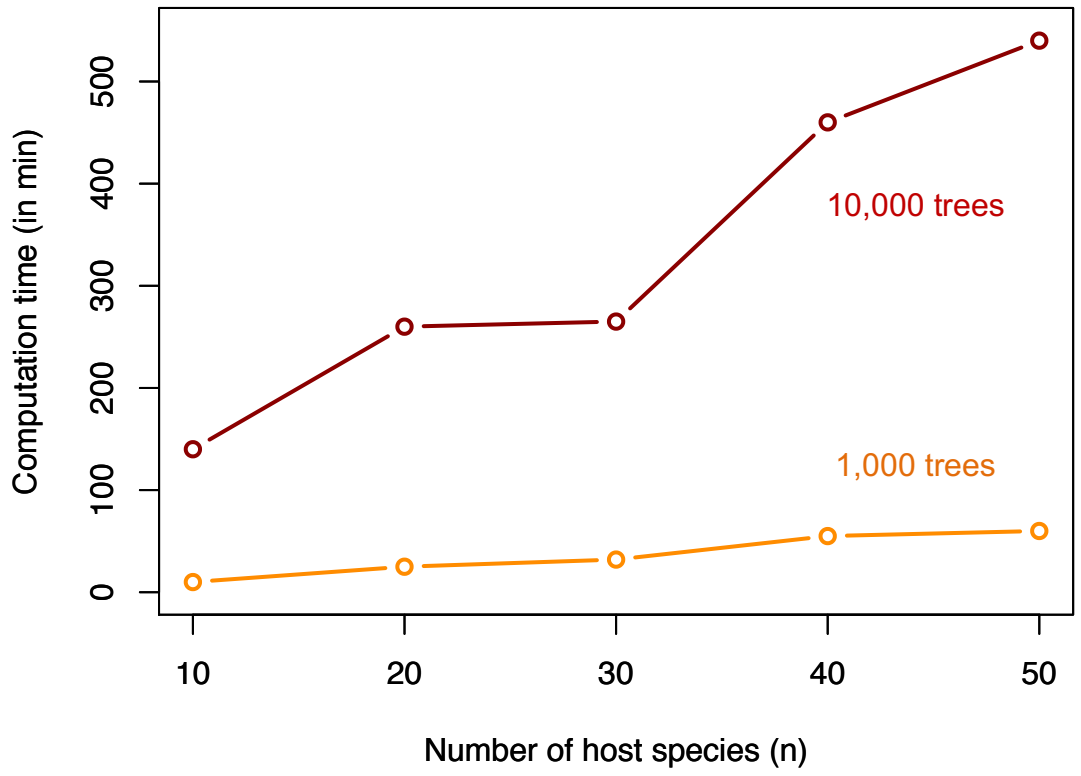
Problems faced when trying to delineate the "right" biological unit across all microbial groups: "over-splitting" (*right*) a biological unit using a similarity threshold that is too high, and "over-merging" (*left*) OTUs using a similarity threshold that is too low.



MOLECULAR ECOLOGY
RESOURCES

Supp. Figure S3: Computational time of HOME

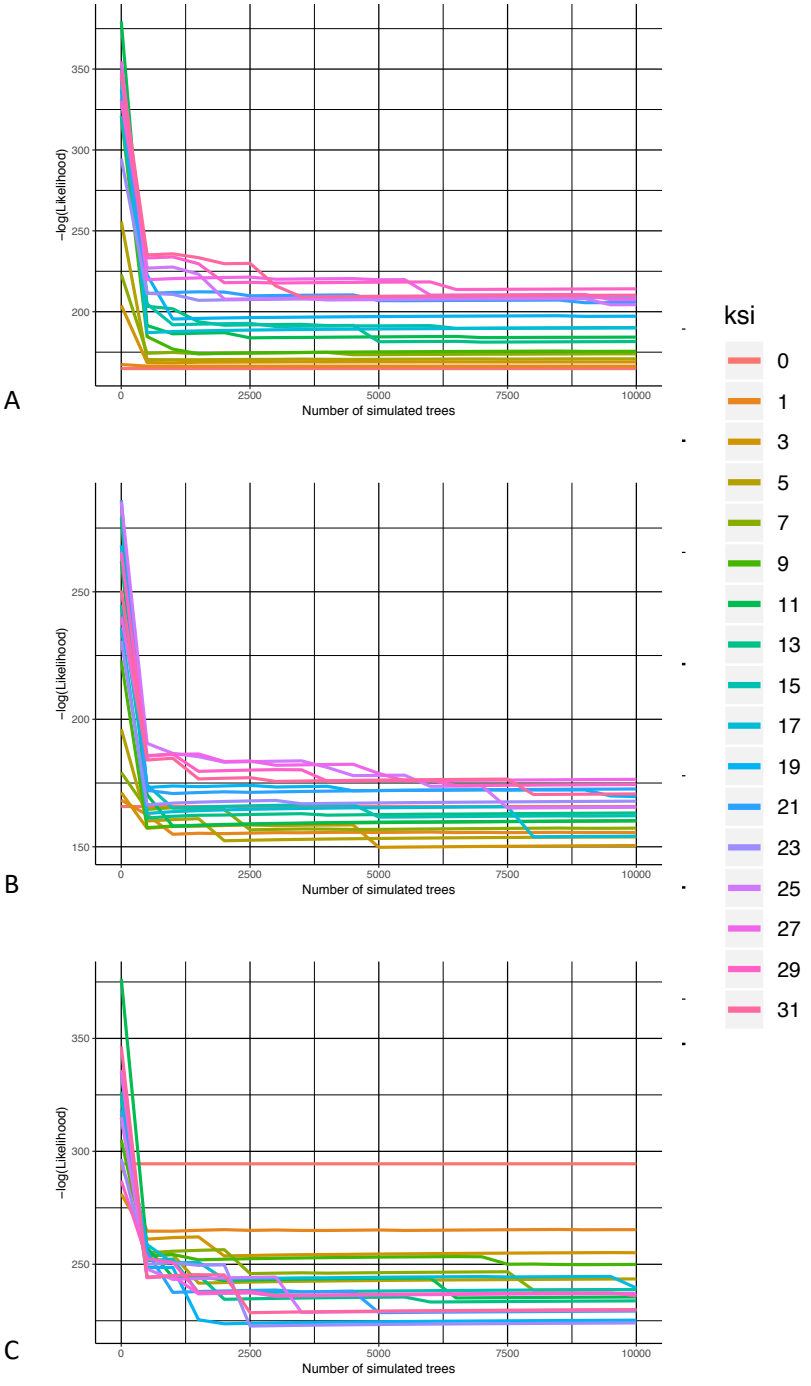
The computational time of HOME increases linearly with the number of hosts n (here n=10, 20, 30, 40, or 50 species) and the number of trees S used in the likelihood inference (here S=1,000 in orange, and S=10,000 trees in red). The computation time reported here for each n and S corresponds to the cumulative time required to fit HOME and to perform the tests of strict vertical transmission and independent evolution on 3 alignments, one simulated with strict vertical transmission, the other one vertical transmission with 5 host-switches, and the last environmental acquisition. Other parameters were similar to the parameters used in the section "*Testing our approach with simulations*". Simulations were performed on a multi-cores cluster, using 16 cores.



MOLECULAR ECOLOGY
RESOURCES

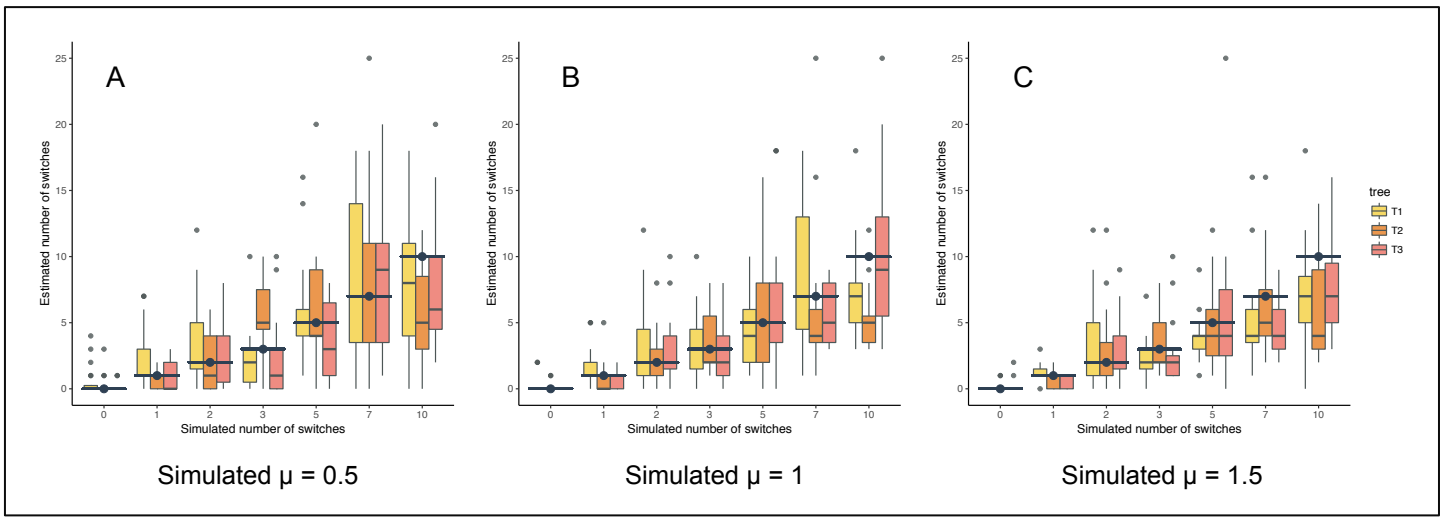
Supp. Figure S4: Rarefaction curves

Approximation of the minus log-likelihood of the model as a function of the number S of trees T_B used in the computation of the likelihood for scenarios of: A) strict vertical transmission, B) vertical transmission with 5 host-switches, and C) environmental acquisition. Likelihoods computed on alignments simulated on a 20 tips host tree (simulated under the Yule process), with parameters similar to those used in the section "*Testing our approach with simulations*". Each line corresponds to a given number of host switches (ξ). At a given S, the lower line corresponds to the most likely number of switches.

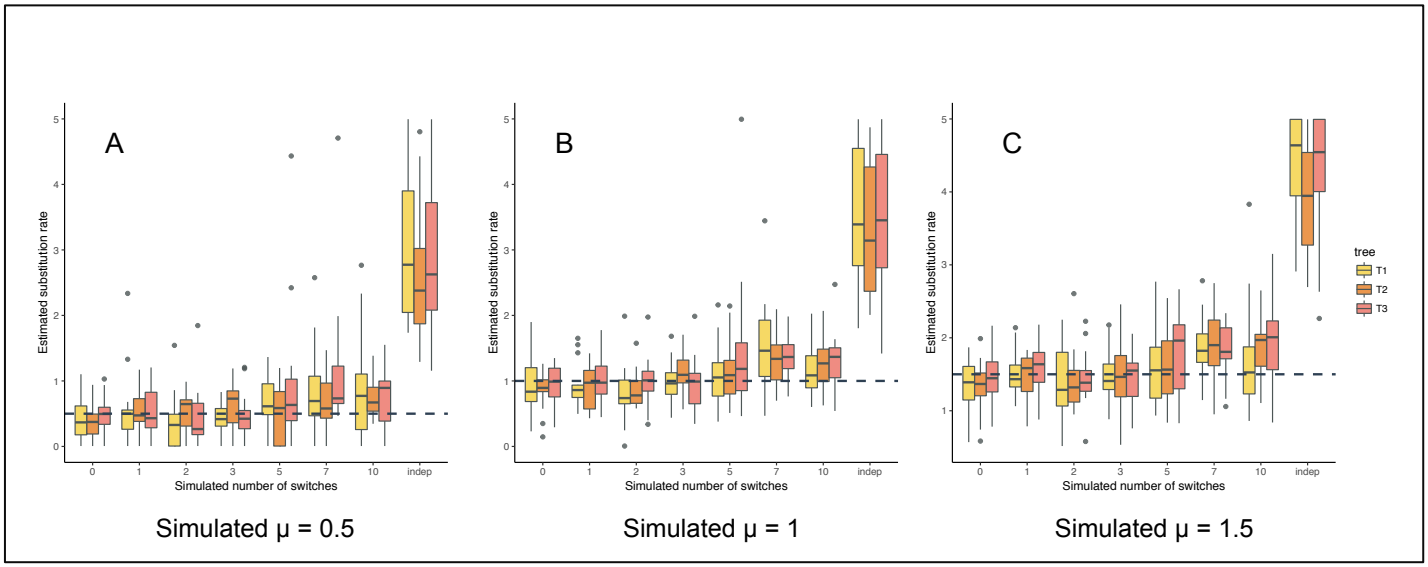


MOLECULAR ECOLOGY
RESOURCES

Supp. Figure S5: Estimated *versus* simulated number of switches ξ on the three different host trees T1, T2 and T3 (represented by three different colors) and for three different substitution rates. Blue ticks represent the simulated values.

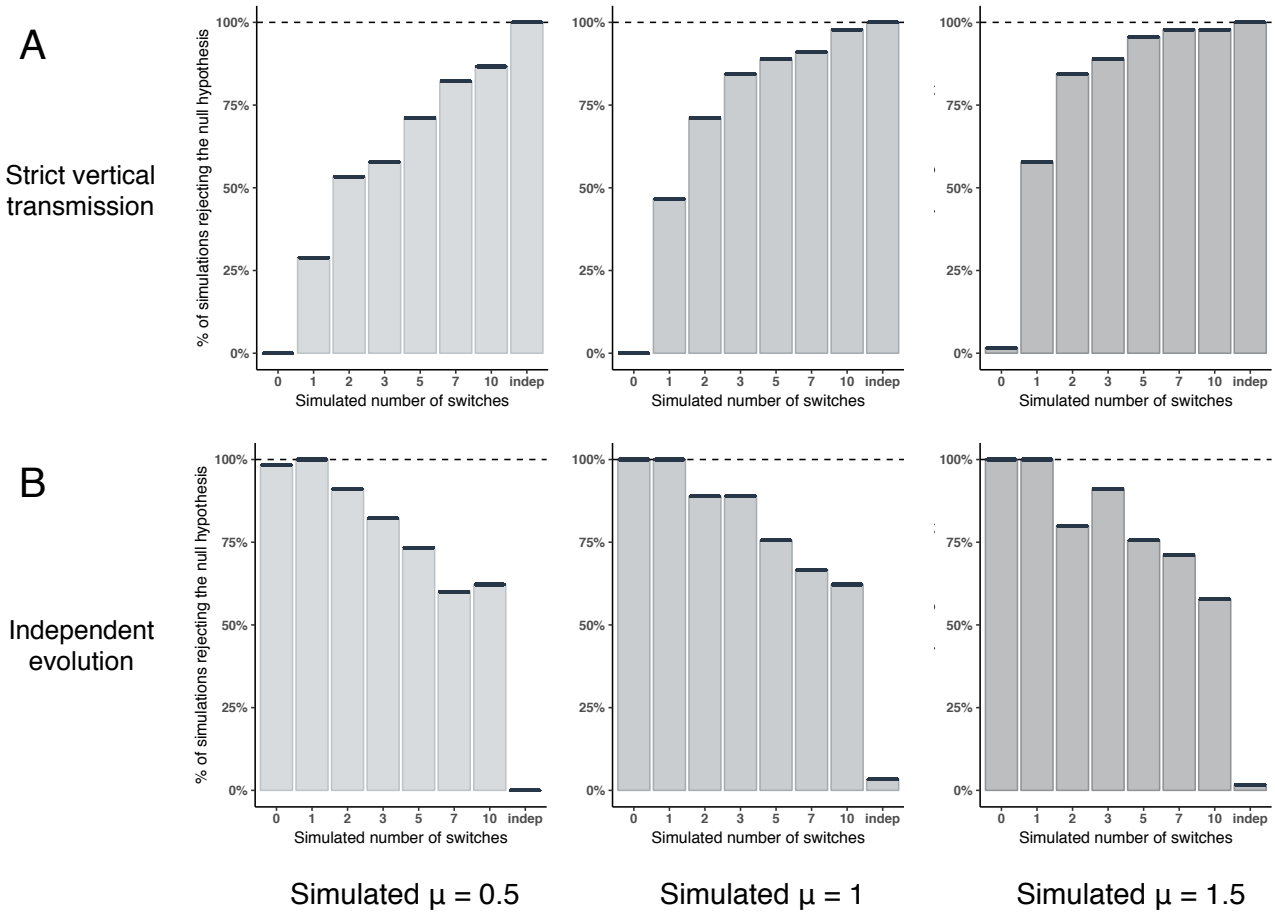


Supp. Figure S6: Estimated *versus* simulated substitution rate μ on the three different host trees T1, T2 and T3 (represented by three different colors). Blue dashed lines represent the simulated values.



MOLECULAR ECOLOGY
RESOURCES

Supp. Figure S7: Percentage of simulated alignments for which the null hypothesis of strict vertical transmission **(A)** or independent evolution **(B)** is rejected under various evolutionary scenarios (strict vertical transmission, vertical transmission with a given number of switches, and independent evolution) and different simulated substitution rates.



MOLECULAR ECOLOGY
RESOURCES

Supp. Figure S8: Effect of intraspecific variability on parameter estimation and model selection.

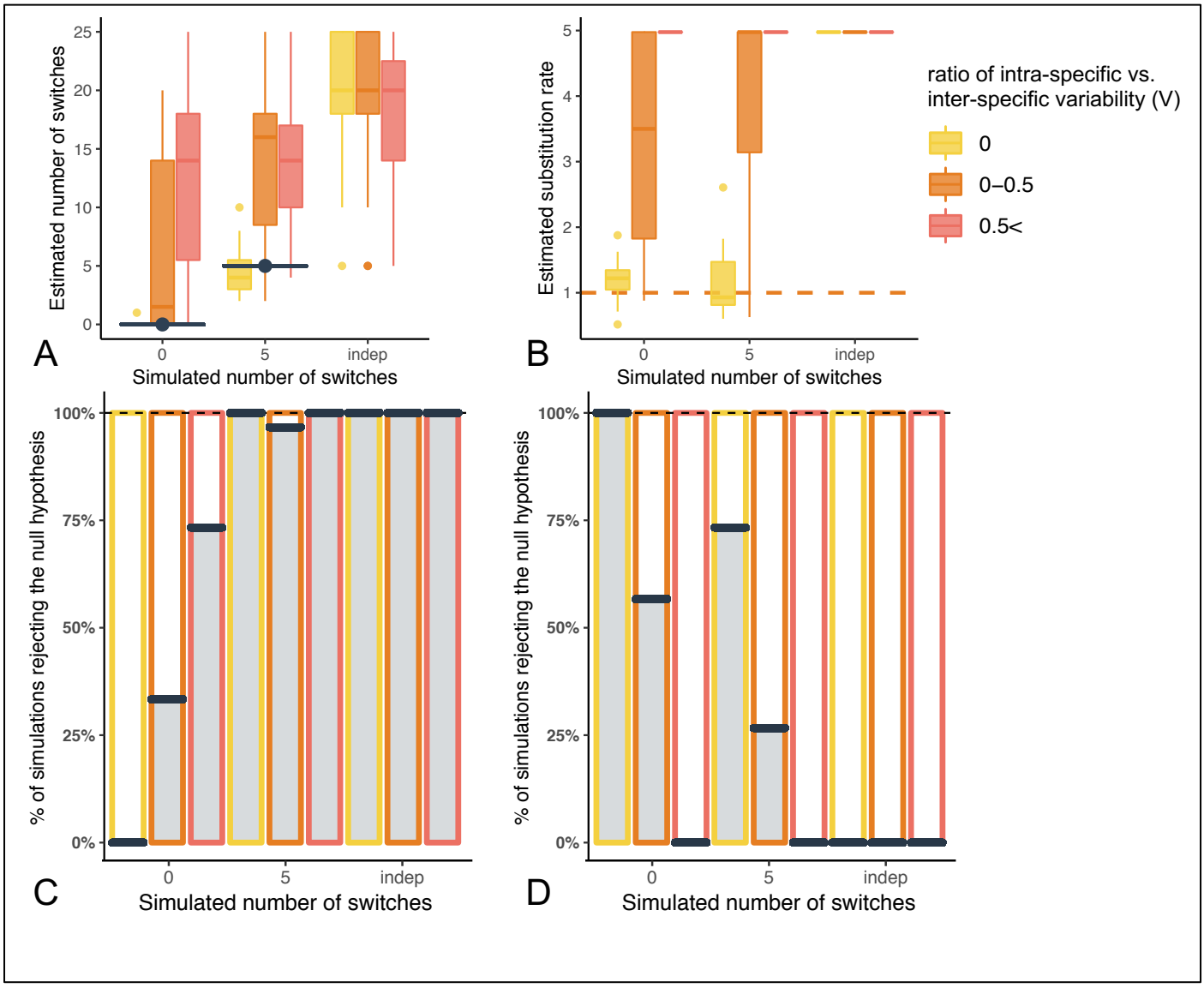
Alignments have been grouped according to their intraspecific variability.

A: Estimated *versus* simulated number of switches

B: Estimated *versus* simulated substitution rates

C: % of simulations rejecting the hypothesis of strict vertical transmission

D: % of simulations rejecting the hypothesis of environmental acquisition



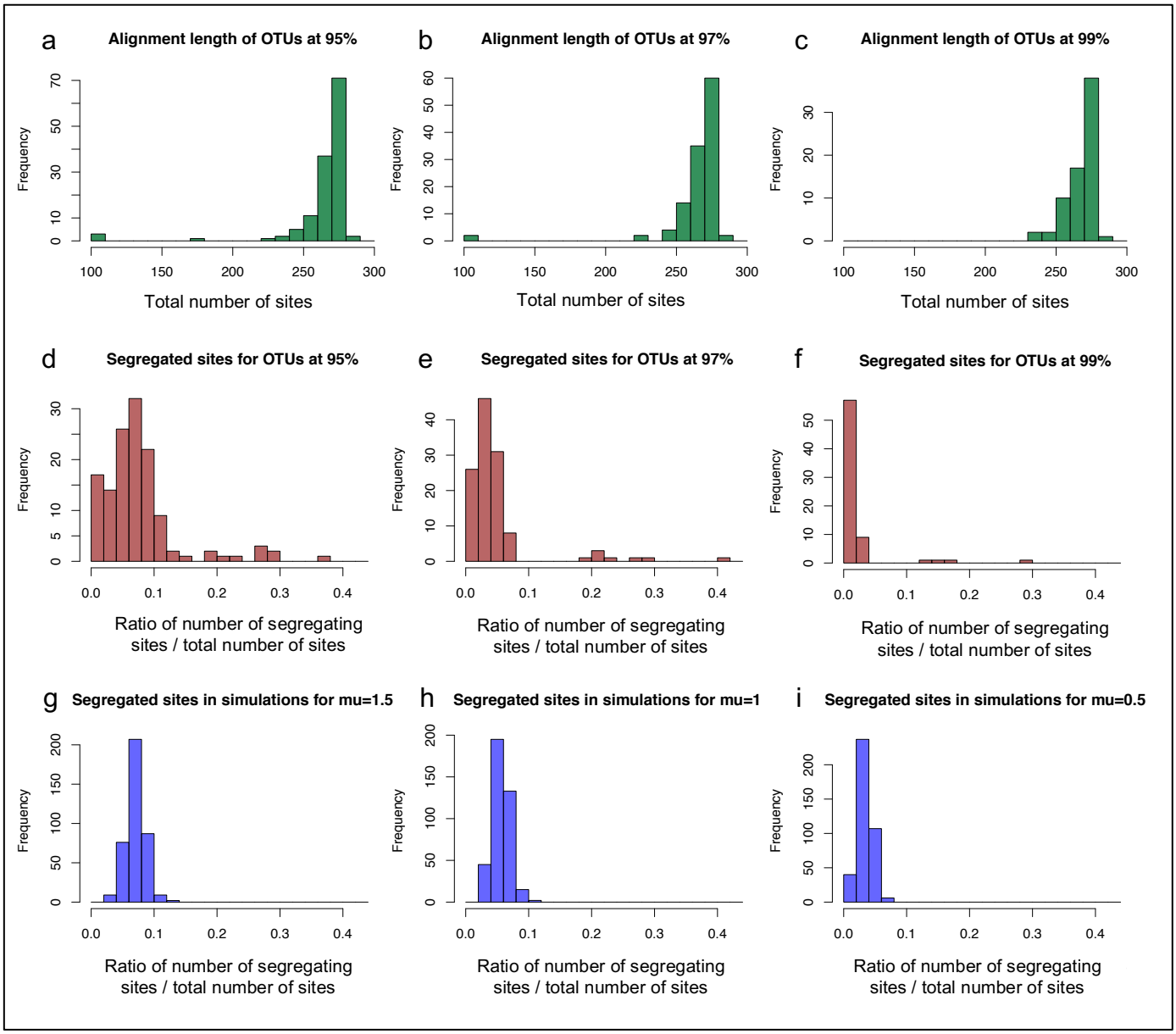
MOLECULAR ECOLOGY
RESOURCES

Supp. Figure S9: Characteristics of the empirical alignments.

a-c: distribution of the lengths of the empirical alignments for the different clustering thresholds

d-f: distribution of ratio of segregating sites of the empirical alignments for the different clustering threshold

g-i: distribution of ratio of segregating sites of the simulated alignments for the different simulated substitution rates



MOLECULAR ECOLOGY
RESOURCES

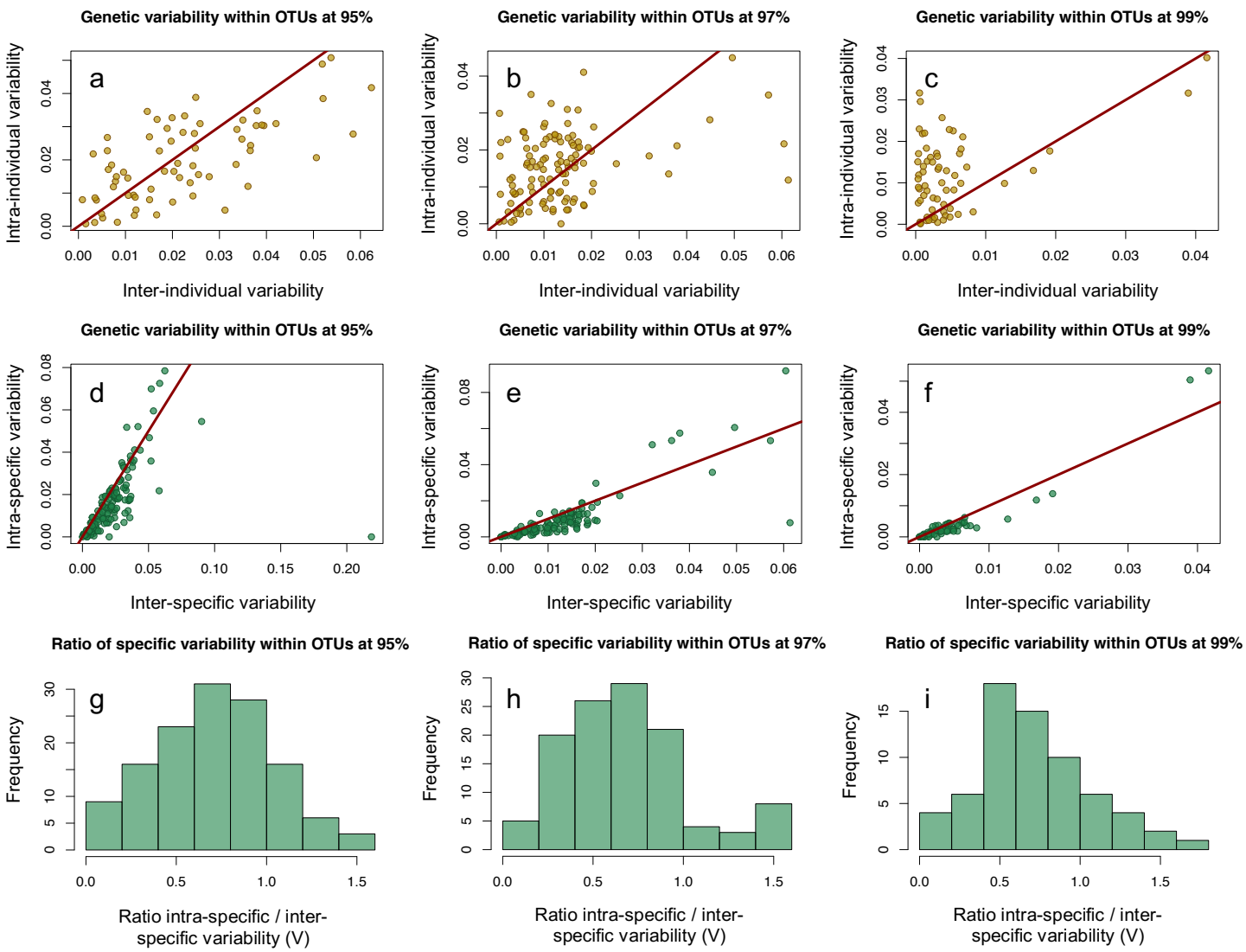
Supp. Figure S10: Genetic variability within OTUs

a-c: intra-individual variability versus inter-individual variability within each OTU for the different clustering thresholds. Intra-individual variability can be high compared to inter-individual variability (e.g. c), suggesting that it is driven by sequencing and PCR errors.

d-f: intra-specific variability versus inter-specific variability within each OTU for the different clustering thresholds. (NB: we only consider the most abundant sequence per individual)

g-i: histogram of the ratio of intra-specific versus inter-specific variability (V) within each OTU for the different clustering thresholds.

The red lines correspond to the first bisector. Every dot corresponds to one OTU.



MOLECULAR ECOLOGY
RESOURCES

Supplemental tables:

Supp. Table S1. Number of OTUs from the great apes microbiota for the different clustering thresholds. HOME can only be applied on the core variant OTUs. The two last rows indicate the total number of reads corresponding to the core OTUs (resp. variant core OTUs) for the different clustering thresholds (as a reference, the total number of raw reads is 1,292,542).

	Clustering threshold		
	95%	97%	99%
Total number of OTUs	1,074	1,793	4,935
Number of core OTUs (present in more than 75% individuals)	134	120	71
Number of core variant OTUs (at least one segregating site)	130	110	66
Total number of reads corresponding to core OTUs	749,605	611,193	241,666
Total number of reads corresponding to core variant OTUs	720,633	554,230	239,236

MOLECULAR ECOLOGY
RESOURCES

Supp. Table S2. Comparison of the characteristics of all the empirical alignments, the alignments inferred as corresponding to transmitted OTUs, and the simulated alignments (when applicable):

	All OTUs			Transmitted OTUs			Simulations		
	Threshold OTUs						Substitution rate		
	95%	97%	99%	95%	97%	99%	1.5	1	0.5
Average length of the alignments	263	264	267	265	264	262	300	300	300
Average number of segregating sites	20.1	12.8	5.6	18.2	17.6	3.5	21.9	17.4	10.7
Average ratio number segregating sites / alignment length	0.076	0.048	0.021	0.068	0.066	0.013	0.073	0.058	0.036
Average intra-individual variability	0.020	0.015	0.012	0.014	0.012	0.012	NA	NA	NA

MOLECULAR ECOLOGY
RESOURCES

Supp. Table S3. Taxonomic information and estimated parameters from the transmitted OTUs. "Threshold" stands for the percent similarity cut-off used for OTU delimitation. "Relative abundance" is the total number of sequences in the corresponding OTU divided by the total number of sequences in the study. "Number of non-overlapping reads" is the number of sequences in the OTU that do not occur in another OTU (at a different threshold).

OTU Name	Taxonomic family	Estimated μ	Estimated ξ	Strict vertical transmission	Threshold	Relative abundance	Total number of reads	Number of non-overlapping reads
OTU137396942	Alcaligenaceae	0.006	0	Not rejected	95	0.08%	1,089	6
OTU284019116	Alcaligenaceae	0.006	0	Not rejected	97	0.08%	1,083	0
OTU382421569	Coriobacteriaceae	1.400	4	Rejected	95	0.09%	1,111	1,111
OTU714176148	Coriobacteriaceae	0.006	0	Not rejected	97	0.15%	1,940	1,940
OTU910924283	Coriobacteriaceae	0.006	0	Not rejected	99	0.03%	339	339
OTU329886714	Desulfurococcaceae	0.006	1	Rejected	95	0.10%	1,279	431
OTU114691526	Desulfurococcaceae	0.066	1	Rejected	97	0.07%	848	0
OTU322547943	Eubacteriaceae	0.006	0	Not rejected	97	0.10%	1,300	1,300
OTU548957525	Lachnospiraceae	0.006	0	Not rejected	95	0.02%	297	5
OTU693717586	Lachnospiraceae	4.994	4	Rejected	95	0.04%	551	551
OTU777004095	Lachnospiraceae	0.006	0	Not rejected	95	0.06%	771	771
OTU516660135	Lachnospiraceae	0.006	0	Not rejected	97	0.02%	292	0
OTU657732334	Lachnospiraceae	0.066	0	Not rejected	97	0.22%	2,834	2,834
OTU908720582	Lachnospiraceae	0.006	0	Not rejected	97	0.10%	1,343	1,343
OTU234421667	Lachnospiraceae	0.006	0	Not rejected	99	0.08%	1,034	1,034
OTU469780863	Moraxellaceae	0.653	6	Not rejected	97	3.83%	49,508	49,508
OTU843396479	Paraprevotellaceae	0.006	0	Not rejected	95	0.17%	2,176	53
OTU347786903	Paraprevotellaceae	0.006	0	Not rejected	97	0.16%	2,123	0
OTU728699596	Paraprevotellaceae	0.006	0	Not rejected	99	0.03%	334	334
OTU113078451	Pelobacteraceae	0.006	0	Not rejected	95	0.06%	712	712
OTU257931929	Prevotellaceae	0.006	0	Not rejected	95	0.13%	1,665	1,665
OTU892624276	Prevotellaceae	3.175	9	Rejected	95	2.23%	28,843	28,843
OTU559296426	Rhodocyclaceae	0.006	0	Not rejected	95	0.05%	680	1
OTU735260590	Rhodocyclaceae	0.066	0	Not rejected	97	0.05%	679	0
OTU733943228	Ruminococcaceae	4.993	2	Rejected	97	0.18%	2,307	2,307
OTU704142964	Veillonellaceae	0.006	0	Not rejected	95	0.63%	8,092	21
OTU314436093	Veillonellaceae	0.006	0	Not rejected	97	0.62%	8,072	1
OTU465803492	Veillonellaceae	0.006	0	Not rejected	99	0.53%	6,881	0

MOLECULAR ECOLOGY
RESOURCES

Supp. Table S4: Taxonomic repartition of transmitted OTUs.

For each threshold, we give the number of transmitted OTUs divided by the total number of core OTUs corresponding to a given bacterial family. We also give the percentage of reads corresponding to transmitted OTUs.

Bacterial family	Ratio of transmitted OTUs over the total number of OTUs in the family and percentage of reads			
	95%	97%	99%	Total
Alcaligenaceae	1/2 (73%)	1/2 (73%)	0/1 (0%)	65%
Coriobacteriaceae	1/4 (4%)	1/8 (9%)	1/10 (5%)	6%
Desulfurococcaceae	1/1 (100%)	1/1 (100%)	0/0 (NA)	100%
Lachnospiraceae	3/6 (1%)	3/24 (4%)	1/33 (2%)	2%
Paraprevotellaceae	1/3 (14%)	1/2 (15%)	1/2 (3%)	11%
Pelobacteraceae	1/1 (100%)	0/0 (NA)	0/0 (NA)	100%
Prevotellaceae	2/4 (38%)	0/10 (0%)	0/11 (0%)	21%
Rhodocyclaceae	1/1 (100%)	1/1 (100%)	0/0 (NA)	100%
Veillonellaceae	1/5 (19%)	1/6 (22%)	1/4 (27%)	22%
Eubacteriaceae	0/0 (NA)	1/1 (100%)	0/0 (NA)	100%
Moraxellaceae	0/0 (NA)	1/2 (39%)	0/1 (0%)	23%
Ruminococcaceae	0/24 (0%)	1/21 (2%)	0/13 (0%)	1%

MOLECULAR ECOLOGY
RESOURCES

Supp. Table S5: Ratio of intra-specific versus inter-specific variability (V) within each OTUs of the great apes microbiota for the different clustering threshold. We computed (V) as the average of the ratio of specific variability for every (sub)-species among the 7 (sub)-species of great apes.

	All OTUs			Transmitted OTUs		
	Threshold OTUs					
	95%	97%	99%	95%	97%	99%
Median of ratio intra-/inter-specific variability (V)	0.71	0.64	0.68	0.37	0.36	0.47
Quantile 2.5% of ratio intra-/inter-specific variability (V)	0.01	0.13	0	0.05	0.05	0.34
Quantile 97.5% of ratio intra-/inter-specific variability (V)	1.32	1.53	1.49	1.08	1.34	0.58