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

Benoît Perez-lamarque, Hélène Morlon

Resource Article – Computer program

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

## Modeling host-microbiota evolution

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, 75005 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, 75005 Paris, France*

*Corresponding authors:* Benoît Perez-Lamarque (*benoit.perez@ens.fr);* Hélène Morlon *(morlon@biologie.ens.fr)*

## *Abstract* (250 words)

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 populations or species by host-switch (i.e. horizontal transmission), or transmitted across generations with a faithful association with their hosts (i.e. vertical transmission). 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 microbial data.

Here, we develop a model-based framework for quantifying the proportion of environmental acquisition, horizontal transmission, and vertical transmission during the evolution of host-associated 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 environmental acquisition. 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.

**Key words:** symbiont transmission, microbiota, molecular evolution, likelihood-based framework, holobiont, great apes

## *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, behavior, 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: i) 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) ii) horizontal transmission between unrelated host lineages (Henry et al., 2013), which can for example happen through direct interactions, via vectors or via shared habitats (Engel & Moran, 2013), and iii) 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 transmission and environmental acquisition 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 coevolution remains a major challenge. Patterns of "phylosymbiosis", i.e. a pattern of concordance between a given host phylogeny and the dendrogram reflecting the similarity of microbial communities across these hosts, is 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 coevolution 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 (de Vienne et al., 2013; Page & Charleston, 1998), while processes such as host-switches disrupt this congruence. Cophylogenetic tools either quantify the congruence between symbiont and host trees using distance-based methods (e.g. ParaFit (Legendre, Desdevises, & Bazin, 2002), 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. (Groussin et al., 2017) and Bailly-Bechet et al. (Bailly-Bechet et al., 2017) have used the ALE program (Szöllõsi, 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 a reconstruction of the symbionts' trees, which can be problematic for microbiota data typically generated using Next Generation Sequencing (NGS) metabarcoding techniques.

Here, we develop a probabilistic model of host-symbiont evolution, which aims at studying modes of inheritance in the microbiota; our framework does not require building a symbiont phylogeny and allows model comparison. Huelsenbeck et al. (Huelsenbeck et al., 2000) developed a similar approach, focused on host-parasite associations, with a model of cospeciation and host-switches. However, the authors

developed an inference framework associated with the joint reconstruction of both host and parasite phylogenetic trees, 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. 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 strict environmental acquisition and strict vertical transmission. We test our approach using simulations and apply it to gut microbiota high-throughput sequencing data of the family Hominidae.

## *Materials & Methods*

### *HOME: A general framework for studying Host-Microbiota Evolution*

*From metabarcoding microbiota data to independent alignments*

Given a host species tree and metabarcoding microbiota data (e.g. rRNA 16S sequences) sampled from each host species, 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 (Fig. 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 consists of a series of microbial alignments $A$, each corresponding to a specific OTU; a given alignment is composed of N-nucleotide long sequences (with potential gaps corresponding to missing data), each corresponding to a specific host. In each alignment, we 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 independently to each alignment.

*Modeling the evolution of an OTU on a host phylogeny*

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

(i) Vertical transmission: From an ancestral microbial population at the root of the host phylogeny represented by a N-nucleotide 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 nucleotide evolves independently from the others according to a substitution model with a rate μ that is supposed to be the same for all variable nucleotides 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.

(ii) Host-switches: A discrete number (ξ) of host-switches happens 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 (Fig. 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. Fig. 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.

*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:

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

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

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

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

We compute the probability $L(A_S|\mu,T_B)$ of the sequence alignment $A_S$ on a given 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 nucleotide position 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 $P(0) = (1_A, 1_C, 1_G, 1_T)$ 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 nucleotide ν. We have:

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

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. We iterate this pruning calculation from the leaves to the root of the tree, and obtain the probability of the alignment at the nucleotide position 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 $(1 - e^{-\mu B})$. Finally, the probability of the alignment $A_S$ is obtained by multiplying the probabilities corresponding to each nucleotide. Hence the probability of the variable alignment $A_S$ is given by:

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

where $N_S$ is the number of segregating nucleotides.

In practice, we used $S = 10^4$ and plotted the resulting value of $L(A_S|\mu, \xi)$ 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 $L(A_S|\mu, \xi)$. 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. 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.

*Simulations of host-switches: from* T *to* $T_B$

Each switch is characterized by its "donor" branch, by its position on the branch, and by its "receiving" branch. 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 (Fig. 2):

(i) 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).

(ii) the switch occurs from an internal branch to a branch directly related to the root, i.e. 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 ancestor to all the current microbial sequences. This switch changes both the topology of the tree and the branch lengths.

(iii) the switch occurs between 2 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.

(iv) the switch occurs between 2 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 $\xi$ switches on the host tree and use these same trees in our different analyses.

*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 frequent horizontal transmissions with high substitution rates instead of environmental transmission).

In order to test support for a scenario with horizontal host-switches *versus* strict vertical transmission, we compute $L_0 = \mathrm{L}(A_S|\hat{\mu}, \mathrm{T})$, 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 $L_1 = \mathrm{L}\left(A_S\middle|\hat{\mu}, \hat{\xi}\right)$ corresponding to the best scenario with horizontal transmission, using a likelihood ratio test. In order to test support for a scenario 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 analyze the rank of $\hat{\xi}$ and $\hat{\mu}$ estimated from the original alignment in the distribution of $\xi_R$ and $\mu_R$ 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 $\xi$ or the substitution rate $\mu$ are ranked within the distribution of $\xi_R$ and $\mu_R$, we consider that a scenario of independent evolution cannot be rejected.

*Detecting transmitted OTUs*

Based on the analyses above and our definition of modes of inheritance, we sort the OTUs into two different categories: the vertically and/or horizontally transmitted OTUs called *transmitted* OTUs (those that reject the hypothesis of independent evolution), and the environmentally acquired OTUs called *independent* OTUs (those that do not reject the hypothesis of independent evolution). In practice, there is no universal similarity threshold that will provide the "right" biological unit delineation across all microbial groups (Sanders et al., 2014) (Fig. S1). "Over-splitting" a biological unit using a similarity threshold that is too high for that biological unit will reduce statistical signal (each sub-

unit will be represented in fewer hosts) and will miss host-switches between sub-units (given that sub-units will be analyzed independently). “Over-merging” OTUs using a similarity threshold that is too low will tend to blur a signal of transmission, and will over-estimate mutation 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.

*Implementation*

All the scripts of our model are written in R (R Core Team 2018), 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). We also used the packages parallel, expm, ggplot2, reshape2 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.

***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 (Fig. S6 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 ($\xi=0$), scenarios with host-switches $\xi=[1, 2, 3, 5, 7, 10]$, and a scenario of environmental acquisition; each of these scenarios were 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 $T_B$ 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 with a proportion of variable nucleotides x=0.1, using our own codes. 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). 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 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 environmental acquisition.

***Empirical application: great apes microbiota***

We illustrate our approach using data from Ochman et al. (Ochman et al., 2010); this paper is one of the first paper testing hypotheses about co-diversification in the well-studied clade of great apes (using phylosymbiotic patterns), and the associated data has been used in other papers aimed at studying codiversification (Sanders et al., 2014). The dataset consists of fecal samples collected from 26 wild-living hominids, including eastern and western African gorillas (2 individuals of *G. gorilla* and 2 individuals of *G. beringei*), bonobos (6 individuals of *P. paniscus*), and three subspecies of chimpanzees (5 individuals of *P. t. schweinfurthii,* 7 individuals of *P. t. troglodytes* and 2 individuals of *P. t. ellioti*), as well as two humans from Africa and America (*H. sapiens*).

Ochman et al. (Ochman et al., 2010) extracted DNA from the fecal 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 resolution than was possible at the time of the Ochman paper, but not necessarily for entire clades. These data provide a good illustration of our approach.

We obtained the reads from Dryad (http://datadryad.org/resource/doi: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:

(i) Dereplication: we discarded all the singletons and sorted the sequences by abundance using USEARCH commands derep_fulllength and sortbysize

(ii) 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 an OTU member sequence and the representative sequence of that OTU), 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).

(iii) 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.

(iv) 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).

(v) 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.

(vi) Making intra-OTU alignments: discarding few OTUs that had unvaried alignments, we obtained 130 OTUs at 95%, 110 OTUs at 97%, and 66 OTUs at 99% similarity thresholds (Table S1). For each OTU, we built the bacterial alignment by selecting for each host individual the most abundant sequence among all the reads mapped to that OTU. We considered that the microbial genetic variability within each host individual (hereafter referred to as "intra-individual variability") is mainly due to PCR and sequencing artefacts, so we neglected it (Fig. S7).

Finally, we applied our approach to each core OTU independently, 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 sub-species level tree in order to increase tree size (there are only 7 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 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.

***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

448 a single representative microbial sequence per host species), it can be run on an
449 individual-level host tree, with arbitrarily small intra-specific branch-lengths. Because
450 this usage of HOME is slightly different from the case envisioned in our description of
451 the approach, we tested its behavior. We simulated the evolution of microbial
452 alignments on the great apes sub-species tree with a range of intraspecific variability
453 similar to the range observed in the great apes alignments. For each OTU alignment, we
454 defined intraspecific variability (V) as the mean nucleotidic diversity within host
455 subspecies (computed using Nei's estimator (Ferretti, Raineri, & Ramos-Onsins, 2012))
456 divided by the total nucleotidic diversity computed on the entire alignment. We
457 simulated a total of 180 alignments according to 3 scenarios: strict vertical transmission
458 ($\xi$=0), transmission with 5 host-switches ($\xi$=5), and environmental acquisition. For
459 every scenario, we simulated intraspecific variability by extending the stochastic
460 process generating nucleotidic substitution on every sequence for a time range that
461 allowed to obtain levels of intraspecific variability that corresponded to the empirical
462 level of intraspecific variability. We ran HOME on each of these simulated alignments
463 and evaluated its performance, in terms of parameter estimation and model selection,
464 when there was no intraspecific variability (V=0), low and intermediate intraspecific
465 variability ($0<V<0.5$), and high intraspecific variability ($V>0.5$).
466
467
468

## *Results*

### *Performance of HOME*

Testing the performance of HOME using intensive simulations, we find a reasonable ability to recover simulated parameter values (Fig. 3). Estimates of the number of switches $\hat{\xi}$ are highly correlated with simulated values ξ, although the approach tends to overestimate the true number of switches when there are very few (less than 2) and to underestimate this number when there are many (Fig. 3A). The linear regression confirms these results $\hat{\xi}$ = 2.15 ($F_{dl=606}$=1015, p-value <0.0001) + ξ * 0.58 ($F_{dl=606}$=141, p-value <0.0001). The ability to recover the true number of switches does not depend on the simulated substitution rate ($F_{dl=606}$=0.2601, p-value=0.61; Fig. S2). The substitution rate is rather well estimated (Fig. 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<0.0001; Fig. 3B). The simulated transition/transversion rate ratio κ is well estimated (median ± s.d. =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=0.0007). For alignments simulated independently from the host tree, the approach estimates a high number of switches (median ± s.d. = 16 ± 6.2, Fig. 3A), and highly overestimates the substitution rate (Fig. 3B). The type of host tree (T1, T2 or T3) has little impact on the estimation of ξ (it explains less than 3% of the total variance, Fig. S2), μ (around 10%, Fig. S3) and κ (less than 0.01%).

Our model selection procedure has very low type I error rates, and type II error rates that depend on the situation (Fig. 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 (Fig. 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%, Fig. 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 (Fig. 4A). In the

502 case of the test of environmental acquisition, the power ranges from 100% for ξ=1 to
503 60% for ξ=10, and it would decrease further with more switches (Fig. 4B). In both cases,
504 the power increases when the substitution rate μ is larger (Fig. S4).
505
506 When HOME is applied to an individual-level host tree in order to account for
507 intraspecific microbial genetic variability, Type I error rates associated to the test of
508 environmental acquisition remain very low regardless of the magnitude of the
509 variability (Fig. S5). The confidence in the estimation of the parameters (ξ and μ)
510 remains good for low values of intraspecific variability (V<0.5), but decreases with
511 increasing variability (V>0.5). The type I error rate associated to the test of strict vertical
512 transmission increases with increasing variability, and the power of the two tests
513 decreases with increasing variability.
514
515 ***Modes of inheritance in the great apes microbiota***
516
517 Applying HOME to great apes gut microbiota data, we found that among the core OTUs
518 with at least one segregating site, approximately 9 in 10 OTUs are environmentally
519 acquired while 1 in 10 is transmitted (Fig. 5A); more specifically, the ratios of
520 transmitted OTUs (and strictly vertically transmitted OTUs) were the following:
521 12(8)/130 at 95%, 12(10)/110 at 97%, and 4(4)/66 at 99%. In terms of relative
522 abundance, 108,206 unique sequences in 1,292,542 (8.4%) belonged to transmitted
523 OTUs (and 1,184,336 sequences, 91.6%, to strictly vertically transmitted ones, Table
524 S3). Almost half of these sequences (49,508) were from an *Acinetobacter* bacterium
525 (Moraxellaceae family); another important pool of these sequences was from the family
526 Prevotellaceae (28,843 reads). In total, 12 bacterial families (in 27) contained OTUs that
527 were transmitted, including Veillonellaceae, Lachnospiraceae, Ruminococcaceae and
528 Paraprevotellaceae (Fig. 5B, Table S4). Some of these families (e.g. Desulfurococcaceae,
529 Pelobacteraceae, Rhodocyclaceae and Eubacteriaceae) were entirely made of a
530 transmitted OTU, while others also had many OTUs and/or sequences that were not
531 transmitted (e.g. Ruminococcaceae, Lachnospiraceae and Coriobacteriaceae).
532
533 The sequence length and proportion of segregating sites of OTUs inferred as transmitted
534 were similar to those of other OTUs (Fig. S6 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 tend to be smaller than that of other OTUs (Table S5 and Fig. S7), which is consistent with our simulation results showing that the power to detect vertical transmission decreases with increasing intraspecific variability.

## *Discussion*

We developed 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 our model to great apes microbiota data, we identified a set of transmitted gut bacteria that account on average for 8.4% of the total 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 faithful vertical transmission for low ξ values to horizontal transmission and 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 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.

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 (Louca et al., 2016), and this also prevents incorporating in our model horizontal host-switches without replacement (i.e. the persistence of both ancestral and newly-acquired symbionts in a lineage). 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 nucleotide -- 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).

There are several other developments that would significantly improve the approach. For example, accounting for extinction and missing species in the host phylogeny would provide a better representation of past host-switches. Also, rather than considering each OTU as an independently 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. Finally, incorporating dynamics of extinctions and recolonizations of a symbiont across host clades would extend the time scale of application of the approach to hundreds of millions of years (Shapira, 2016). Indeed, while ignoring such dynamics is reasonable for studying microbial evolution at small evolutionary scales such as within great apes (Ochman et al., 2010), it would not be reasonable at larger evolutionary timescales such as across invertebrate or vertebrate species (Brooks, Kohl, Brucker, van Opstal, & Bordenstein, 2016).

In the great apes gut microbiota, 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. 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 et al. (2014). Still, the major part of the microbiota is constituted of bacteria that are environmentally acquired and therefore evolving independently from the great apes phylogeny (Moeller et al., 2013). We found transmitted OTUs in 12 microbial families, including Lachnospiracea, Coriobacteriaceae, Paraprevotellaceae, Rhodocyclaceae, and Alcaligenaceae. This illustrates the utility of our approach, which offers the advantage of investigating the whole microbiota without an *a priori* on which families might be transmitted; 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 amplified 3 primer-specific families (Bacteroidaceae, Bifidobacteriaceae, and

638 Lachnospiracea) and showed that phylogenies representing the Bifidobacteriaceae and
639 Bacteroidaceae were congruent with the apes phylogeny, suggesting that co-
640 diversification occurred in these two families. Unfortunately, neither Bifidobacteriaceae
641 nor Bacteroidaceae were represented in the core OTUs in Ochman et al.'s data, even with
642 a 95% similarity threshold: those bacteria were either not sampled, badly processed
643 during DNA extraction and PCR, wrongly taxonomically annotated, or too divergent to
644 be merged into core OTUs defined at 95%. Conversely, while Moeller et al. did not find
645 any signal of co-phylogeny in the Lachnospiraceae family, we found 3 transmitted OTUs
646 belonging to this family. However, they investigated the phylogenetic relationships
647 between all the strains of Lachnospiraceae and whether they match the phylogenetic
648 tree of great apes. This illustrates the utility of our approach, which investigates
649 transmission modes of separate OTUs within bacterial families, rather than considering
650 in a single evolutionary framework all the sequences from the same family.
651
652 Among the families in which we found transmitted OTUs, some are well known for
653 having mutualistic properties. For example, the Lachnospiraceae, Paraprevotellaceae
654 and Rhodocyclales families are involved in breaking down complex carbohydrates in the
655 gut; they have even evolved to a fibrolytic specialization in gut communities (Biddle,
656 Stewart, Blanchard, & Leschine, 2013). These vertically transmitted fibrolytic bacteria,
657 which have been co-evolving for millions of years with the great apes, may be one of
658 factors that allowed frequent and rapid dietary shifts during the evolutionary history of
659 hominids (Hardy, Brand-Miller, Brown, Thomas, & Copeland, 2015; Head, Boesch,
660 Makaga, & Robbins, 2011; Muegge et al., 2011). However, why these particular bacteria
661 are faithfully vertically transmitted while other digesting gut bacteria seem largely
662 environmentally acquired remains unclear.
663
664 Microbiota data is being collected across multiple hosts at an unprecedented scale. Our
665 approach allows identifying, among numerous microbial units most of which are
666 environmentally acquired, those that are vertically transmitted and potentially
667 coevolving with their hosts. The current implementation of our model is entirely
668 adapted to applications to other datasets using different sequencing techniques,
669 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.

## *Acknowledgments*

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.

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.

## *References*

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. doi: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). doi: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. doi: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. doi:10.1371/journal.pbio.1002226

Bright, M., & Bulgheresi, S. (2010). A complex journey : transmission of microbial symbionts. *Nat Rev Microbiol.*, *8*(3), 218–230. doi: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. doi:10.1371/journal.pbio.2000225

706 Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., & Holmes, S. P.
707 (2016). DADA2: High-resolution sample inference from Illumina amplicon data.
708 *Nature Methods*, *13*(7), 581–583. doi:10.1038/nmeth.3869
709 Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., …
710 Knight, R. (2010). QIIME allows analysis of high- throughput community
711 sequencing data. *Nature Methods*, *7*(5), 335–336. doi:10.1038/nmeth0510-335
712 Conow, C., Fielder, D., Ovadia, Y., & Libeskind-Hadas, R. (2010). Jane: A new tool for the
713 cophylogeny reconstruction problem. *Algorithms for Molecular Biology*, *5*(1), 1–10.
714 doi:10.1186/1748-7188-5-16
715 de Vienne, D. M., Refrégier, G., López-Villavicencio, M., Tellier, A., Hood, M. E., & Giraud, T.
716 (2013). Coespeciation vs hos-shift speciation: methodsfor testing, evidence from
717 natural associations and ralation to coevolution. *New Phytologist*, *198*(2), 347–385.
718 Edgar, R. C. (2013). UPARSE: Highly accurate OTU sequences from microbial amplicon
719 reads. *Nature Methods*, *10*(10), 996–998. doi:10.1038/nmeth.2604
720 Engel, P., & Moran, N. A. (2013). The gut microbiota of insects - diversity in structure and
721 function. *FEMS Microbiology Reviews*, *37*(5), 699–735. doi:10.1111/1574-
722 6976.12025
723 Felsenstein, J. (1981). Evolutionary trees from DNA sequences: A maximum likelihood
724 approach. *Journal of Molecular Evolution*, *17*(6), 368–376.
725 doi:10.1007/BF01734359
726 Ferretti, L., Raineri, E., & Ramos-Onsins, S. (2012). Neutrality tests for sequences with
727 missing data. doi:10.1534/genetics.112.139949
728 Groussin, M., Mazel, F., Sanders, J. G., Smillie, C. S., Lavergne, S., Thuiller, W., & Alm, E. J.
729 (2017). Unraveling the processes shaping mammalian gut microbiomes over
730 evolutionary time. *Nature Communications*, *8*. doi:10.1038/ncomms14319
731 Hacquard, S., Garrido-Oter, R., González, A., Spaepen, S., Ackermann, G., Lebeis, S., …
732 Schulze-Lefert, P. (2015). Microbiota and host nutrition across plant and animal
733 kingdoms. *Cell Host and Microbe*, *17*(5), 603–616. doi:10.1016/j.chom.2015.04.009
734 Hardy, K., Brand-Miller, J., Brown, K. D., Thomas, M. G., & Copeland, L. (2015). The
735 importance of dietary carbohydrate in Human evolution. *The Quarterly Review of*
736 *Biology*, *90*(3), 251–268. doi:10.1086/682587
737 Head, J. S., Boesch, C., Makaga, L., & Robbins, M. M. (2011). Sympatric chimpanzees (Pan
738 troglodytes troglodytes) and gorillas (Gorilla gorilla gorilla) in Loango National

739 Park, Gabon: dietary composition, seasonality, and intersite comparisons.
740 *International Journal of Primatology*, *32*(3), 755–775. doi:10.1007/s10764-011-
741 9499-6
742 Henry, L. M., Peccoud, J., Simon, J.-C., Hadfield, J. D., Maiden, M. J. C., Ferrari, J., & Godfray,
743 H. C. J. (2013). Horizontally transmitted symbionts and host colonization of
744 ecological niches. *Current Biology*, *23*(17), 1713–1717.
745 doi:10.1016/J.CUB.2013.07.029
746 Huelsenbeck, J. P., Rannala, B., & Larget, B. (2000). A Bayesian framework for the
747 analysis of cospeciation. *Evolution; International Journal of Organic Evolution*, *54*(2),
748 352–64. doi:10.1111/j.0014-3820.2000.tb00039.x
749 Koeppel, A. F., & Wu, M. (2014). Species matter: the role of competition in the assembly
750 of congeneric bacteria. *The ISME Journal*, *8*, 531–540. doi:10.1038/ismej.2013.180
751 Legendre, P., Desdevises, Y., & Bazin, E. (2002). A statistical test for host–parasite
752 coevolution. *Systematic Biology*, *51*(2), 217–234.
753 doi:10.1080/10635150252899734
754 Linz, B., Balloux, F., Moodley, Y., Manica, A., Liu, H., Roumagnac, P., … Achtman, M. (2007).
755 An African origin for the intimate association between humans and Helicobacter
756 pylori. *Nature*, *445*(7130), 915–918. doi:10.1038/nature05562
757 Louca, S., Jacques, S. M. S., Pires, A. P. F., Leal, J. S., Srivastava, D. S., Parfrey, L. W., …
758 Doebeli, M. (2016). High taxonomic variability despite stable functional structure
759 across microbial communities. *Nature Ecology & Evolution*, *1*(1), 0015.
760 doi:10.1038/s41559-016-0015
761 McFall-Ngai, M., Hadfield, M. G., Bosch, T. C. G., Carey, H. V., Domazet-Lošo, T., Douglas, A.
762 E., … Wernegreen, J. J. (2013). Animals in a bacterial world, a new imperative for the
763 life sciences. *Proceedings of the National Academy of Sciences*, *110*(9), 3229–3236.
764 doi:10.1073/pnas.1218525110
765 McKenney, E. A., Maslanka, M., Rodrigo, A., & Yoder, A. D. (2018). Bamboo specialists
766 from two mammalian orders (Primates, Carnivora) share a high number of low-
767 abundance gut microbes. *Microbial Ecology*, *76*(1), 272–284. doi:10.1007/s00248-
768 017-1114-8
769 Moeller, A. H., Caro-Quintero, A., Mjungu, D., Georgiev, A. V, Lonsdorf, E. V, Muller, M. N.,
770 … Ochman, H. (2016). Cospeciation of gut microbiota with hominids. *Science (New*
771 *York, N.Y.)*, *353*(6297), 380–2. doi:10.1126/science.aaf3951

772 Moeller, A. H., Peeters, M., Ndjango, J. B., Li, Y., Hahn, B. H., & Ochman, H. (2013).
773 Sympatric chimpanzees and gorillas harbor convergent gut microbial communities.
774 *Genome Research*. doi:10.1101/gr.154773.113
775 Morlon, H., Lewitus, E., Condamine, F. L., Manceau, M., Clavel, J., & Drury, J. (2015).
776 RPANDA: an R package for macroevolutionary analyses on phylogenetic trees.
777 *Methods in Ecology and Evolution*. doi:10.1111/2041-210X.12526
778 Morris, J. J., Lenski, R. E., & Zinser, E. R. (2012). The Black Queen Hypothesis: Evolution
779 of dependencies through adaptive gene loss. doi:10.1128/mBio.00036-12
780 Muegge, B. D., Kuczynski, J., Knights, D., Clemente, J. C., Fontana, L., Henrissat, B., …
781 Gordon, J. I. (2011). Diet drives convergence in gut microbiome functions across
782 mammalian phylogeny and within humans. *Science*, *332*(6032), 970–974.
783 doi:10.1126/science.1198719.Diet
784 Ochman, H., Worobey, M., Kuo, C. H., Ndjango, J. B. N., Peeters, M., Hahn, B. H., &
785 Hugenholtz, P. (2010). Evolutionary relationships of wild hominids recapitulated by
786 gut microbial communities. *PLoS Biology*, *8*(11), 3–10.
787 doi:10.1371/journal.pbio.1000546
788 Page, R. D. M., & Charleston, M. A. (1998). Trees within trees: Phylogeny and historical
789 associations. *Trends in Ecology and Evolution*, *13*(9), 356–359. doi:10.1016/S0169-
790 5347(98)01438-4
791 Paradis, E., Claude, J., & Strimmer, K. (2004). APE: Analyses of phylogenetics and
792 evolution in R language. *Bioinformatics*, *20*(2), 289–290.
793 doi:10.1093/bioinformatics/btg412
794 Philippot, L., Raaijmakers, J. M., Lemanceau, P., & van der Putten, W. H. (2013). Going
795 back to the roots: the microbial ecology of the rhizosphere. *Nature Reviews*
796 *Microbiology*, *11*(11), 789–799. doi:10.1038/nrmicro3109
797 Pylro, V. S., Roesch, L. F. W., Ortega, J. M., do Amaral, A. M., Tótola, M. R., Hirsch, P. R., …
798 Brazilian Microbiome Project Organization Committee. (2014). Brazilian
799 Microbiome Project: Revealing the unexplored microbial diversity - challenges and
800 prospects. *Microbial Ecology*, *67*(2), 237–241. doi:10.1007/s00248-013-0302-4
801 Revell, L. J. (2012). phytools: An R package for phylogenetic comparative biology (and
802 other things). *Methods in Ecology and Evolution*, *3*(2), 217–223. doi:10.1111/j.2041-
803 210X.2011.00169.x
804 Rosenberg, E., & Zilber-Rosenberg, I. (2016). Microbes Drive Evolution of Animals and

Plants: the Hologenome Concept. doi:10.1128/mBio.01395-15

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. doi:10.1111/mec.12611

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

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

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

Szöllõsi, G. J., Rosikiewicz, W., Boussau, B., Tannier, E., & Daubin, V. (2013). Efficient exploration of the space of reconciled gene trees. *Systematic Biology*, *62*(6), 901–12. doi: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. doi: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. doi: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. doi:10.1111/j.1574-6976.2008.00123.x

## *Data Accessibility 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.

## *Data citation*

Sanders JG, Powell S, Kronaue DJC, Vasconcelos HL, Fredrickson ME, Pierce NE (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

## *Author Contributions*

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

The authors declare no conflicts of interest.

# *Figures*

**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 (ξ=0**, B)**, the microbial is the same as the host tree; under a model with vertical transmission and host-switches **(**ξ>0, **C),** microbial trees are simulated from the host tree with various numbers of switches ξ. We find the mutation rate $\hat{\mu}$ and the number of switches $\hat{\xi}$ that maximize the probability of the alignment.

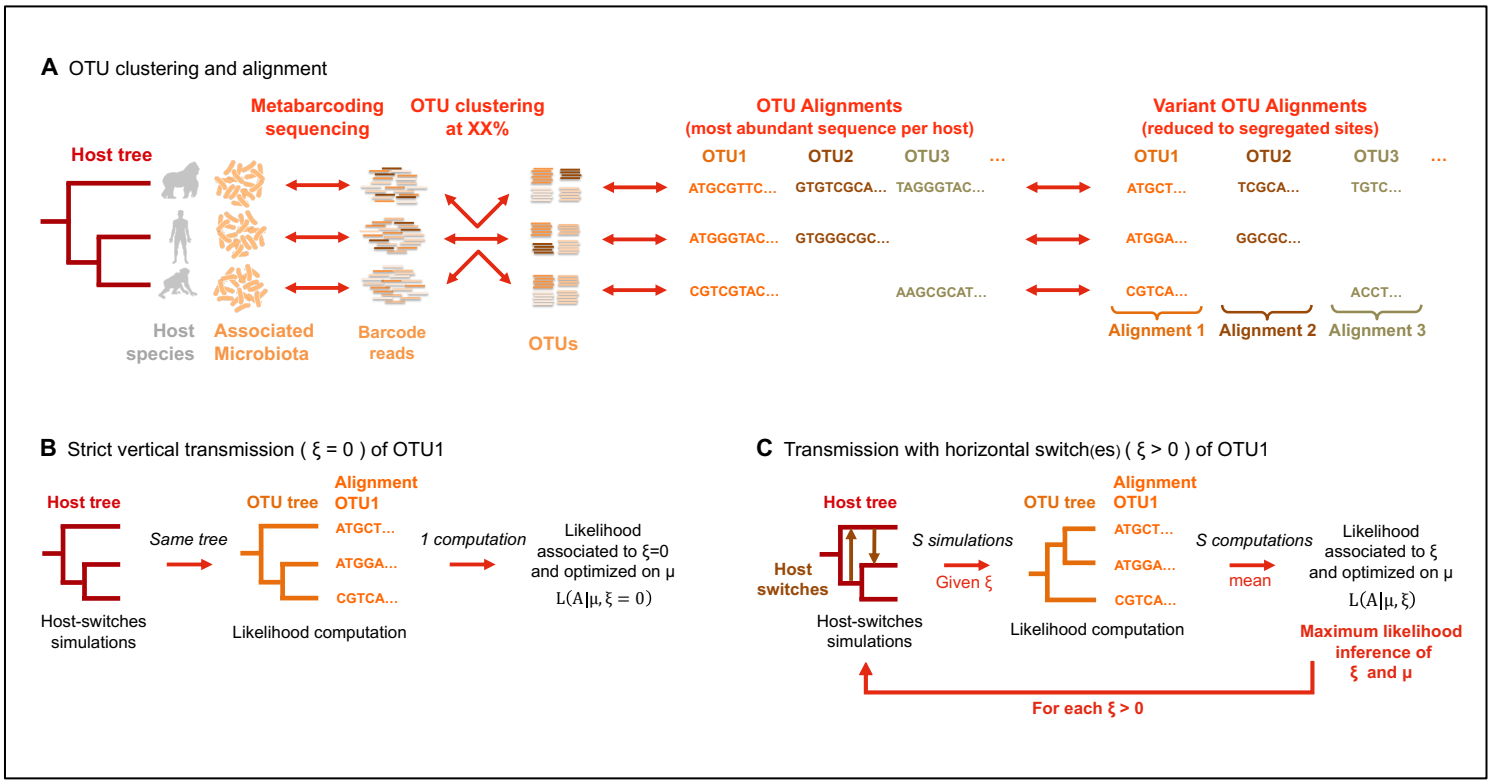

**Figure 2: Host-switch simulations**

**(A)** Four types of host-switch can occur on the host tree T **(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.

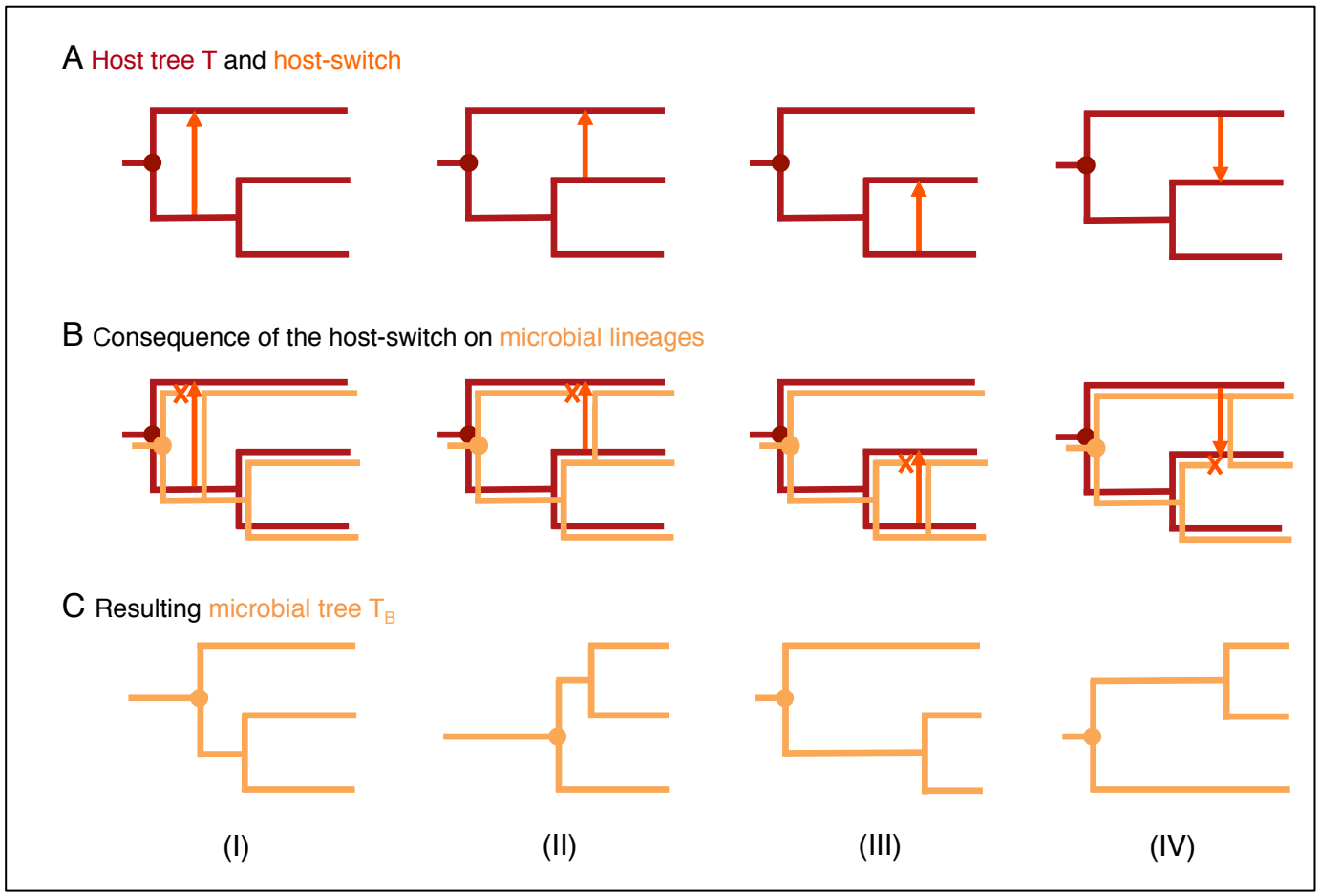

**Figure 3: Parameter estimation**

Estimated *versus* simulated number of switches ξ **(A)** and mutation rate μ **(B)** under various evolutionary scenarios (strict vertical transmission, vertical transmission with a given number of switches, and independent evolution). 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.

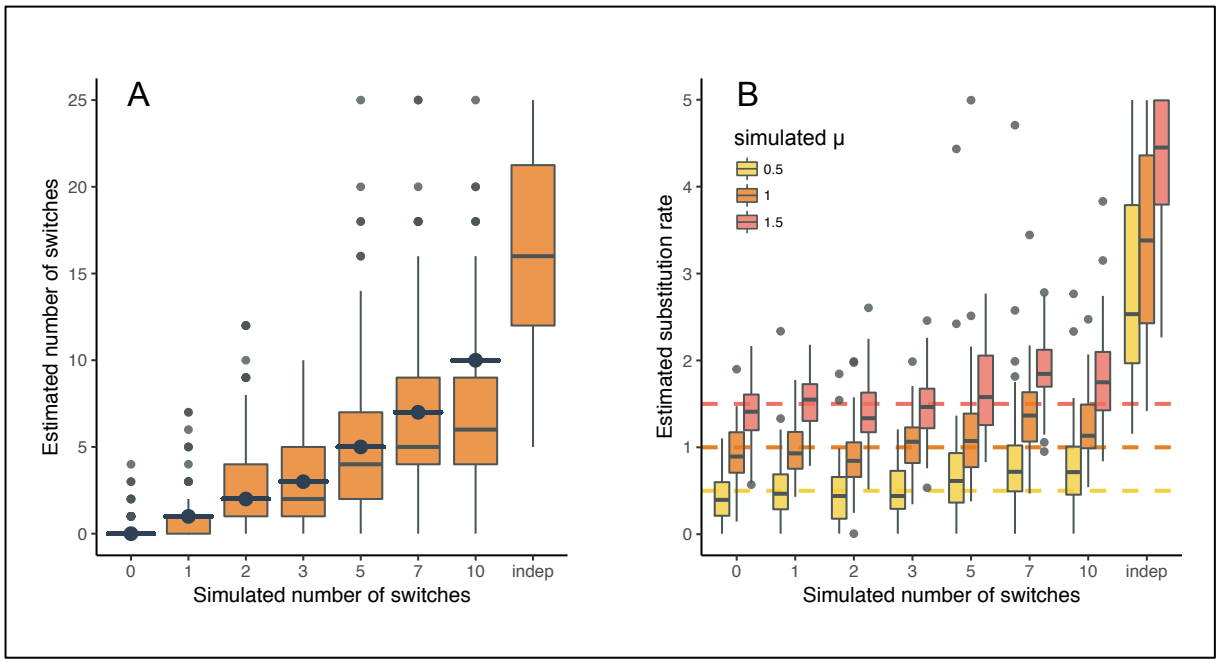

**Figure 4: Model selection**

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).

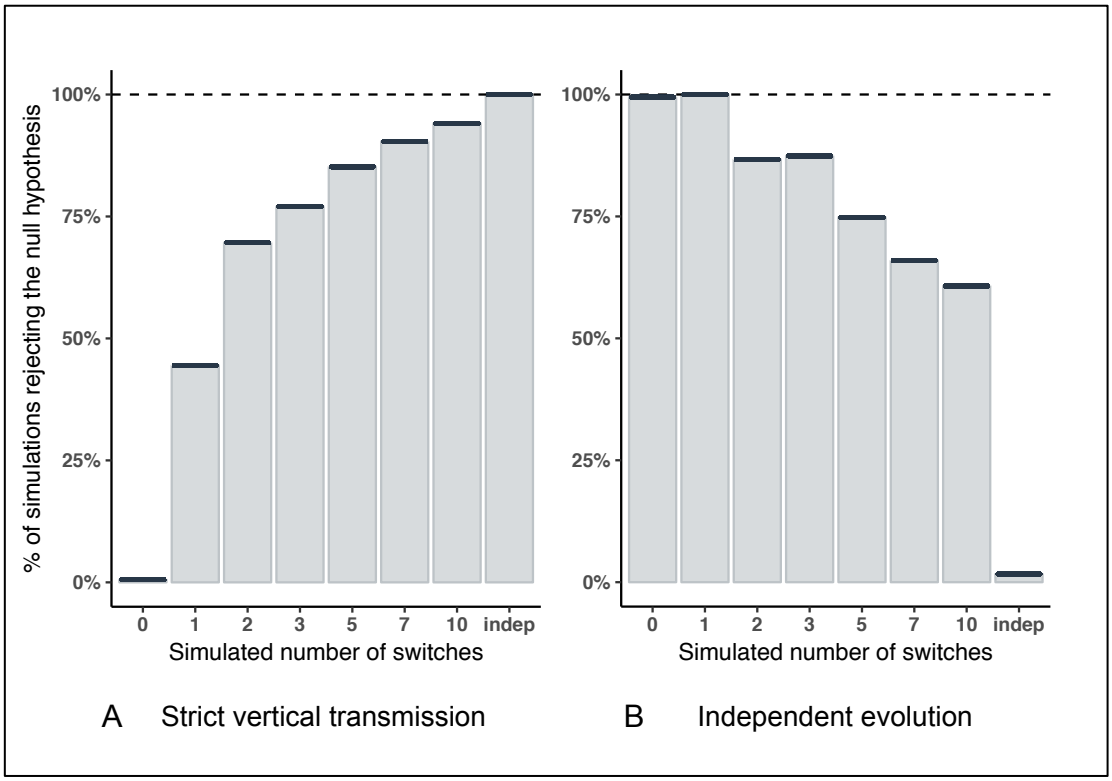

**Figure 5: Transmitted OTUs in the great ape microbiota:**

**(**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. The size of the dots represents the absolute number of reads (on a log scale) of the corresponding OTU found in each host.