HOME Tutorial - HOst-Microbiota Evolution

Benoît Perez-Lamarque

November 2018

This document indicates how to use our model of **HO**st-**M**icrobiota **E**volution. The first part presents the simulations of mock microbiota. The second part present how to deal with empirical application (example on the great apes microbiota). The last part gives examples of how to interpret the HTML outputs of HOME.

Citation: Perez-Lamarque B., Morlon H., A modeling approach for characterizing symbiont inheritance during host-microbiota evolution using maximum likelihood inference, 2018. http://rmarkdown.rstudio.com

Contact: Benoît Perez-Lamarque, benoit.perez.lamarque@gmail.com

Contents:

Installation:

Our model is part on the R package RPANDA (Morlon et al., 2016) availbale on the CRAN or from gitHub.

```
library(devtools)
install_github("hmorlon/PANDA")
```

Run Simulations:

Simulated host tree
Simulations can be done on a simulated host tree:

```
########  Parameters  #######

name <- "simulation_tree_1" # name of the simualtions on "tree 1"
n <- 20  # number of host individuals
name_index <-  c("S1","S2","S3","S4","S5","S6") # name of the different simulations
simul <- c(0,1,3,5,"indep","indep") # simulated scenarii:
# i) 0 for strict vertical   transmission
# ii) any positive integer for the number of host-switches transmissions
```

```
# iii) "indep" to simulate independent evolutions and environmental acquisitions.

simulated_mu <- 1 # simulated relative substitution rate
# NB: simulated_mu=1 corresponds to on average 1 mutation per nucleotide)
N <- 300 #total number of nucleotides in the alignment
proportion_variant <- 0.1 # proportion of variant nucelotides in the alignment

lambda <- c(1:n) # values of number of switches to test during estimations.
nb_tree <- 10000 # number of trees (for Monte-Carlo estimation of the number of switches)
raref <- F # if TRUE rarefactions on the number of trees are performed

path <- "~/"  # path toward the folder of simulations
nb_cores <- 1 # number of cores to run the analyses
seed <- 1 # seed for simulations


########  Simulation  #######
simulate_data(model="uniform", name=name, mu=simulated_mu, n=n, nb_cores=nb_cores,
              name_index=name_index, simul=simul, N=N,
              proportion_variant=proportion_variant,
              path=path, seed=seed)
```

Empirical host tree
Or you can *provide a host tree* (e.g. an empirical tree) and simulate the evolution of a mock microbiota on it. in that case, the filename of the host tree sould be well-formated **host_tree_NAME.tre** (Newick format) and saved in your PATH.

```
name <- "simulation_tree_1" # name of the simualtions (sould match the name of your host tree)

host_tree <- host_tree <- read.tree(file=paste("host_tree_",name,".tre",sep=""))

########  Simulation  #######
simulate_data(model="uniform", name=name, mu=simulated_mu, n=n, nb_cores=nb_cores,
              name_index=name_index, simul=simul, N=N,
              proportion_variant=proportion_variant,
              path=path, seed=seed, provided_tree=host_tree)
```

Then, you can proceed to the **parameters estimation**.

```
########  Run HOME  #######

HOME_model(name=name,name_index=name_index,path=path,path_alignment=path_alignment,
           nb_cores=nb_cores,seed=seed,nb_tree=nb_tree,lambda=lambda,raref=raref,
           empirical=FALSE,randomize=TRUE,nb_random=10)
# nb_random: number of randomizations in the model selection on indep. evolutions
```

Run Empirical applications:

Let's run HOME for 3 the great apes microbiota. Previously, a folder ("path") must contain the host tree with a filename **host_tree_NAME.tre** (Newick format). An another folder ("path_alignment") must contain all the OTU alignments with the filenames **alignment_NAME_OTU.fas** (FASTA format).

For instance, in this empirical application, you must provide:
– */path/host_tree_great_apes_97.tre*
– */path_alignment/alignment_great_apes_97_OTU880092397.fas*
– */path_alignment/alignment_great_apes_97_OTU838728681.fas*
– */path_alignment/alignment_great_apes_97_OTU934380954.fas*

```
########  Parameters  #######

name <- "great_apes" # name of the empirical application
# Great apes microbiota with OTUs defined at 97%
name_index <-  c("OTU892624276","OTU47610657","OTU733943228")
# name of the different OTUs

lambda <- c(1,2,3,4,5,6,7,8,9,10,12,14,16,18,20,25)
# values of number of switches to test during estimations.
nb_tree <- 5000 # number of trees (for Monte-Carlo estimation of the number of switches)
raref <- F # if TRUE rarefactions on the number of trees are performed


path <- "~/"  # path toward the main folder
path_alignment <- "~/" # path toward the folder containing the different OTU alignments


nb_cores <- 1 # number of cores to run the analyses
seed <- 1 # seed for simulations

########  Download the data of this example  #######
##  OTU892624276, OTU47610657, and OTU733943228 from the great apes microbiota ##
example_great_apes_microbiota(name,path)


########  Run HOME  #######

HOME_model(name=name,name_index=c("OTU892624276","OTU47610657","OTU733943228"),
           nb_tree=nb_tree,lambda=lambda, empirical=TRUE,randomize=TRUE,
           nb_random=10,path=path)
```

Interpret Results:

Example 1: Results from a simulation with horizontal transmission

Description of the data - Simulation

Parameters	Values
Number of symbiont-host association:	20
Simulated substitution rate:	1
Number of simulated switches:	3
Seed for simulations:	30
Sequences length:	300
Probability variant sites:	0.1
Number of invariant sites:	281
Number of strictly variant sites:	19

Summary of the most likely scenario:

Parameters	Values
Inferred substitution model:	K80
Transition/Transversion ratio:	0.6558
Estimated substitution rate:	0.8519
Estimated number of switches:	3
Associated likelihood:	137.151
p-value: Strict vertical transmission:	6e-04
p-value: Independent evolutions (nb switches):	0
p-value: Independent evolutions (subs. rate):	0

Host-switches inference:

Most likely scenario estimated by the host-switches estimation (see Figure 1 & 2):

ksi	mu	-log(Likelihood)
3	0.8519	137.151

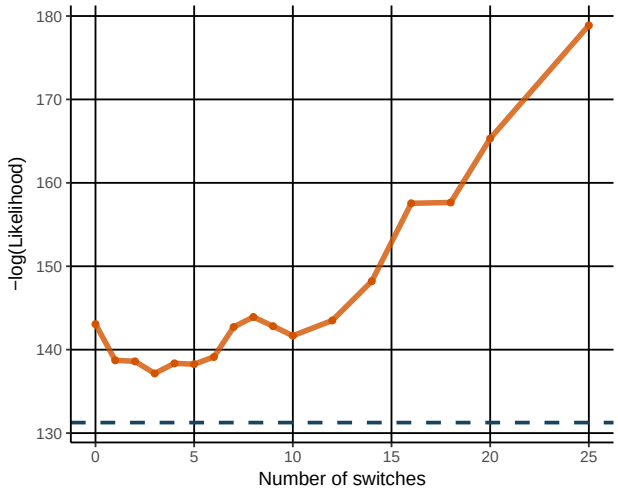

Figure 1: Profil of minus log likelihood as a function of the number of switches.

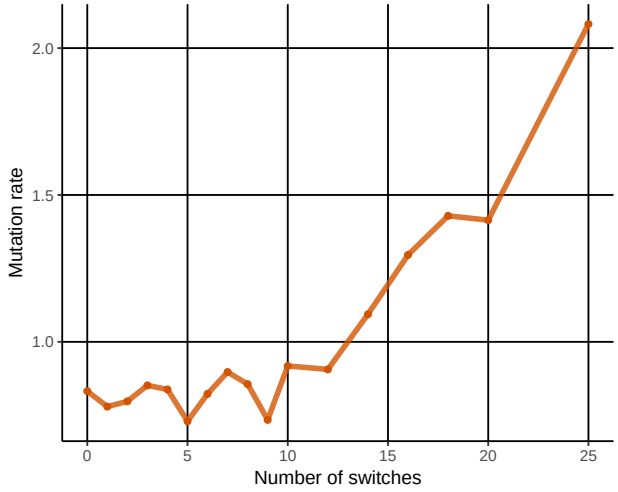

Figure 2: Estimated substitution rate as a function of the number of switches.

Strict vertical transmission model:

Likelihood ratio test testing the model of strict vertical transmission (ksi=0). Strict vertical transmission is rejected if p-value < 0.05 (see Figure 3).

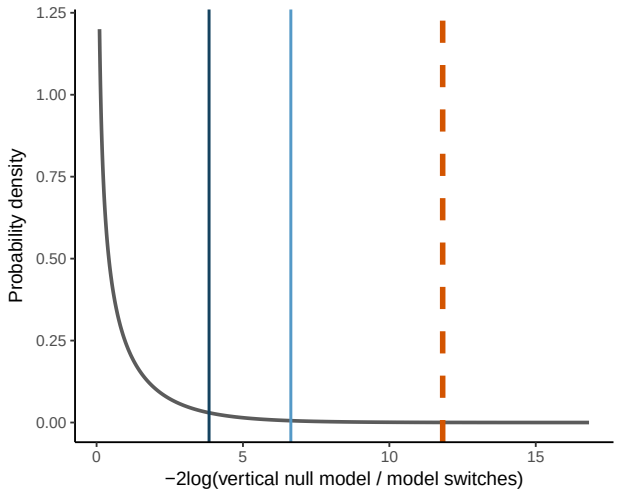

Figure 3: Results of the likelihood ratio test. The grey curve correponds to the Chi2 distribution with df=1. The dark blue line (resp. light) stands for the 0.05 (resp. 0.01) p-value threshold and the dashed orange line is the observed LRT ratio.

Host-symbiont independent evolutions:

Model selection on independent evolutions (see Figure 4).

Test	p-values
Empirical ranking (ksi distribution)	0
Empirical ranking (mu distribution)	0

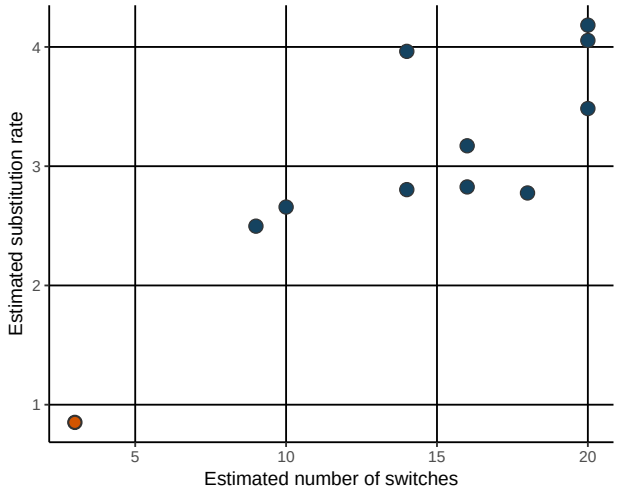

Figure 4: Representation of the estimated numbers of switches and the estimated substitution rates for the randomized alignments (in blue) and the empirical alignment (in orange). Independent evolutions can be rejected if the orange dot stands alone in the bottom left corner (i.e. rejected if p-values < 0.05).

Estimated substitution model

Estimated rate matrix:

Rates	A	C	G	T
A	-1	0.17	0.66	0.17
C	0.17	-1	0.17	0.66
G	0.66	0.17	-1	0.17
T	0.17	0.66	0.17	-1

Nucleotide frequencies:

A	C	G	T
0.25	0.25	0.25	0.25

Simulated switches:

Simulated switch(es):	1	2	3
Branch origin	1	36	35
Branch arrival	35	2	28
Absolute position	0.022	0.089	0.13

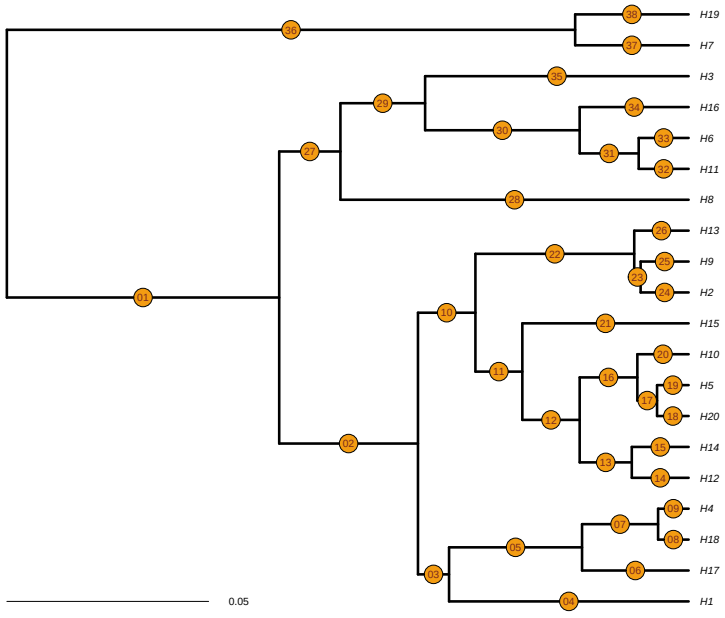

Figure 5: Host tree.

Example 2: Results from a simulation with strict vertical transmission

Description of the data - Simulation

Parameters	Values
Number of symbiont-host association:	20
Simulated substitution rate:	1
Number of simulated switches:	0
Seed for simulations:	30
Sequences length:	300
Probability variant sites:	0.1
Number of invariant sites:	279
Number of strictly variant sites:	21

Summary of the most likely scenario:

Parameters	Values
Inferred substitution model:	K80
Transition/Transversion ratio:	0.70
Estimated substitution rate:	0.70
Estimated number of switches:	0
Associated likelihood:	149.8
p-value: Strict vertical transmission:	1
p-value: Independent evolutions (nb switches):	0
p-value: Independent evolutions (subs. rate):	0

CONCLUSION: STRICT VERTICAL TRANSMISSION.

Host-switches inference:

Most likely scenario estimated by the host-switches estimation (see Figure 6 & 7):

ksi	mu	-log(Likelihood)
3	0.70	149.8

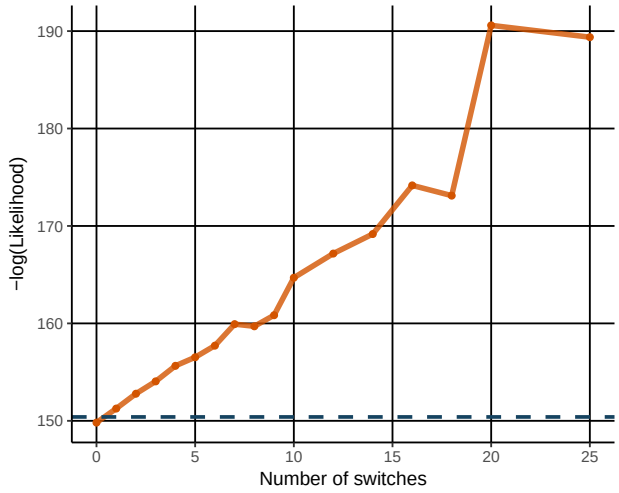

Figure 6: Profil of minus log likelihood as a function of the number of switches.

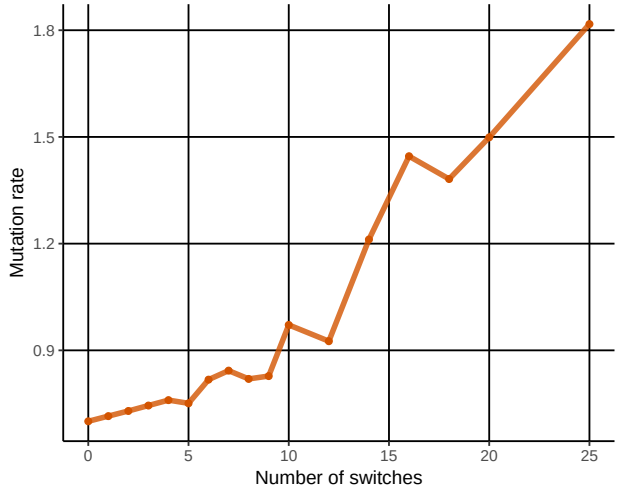

Figure 7: Estimated substitution rate as a function of the number of switches.

Strict vertical transmission model:

Likelihood ratio test testing the model of strict vertical transmission (ksi=0). Strict vertical transmission is rejected if p-value < 0.05 (see Figure 8).

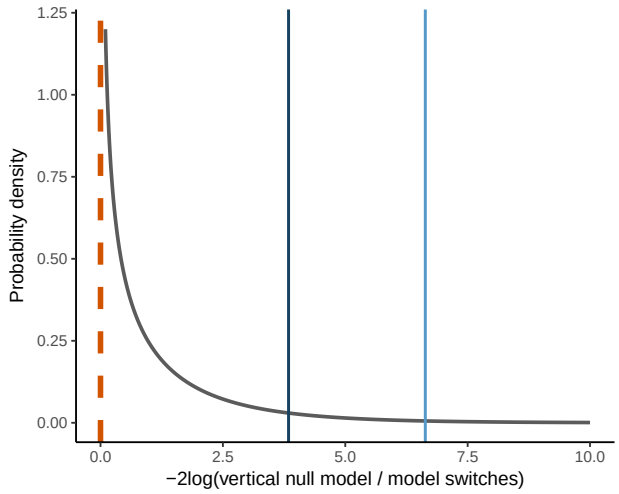

Figure 8: Results of the likelihood ratio test. The grey curve correponds to the Chi2 distribution with df=1. The dark blue line (resp. light) stands for the 0.05 (resp. 0.01) p-value threshold and the dashed orange line is the observed LRT ratio.

Host-symbiont independent evolutions:

Model selection on independent evolutions (see Figure 9).

Test	p-values
Empirical ranking (ksi distribution)	0
Empirical ranking (mu distribution)	0

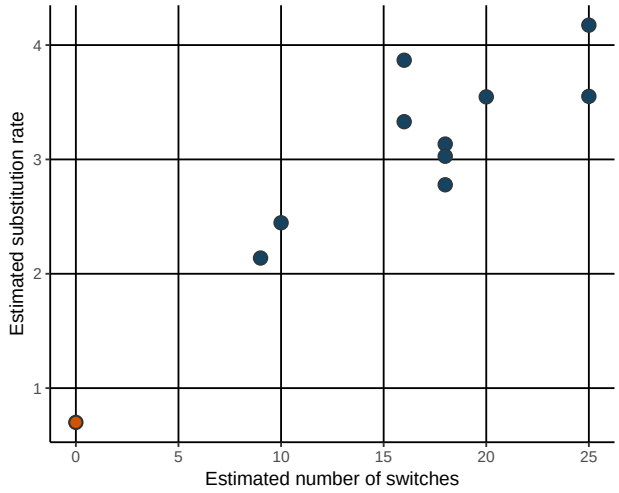

Figure 9: Representation of the estimated numbers of switches and the estimated substitution rates for the randomized alignments (in blue) and the empirical alignment (in orange). Independent evolutions can be rejected if the orange dot stands alone in the bottom left corner (i.e. rejected if p-values < 0.05).

Estimated substitution model

Estimated rate matrix:

Rates	A	C	G	T
A	-1	0.15	0.7	0.15
C	0.15	-1	0.15	0.7
G	0.7	0.15	-1	0.15
T	0.15	0.7	0.15	-1

Nucleotide frequencies:

A	C	G	T
0.25	0.25	0.25	0.25

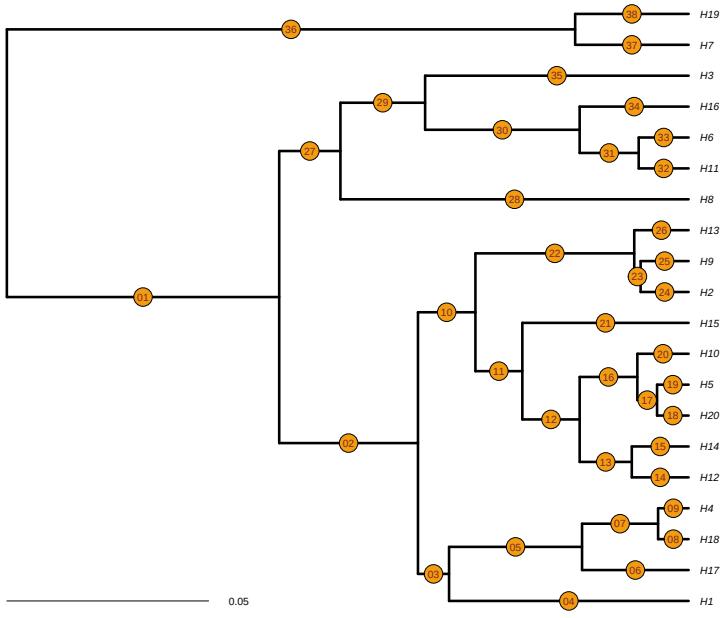

Figure 10: Host tree.

Example 3: Results from a simulation with environnemental acquisition (independent evolution)

Description of the data - Simulation

Parameters	Values
Number of symbiont-host association:	20
Simulated substitution rate:	1
Number of simulated switches:	independent
Seed for simulations:	30
Sequences length:	300
Probability variant sites:	0.1
Number of invariant sites:	278
Number of strictly variant sites:	22

Summary of the most likely scenario:

Parameters	Values
Inferred substitution model:	K80
Transition/Transversion ratio:	0.54
Estimated substitution rate:	3.3793
Estimated number of switches:	25
Associated likelihood:	270.3
p-value: Strict vertical transmission:	0
p-value: Independent evolutions (nb switches):	0.56
p-value: Independent evolutions (subs. rate):	0.89

CONCLUSION:INDEPENDENT EVOLUTIONS (BASED ON THE NUMBER OF SWITCHES AND THE SUBSTITUTION RATE).

Host-switches inference:

Most likely scenario estimated by the host-switches estimation (see Figure 11 & 12):

ksi	mu	-log(Likelihood)
25	3.38	270.3

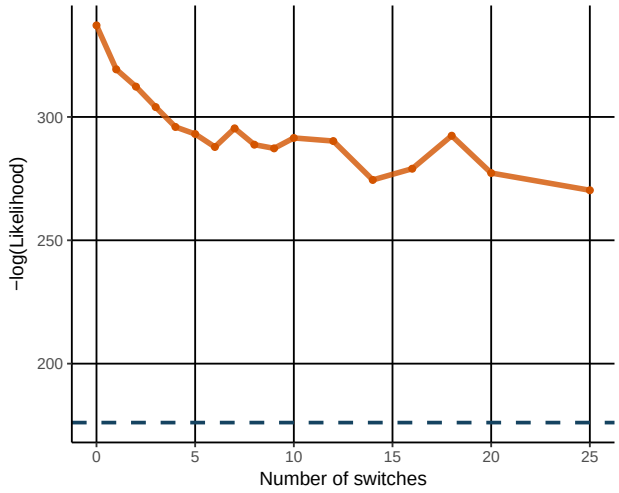

Figure 11: Profil of minus log likelihood as a function of the number of switches.

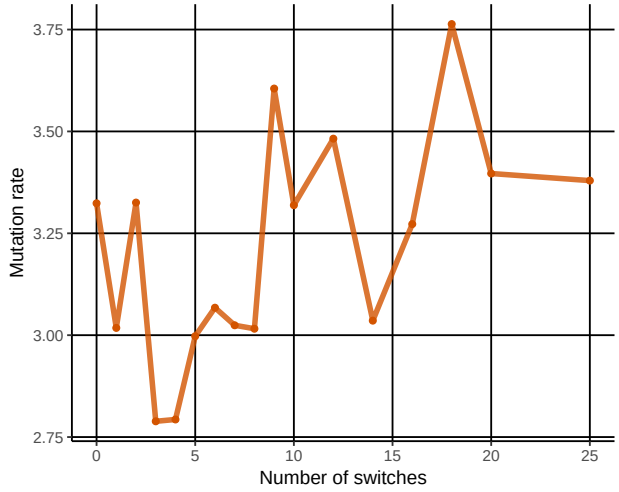

Figure 12: Estimated substitution rate as a function of the number of switches.

Strict vertical transmission model:

Likelihood ratio test testing the model of strict vertical transmission (ksi=0). Strict vertical transmission is rejected if p-value < 0.05 (see Figure 13).

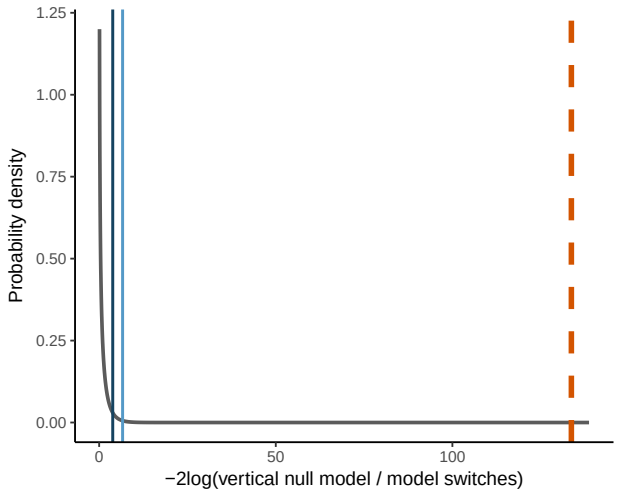

Figure 13: Results of the likelihood ratio test. The grey curve correponds to the Chi2 distribution with df=1. The dark blue line (resp. light) stands for the 0.05 (resp. 0.01) p-value threshold and the dashed orange line is the observed LRT ratio.

Host-symbiont independent evolutions:

Model selection on independent evolutions (see Figure 14).

Test	p-values
Empirical ranking (ksi distribution)	0.56
Empirical ranking (mu distribution)	0.88

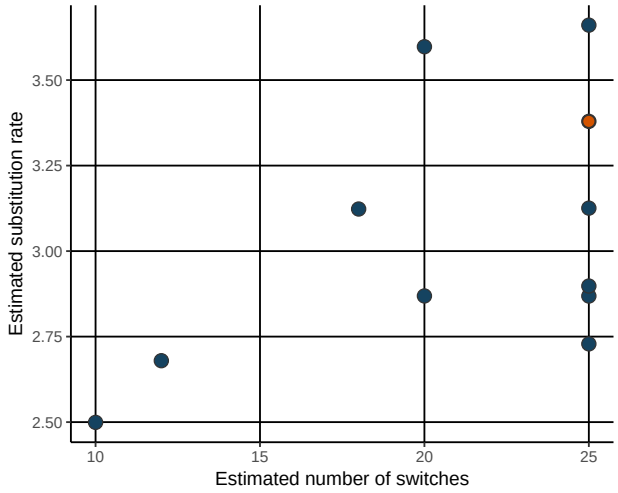

Figure 14: Representation of the estimated numbers of switches and the estimated substitution rates for the randomized alignments (in blue) and the empirical alignment (in orange). Independent evolutions can be rejected if the orange dot stands alone in the bottom left corner (i.e. rejected if p-values < 0.05).

Estimated substitution model

Estimated rate matrix:

Rates	A	C	G	T
A	-1	0.23	0.54	0.23
C	0.23	-1	0.23	0.54
G	0.54	0.23	-1	0.23
T	0.23	0.54	0.23	-1

Nucleotide frequencies:

A	C	G	T
0.25	0.25	0.25	0.25

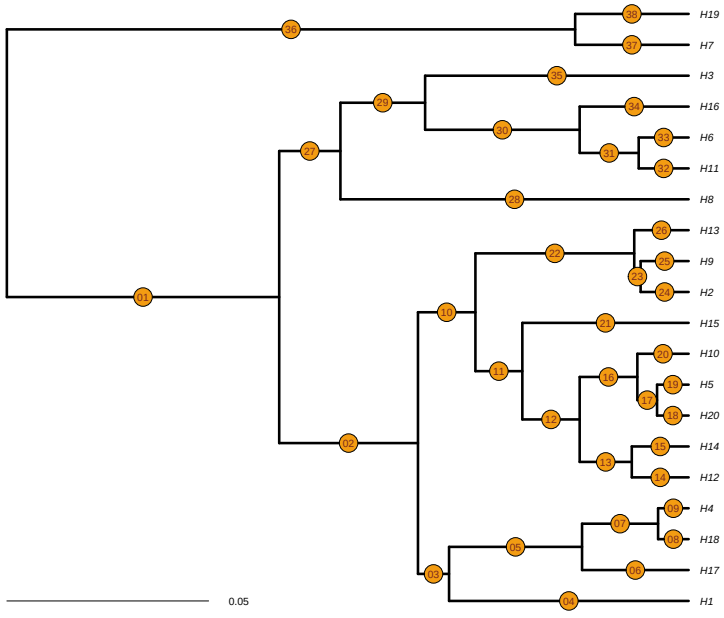

Figure 15: Host tree.