Let's start by considering the ARD model. This model has three parameters:

1. The rate of going from N to T (let's call it r_1).
2. The rate of going from T to N (r_2).
3. The root prior for state N (let's call it π_N).

Note that the root prior for state T is not a free parameter, because $\pi_T = 1 - \pi_N$. The ER model is equivalent, except that you only have one rate parameter.

Now, given a certain set of values for r_1, r_2, and π_N (let's say that $r_1 = x, r_2 = y$, and $\pi_N = z$), the *likelihood* is the probability of obtaining the data (in this case, your observations for the tip states), given that set of parameter values:

$$\mathbb{L}(r_1 = x, r_2 = y, \pi_N = z) = \mathbb{P}(\mathrm{Data} \mid r_1 = x, r_2 = y, \pi_N = z)$$

This is what you get as the "current ln-likelihood" when you run the analyses with fixed values for the rates and for π_N: since there is no uncertainty on the parameter values, you get just a single likelihood value.

However, you can consider additional restrictions on these. For example, if you're interested in the root node (ρ), you can consider the "likelihood that the root node is in state N, given that $r_1 = x$, $r_2 = y$, and $\pi_N = z$ ". This would be:

$$\mathbb{L}(r_1 = x, r_2 = y, \pi_N = z, \rho = N) = \mathbb{P}(\mathrm{Data} \mid r_1 = x, r_2 = y, \pi_N = z, \rho = N)^{1}$$

These likelihoods are all relatively easy to compute, using Felsenstein’s pruning algorithm. However, what we are actually interested in is not much the likelihood, but the “posterior probability that the root node is in state N, given the observed data and that $r_1 = x$, $r_2 = y$, and $\pi_N = z$”. This can be expressed as:

$$\mathbb{P}(\rho = N \mid Data, r_1 = x, r_2 = y, \pi_N = z)$$

We cannot compute this directly. However, using Bayes’ theorem, we can rearrange this:

$$\mathbb{P}(\rho = N \mid Data, r_1 = x, r_2 = y, \pi_N = z)$$
$$= \frac{\mathbb{P}(Data \mid \rho = N, r_1 = x, r_2 = y, \pi_N = z) \cdot \mathbb{P}(\rho = N \mid r_1 = x, r_2 = y, \pi_N = z)}{K_1}$$

Here, $\mathbb{P}(\rho = N \mid r_1 = x, r_2 = y, \pi_N = z)$ is the *a priori* (prior) probability that the root node is in state N, given the parameter values, but without considering the data. This is a known quantity.[2] K_1 instead is a normalisation constant; in this case, it can be easily expressed as:

$$K_1 = \mathbb{P}(Data \mid \rho = \boldsymbol{N}, r_1 = x, r_2 = y, \pi_N = z) \cdot \mathbb{P}(\rho = \boldsymbol{N} \mid r_1 = x, r_2 = y, \pi_N = z) +$$
$$+ \mathbb{P}(Data \mid \rho = \boldsymbol{T}, r_1 = x, r_2 = y, \pi_N = z) \cdot \mathbb{P}(\rho = \boldsymbol{T} \mid r_1 = x, r_2 = y, \pi_N = z) =$$

[1] Note that, because of how the model works, the value of π_N becomes irrelevant here (because you’re conditioning the root node), but this is beside the point.

[2] In our model, $\mathbb{P}(\rho = N \mid r_1 = x, r_2 = y, \pi_N = z) = z$; but again, this is beside the point.

$$= \mathbb{P}(\text{Data} \mid r_1 = x, r_2 = y, \pi_N = z)$$

K_1 is, in fact, a marginal likelihood: it is a likelihood (that is, the probability of the data given a certain set of parameter values), *marginalised* (that is, summed over/integrated) over all the possible values for an uncertain quantity (in this case, the state of the root node). This likelihood, though, is still conditioned on a certain set of fixed parameters.

Things get more complicated when the values of r_1, r_2 and/or π_N are not known with certainty. We represent this uncertainty with a prior distribution on the parameter values. For example, if you have a prior with a $Gamma(2,2) = \Gamma(2,2)$ distribution for r_1, you're saying: "If I didn't have any data to look at, I would think that a value of $r_1 \leq 1$ has about a 9% probability of being correct."

What you're now interested in, is the "posterior probability that the root node is in state N, given the observed data, that $r_2 = y$ and $\pi_N = z$, and a prior distribution of $\Gamma(2,2)$ for r_1" (basically, r_1 no longer has a known fixed value, but many possible values with various degrees of plausibility). This is:

$$\mathbb{P}(\rho = N \mid Data, r_1 \sim \Gamma(2,2), r_2 = y, \pi_N = z)$$

Applying again Bayes' theorem, you get:

$$\mathbb{P}(\rho = N \mid Data, r_1 \sim \Gamma(2,2), r_2 = y, \pi_N = z) \\ = \frac{\mathbb{P}(Data \mid \rho = N, r_1 \sim \Gamma(2,2), r_2 = y, \pi_N = z) \cdot \mathbb{P}(\rho = N \mid r_1 \sim \Gamma(2,2), r_2 = y, \pi_N = z)}{K_2}$$

$\mathbb{P}(\rho = N \mid r_1 \sim \Gamma(2,2), r_2 = y, \pi_N = z)$ is still a known quantity[3], but we cannot analytically compute the term $\mathbb{P}(Data \mid \rho = N, r_1 \sim \Gamma(2,2), r_2 = y, \pi_N = z)$. This can be expressed as:

$$\mathbb{P}(Data \mid \rho = N, r_1 \sim \Gamma(2,2), r_2 = y, \pi_N = z) \\ = \int_0^{+\infty} \mathbb{P}(Data \mid \rho = N, r_1 = x, r_2 = y, \pi_N = z)\, \mathrm{d}\mathbb{P}(r_1 = x \mid r_1 \sim \Gamma(2,2))$$

This is a Lebesgue integral, but it doesn't really matter. The point is that, since this cannot be computed analytically, some clever people came up with the idea of using the MCMC approach to approximate its value. You could again think of this term as a marginal likelihood: it is a likelihood that has been marginalised (integrated) over the uncertainty in the value of the r_1 parameter. However, this is still conditioned on the state of the root node, which is not a model parameter.

The really interesting term here is the normalisation constant K_2. Like before:

$$K_2 = \mathbb{P}(Data \mid \rho = \boldsymbol{N}, r_1 \sim \Gamma(2,2), r_2 = y, \pi_N = z) \cdot \mathbb{P}(\rho = \boldsymbol{N} \mid r_1 \sim \Gamma(2,2), r_2 = y, \pi_N = z) + \\ + \mathbb{P}(Data \mid \rho = \boldsymbol{T}, r_1 \sim \Gamma(2,2), r_2 = y, \pi_N = z) \cdot \mathbb{P}(\rho = \boldsymbol{T} \mid r_1 \sim \Gamma(2,2), r_2 = y, \pi_N = z) = \\ = \mathbb{P}(\text{Data} \mid r_1 \sim \Gamma(2,2), r_2 = y, \pi_N = z)$$

[3] And it's still equal to z.

K_2 is again a marginal likelihood, integrated over the uncertainty in the model parameter r_1 and the state of the root node ρ. Computing this is again not analytically possible; furthermore, thanks to the MCMC approach, we don't normally need its value.
You could then add uncertainty on the other parameters too (e.g., another $\Gamma(2{,}2)$ prior for r_2 and a $Beta(1{,}1)$), obtaining:

$$K_3 = \mathbb{P}(\text{Data} \mid r_1 \sim \Gamma(2{,}2), r_2 \sim \Gamma(2{,}2), \pi_N \sim Beta(1{,}1))$$

This is "the" marginal likelihood, because this has been integrated over all possible sources of uncertainty.

As I mentioned above, the marginal likelihood cannot be computed analytically; to estimate it, sMap uses the stepping-stone algorithm. When you ask sMap to compute the marginal likelihood, the program will try to integrate over all possible sources of uncertainty (i.e., internal nodes and parameters for which you have defined a prior distribution).

If there is no uncertain parameter (e.g., all of them have fixed values), the stepping-stone algorithm will not be actually executed, as the result in this case can be computed analytically. Therefore, you don't get the .marginal.likelihood.txt output file, but just the file with the "regular" likelihood (integrated over node states and not parameter values).

Why do we care about the marginal likelihood?

Let's say, again, that we are interested in the state of the root node. In particular, consider the "odds" that the state of the root node is N:

$$\mathbb{O}(\rho = N) = \frac{\mathbb{P}(\rho = N)}{1 - \mathbb{P}(\rho = N)} = \frac{\mathbb{P}(\rho = N)}{\mathbb{P}(\rho = T)}$$

You can think of a Bayesian analysis as a way of "updating" these odds. Before looking at the data, you had some "prior" thoughts about these odds. For example, maybe you didn't care whether ants do trophallaxis or not and you thought these two hypotheses were more or less equivalent (odds ≈ 1); or maybe you thought that trophallaxis is really gross, thus it's very unlikely that ants actually do it (odds $\gg 1$). Then, you looked at the data, and maybe you changed your mind (or maybe not): these are your "posterior" odds:

$$\mathbb{O}(\rho = N \mid Data) = \frac{\mathbb{P}(\rho = N \mid Data)}{\mathbb{P}(\rho = T \mid Data)}$$

Applying Bayes' theorem, we can express this as:

$$\mathbb{O}(\rho = N \mid Data) = \frac{\dfrac{\mathbb{P}(Data \mid \rho = N) \cdot \mathbb{P}(\rho = N)}{\mathbb{P}(Data)}}{\dfrac{\mathbb{P}(Data \mid \rho = T) \cdot \mathbb{P}(\rho = T)}{\mathbb{P}(Data)}} = \frac{\mathbb{P}(Data \mid \rho = N)}{\mathbb{P}(Data \mid \rho = T)} \cdot \frac{\mathbb{P}(\rho = N)}{\mathbb{P}(\rho = T)}$$

This means that the posterior odds are equal to the prior odds, times a "correction factor":

$$\mathbb{O}(\rho = N \mid Data) = \frac{\mathbb{P}(Data \mid \rho = N)}{\mathbb{P}(Data \mid \rho = T)} \cdot \mathbb{O}(\rho = N)$$

This “correction factor” is the Bayes factor (BF), which expresses how much the data have caused us to change our mind about the state of the root node. The BF is a number between 0 and $+\infty$. If $BF \gg 1$, the data strongly support the hypothesis that the root node was in state N. If $BF \ll 1$, the data strongly support the hypothesis that the root node was in state T. If $BF \approx 1$, the data are not able to say anything about this question.

Now, let’s consider two “models”. In this context, a “model” is a set of uncertainties/restrictions on the parameter values for the reconstruction. For example, in your case you could have models M_1 and M_2:

$$M_1 \coloneqq \begin{cases} r_1 \sim \Gamma(2,2) \\ r_2 \sim \Gamma(2,2) \\ \pi_N = 0 \end{cases} \qquad M_2 \coloneqq \begin{cases} r_1 \sim \Gamma(2,2) \\ r_2 \sim \Gamma(2,2) \\ \pi_N = 1 \end{cases}$$

The models are the same, except for the value of the parameter π_N. What you want to find is which of these is the “true model” M. Thus, you ask the data: “How should I update my prior odds in favour of one of these models?”. The Bayes factor here is:

$$BF = \frac{\mathbb{P}(Data \mid M = M_1)}{\mathbb{P}(Data \mid M = M_2)}$$

It’s exactly the same as above, except I’ve replaced $\rho = N$ with $M = M_1$. But if you look closely, you should realise that:

$$\mathbb{P}(Data \mid M = M_1) = \mathbb{P}(Data \mid r_1 \sim \Gamma(2,2), r_2 \sim \Gamma(2,2), \pi_N = 0)$$

This is basically the same as K_1, K_2, and K_3 above! In this case the two rates have a certain amount of uncertainty, while the π_N is fixed.

What this means is that you can use sMap to compute the marginal likelihood for M_1, then use it again to compute the marginal likelihood for M_2, and then compute their ratio to get the Bayes Factor.[4]

Final important note: since you have blended analyses, what you are really comparing here are four models:

$$M_{1,ARD} \coloneqq \begin{cases} r_1 \sim \Gamma(2,2) \\ r_2 \sim \Gamma(2,2) \\ \pi_N = 0 \end{cases} \qquad M_{2,ARD} \coloneqq \begin{cases} r_1 \sim \Gamma(2,2) \\ r_2 \sim \Gamma(2,2) \\ \pi_N = 1 \end{cases}$$

[4] You can find tables with rule-of-thumb values for interpreting Bayes Factors e.g. on Wikipedia (or in the original publications that are cited there).

$$M_{1,ER} := \begin{cases} r_1 \sim \Gamma(2,2) \\ r_2 = r_1 \\ \pi_N = 0 \end{cases} \qquad M_{2,ER} := \begin{cases} r_1 \sim \Gamma(2,2) \\ r_2 = r_1 \\ \pi_N = 1 \end{cases}$$

What you now want to do is to marginalise over the rate model, i.e., compute the odds for "$M = M_{1,ARD}$ or $M = M_{1,ER}$". Assuming that you have equal priors over the ER and ARD models, you can compute the Bayes Factor as:

$$BF = \frac{\mathbb{P}(Data \mid M = M_{1,ARD}) + \mathbb{P}(Data \mid M = M_{1,ER})}{\mathbb{P}(Data \mid M = M_{2,ARD}) + \mathbb{P}(Data \mid M = M_{2,ER})}$$

So, basically take the marginal likelihood for the ARD model and ER model where π_N is fixed to a certain value and sum them, then divide this sum by the sum of the marginal likelihoods for the ARD and ER model where π_n is fixed to the other value.

Note that you need to sum the likelihoods and not their logarithms!