

Migrate tutorial 2015

Peter Beerli



Tutorial Overview

How to get MIGRATE

Locally, the most recent copy of the software on the server is in the `/class/molevol-software/migrate` directory, you can either use the program that is already installed on the class server or use these files in the *local_distribution* directory to install MIGRATE on your own computer. if this fails, download MIGRATE from the Migrate download-website (<http://popgen.sc.fsu.edu/Migrate/Download.html>).

Comparison of gene flow models using Bayes Factors with MIGRATE

Most are familiar with the concept of likelihood ratio tests, or Akaike's Information criterion for model comparison. This tutorial describes how to compare population models using Bayes Factors. These allow comparing nested and un-nested models, without assuming Normality, or large samples. Bayes factors are ratios of marginal likelihoods. In contrast to the maximum likelihood, the marginal likelihood is the integral of the likelihood function over the complete parameter range. MIGRATE can calculate such marginal likelihoods for a particular migration model (Beerli and Palczewski 2010).

Executive Summary

This tutorial steps through all necessary program runs to calculate Bayes factors for comparing different gene flow models. We need to do the following:

- Decide on the models that are interesting for a comparison. The method does not work well for a fishing expedition where one would try to evaluate all possible models; this is only possible for population models with very few number of populations. It will be possible to enumerate all models for two or three populations but more will be very daunting.
- Run each model through MIGRATE. Use the same prior settings for each of them because the prior distribution has some influence on the Bayes factors. Use the heating menu to allow

for at least four chains. The menu supplies a shortcut to specify the temperatures, it is `#`. It generates temperatures that are spaced in a particular way: they are spaced so that the inverse of the temperature are regularly spaced on the interval 0 to 1. For example, the 4 different chains have temperatures 1.0, 1.5, 3.0, 100,000.0, this results in the spacing 1.0, 0.666, 0.333, and 0.0.

- Compare the marginal likelihood of the different runs and calculate the Bayes factor and calculate the probability for each model.

More Details



The following pages detail all steps using a small example. We use a simulated dataset that was generated using parameters that force a direction of migration from the population Ascona (A) to the population Brissago (B). The Brissago population is larger than the Ascona population and no individual from Brissago ever goes to Ascona, but Brissago receives about 1 migrant every four generation from Ascona. The dataset name is **twoswisstowns** (if you are not at the workshop then download here (<http://www.peterbeerli.com/downloads/twoswisstowns>)) We will evaluate a total of 7 population models.

1. a full migration model with two population sizes and two migration rates (from A to B and from B to A);
2. a model with two population sizes and one migration rate to Brissago;
3. a model with two population sizes and one migration rate to Ascona;
4. a model where Ascona and Brissago are part of the same panmictic population.
5. a model where Ascona is an ancient city and Brissago was built a new from people who left Ascona; but Brissago is still attractive and people migrate to there ever since.
6. a model where Brissago is an ancient city and Ascona was built a new from people who left Brissago; but Ascona is still attractive and people migrate to there ever since.
7. a model where Ascona is an ancient city and Brissago was built a new from people who left Ascona and they became hostile towards each other and nobody moves between the cities.

We know the truth therefore we have some prejudice about the ranking of the models, **model 5** should be *best*, models with the same migration pattern as the truth should work better than those with alternative migration patterns, or those with many parameters.

First we need to figure out how to run the dataset efficiently in MIGRATE. For that we pick a parameter-rich **model 1** and experiment with run conditions until we are satisfied that the run converges and delivers posterior distributions that look acceptable. Here are now the detailed instructions how to rank population genetics models for a particular dataset.

Familiarize with MIGRATE [Tutorial Start]

Make a new directory and download or copy the datafile

```
#if you are AT THE WORKSHOP use this
rsync -avz /class/molevol-software/migrate/migrate_tutorial .
cd migrate_tutorial

#if you are NOT at the workshop use this
mkdir migrate_lab
cd migrate_lab
wget http://popgen.sc.fsu.edu/tutorials/BF_migrate_tutorial2/twoswisstowns
```

Start the program: the regular distribution comes in two flavors the single cpu processor version called `migrate-n` and the parallel processing version that runs on cluster or computers with multiple cores is called `migrate-n-mpi`. We use a new version of MIGRATE: version 4.2.2. On the cluster you should be able to call `migrate-n`. (In this text I will call the program from now on simply MIGRATE).

We will run the exercise on the server. On the server type `migrate-n`

On your laptop you may need to use `./migrate-n`, if the program is in the same directory. The main menu will appear, looking like this

```
[pbeerli@class-02 migrate_lab]$ migrate-n
+++++
+                                     +
+  POPULATION SIZE, MIGRATION, DIVERGENCE, ASSIGNMENT, HISTORY  +
+  Bayesian inference using the structured coalescent           +
+                                     +
+++++
PDF output enabled [Letter-size]
Version 4.2.2a    [July-18-2015]
Program started at   Tue Jul 21 15:03:15 2015
```

```
=====
MAIN MENU
=====

D      Data type currently set to: DNA sequence model
I      Input/Output formats and Event reporting
P      Parameters  [start, migration model]
S      Search strategy
W      Write a parmfile
```

Q Quit the program

To change the settings type the letter for the menu to change

Start the program with typing Yes or Y

==>

Go to the Input/Output formats menu (press I and hit Enter), in the INPUT section change the Datafile name to twoswisstowns, Return to the main menu by typing Y.

In the Search strategy menu: Change the Number of recorded steps in chain to 1000, and also change the Burn-in for each chain to 1000. Do not worry about priors or other runtime options for the moment. Return to the main menu.

Save the changes by using the menu item Write a parmfile. This will write a file named parmfile. Now run the program (pressing Y will start the run if you are in the main menu). For this dataset the runtime will be very short. On a modern computer this will take under a minute. If this takes more than 3 minutes, something is not set up correctly! On the server this takes about 29.6 seconds (on my macOS 10.10.4: 8.8 seconds).

The program writes considerable information during the run to the screen, that gives some information about the run. Most interesting are the acceptance ratio for the genealogy and the autocorrelations of the parameter and the genealogy. If the autocorrelation is high and the effective sample size is low (< 500) then a longer run may be needed. If the priors boundaries are too tight, then you will see that the values reported are either very close or exactly at the upper prior boundary, in these cases you need to extend the prior range. See prior problems in the output, but for this dataset we will have no such problems.

Look at the outfile.pdf, you will need to transfer the pdf file to your computer and use preview or acrobat or another PDF viewer. In the outputfile look at the figures labeled Bayesian analysis: Posterior distribution, you see histograms similar to the ones in Figure 1. We expect single peaks where the shading of the histogram shows one dark block in the center (50% credibility set), two light gray bars indicating the extent of the 95% credibility set, and two lighter gray bars indication the 99% credibility set. We expect a histogram that looks smooth that usually has a single peak, commonly similar to an exponential or normal distribution with a heavy right tail.

In your investigation of Figure 1 you recognize that the histogram has a few kinks because our run was too short. Now restart MIGRATE and set in the strategy menu the setting for change the number of recorded steps in chain from 1,000 to 10,000. This will lengthen the run by a factor of 10.

The next step would take 4x longer than before run the same way, but we could run this in parallel, by calculating the parameters for the loci in parallel. After saving your change to 10,000 steps, close down the program and restart, but this time instead of `migrate-n parmfile` execute this (these instructions will work on the MBL cluster and may most likely not work on your computer or other clusters).

```
module load openmpi
mpirun -np 6 migrate-n-mpi parmfile
```

Run this way, it will set up a computer with 6 virtual nodes (one master and 5 workers); the workers will do most of the work and at the end report back to the master who then assembles and prints the results. On the cluster, this step takes 339 seconds (on my mac 149 seconds).

For the MBL tutorial: USE THE PARALLEL VERSION! Because we want to use the thermodynamic integration method for our marginal likelihood calculations, we need to turn on

heating. Once you start MIGRATE with the mpi command above, use the Strategy menu to turn on heating. MIGRATE will tell what to do next, you will need to enter 4 chains sampling at every tenth (10) interval using the temperature scheme that is suggested with the character #. Save the parmfile, and run. Here the sequence of entries that you will need to enter:

```

.....
13   Heating:                                NO
16   Run analysis without data:              NO

Are the settings correct?
(Type Y to go back to the main menu or the number for a menu to change)
==>13
Heating scheme? < NO | YES | STATIC | BOUNDED_ADAPTIVE >

==> Static
Enter the number of different "heated" chains.
Minimum is 4
==> 4
Enter the interval between swapping trees
Enter 0 (zero) for NO swapping
[Current interval is 1]
==> 1
Enter 4 "temperatures"
[The coldest temperature, which is the first, has to be 1
For example: 1.0 1.5 3.0 1000000.0]
OR give a range of values [linear increase: 1 - 10]
[exponential increase: 1 @ 10]
or, most lazily, let me suggest a range [simply type a #]
@@@@ For model comparison, the range of temperatures @@@@
@@@@ MUST include a very hot chain (>100000.0)          @@@@
==> #
Chain   Temperature
-----
4       1000000.00000
3       3.00000
2       1.50000
1       1.00000
Is this correct? [YES or NO]
==> yes

```

Don't forget to write the parmfile to save the settings. It should give a better posterior distribution histogram and will add a full table of (natural) log marginal likelihoods is shown towards the end of the outfile.pdf, but you may still see double peaks.

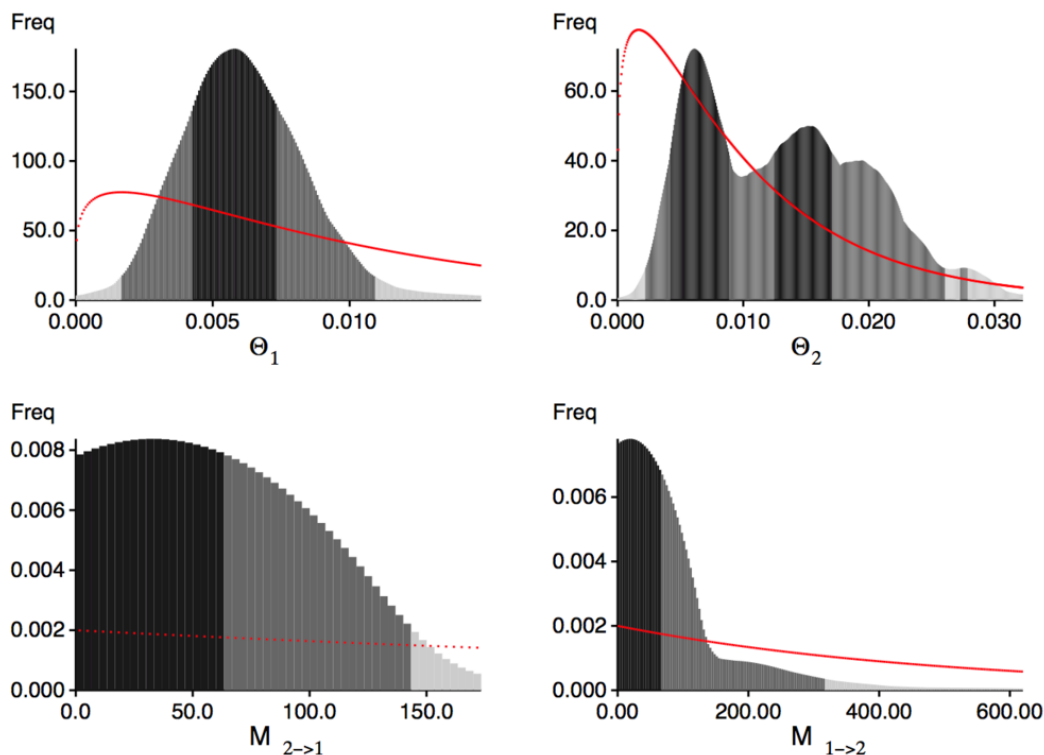


Figure 1: Example of a short run showing an unsatisfactory posterior histogram; red lines are the prior distributions used.

With your own data you may want to do another round of refinements, but eventually, by comparing the medians and modes of the parameters in the table and the shape of the histograms you should see a good agreement on similar values, if the modes of the different runs are not within the 50% credibility intervals you certainly need to run longer.

Finally, let's work on the different models

We now start to work on the different models. Remember, we want to run 7 models, and this will take some time, I suggest that you work with your neighbor and use the parallel version. To make sure that we have examples of all different models for the discussion at the end of tonight, start with the model that is equivalent to the class computer name you use, for example I am on class07 and would start with **model 7**. If you are on class08, class09, class10 pick model 5, 6, or 7, respectively. Once you finished your designated model work on the other ones, the more models the better.

Create directories for all model

I have created a small script that assists with the creation of all model directories, execute at the prompt:

```
create_migmodels
```

It will create the directories and will also copy the parmfile into all these directories because we will need to change each parmfile to accommodate the population model. The parmfile in the

directories will be correctly set up, except for the population model. You will also need to copy the data file ‘twoswisstowns’ into these new directories.

Model 1: xxxx



This model allows migration between A and B, the populations are assumed to exist since a very long time. The xxxx in the title is a shortcut for this model it means that we estimate migration rates and population sizes. I use a particular notion to define the model, assume there is a migration matrix (or call it population interaction matrix), on the diagonal we have instructions for the population size, for example, estimate (x or *) or constant (c), there are several other possibilities which you could look up in the manual. The off-diagonal are the immigration rates, a * or ‘x’ means that this parameter is estimated. This particular model is the default and, in principle does not need any further input and you can simply run MIGRATE, to report follow now the instructions in section Reporting values.

Model 2: x0xx



This model allows migration from A to B. The populations are assumed to exist since a very long time. The x0xx in the title is a shortcut for this model it means that we estimate an immigration rate into the second population and population sizes. I use a particular notion to define the model, read the instructions about the migration/divergence matrix completely before you start typing. Once you have changed the population model, exit the parameter menu, write the parmfile, run MIGRATE to report follow now the instructions in section Reporting values.

Model 3: xx0x



This model allows migration from B to A. The populations are assumed to exist since a very long time. The xx0x in the title is a shortcut for this model; it means that we estimate an immigration rate into the first population and population sizes. I use a particular notion to define the model, read the instructions about the migration/divergence matrix completely before you start typing. Once you have changed the population model, exit the parameter menu, write the parmfile, run MIGRATE to report follow now the instructions in section Reporting values.

Model 4: x



Start MIGRATE and choose the menu **Parameter settings**. Choose the entry about sampling locations. We want to use the data as if we would have sampled a single population, therefore we need to claim that the two locations Ascona and Brissago belong to the same panmictic population. MIGRATE's default is to assume that every location is an individual population. The dialog (figure below gives an example of this interaction with the menu) will ask first how many locations are in the dataset (for our example we have 2). After that, you will need to assign the locations to a population. For this model we need to assign each location to the same population, so location 1 (Ascona) belongs to population 1 and location 2 (Brissago) also belongs to population 1. You need to enter 1 1 (one space one). With multiple populations more complicated settings are possible. Save the parmfile and then run MIGRATE. Follow now the instructions in section Reporting values.

```
Associate sampling locations with populations
-----
This menu allows to combine sample locations into populations
For example there are 4 locations: 1, 2, 3, 4
They can be combined into 2 populations
by mapping the 4 positions 1, 2, 3, 4 to 1, 1, 1, 2
Migrate will now combine the first three locations

Give (1) the number of populations <return> then (2) the mappi

How many localities are in the data set?
[Default: every sampling location is a population]
> 2
Enter now the remappings (little checking is done with this, e
  1 2
>1 1
```

Model 5: x0Bx



This model allows divergence, B splits off from A, with migration from A to B after the split. The first population (A: Ascona) exists for a long time and the population B: Brissago splits off at the time we want to estimate. The x0Dx in the title is a shortcut for this model; it means that we estimate a divergence time and an immigration rate into the second population and estimate also population sizes. I use a particular notion to define the model, read the instructions about the migration/divergence matrix completely before you start typing. Once you have changed the population model, exit the parameter menu, write the parmfile, run MIGRATE to report follow now the instructions in section Reporting values.

Model 6: xB0x



This model is the mirror image of Model 5. This model allows divergence, A splits off from B, with migration from B to A after the split. For details read model 5 (simply exchange A with B). Read the instructions about the migration/divergence matrix completely before you start typing. Once you have changed the population model, exit the parameter menu, write the parmfile, run MIGRATE to report follow now the instructions in section Reporting values.

Model 7: x0dx



This model allows divergence, B splits off from A. The first population (A: Ascona) exists for a long time and the population B: Brissago splits off at the time we want to estimate. After that split there is no interaction between the populations. This scenario is equivalent to a species split. The x0dx in the title is a shortcut for this model; it means that we estimate a divergence time and estimate also population sizes. I use a particular notion to define the model, read the instructions about the migration/divergence matrix completely before you start typing. Once you have changed the population model, exit the parameter menu, write the parmfile, run MIGRATE to report follow now the instructions in section Reporting values.

Report values

We want to compare models, but the tutorial time is too short to do a thorough job and thus our runs will be too short and the posterior distributions of the parameters may still be in bad shape. For our exercise we are mostly interested in the model selection/ordering using marginal likelihoods.

Come to the front and write down the log marginal likelihood into the spreadsheet (look at the example figure labeled Log-Probability of the data given model (marginal likelihood)) [the figure is from a different dataset, so be not alarmed that the marginal likelihoods are different]. You will need to report a number from the row labeled All. There are three columns, report the values for the Bezier approximation column.

<i>Log-Probability of the data given the model (marginal likelihood)</i>			
Use this value for Bayes factor calculations: $BF = \text{Exp}[\ln(\text{Prob}(D \mid \text{thisModel}) - \ln(\text{Prob}(D \mid \text{otherModel}))]$ or as $LBF = 2 (\ln(\text{Prob}(D \mid \text{thisModel}) - \ln(\text{Prob}(D \mid \text{otherModel})))$ shows the support for thisModel]			
Locus	Raw thermodynamic score(1a)	Bezier approximation score(1b)	Harmonic mean(2)
1	-1963.93	-1813.58	-1802.41
2	-1934.41	-1804.46	-1799.72
3	-2091.58	-1912.06	-1901.83
4	-2606.12	-2202.97	-2149.12
5	-2138.61	-1913.70	-1890.85
All	-10726.56	-9638.69	-9535.84
(1a, 1b and 2) are approximations to the marginal likelihood, make sure that the program run long enough! (1a, 1b) and (2) should give similar results, in principle. But (2) is overestimating the likelihood, it is presented for historical reasons and should not be used (1a, 1b) needs heating with chains that span a temperature range of 1.0 to at least 100,000. (1b) is using a Bezier-curve to get better approximations for runs with low number of heated chains [Scaling factor = 8.087035]			

Migration/Divergence matrix modifications

Start MIGRATE, choose the **Parameter** menu.

Choose the entry labeled **Model is set to**. MIGRATE will now show a dizzying list of options, don't panic, we will only use a few of them. MIGRATE will ask you how many populations are used: enter 2. For a 2-population model we can have 4 parameters. For example, two population sizes and two migration rates. A * or x means that that particular parameter will be unrestrictedly estimated, a zero (0) means that that particular parameter will not be estimated (is not used). Our goal is to set one of the migration/divergence parameters to 0. MIGRATE needs to know how to treat all connections between the populations. The connection matrix is square so we can label it like it is shown in Table 1.

MIGRATE asks now that you input each row, this can be done by either specifying * 0 (see the lower subtable of Table 1) and then return and then entering the next line * * return (second row in second table), or you can enter the whole matrix as * 0 * *.

For 'Divergence' models instead of the '*' you need to enter a 'd' for divergence without migration, and a 'D' for divergence with migration.

Table 1: Model for uni-directional migration: top table shows the parameters, the bottom table shows the values that need to be assigned to the custom-migration option.

To\From	Ascona	Brissago
Ascona	Θ	$M_{B \rightarrow A}$
Brissago	$M_{A \rightarrow B}$	Θ

To\From	Ascona	Brissago
Ascona	*	0
Brissago	*	*

Table 2: Model for divergence where a population splits off another looking forward in time: top table shows the parameters, the bottom table shows the values that need to be assigned to the custom-migration option, a D means divergence with migration and a d means divergence only.

To\From	Ascona	Brissago
Ascona	Θ	0
Brissago	$\Delta_{A \rightarrow B}$	Θ

To\From	Ascona	Brissago
Ascona	*	0
Brissago	D	*

Comparing models

How to calculate Bayes factors? In Table 3 I summarized all log marginal likelihoods, $\ln(mL)$, the Bayes factors are often calculated in very different ways. Here, I report the natural log Bayes factors where

$$LBF = 2(\ln mL(\text{model}_1) - \ln mL(\text{model}_2))$$

Using the guidelines of Kass and Raftery (1995), values smaller than -2 suggest preference for ‘model 2’, values larger than 2 suggest preference for ‘model 1’. We can use the log marginal likelihoods or the BF to order the models (see column Choice in Table 3).

We also can calculate the model probability. It is calculated by dividing each marginal likelihood by the sum of the marginal likelihoods of all used models:

$$\text{Prob}(\text{model}_i) = \frac{mL_{\text{model}_i}}{\sum_j mL_{\text{model}_j}}$$

Note that for the above formula uses the marginal likelihoods, *not* the *log* marginal likelihoods (which is what the program reports). The calculation of model probabilities from the reported log likelihoods is easy with computer programs that have variable precision (for example Maple or Mathematica). Calculations on a desk calculator often fail, for example the likelihood of model 1 is a remarkably small number because the likelihood is $\exp(-4803.07) = 1.130323625060 \times 10^{-2086}$, my emulated HP sci 15C calculator delivers 0.0000. But you can calculate the above quantities using this recipe: (1) find the largest log likelihood (-4795.23), (2) subtract that number from each log likelihood in the list (result: $-2.27, 0.0, 2.5, -26.67$), (3) exponentiate each element in the new list (result: $0.1033, 1.0, 0.0821, 2.6144 \times 10^{-12}$), (4) sum all elements in the list up ($0.1033 + 1.0 + 0.0821 + 2.6144 \times 10^{-12}$), this is the denominator (1.1854). (5) now divide each element in the list by that sum and the numbers will look like the one in table 3 last column.

I have added a little python script that calculates the model probabilities from the text outfiles: If you want to see the results of the model exercise do:

```
grep " All " */outfile_* | sort -n -k 4,4 | migbf.py
```

Table 3: Showing log marginal likelihoods for all models tested (and three more, but without the panmictic model) and model probabilities

Model	Log(mL)	LBF	Model-probability
1:xxxx:	-9662.42	-23.73	0.0000
2:xBxx:	-9661.98	-23.29	0.0000
3:xxBx:	-9661.52	-22.83	0.0000
4:xd0x:	-9656.51	-17.82	0.0000
5:xB0x:	-9649.33	-10.64	0.0000
6:xx0x:	-9648.93	-10.24	0.0000
7:x0dx:	-9641.77	-3.08	0.0402
8:x0xx:	-9641.01	-2.32	0.0859
9:x0Bx: [your model 5]	-9638.69	0.00	0.8739

The MacOS file system is capitalization-blind, D and d are the same, for that you see capital B which are placeholders for capital D.

Looking at the model probabilities we can see that the “true” model has considerably higher support than the full model or the model that suggests a wrong direction of gene flow.

Summary of results

The best model (the one with the highest marginal likelihood) is model 5 (custom-migration={*0D*}). if we use the thermodynamic approximation of marginal likelihood. MIGRATE also reports the harmonic mean, but I suggest to ignore it and use thermodynamic integration (as we did) although it will be more costly to run. It was shown several times (for example, Beerli and Palczewski 2010, Xie et al. 2011) that the harmonic mean estimator is not a good estimator and may be misleading and prefer the more complex model. The picture below is a sample from the class tutorial done on August 1st 2011.

	A	B	C	D	E	F	G	H	I
1			Migrate Bayes Factors						
2			write down the marginal likelihood						
3			Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7
4	Model prob (max		3.1879E-09	0.996088827	0.00391117	6.24468E-37	178.2095117	9.33635E+30	1.70924E-37
5	Model probability		8.1549E-06	0.000639798	0.00149262	1.54915E-49	0.999359246	0.006539495	1.62258E-46
6	Parameters		4	3	3	1	3	3	3
7	Specification		xxxx	x0xx	xx0x	x	x0Dx	x0Dx	x0dx
8	N		10	13	9	7	6	6	5
9	Maximum		-9663.07	-9643.51	-9649.05	-9726.87	-9642.66	-9648.81	-9642.09
10	Mean		-9670.768	-9652.54538	-9659.0511	-9764.93286	-9645.19167	-9650.215	-9750.62
11	Standard dev.		10.953394	14.98442114	17.2543548	24.89945295	4.93516126	0.971365019	151.0453535
14			-9665.64	-9644.18	-9649.68	-9730.16	-9643.6	-9650.38	-9642.2
15			-9700.89	-9644.32	-9683.38	-9778.86	-9642.7	-9648.81	-9642.09
16			-9666.11	-9681.96	-9649.39	-9780.09	-9643.47	-9651.84	-9954.09
17			-9665.99	-9643.7	-9649.95	-9726.87	-9655.23	-9649.97	
18			-9663.07	-9643.51	-9649.05	-9779.27			
19			-9665.95	-9643.73	-9649.76				
20			-9666.5	-9665.95	-9655.81				
21			-9671.56	-9644.34					
22				-9643.9					
23				-9644.8					
24				-9654.45					
25									

References

- Beerli, P. and M. Palczewski. 2010. Unified framework to evaluate panmixia and migration direction among multiple sampling locations. *Genetics* 185: 313–326.
- Kass, R. E. and A. E. Raftery. Bayes factors. 1995. *Journal of the American Statistical Association* 90(430): 773–795.

- Xie, W., P. O. Lewis, Y. Fan, L. Kuo, and M.-H. Chen. 2011. Improving marginal likelihood estimation for Bayesian phylogenetic model selection, *Systematic Biology*, 60: 150–160.

Retrieved from https://molevol.mbl.edu/index.php?title=Migrate_tutorial_2015&oldid=4845

This page was last modified on 14 September 2015, at 19:41.