

Bacmeta:
Simulator for genomic evolution
in bacterial metapopulations
Supplementary material

Alexi Sipola, Pekka Marttinen and Jukka Corander

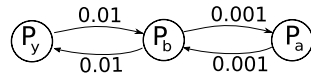


Figure S1: Migration rate parameters for second example

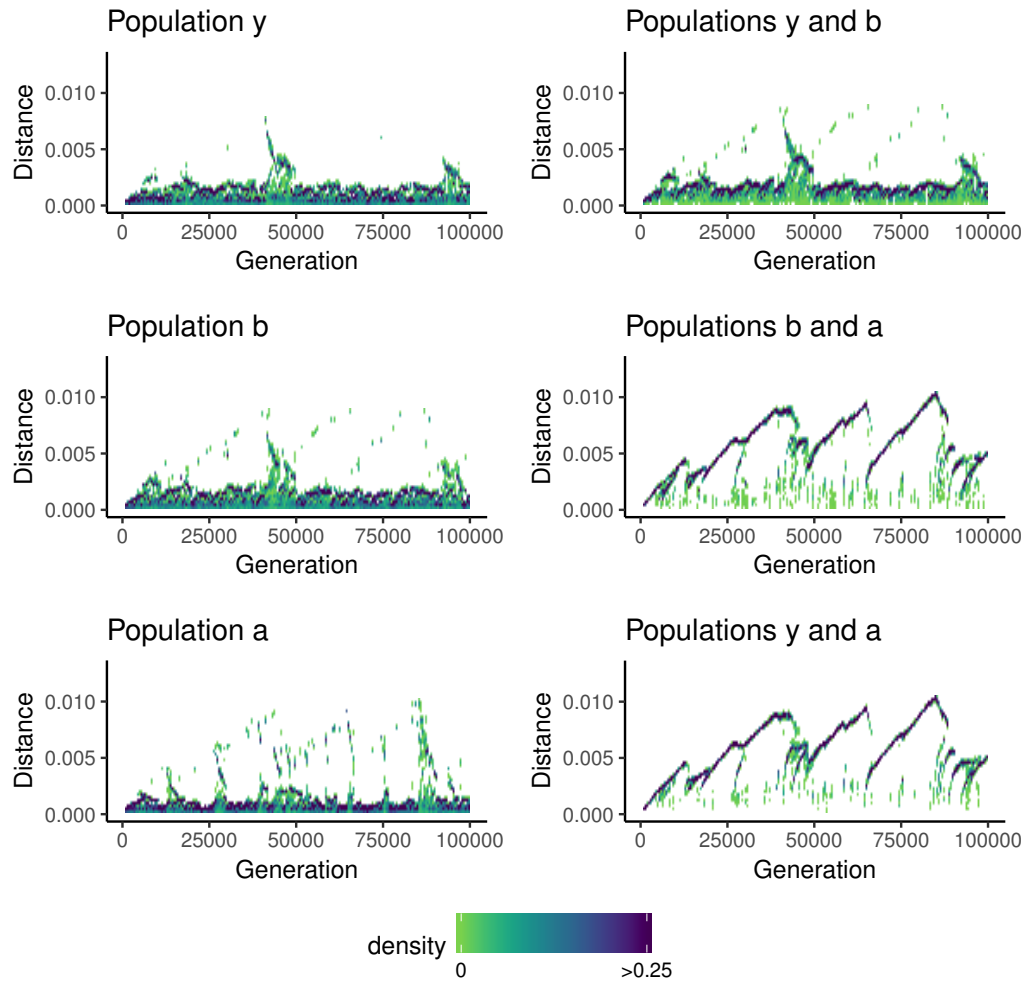


Figure S2: Intra- and interpopulation pairwise distances across generations in the second example simulation run, computed every 500 generations.

1 Simulator usage

To give a quick overview of using the Bacmeta software a short summary is given here. For running the program and further information, it is recommended to view the full and up-to-date descriptions in the manual "InputOutputManual.pdf", which can be found in the folder "manuals" of Bacmeta repository at: www.bitbucket.org/aleksisipola/bacmeta (under the "source" section).

1.1 Setting up

Settings for simulations are set with plain text files, which can be edited with standard text editors or command line programs such as *sed*. The general input parameters are set in a file "simu.input", or in a file with custom filename of form "simu123.input", where "123" can be replaced with some alphanumeric string. In addition to the general parameters, if one wants to use non-homogenous migration rate network, it can be set with a file migration.input, or again with a custom filename of form "migration123.input".

For the general simulation settings, each desired parameter is set with a line in the "simu.input" file. In essence, such line includes the corresponding four letter parameter key, followed by colon, whitespace and the desired value. Additional short comment can follow after #-symbol. For example, to set number of populations to 5 and number of bacteria in each population to 1000, one would include lines:

```
NPOP: 5      #Number of populations in metapopulation
NBAC: 1000   #Number of bacteria in population
```

Full list of the parameters and their parameter keys is given with descriptions in the "InputOutputManul.pdf". Furthermore, "ModelManual.pdf" describes the role of these parameters in the model.

The possible non-homogenous migration rate network can be given in file migration.input, as an adjacency matrix of weighted, directed graph, so that edge weights represent migration rates. In essence, the input file starts with tab-separated migration rate values forming a matrix, where value in column j of row i defines migration rate from population i to population j, as a proportion of population size.

1.2 Running simulation

When the input files are set, the simulator is run with simple terminal command. The program then reads the input file(s) and runs the simulation. No user input is done after running the command and output files are created automatically, according to the given settings.

1.3 Examples

See brief examples of inputs in tables S1 and S2, or full examples of general- and migration-input files, with the resulting output files, in the FeatureExample.zip and CaseExample.zip. These are included in the "examples" folder of Bacmeta.

2 Model

Here a general outline of the model is given. Full and up-to-date description is included in the manual "ModelManual.pdf", which can be found in the folder "manuals" of Bacmeta repository at: www.bitbucket.org/aleksisipola/bacmeta (under "source" section).

Bacmeta provides simulations of a forward-time, finite metapopulation Wright-Fisher, neutral evolution model with explicit genome sequences. The simulations begin with single genotype and progress in forward time with mutations, recombinations, indels, migrations and microepidemics occurring stochastically in the populations across discrete generations.

In essence, simulation runs generation-by-generation and in each generation, events are sampled randomly according to the given parameters. For example, the amount of mutations in each generation is sampled for each population from Poisson distribution with rate:

$$\lambda_{mutations} = \theta n_{bacs} n_{loci} L = [MUTR] \cdot [NBAC] \cdot [NLOC] \cdot [LOLE] \quad (1)$$

Here n_{bacs} is amount of bacteria in population, n_{loci} is the amount of loci simulated in each bacterium, L is the length of one locus and θ is mutation rate parameter. The terms in brackets at the right hand side show the corresponding input parameter keys that are used to set these parameters.

For each mutation, the site is randomly sampled from a random bacterium. The outcome of a mutation is sampled based on the initial base

at the site and the corresponding probabilities in a mutation matrix that is defined by parameters $[ATOT]$, $[ATOG]$, $[ATOC]$, $[TTOA]$...

Other events are modelled in similar manner, with amounts of specific events modelled as Poisson processes, but naturally every event has specific modelling they follow. See "ModelManual.pdf" for full descriptions.

3 Parameter inputs for example case

Table S1: Non-default parameter values used in the the example case:

```
RSEL: 1
RORD: 1
GENR: 20000      # Amount of generations
LOLE: 1000      # Length of one locus
NLOC: 300       # Amount of loci
NBAC: 1000      # Amount of bacteria in one population
NPOP: 3         # Amount of populations
MIGI: 1
MIGP: 0.05
MUTR: 2e-7      # Mutation rate
RECR: 4         # Recombination rate
RECL: 0
RECA: 18
PROA: 0.31
PROT: 0.30
PROG: 0.18
PROC: 0.21
ATOT: 0.127
ATOG: 0.655
ATOC: 0.218
TTOA: 0.278
TTOG: 0.165
TTOC: 0.557
GTOA: 0.686
GTOT: 0.279
GTOC: 0.035
CTOA: 0.261
CTOC: 0.622
CTOG: 0.117
SUMI: 20000
GWDS: 0.2
GWDI: 20000
RECT: 20
MIGD: 1
```

For example case displayed in figure S2, same general parameters were used, with two exceptions: Generation amount was raised to 100000 and both intra- and interpopulation pairwise distance computing were set to be done in intervals of 500 generations and with sample size 0.1.

Migration rate matrix of *migration.input* file that was used for example cases with low $\lambda = 0.005$ value is displayed below. For the high $\lambda = 0.05$ case, elements with value 0.005 were switched to 0.05.

Table S2: Migration rate inputs

0	0	0.005
0.005	0	0.01
0	0.01	0