HELP FILE FOR quedo

VERSION: 0.6
LAST UPDATE: 22.02.2019

1. INTRODUCTION

quedo was designed to estimate effects of ecological variables on outcrossing rates. Estimation procedure is based on a hierarchical Bayesian approach and uses the Reversible Jump Markov Chain Monte Carlo algorithm. In result, quedo estimates slopes (regression parameters) of effects of variables measured within natural populations on either single- or multi-locus outcrossing rates. quedo takes explicitly into account both genetic structure among populations and genotyping errors, including random allele dropout and random allele misclassification.

2. COMPILATION

The code (quedo.f90) can be compiled using the GFortran compiler that is freely available for multiple platforms, including Windows, Linux or iOS. Because the program uses OpenMP interface to parallel computing, in order generate a properly-working executable binary one needs to use -fopenmp flag. Adding -O3 flag can result in further optimisations. More information can be found at https://gcc.gnu.org/fortran/.

Note that quedo.f90 uses a few 'non-standard' Fortran routines, that are specific to the GFortran dialect. Hence, another compiler (e.g. ifort, F95) may fail to compile the code.

3. DATA FILE

Data are stored in a single text file. Data entries are separated with a white space (or tab). A valid yet very simple example is shown below.

```
3       8       3       2
1       120     120     140     142     -1      -1      23.1    11.0    1
2       122     126     140     140     201     203     14.1    12.8    8
5       120     122     140     144     203     205     21.7    11.7    2
1       120     122     140     140     201     201
1       120     122     142     144     -1      -1
1       120     120     140     140     201     201
2       120     122     140     140     201     207
2       122     122     140     144     203     203
5       120     122     -1      -1      201     205
5       120     122     140     144     205     205
5       122     122     142     144     203     205
```

The first line contains information about a number of maternal individuals (N, here N = 3), a number of progeny individuals (J, here J=8), number of markers (L, here L = 3) and a number of ecological variables (M, here M = 2). Next, N rows follows, each starting with an integer indicator of maternal individual. Subsequently, a diploid genotype is given at L markers (2*L integer values are expected, with -1 reserved for missing data). Next, M values of explanatory variables are expected. Each line for maternal data ends with an integer indicator of sampling location (any ordinal number).

Finally, a block of J lines are expected to provide genotypes of progeny individuals. Each line starts with a family indicator that informs which maternal family a given progeny belongs to.

4. THE 'INFO' FILE

In order to run the analysis, the program requires a text file called 'info.txt' that provides settings for the algorithm. The example file looks like as follows:

```
10000   10000   100000  50
1       1       1       0       1       1
0.5     0.01    0.025   0.01
23125
MyData.txt
MyData
```

The first line contains:
- number of initial MCMC iterations (here 10000). During the initial stage, the program runs the saturated regression model in order to adjust proposal distributions for parameters. Proposals are adjusted every 1000 cycles to get 25-40% acceptance rates. Therefore, it is advisable to set this number to c*1000 (c=1,2,...,10,...,20).
- number of MCMC iterations for pilot adjustments of sampling distributions for regression coeficients (here 10000). During this stage, the program adjusts proposal distributions for regression parameters that is used in between-model jumps. Still, the saturated regression model is used. The number of pilot cycles should be d*1000 (d=1,2,...,10,...,20).
- number of MCMC iterations used for final sampling (here 100000). Here, the program runs the RJ-MCMC algorithm.
- number of iterations used to thin the chain (here, every 50th iteration is only kept). It is advisable to keep this number >20.

The second line contains:
- binary indicator equal 1 if single-locus outcrossing model is assumed and 0 for the multi-locus model (here 1)
- binary indicator equal 1 if genotyping errors are treated as estimable parametes and 0 if they are fixed at initial values (here 1)
- binary indicator equal 1 if maternal genotypes are treated as estimable parameters and and 0 if they are fixed along MCMC (in this case, maternal genotypes must be provided in a data file) (here 1)
- binary indicator equal 1 if the null regression model is assumed (in this case effects of ecological variables are fixed at zeros and a constant is only estimated)
- binary indicator equal 1 if ecological variables are to be standardized in the flow, and 0 if they are to be treated as provided in the datafile
- number of cores/threads to be used for computations (as integer). Note that multi-threading attempts to use the requested number of cores in order to speed-up the analysis. It would be especially useful when non-zero genotyping errors are set.

The third line contains:
- initial value for outcrossing rates (here 0.5; same across families)
- initial value for population divergence rates (here 0.01; same across locations)
- initial value for allele dropout (here 0.025; same across markers)
- initial value for allele mistyping (here 0.01; same across markers)

The fourth line contains a single integer number to initialise a random number generator (here 23125). This value should vary for replicated runs.

The fifth line contains a data file name (here MyData.txt). If data file is not located in a directory of quedo.exe file, the full path is needed. Note that the string of characters must have length <256.

The sixth line contains the prefix of a name for output files (here MyData). The full path is needed in order to save output files in a specific directory. The string of characters must have a length <256.

5. OUTPUT FILES

By default, quedo generates two output files: *.sum and *.run, where * is replaced by a prefix given in the info file.

*.sum is a summary file that contains some basic information about the settings as well as summary output, including acceptance rates for estimated parameters.

*.run is a main output file, where sequences of MCMC samples for parameters are saved. If a non-zero number of initial and pilot cycles is set, the *.run can be used without any "burn-in" treatment in order to compute posterior means as well as credible intervals.

Description of parameter symbols used in the *.run file:

Symbol	Description
It	number of MCMC cycle (iteration)
LogL	log-likelihood of the mating model in the It-th iteration
b_0	constant term in the regression model in the It-th iteration
b_w	slope for for the w-th variable in the It-th iteration
y_t	dispersion of outcrossing rates around a regression line
m_F	mean of Gamma prior distribution for divergence rates (F_k)
k_F	shape of Gamma prior distribution for divergence rates
t_i	outcrossing rate for the i-th family
F_k	divergence rate for the k-th location
m_e1	mean of Beta prior distribution for allele dropout
y_e1	dispersion of Beta prior distribution for allele dropout
e1_l	rate of allele dropout for the l-th marker
m_e2	mean of Beta prior distribution for allele misclassification
y_e2	dispersion of Beta prior distribution for allele misclassification
e2_l	rate of allele misclassification for the l-th marker

When maternal genotypes are set to be inferred during MCMC algorithm, quedo creates a subfolder where a number of files with the inferred maternal genotypes are written. File names for maternal genotypes are generated following the rule:

*_n.mom

where * is the output prefix specified in the info file and n is the integer indicator number of a given maternal individual.

When error rates are set to zero, quedo performs a review of parent-offspring genetic compatibility and removes progeny genotypes that are incompatible with a maternal genotype. The results of data review are then written to *.rev file, where * is the output prefix.

6. ANALYSIS OF THE EXAMPLE DATA

quedo is provided with an example data that were generated in computer simulations. The expected values of parameters are given in the data file, just below data entries. Here, it is worth to mention that the variable Z_2 and Z_3 has a positive and negative effect on outcrossing, respectively, and that Z_2 acts as a factor of individual outcrossing while Z_3 acts as a factor of a location-level outcrossing.

Here, the example run is described. The number of init. and pilot runs was set to 10,000. The number of MCMC cycles for final sampling were set to 100,000. The mating-model was set to single-locus outcrossing. Genotyping errors were not estimated and set to zero. Also, maternal genotypes were not inferred. The results of the example MCMC run are given below.

The frequency of the regression model M in the generated *.run output file is the estimate of the posterior probability of regression models (Pr(M)). Here, the model 6, contaning two variables (2nd and 3rd) is characterised by the highest probability of 0.87. The second best model has the posterior probability of 0.05.

```
Posterior proabilities of regrssion models (M):
-----------------------------------------------
M       Structure       Pr(M)
-----------------------------------------------
6       { 01100 }       0.8675
7       { 11100 }       0.0485
14      { 01110 }       0.0330
22      { 01101 }       0.0320
4       { 00100 }       0.0075
23      { 11101 }       0.0030
5       { 10100 }       0.0025
12      { 00110 }       0.0015
15      { 11110 }       0.0015
20      { 00101 }       0.0010
3       { 11000 }       0.0005
21      { 10101 }       0.0005
30      { 01111 }       0.0005
13      { 10110 }       0.0005
-----------------------------------------------
```

Using a series of model indicators in the output file, it is possible to estimate the posterior probability for each variable to be in the model. Model indicators are coded as an interger representation of a binary system. Let y_w denote a binary indicator for the w-th variable equal 1 if the variable is in the model, and 0 otherwise. Then,

M = y_1*2^0 + y_2*2^1 + ... + y_W*2^(W-1),

where W is the total number of variables. The backward translation of M into a series of y_w indicators follows a rule:

y_w|M = (M div 2^(w-1)) mod 2,

where y_w|M is a conditional indicator for a given model M, div is an integer division and mod is a remider after division (modulo operation). Then, the probability of the w-th variable to be in a model is

Pr(Z_w) = Pr(M=1)*y_w|1 + Pr(M=2)*y_w|2 + Pr(M=3)*y_w|3 + ...

The resulting probabilities are shown below. Out of 5 variables, two have remarkably high posterior probabilities.

```
Posterior proabilities of variables (Z):
-----------------------------------------------
w       Pr(Z_w)
-----------------------------------------------
1       0.0570
2       0.9865
3       0.9995
4       0.0370
5       0.0370
-----------------------------------------------
```

Estimates of parameters were computed based on a subsequence of values for the highest probability model (6). Results are shown below. In every case, limits of the highest posterior density interval (HPD-l and HPD-h) were computed as the shortest interval that contains 95% MCMC samples. Mode was estimated using the half-sample method by Bickel and Frühwirth (Comput. Stat. Data Anal. 50: 3500-3530).

Because the highest-probability model includes two variables, only b_2 and b_3 are different from zero. As expected, in both cases, the HPD intervals do not include zero. The constant 2.53 refers to the base outcrossing of 1/(1+exp(-b_0))=0.926, i.e. to the outcrossing rate of an average individual in an average population with respect to the two significant ecological variables.

```
Regression parameters (b_w):
-----------------------------------------------
Param   Mode    Mean    Median  HPD-l   HPD-h
-----------------------------------------------
b_0     2.5267  2.5266  2.5258  2.3324  2.7415
b_1     0.0000  0.0000  0.0000  0.0000  0.0000
b_2     0.3727  0.3499  0.3532  0.1899  0.5115
b_3     -0.4755 -0.4595 -0.4605 -0.6266 -0.2781
b_4     0.0000  0.0000  0.0000  0.0000  0.0000
b_5     0.0000  0.0000  0.0000  0.0000  0.0000
-----------------------------------------------
b_0 - constant term
b_w - regression slope for w-th variable (w>0)
```

In addition, it may be interesting to compute the posterior estimates of genetic parameters such as outcrossing rates across families and divergence rates across locations.

```
Outcrossing rates (t_i) across families:
-----------------------------------------------
Param   Mode    Mean    Median  HPD-l   HPD-h
-----------------------------------------------
t_1     0.9017  0.9014  0.9028  0.8379  0.9651
t_2     0.9270  0.9264  0.9291  0.8625  0.9865
t_3     0.9513  0.9398  0.9442  0.8778  0.9954
t_4     0.9786  0.9635  0.9667  0.9229  0.9988
t_5     0.9785  0.9653  0.9689  0.9228  0.9994
t_6     0.9433  0.9407  0.9426  0.8950  0.9874
t_7     0.9571  0.9470  0.9512  0.8928  0.9981
t_8     0.9663  0.9638  0.9671  0.9244  0.9986
t_9     0.8748  0.8873  0.8879  0.8252  0.9451
t_10    0.8661  0.8529  0.8533  0.7899  0.9190
t_11    0.8317  0.8347  0.8340  0.7640  0.9016
t_12    0.8904  0.8973  0.8980  0.8384  0.9565
t_13    0.7974  0.7898  0.7907  0.7158  0.8567
t_14    0.9040  0.8934  0.8942  0.8344  0.9565
```

```
t_15    0.9351  0.9328  0.9343  0.8859  0.9809
t_16    0.8473  0.8381  0.8387  0.7740  0.9022
t_17    0.9825  0.9717  0.9741  0.9390  0.9998
t_18    0.9479  0.9459  0.9480  0.8944  0.9964
t_19    0.9575  0.9456  0.9486  0.8981  0.9966
t_20    0.9673  0.9606  0.9638  0.9173  0.9998
t_21    0.9795  0.9663  0.9702  0.9274  0.9994
t_22    0.9634  0.9590  0.9613  0.9190  0.9984
t_23    0.9560  0.9403  0.9426  0.8809  0.9934
t_24    0.9787  0.9633  0.9661  0.9232  0.9993
t_25    0.9379  0.9307  0.9334  0.8690  0.9970
t_26    0.9379  0.9347  0.9357  0.8800  0.9944
t_27    0.9988  0.9851  0.9885  0.9589  1.0000
t_28    0.9734  0.9539  0.9575  0.9033  0.9997
t_29    0.9930  0.9694  0.9740  0.9263  1.0000
t_30    0.9804  0.9636  0.9673  0.9155  0.9999
t_31    0.9717  0.9604  0.9645  0.9142  0.9999
t_32    0.9960  0.9792  0.9841  0.9446  1.0000
t_33    0.9963  0.9764  0.9803  0.9422  1.0000
t_34    0.9381  0.9269  0.9291  0.8672  0.9903
t_35    0.9803  0.9615  0.9650  0.9125  0.9997
t_36    0.9488  0.9435  0.9457  0.8909  0.9993
t_37    0.9849  0.9667  0.9710  0.9210  1.0000
t_38    0.9495  0.9386  0.9397  0.8859  0.9937
t_39    0.9992  0.9880  0.9906  0.9670  1.0000
t_40    0.9931  0.9838  0.9868  0.9569  1.0000
t_41    0.8601  0.8529  0.8550  0.7797  0.9257
t_42    0.8614  0.8548  0.8570  0.7756  0.9328
t_43    0.6729  0.6900  0.6906  0.6065  0.7807
t_44    0.8785  0.8759  0.8771  0.8099  0.9441
t_45    0.7509  0.7476  0.7486  0.6706  0.8312
t_46    0.8483  0.8540  0.8542  0.7809  0.9245
t_47    0.8162  0.8165  0.8163  0.7317  0.9060
t_48    0.8417  0.8330  0.8337  0.7562  0.9133
t_49    0.9160  0.9088  0.9106  0.8511  0.9683
t_50    0.8616  0.8594  0.8602  0.7902  0.9321
t_51    0.9637  0.9526  0.9545  0.9098  0.9942
t_52    0.9072  0.9192  0.9198  0.8590  0.9708
t_53    0.8458  0.8388  0.8399  0.7657  0.9129
t_54    0.9627  0.9553  0.9568  0.9142  0.9957
t_55    0.9581  0.9576  0.9596  0.9181  0.9965
t_56    0.8768  0.8764  0.8769  0.8151  0.9378
t_57    0.9594  0.9542  0.9569  0.9074  0.9949
t_58    0.9389  0.9250  0.9267  0.8653  0.9839
t_59    0.9750  0.9690  0.9718  0.9329  0.9996
t_60    0.9579  0.9487  0.9515  0.8983  0.9978
t_61    0.9062  0.9034  0.9046  0.8333  0.9697
t_62    0.9571  0.9396  0.9415  0.8839  0.9925
t_63    0.9916  0.9805  0.9831  0.9544  0.9999
t_64    0.9247  0.9178  0.9199  0.8511  0.9822
t_65    0.8282  0.8278  0.8286  0.7598  0.8972
t_66    0.8900  0.8783  0.8791  0.8119  0.9359
t_67    0.9365  0.9293  0.9309  0.8752  0.9804
t_68    0.9338  0.9241  0.9263  0.8690  0.9722
t_69    0.9076  0.8937  0.8954  0.8342  0.9565
t_70    0.8663  0.8605  0.8615  0.7958  0.9212
t_71    0.7838  0.7832  0.7840  0.7056  0.8484
t_72    0.8450  0.8466  0.8479  0.7804  0.9106
t_73    0.9276  0.9111  0.9121  0.8485  0.9755
```

```
t_74    0.9478  0.9362  0.9378  0.8829  0.9847
t_75    0.9443  0.9377  0.9407  0.8840  0.9970
t_76    0.9618  0.9554  0.9585  0.9112  0.9973
t_77    0.9604  0.9575  0.9596  0.9140  0.9961
t_78    0.9239  0.9264  0.9282  0.8662  0.9814
t_79    0.9708  0.9563  0.9594  0.9128  0.9962
t_80    0.9174  0.9100  0.9113  0.8511  0.9769
-----------------------------------------------
```

```
Divergence rates (F_k) across locations:
-----------------------------------------------
Param   Mode    Mean    Median  HPD-l   HPD-h
-----------------------------------------------
F_1     0.1695  0.1639  0.1636  0.1308  0.1982
F_2     0.0371  0.0388  0.0385  0.0267  0.0517
F_3     0.0433  0.0452  0.0445  0.0321  0.0595
F_4     0.1240  0.1243  0.1240  0.0983  0.1542
F_5     0.0446  0.0464  0.0458  0.0334  0.0612
F_6     0.1505  0.1494  0.1485  0.1162  0.1778
F_7     0.0428  0.0456  0.0451  0.0311  0.0590
F_8     0.0784  0.0821  0.0814  0.0609  0.1019
F_9     0.0837  0.0832  0.0826  0.0632  0.1051
F_10    0.1028  0.1087  0.1078  0.0842  0.1341
-----------------------------------------------
```

Finally, estimates of hyper-parameters informs about a variation in outcrossing and divergence rates.

```
Hyper-parameters:
-----------------------------------------------
Param   Mode    Mean    Median  HPD-l   HPD-h
-----------------------------------------------
y_t     0.0279  0.0305  0.0292  0.0147  0.0496
m_F     0.0895  0.0939  0.0919  0.0659  0.1270
k_F     3.4097  4.5819  4.2013  1.2110  8.3831
-----------------------------------------------
```

y_t - dispersion of t_i around a regression model
m_F - mean of Gamma distribution for F_k
k_F - shape of Gamma istribution for F_k

7. FINAL REMARKS

The software is provided "as is" with no warranty. Bugs or any signs of irrational behaviour can be reported to the author via e-mail: igorchy@ukw.edu.pl.