Major improvements included in the program DIYABC v2.1.0 (July 2015)

The latest version of the program (DIYABC v2.1.0) includes the following major improvements: (i) new analysis options to compute error / accuracy indicators conditionally to the observed dataset, (ii) possibility to specify a MAF (minimum allele frequency) criterion on the analyzed SNP datasets, and (iii) optimization of the simulation process of SNP datasets that include a substantial amount of missing data.

1. New analysis options to compute error / accuracy indicators conditionally to the observed dataset.

The program DIYABC allows evaluating the confidence in scenario choice and the accuracy of parameter estimation under a given scenario using simulated pseudo-observed datasets (pods), for which the true scenario ID and parameter values are known. So far such pods were drawn randomly into prior distributions for both the scenario ID and the parameter values. By doing so, we estimate global error/accuracy levels computed over the whole (and usually huge) data space defined by the prior distributions. These indicators hence actually correspond to "prior" error rates (when evaluating the confidence in scenario choice) or "prior" precision measures (when evaluating the accuracy of parameter estimation under a given scenario). The levels of error/accuracy may be substantially different depending on the location of an observed or pseudo-observed dataset in the prior data space. Indeed, some peculiar combination of parameter values may correspond to situations of strong (weak) discrimination among the compared scenarios or of accurate (inaccurate) estimation of parameter values under a given model. Aside from their use to select the best classifier and set of summary statistics, prior-based indicators are, however, poorly relevant since, for a given dataset, the only point of importance in the data space is the observed dataset itself. Computing error / accuracy indicators conditionally to the observed dataset (i.e. focusing around the observed dataset by using the posterior distributions) is hence clearly more relevant than blindly computing indicators over the whole prior data space as done so far. This is basically what DIYABC v2.1.0 proposes to do with several new analysis sub-options available within the options "Evaluate confidence in the scenario choice" and "Compute bias and precision on parameter estimations". Indeed, one can now choose to compute a "posterior" error rate (when evaluating the confidence in scenario choice) by drawing the scenario ID and parameter values of a large number of pods from the *s* simulated datasets closest to the observed dataset (i.e. the s datasets with the smallest Euclidean distance). Typically, *s* = 500 (when simulating 10,000 to 1 million datasets per compared scenario) but this number can be lowered to 100. In the same vein, one can now choose to compute "posterior" accuracy indicators (when evaluating the accuracy of parameter estimation under a given scenario) by drawing the parameter values of a large number of pods among the parameter posterior distributions estimated under a given scenario using a standard ABC procedure. Note that we found, using controlled genetics experiments, that posterior error (accuracy) measures could strongly differ from prior error (accuracy) measures, hence making a case of the significance of computing error (accuracy) measures conditionally to the observed dataset rather than blindly computing such measures over the whole prior data space (unpublished results and see Pudlo et al. 2015).

2. Possibility to specify a MAF (minimum allele frequency) criterion on the analyzed SNP datasets.

Compared to other types of molecular markers, SNP loci have low mutation rates, so that polymorphism at such loci results from a single mutation during the whole population(s) gene tree and genotypes are bi-allelic. To generate a simulated polymorphic dataset at a given SNP locus, we proceeded following the algorithm proposed by Hudson (2002) (cf –s 1 option in the program ms associated to Hudson, 2002). Briefly, the genealogy at a given locus of all genes sampled in all populations of the studied dataset is simulated until the most recent common ancestor according to coalescence theory. Then a single mutation event is put at random on one branch of the genealogy (the

branch being chosen with a probability proportional to its length relatively to the total gene tree length). This algorithm provides the simulation efficiency and speed necessary in the context of ABC, where large numbers of simulated datasets including numerous SNP loci have to be generated (Cornuet et al. 2014). Most importantly, using the Hudson's simulation algorithm is equivalent to applying a default MAF (minimum allele frequency) criterion on the simulated dataset. As a matter of fact, each locus in both the observed and simulated datasets will be characterized by the presence of at least a single copy of a variant over all genes sampled from all studied populations (i.e. pooling all genes genotyped at the locus). In DIYABC v2.1.0, it is possible to impose a different MAF criterion for each locus on the observed and simulated datasets. This MAF is computed pooling all genes genotyped over all studied population samples. For instance, the specification of a MAF equal to 5% will automatically select a subset of m loci characterized by a minimum allele frequency > 5% among the *l* locus of the observed dataset. In agreement with this, only *m* locus with a MAF>5% will be retained in a simulated dataset (simulated loci with a MAF=5% will be discarded). In practice, the instruction for a given MAF has to be indicated directly in the headline of the observed dataset. For instance, if one wants to consider only loci with a MAF equal to 5% one will write <MAF=0.05> in the headline. Writing <MAF=hudson> (or omitting to write any instruction with respect to the MAF) will bring the program to use the standard Hudson's algorithm without further selection as done so far in the previous version of DIYABC. The selection with DIYABC v2.1.0 of a subset of loci fitting a given MAF allows: (i) to remove the loci with very low level of polymorphism from the dataset and hence increase the mean level of genetic variation of both the observed and simulated datasets, without producing any bias in the analyses; and (ii) to reduce the proportion of loci for which the observed variation may corresponds to sequencing errors. In practice MAF values =10% are considered. To check for the consistency/robustness of the ABC results obtained, it may be useful to treat a SNP dataset considering different MAFs (for instance MAF=hudson, MAF=1% and MAF=5%).

3. Optimization of the simulation process of SNP datasets that include a substantial amount of missing data.

We have radically changed our way to take into account missing data for SNP datasets (i.e. missing genotypes denoted "9" in the data file). The initial way to deal with missing data turned out to be poorly efficient in term of computation time, especially when the number of SNP missing data was large which seems to be the case for many real SNP datasets. The new code we have implemented to deal with this issue is particularly efficient and makes it feasible to simulate in a reasonable time large SNP datasets including (or not) numerous missing data.

References cited

Cornuet, J-M., Pudlo, P., Veyssier, J., Dehne-Garcia A., Gautier M., Leblois R., Marin J-M, Estoup A. (2014) DIYABC v2.0: a software to make approximate Bayesian computation inferences about population history using single nucleotide polymorphism, DNA sequence and microsatellite data. *Bioinformatics*, 10.1093/bioinformatics/btt1763.

Hudson, R. (2002) Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics*, 18, 337-338.

Pudlo, P., Marin, J-M., Estoup, A., Cornuet, J-M., Gautier, M., and C.P. Robert C.P. (2014) Reliable ABC model choice via random forests. *Bioinformatics*, in revision.