DILS user manual

version: 1.0

June 16, 2020

Contents

1 Introduction **2**

2 Installation **2**
- 2.1 Singularity container . . . 2
- 2.2 Manual installation . . . 2

3 Execution **3**

4 Example **3**
- 4.1 Execute DILS in a web browser . . . 3
- 4.2 Running ABC inferences . . . 3
- 4.3 Exploring the results . . . 4

5 Data **5**
- 5.1 Input file format . . . 5
- 5.2 Data Filtering . . . 6
- 5.3 Statistics calculated from the data . . . 6
  - 5.3.1 Site frequency spectrum . . . 7
  - 5.3.2 Summary Statistics . . . 7
  - 5.3.3 Other statistics . . . 7

6 Models **8**
- 6.1 Demographic models . . . 8
- 6.2 Genomic models . . . 9
- 6.3 Model comparisons . . . 9
- 6.4 Parameters . . . 10

7 Results **11**
- 7.1 Model Comparisons . . . 12
- 7.2 Parameter Estimations . . . 12
- 7.3 Quality controls . . . 12
- 7.4 General information . . . 13
- 7.5 YAML file . . . 13

8 How to access DILS online? **13**
- 8.1 Dependencies installed by Singularity . . . 13

9 Citations **14**

10 Help and support **14**

11 References **15**

1 Introduction

DILS is a statistical analysis platform for conducting demographic inferences with linked selection from population genomic data using an Approximate Bayesian Computation framework.

It takes as input single-population or two-population datasets and performs three types of analyses in a hierarchical manner, identifying:

1. the best demographic model to study the importance of gene flow and population size change on the genetic patterns of polymorphism and divergence
2. the best genomic model to determine whether the effective size Ne and migration rate $N.m$ are heterogeneously distributed along the genome
3. loci in genomic regions most associated with barriers to gene flow

DILS will also provide parameter estimates of the best demo-genomic model, and measure the robustness of the analyses.

DILS is composed of a web interface developed in *Shiny* which will execute a workflow managed by *Snakemake*. To ensure full reproducibility and portability on any server, DILS is packaged in a singularity container freely available at https://github.com/popgenomics/DILS_web. However, the workflow can also simply be executed from the command line without going through the web interface: https://github.com/popgenomics/DILS.

You can join the DILS Google group (DILS-gg) for asking questions and getting help. It is available at https://groups.google.com/forum/#!forum/dils---demographic-inferences-with-linked-selection. If you discover a bug in DILS, please contact camille.roux@univ-lille.fr or report it to the Google group.

2 Installation

You can either install DILS by using Singularity or manually.

2.1 Singularity container

1. clone git repository

```
git clone https://github.com/popgenomics/DILS_web
```

2. move into the DILS repertory

```
cd DILS_web
```

3. build singularity image (could take ≈ 20minutes)

```
sudo singularity build DILS.sif DILS.def
```

2.2 Manual installation

Shortly here

3 Execution

This step may need the root permissions.

```
sudo singularity exec --bind DILS/:/mnt DILS.sif host=[ip adress of your server] port=[port number where shiny is reachable] nCPU=[maximum number of CPUs to use simultaneously]
```

eg on a powerful machine with 100 CPUs:

```
sudo singularity exec --bind DILS/:/mnt DILS.sif webinterface/app.R host=127.0.0.9 port=8912 nCPU=100
```

Please keep in mind that the max number of CPUs is the maximum number of CPUs **DILS will use at certain times**, but that DILS will not use 100% of the indicated number of CPUs throughout its whole run. This maximum usage will be punctual.

The computer may not appreciate a value of nCPU greater than the number of cores it has.

shiny app is now available in your web browser at

```
http://[ip adress of your server]:[port number]
```

eg:

```
http://127.0.0.9:8912/
```

But chose the IP adress and port number you want

4 Example

All example files are in the **example** sub-directory. It contains two files with an example dataset (mytilus.tar.xz; to be extracted in this case) and an example results to view (byWQxjtdnp.tar.gz).

4.1 Execute DILS in a web browser

In your terminal, from the **DILS_web directory**, you can execute the Shiny application as follows:

```
singularity exec --bind DILS/:/mnt DILS.sif webinterface/app.R host=127.0.0.9 port=8912 nCPU=100
```

However, depending on your system, it may requires a ‘sudo‘ to be run:

```
sudo singularity exec --bind DILS/:/mnt DILS.sif webinterface/app.R host=127.0.0.9 port=8912 nCPU=100
```

Then in a web brower, copy/paste the following address:
`http://127.0.0.9:8912/`

4.2 Running ABC inferences

Before getting your hands on the example files, it’s important to extract the sequences that are currently compressed for storage on GitHub. The example file is a tar.xz file only to store it on GitHub, but not to be read in DILS. DILS **only works with fasta files**. In this case here, you must first decompress the archive as follows:

```
tar -Jxvf mytilus.tar.xz
```

It will generates the following example input file:

```
mytilus.fas
```

This file can be uploaded in DILS by clicking as follows:

1. ABC
2. Upload data
3. Browse (Input file upload)

The input file (in fasta format) is fully uploaded when the box entitled **Information extracted from the uploaded file** contains quantitative values describing the uploaded dataset as well as tables describing the populations, genes and individuals.

The **ABC** feature comprises four tabs to configure the ABC analysis and a last one to run execute the analysis:

1. Upload Data
2. Data Filtering. Allows the user to define thresholds considered by DILS to filter data (on locus length, number of individuals, proportion of missing data per locus (see section 5.2).
3. Populations/Species. Configures the type of models to study (single population, two populations), which populations are studied, whether an outgroup is to be considered, whether populations are assumed to be constant or variable in two-population models (see section 6.1).
4. Prior Distributions. Defines the parameter space to explore, mutation rates and intralocus recombination rates (see section 6.4).
5. Run ABC will execute the ABC analyses

In order to run the ABC analysis, it is necessary to validate your choices of configurations by clicking on the Please check/validate your choices button at the bottom of the four configuration pages. To avoid saturating the computer server, we have limited the number of different ABC analyses that a given user can run simultaneously. So you will need to specify the Number of ABC analysis to run from the same input file (from 1 to 5), then proceed to the configuration of each analysis, and run them sequentially.

4.3 Exploring the results

Upload the archive produced by DILS in a **tar.gz** format (doesn't need to be extracted) by clicking on:

1. Results visualization
2. Upload results
3. Browse (Results to upload)

In this example, the file to upload is **mytilus.fas**.

The **Results visualization** feature comprises three tabs:

1. Upload results to upload and preview the archive produced by DILS after the ABC analysis. This archive is either sent to your email address if you use the online version of DILS (available soon) or produced on the machine on which you installed DILS. Once the archive has been uploaded here, a table will be displayed containing summarized statistics (described in section 5.3) and locus-specific inferences. These locus-specific inferences concern only the analyses for two-populations models and indicate:
   (a) the best model (linked or not to a species barrier)
   (b) the posterior probability of the best model
2. User dataset to explore the observed summary statistics from the genomic data and the results of the demographic inferences obtained by the ABC

(a) Observed summary statistics (see section 5.3.2)
   i. Summarized SFS
   ii. Polymorphism
   iii. Tajima's D
   iv. Differentiation and divergence
(b) Demographic inferences (see section 7)
   i. Multilocus model comparison
   ii. Locus specific model comparison
   iii. Estimated parameters
   iv. Goodness-of-fit test
(c) Definitions of statistics and parameters
   i. Summary statistics (names and descriptions of the used statistics describing the genetic patterns).
   ii. Model parameters (names and descriptions of the parameters defining the explored models).

3. Collaborative science to record on the web-platform where the user's inference falls within a global speciation picture of transition from gene flow to no gene flow (Roux et al. 2016).

5 Data

5.1 Input file format

DILS requires a single fasta file containing all sequences obtained from all populations/species, and for all sequenced genes or DNA fragments. Even sequences obtained from non-studied species can be included in the file. The user will specify in drop-down menus the names of the species to consider after the upload.

In the current version, there are no haplotype-based statistics being used in the inference; therefore the SNPs do not need to be phased (i.e. the association of alleles across distinct heterozygous positions in a sequence can be arbitrary). Haploid data may also be used.

The input file format is largely inspired by the output of Reads2snp, a SNP and genotype caller available at https://kimura.univ-montp2.fr/PopPhyl/. All sequences have to respect the following format, with missing data only encoded by 'N':

```
>"gene"|"species or population"|"individual"|"allele1 or allele2"
ATGTCTCGGCCAGTATTATCTACGCACGTGTTAGACACTTCNACTGGTCAGCCAGGAAATGGAATTTTCGTGGAATTATACAAA
```

Example:

```
>Hmel210004_196|chi|chi.CJ560|allele1
NNNNNNNGGCCAGTATTATCTACGCACGTGTTAGACACCTCNACTGGTCAGCCAGGAAGTGGAATTTTCGTCGAATTATACAAA
>Hmel210004_196|chi|chi.CJ560|allele2
NNNNNNNGGCCAGTATTATCTACGCACGTGTTAGACACTTCNACTGGTCAGCCAGGAAGTGGAATTTTCGTCGAATTATACAAA
>Hmel210004_196|chi|chi.CJ564|allele1
NTGTCTCGGCCAGTATTATCTACGCACGTGTTAGACACTTCNACTGGTCAGCCAGGAAGTGGAATTTTCGTCGAATTATACAAA
>Hmel210004_196|chi|chi.CJ564|allele2
NTGTCTCGGCCAGTATTATCTACGCACGTGTTAGACACTTCNACTGGTCAGCCAGGAAGTGGAATATTCGTCGAATTATACAAA
>Hmel210004_196|flo|flo.CS2338|allele1
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
>Hmel210004_196|flo|flo.CS2338|allele2
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
>Hmel210004_196|flo|flo.CS2341|allele1
ATGTCTCGGCCAGTATTATCTACGCACGTGTTAGACACTTCNACTGGTCAGCCAGGAAGTGGAATTTTCGTCGAATTATACAAA
>Hmel210004_196|ros|ros.CJ2071|allele1
ATGTCTCGGCCAGTATTATCTACGCACGTGTTAGACACCTCNACTGGTCAGCCAGGAAGTGGAATTTTCGTCGAATTATACAAA
>Hmel210004_196|ros|ros.CJ2071|allele2
ATGTCTCGGCCAGTGTTATCTACGCACGTGTTAGACACTTCNACTGGTCAGCCTGGAAGTGGAATTTTCGTCGAATTATACAAA
>Hmel210004_196|ros|ros.CJ531|allele1
ATGTCTCGGCCAGTATTATCTACGCACGTGTTAGACACCTCNACTGGTCAGCCTGGAAGTGGAATTTTCGTCGAATTATACAAA
>Hmel210004_196|num|nu_sil.MJ09-4125|allele1
ATGTCTCGGCCAGTATTATCTACGCACGTGTTAGACACCTCNACTGGTCAGCCAGGAAGTGGAATTTTCGTGGAATTATACAAA
>Hmel210004_196|num|nu_sil.MJ09-4125|allele2
```

```
ATGTCTCGGCCAGTATTATCTACACACGTGTTAGACACTTCNACTGGTCAGCCAGGAAGTGGAATTTTCGTGGAATTATACAAA
>Hmel210004_196|num|nu_sil.MJ09-4184|allele1
ATGTCTCGGCCAGTATTATCTACGCACGTGTTAGACACTTCNACTGGTCAGCCAGGAAGTGGAATTTTCGTCGAATTATACAAA
>Hmel210004_196|num|nu_sil.MJ09-4184|allele2
ATGTCTCGGCCAGTATTATCTACGCACGTGTTAGACACTTCNACTGGTCAGCCAGGAAATGGAATTTTCGTGGAATTATACAAA
>Hmel219015_26|chi|chi.CAM25091|allele1
GGAAATNNAAACTTTTGTATCAAGTGTGTTACGGCGATTTCGTCTAGAAGCTGTAACGAAGCCATCTGATCTGGTNTTCCGCAC
TGATATTATATTGCGAACTATGGGACAACCAATTTACGTTAAATTTCACANGAGAAAATAA
>Hmel219015_26|chi|chi.CAM25091|allele2
GGAAATNNAAACTTTTGTATCAAGTTTGTTACGGCGATTTCGTCTAGAAGCTGTAACGAAGCCATCTGATCTGGTNTTCCGCAC
TGATATTATATTGCGAACTATGGGACAACCAATTTACGTAAAATTTCACANGAGAAAATAA
>Hmel219015_26|chi|chi.CAM25137|allele1
GGAAATNNAAACTTTTGTATCAAGTGTGTTACGGCGATTTCGTCTAGAAGCTGTAACGAAGCCATCTGATCTGGTCTTCCGCAC
TGATATTATATTGCGAACTATGGGACAACCAATTTACGTTAAATTTCANAGGAGAAAATAA
>Hmel219015_26|chi|chi.CAM25137|allele2
GGAAATNNAAACTTTTGTATCAAGTTTGTTACGGCGATTTCGTCTAGAAGCTGTAACGAAGCCATCTGATCTGGTCTTCCGCAC
TGATATTATATTGCGAACTATGGGACAACCAATTTACGTAAAATTTCANAAGAGAAAATAA
```

Two genes are displayed in this example: Hmel210004_196 and "Hmel219015_26". There are four populations, named: "chi", "flo", "ros" and "num". Only populations whose names are specified in the Populations/species menu are considered. Two diploid individuals are sequenced for each population. For example for chi: chi.CJ560 and chi.CJ564. This number can obviously vary between populations, according to the sequencing strategy and its success.

A full example of correct input file to run ABC inferences is provided in the **sub-directory example** of the GitHub repository (file: **mytilus.fas**)

5.2 Data Filtering

Several filtering options are available:

max_N_tolerated (Float between 0.0 and 1.0): defines the maximum proportion of N/gaps in the sequence of a gene beyond which this sequence is not considered.

Lmin (Positive integer): defines the minimum number of treatable sites (monomorphic + biallelic positions) below which a sequence is removed from the analysis.

1. In a noncoding sequence: a site is an alignment of nucleotides for a single given nucleotide position, including all individuals among the species considered.
2. In a coding sequence: a site is an alignment for a given codon, comprising all the individuals among the species considered. A coding position is not considered if:
   (a) a codon alignment contains a non-synonymous polymorphism.
   (b) more than two codons segregate (even synonyms).
   (c) at least one N is found in one codon, in one individual.

nMin (Positive integer): defines the number of sequences per gene and per population/species than DILS will use. DILS starts for each gene by eliminating individual sequences containing too many N and gaps (**max_N_tolerated**). If for a gene and within a population/species, there are fewer than **nMin** sequences left, then the gene is not considered in the ABC analysis. If an outgroup is specified, then **nMin** becomes the number of sequences sampled for each species at each locus, to produce a standardized site frequency spectrum.

5.3 Statistics calculated from the data

DILS will produce different **output files** describing the observed data. It's worth noting that two-population statistics will not be calculated from single-population data. The Site Frequency Spectrum (SFS) is always used for single-population models, but it is optionally used for two-populations models. Each bin (except the first one counting the number of SNPs that are singletons) is considered as a summary-statistic to be adjusted during the model comparison and the estimation of parameters.

5.3.1 Site frequency spectrum

ABCjsfs.txt: observed site frequency spectrum. The first line is the name of each bin of the SFS (*fAX_fBY*, where X and Y correspond to the allele count in species A and B, respectively). The second line is the number of SNPs observed in each bin. A graphical representation is provided in the user interface by clicking as follows:

1. Results visualization (left-hand menu)
2. User's dataset (left-hand menu)
3. Demographic inferences (tab)
4. Goodness-of-fit test (tab)
5. SFS (tab)

5.3.2 Summary Statistics

ABCstat_global.txt: observed summary statistics for the global dataset. The average (*avg*) and standard variation (*std*) over loci is provided for each statistic, listed below.

General
dataset = index
bialsites = number of sites per locus

Summarized SFS
s_{f} = fraction of sites with a fixed difference between populations
s_{xA}, s_{xB} = fraction of sites with a polymorphism specific to each population
s_{s} = fraction of sites with a polymorphism shared between populations

Polymorphism
pi_{A}, pi_{B} = pairwise nucleotide diversity (π) for each population (Tajima, 1983)
$theta_{\mathrm{A}}$, $theta_{\mathrm{B}}$ = Watterson's θ for each population (Watterson, 1975).

Tajima's *D*
D_{TajA}, D_{TajB} = Tajima's D for each population (Tajima, 1989).

Differentiation and divergence
div_{AB} = raw divergence (D_{xy}) between populations (Nei, 1987)
$netdiv_{\mathrm{AB}}$ = net divergence (D_a) between populations, measured by $D_{xy} - (\pi_A + \pi_B)/2$ (Nei & Li, 1979)
$FST = F_{ST}$ measured by $1 - \pi_S/\pi_T$; where π_S is the average nucleotide diversity in each population and π_T is the total nucleotide diversity over the populations (Wright, 1943).

Pearson's correlation between species/populations and across loci
pearson_r_pi = between π_A and π_B
pearson_r_theta = between θ_A and θ_B
pearson_r_divAB_netDivAB = between div_{AB} and $netDiv_{AB}$
pearson_r_divAB_FST = between div_{AB} and F_{ST}
pearson_r_netDivAB_FST = between $netDiv_{AB}$ and F_{ST}

ABCstat_loci.txt: observed summary statistics for each locus. Same as ABCstat_global.txt, except that Pearson's correlations are not computed.

5.3.3 Other statistics

spA_spB_infos.txt: length and sample size for each locus.

General
locusName = locus name
L_including_N = total length
L = length without N
nSegSite = number of segregating sites
nsamA,*nsamB*= number of sequences in species A (resp.B)

results_recombination.txt: output of the *RNAseqFGT* program which searches for recombination events using the four-gamete test on unphased sequences.

General
LocusID = locus name
Species = species name
N_individuals = number of individuals sampled
AlnSeqLength = length of the sequence alignment
N_sites_noN = number of alignment sites at which there are at least two informative (i.e. non-N) sequences
AvgSeqLg_noN = average number of informative (i.e. non-N) base per sequence
N_SNPsBi= number of bi-allelic SNPs
N_SNPs_Mult= number of SNPs with more than two alleles

Features
GC_content = GC-content (average over all sequences, excluding N's)
NbRec = number of non-overlapping recombinant intervals (or NA if less than two SNPs or less than four identifiable haplotypes)
SNP_coverage = average fraction of individuals with known genotype (i.e. non-N) at bi-allelic SNP sites

6 Models

DILS uses a coalescent-based machinery (*msnsam*, Hudson 2002, Ross-Ibarra et al. 2008) to simulate demo-genomic models.

6.1 Demographic models

A demographic model specifies population sizes, migration rates and times of discrete events (i.e. time of split, time of change in population sizes, time of change in migration regime).

Two main settings are currently implemented in DILS:

If Populations/Species is set to one, then a model for a single panmictic population is evaluated. Three demographic 1-population models are implemented in DILS:

Constant = a single panmictic population of effective size *Ne* constant over time.

Expansion = the size of the current population has suddenly become larger than in the past *T_dem* generations.

Contraction = the current population experienced a decline in its size *T_dem* generations ago.

If Populations/Species is set to two, then a model of divergence between two populations/species is evaluated. Four demographic 2-population models are implemented in DILS:

SI = strict isolation: subdivision of an ancestral diploid panmictic population (of size *Na*) in two diploid populations (of constant sizes *N1* and *N2*) at time *T_split*.

AM = ancestral migration: the two newly formed populations continue to exchange alleles until time *T_AM*.

IM = isolation with migration: the two daughter populations continuously exchange alleles until present time.

SC = secondary contact: the daughter populations first evolve in isolation (forward in time), then experience a secondary contact and start exchanging alleles at time *T_SC*.

Allowing changes in population size for each sister population/species is optional in two-populations models.

If the Presence of an outgroup is set to yes, then the site frequency spectrum is unfolded (if chosen to be used), and the mutation rate for each locus is corrected by its divergence with the outgroup. Otherwise, the site frequency spectrum is folded, and the mutation rate is the same for all loci.

Note that it is **better NOT to use** an outgroup than to use a bad outgroup, i.e. one with a large amount of incomplete lineage sorting.

6.2 Genomic models

A major feature of DILS is to decouple the effect of linked selection and neutral history by relaxing the assumption that all loci share the same demography.

Therefore, all demographic models exist under two alternative genomic models concerning the effective size:

NE_homo = genomic homogeneity: the effective size *Ne* is genomically homogeneous, i.e. all locus are simulated by sharing the same *Ne* value. In this model DILS will try to estimate the value of *Ne* best explaining the observed data. *Ne* being independently estimated in all populations (current, past).

NE_hetero = genomic heterogeneity: the effective size *Ne* is genomically heterogeneous, i.e. all locus are simulated with a value of *Ne* drawn in a *Beta* distribution. We model variation in the rate of drift among loci to account for background selection (due to purifying selection) and adaptive sweeps. In this model DILS will try to estimate the value of *Ne* as well as the two shape parameters α and β that best explain the observations. DILS assumes that all populations (current and past) share the same *Beta* distribution but are independently re-scaled by different *Ne* values.

In addition, all demographic models with migration have two alternative models of introgression:

M_homo = genomic homogeneity: all loci share the same introgression rate for a given direction. DILS will simply try to independently estimate the introgression rate of each direction.

M_hetero = genomic heterogeneity: introgression rates vary throughout the genome. We model variation in migration rates among loci to capture the effect of selection against migrants at neutral markers linked to species barriers. This variation can follow either a *Beta* distribution or a *Bimodal* distribution. The *Beta* model is set by using two shape parameters (α and β); while the *Bimodal* model requires a number of selected loci and a proportion of barriers (in percent).

6.3 Model comparisons

DILS will perform the following hierarchical comparisons for 1-population models:

1) comparison between all models with Expansion (*Ne_homo*; *Ne_hetero*) *vs* Constant size (*Ne_homo*; *Ne_hetero*) *vs* Contraction (*Ne_homo*; *Ne_hetero*).

2) determine whether effective size *Ne* is homogeneously or heterogeneously distributed in genomes for the best supported model in step 1.

DILS will perform the following hierarchical comparisons for 2-population models:

1) comparison between all models with current isolation ([SI; AM] x [*Ne_homo*; *Ne_hetero*] x [*M_homo*; *M_hetero*]) *vs* ongoing migration ([IM; SC] x [*Ne_homo*; *Ne_hetero*] x [*M_homo*; *M_hetero*]).

2a) if current isolation: comparison between SI ([*Ne_homo*; *Ne_hetero*]) *vs* AM ([*Ne_homo*; *Ne_hetero*] x [*M_homo*; *M_hetero*]).

2b) if ongoing migration: comparison between IM ([*Ne_homo*; *Ne_hetero*] x [*M_homo*; *M_hetero*]) *vs* SC ([*Ne_homo*; *Ne_hetero*] x [*M_homo*; *M_hetero*]).

3) the last step is to determine whether effective size (Ne) and migration rates ($N.m$) are homogeneously or heterogeneously distributed in genomes.

6.4 Parameters

Parameters in DILS are expressed in demographic units: 1) times (T) are given in number of generations; 2) population sizes (N) are given in number of diploid individuals; 3) migration rates (i.e. the number of migrants per generation, M) are given in units of $4.N.m$. If your fits are often at the prior bounds of one or more parameters, this indicates a problem: 1) it may be that your bounds are too conservative, so try widening them, 2) it may also be that the model does not account for biases in your data.

Mutation and recombination rates:

The average mutation rate, mu (to be defined by the user, default value: 0.000000003), is the probability per generation and per nucleotide that an allele will not be properly replicated. If an external group is not specified then all genes/contigs/locus share the same μ. If an external group is specified then the local μ, for a locus i is corrected by μ * *div_i* / *div_avg* where *div_i* is the local divergence between the ingroup and the outgroup at that locus, and *div_avg* is the divergence averaged over all loci.

The *rho_over_theta* ratio (to be defined by the user, default value: 0.1) is the ratio of recombination ($\rho = 4.N.r$ in /bp /generation) over mutation ($\theta = 4.N.\mu$ in /bp /generation). Its exact value does not matter as long as it is greater than zero (because there are no haplotype-based statistics being used in the inference).

Population sizes:

The population size Ne is the number of diploid individuals within current and ancestral populations. The prior distribution is uniform between *N_min* (default value: 100) and *N_max* (default value: 1000000), and they are drawn from the prior distributions independently set for current and ancestral populations.

If Ne is constant through time: if a single population is modeled, its size is simply estimated as *N*. If two populations are modeled, three sizes are estimated: the ancestral *Na*, the current populations 1 (*N1*) and 2 (*N2*)

If Ne if variable through time: if a single population is modeled, two sizes are estimated: one before (*Na*) and one after (*N*) the time of demographic change. If two populations are modeled, each of the two daughter populations has two distinct sizes corresponding to two epochs: 1) one between *T_split* and *T_dem* generations ago (*N_founders1* and *N_founders2*); and 2) a current one for the last *T_dem* generations (*N1* and *N2*). This makes a total of five population sizes (including *Na*). The values of *T_dem* are independent between populations.

If Ne is genomically heterogeneous: all locus are simulated with a value of Ne drawn in a *Beta* distribution with two shape parameters drawn from uniform distributions: shape_N_α and shape_N_β (default values: min=1 and max=20). All populations (current and past) share the same *Beta* distribution but are independently re-scaled by different *Ne* values drawn from uniform distributions (default values: min=0 and max=1000000).

Times:
The speciation time *T_split* is the time since the two populations/species split from each others. It is expressed in number of generations. For annual organisms: one generation = one year. For perennial organisms: one generation = average age for an individual to transmit a descendant (which is different from the age of sexual maturity). The prior distribution is uniform between *T_split_min* (default value: 100) and *T_split_max* (default value: 1000000).

For each simulation in the SC and AM models, the time of secondary contact between populations/species (*T_SC*) and and arrest of ancient migration (*T_AM*) are drawn uniformly between *T_split_min* and the sampled *T_split*.

The time of demographic change *T_dem* represents the number of generations since population expansion or contraction. The prior distribution is uniform between *T_dem_min* and *T_dem_max*.

Migration:
Migration rates are expressed in units $4.N.m$ where m is the fraction of each population made up of new migrants each generation. The prior distribution is uniform between *M_min* (default value: 0.4) and *M_max* (default value: 20).

If M is genomically homogeneous: all loci share the same introgression rate for a given direction. Two introgression rates are estimated: 1) one from population 1 to 2 (*M12*), and 2) one in the opposite direction (*M21*).

If M is genomically heterogeneous (beta model): introgression rates vary throughout the genome and follow a *Beta* distribution with two shape parameters drawn from uniform distributions: shape_M_α and shape_M_β (default values: min=1 and max=20). Both directions of introgression share the same *Beta* distribution but are independently re-scaled by different *M* values drawn from uniform distributions (default values: min=0 and max=20).

If M is genomically heterogeneous (bimodal model): introgression rates vary throughout the genome and follow a bimodal distribution. This model is defined by 1) two introgression rates (one in each direction) drawn from uniform distributions (default values: min=0 and max=20), and 2) the proportion of barriers for which M=0 (in percent, default values: min=0 and max=100). Combined with the number of studied loci, this gives the number of loci in genomic regions most associated with barriers to gene flow (*nBarriers*).

7 Results

The raw data can be easily visualized with DILS as a site frequency spectrum and summary statistics across loci. DILS produces comprehensive result files for each inference step:

1. the global model comparison to estimate the best demo-genomic model
2. the locus-specific model comparison to identify barrier loci
3. the estimation of parameter values for the best model

To help users interpreting these results and assessing the robustness of inferences, DILS produces a series of goodness-of-fit tests to the data. These goodness of fit tests are performed both on population genetics statistics and on (the bins of) SFS:

1. Results visualization (left-hand menu)
2. User's dataset (left-hand menu)
3. Demographic inferences (tab)
4. Goodness-of-fit test (tab)

7.1 Model Comparisons

Global model:

modelComp/hierarchical_models.txt: step-by-step comparison that leads to the best model. A complete report is provided in report_spA_spB.txt.

Locus-specific model:

locus_modelComp/locus_specific_modelComp.txt: for each locus, it gives the value of each summary statistic, as well as its posterior probability (*post_proba*) to be a barrier (*allocation*=isolation) or not (*allocation*=migration).

7.2 Parameter Estimations

Note that the files in the best_model directory use expectations based on the *posterior*, while files in the best_model_5 directory are based on the *optimized posterior*.

posterior_bestModel.txt: posterior parameters of the best model using a neural network procedure. A graphical representation is provided in posterior_bestModel.pdf

posterior_summary_RandomForest_bestModel.txt: posterior parameters of the best model using a random forest procedure.

priorfile.txt: sample of 10,000 prior parameters.

7.3 Quality controls

DILS is transparent on the ability of its inferences to reproduce the observed data or not. Therefore goodness-of-fit tests are performed by simulating under the best model each population genetic statistic listed in subsection "Statistics calculated from the data" (genomic mean and variance of π, θ, F_{ST}, etc.), as well as for each bin of the SFS.

In addition to an individual test for each summary statistic, a test is also performed from statistics transformed by a PCA. DILS also provides values for each locus of:

1. each summary statistic
2. the approximated recombination rate calculated based on the four-gamete rule (Galtier et al. 2018)
3. the posterior probability of being genetically linked to a barrier to gene flow (for two-population models only)

PCA

distribution_PCA.txt: input for the PCA containing the summary statistics and SFS for the observed dataset, 2000 prior simulations, 2000 posterior simulations and 2000 optimized posterior simulations.

table_contrib_PCA_SS: contribution of each statistic to the three first principal components (*Dim.1*; *Dim.2*; *Dim.3*).

table_coord_PCA_SS: coordinates of each dataset (*origin*) on the three first principal components (*Dim.1*; *Dim.2*; *Dim.3*).

table_eigenvalues_PCA_SS: features of each principal component (*eigenvalue*; *percentage of variance*; *cumulative percentage of variance*).

Note that the files in the gof directory use expectations based on the *posterior*, while files in the gof_2 directory are based on the *optimized posterior*.

GOF_SFS

gof_sfs.txt: goodness-of-fit for each cell of the SFS. The first line is the name of each bin of the SFS (*fAX_fBY*, where X and Y correspond to the allele count in species A and B, respectively). The second line is the observed SFS, i.e. the number of SNPs observed in each bin. The third class is the predicted SFS under the best model, i.e. the number of SNPs expected in each bin. The fourth line corresponds to the p-value for each bin when comparing the expected and observed SFS.

GOF_summary_statistics

goodness_of_fit_test.txt: goodness-of-fit for each summary statistic (*stats*= summary statistic; *mean_exp* = expected mean; *mean_obs*= observed mean; *pvals_fdr_corrected*= p-value corrected for multiple testing).

7.4 General information

general_infos.txt: information related to the user's dataset.
Nref.txt: effective size of the reference population in the coalescent simulations (arbitrarily fixed); and used to convert in demographic units.
config.yaml: settings of the DILS inference (see next section 7.5).

7.5 YAML file

The ABC analysis executed with DILS consists of a Snakemake pipeline which will have as its only input file a **configuration file in yaml format**, generated by the DILS GUI. This analysis will take care of the **input file in fasta** format uploaded by the user. Essentially, an ABC analysis needs two input files to be performed:

1. the fasta file containing the sequences (provided by the user, see section 5.1)
2. the configuration file in yaml format (generated by the GUI; not manipulated by the user)

To encourage customization of DILS by different users, the **.yaml** file contains the following information:

mail_address: your email address used for the collaborative project.
infile: pathway to the fasta file.
region: type of genomic region sequenced [coding or noncoding].
nspecies: number of species/populations to analyse [1 or 2].
nameA: name of species/population 1 [nameA].
nameB: name of species/population 2 [nameB].
nameOutgroup: name of the outgroup [nameO]; if absent [NA].
useSFS: use the SFS in the 2-population inference [1]; do not use it [0].
config_yaml: pathway to this yaml file.
timeStamp: name of the output directory [DILS_output].
population_growth: temporal variation in the population size [constant or variable].
modeBarrier: type of modelling for the genomically heterogeneous Me [bimodal or beta].
max_N_tolerated: maximum proportion of N/gaps allowed in the sequence of a gene [0.2].
Lmin: minimum number of treatable sites allowed in the sequence of a gene [100].
nMin: minimum number of sequences allowed per gene and per population/species [6].
mu: mutation rate per site per generation [0.00000002763].
rho_over_theta: the ratio of recombination over mutation [0.5].
N_min: minimum prior number of diploid individuals [0].
N_max: maximum prior number of diploid individuals [500000].
Tsplit_min: minimum prior number of generations for times [0].
Tsplit_max: maximum prior number of generations for times [1750000].
M_min: minimum prior number of migrants per generation [1].
M_max: maximum prior number of migrants per generation [40].

8 How to access DILS online?

The easiest way to use DILS is *via* the online platform hosted by the CC LBBE/PRABI (France). **Website coming soon**

8.1 Dependencies installed by Singularity

Snakemake:

- Snakemake, version 5.10

Python/PyPy:

- Python, version 3.6
- PyPy, version 7.3
- NumPy, version 1.18

R:

- R, version 3.6
- abcrf
- data.table
- FactoMineR
- ggplot2
- ggpubr
- nnet
- plotly
- tidyverse
- viridis

C programs:

- msnsam
- RNAseqFGT

9 Citations

DILS : Demographic Inferences with Linked Selection by using ABC
Christelle Fraisse, Iva Popovic, Jonathan Romiguier, Etienne Loire, Alexis Simon, Nicolas Galtier, Laurent Duret, Nicolas Bierne, Xavier Vekemans, Camille Roux
doi: https://doi.org/10.1101/2020.06.15.151597

Please, if you use this online version of DILS, do not forget to recognize and acknowledge the free provision of calculation cores by CC LBBE/PRABI: "The demographic inferences were conducted on the LBBE/PRABI cluster".

10 Help and support

For problems of installation, use or interpretation, do not hesitate to post a message on the following group:
https://groups.google.com/forum/#!forum/dils---demographic-inferences-with-linked-selection

11 References

- Galtier, N., Roux, C., Rousselle, M., Romiguier, J., Figuet, E., Glémin, S., Bierne, N. Duret, L. (2018). Codon usage bias in animals: disentangling the effects of natural selection, effective population size, and GC-biased gene conversion. *Mol. Biol. Evol.*, 35(5): 1092-1103.
- Hudson, R. (2002). Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics*, 18: 337-338
- Nei, M. (1987). Molecular Evolutionary Genetics. Columbia University Press, New York.
- Nei, M. & Li, W-H. (1979). Mathematical model for studying genetic variation in terms of restriction endonucleases. *PNAS*, 76: 5269–5273.
- Ross-Ibarra J, Wright SI, Foxe JP, Kawabe A, DeRose-Wilson L, et al. (2008). Patterns of Polymorphism and Demographic History in Natural Populations of Arabidopsis lyrata . *PLoS ONE* 3(6): e2411
- Roux, C., Fraisse, C., Romiguier, J., Anciaux, Y., Galtier, N., and Bierne N. (2016) Shedding light on the grey zone of speciation along a continuum of genomic divergence. *PLoS Biol.*, 14(12): e2000234.
- Tajima, F. (1989). The effect of change in population size on DNA polymorphism. *Genetics*, 123(3): 597-601.
- Tajima, F. (1983). Evolutionary relationship of DNA sequences in finite populations. *Genetics*, 105(2): 437-460.
- Watterson, G. A. (1975). On the number of segregating sites in genetical models without recombination. *Theor. Popul. Biol.*, 7(2): 256-276.
- Wright, S. (1943). Isolation by distance. *Genetics*, 28: 114–138.