Physalia-course: Population genomics

Thibault Leroy (thibault.leroy@inrae.fr)

Part 3: Introduction to demographic modelling
November 27, 2024

General context & dataset

Demographic modelling is a powerful approach to reconstruct the evolutionary history of populations, such as changes in population size, migration patterns, and divergence times. Leveraging large-scale genomic datasets enhances the accuracy and resolution of such analyses, providing insights into complex population dynamics. However, it is important to remember that these analyses rely on simplified representations of reality. Consequently, results from demographic modelling should be interpreted with caution. As George Box famously noted, "*All models are wrong, but some are useful. However, the approximate nature of the model must always be borne in mind.* Applied to demographic modelling with population genomics data, this means that no model can fully capture the exact real demographic history. The goal is to approximate reality as closely as possible, providing a useful framework for understanding evolutionary processes, therefore providing a window into the past.

The practical implementation of demographic modelling is also highly challenging due to the computational demands of handling extensive datasets. Many tools require simulation-based approaches, which can take days or weeks to complete for realistic parameter ranges. Moreover, demographic modelling is time-intensive for researchers, demanding careful parameter selection, algorithmic understanding, and result interpretation.

In this practical, due to all these constraints, the objective will be to provide an overview of the key steps in demographic modelling using large genomic data. To address the constraints of time and computation, we will mostly evaluate the

results of some inferences.

Regarding the data, we will first use a subset of the same dataset than in yesterday's practical (Flowers et al. 2019 PNAS), which details genomic data from date palms, ensuring continuity in our analysis. It is indeed very important to understand a solid understanding of the population structure within the dataset is crucial before. Exploring population structure through principal component analysis (PCA), Bayesian clustering methods (e.g., STRUCTURE, ADMIXTURE), and other population genetics approaches are essential to infer genetic boundaries and identify distinct genetic units within the data. Demographic models indeed make assumptions (e.g. panmixia within populations), which therefore require accurate definitions of genetic populations to perform meaningful demographic inferences. Therefore, performing these initial population structure analyses is critical before beginning demographic modelling.

Before to start, copy the content of the repository:
`cp -r ~/Share/Day3_demography/ ~`

Briefly explore the VCF file and determine how many high-quality SNPs are present in this dataset.
Remember, you can examine a compressed file without fully uncompressing it by using:
```
cd ~/Day3_demography/data-datepalm/Calling/
zmore Flowers_et_al_2019.SNPs.tronq.vcf.gz
```
OPTIONAL: Can you estimate a SNP density (# high quality SNPs / reference genome size)? See Monday's practical. Do you see some limits regarding the accuracy of this SNP density estimate?

Step 1: identify non-admixed individuals

When performing demographic modeling, it is crucial to filter out individuals with evidence of recent admixture. Recent patterns of gene flow can mask more ancient demographic signals, making it difficult to accurately infer historical population dynamics. By removing most admixed individuals (e.g. Q-values $>$ 0.9), we ensure that the analysis focuses on distinct genetic units, reflecting the deeper evolutionary history rather than being confounded by recent events.

First, we will identify the individuals with the most unambiguous genetic assignments using the file samples_Qvalues.txt. This file provides Q-values for each individual, representing their proportional membership in different

genetic clusters. To ensure robust analyses, we will retain only individuals with Q-values greater than 0.9, indicating clear assignment to a single cluster.

```
cd ~/Day3_demography/data-datepalm/
grep "P._dactylifera_ME" samples_Qvalues.txt | awk '$4 > 0.90
{print $0}' >  samples_Qvalues.PdactyliferaME.txt
grep "P._dactylifera_NAfr" samples_Qvalues.txt | awk '$4 > 0.90
{print $0}' >  samples_Qvalues.PdactyliferaNAfr.txt
grep "P._theophrasti" samples_Qvalues.txt | awk '$5 > 0.90 {print
$0}' >  samples_Qvalues.Ptheophrasti.txt
```

Print only the name of the individuals

```
awk '{print $1}' samples_Qvalues.PdactyliferaME.txt > samples_Qvalues.PdactyliferaME.list
awk '{print $1}' samples_Qvalues.PdactyliferaNAfr.txt >
samples_Qvalues.PdactyliferaNAfr.list
awk '{print $1}' samples_Qvalues.Ptheophrasti.txt > samples_Qvalues.Ptheophrasti.list
```

How many individuals are identified as "relatively pure" in each population or species based on this criterion?

Step 2: Perform inferences of N_e with SMC++

Many softwares are available to perform demographic inferences of N_e, including PSMC, MSMC, poolsizeABC. Here we will use the Sequentially Markovian Coalescent (SMC++), a powerful tool for reconstructing historical population sizes using whole-genome data from multiple individuals. Unlike PSMC, SMC++ can incorporate multiple genomes, increasing resolution and robustness in demographic inference. Even if MSMC can combines multiple diploid genomes to jointly infer demographic history, smc++ was specifically designed to handle large number of genomes efficiently (less computationally intensive and more efficient as the genomes increases). Preparing input file can be a bit time-consuming (vcf2smc command), but the script below can help you.

IMPORTANT: The commands provided below have the objective to give you an idea of the expected scripts.
1/ Please again, review the code carefully to understand the rationale, but DO NOT EXECUTE it on the Physalia cluster today! To prevent accidental execution, I have deliberately chosen not to directly provide the exact input file ;) Please only execute the code from the section corresponding to the "smc++ plot (...)", see below (from line starting with conda activate...).
2/ please DO NOT OVERINTERPRET the results of the inferences that will be shown in this section, since these resulrs will be only baed on genetic

data from a few scaffolds, without masking regions of the genome (masking is recommended, see also below)

```
vcftools --vcf ./Calling/Flowers_et_al_2019.SNPs.vcf \
--remove-indels \
--max-alleles 2 \
--min-alleles 2 \
--recode \
--out ./Calling/Flowers_et_al_2019.SNPs.bialleliconly.vcf
```

Extract the 3 names of the 3 first scaffolds in the vcf (we focus on the first three to speed up the computations, but in general we work on all scaffolds (at least those > 100kb))

```
awk '{print $1}' ./Calling/Flowers_et_al_2019.SNPs.bialleliconly.vcf
| grep -v "#" | uniq | head -3 >
./Calling/Flowers_et_al_2019.SNPs.bialleliconly.vcf.top3scaffoldsID
```

Try to understand the command above, what is perform at each step (i.e. from each side of the pipe).

Ok so now, the objective will be to create the input file for smc++ for each of the three scaffolds and for each of the three populations composed by the individuals with Q-values > 0.9. Given that the inputfiles will be generated in different directories for each population, so we have to first create the directories with mkdir.

```
mkdir ./Calling/PdactyliferaMe ./Calling/PdactyliferaNAfr ./Calling/Ptheophrasti
grep "#" ./Calling/Flowers_et_al_2019.SNPs.bialleliconly.vcf >
./Calling/Flowers_et_al_2019.SNPs.bialleliconly.vcf.header while
read line; do
grep "$line" ./Calling/Flowers_et_al_2019.SNPs.bialleliconly.vcf
| grep -v "#"> ./Calling/Flowers_et_al_2019.SNPs.bialleliconly.$line.vcf.1
cat ./Calling/Flowers_et_al_2019.SNPs.bialleliconly.vcf.header
./Calling/Flowers_et_al_2019.SNPs.bialleliconly.$line.vcf.1 >
./Calling/Flowers_et_al_2019.SNPs.bialleliconly.$line.vcf
rm ./Calling/Flowers_et_al_2019.SNPs.bialleliconly.$line.vcf.1
rm ./Calling/Flowers_et_al_2019.SNPs.bialleliconly.$line.vcf.gz
bgzip ./Calling/Flowers_et_al_2019.SNPs.bialleliconly.$line.vcf
tabix ./Calling/Flowers_et_al_2019.SNPs.bialleliconly.$line.vcf.gz

smc++ vcf2smc --missing-cutoff 500
-v ./Calling/Flowers_et_al_2019.SNPs.bialleliconly.$line.vcf.gz
./Calling/PdactyliferaMe/Flowers_et_al_2019.SNPs.bialleliconly.smc.PdactyliferaMe.$line.gz
$line PdactyliferaME:Abouman,Ajwa,Amir_haj,Azraq_azraq,Began,Braim,Chichi,
Dajwani,Dedhi,Dibbas,Ebrahimi,Ewent_ayob,Fard4,Faslee,Gajar,Halawy,Hawawiri,
Helwa,Hilali,Hiri,Kabkab,Karbali,Kashoowari,Khadrawy,Khastawi,Khenezi,Khisab,
Kuproo,Lulu,Maktoumi,Manjouma,Mazafati,Nagal,Naquel_khuh,Nebeit_seif,Otaquin,
```

```
Piavom,Rothan,Shagri,Silani,Sultana,Um_al_blaliz,Um_al_hamam,Zahidi,Rabee

smc++ vcf2smc --missing-cutoff 500
-v ./Calling/Flowers_et_al_2019.SNPs.bialleliconly.$line.vcf.gz
./Calling/PdactyliferaNAfr/
Flowers_et_al_2019.SNPs.bialleliconly.smc.PdactyliferaNAfr.$line.gz
$line PdactyliferaNAfr:Hayany,Saidi,Samany,Hamria,Zagloul,Barmel,Atlantica_CAP1_POPMAL1,
Kamla,Jihl

smc++ vcf2smc --missing-cutoff 500 -v
./Calling/Flowers_et_al_2019.SNPs.bialleliconly.$line.vcf.gz
./Calling/Ptheophrasti/
Flowers_et_al_2019.SNPs.bialleliconly.smc.Ptheophrasti.$line.gz
$line Ptheophrasti:Theophrasti_02a,Theophrasti_05a,Theophrasti_A1,Theophrasti_A5,
Theophrasti_B1,Theophrasti_B3,Theophrasti_B5,Theophrasti_C1,Theophrasti_C4,
Theophrasti_D1,Theophrasti_D3,Theophrasti_D5,Theophrasti_E2,Theophrasti_F1,
Theophrasti_F2,Theophrastis_THE83_91051

done < ./Calling/Flowers_et_al_2019.SNPs.bialleliconly.vcf.top3scaffoldsID
```

Again, on the code above, we used "--missing-cutoff" with a quite arbitrary value but probably very conservative value (500), but please rather use a masked version of your reference genome on your own analysis. This will mask regions prone to biases, such as regions of high repetitive content, and therefore focus on reliable regions of the genome, masking ensures more accurate inference of effective population size.

Now it's time to use smc++ estimate, a key step for demographic inference. This command estimates effective population size changes over time using coalescent patterns inferred from the data.

```
cd Calling/
for pop in PdactyliferaMe PdactyliferaNAfr Ptheophrasti; do
inputdir=$pop
outdir="res_SMCpp_missing50kb"
mkdir -p "$outdir"/"$pop"
smc++ estimate -o ./"$outdir"/"$pop" 2.5e-08 $inputdir/*.gz
done
```

One critical aspect of inferring effective population size (N_e) is translating coalescent units into real time using mutation rates and generation time. This conversion allows estimates to be expressed in terms of generations and years. While coalescent rates are generally well inferred with sufficient data—such as whole-genome sequences of moderate coverage for tens of individuals and low reference genome fragmentation (typically a chromosome-scale reference). Of course, this is not our case since we use only 3 scaffolds to speed up the

computations.
A major challenge in the community lies in the limited knowledge of mutation rates and generation times in natural populations. Many researchers still default to using the mutation rate of Arabidopsis for plants or that of humans for animals, despite evidence of substantial variability across species (e.g. Bergeron et al., 2023, Nature). This generalized approach can lead to inaccuracies, as mutation rates are highly context-specific. This knowledge gap emphasizes the importance of caution when interpreting Ne estimates, as uncertainties in mutation rates and generation times can significantly impact the results. Further research into species-specific mutation rates and life histories is essential for improving the accuracy of demographic inferences. As you can see, here I used a mutation rate of 2.5e-8 (see the value in the command above) and a generation time of 10 years (see below) as assumed by Gros-Balthazard et al. 2017 (Current biology).

Until now, due to the computational constraints, it was impossible for your launch the commands, but now you will be able to use a smc++ function to generate the final output files! To do so, use the following command:

```
conda activate Workshop_TL_YB_Demography
cd /home/user2/Day3_demography/data-datepalm/Res_SMCpp

smc++ plot -g 10 -c physalia_smcpp_plot.pdf PdactyliferaMe/model.final.json
PdactyliferaNAfr/model.final.json Ptheophrasti/model.final.json
```

The command smc++ plot generates a demographic history plot from SMC++ results, providing a visual representation of effective population size (N_e) over time. We can refine a bit the output using R. **The results are available on the following directory:** `cd ./Day3_demography/data-datepalm/Res_SMCpp/`

Then you can inspect and start the R script:

```
less ~/Day3_demography/data-datepalm/Res_SMCpp/script_generate_smcpp.R
Rscript script_generate_smcpp.R
```

This R script is expected to generate a PDF file titled "Rplot_ggplot2_smcpp_missingcutoff500.pdf", illustrating the variation in inferred effective population size (N_e) over time.

Modify the R script to perform the same analysis using an inference which assumed a different –missing-cutoff option (namely 50000). The models to use for the smc++ plot function are available on the following directory: ~/Day3_demography/data-datepalm/Res_SMCpp/res_SMCpp_missing50kb/. Compare the two results. Are there significant differences? If so, identify the population(s) where these differences occur.
OPTIONAL: Create a single plot combining the results from both inferences to better visualize the differences.

Again, remember that the inferences are based on a restricted dataset, do not interpret here the results biologically. More broadly, always interpret the result of demographic inferences critically, especially those thare are based on inferences of N_e only, since these methods make a series of assumptions (e.g. panmixia). Here, we recommend first using data from the entire genome and applying a mask to exclude potentially problematic regions.

Step 3: Perform inferences of recent N_e

As you have observed, one limitation of coalescent-based methods is their focus on inferring relatively ancient effective population sizes (N_e), since recent effective population sizes are generally poorly accurate and sensitive to different biases (e.g. phasing errors for methods requiring phased data). In contrast, increasingly popular approaches, such as the one implemented in GONE (Santiago et al. 2020 MBE), utilize patterns of linkage disequilibrium (LD) across the genome to infer N_e.

This LD-based strategy offers a significant advantage by providing robust estimates of recent N_e, often within just a few generations. This is particularly useful for detecting recent demographic events, such as population bottlenecks, expansions, or declines, that are crucial for understanding contemporary population dynamics.

From now to the end of this practical, we will focus on a large dataset of whole-genome sequencing data in honey bees (more than 300 genomes) corresponding to 9 genetic clusters (for details see Leroy et al. 2024 bioRxiv, 2024.09.04.611184v1).

A typical question one might ask is whether species have been affected by recent environmental changes (e.g. since the Industrial Revolution). This type of question is nearly impossible to answer using coalescent-based approaches, at least for most organisms. However, methods based on patterns of linkage disequilibrium (LD) hold much more promise in addressing such issues. Let's take a concrete example: the black honey bee, native to Western Europe, is expected to have been heavily impacted by the widespread use of bees from other regions (notably Central Europe) by beekeepers and has been confined to isolated refuges in order to limit their introgression. We can therefore ask how its recent effective population size has recently evolved.

Here, we assume that the script_GONE.sh file has just been downloaded from https://github.com/esrud/GONE/tree/master/Linux/, and that it has been executed on a subset of approximately 500,000 SNPs using the command

below. Note that while this simplified approach is used for the purpose of this demonstration, the process described in our preprint was a bit more complex than this.
We could therefore use the following command:

```
cd ~/Day3_demography/data-honeybees/recentNe_GONE
bash script_GONE.sh seqapipop870_Ouessant_500kSNPs.withheader
```

Note: I chose this specific population as a model due to its biological significance (the black honey bee) and because it allows for faster computations—around 10-20 minutes compared to the 5-10 hours required for some other honey bee populations. However, it will still take some time. You can either launch the command and take a 15-20 minute break (!) or directly explore the results provided below. In any case, I encourage you to review the parameter file "INPUT_PARAMETERS_FILE" before. Do you notice anything particularly unique about honey bees in the file?
Also note that if you need to generate the ped and map files on your own data, Plink can be a relevant program to consider.

In case, you have launched the command on the cluster, the results regarding the effective population sizes will be in the file: "Output_Ne_seqapipop870_Ouessant_500kSNPs.withheader", in case you were too impatient to have a look at the results, they are in the file "Output_Ne_seqapipop870_Ouessant_500kSNPs.withheader_FOR_IMPATIENT_PERSON_ONLY.txt"!

Ok if you want to know to what corresponds the values reported in the file, just execute this:

```
head -1 Output_Ne_seqapipop870_Ouessant_500kSNPs.withheader
```

OR *(in case, you were too impatient!)*

```
head -1 Output_Ne_seqapipop870_Ouessant_500kSNPs.withheader_FOR_IMPATIENT_PERSON_ONLY.txt
```

Since you know to what corresponds the value now, let's remove this sentence of this file:

```
grep -v "independent" Output_Ne_seqapipop870_Ouessant_500kSNPs.withheader
> Output_Ne_Ouessant.txt
```

OR *(in case, you were too impatient!)*

```
grep -v "independent"
Output_Ne_seqapipop870_Ouessant_500kSNPs.withheader_FOR_IMPATIENT_PERSON_ONLY.txt
> Output_Ne_Ouessant.txt
```

Use the Rscript script_GONE_ggplot2.R to generate the R plot based on the GONE-inferred N_e

```
Rscript script_GONE_ggplot2.R
```

The script is expected to output a pdf containing the results.
How has evolved the effective population sizes over the last 100 generations?
Note: GONE provides results for the last ~600 generations, but this can

correspond to very small genomic regions influenced by recombination over these generations—especially in species like honey bees, which exhibit unique characteristics (as you have perhaps noted earlier ;)). Therefore, I strongly recommend avoiding plots that extend beyond the last 200 generations for most species and potentially even fewer for others (e.g. honey bees). Most importantly, refrain from overinterpreting changes in the more distant past, as these estimates become increasingly uncertain with the number of generations (i.e. which can be considered as the opposite of coalescent-based approaches).
OPTIONAL: Generate a plot that combines the results from two separate inferences. Do the results appear consistent ?

Step 4: Perform relatively simple ABC demographic modelling under DILS)

For the end of this practical, I will introduce the DILS software (Demographic Inference with Linked Selection, Fraïsse et al. 2021 Mol Ecol Res), a tool designed for demographic modeling within a single population or between two populations. One key advantage of DILS is its accessibility and speed, especially compared to many other demographic methods. Additionally, DILS incorporates the effects of linked selection—such as background selection and selective sweeps—into its analyses. This allows for more robust estimates of population size changes, migration rates, and other key demographic parameters compared to most currently available software. This feature is particularly valuable for species with high levels of genetic linkage or strong selection pressures, where neglecting these factors could result in misleading inferences.

Today, we will use the online version available at: http://dils.univ-lyon1.fr/. Visit the website from your own computer (!), the user name is "dils" (I will provide the password independently on Slack). Note that it is also possible to install a version on your computer or computing cluster. It is even recommended if you want to perform relatively large projects.

There are two important sections available under the menu in DILS, the "ABC" section (this section) and the "results visualization" one (see step 5).

- **ABC section**: it allows you to import some data (Upload data), filter the data (Data filtering), define the populations for which inferences should be done (one- or two-population tests, outgroups = here "Laboriosa"), and define some priors (by default, broad priors are highly recommended). Explore all these different menus. Each time, you are expected to tick the "please check/validate your choice,"so that everything turns green for the "Run ABC" step.
  But of course, you won't be able to run the computations!! First, because it is expected to be relatively long (between 3 and 12 hours per

run... yes it is still extremely fast for ABC!), and second, imagine 30 participants connecting to the same cluster and launching up to 5 ABC analyses (which remains highly demanding in resources, even under DILS) simultaneously! No, impossible! But at least, you can perform all the steps except the last one that would have allowed you to launch the job on the cluster. And you will know how to do, if you want to test on your own ;)

You can try with one of the two input files I provided and that you can directly download from the physalia cluster (see ~/Day3_demography/data-honeybees/ABC-based_DILS/inputfiles).

Thanks to this first step, you should be able to answer the following questions:
How many individuals are present in the dataset?
How many loci?
Why does this file contain sequence data from three species even if demographic modeling can be performed with one- or two-population models?

Step 5: Explore results of inferences under DILS

This section will be associated with the other menu in DILS:

- **Results visualization**. This section allows you to explore the results of the various inferences performed using DILS and gain valuable insights into the summary statistics, the best-inferred model, the optimal parameter values under the model, and to conduct essential model checking.
  All the models provided here assume a constant population size over TIME. Note that effective population size (N_e) can vary both through time (i.e. during divergence, as we aimed to capture in steps 2 and 3 of this practical) and along the genome due to the effects of linked selection. While DILS did not model temporal changes in N_e in the results provided here, it has still accounted for heterogeneities in effective population sizes across the genome.

First, upload the tar.gz archive (Results visualization > Upload results).

5.1 Observed summary statistics

In the summary statistics (User's dataset > Observed summary statistics), A and B corresponds to population A and B, with A corresponding to the first population and B the second in the name of the archive, e.g. "Carnica_Caucasia_..." Carnica and Caucasia are A and B, respectively.

For instance, piA corresponds to the nucleotide diversity of population A (Carnica in my example immediately above)
For detailed information regarding the summary statistics, click on "Definitions of statistics and parameters"

Questions:
What were your two focal populations? Please make sure to write this information down somewhere!
Do the populations exhibit a high proportion of SNPs with fixed differences (two distinct alleles)?
Based on this information, which of the two populations is the most genetically diverse? What are the median diversity values, as well as the lowest and highest diversity values for each population?
Are the two populations highly differentiated? What is their mean Fst value? How variable is Fst across the genomic windows (lowest/highest values)?

5.2 Model choices

Assuming two-population inferences, DILS is able to hiearchically performs four comparisons:

/* 1: **Current isolation vs. ongoing migration?**i.e. determine whether populations support models with ongoing gene flow or complete isolation

/* 2: **Best model assuming current isolation or ongoing migration?**
Based on the outcome of the first test, DILS refines the comparison, DILS refines the comparison:
If no ongoing migration is detected, it tests whether SI (Strict Isolation) or AM (Ancient Migration) is the better fit.
If ongoing migration is supported, it compares IM (Isolation with Migration) and SC (Secondary Contact).

Briefly, these represent the four classic speciation scenarios.
SI (strict isolation) corresponds to a population split at time T_{SPLIT} (divergence time) from an ancestral diploid panmictic population of N_a individuals. The two diploid populations evolve under constant sizes N_1 and N_2.
AM (ancient migration) is similar to SI but assumed that the two newly formed populations continue to exchange alleles until time T_{AM} (then gene flow stopped until present).
IM (isolation with migration or island model): the two populations continuously exchange alleles throughout the divergence (until present time).

SC (secondary contact): the two populations first evolve in isolation, then experience a secondary contact and start exchanging alleles at time T_{SC}.

/* 3: **Homogeneous or heterogeneous N_e along the genome?**
Test if effective population sizes (N_e) vary along the genome due to linked selection. Models may support homogeneous (Nhomo) or heterogeneous (Nhetero) N_e.

/*4 (OPTIONAL): **Homogeneous or heterogeneous migration rates along the genome?**
For models supporting ongoing migration, this step evaluates whether migration rates are consistent (Mhomo) or variable (Mhetero) across loci. Heterogeneous migration rates can identify loci that suggest isolation, enabling genome scans for selection through locus-specific ABC-based model comparisons. Unlike classical approaches relying on a single summary statistics, this method directly compares locus-specific models based on a large set of summary statistics. However, note that the potential of this approach for detecting barriers to gene flow still needs to be further assessed, both through simulations and empirical studies. This is why our practicals, focused on genome scan approaches, will not rely on such methods (the practicals on Thursday and Friday will use more common approaches). However, we can anticipate that in the near future, methods integrating the detection of both demography and selection will become increasingly available. This development is crucial, as selection within genomes can bias the accurate inference of past demography, while, reciprocally, past demographic events can bias the detection of selection footprints.

Questions:
After navigating to User's dataset > Demographic inferences > Multilocus model comparison, which models are best supported for your specific comparison? Are these model choices strongly supported based on the posterior probabilities?
When you explore Locus specific model comparison, do you find loci that support models different from the one identified in the multilocus model comparison? Are these results consistent with your expectations derived from the summary statistics?

5.3 Estimated parameters

Identifying the best model is an essential step, but exploring the parameter estimates under this model is equally crucial..
First, examining parameter estimates helps diagnose potential issues. For example, if posterior values for a parameter (e.g. divergence time T_{SPLIT}) are close to the limits of the prior, this might indicate that the priors are constraining the model too strongly. Such constraints could influence conclusions,

suggesting the need to redo the analysis with expanded priors. To investigate this, check the distributions in User's dataset > Demographic inferences > Estimated Parameters > Posterior."
Second, click on Highest Posterior Density to view the median and 95% confidence intervals for the posterior estimates.

Questions:
Using the neural network and direct posterior (purple), are the posterior probabilities of effective population sizes large? The values represent the number of diploid individuals. Which population has the largest distribution? Which effective population size seems the most challenging to infer? Is this consistent with your observations?
Beyond effective population sizes, focus on the inferred divergence time (TSPLIT). If AM or SC scenarios are supported, also consider TAM or TSC. These values are expressed in generations. Assuming one generation per year for honey bees (prior to domestication and hive use), is the inferred divergence recent? For AM or SC scenarios, are the inferred timings of these events recent?

5.4 Model checking

A critical step in demographic modeling is assessing how well the model fits the observed data. This involves verifying whether the simulations effectively reproduce the observed values of the summary statistics. While this might seem straightforward, this crucial step is too often neglected.
First, I encourage you to use the PCA (User's dataset > Demographic inferences > Goodness-of-fit test > plot) to check whether the observed data (yellow dot) falls within the scatterplot of simulations. If the yellow dot is distant from all other dots, it suggests the simulations cannot replicate the observed dataset. This indicates that the inference results are likely wrong and should be reevaluated (e.g. check for issues regarding the data, the inputfile, too narrow priors, etc.).
Next, examine which summary statistics from the real data align well with those from the simulations and identify any that deviate. This information is available in the "P-values" section. Summary statistics with significant p-values are those not well captured by the simulations.
Finally, evaluate the joint Site Frequency Spectrum (jSFS), which is particularly important. Here, the jSFS (a two-dimensional SFS) was not directly used as a summary statistic for the ABC (as indicated by "noSFS" in the file name). From my perspective, this makes it even more interesting to check how well simulations reproduce the observed SFS independently of the ABC. Even when the SFS is used in simulations, examining the fit of the jSFS remains valuable (similarly to what is generally done for methods based on composite likelihood).

In the "SFS" section, the observed jSFS corresponds to real data (top-left), while the expected SFS (top-right) is derived from the posterior set of simulations. The difference between the two former SFS are also shown (bottom-left) as well as the significance of this difference (bottom-right).

Questions:
Does the PCA suggest that simulations (or a subset of them) reproduce the observed summary statistics?
According to you, why might some gray points (not used for the posterior) be closer to the yellow dot than some purple points in the PCA?
How many summary statistics are well captured versus not well captured for the ABC analysis you have selected?
Understand the SFS. Why is there missing information in certain corners of the SFS (e.g. x=0 & y=0)?
How well does the jSFS align with the simulations? Are many bins in the SFS associated with significant p-values?
Based on all the model checking you have now performed, are you confident in the results of these simulations?

Congratulations on completing this practical on demographic modeling!
If you still have time, consider revisiting Step 5 with a few additional tar.gz files to further explore the results. You may also want to have more precise information regarding the methods or the results, including the original GONE paper (Santiago et al. 2020 MBE), the original DILS paper (Fraïsse et al. 2021), the DILS manual, as well as my recent preprint on honey bees, in which I describe a bit more the results of the sections 3 to 5. These three files are available on ~/Day3_demography/data-honeybees/References.