gtImputation (Genotype Imputation)

v0.19

GI en Desarrollo de Especies y Comunidades Leñosas (WooSp)
Dpto. Sistemas y Recursos Naturales
ETSI Montes, Forestal y del Medio Natural
Universidad Politécnica de Madrid

https://github.com/ggfhf/

Table of contents

Disclaimer

The software package gtImputation (Genotype Imputation) is available for free download from the GitHub software repository (https://github.com/GGFHF/gtImputation) under GNU General Public License v3.0.

Disclaimer

Genotype imputation: why do we need it?

Reduced representation sequencing methodologies allow genotyping of many specimens at once and are extensively used in many fields of genomics. The output of sequencing platforms is processed to call variants at specific polymorphic sites of the individual genomes, ant all the information is stored in VCF (Variant Call Format) files. In these files, the format of the records with variant information consists of a series of tab-delimited fields including: the REF field that indicates the base or bases of the reference allele, the ALT field that includes the base or bases of the alternative alleles, the FORMAT field that contains a list of the data reported for each sample and n SAMPLE fields with the values of the data detailed in FORMAT for each sample. The genotype is one of the data reported and is represented by the two encoded alleles separated by a '/' character (non-phased genotypes) or the '|' character (phased genotypes). The allele coding is denoted by a '0' for the reference allele, an '1' for the first alternative allele, a '2' for the second alternative allele, and so on. In the case where an allele in a sample is missing data it is represented by a '.' character.

One of the caveats of current methodologies of reduced representation genome approaches is that they produce large amounts of missing data that may impact statistical inference and introduce bias in the outcome of experiments that use these methods. There are several strategies to reduce the impact of missing data in genomic datasets. One solution is to remove variants containing missing data, but, depending on the outcome of the experiment, this can lead to loss of accuracy in genomic inference. Another approach is to deduce (impute) the genotype that each missing data should have using validated references through frequentist or machine learning approaches.

In the GBS context, imputation is the process of inferring untyped polymorphic markers (usually only SNPs) in a genotyped population. Missing genotype imputation is commonly used in several applications of genomics and population genetics, such as GWAS (Browning, 2008), where researchers analyze the genetic variations across a large population to identify associations between genetic markers and particular traits. Additionally, population genetics studies, such as inferring population history or demographic events, often require imputation of missing genotype data to accurately reconstruct the genetic landscape and evolutionary dynamics of different populations (Fu, 2014).

While haplotype-clustering algorithms using EM models (Scheet & Stephens, 2006), and hidden Markov models (Marchini et al., 2007) produce accurate genotype imputation in model species that have haplotype reference panels, genotype imputation in non-model species with less genomic resources can also be approached using unsupervised machine learning algorithms, such as random forest, K-NN (Money et al., 2015) or, as is the case in the present manual, self-organizing maps (SOM) .

gtImputation

gtImputation is an application designed to determine the genotypes of missing data in a VCF or tabular format file using a machine learning procedure based on self-organizing maps (SOM) called **SOM imputation**. In addition, the application is also designed to perform a **naive imputation** method.

For the case of genotype bi-allelic loci, such as most SNP data in VCF files, a vector of genotypes representing all the possible states in a variant calling file can be easily defined by using the IUPAC nucleotide codes:

- **A** for adenine/adenine homozygotes
- **C** for cysteine/cysteine homozygotes
- **G** for guanine/guanine homozygotes
- **T** for thymine/thymine homozygotes
- **E** for adenine/guanine heterozygotes
- **K** for guanine/thymine heterozygotes
- **M** for adenine/cytosine heterozygotes
- **S** for cytosine/guanine heterozygotes
- **W** for adenine/thymine heterozygotes
- **Y** for cytosine/thymine heterozygotes
- **N** in the case of missing data.

Naive imputation

Naive imputation is a frequentist procedure that determines which is the most frequent genotype in the affected variant and assigns the obtained genotype to the missing variant data.

SOM imputation

Kohonen's self-organizing map (SOM; Kohonen, 2001), is a commonly used method for data clustering and visualization that fits to the analysis of genomic sequence datasets. SOM performs an unsupervised clustering process that spatially organizes the patterns found in a dataset considering only their homology, without knowing the class to which they belong. SOM is able to conduct exploratory data analysis and visualization without the need for calibration or classification of the information to be processed. In the context of genotype data, SOM orders spatially the vectorized genotypes on a two-dimensional (2D) map, providing the similarity or dissimilarity with all other groups of sequences and the clustering of the input into groups.

SOM networks have two layers of neurons with connections between them: n neurons in the input layer and m neurons in the output layer, with n << m and n*m connections. The data represent information obtained from several dimensions (2 in our case). No supervision is required to perform the classification.

The SOM imputation performed by gtImputation has two processes (Figure 1) that store the required information in three SQL tables (Box 1):

1. **Process 1**: Creation a genotype database from the VCF or tabular format file
2. **Process 2**: Run of the imputation process itself using the genotype database

Process 1. Fetching the information from the VCF or tabular format (see Appendix A) file and creation of SQL databases. The linkage disequilibrium between each pair of SNPs according to Ragsdale & Gravel (2020) and sample kinship according to Goudet et al. (2018) are calculated and saved in a Sqlite database (Box 2). This process should be run once per VCF file.

Process 2. Imputation of the missing data from the VCF file whose genomic SQL database was built in Process 1 (Box 3). In order to create SOM objects, gtImputation uses MiniSom software (Vettigli, 2018). Process 2 can be run several times by varying the input parameters.
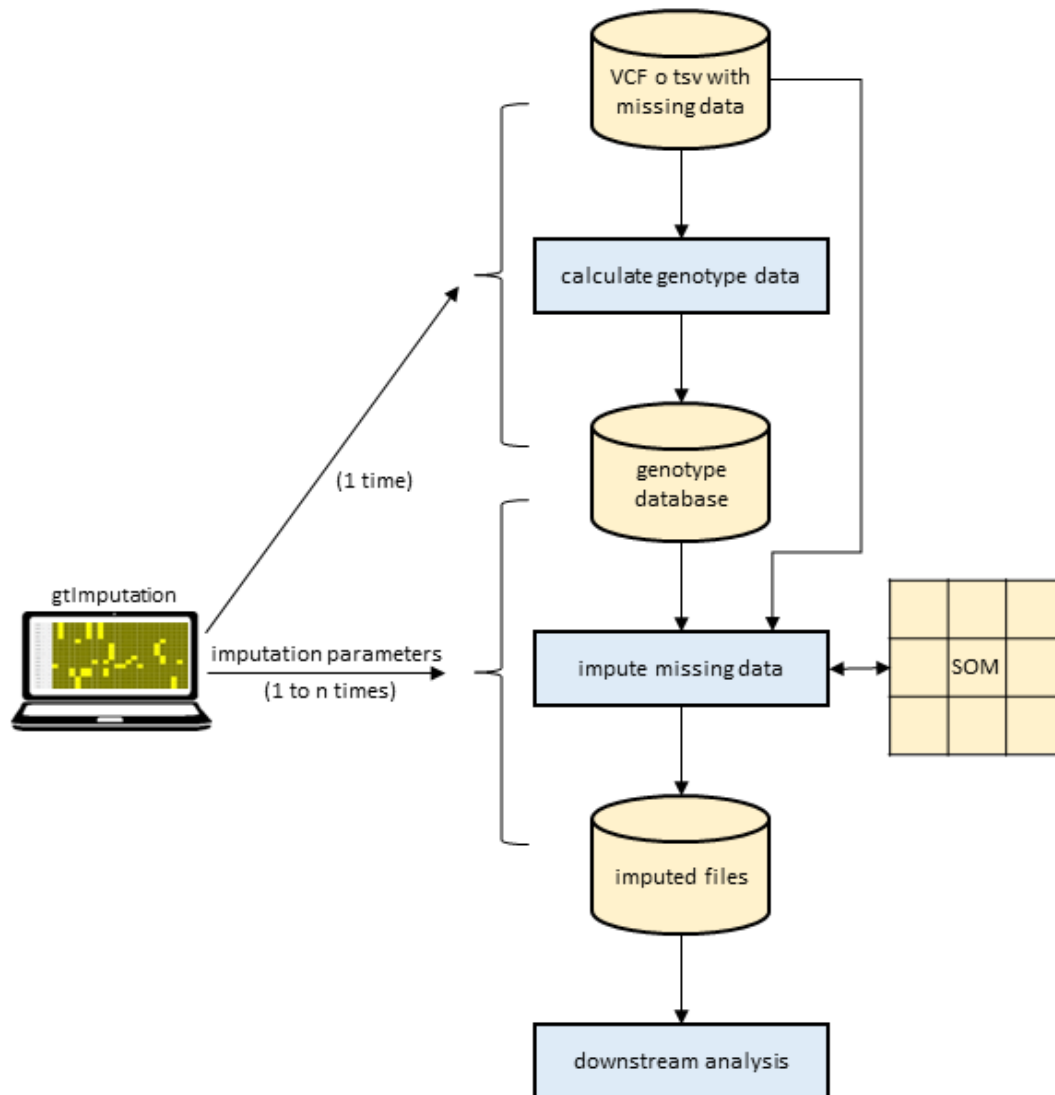
Figure. 1. Flow-chart of gtImputation.

The user can modify the following MiniSom input parameters:

- x and y dimensions of the SOM.
- sigma: spread of the neighborhood function, needs to be adequate to the dimensions of the map.

- initial learning rate.
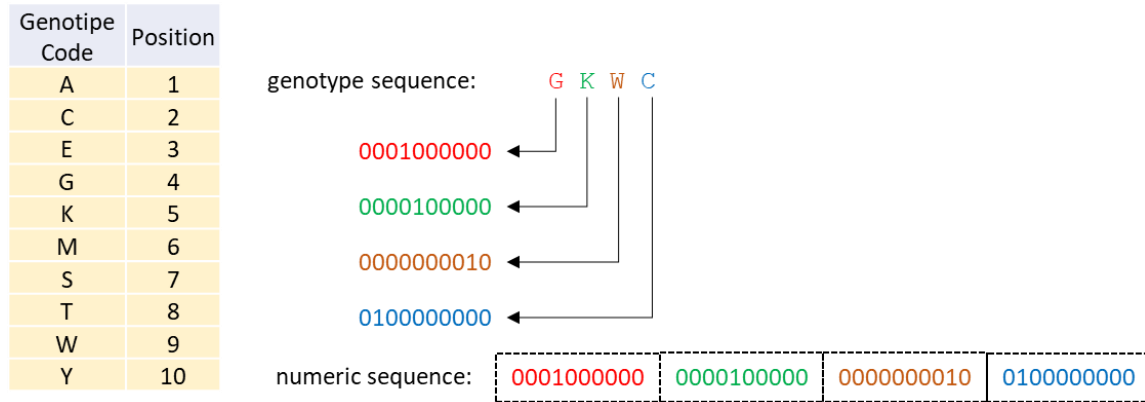- maximum number of iterations for the training.

The other MiniSOM parameters are set by default or by running the process.

Other mandatory input parameters of gtImputation related with the selection of SNP to construct the vector sequences that the SOM will consider to assign the samples to their corresponding neurons are the following:

- a minimum squared correlation (r^2) threshold between SNP.
- a maximum number of SNP.

Notice that while running the SOM, in case the maximum number of SNP is lower than the number of SNP obtained by applying the r^2 threshold, the maximum number of SNP with the higher r^2 are considered. To fix a number of SNP the r^2 should be fixed to 0.

Once the SNP vector sequences are constructed, they are converted to 0-1 vectors. The traditional SOM works with a continuous numerical input space and Euclidean distances. In order to work with nucleotide sequences, gtImputation made a conversion to numerical values using an array of 10 positions such that the first position corresponds to the character 'A', the second to the character 'C' and so on up to the tenth which corresponds to the character 'Y'. Each coded genotype assigns a 1 to the element of the array corresponding to its character and 0 to the rest. In the case of missing data, all the elements of the array have a value of 0 (Figure 2).

Genotipe Code	Position
A	1
C	2
E	3
G	4
K	5
M	6
S	7
T	8
W	9
Y	10

genotype sequence: G K W C

0001000000

0000100000

0000000010

0100000000

numeric sequence: | 0001000000 | 0000100000 | 0000000010 | 0100000000 |

**Figure. 2**. From a genotype sequence to its numeric sequence.

Once the SOM is created, it is trained with the numeric vectors of each sample and each sample is assigned to a specific neuron (Figure 3). The neuron that corresponds to a sample with missing data is identified and there are two possibilities to impute the genotype:

- imputing the most frequent genotype of the samples found in the neuron hosting the sample with missing data.
- imputing the genotype of the sample with the closest kinship found in the neuron.

For further information refer to the manuscript:

Mora-Márquez F, Nuño J. C., Soto A. & López de Hereda U. (2024). Missing genotype imputation in non-model species using self-organizing maps. *Molecular Ecology Resources*, e13992. DOI: https://doi.org/10.1111/1755-0998.13992
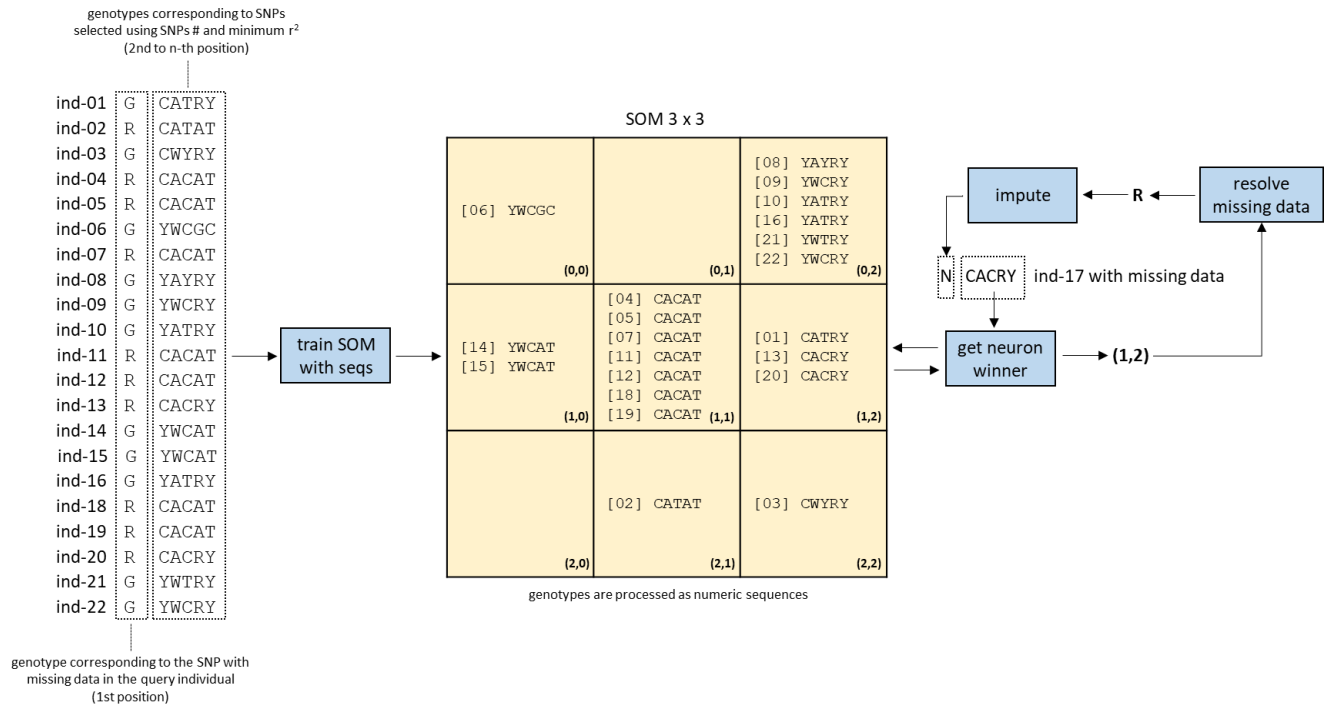
Figure. 3. SOM training and imputation process in gtImputation.

Box 1: Description of the three SQL tables used by gtImputation to store the information fetched from the VCF file in Process 1 and employed in Process 2

```
###################
# Table vcf_snps #
###################

Column                   Type    Comment
===================---==== ========================================================================
variant_id               TEXT    identification of the variant (fields CHROM-POS of VCF file)
ref                      TEXT    reference allele (field REF of VCF file)
alt                      TEXT    alternative allele(s) (field ALT of VCF file)
sample_gt_list           TEXT    comma-separated of sample (individual) genotypes in the variant
sample_withmd_list       TEXT    list of samples (individuals) with missing data


######################
# Table vcf_kinship #
######################

Column           Type      Comment
============---======= ====================================================================
individual_i     INTEGER   identification of an individual
individual_j     TEXT      identification of another individual
ru               REAL      unweighted average estimator of kinship (formula 4, Goudet, 2018)


####################################
# Table vcf_linkage_disequilibrium #
####################################

Column                  Type    Comment
==================== ==== ================================================================
snp_id_1                TEXT    identification of a SNP with missing data (fields CHROM-POS)
snp_id_2                TEXT    identification of one of the other SNP (fields CHROM-POS)
dhat                    REAL    unbiased estimator for the covariance of alleles co-occurring on a
                                haplotype
r_2                     REAL    squared correlation
sample_withmd_list_2    TEXT    list of samples (individuals) with missing data


Note:

A sample (individual) is identified by the order in which it appears in the fields SAMPLE: the first
sample is 0; the second one, 1; and so on.
```

Box 2: Pseudocode of the creation of a genotype database from the information fetched from a VCF file.

```
##################################################
# Creation a genotype database from the VCF file #
##################################################

function calculate_genotype_data (database, vcf_file)

    initialize sample_list and sample_number
    initialize kinship_dictionary (structure: {snp_i: {snp_j: {summation, l}})

    do connection to database

    drop the table vcf_snps (if it exists)
    create the table vcf_snps

    drop the table vcf_linkage_disequilibrium (if it exists)
    create the table vcf_linkage_disequilibrium

    drop the table vcf_kinship (if it exists)
    create the table vcf_kinship

    open the vcf_file
    read the first record of vcf_file
    while there are records in vcf_file

        while there are record in vcf_file and they corresponding to metadata
            read the next record of vcf_file

        if the record corresponding to the column description
            build sample_list and sample_dictionary
            calculate samples_number
            set 0 in the summation and l values in kinship_dictionary
                    (used in the calculation of the unweighted average estimator of kinship)
            read the next record of vcf_file

        while there are records in vcf_file and they corresponding to variants
            variant_id = f'{field CHROM}-{field CHROM}'
            get the reference allele (field REF) and alternative alleles (field ALT)
            if the variant has more than one alternative allele
                end the process with error
            get the genotype_position (subfield GT in the field FORMAT)
            build sample_genotypes_list considering genotype_position in fields SAMPLE
            build samples_with_missing_data_list
            update kinship_dictionary adding this calculation in the summation for each
                    samples i and j:
                        (Xi - 2 * p) * (Xj - 2 * p) / (2 * p * (1 - p))
                    where Xi and Xj are the dosage of reference allele for samples i and j
                    respectively and p is the frequency of reference allele in the current variant
                    (do not-update when there is missing data in samples i or j, or p value
                    is 0 or 1)
            update kinship_dictionary adding 1 in l for samples i and j
                    (do not-update when there is missing data in samples i or j, or p value
                    is 0 or 1)
            insert SNP data into table vcf_snps if there is more than one genotype
            read next record of vcf_file

    close vcf_file

    for each samples i and j:
        calculate the unweighted average estimator as summation / l
        insert kinship data into the table vcf_kinship from kinship_dictionary

    get snp_ids_list from the table vcf_snps
    get snp_with_missing_data_ids_list from the table vcf_snps
```

```
    for each snp_id in snp_with:missing_data_ids_list:
        calculate_snp_linkage_disequilibrium (connection, sample_number, snp_id, snp_ids_list)

    save changes into database
    close connection to database

    close vcf_file
```

```
####################################################
# Calculation the linkage disequilibrium of a SNP #
####################################################


function calculate_snp_linkage_disequilibrium (connection, sample_number, snp_id_1, snp_id_2_list)

    get sample_genotypes_1_list of snp_id_1 from the table vcf_snps

    for each snp_id_2 in snp_id_2_list (except snp_id_1):

        get sample_genotypes_2_list of snp_id_2 from the table vcf_snps

        calculate the observed genotype counts and allele frequencies considering:

                ri: reference allele of SNPi
                ai: alternative allele of SNPi
                ni: count of observed genotype pair i
                n: summation of ni

                    r2/r2 r2/a2 a2/a2
                +------------------
        r1/r1 |   n1     n2     n3
        r1/a1 |   n4     n5     n6
        a1/a1 |   n7     n8     n9

        calculate the unbiased estimator for the covariance of alleles co-occurring on a haplotype
                according to:
                dhat = ((n1 + n2/2 + n4/2 + n5/4) * (n5/4 + n6/2 + n8/2 + n9) –
                        (n2/2 + n3 + n5/4 + n6/2) * (n4/2 + n5/4 + n7 + n8/2)) / (n * (n - 1))

        calculate the squared correlation according to:

                rfi: reference allele frequency of SNP i
                afi: alternative allele frequency of SNP i

                r_2 = (dhat ** 2) / (rf1 * af1 * rf2 * af2)

        insert linkage disequilibrium data into the table vcf_linkage_disequilibrium
```

Box 3: Pseudocode of the missing data imputation using Self-Organizing Maps and the genotype database created previously with the input VCF file

```
################################################################################
# Imputation of genotypes with missing data in a VCF file using Self-Organizing Maps #
################################################################################

function impute_md_som (database, input_vcf_file, imputed_vcf_file, minimum_r2, snps_num, xdim, ydim,
                        sigma, learning_rate, num_iteration, genotype_imputation_method)

    initialize sample_list and sample_number

    do connection to database

    get kinship_dictionary (structure: {snp_i: {snp_j: {summation, l}}) from table vcf_kinship

    get snp_ids_list with linkage disequilibrium data from the table vcf_linkage_disequilibrium

    open input_vcf_file
    open imputed_vcf_file

    read the first record of input_vcf_file
    while there are record in input_vcf_file

        while there are records in input_vcf_file and they corresponding to metadata
            read the next record of input_vcf_file
            write the metadata record in imputed_vcf_file

        if the record corresponding to the column description
            build sample_id_list
            calculate the samples_number
            write column description record in imputed_vcf_file
            read the next record of input_vcf_file

        while there are records in input_vcf_file and they corresponding to variants
            variant_id = f'{field CHROM}-{field CHROM}'
            get the reference allele (field REF) and alternative alleles (field ALT)
            if the variant has more than one alternative allele
                end the process with error
            get the genotype_position (subfield GT in the field FORMAT)
            build sample_genotypes_list considering genotype_position in fields SAMPLE
            if variant_id in snp_ids_list:
                get samples_with_missing_data_list of the variant from table vcf_snps
                get best_snps_ids_list with SNP ids with the highest r^2 values regarding variant_id
                selected_snp_ids_list = [variant_id] + best_snps_ids_list
                build sample_sequence_list corresponding to selected_snp_ids_list
                build sample_numeric_haplotypes_list from sample_sequence_list
                calculate most_frequent_genotype
                if there is one genotype:
                    set most_frequent_genotype as genotype of samples with missing data
                else:
                    build the input data to the SOM algorithm
                    build the training data corresponding to samples without missing data
                            from the input data
                    build the test data corresponding to samples with missing data from the input data
                    create a new SOM xdim x ydim instance with sigma, learning_rate, num_iteration
                            (default values for all other parameters)
                    initialize the weights to span the first two principal components
                    train the SOM
                    for each sample_with_missing_data in samples_with_missing_data_list:
                        get winning_neuron for samples_with_missing_data
                        if genotype_imputation_method is the most frequent genotype in the neuron:
                            get samples_in_neuron_list in wining_neuron
                            calculate most_frequent_genotype_in_neuron from samples_in_neuron_list
                            set most_frequent_genotype_in_neuron in genotype of
```

```
                        sample_with_missing_data
            elif genotype_imputation_method is the genotype of closest kinship individual:
                get the most_related_sample_id from kinship_dictionary
                set genotype most_related_sample_id of in genotype of
                        sample_with_missing_data
            build the genotype text after imputation
        rebuild the sample genotype data list and their corresponding record data
        write the variant record in imputed_vcf_file
        read the next record of input_vcf_file

    close connection to database

    close input_vcf_file
    close imputed_vcf_file
```
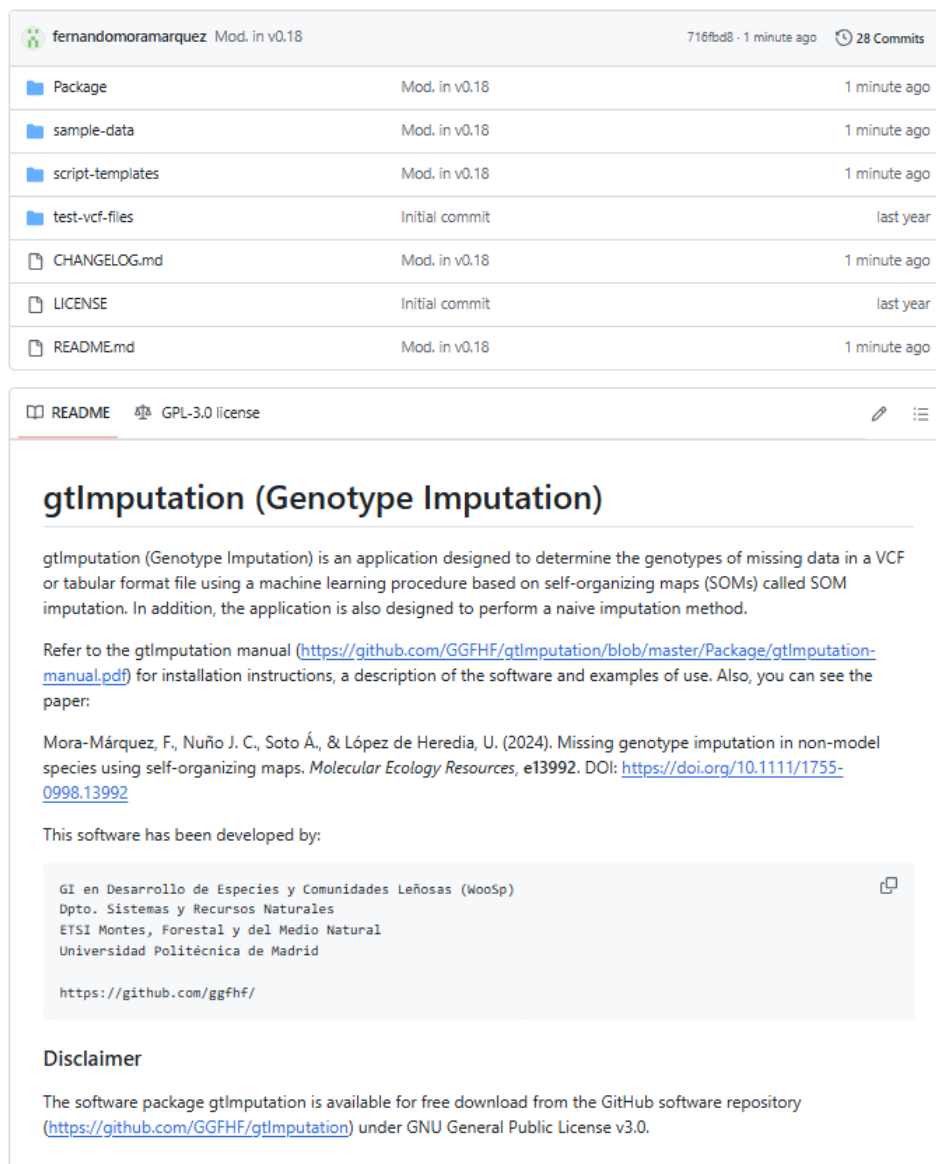
Installation

gtImputation installation

gtImputation was programmed in Python 3 and it generates dynamic Bash scripts to perform annotation pipelines. gtImputation runs in any computer with Linux, macOS or Windows with WSL (Windows Subsystem for Linux).

gtImputation is available from the GitHub software repository in the URL https://github.com/GGFHF/gtImputation/, and it is distributed under GNU General Public License Version 3 (Figure 4).

Figure. 4. gtImputation home at GitHub software repository.

To download gtImputation, click in *Code* and in the pup-up window click in *Download ZIP*.

To install gtImputation on Linux and macOS, simply decompress the gtImputation-main.zip into a directory, typing the following command in a terminal window:

```
$ unzip gtImputation-master.zip
```

Then, the execution permissions of the programs must be set by using this command:

```
$ chmod u+x *.py
```

To install gtImputation on Windows use, e.g., "Extract All.." of the File Explorer on gtImputation-master.zip.

Additional infrastructure software installation

Python 3, version 3.11 or higher, is necessary for the correct functioning of gtImputation. If Python 3 is not installed in your computer, you can download it from the official website (https://www.python.org/), or use one of the several distributions that include Python along with other software packages for standard bioinformatic analysis such as Anaconda (https://www.continuum.io/).

To work properly, gtImputation needs the following Python modules:

- PyQt5 (https://www.riverbankcomputing.com/static/Docs/PyQt5/), a Python interface for QT software package.
- Pandas (https://pandas.pydata.org/), a Python library for data analysis and manipulation tool
- Matplotlib (https://matplotlib.org/), a software for creating static, animated, and interactive visualizations in Python.
- Plotnine (https://plotnine.readthedocs.io/en/stable/), an implementation of a grammar of graphics in Python based on ggplot2

Next, we present how to install this additional software in five environments: a) an Ubuntu Linux 22.04 LTS where Python3 is pre-installed in the OS; b) an Ubuntu Linux 24.04 LTS where Python3 is pre-installed in the OS; c) an Ubuntu 22.04 LTS or an Ubuntu 24.04 LTS using Miniconda3; d) a macOS 12.6.1 where Python is installed using Miniconda3; e) a Windows 10 using WSL and Miniconda3.

Installation on Ubuntu 22.04 LTS using the O.S. Python

On Ubuntu 22.04 LTS, Python 3 is pre-installed. To install PyQt5, if necessary, open a terminal window and type the following commands:

```
$ [sudo apt install --yes python3-pip]

$ sudo pip3 install pyqt5

$ sudo apt install libxcb-xinerama0
```

And finally, install the Pandas, Matplotlib and Plotnine libraries, if necessary, typing the following commands in the terminal window:

```
$ sudo pip3 install pandas

$ sudo pip3 install matplotlib

$ sudo pip3 install plotnine
```

Installation on Ubuntu 24.04 LTS using the O.S. Python

On Ubuntu Ubuntu 24.04 LT, Python 3 is pre-installed. To install PyQt5, if necessary, open a terminal window and type the following commands:

```
$ [sudo apt install --yes python3-pip]

$ sudo pip3 install --break-system-packages pyqt5

$ sudo apt install libxcb-xinerama0
```

And finally, install the Pandas, Matplotlib and Plotnine libraries, if necessary, typing the following commands in the terminal window:

```
$ sudo pip3 install --break-system-packages pandas

$ sudo pip3 install --break-system-packages matplotlib

$ sudo pip3 install --break-system-packages plotnine
```

Installation on Ubuntu 22.04 LTS or Ubuntu 24.04 LTS using Miniconda3

On Ubuntu another way to install Python and the additional software is to use Miniconda3. First, install Miniconda3, e.g. in the user root directory ($HOME). Open a terminal window and type the following commands:

```
$ cd $HOME
```

```
$ wget https://repo.continuum.io/miniconda/Miniconda3-latest-Linux-x86_64.sh
```

```
$ chmod u+x Miniconda3-latest-Linux-x86_64.sh
```

```
$ ./Miniconda3-latest-Linux-x86_64.sh -b -p $HOME/Miniconda3
```

```
$ rm Miniconda3-latest-Linux-x86_64.sh
```

```
$ $HOME/Miniconda3/condabin/conda init bash
```

Then close the terminal and open a new terminal and type the following commands to configure the download environment:

```
$ conda config --add channels defaults
```

```
$ conda config --add channels bioconda
```

```
$ conda config --add channels conda-forge
```

```
$ conda config --set channel_priority strict
```

```
$ conda install --yes --name base mamba
```

Now, install the PyQt5 library:

```
$ mamba install --yes --name base pyqt
```

And finally, install the Pandas, Matplotlib and Plotnine libraries, and update SQLite3 typing the following commands in the terminal window:

```
$ mamba install --yes --name base pandas
```

```
$ mamba install --yes --name base matplotlib
```

```
$ mamba install --yes --name base plotnine
```

```
$ mamba update --yes --name base sqlite
```

Installation on macOS 15.0.1 using Miniconda3

First, install Homebrew and wget command, if necessary, typing the following commands in a terminal window:

```
$ /bin/bash -c "$(curl -fsSL
https://raw.githubusercontent.com/Homebrew/install/HEAD/install.sh)"
```

```
$ brew install wget
```

Now download the Miniconda3 software file typing the command:

```
$ cd $HOME
```

```
$ wget https://repo.anaconda.com/miniconda/Miniconda3-latest-MacOSX-x86_64.sh
```

And provide execution permission to this file and run it typing the commands:

```
$ chmod u+x Miniconda3-latest-MacOSX-x86_64.sh
```

```
$ ./Miniconda3-latest-MacOSX-x86_64.sh -b -p $HOME/Miniconda3
```

```
$ rm Miniconda3-latest-MacOSX-x86_64.sh
```

Now modify the PATH adding the directory bin of Miniconda3 typing the command:

```
$ echo "export PATH=\"$HOME/Miniconda3/bin:\$PATH\"" >> ~/.zshrc
```

Then close the terminal and open a new terminal and type the following commands to configure the download environment:

```
$ conda config --add channels defaults
```

```
$ conda config --add channels bioconda
```

```
$ conda config --add channels conda-forge
```

```
$ conda config --set channel_priority strict
```

```
$ conda install --yes --name base mamba
```

Now, install the PyQt5 library:

```
$ mamba install --yes --name base pyqt
```

And finally, install the Pandas, Matplotlib and Plotnine libraries, and update SQLite3 typing the following commands in the terminal window:

```
$ mamba install --yes --name base pandas
```

```
$ mamba install --yes --name base matplotlib
```

```
$ mamba install --yes --name base plotnine
```

```
$ mamba update --yes --name base sqlite
```

Installation on Microsoft Windows 10 (64 bits) using WSL and Miniconda3

gtImputation uses the Windows Subsystem for Linux (WSL) and Ubuntu to run some scripts coded in Bash. WSL has to be installed before using gtImputation. For further clarification about WSL, you can see the URL https://learn.microsoft.com/en-us/windows/wsl. In order to install WSL and Ubuntu 24.04, open a command prompt as administrator and type the following command:

```
> wsl --install --distribution Ubuntu-24.04
```

Restart Windows. Then a window will appear installing Ubuntu. This process may take a few minutes. You will have to enter an username and its password. When the process ends, close this window.

In order to install Miniconda3, e.g. in the user root directory (%USERPROFILE%), open a command prompt and type the following commands:

```
> cd %USERPROFILE%
```

```
> curl -O https://repo.anaconda.com/miniconda/Miniconda3-latest-Windows-x86_64.exe
```

```
> Miniconda3-latest-Windows-x86_64.exe /InstallationType=JustMe /AddToPath=1 /RegisterPython=1 /S /D=%USERPROFILE%\Miniconda3
```

```
> del Miniconda3-latest-Windows-x86_64.exe
```

Then close the terminal and open a new terminal and type the following commands to configure the download environment:

```
> conda config --add channels defaults

> conda config --add channels bioconda

> conda config --add channels conda-forge

> conda config --set channel_priority strict

> conda install --yes --name base mamba
```

Now, install the PyQt5 library:

```
> mamba install --yes --name base pyqt
```

And finally, install the Pandas, Matplotlib and Plotnine libraries, and update SQLite3 typing the following commands in the terminal window:

```
> mamba install --yes --name base pandas

> mamba install --yes --name base matplotlib

> mamba install --yes --name base plotnine

> mamba update --yes --name base sqlite
```

Starting gtImputation

On Ubuntu or macOS computers, gtImputation starts typing the following command in a terminal window in the directory where the package of gtImputation is downloaded:

$ ./gtImputation.py

On Windows computer, the command to start gtImputation is:

> python gtImputation.py

Initial appearance of gtImputation at application startup is shown in Figure 5.
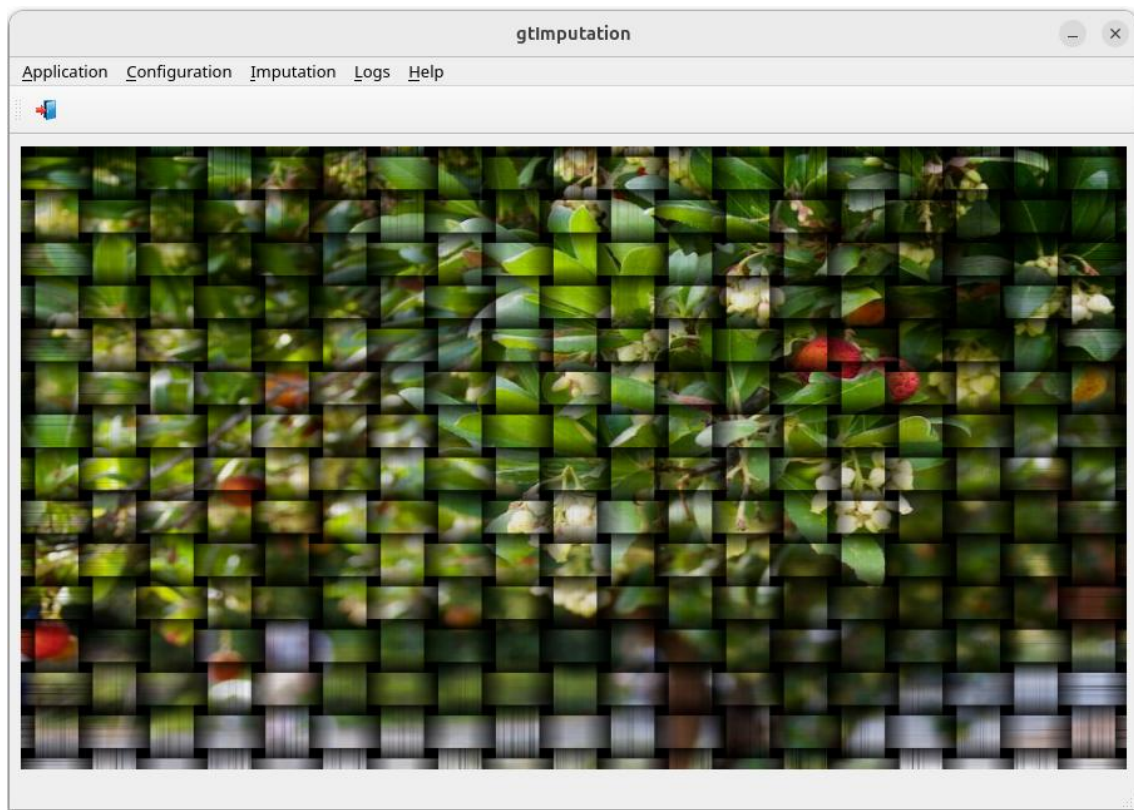
Figure. 5. Aspect of the gtImputation interface at startup.

First steps

gtImputation menus

gtImputation is structured in several menus:

System

Just to exit the application.

Configuration

This menu contains all the items related to:

- Recreate gtImputation config file
- View gtImputation config file
- Install Bioinfo software

Imputation menu

The options included here allow to run scripts to perform imputation process:

- Naive imputation
  - Run imputation process
  - Review imputation process
- SOM imputation
  - Build a genotype database
  - Run imputation process
  - Review imputation process

Logs menu

This menu allows the access to the application logs:

- View submission logs
- View result logs

Help menu

It contains the documentation of the application.

First steps

Configuring the gtImputation environment

When gtImputation starts for the first time it is required to configure the gtImputation environment. To do so, we select the menu item with the following path:

Main menu > Configuration > Recreate gtImputation config file

The figure 6 shows the window corresponding to this menu item. Default values are presented for *Miniconda directory*, *Genotype database directory* and *Result directory*. If necessary, modify them and press the button *[Execute]*.
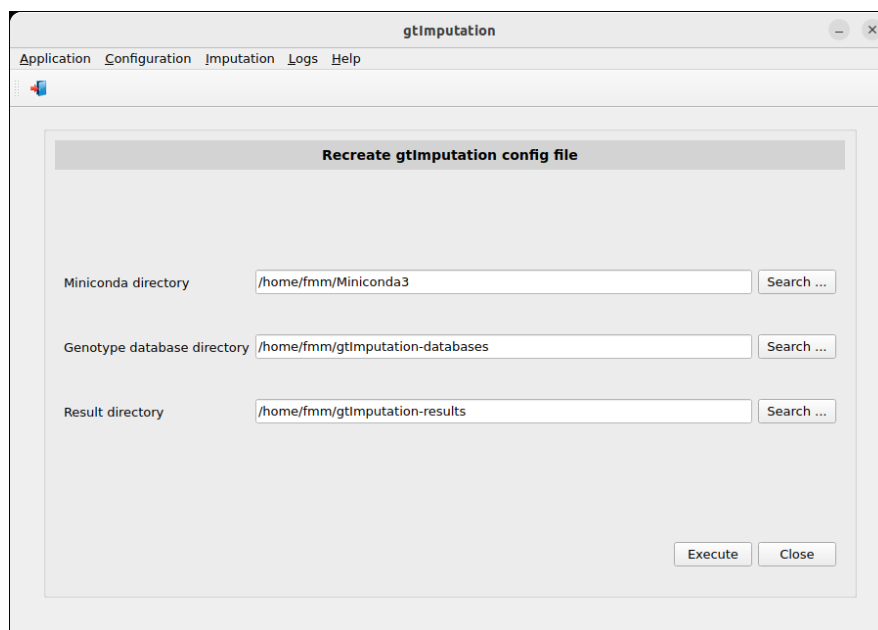
Figure. 6. gtImputation window corresponding to the menu item *Recreate gtImputation config file*.

Installing infrastructure software

gtImputation needs the following software:

- MiniSom: a minimalistic and NumPy-based implementation of the Self Organizing Map (Vittigli, 2018).

The installation of this software is automatic when Miniconda is installed Then, install Miniconda·selecting the menu item with this path:

Main menu > Configuration > Bioinfo software installation > Miniconda3 and additional infrastructure software

A new window is shown (Figure 7).
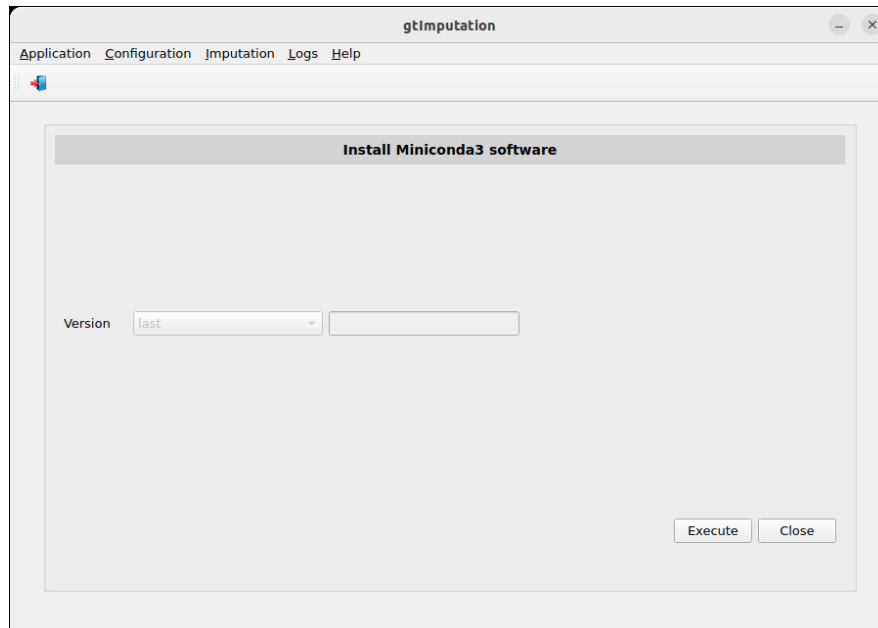
Figure. 7. gtImputation window corresponding to the menu item *Miniconda3 and adiccional infrastructure software*.

Press the bottom *[Execute]*. A pop-up window will display the submission log (Figure 8).
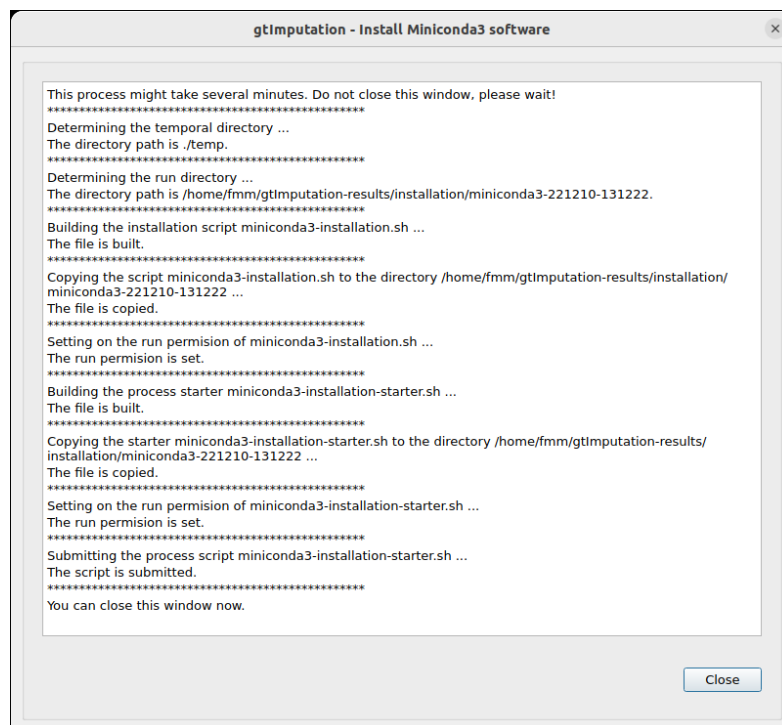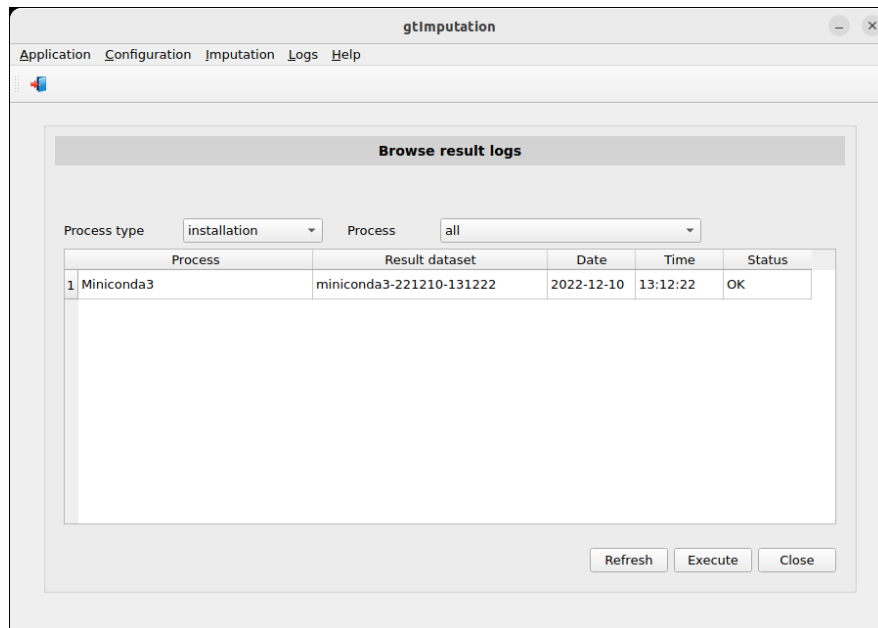

Figure. 8. The log dialog showing the submission log of Miniconda3 installation process.

To view the process log during and after the run, select the menu item with this path:

Main menu > Logs > Result logs

In the raised window (Figure 9), select **installation** in the *Process type* combo-box*.*

Figure. 9. gtImputation window corresponding to the menu item *Logs - Result logs* after ended the Miniconda installation.

Since a single installation process was performed, the process **miniconda3-20221210-131222** corresponding to the last (and unique) Miniconda3 installation. Double click on it, or click on it and on the button Execute. Then, a pop-up window appears with its corresponding log (Figure 10).

All the process logs have:

- A header with the time when it started.
- At the bottom, a summary with the status (OK, if all the programs have ended without errors; WRONG, otherwise), the end time, and the duration of the script run.

In the log dialog, there is a button to refresh the run status. If the process has not ended, click on it and the log will be updated.
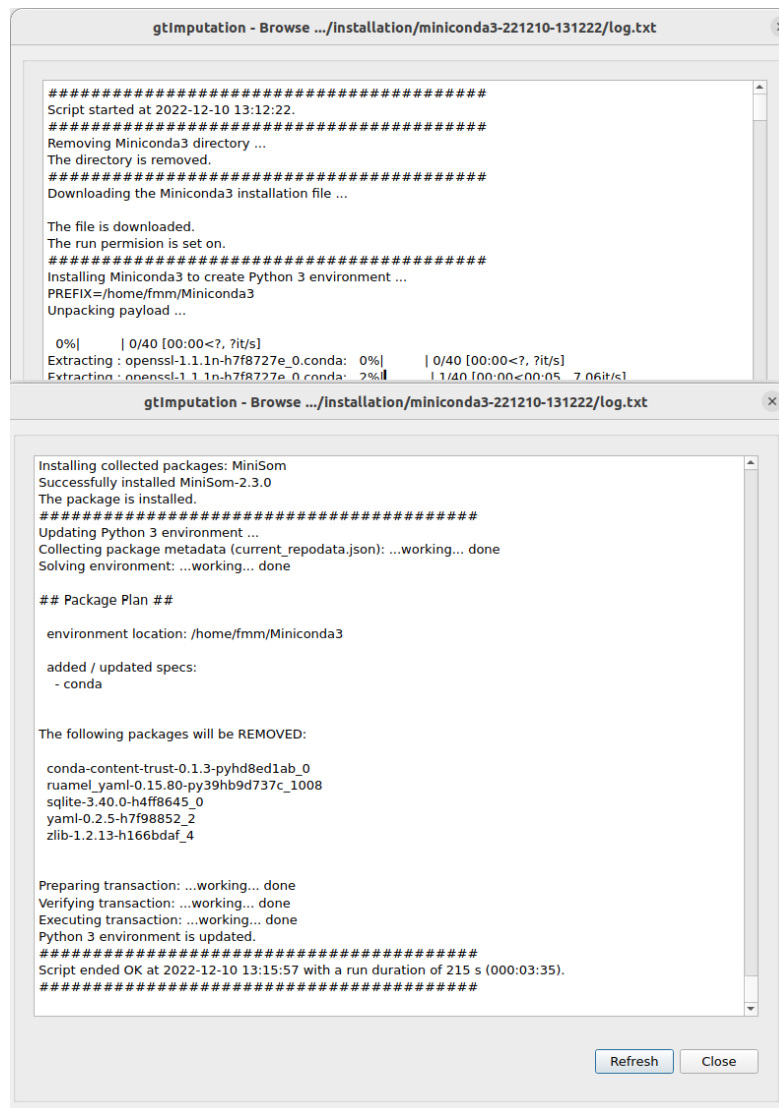
Figure. 11. Begin and end of the log installation of Miniconda3 software.

Consulting submitted processes and troubleshooting

The correct operability of the submitted processes is controlled by logs similar to the Miniconda3 installation (Figure 8). So, the user can monitor the performance of the process at any time and detect problems. Please, confirm that each process is ended before submitting other one.

# A step-by-step example

Here is a brief tutorial on how to use gtImputation.

SOM imputation

Build a genotype database

A genotype database holds sample genotypes of SNPs, linkage disequilibrium between each pair of SNPs and sample kinship got of a VCF file. These data are used by SOM imputation processes. First, select menu item with this path:

Main menu > Imputation > SOM imputation > Buid a genotype database

Then, gtImputation presents a window (Figure 11) where you indicate the *VCF file* typing its path or selecting it with the button *Search*. In this example, the file **variants-selected.vcf** is used. This file is included in the subdirectory sample-data of the gtImputation software.
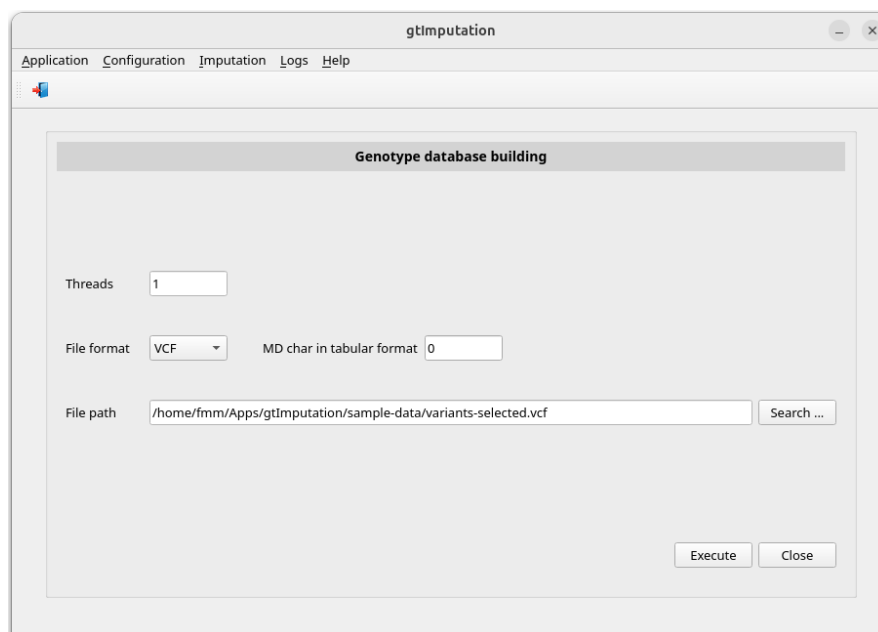
Figure. 11. gtImputation window corresponding to the menu item *Imputation > SOM imputation > Buid a genotype database*.

Once this is done, click on button *Execute*. Then a log dialog is opened with the submission information. (Figure 12).
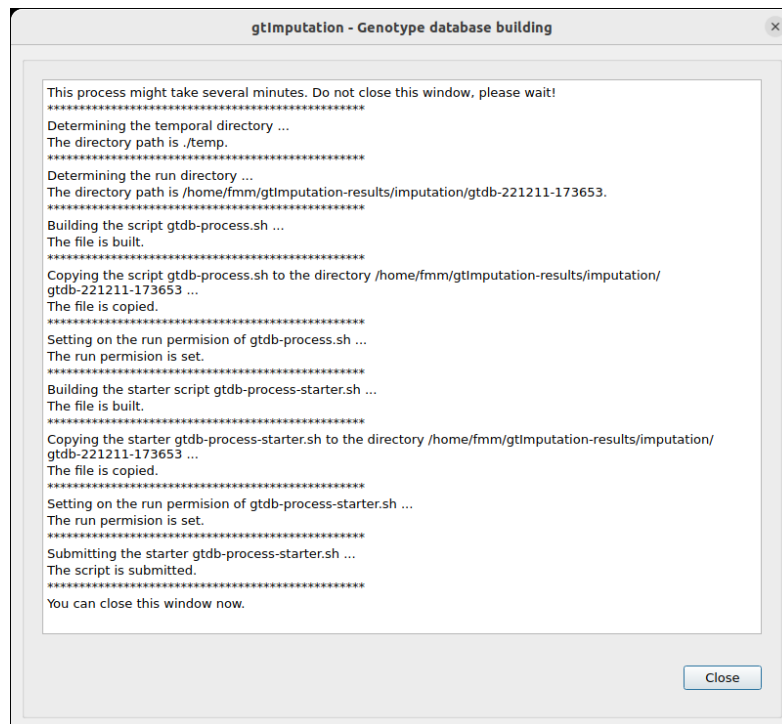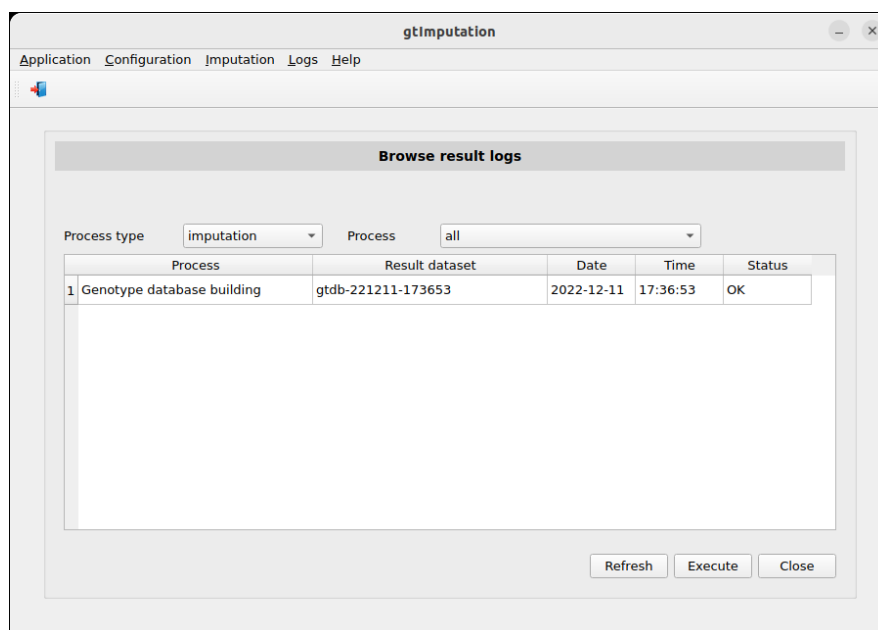
Figure. 12. The dialog showing the submission log of genotype database building.

To view the process log during and after the run, select the menu item with this path:

Main menu > Logs > Result logs

In the raised window (Figure 13), select **imputation** in the *Process type* combo-box*.*

Figure. 13. gtImputation window corresponding to the menu item *Logs - Result logs* after ended the genotype database building.

Double clicking on the row corresponding to **gtdb-221211-173653** or clicking on it and on the button Execute. Then a pop-up window appears with its corresponding log (Figure 14).
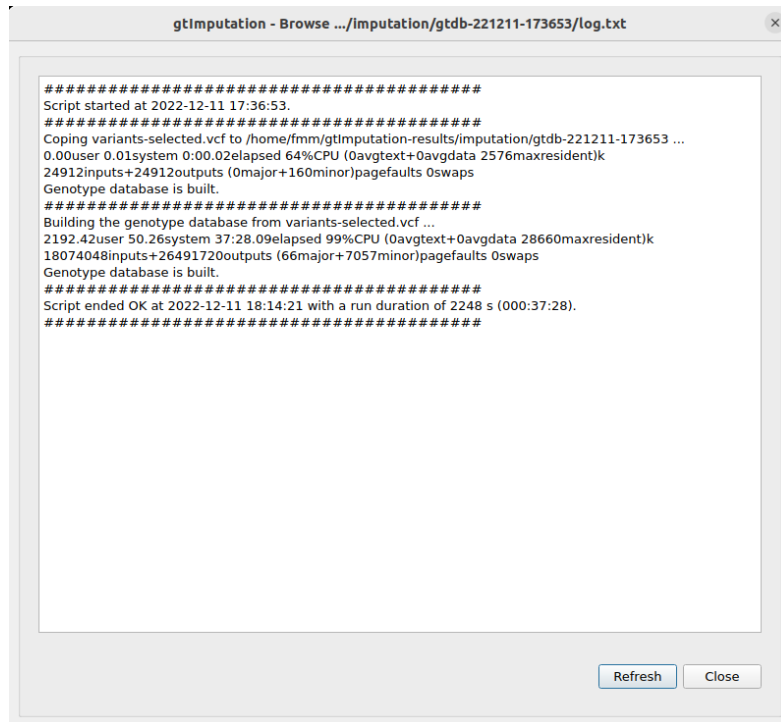
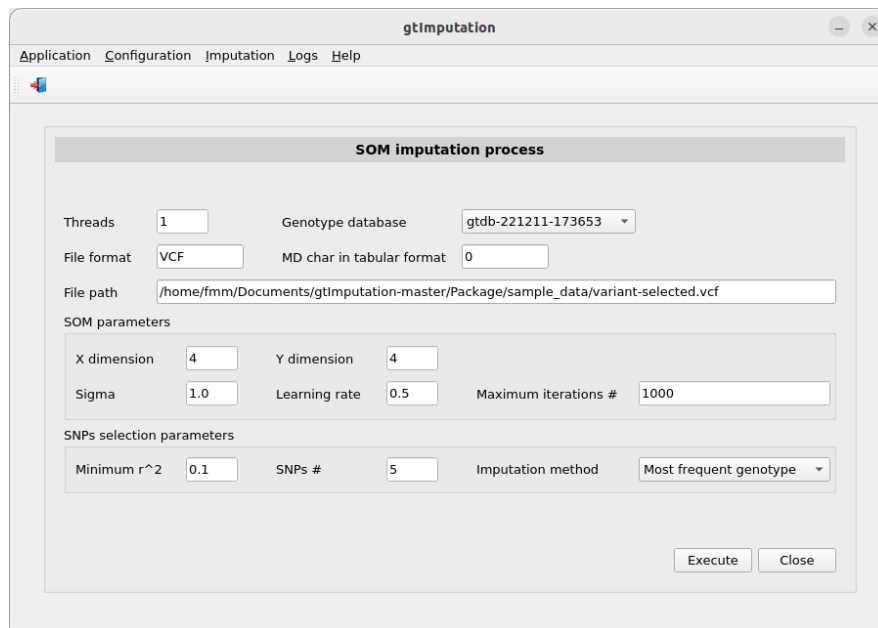**Figure. 14**. The dialog showing the process log of genotype database building.

Run imputation process

Once the genotype dataset is built, you can run one or several SOM imputation processes with the appropriate parameters. To do so, select the menu item with this path:

Main menu > Imputation > SOM imputation > Run imputation process

Then, gtImputation presents a window (Figure 15) where the application shows default parameters. Select the genotype database corresponding to the VCF file you need to impute by clicking its identification in the corresponding combo-box and, below, set the appropriate parameters.

The combo box *Genotype database* is loaded with the process identifications of genotype database building ended OK. In this tutorial, we will use the database built previously. Select the identification **gtdb-221211-173653** in the combo box *Genotype database* and then set **5** in the *X dimension* and *Y dimension* of *SOM parameters*.

Figure. 15. gtImputation window corresponding to the menu item *Imputation > SOM imputation > Run imputation process*.

Then, click the button *Execute*. A log dialog is opened with the submission information. (Figure 16).
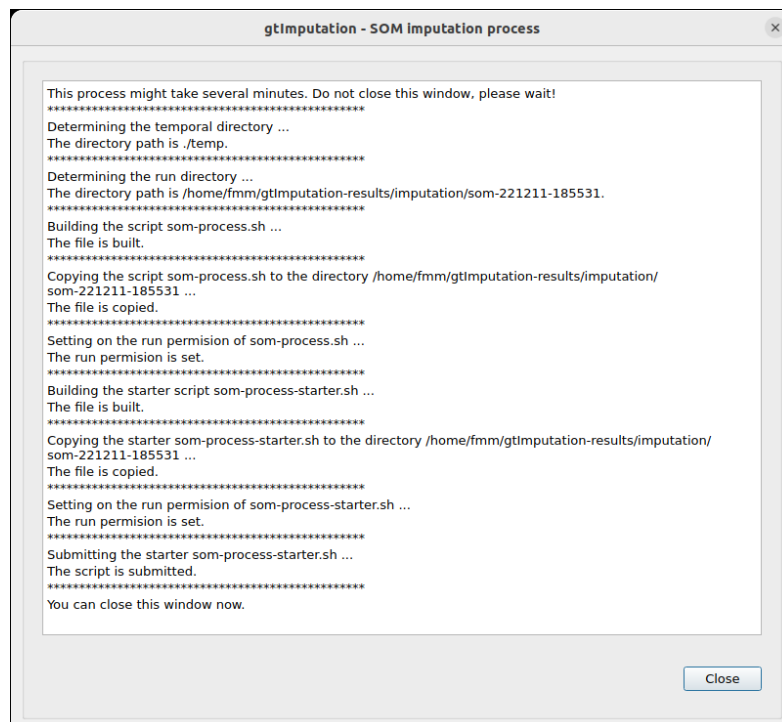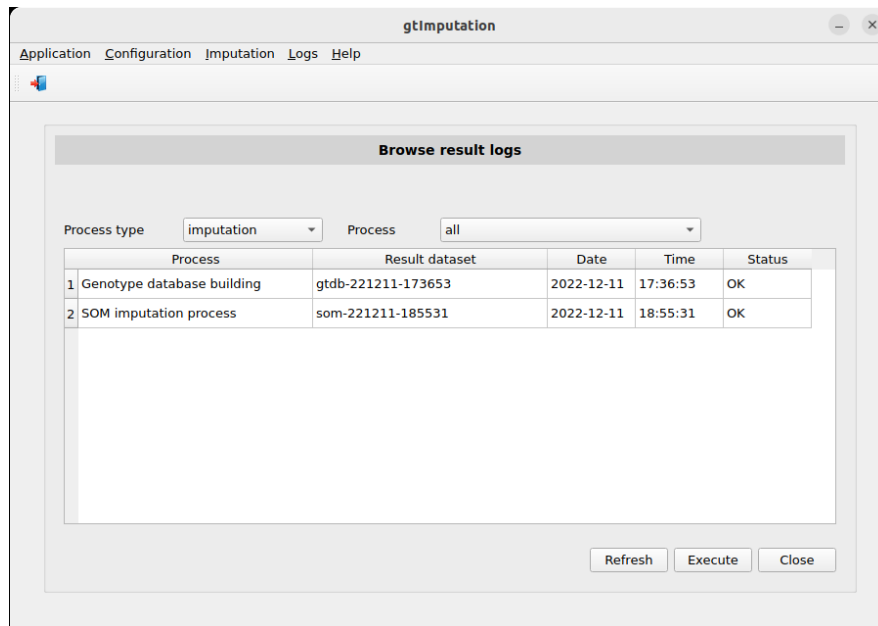

Figure. 16. The dialog showing the submission log of SOM imputation process.

To view the process log during and after the run, select the menu item with this path:

Main menu > Logs > Result logs

In the raised window (Figure 17), select **imputation** in the *Process type* combo-box*.*

Figure. 17. gtImputation window corresponding to the menu item *Logs - Result logs* after ended the SOM imputation process.

Double clicking on the row corresponding to **som-221211-185531** or clicking on it and on the button Execute. Then a pop-up window appears with its corresponding log (Figure 18).
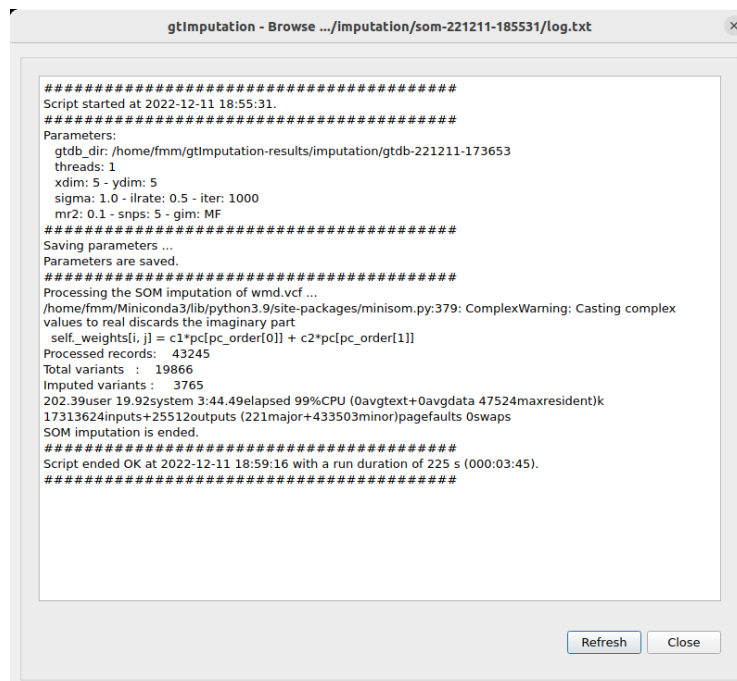
**Figure. 18**. The dialog showing the log of SOM imputation process.

Review imputation process

In order to review the imputation done by the SOM process, click the following menu item:

Main menu > Imputation > SOM imputation > Review imputation process

Then, gtImputation presents a window (Figure 19) where you must select the SOM process identification clicking its identification in the combo box *SOM process* which is loaded with the identifications of SOM processes ended OK. In this example, the identification of the process is **som-221211-185531**. When the identification is selected, the parameters used in the imputation process are shown.
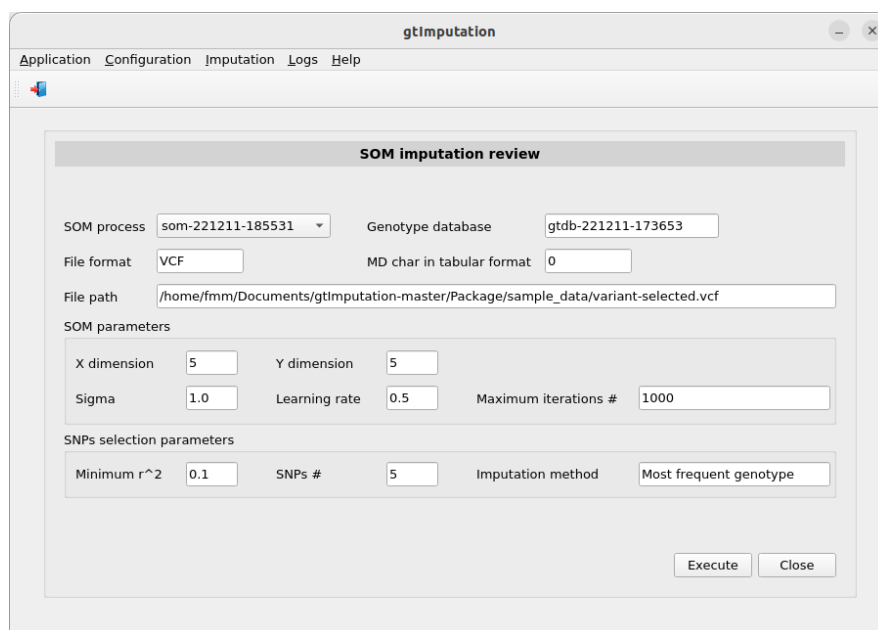
Figure. 19. gtImputation window corresponding to the menu item *Imputation > SOM imputation > Review imputation process*.

Click the execute button. Then gtImputation shows a window with the detailed imputation (Figure 20). Each row corresponds to the position of a sequence and each column represents a sample. The genotypes with missing data in the original VCF file are shown in yellow and the genotypes with data in the original VCF file in dark yellow.

If you click the button *Show map*, a summary map is shown on the left side (Figure 21) and the button changes its title to *Hide map*. Now, if you click on it, the summary map is hidden.

If you click on a sequence identification or on a position of the summary map (Figure 22), the detailed imputation will show only the positions of this sequence and the buttons *Prev sequence* and *Next sequence* will be available. These buttons allow you to browse sequence by sequence the positions of each sequence. To show all sequences, click on the button *All Sequences*
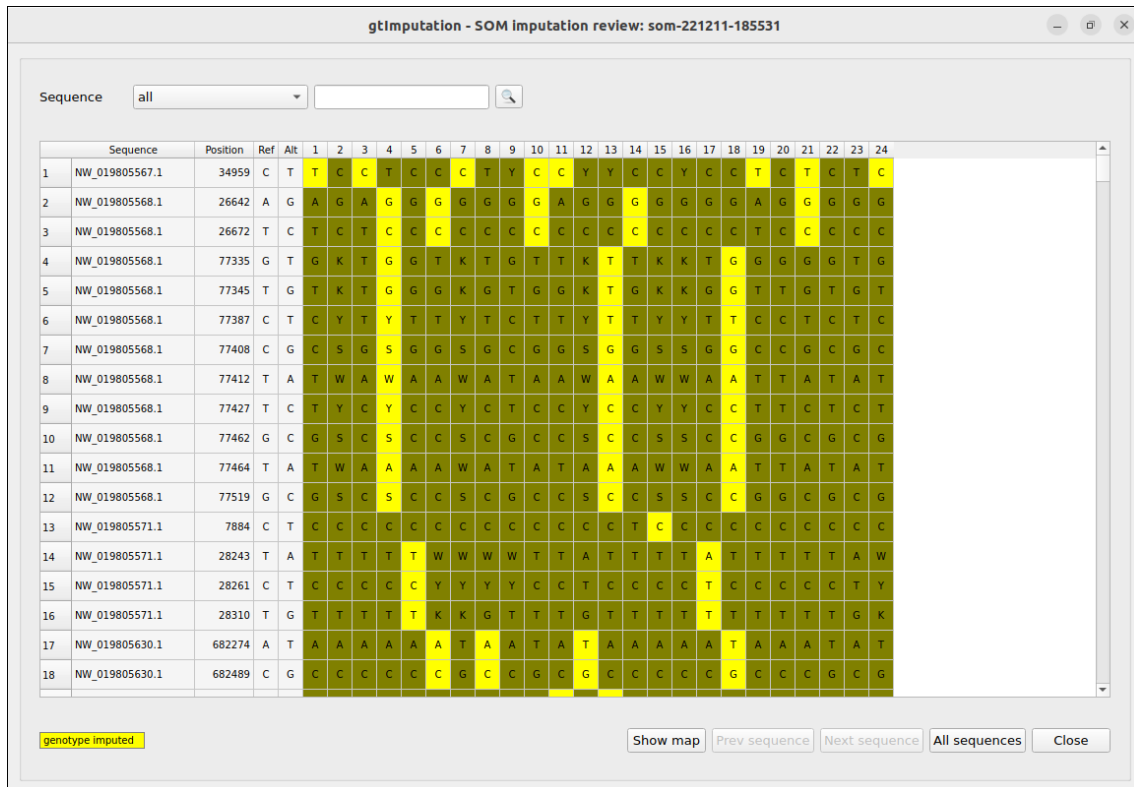
Figure.20. gtImputation window showing the detailed imputation.
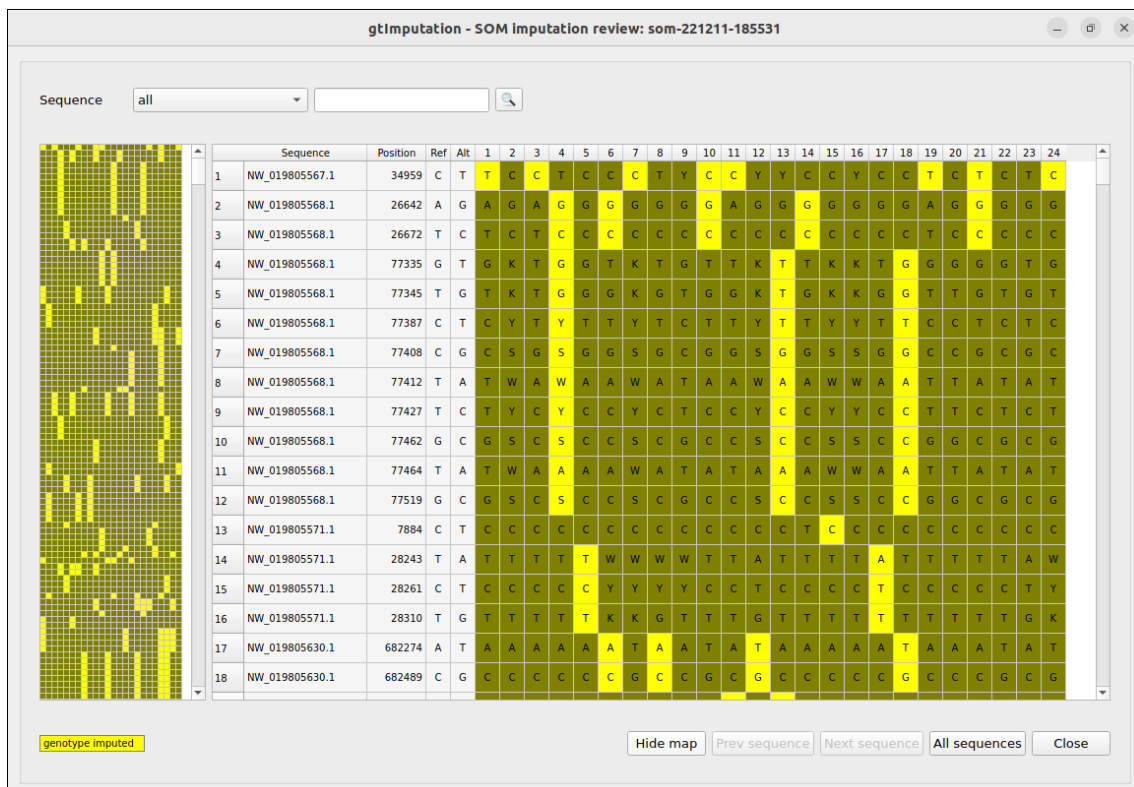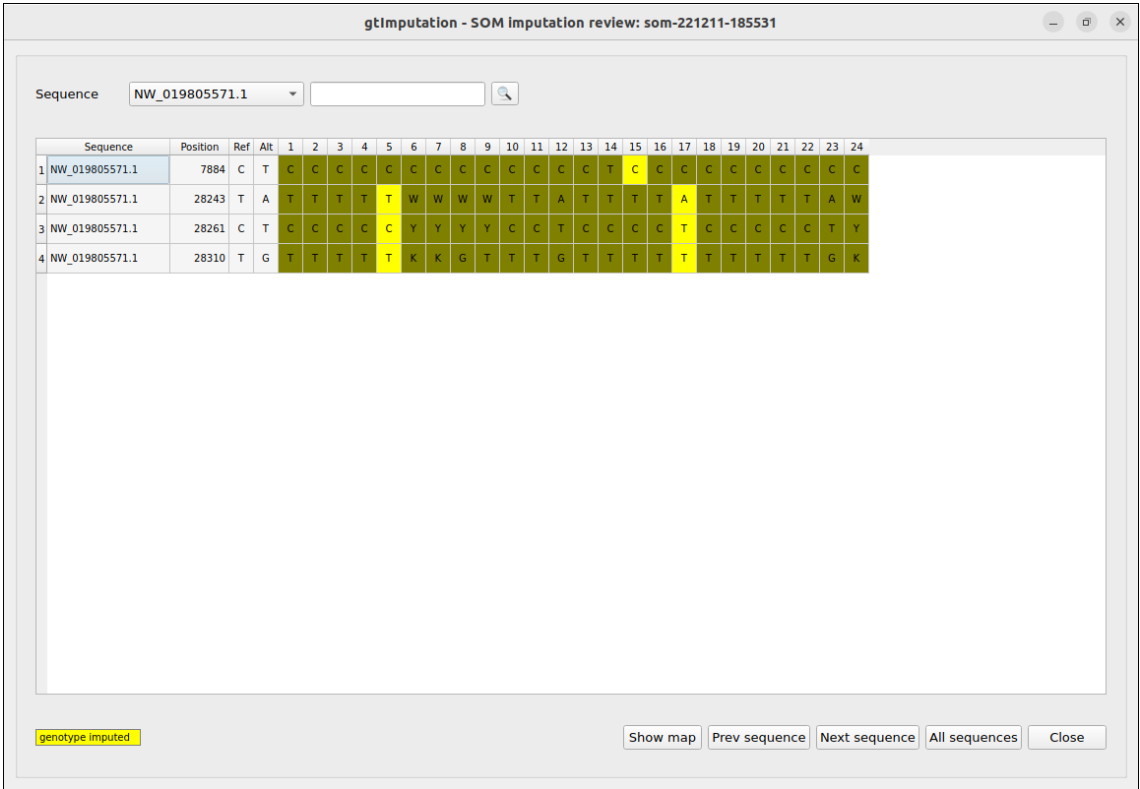

Figure. 21. gtImputation window showing both the detailed imputation and a summary map.

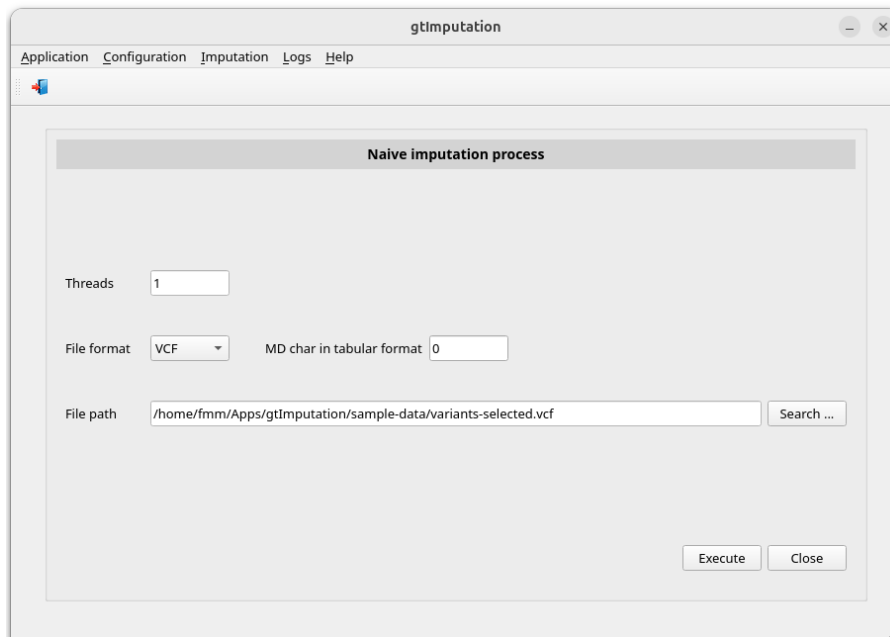**Figure. 22**. gtImputation window showing the positions of a sequence.

Naive imputation

Run imputation process

A naive imputation processes are run from the following menu item:
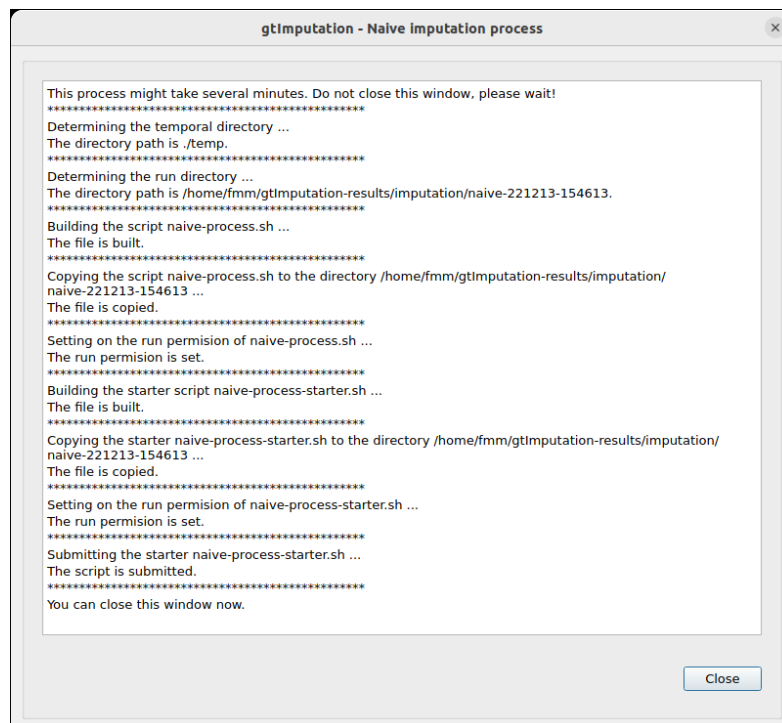
Main menu > Imputation > Naive imputation > Run imputation process

Click this menu item. gtImputation presents a window (Figure 23) where you indicate the *VCF file* typing its path or selecting it with the button *Search*. Like the SOM imputation example, the file **variants-selected.vcf** is used which is included in the subdirectory sample-data of the gtImputation software.

Figure. 23. gtImputation window corresponding to the menu item *Imputation > Naive imputation > Run imputation process*.

Then, click the button *Execute*. A log dialog is opened with the submission information. (Figure 24).
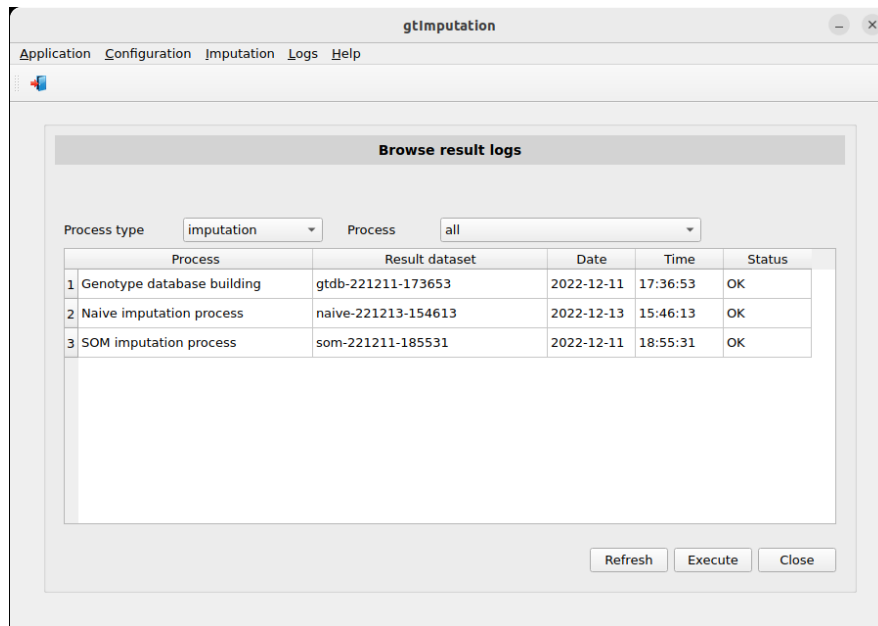

Figure. 24. The dialog showing the submission log of naive imputation process.

To view the process log during and after the run, select the menu item with this path:

Main menu > Logs > Result logs

In the raised window (Figure 25), select **imputation** in the *Process type* combo-box.

Figure. 25. gtImputation window corresponding to the menu item *Logs - Result logs* after ended the naive imputation process.

Double clicking on the row corresponding to **naive-221213-154613** or clicking on it and on the button Execute. Then a pop-up window appears with its corresponding log (Figure 26).
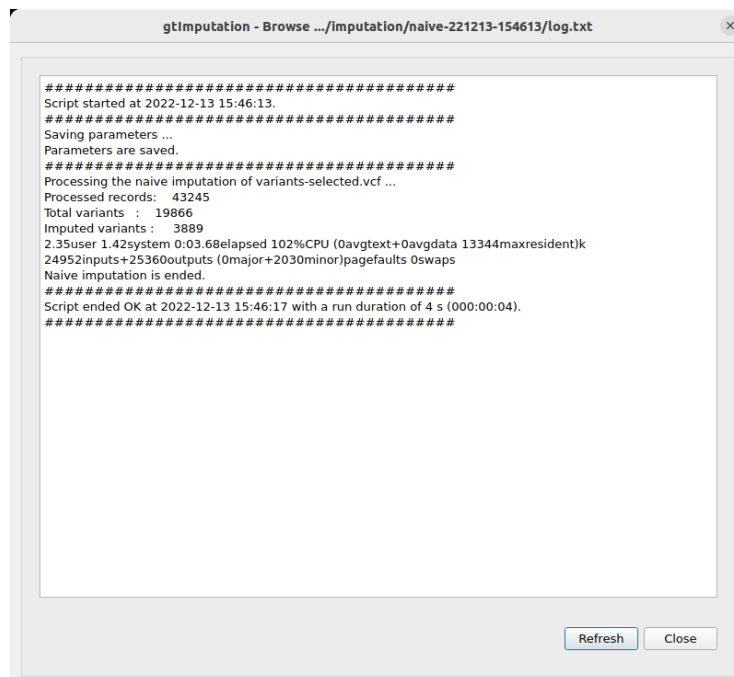
Figure. 26. The dialog showing the log of naive imputation process.

Review imputation process

The review of a naive process is done clicking the following menu item:

Main menu > Imputation > Naive imputation > Review imputation process

Then, gtImputation presents a window (see Figure 27) where you must select the naive process identification clicking its identification in the combo box *Naive process* which is loaded with the identification of the naive processes ended OK. In this example, the identification of the process is **naive-221213-154613**. When the identification is selected, the VCF file used in the imputation process is shown.
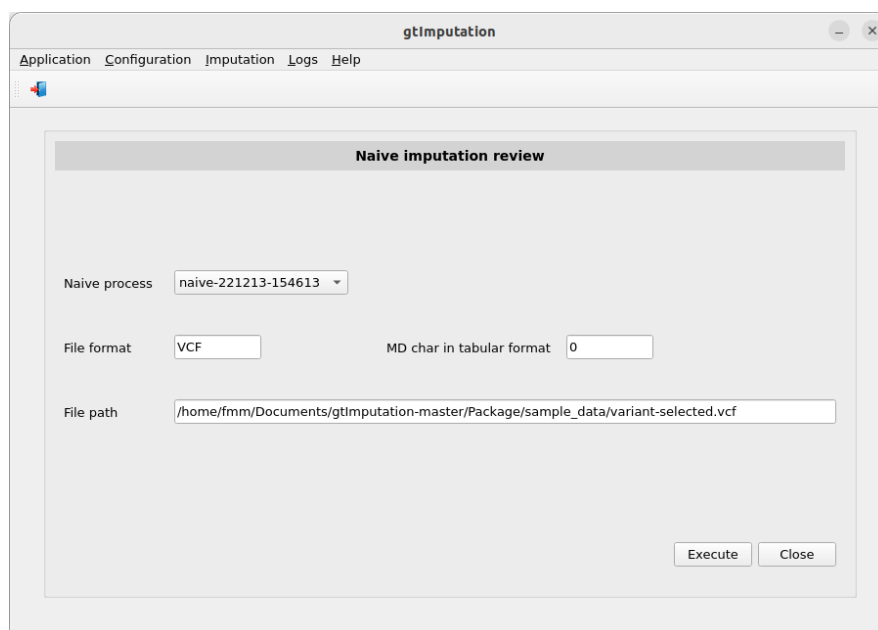
Figure. 27. gtImputation window corresponding to the menu item *Imputation > Naive imputation > Review imputation process*.

Click the execute button. Then gtImputation shows the imputation window. The review of a naive imputation process is similar to the procedure described for the SOM imputation.

Where are the imputed files?

An imputation process generates two files in the result dataset directory whose names are **imputed.vcf** and **imputed.tsv** (in tabular format) with the imputed genotypes in the missing data of the original VCF file. For instance, in this tutorial, the imputed.vcf in the example of the SOM process is in the directory path (see Figure 28):

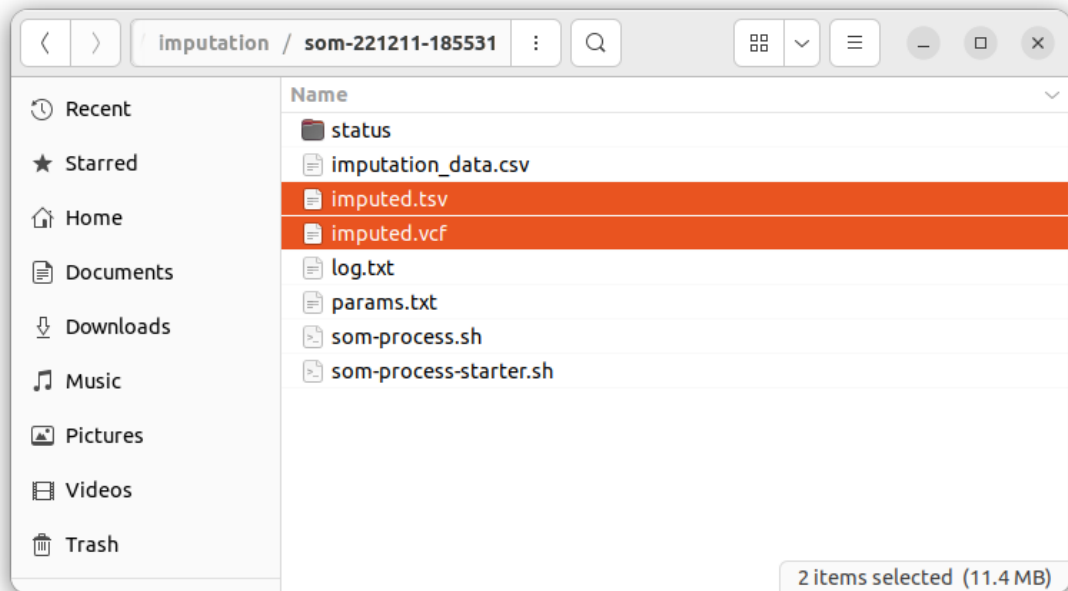**result_dataset_directory/imputation/som-221211-185531**

where *result_dataset_directory* is set in the menu item:

Main menu > Configuration > Recreate gtImputation config file

You can check the current value of the in the following menu item:

Main menu > Configuration > Browse gtImputation config file

Figure. 28. Files generates in the result dataset directory **som-221211-185531**.

Tips about parameter setting

After examining several genotype files with varying missing data levels, we can conclude that for the majority of datasets GTIMPUTATION will offer accurate imputation using the following default parameters:

- SOM size =3x3

- $\sigma = 1$

- $mr^2 = 0.001$

- snps = 5.

However, for datasets with individuals without strong familiar structures, we suggest to consider a 0.1 mr^2 threshold in order to include less SNP genotypes in the SOM vectors.

We do not recommend using rectangular maps bigger than 4x4 because larger maps are expected to produce worse results because higher variability in the quantization error is observed for big maps (De Bodt et al., 2002). Other parameter setting configurations can be easily tested by the user running SOM several times (remember that genotype database is only computed once).

SOM imputation accuracy

Imputation accuracy is largely dependent on the genetic structure / composition of the population(s and on the level of missing data) in our genotype file. SOM imputation has proven to be accurate and precise, and provides overall good results irrespective of the dataset and missing data conditions. Figure 29 presents the results of imputation in four different genotype files from a real experiment in Mediterranean oaks (López de Heredia et al., 2020) using six different imputation methods: (1) a naïve imputation of the most frequent value in the dataset (NAIVE); (2) an imputation based on inverse non-linear principal component analysis (Scholz et al., 2005) as implemented in the PCAMETHODS Bioconductor package (Stacklies et al., 2007) (NLPCA) (3) an imputation based on the random forest algorithm implemented in the R package MISSFOREST (Stekhoven & Bühlmann, 2012) (MISSFOREST); (4) the LD K-nearest neighbor imputation method (Money et al., 2015) implemented in the software TASSEL (Bradbury et al., 2007) (TASSEL LD-KNN); (5) the SOM-based imputation algorithms implemented in GTIMPUTATION; and (6) a haplotype phasing, clustering and imputation method estimating clustering parameters with the EM algorithm using the version of FASTPHASE (Scheet & Stephens, 2006) that does not consider haplotype reference panels (FASTPHASE).
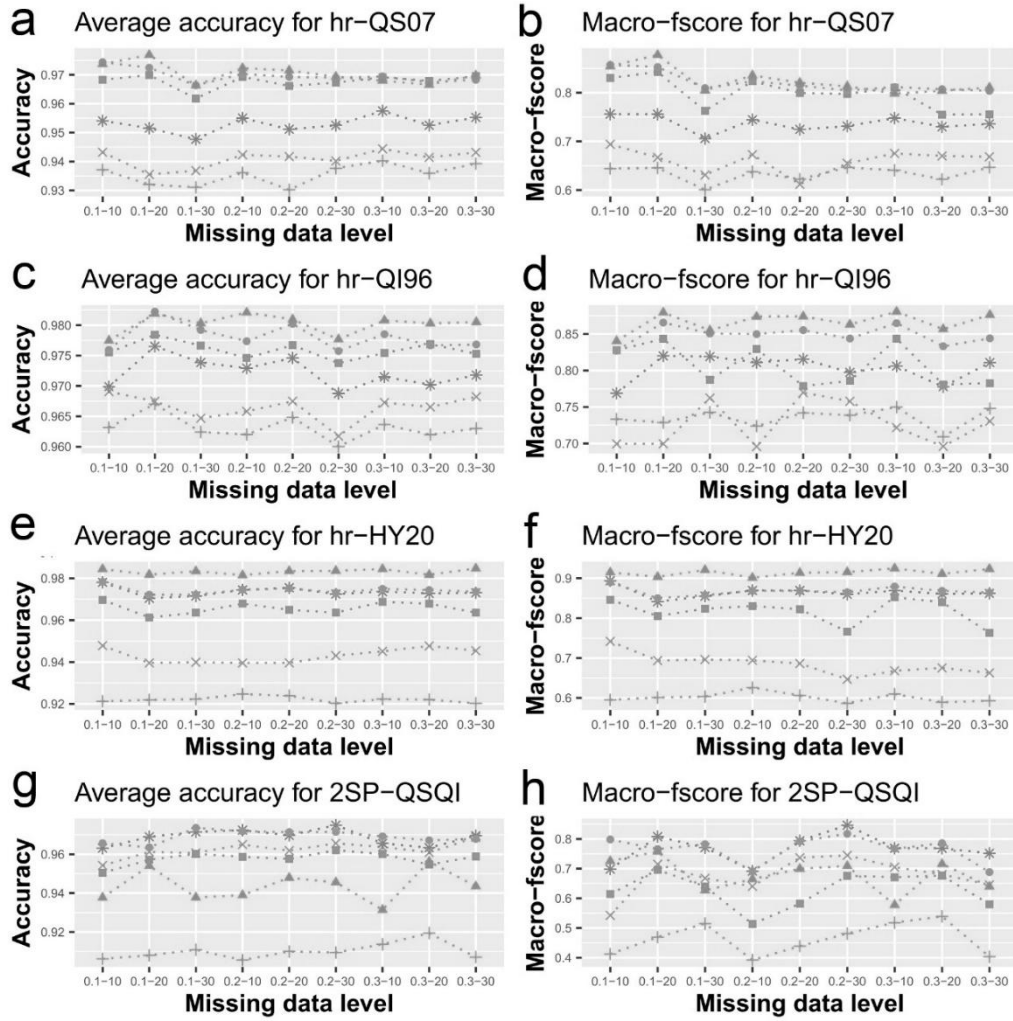
Figure. 29. Comparison of the best average accuracy and maro-F1-score for Dataset 4 and Dataset5 with the four algorithms under several missing data conditions. NAIVE: crosses; NLPCA: blades; MISSFOREST: asterisks; TASSEL LD-KNN: squares; SOM: circles; FASTPHASE: triangles. **a**, **b**: hr-QS07 (94 samples; 1451 SNPs); **c**, **d**: hr-QI96 (93 samples, 2090 SNPs); **e**, **f**: hr-HY20 (30 samples, 4424 SNPs); **g**, **h**: 2SP-QSQI (2 species, 197 samples, 400 SNPs).

Limitations of the method

The current version of the software presents the following limitations:

- Imputation of bi-allelic SNPs only is supported. Multi-allelic loci and other types of variants (indels) are not considered.

- Imputation accuracy may drop in samples with few individuals or in highly heterozygous populations. Also, large proportions of missing data (>50%), may provoke less accurate imputation.

- Samples from more than a single species in the same genotype file can be used as an input. Providing sufficient number of individuals per population are present in the genotype file, SOM provides accurate imputation. Imputation of outgroup samples to the focal species is not recommended, unless several individuals are considered for the outgroup.

- For large datasets with many samples and several thousand SNPs, it must be stressed out that SOM computation times will scale to hours, particularly at the genotype database building stage, but still remaining in reasonable times.

How to cite

If you are using gtImputation, you should cite the following paper:

Mora-Márquez F, Nuño J. C., Soto A. & López de Hereda U. (2024). Missing genotype imputation in non-model species using self-organizing maps. *Molecular Ecology Resources*, e13992. DOI: https://doi.org/10.1111/1755-0998.13992

References

Bradbury, P.J., Zhang, Z., Kroon, D.E., Casstevens, T.M., Ramdoss, Y. & Buckler, E.S. (2007). TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics*, 23(19), 2633-5. doi:10.1093/bioinformatics/btm308

Browning, S.R. (2008). Missing data imputation and haplotype phase inference for genome-wide association studies. *Human Genetics*, 124, 439-450. doi:10.1007/s00439-008-0568-7

De Bodt, E., Cottrell, M., & Verleysen, M. (2002). Statistical tools to assess the reliability of self-organizing maps. *Neural Networks*, 15(8-9), 967-978. doi:10.1016/s0893-6080(02)00071-0

Fu, Y.B. (2014). Genetic Diversity Analysis of Highly Incomplete SNP Genotype Data with Imputations: An Empirical Assessment. *G3 Genes|Genomes|Genetics*, 4(5), 891-900. doi:10.1534/g3.114.010942

Goudet, J., Kay, T., & Weir, B. S. (2018). How to estimate kinship. *Molecular Ecology*, 27(20), 4121–4135. DOI: 10.1111/mec.14833

Kohonen, T. (2001). *Self-Organizing Maps*. 3th edn. Berlin, Heidelberg: Springer.

López de Heredia, U., Mora-Márquez, F., Goicoechea, P.G., Guillardín-Calvo, L., Simeone, M.C. & Soto, Á. (2020). ddRAD Sequencing-Based Identification of Genomic Boundaries and Permeability in Quercus ilex and Q. suber Hybrids. *Frontiers in Plant Science*, 11, 564414. doi: 10.3389/fpls.2020.564414

Marchini, J., Howie, B.N., Myers, S.R., McVean, G., & Donnelly, P. (2007). A new multipoint method for genome-wide association studies by imputation of genotypes. *Nature Genetics*, 39, 906-913. doi:10.1038/ng2088

Money, D., Gardner, K., Migicovsky, Z., Schwaninger, H., Zhong, G.Y. & Myles, S. (2015). LinkImpute: Fast and Accurate Genotype Imputation for Nonmodel Organisms. *G3 Genes|Genomes|Genetics*, 5(11), 2383-2390. doi:10.1534/g3.115.021667

Ragsdale, A. P., & Gravel, S. (2020). Unbiased Estimation of Linkage Disequilibrium from Unphased Data. *Molecular Biology and Evolution*, 37(3), 923–932. DOI: 10.1093/molbev/msz265)

Scheet, P. & Stephens, M. (2006). A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *American Journal of Human Genetics*, 78(4), 629-44. doi:10.1086/502802

Scholz, M., Kaplan, F., Guy, C.L., Kopka, J. & Selbig, J. (2005). Non-linear PCA: a missing data approach. *Bioinformatics*, 21(20), 3887-3895. doi:10.1093/bioinformatics/bti634

Stacklies, W., Redestig, H., Scholz, M., Walther, D. & Selbig, J. (2007). pcaMethods – a Bioconductor package providing PCA methods for incomplete data. *Bioinformatics*, 23, 1164-1167. doi: 10.1093/bioinformatics/btm069

Stekhoven, D.J. & Bühlmann, P. (2012). MissForest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1), 112–118. doi:10.1093/bioinformatics/btr597

Vettigli, G. (2018). MiniSom: minimalistic and NumPy-based implementation of the Self Organizing Map. URL: https://github.com/JustGlowing/minisom/

Appendix A: File formats

gtImputation supports two types of file formats: VCF and tabular format.

The VCF specification can be found at the URL https://samtools.github.io/hts-specs/VCFv4.3.pdf.

In the tabular format, records are separated by newlines and record data are separated by tab characters. In the header record, the first data is a literal like "ID/SNP" and then the SNPS identification. Sample records are formed by the sample identification and the sample genotypes of each SNP. The two alleles of each genotype are separate data. Missing data can be represented by a character other than those used with genotypes, e.g., "0". Usually, ".tsv" and ".txt" extensions are used for tabular format. A tabular format example is shown in Figure A-1.

```
ID/SNP» NW_019805647_1-36332_3»   NW_019805656_1-330489_3» NW_019805923_1-290863_3» NW_019805955_1-9492_3¬
A05-17» C»  C»  A»  A»  A»  A»  C»  C¬
A05-27» C»  C»  A»  A»  A»  A»  C»  C¬
A05-67» C»  C»  A»  A»  A»  A»  C»  C¬
A07-05» C»  C»  A»  A»  A»  A»  C»  C¬
A07-62» C»  C»  A»  A»  A»  A»  C»  C¬
A07-94» C»  C»  A»  A»  A»  A»  C»  C¬
A09-07» C»  C»  A»  A»  A»  A»  C»  C¬
A09-17» C»  C»  A»  A»  A»  A»  C»  C¬
A10-19» C»  C»  A»  A»  A»  A»  C»  C¬
A10-69» C»  C»  A»  A»  A»  A»  C»  C¬
E28-40» G»  G»  G»  G»  C»  C»  T»  T¬
E28-96» G»  G»  G»  G»  C»  C»  T»  T¬
E31-03» G»  G»  G»  G»  C»  C»  T»  T¬
E31-87» G»  G»  G»  G»  C»  C»  T»  T¬
E41-26» G»  G»  G»  G»  C»  C»  T»  T¬
E41-61» G»  G»  G»  G»  C»  C»  T»  T¬
E96-07» G»  G»  G»  G»  C»  C»  T»  T¬
E96-56» G»  G»  G»  G»  C»  C»  T»  T¬
IDE_01» G»  G»  G»  G»  C»  C»  T»  T¬
IDE_07» G»  G»  0»  0»  0»  0»  0»  0¬
IDE_08» 0»  0»  0»  0»  0»  0»  0»  0¬
IDE_10» G»  G»  G»  G»  C»  C»  T»  T¬
IDE_11» G»  G»  G»  G»  C»  C»  T»  T¬
IDE_12» G»  G»  G»  G»  C»  C»  T»  T¬
IDE_14» 0»  0»  0»  0»  0»  0»  0»  0¬
IDE_16» G»  G»  G»  G»  C»  C»  T»  T¬
```

Figure. A-1. Example of tabular format file with 4 SNPs and 26 samples