# EpiMolBio

## Analysis of Genetic Variability

## User Manual v 1.0

**HIV-1 MOLECULAR EPIDEMIOLOGY LABORATORY**
IRYCIS-Microbiology Department
Ramón y Cajal Hospital

# EpiMolBio

## Analysis of Genetic Variability

License: Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 (registry number 2305114294344)

Developer: Roberto Reinosa

Collaborator: Paloma Troyano-Hernáez

Coordination and Supervision: África Holguín

Copyright: Roberto Reinosa Fernández and Biomedical Research Foundation of Ramón y Cajal University Hospital (FIBioHRC)

# INTRODUCTION

Welcome to EpiMolBio, a free bioinformatics program for the analysis of genetic and protein sequences. This software allows you to analyze genetic sequences and obtain information about their structure, function, and evolution. Whether you are an experienced bioinformatician or a beginner biologist, this program is designed to be intuitive and user-friendly, with powerful and customizable tools and workflows to suit your needs.

EpiMolBio version 0.1 allows you to:

• Identify resistance mutations in HIV-1 and HIV-2, their frequency, and the conservation of the three HIV Pol proteins: Protease, Reverse Transcriptase, and Integrase.

• Track specific proteins within the SARS-CoV-2 genome.

• Analyze the genetic variability of any group of sequences, providing information about their conservation, frequency of polymorphisms and mutations, generating consensus sequences, and Wu-Kabat index tables.

• Analyze the similarity of any group of genetic sequences.

• Track specific sequences of interest within one or multiple complete genetic sequences.

• Perform simple or multiple alignments of over 100,000 sequences.

• Perform various other functions on the sequences using the Tools function: from nucleotide to amino acid translation, to file editing.

• Program functions to automate the execution of multiple chained functions.

EpiMolBio has implemented multiple functions for the study of genetic variability of two of the pathogens currently causing pandemics: the human immunodeficiency virus (HIV) causing AIDS, and the SARS-CoV-2 virus causing COVID-19, responsible for more than 85 and 450 million infections worldwide, respectively.

The program is continuously improved to meet the needs of new research projects. EpiMolBio can be applied to the study of genetic variability of pathogens and the analysis of genetic markers associated with diseases and genes or proteins of biomedical or biological interest. This program has been designed to support research on new diagnostic, prognostic, or therapeutic strategies for pathogens or diseases, as well as epidemiological studies of biomarker presence associated with diseases or biological processes of interest.

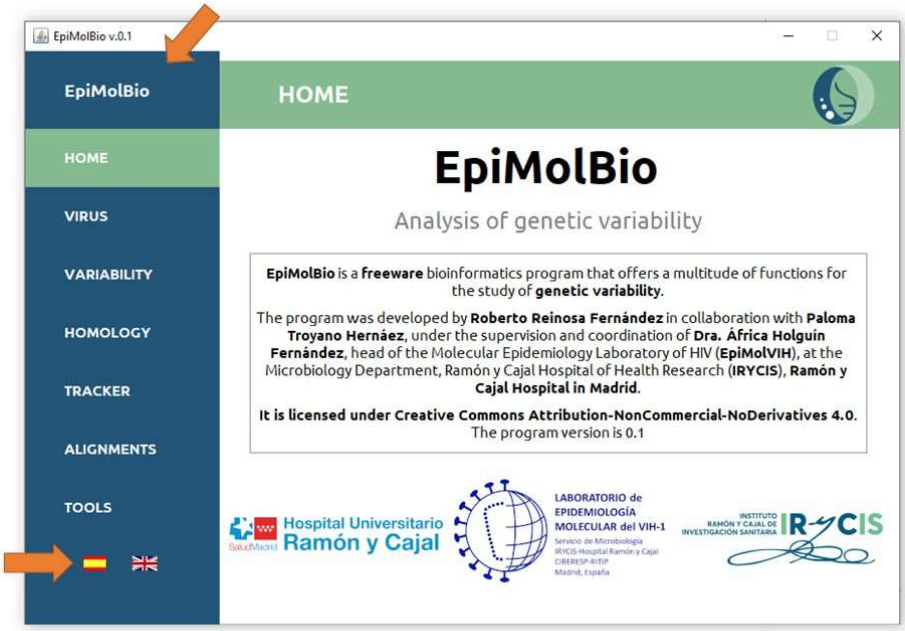
This User Manual provides detailed instructions on how to carry out all these tasks. At the website www.epimolbio.com, we provide comprehensive documentation, video tutorials, and user assistance to help you make the most of the program and achieve your research goals.

This program is cross-platform and can be used on Windows, Mac, and Linux. It is also portable, meaning it can be simply copied and pasted to the desktop or a folder without the need for installation. It only requires Java 11 or higher, which is available for free at https://www.java.com.

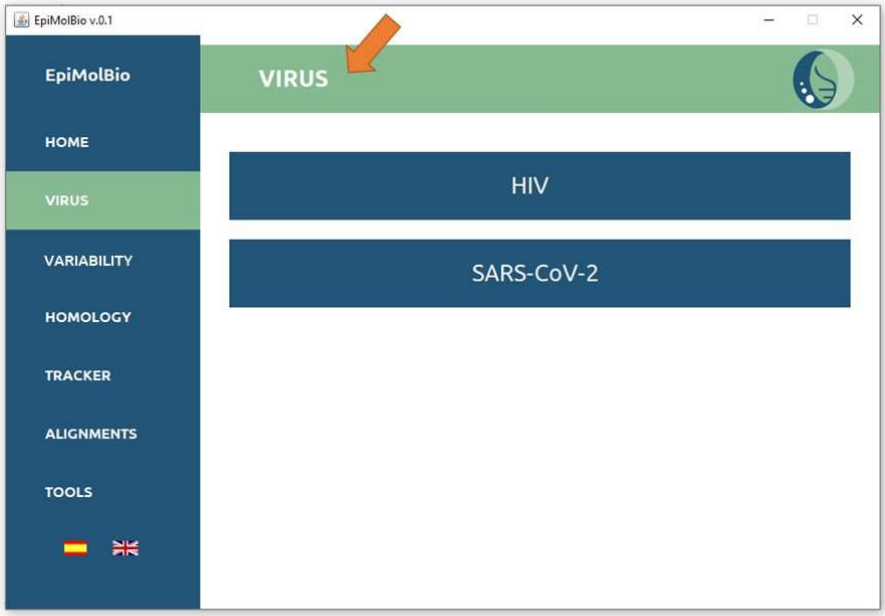
# OVERVIEW

The EpiMolBio **interface** is designed to be as simple and intuitive as possible, so that prior knowledge of programming or bioinformatics is not necessary.

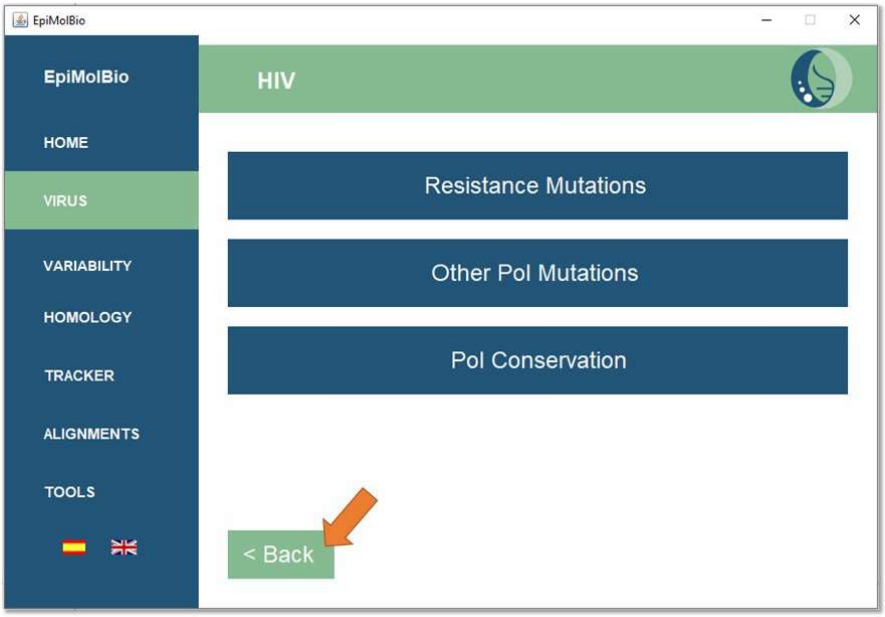
On the left, you will find the **Menu** through which you can access various functions. At the bottom, there are options to select the program's **language** (Spanish or English).




The header will always indicate the last option that has been chosen.



The '**Back**' button allows you to go back to the previous menu.




In most functions, you will find the following buttons:

**Input**: Allows you to choose the folder or subfolders containing the elements to be analyzed.

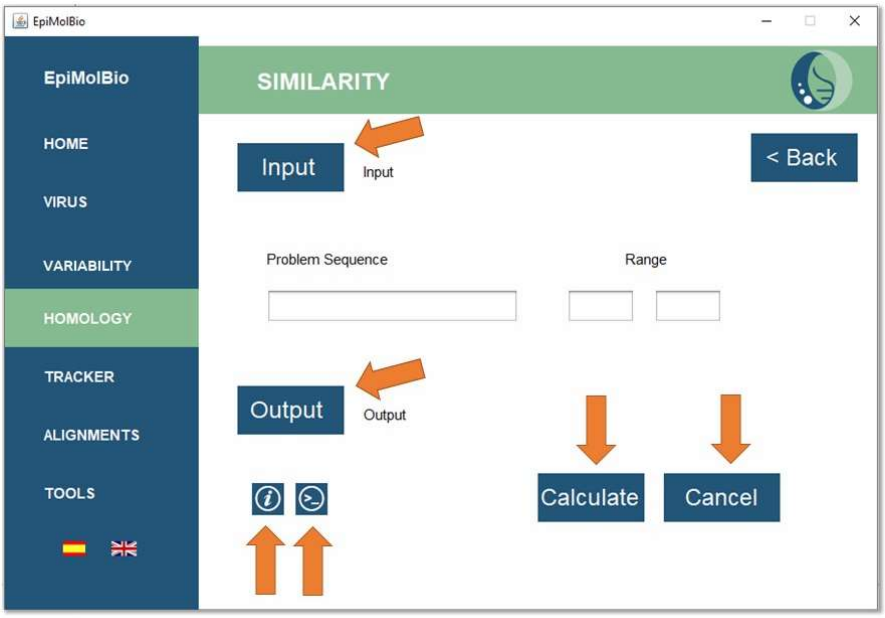
**Output**: Enables you to select the name of the file or folder to store the analysis results.

**Calculate**: Executes the analysis based on the selected options.

**Cancel**: Interrupts an ongoing analysis if needed.

**Information Button**: Provides a summary of the steps required to carry out the analysis and the output formats supported by each function.

**Function Programming Button**: Records the selected options, which can be copied as input for the 'Function Programming' tool.



In most of EpiMolBio's functions, the **input file** types are sequence files in the **.fasta** format, which can be located within a folder (typically) or in subfolders within a folder when certain functions are required. Only in specific cases will another file format be needed, such as .txt text files.

While some functions are designed to work with amino acid .fasta files, they can also be used with nucleotide files if you first use the **Find and Replace** tool in **File Editing** to replace 'N' (unknown nucleotide) with '?' (unknown amino acid). This will exclude these from the analysis, as explained in detail in the **Tools** section of this manual.

The majority of **output files** from EpiMolBio analyses are in **.html**, **.csv**, or **.fasta** formats. In many functions, you can choose different .html output **formats**. These can be viewed in a web browser and copied and pasted into Excel or Word for modifications.

Each section of the manual includes a **Step-by-Step** guide, which explains graphically how to perform the analyses with each function.

**Color code**: In several output files, residues, percentages, or table cells will be colored based on their frequency percentage, following the color code below:

Purple: x = 100%*

Red: 100% ≤ x ≥ 90%

Orange: 90% < x > 75%

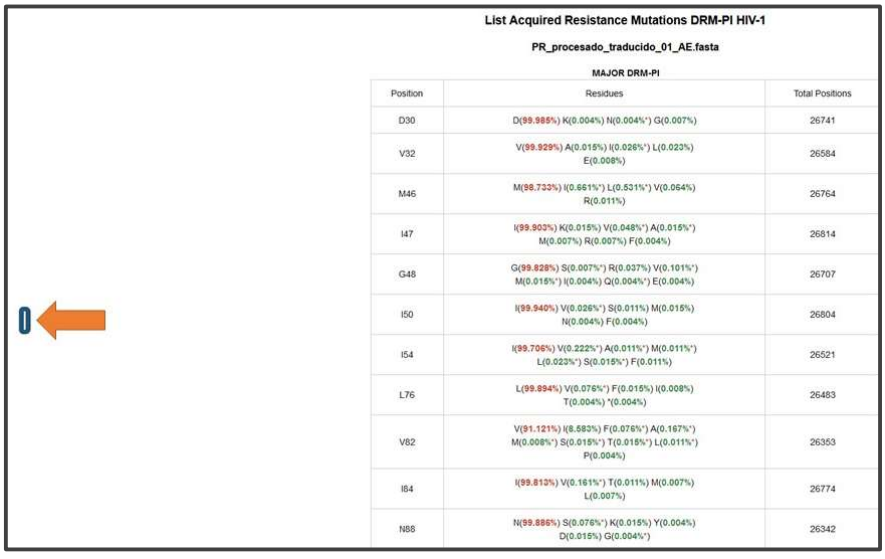
Yellow: 75% ≤ x > 50%

Blue: 50% ≤ x ≥ 10%

Green: x < 10%

* Purple color corresponds to a true 100%, meaning that absolutely all sequences have that residue at that position. If 100% was represented in red, the percentage would be rounded (e.g., from 99.9995 to 100%). EpiMolBio provides percentages with a varying number of decimal places, ranging from 3 to 5, depending on the function.

The color code can be checked in the .html output file by clicking on the blue symbol.



List Acquired Resistance Mutations DRM-PI HIV-1

PR_procesado_traducido_01_AE.fasta

MAJOR DRM-PI

| Position | Residues | Total Positions |
|---|---|---|
| D30 | D(99.985%) K(0.004%) N(0.004%*) G(0.007%) | 26741 |
| V32 | V(99.929%) A(0.015%) I(0.026%*) L(0.023%) E(0.008%) | 26584 |
| M46 | M(98.733%) I(0.661%*) L(0.531%*) V(0.064%) R(0.011%) | 26764 |
| I47 | I(99.903%) K(0.015%) V(0.048%*) A(0.015%*) M(0.007%) R(0.007%) F(0.004%) | 26814 |
| G48 | G(99.828%) S(0.007%*) R(0.037%) V(0.101%*) M(0.015%*) I(0.004%) Q(0.004%*) E(0.004%) | 26707 |
| I50 | I(99.940%) V(0.026%*) S(0.011%) M(0.015%) N(0.004%) F(0.004%) | 26804 |
| I54 | I(99.706%) V(0.222%*) A(0.011%*) M(0.011%*) L(0.023%*) S(0.015%*) F(0.011%) | 26521 |
| L76 | L(99.894%) V(0.076%*) F(0.015%) I(0.008%) T(0.004%) *(0.004%) | 26483 |
| V82 | V(91.121%) I(8.583%) F(0.076%*) A(0.167%*) M(0.008%*) S(0.015%*) T(0.015%*) L(0.011%*) P(0.004%) | 26353 |
| I84 | I(99.813%) V(0.161%*) T(0.011%) M(0.007%) L(0.007%) | 26774 |
| N88 | N(99.886%) S(0.076%*) K(0.015%) Y(0.004%) D(0.015%) G(0.004%*) | 26342 |

Some functions in EpiMolBio require the input of a **reference sequence**, which should be entered without spaces or line breaks. Alignment files in .fasta format downloaded from public sequence databases often include line breaks. An easy way to remove the line breaks is to copy the reference sequence from the alignment file or one downloaded from GenBank using tools like Pinetools (https://pinetools.com/es/eliminar-saltos-linea) or a similar tool.

The **.csv** outputs can be opened with Excel. However, some special characters like accents may not display correctly. To correct this, open the .csv file in Excel, go to Data, select Get Data, choose From File, and then From Text/CSV. Load the .csv file and transform the data. The file will open in the Power Query Editor. Check the characters and select Close & Load. If the characters are still altered, change the data source by selecting Data Source Settings, click Change Source, select Unicode (UTF-8) option, and then click OK, Close, and Load.

**Common abbreviations:**

**DRM**: resistance mutations

**AA**: amino acid

**NT**: nucleotide

# EPIMOLBIO FUNCTIONS

# Summary of EpiMolBio Program Functions

| Main Section | Sub-section | Function |
|---|---|---|
| **VIRUS** | HIV<br>Resistance Mutations (Acquired* and Transmitted) | Calculates the percentage of acquired/transmitted DRM, relative to a reference sequence, from AA sequences of HIV Pol proteins and Capsid. |
| | Other Pol Mutations | Detects any mutation (not only DRM) in HIV-1 or HIV-2 from sequences of Pol proteins, and their percentage relative to the reference sequence. |
| | Pol Conservation | Generates a table displaying the most prevalent AA and its corresponding percentage at each position within the selected Pol protein sequence, facilitating the identification of the protein's most conserved residue for each position |
| | SARS-CoV-2 Protein Tracker | Provides sequences in .fasta format of the chosen proteins from SARS-CoV-2, based on complete genomes, in NT or AA. |
| **VARIABILITY** | Polymorphisms* | **Mutated Positions** and **Mutations Table** enable the detection of polymorphisms, providing information about their location and frequency of occurrence using any sequence introduced by the user as a reference.<br>**Markers** allows the detection of mutations exclusive to each file compared to the rest of the input files.<br>**Multiple Mutations** enables the detection of mutation combinations providing their frequency of occurrence.<br>**Codon** function detects the codons that are different from those in the reference sequence and their frequency of occurrence.<br>**Mutations by Position** allows the detection of residues at one position or several combined positions providing their frequency of occurrence. |
| | Conservation* | Determines the level of conservation of sequences of interest by reporting the most prevalent residue and its corresponding percentage. Additionally, it generates consensus sequences Codon function presents the most conserved codon from an analyzed nucleotide sequence. |
| | Consensus | Provides consensus sequences and consensus of consensuses by performing multiple rounds of analysis. |
| | Wu-Kabat Coefficient | Provides the Wu-Kabat variability coefficient of protein sequences to study the susceptibility of an amino acid position to evolutionary replacements. |
| | Mutation Frequency | Generates a set of parameters related to the frequency of mutations in a group of sequences, such as mutation frequency, conservation percentage, and average mutations per sequence. |
| **HOMOLOGY** | Similarity | Searchs for a user-introduced target sequence among the sequences in the input file, obtaining the proportion of sequences per file that contain the target sequence. |
| | Partial Similarity | Compares a user-introduced sequence with the input sequences to search for similar regions between them, defining the percentage of similarity between the sequences. |
| | Search for Conserved Sequences | Extracts conserved sequence fragments from a set of input sequences. Allows searching within a specific region, choosing the fragment length, and establishing the conservation percentage. |
| **TRACKER** | Similarity | Searchs for target sequences of interest within a set of longer sequences based on a reference sequence. |
| | Flanking | Searchs for proteins within a set of complete genomic sequences using the flanking sequences of the target protein. |
| **ALIGNMENTS** | Multiple Alignments | Aligns AA and NT sequences using the MUSCLE v3.8.31 program. |
| | Dot Plot | Generates a graphic where sequences are compared by plotting points on a two-dimensional matrix, with each axis representing a sequence. |
| | Delete Insertions | Automatically removes insertions from a sequence with respect to a reference with gaps after performing the alignment. |
| **TOOLS** | File Editing<br>Merge Files | Combines multiple .fasta files into a single .fasta file. |
| | Unique Sequences | Removes duplicated sequences from one or multiple input .fasta files. |
| | Sequence Search | Filters sequences from .fasta files that contain one or multiple mutations chosen by the user. |
| | Find and Replace | Replaces a series of characters with others in both the header and the genetic sequence of one or multiple .fasta files. |
| | Filters<br>Header Filtering | Filters one or multiple ".fasta" format files using parameters from their headers. |
| | Specific Filter | Filters sequences from .fasta format files that have a specific set of characters in their headers. |
| | Partial Sequence Filtering | Filters sequences from .fasta format files based on their quality, depending on the quantity of "?" (sequences in AA) or "N" (sequences in NT) they contain. |
| | Translation | Translates .fasta sequences from NT to AA. |
| | Count Sequences | Counts the total number of sequences in one or multiple .fasta files, or how many of those sequences contain mutations. |
| | Function Programming | Automates the program's functions by chaining them together to be executed sequentially without manual intervention. |

* Individual and codon analysis. Abbreviations: DRM, drug resistance mutations;AA, amino acid; NT, nucleotide

# I. VIRUS

In this section, you will find specific tools for the analysis of the Human Immunodeficiency Virus (HIV) and the Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2).

# I.1. HIV

# I.1.A) RESISTANCE MUTATIONS

This function enables the detection of HIV drug resistance mutations (DRM) from Pol protein and Capsid sequences in amino acids. It calculates the percentage of DRM and classifies it according to the reference sequence. EpiMolBio uses HXB2 (NCBI K03455.1) as the reference sequence for HIV-1 and ALI (NCBI AF082339) for HIV-2. The program does not detect deletions or insertions that affect susceptibility to antiretroviral drugs.

The analysis can be performed for **individual mutations**, only in amino acids, or by **codons** or triplets in nucleotides.

**Individual mutation:**

In the analysis of **individual mutations**, both **acquired DRM for HIV-1,** covered in Stanford HIV Drug Resistance Database v9.6, **and acquired DRM for HIV-2,** covered in HIV-2 EU Tool v.2 2015, Charpentier et al. 2015, Tzou et al. 2020, and Troyano-Hernáez P et al. 2021, can be studied. You can also analyze the presence of **transmitted DRM (SDRMs) for HIV-1** according to the WHO list, Bennet et al. 2009, and according to the Stanford HIV Drug Resistance Database v9.6 (https://cms.hivdb.org/prod/downloads/resistance-mutation-handout/resistance-mutation-handout.pdf), with the latest update on 20-03-2024. The complete list of mutations integrated into the EpiMolBio program can be found in **Appendix I**.

The analysis focuses on mutations detected in the HIV *pol* gene's proteins: Protease, Reverse Transcriptase, or Integrase, and in the Capsid. Both gaps (-) and question marks (?) are excluded from the analysis. Stop codons are indicated with a black asterisk (*).

In the case of **acquired mutations**, you need to select the type of HIV to analyze and establish the reference sequence: **HIV-1 or HIV-2**.

For both cases, the **input file** should be a **folder** containing exclusively the aligned sequences of the Pol protein you want to analyze (Protease, Reverse Transcriptase, Integrase, or Capsid) in amino acids and in **.fasta** format. This folder may contain a single file or multiple .fasta files if you want to analyze the same sequences in different groups (e.g., files split by HIV variants, country of origin, year, etc).

The **output file** will be an **.html** file. You need to select the output folder where you want the generated files to appear and name the files with .html extension.

In the '**DRM Type**' field for **Acquired Mutations** or '**SDRM Type**' in **Transmitted Mutations**, you should select the **type of mutation** you want to study, choosing mutations against:

**PI**: Protease Inhibitors (input: Protease)

**NRTI**: Nucleoside Reverse Transcriptase Inhibitors (input: Reverse Transcriptase)

**NNRTI**: Non-Nucleoside Reverse Transcriptase Inhibitors (input: Reverse Transcriptase)

**INSTI**: Integrase Inhibitors (input: Integrase)

**CAI**: Capsid Inhibitors (input: Capsid)

In the '**Output Format**' field, you can choose between three types of output formats: **list, table, and summary table**.

In all three cases, the analysis will provide the detected **mutations** in the input sequences, their **classification** according to the Stanford v9.6 classification (https://hivdb.stanford.edu/page/release-notes/#drm.classification), and their **percentage** with respect to the total valid positions, as indicated by the **color code** described in Overview. You can consult the color code in the .html output file by clicking on the blue symbol.

**1.- List:**

In this format, only positions containing drug resistance mutations are shown. The positions are described under the 'Position' column. In the 'Residues' column, all residues found in the analyzed sequences are displayed, with their percentage color-coded. The DRM or SDRM will be indicated with a red asterisk (*). At the end of each row, in the 'Total Positions' column, the total number of valid sequences for that position present in the analyzed file is shown.

At the top, the analysis title, the input file name, and the classification of DRMs/SDRMs will be displayed.

Example of List output format for the analysis of acquired drug resistance mutations:

**List Acquired Resistance Mutations DRM-PI HIV-1**

**PR_procesado_traducido_01_AE.fasta**

**MAJOR DRM-PI**

| Position | Residues | Total Positions |
|---|---|---|
| D30 | D(99.985%) K(0.004%) N(0.004%*) G(0.007%) | 26741 |
| V32 | V(99.929%) A(0.015%) I(0.026%*) L(0.023%) E(0.008%) | 26584 |
| M46 | M(98.733%) I(0.661%*) L(0.531%*) V(0.064%) R(0.011%) | 26764 |
| I47 | I(99.903%) K(0.015%) V(0.048%*) A(0.015%*) M(0.007%) R(0.007%) F(0.004%) | 26814 |
| G48 | G(99.828%) S(0.007%*) R(0.037%) V(0.101%*) M(0.015%*) I(0.004%) Q(0.004%*) E(0.004%) | 26707 |

Example of List output format for the analysis of transmitted drug resistance mutations:

**List Transmitted Resistance Mutations SDRM-PI OMS**

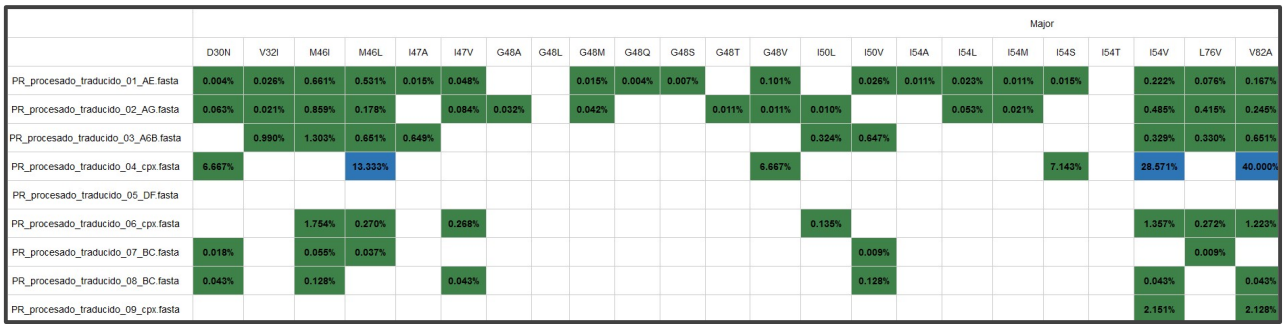
**PR_procesado_traducido_01_AE.fasta**

| Position | Residues | Total Positions |
|---|---|---|
| L23 | L(99.935%) I(0.019%*) F(0.015%) V(0.019%) Q(0.004%) P(0.004%) S(0.004%) | 26251 |
| L24 | L(99.940%) I(0.015%*) V(0.019%) S(0.019%) F(0.007%) | 26714 |
| D30 | D(99.985%) K(0.004%) N(0.004%*) G(0.007%) | 26741 |
| V32 | V(99.929%) A(0.015%) I(0.026%*) L(0.023%) E(0.008%) | 26584 |
| M46 | M(98.733%) I(0.661%*) L(0.531%*) V(0.064%) R(0.011%) | 26764 |

## 2.- Table:

In this format, the title of the analysis is shown in the first row. If analyzing DRM, the type of DRM is displayed below according to its classification. Both in DRM and SDRM, the first column shows the names of the input files used to generate the table. The rest of the columns display the detected DRM or SDRM, with each cell colored according to the color code, and its percentage of occurrence in each input file.

Example of Table output format for the analysis of acquired resistance mutations:



| | Major | | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | D30N | V32I | M46I | M46L | I47A | I47V | G48A | G48L | G48M | G48Q | G48S | G48T | G48V | I50L | I50V | I54A | I54L | I54M | I54S | I54T | I54V | L76V | V82A |
| PR_procesado_traducido_01_AE.fasta | 0.004% | 0.026% | 0.661% | 0.531% | 0.015% | 0.048% | | | 0.015% | 0.004% | 0.007% | | 0.101% | | 0.026% | 0.011% | 0.023% | 0.011% | 0.015% | | 0.222% | 0.076% | 0.167% |
| PR_procesado_traducido_02_AG.fasta | 0.063% | 0.021% | 0.859% | 0.178% | | 0.084% | 0.032% | | 0.042% | | | 0.011% | 0.011% | 0.010% | | | 0.053% | 0.021% | | | 0.485% | 0.415% | 0.245% |
| PR_procesado_traducido_03_A6B.fasta | | 0.990% | 1.303% | 0.651% | 0.649% | | | | | | | | | 0.324% | 0.647% | | | | | | 0.329% | 0.330% | 0.651% |
| PR_procesado_traducido_04_cpx.fasta | 6.667% | | | 13.333% | | | | | | | | | 6.667% | | | | | | 7.143% | | 28.571% | | 40.000% |
| PR_procesado_traducido_05_DF.fasta | | | | | | | | | | | | | | | | | | | | | | | |
| PR_procesado_traducido_06_cpx.fasta | | | 1.754% | 0.270% | | 0.268% | | | | | | | | 0.135% | | | | | | | 1.357% | 0.272% | 1.223% |
| PR_procesado_traducido_07_BC.fasta | 0.018% | | 0.055% | 0.037% | | | | | | | | | | | 0.009% | | | | | | | 0.009% | |
| PR_procesado_traducido_08_BC.fasta | 0.043% | | 0.128% | | | 0.043% | | | | | | | | | 0.128% | | | | | | 0.043% | | 0.043% |
| PR_procesado_traducido_09_cpx.fasta | | | | | | | | | | | | | | | | | | | | | 2.151% | | 2.128% |

Example of Table output format for the analysis of transmitted resistance mutations:

| | Table Transmitted Resistance Mutations SDRM-PI WHO | | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | L23I | L24I | D30N | V32I | M46I | M46L | I47V | I47A | G48V | G48M | I50V | I50L | F53L | F53Y | I54V | I54L | I54M | I54A | I54T | I54S | G73S | G73T | G7 |
| PR_procesado_traducido_01_AE.fasta | 0.019% | 0.015% | 0.004% | 0.026% | 0.661% | 0.531% | 0.048% | 0.015% | 0.101% | 0.015% | 0.026% | | 0.119% | 0.022% | 0.222% | 0.023% | 0.011% | 0.011% | | 0.015% | 0.037% | 0.004% | |
| PR_procesado_traducido_02_AG.fasta | 0.021% | 0.021% | 0.063% | 0.021% | 0.859% | 0.178% | 0.084% | | 0.011% | 0.042% | | 0.010% | 0.115% | 0.115% | 0.485% | 0.053% | 0.021% | | | | 0.159% | | |
| PR_procesado_traducido_03_A6B.fasta | | | | 0.990% | 1.303% | 0.651% | | 0.649% | | | 0.647% | 0.324% | 0.326% | | 0.329% | | | | | | 0.324% | | |
| PR_procesado_traducido_04_cpx.fasta | | 6.667% | 6.667% | | | 13.333% | | | 6.667% | | | | 13.333% | | 28.571% | | | | | 7.143% | | | |
| PR_procesado_traducido_05_DF.fasta | | | | | | | | | | | | | | | | | | | | | | | |
| PR_procesado_traducido_06_cpx.fasta | 0.137% | | | | 1.754% | 0.270% | 0.268% | | | | | 0.135% | 0.269% | 0.269% | 1.357% | | | | | | | | |
| PR_procesado_traducido_07_BC.fasta | | | 0.018% | | 0.055% | 0.037% | | | | | 0.009% | | | | | | | | | | 0.046% | | |
| PR_procesado_traducido_08_BC.fasta | | 0.043% | 0.043% | | 0.128% | | 0.043% | | | | 0.128% | | | | 0.043% | | | | | | | | |
| PR_procesado_traducido_09_cpx.fasta | | | | | | | | | | | | | 1.064% | | 2.151% | | | | | | | | |

### 3.- Summary Table:

In this table, the title of the analysis is shown in the first row. Below, in the first column labeled ‘File,’ the names of the input files used to generate the analysis are listed. In the following columns, the type of SDRM or DRM is displayed according to their classification, with the corresponding residues found for each DRM or SDRM colored based on their percentage following the color code described earlier.

Example of Summary Table output format for the analysis of acquired resistance mutations:

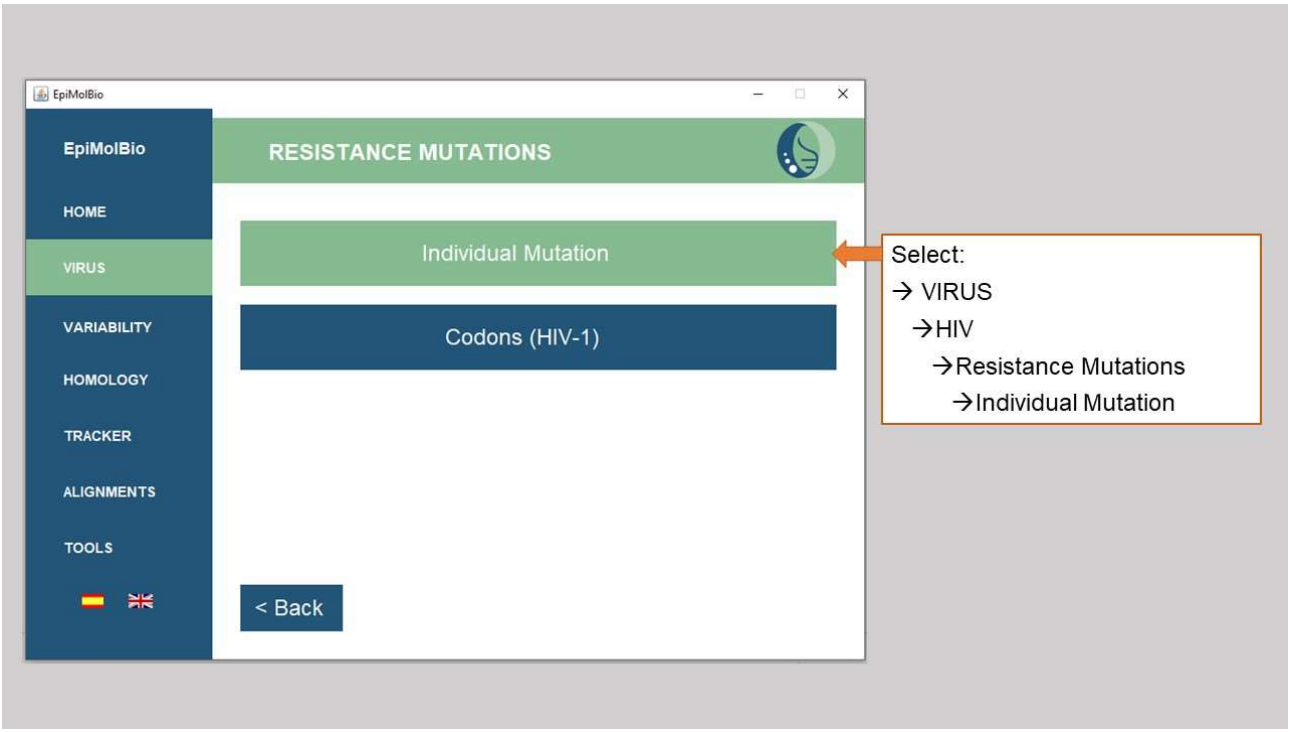
| Summary Table Acquired Resistance Mutations DRM-PI HIV-1 | | | |
|---|---|---|---|
| File | DRM PI | | |
| | Major | Accessory | Other |
| PR_procesado_traducido_01_AE.fasta | D30N, V32I, M46IL, I47AV, G48MQSV, I50V, I54ALMSV, L76V, V82AFLMST, I84V, N88GS, L90M | L10F, K20T, L23I, L24FI, L33F, K43T, M46V, F53LY, Q58E, G73DSTV, T74P, N83D, N88D, L89TV | L10IRVY, V11IL, K20IMRV, L33IV, A71ITV, T74S, V82I, I85V, L89IM |
| PR_procesado_traducido_02_AG.fasta | D30N, V32I, M46IL, I47V, G48AMTV, I50L, I54LMV, L76V, V82ACFLST, I84ACV, N88ST, L90M | L10F, K20T, L23I, L24I, L33F, K43T, M46V, F53LY, Q58E, G73ADSV, T74P, N83D, N88D, L89TV | L10IRVY, V11IL, K20IMRV, L33IV, A71ITV, T74S, V82I, I85V, L89IM |

Example of Summary Table output format for the analysis of transmitted resistance mutations:

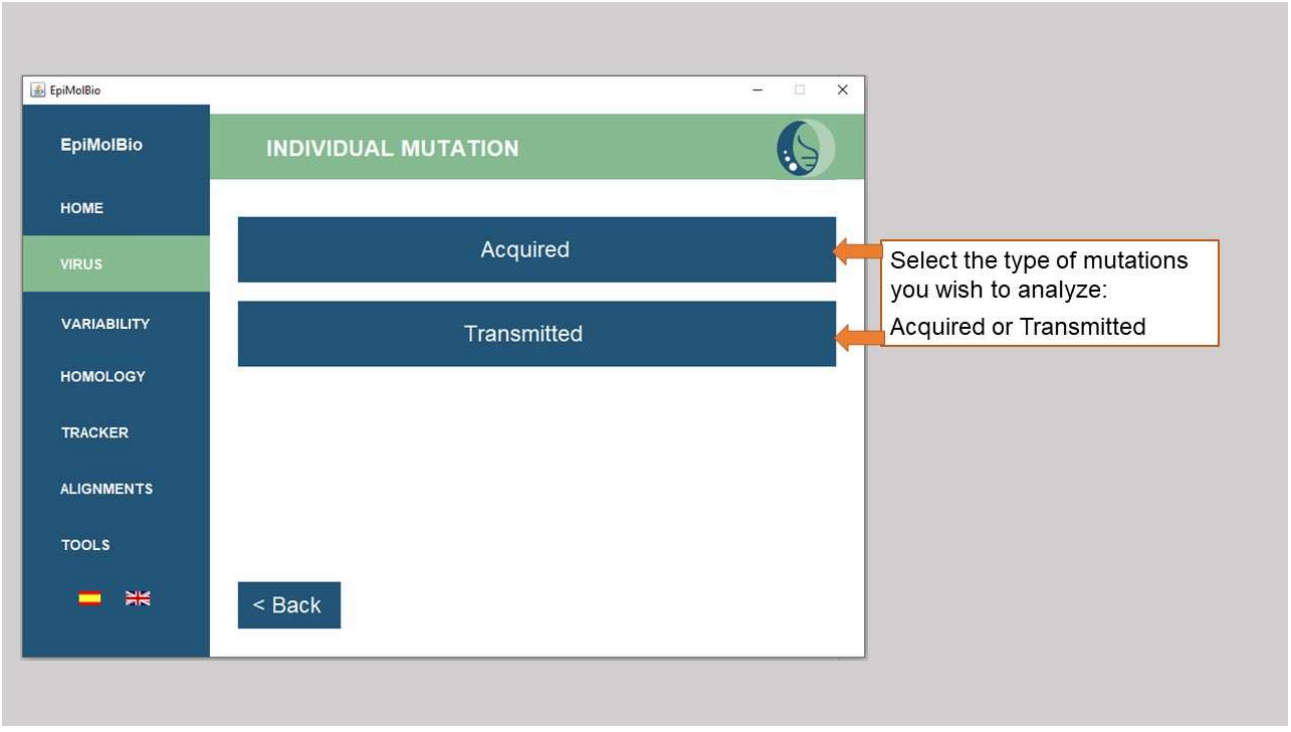
| Summary Table Transmited Resistance Mutations SDRM-PI WHO | |
|---|---|
| File | SDRM PI WHO |
| PR_procesado_traducido_01_AE.fasta | L23I, L24I, D30N, V32I, M46IL, I47VA, G48VM, I50V, F53LY, I54VLMAS, G73ST, L76V, V82ATFSML, N83D, I84V, I85V, N88DS, L90M |
| PR_procesado_traducido_02_AG.fasta | L23I, L24I, D30N, V32I, M46IL, I47V, G48VM, I50L, F53LY, I54VLM, G73SA, L76V, V82ATFSCL, N83D, I84VAC, I85V, N88DS, L90M |
| PR_procesado_traducido_03_A6B.fasta | V32I, M46IL, I47A, I50VL, F53L, I54V, G73S, L76V, V82AT, I84V, L90M |

# Step-by-step:

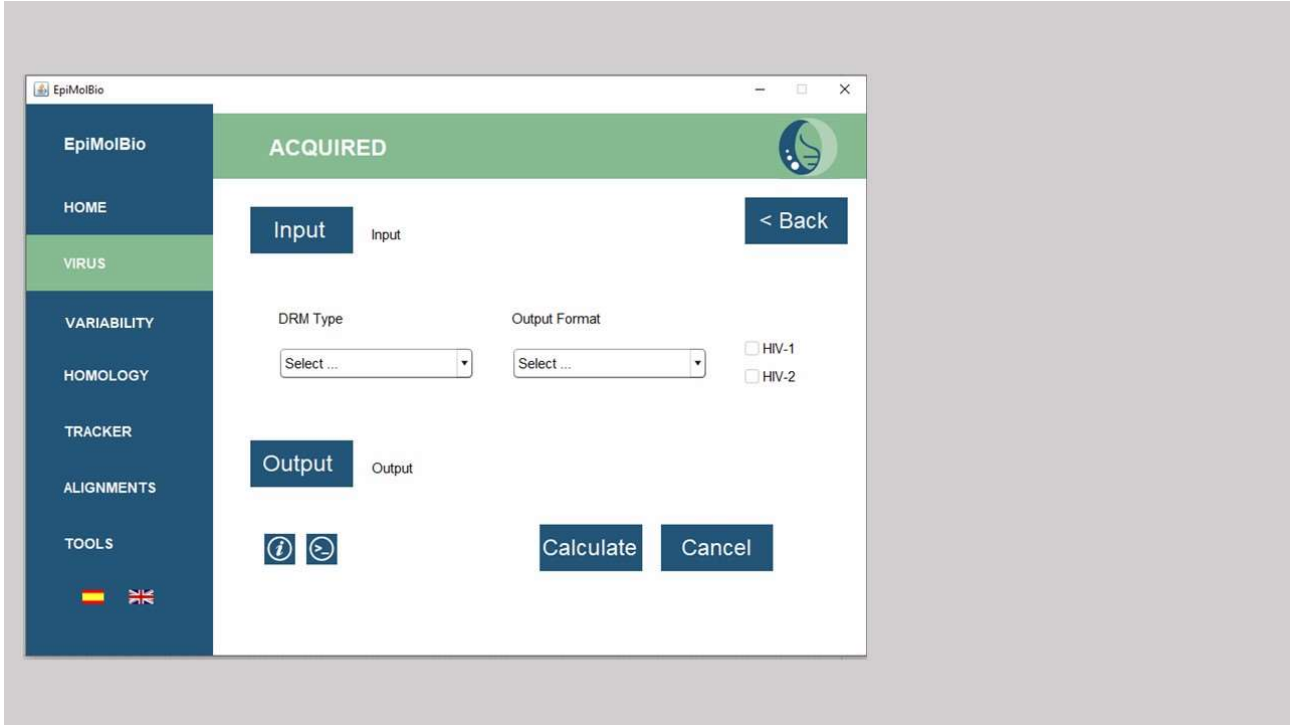
## 1)

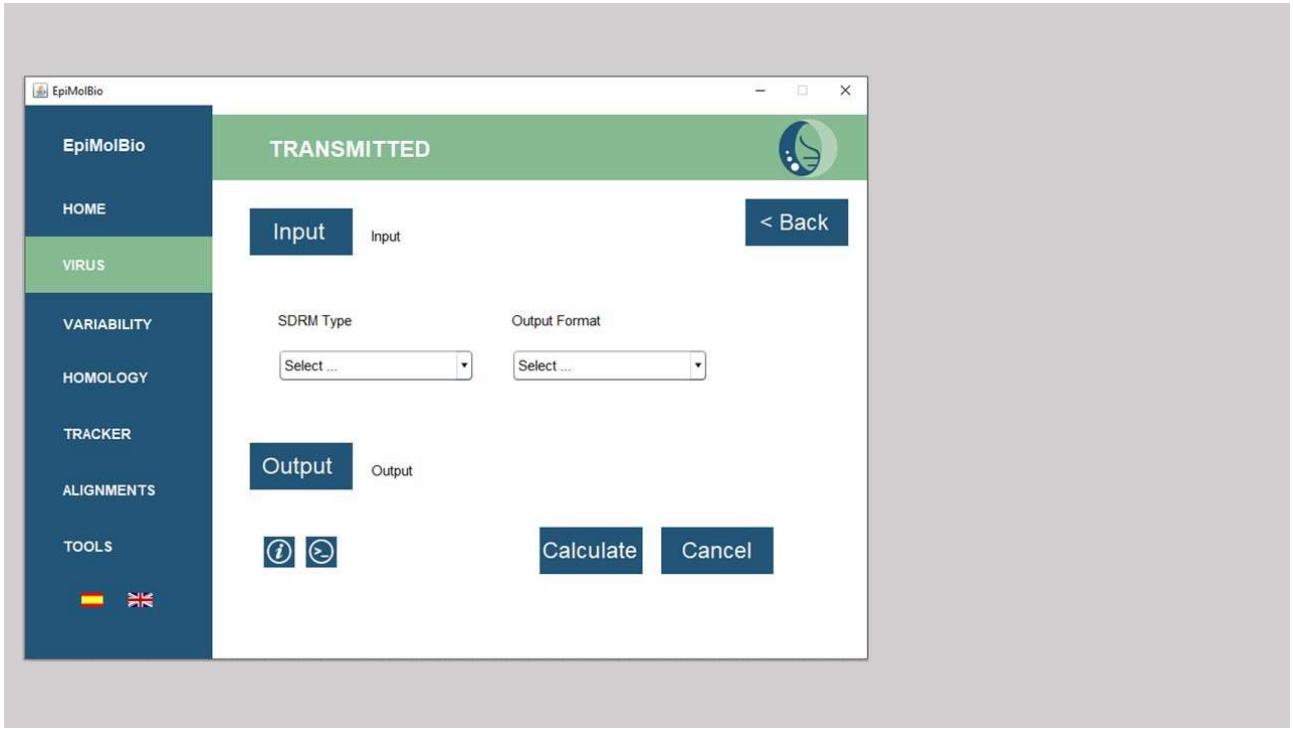



## 2)



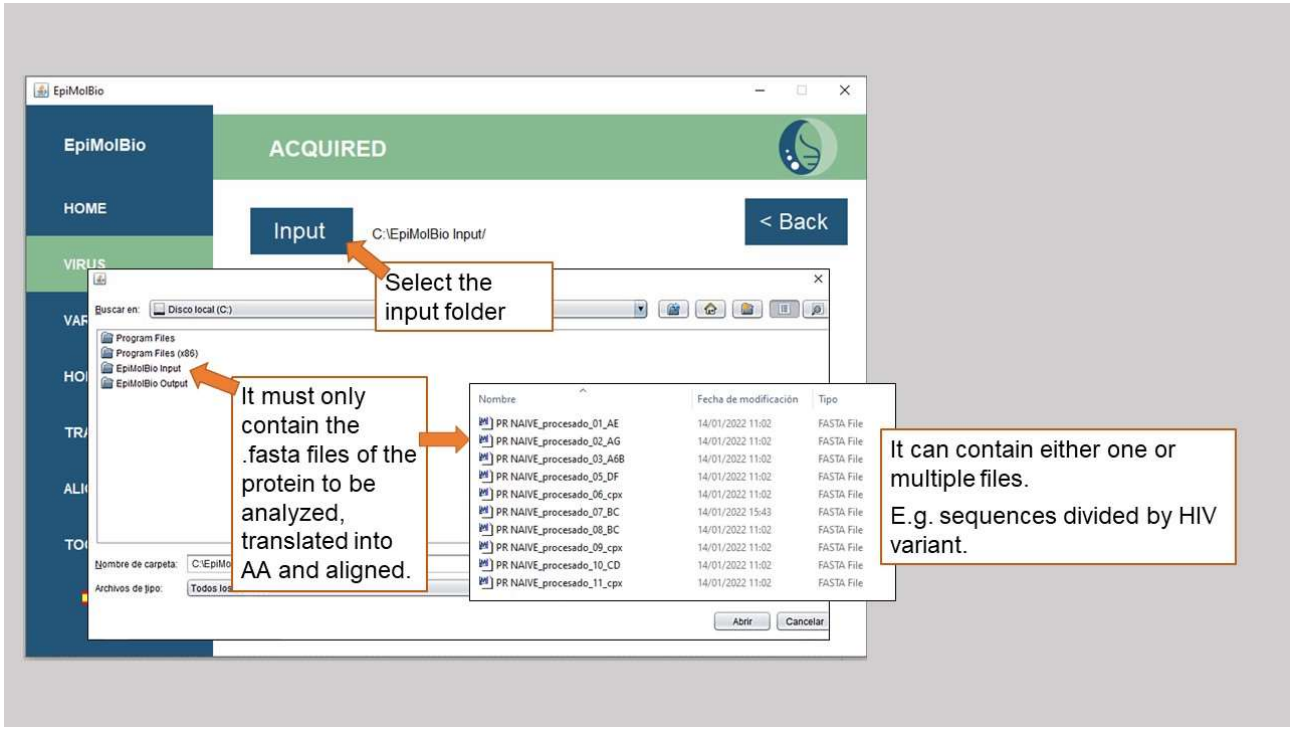
**3)**



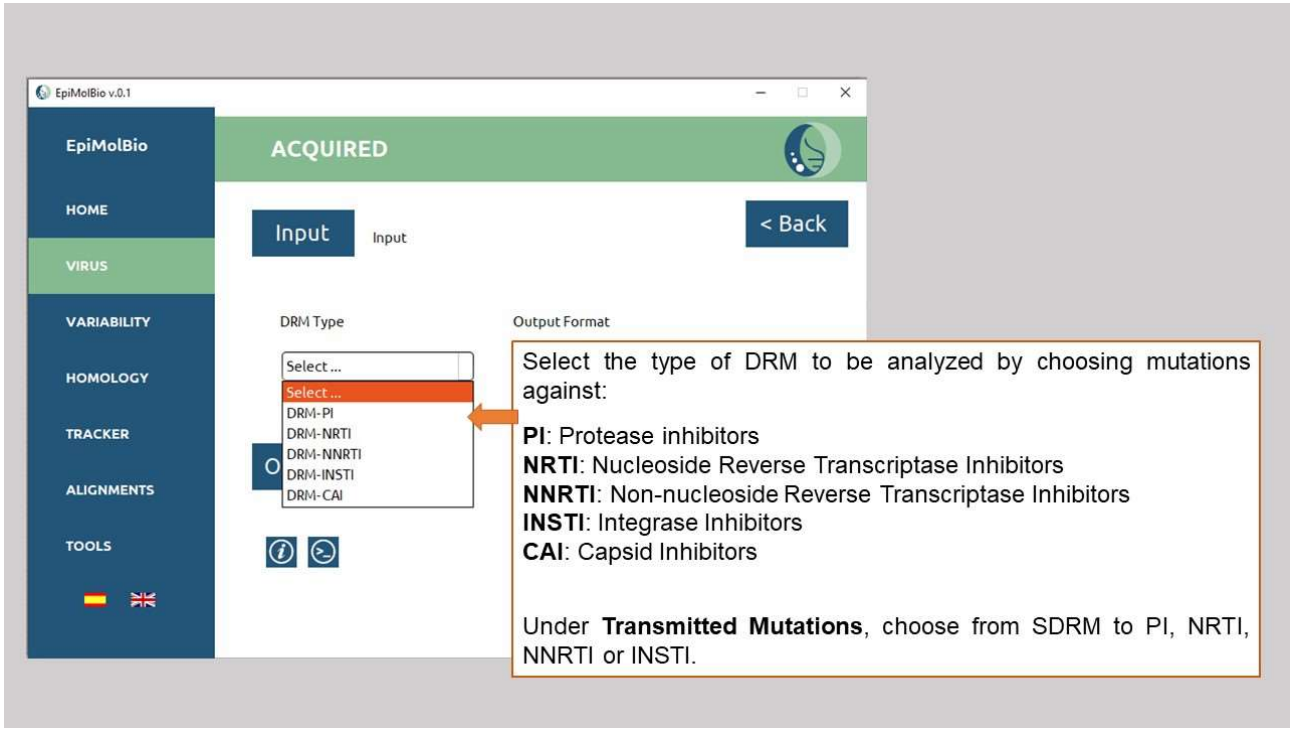




The following steps are identical for both types of mutations, except for step 7, which is exclusive to the detection of acquired DRM.

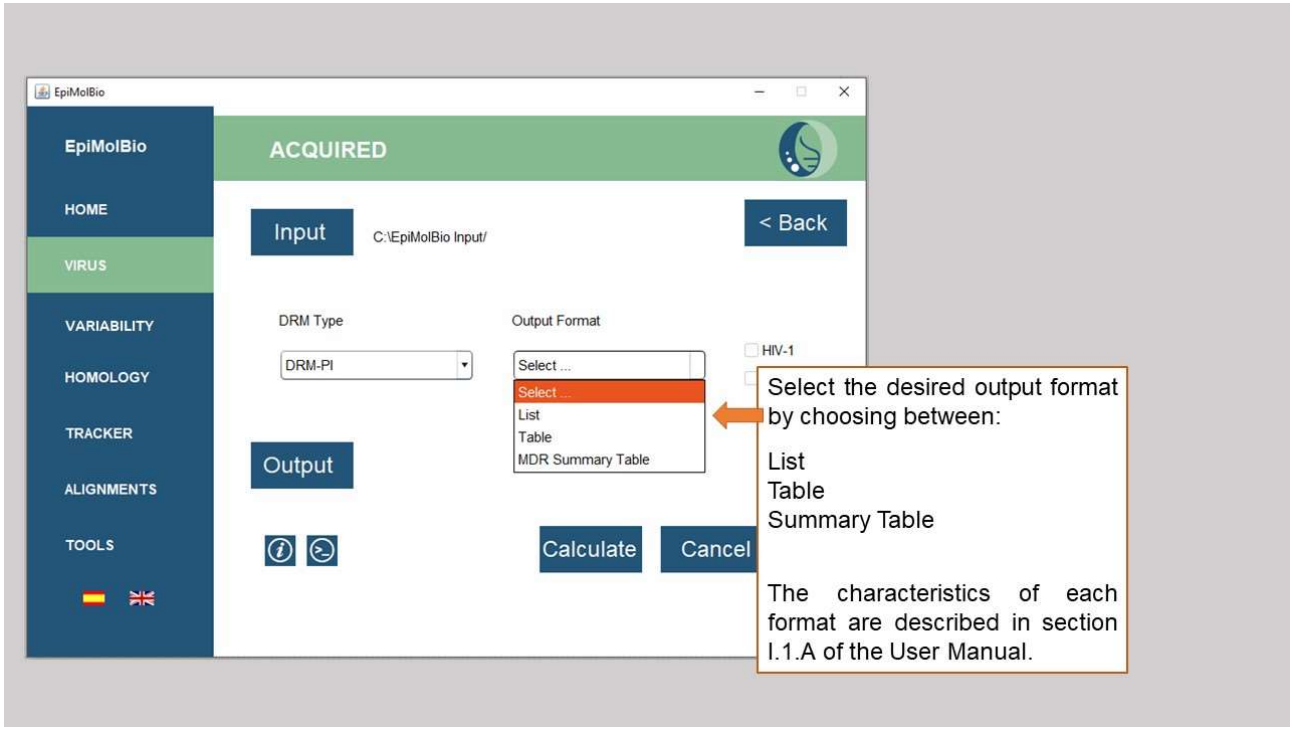
## 4)


EpiMolBio
ACQUIRED
HOME
Input
C:\EpiMolBio Input/
< Back
Select the input folder
It must only contain the .fasta files of the protein to be analyzed, translated into AA and aligned.
It can contain either one or multiple files.
E.g. sequences divided by HIV variant.

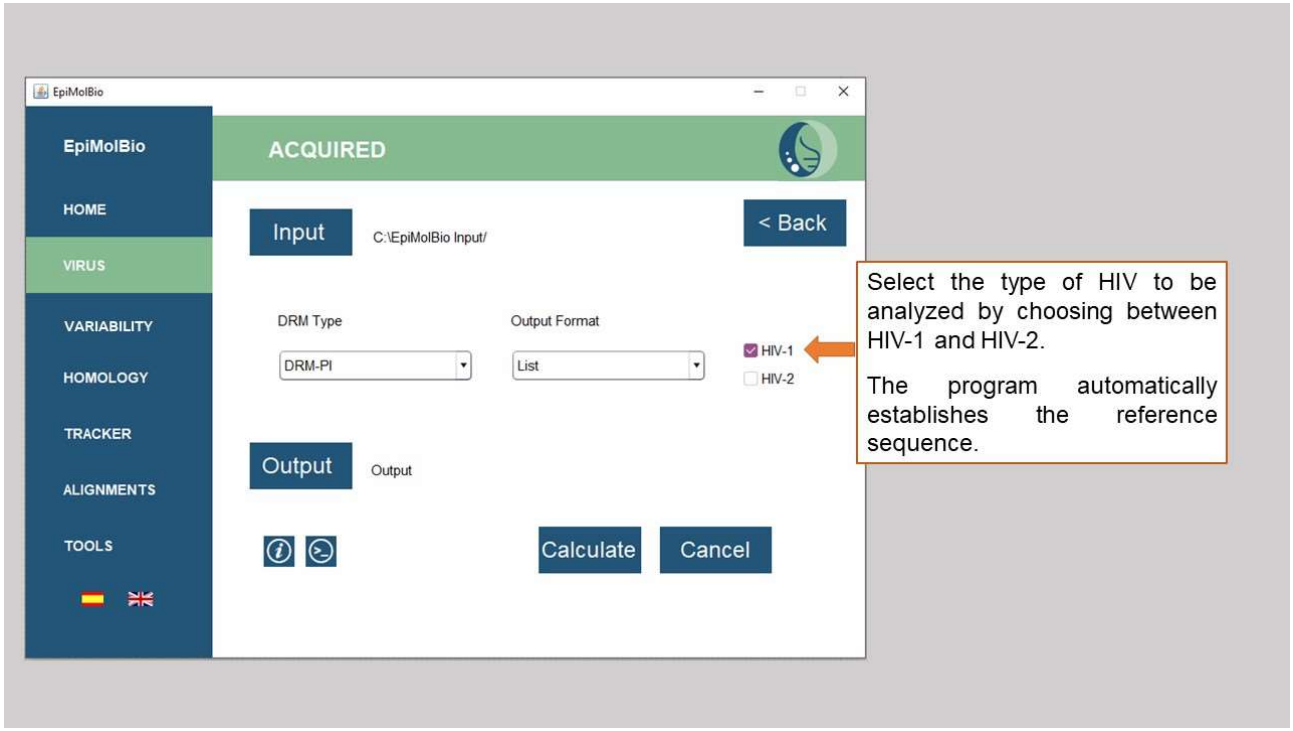
## 5)


EpiMolBio v.0.1
EpiMolBio
ACQUIRED
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
Input
Input
< Back
DRM Type
Output Format
Select ...
DRM-PI
DRM-NRTI
DRM-NNRTI
DRM-INSTI
DRM-CAI
Select the type of DRM to be analyzed by choosing mutations against:
PI: Protease inhibitors
NRTI: Nucleoside Reverse Transcriptase Inhibitors
NNRTI: Non-nucleoside Reverse Transcriptase Inhibitors
INSTI: Integrase Inhibitors
CAI: Capsid Inhibitors
Under Transmitted Mutations, choose from SDRM to PI, NRTI, NNRTI or INSTI.

## 6)

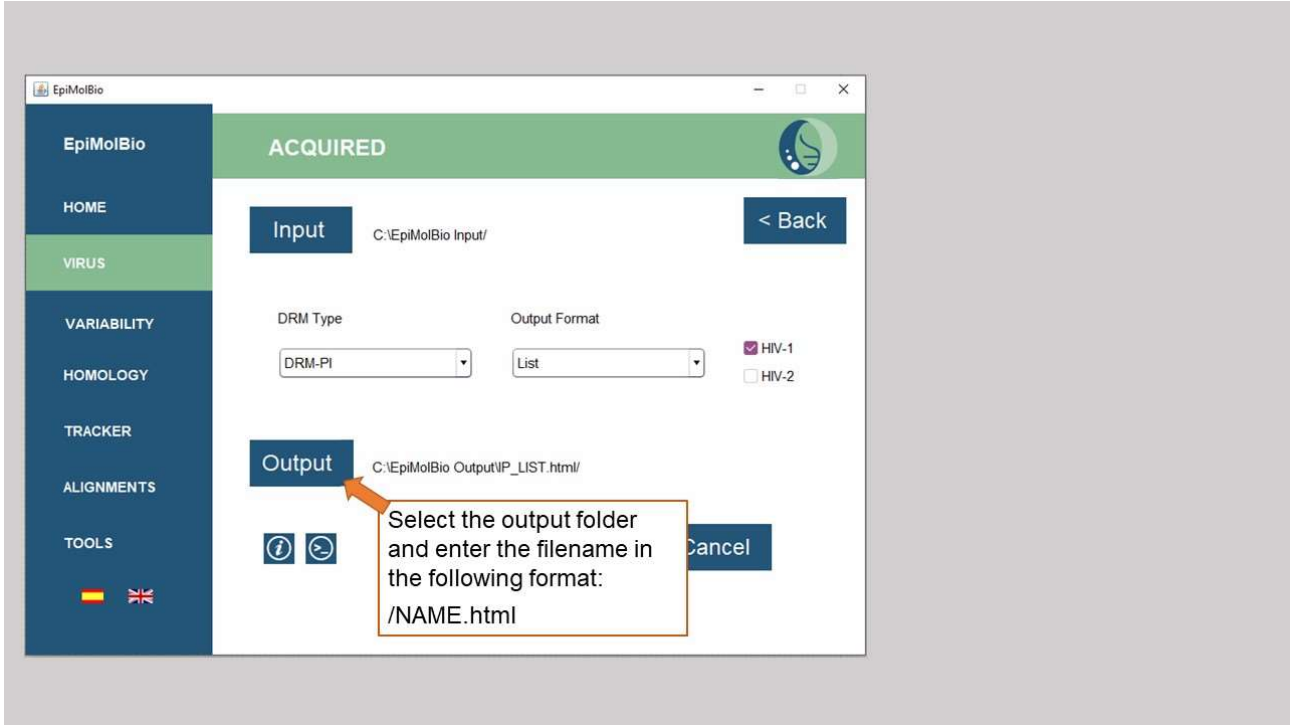



## 7)

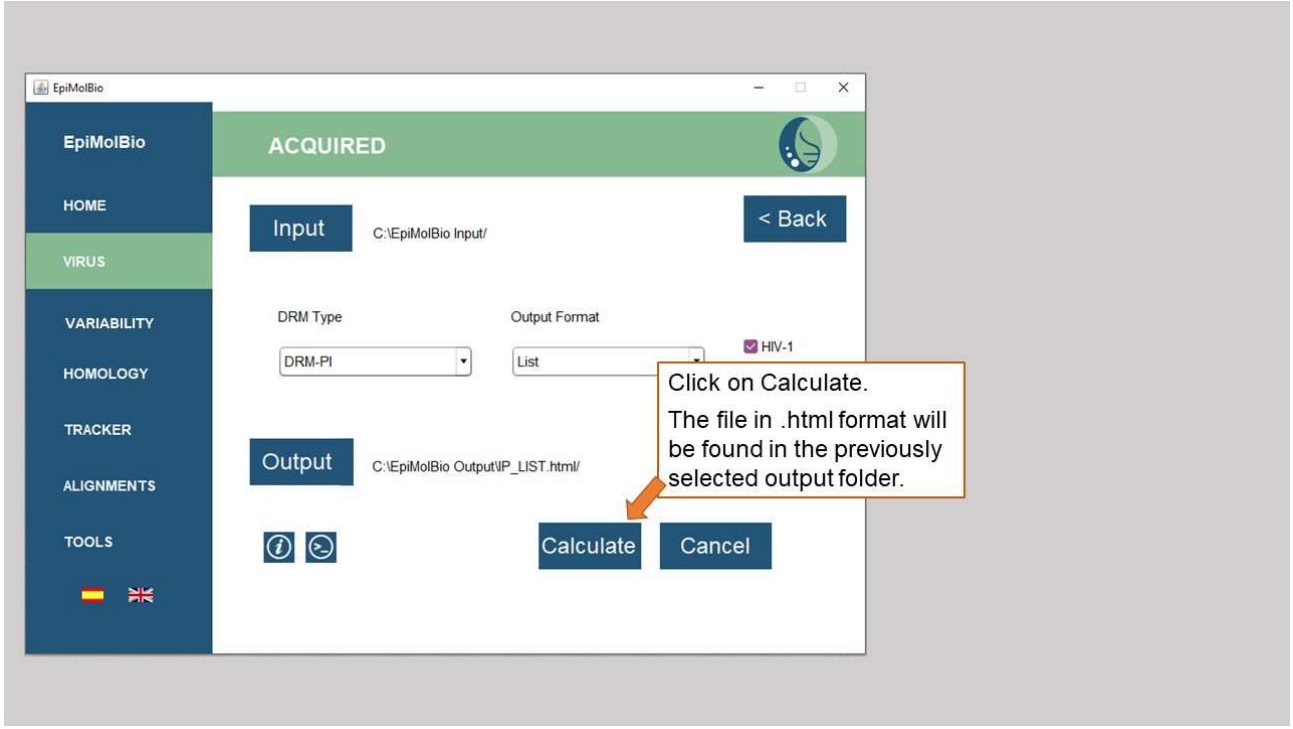



This step is omitted in the analysis of transmitted resistances since the default reference sequence used is that of HIV-1.

## 8)


EpiMolBio
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
ACQUIRED
Input
C:\EpiMolBio Input/
< Back
DRM Type
Output Format
DRM-PI
List
HIV-1
HIV-2
Output
C:\EpiMolBio Output\IP_LIST.html/
Select the output folder and enter the filename in the following format:
/NAME.html
Cancel

## 9)


EpiMolBio
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
ACQUIRED
Input
C:\EpiMolBio Input/
< Back
DRM Type
Output Format
DRM-PI
List
HIV-1
Click on Calculate.
The file in .html format will be found in the previously selected output folder.
Output
C:\EpiMolBio Output\IP_LIST.html/
Calculate
Cancel

**Codons:**

In the **codons** analysis, the aim is to study the codons that, when translated, generate DRM in HIV-1, as defined in the Stanford HIV Drug Resistance Database v9.6. The analysis involves obtaining the frequency of occurrence for each codon and identifying those that generate DRM. The mutations detected in the proteins of the *pol* gene of HIV-1, including Protease, Reverse Transcriptase, or Integrase, and of the Capsid are analyzed. Gaps (-) and 'N' characters are excluded from the analysis. The reference sequence used for HIV-1 by EpiMolBio is HXB2 (NCBI K03455.1).

To conduct the analysis, the **input file** should be a **folder** containing exclusively the aligned sequences of the Pol protein or the Capsid to be analyzed in nucleotide format (.fasta). This folder can contain a single file or multiple .fasta files if the same analysis needs to be performed on different sequence groups (e.g., files divided by HIV variants, country of origin, year, etc.).

In the '**MDR Type**' field, you will need to select the **type of mutation** you wish to study, choosing from mutations related to:

**PI**: Protease Inhibitors (input: Protease)

**NRTI**: Nucleoside Reverse Transcriptase Inhibitors (input: Reverse Transcriptase)

**NNRTI**: Non-Nucleoside Reverse Transcriptase Inhibitors (input: Reverse Transcriptase)

**INSTI**: Integrase Inhibitors (input: Integrase)

**CAI**: Capsid Inhibitors (input: Capsid)

You can perform the study of the entire sequence or search for specific mutations in one or several codons by selecting '**Select Mutation**' and adding one or more valid DRM in the '**Mutations**' field. If you enter more than one mutation, they must be separated by commas without spaces and in amino acids format (e.g., V32I,I50L). If no specific mutation is selected, all positions containing resistance mutations will be shown.

In both cases, the detected **mutations** in the input sequences will be obtained, **classified** according to the Stanford v9.4 classification, and their **percentage** relative to the total analyzed sequences will be calculated according to the **color code** described in 'Overview,' which can be checked in the output .html file by clicking on the blue symbol.

The **output file** will be an .html file. You will need to select the output folder where you want the generated files to appear and name the files by writing .html at the end, except in the case of selecting a mutation, in which case the file will be automatically named based on the DRM provided in the '**Mutations**' field. For each introduced DRM, a separate .html file will be generated.

In the output file, at the top, you will find the title of the analysis, followed by the name of the input file, and the type of DRM analyzed (the latter won't appear if a specific mutation is selected). In the 'Position' column, you'll see the positions containing resistance mutations or the positions of the mutations entered in the 'Mutations' field if a specific mutation is selected. In the 'Residues' column, all the residues found for that position will be listed along with the codon that encodes it, and the percentage of occurrence of that codon will be colored according to the color code. MDRs are indicated with a red asterisk (*). Non-coding codons will be denoted with a question mark (?). The 'Total Codons' column describes the

total number of valid sequences for that position present in the analyzed file. If no mutation is detected in any row, the output file will display the information present in that position, even if it is not mutated.

Example of output format for the analysis of codon resistance mutations without selecting a specific mutation:

**Codon Resistance Mutations DRM-PI**

**PR_procesado_01_AE.fasta**

**MDR-PI Major**

| Position | Residues | Total Codons |
|---|---|---|
| D30N | D[GAT(98.484%)] ?[GAY(0.339%)] D[GAC(1.114%)] ?[RAK(0.007%)] K[AAG(0.004%)] N[AAT*(0.004%)] ?[RAT(0.004%)] ?[GAK(0.004%)] ?[GRT(0.022%)] ?[GAW(0.004%)] ?[GMT(0.004%)] ?[GWT(0.004%)] G[GGT(0.007%)] | 26845 |
| V32I | V[GTA(95.128%)] V[GTG(3.356%)] V[GTC(0.253%)] ?[GTR(0.756%)] ?[GTM(0.082%)] ?[GTW(0.063%)] V[GTT(0.205%)] ?[GTY(0.007%)] A[GCA(0.015%)] I[ATA*(0.026%)] ?[RTA(0.004%)] ?[GTD(0.007%)] L[TTA(0.015%)] ?[KTA(0.011%)] ?[GWA(0.026%)] E[GAA(0.007%)] ?[GYA(0.015%)] L[CTA(0.007%)] ?[STA(0.004%)] ?[KTW(0.004%)] ?[GKA(0.007%)] | 26849 |
| M46I | M[ATG(98.424%)] I[ATA*(0.648%)] L[TTG(0.466%)] V[GTG(0.030%)] ?[WTG(0.071%)] ?[AYG(0.007%)] ?[ATR(0.142%)] ?[TTR(0.011%)] ?[RTG(0.045%)] L[CTG(0.030%)] I[ATT*(0.011%)] L[TTA(0.030%)] ?[AKG(0.011%)] ?[MTG(0.007%)] ?[TYG(0.004%)] L[CTA(0.004%)] R[AGG(0.007%)] R[AGA(0.004%)] ?[AWG(0.004%)] ?[RTA(0.004%)] V[GTA(0.034%)] ?[WWG(0.004%)] ?[RTR(0.004%)] | 26848 |

Example of output format for the analysis of codon resistance mutations selecting M46I mutation:

**Codon Resistance Mutations DRM-PI**

**PR_procesado_01_AE.fasta**

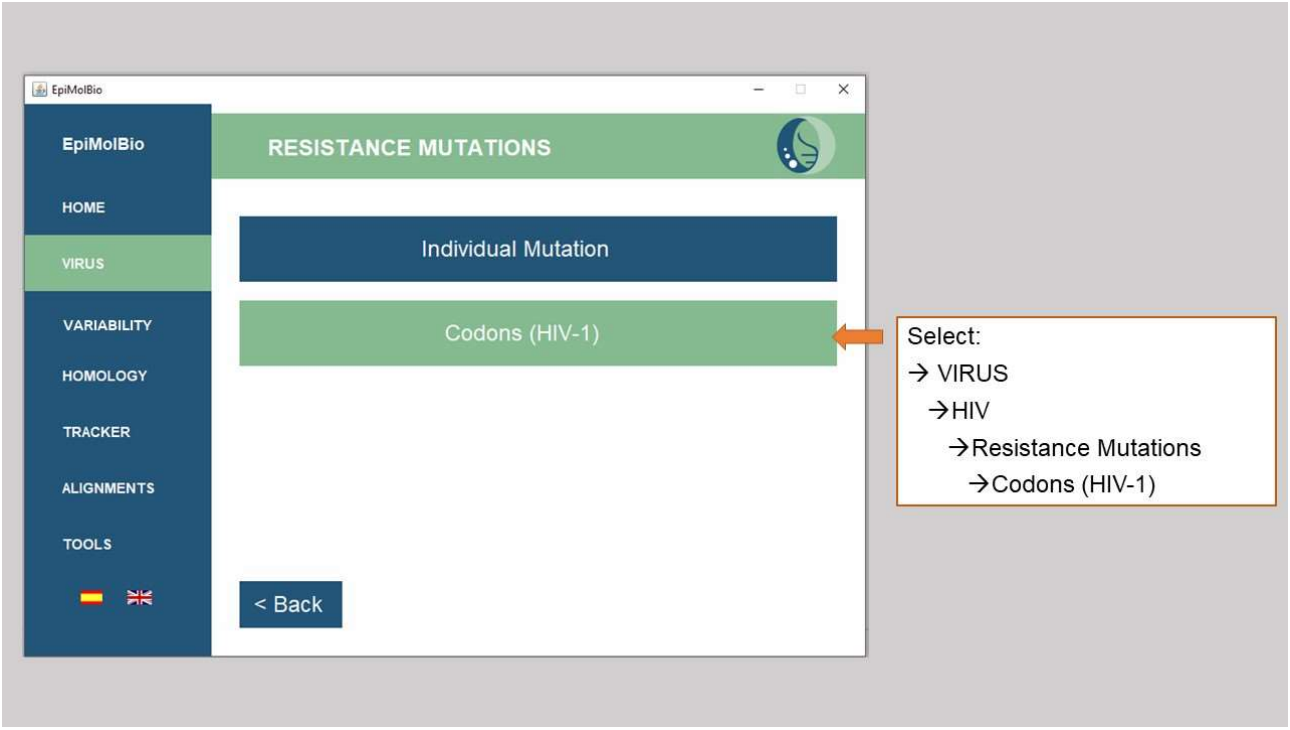
| Position | Residues | Total Codons |
|---|---|---|
| M46I | M[ATG(98.424%)] I[ATA*(0.648%)] L[TTG(0.466%)] V[GTG(0.030%)] ?[WTG(0.071%)] ?[AYG(0.007%)] ?[ATR(0.142%)] ?[TTR(0.011%)] ?[RTG(0.045%)] L[CTG(0.030%)] I[ATT*(0.011%)] L[TTA(0.030%)] ?[AKG(0.011%)] ?[MTG(0.007%)] ?[TYG(0.004%)] L[CTA(0.004%)] R[AGG(0.007%)] R[AGA(0.004%)] ?[AWG(0.004%)] ?[RTA(0.004%)] V[GTA(0.034%)] ?[WWG(0.004%)] ?[RTR(0.004%)] | 26848 |

**PR_procesado_02_AG.fasta**

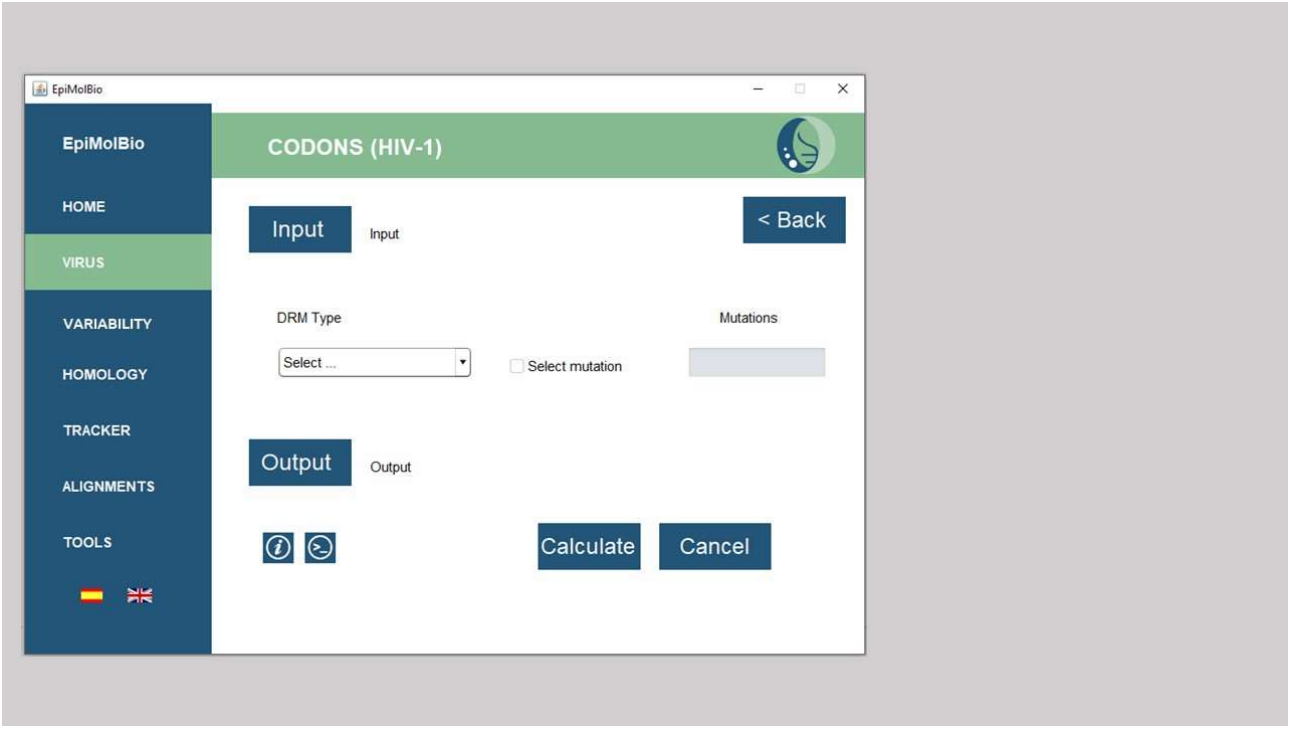
| Position | Residues | Total Codons |
|---|---|---|
| M46I | M[ATG(98.528%)] I[ATA*(0.835%)] L[TTG(0.157%)] V[GTG(0.063%)] ?[ATR(0.198%)] ?[WTG(0.042%)] I[ATT*(0.010%)] ?[TTR(0.021%)] V[GTA(0.021%)] I[ATC*(0.010%)] ?[AKG(0.010%)] ?[MTG(0.010%)] ?[AYG(0.010%)] ?[RTG(0.010%)] ?[RTA(0.010%)] L[TTA(0.010%)] ?[AWG(0.010%)] L[CTG(0.010%)] K[AAG(0.021%)] ?[WTR(0.010%)] | 9577 |

# Step-by-step:

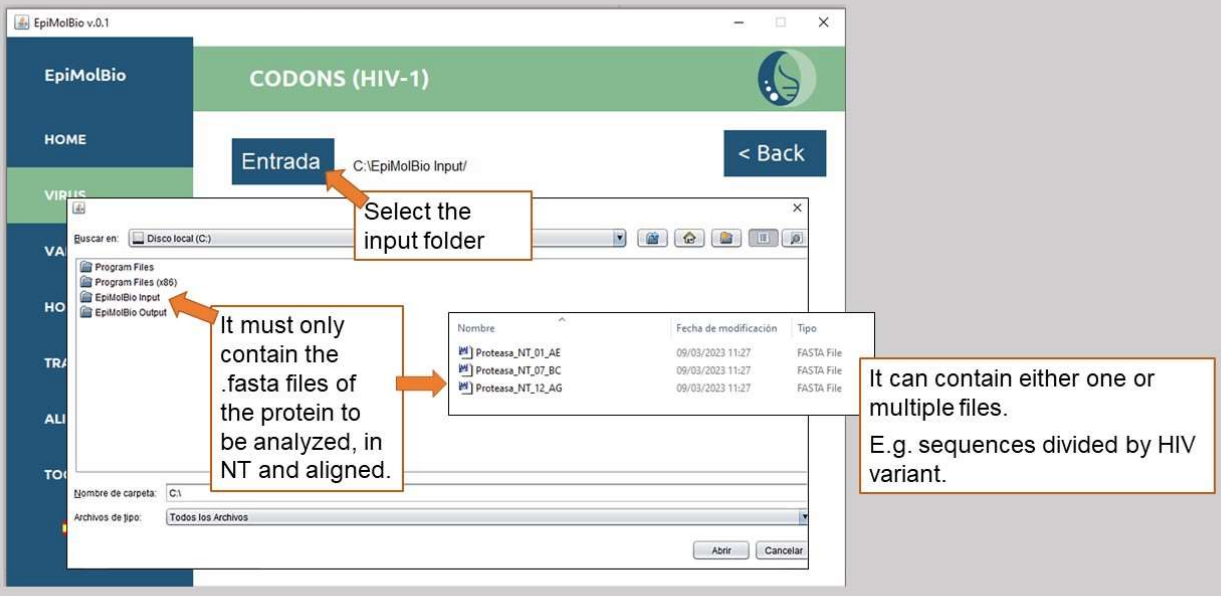
## 1)



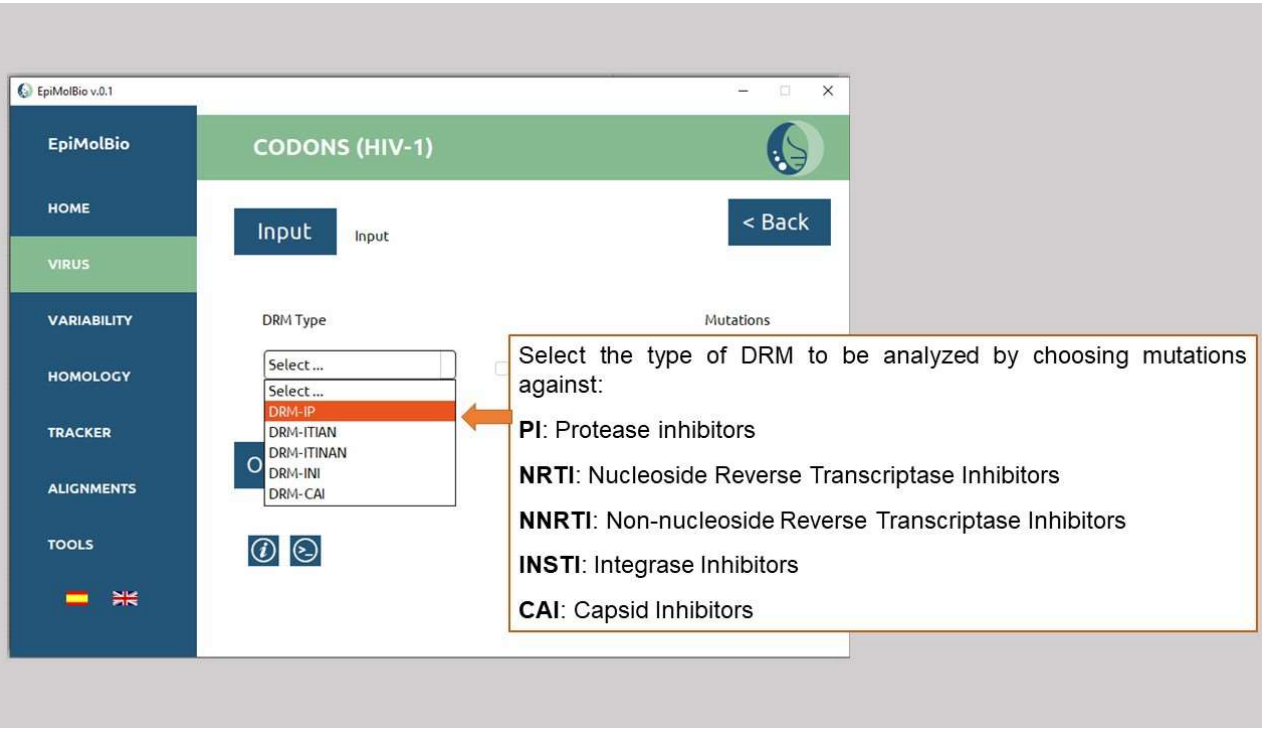

## 2)



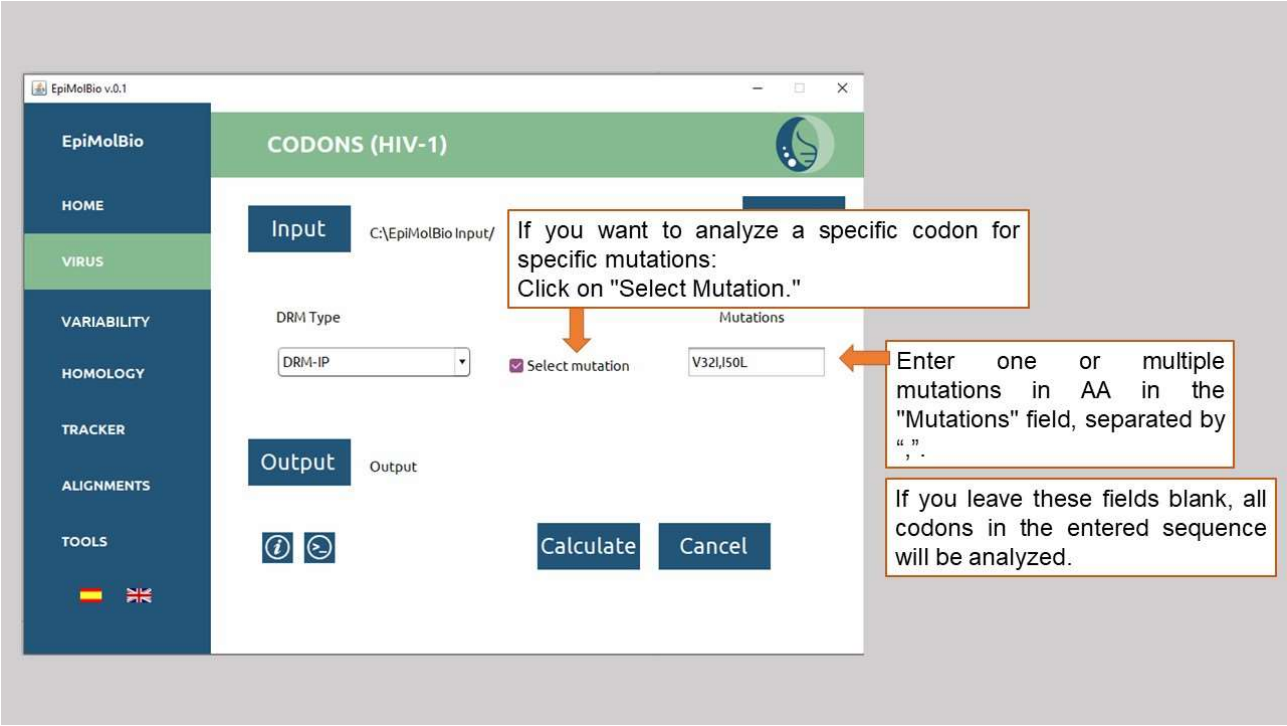
## 3)


EpiMolBio v.0.1
EpiMolBio
HOME
CODONS (HIV-1)
Entrada
C:\EpiMolBio Input/
< Back
Select the input folder
Program Files
Program Files (x86)
EpiMolBio Input
EpiMolBio Output
It must only contain the .fasta files of the protein to be analyzed, in NT and aligned.
Proteasa_NT_01_AE
Proteasa_NT_07_BC
Proteasa_NT_12_AG
It can contain either one or multiple files.
E.g. sequences divided by HIV variant.

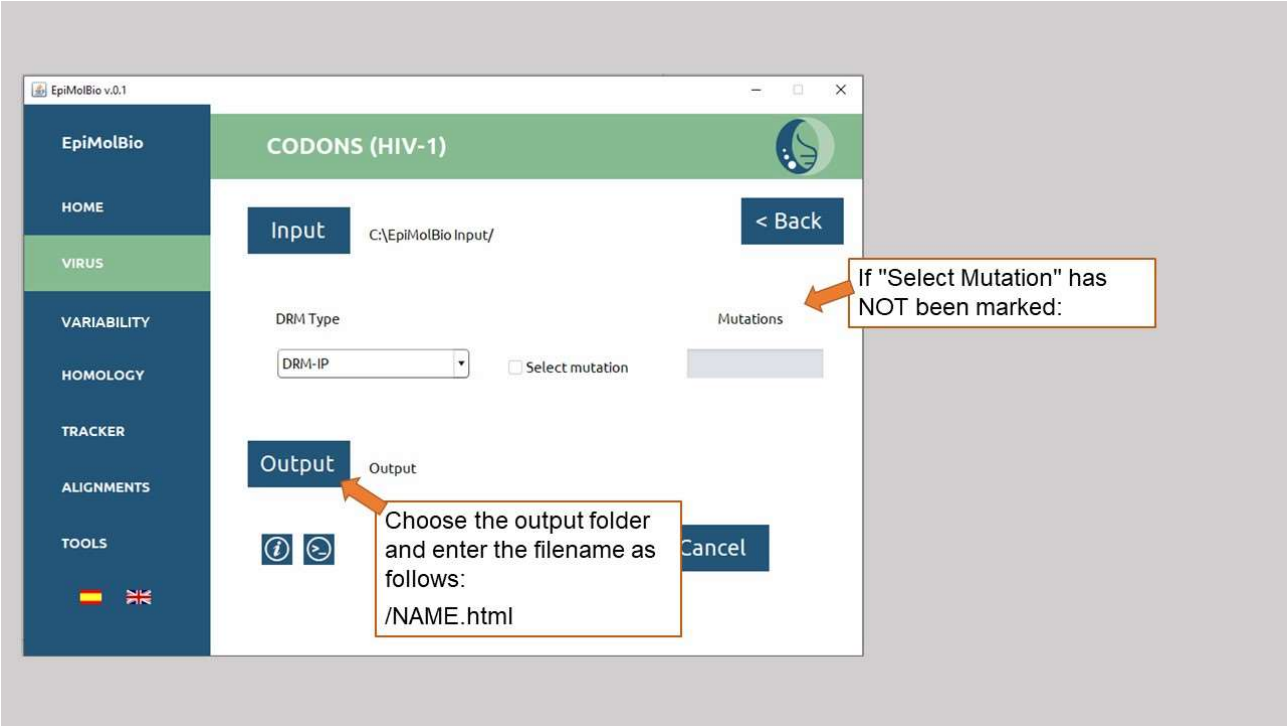
## 4)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
CODONS (HIV-1)
Input
Input
< Back
DRM Type
Mutations
Select ...
DRM-IP
DRM-ITIAN
DRM-ITINAN
DRM-INI
DRM-CAI
Select the type of DRM to be analyzed by choosing mutations against:
PI: Protease inhibitors
NRTI: Nucleoside Reverse Transcriptase Inhibitors
NNRTI: Non-nucleoside Reverse Transcriptase Inhibitors
INSTI: Integrase Inhibitors
CAI: Capsid Inhibitors

## 5)

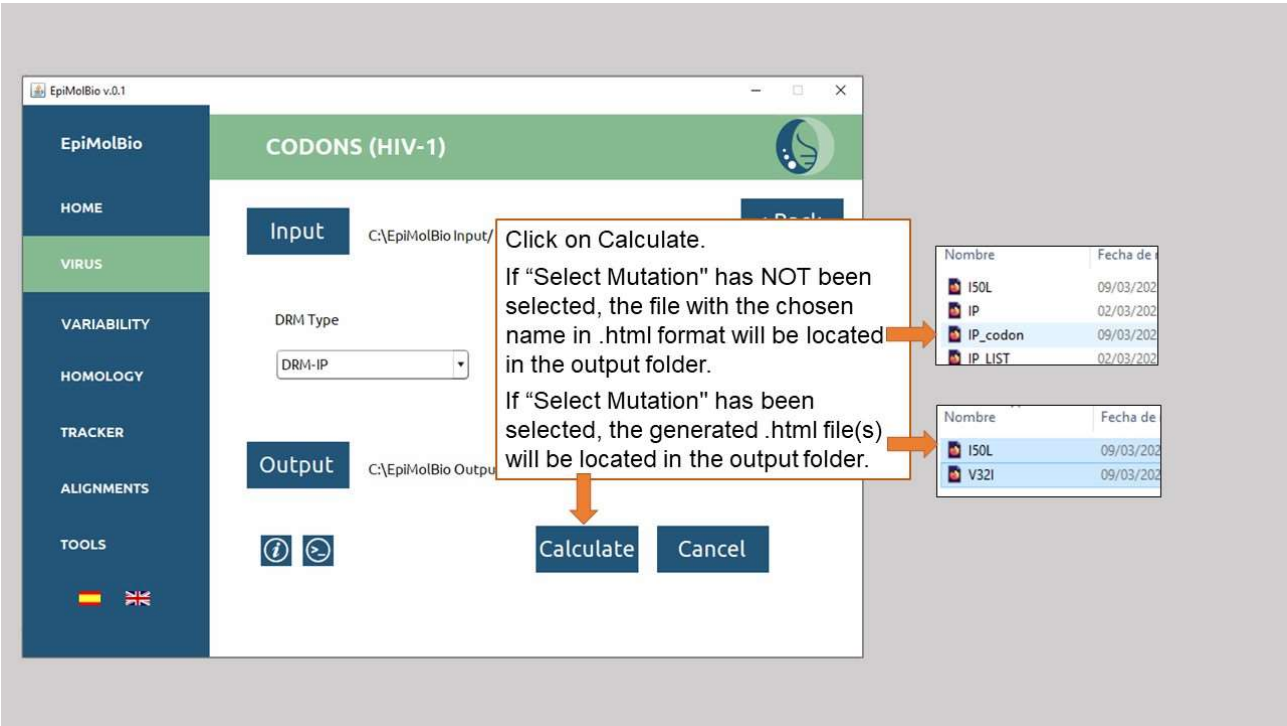

EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
CODONS (HIV-1)
Input
C:\EpiMolBio Input/
If you want to analyze a specific codon for specific mutations:
Click on "Select Mutation."
DRM Type
Mutations
DRM-IP
Select mutation
V32I,I50L
Enter one or multiple mutations in AA in the "Mutations" field, separated by " , ".
Output
Output
If you leave these fields blank, all codons in the entered sequence will be analyzed.
Calculate
Cancel

## 6)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
CODONS (HIV-1)
Input
C:\EpiMolBio Input/
< Back
If "Select Mutation" has NOT been marked:
DRM Type
Mutations
DRM-IP
Select mutation
Output
Output
Choose the output folder and enter the filename as follows:
/NAME.html
Cancel


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
CODONS (HIV-1)
Input
C:\EpiMolBio Input/
< Back
DRM Type
DRM-IP
Select mutation
Mutations
V32I,I50L
If "Select Mutation" has been marked:
Output
Output
Choose the output folder. There's no need to name the files. They will take the name of the analyzed DRM.

## 7)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
CODONS (HIV-1)
Input
C:\EpiMolBio Input/
DRM Type
DRM-IP
Output
Click on Calculate.
If “Select Mutation" has NOT been selected, the file with the chosen name in .html format will be located in the output folder.
If “Select Mutation" has been selected, the generated .html file(s) will be located in the output folder.
Calculate
Cancel
Nombre
I50L
IP
IP_codon
IP_LIST
V32I

# I.1.B) OTHER POL MUTATIONS

This function allows the detection of **any mutation** (not only DRM) of HIV-1 or HIV-2 from Pol protein sequences, obtaining their percentage of occurrence with respect to the reference sequence.

AA changes and mutations in the HIV Pol proteins are analyzed: Protease, Reverse Transcriptase, or Integrase. Both gaps (-) and question marks (?) are excluded from the analysis. EpiMolBio does not detect deletions or insertions.

The **input file** should be a folder containing exclusively aligned sequences of the Pol protein to be analyzed (protease, reverse transcriptase, or integrase) in amino acids and in .fasta format. This folder may contain a single file or multiple .fasta files if we want to analyze the same in different sequence groups (e.g., files divided by HIV variants, country of origin, year, etc.).

The **output file** will be an .html file. You will need to select the output folder where you want the .html files to appear and name the files by writing .html at the end.

In the '**Protein**' field, you will need to select the protein you want to analyze:

**PR**: Protease

**RT**: Reverse Transcriptase

**IN**: Integrase

Next, you will have to choose the **type of HIV** to be analyzed to establish the reference sequence: **HIV-1 or HIV-2**. The reference sequence for HIV-1 used by EpiMolBio is HXB2 (NCBI K03455.1), and for HIV-2, it is ALI (NCBI AF082339).

In the '**Output Format**' field, select the desired output format, choosing from three types of output formats: **list, table, and summary table**.

In the '**List**' output format, there are two different screening percentages: 100% and >75%. This means that in the first option (100%), all mutations found will be shown, and in the second option (>75%), only mutations with an occurrence frequency above 75% will be displayed.

In the other output formats, '**Table**' and '**Summary Table**,' the default screening is set to >75%.

In all formats, the detected **mutations** in the input sequences will be obtained along with their **percentage** relative to the total number of analyzed sequences, following the color code described in the Overview, which can be consulted in the .html output file by clicking on the blue symbol.

**1.- List:**

In this output format, at the top, you will find the title of the analysis and the applied screening, followed by the name of the input file. In the 'Position' column, all positions with their reference amino acid are listed. In the 'Residues' column, all detected residues for that

position are shown, based on the applied screening, along with their occurrence percentage, colored according to the color code. The ‘Total Positions’ column describes the total number of valid sequences for that position present in the analyzed file.

Example of List output format for the analysis of Other Pol Mutations with 100% screening:

**List Other Pol Mutations Protease HIV-1 100%**

**PR_procesado_traducido_01_AE.fasta**

| Position | Residues | Total Positions |
|---|---|---|
| P1 | P(99.896%) S(0.078%) A(0.004%) L(0.007%) T(0.007%) H(0.004%) V(0.004%) | 26838 |
| Q2 | Q(99.782%) E(0.071%) S(0.019%) H(0.056%) D(0.004%) K(0.023%) L(0.015%) P(0.008%) R(0.011%) T(0.004%) *(0.008%) | 26649 |
| V3 | I(99.858%) V(0.078%) N(0.015%) L(0.041%) T(0.007%) | 26831 |
| T4 | T(99.858%) M(0.004%) I(0.048%) N(0.019%) P(0.022%) S(0.034%) F(0.004%) A(0.007%) H(0.004%) | 26816 |
| L5 | L(99.888%) F(0.075%) V(0.015%) S(0.004%) R(0.007%) I(0.007%) T(0.004%) | 26780 |
| W6 | W(99.929%) G(0.030%) R(0.022%) *(0.007%) C(0.011%) | 26836 |

Example of List output format for the analysis of Other Pol Mutations with >75% screening:

**List Other Pol Mutations Protease HIV-1 > 75%**

**PR_procesado_traducido_01_AE.fasta**

| Position | Residues | Total Positions |
|---|---|---|
| V3 | I(99.858%) | 26831 |
| E35 | D(86.051%) | 26160 |
| M36 | I(99.177%) | 26743 |
| S37 | N(92.755%) | 26461 |
| R41 | K(97.554%) | 26572 |
| H69 | K(97.972%) | 26531 |
| L89 | M(96.625%) | 26547 |

**PR_procesado_traducido_02_AG.fasta**

| Position | Residues | Total Positions |
|---|---|---|
| V3 | I(99.529%) | 9557 |
| I13 | V(91.362%) | 9308 |
| K20 | I(94.888%) | 9448 |
| M36 | I(98.475%) | 9511 |
| R41 | K(92.213%) | 9374 |
| H69 | K(96.945%) | 9394 |
| L89 | M(92.325%) | 9407 |

## 2.- Table:

In the Table output format, the default screening is set to >75%, so mutations with a frequency ≤ 75% will not appear in this format.

At the top of the table, you will find the title of the analysis. In the first column, 'File,' the names of the input files used to generate the table are listed. In the following columns, each position is displayed with its reference amino acid and the mutated residue, with the cell colored according to the color code, indicating the occurrence percentage for that position.

Example of Table output format for the analysis of Other Pol Mutations:

| File | P1 | Q2 | V3 | T4 | L5 | W6 | Q7 | R8 | P9 | L10 | V11 | T12 | I13 | K14 | I15 | G16 | G17 | Q18 | L19 | K20 | E21 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| PR_procesado_traducido_01_AE.fasta | | | I | | | | | | | | | | | | | | | | | | |
| PR_procesado_traducido_02_AG.fasta | | | I | | | | | | | | | | V | | | | | | | I | |
| PR_procesado_traducido_03_A6B.fasta | | | I | | | | | | | | | | | | | | | | | | |
| PR_procesado_traducido_04_cpx.fasta | | | I | | | | | | | | | | | | | | | | | | |
| PR_procesado_traducido_05_DF.fasta | | | I | | | | | | | | | | | | | | | | | | |
| PR_procesado_traducido_06_cpx.fasta | | | I | | | | | | | | | | V | | | | | | | I | |
| PR_procesado_traducido_07_BC.fasta | | | I | | | | | | | | | | | | | | | | | | |
| PR_procesado_traducido_08_BC.fasta | | | I | | | | | | | | | S | | | V | | | | I | | |
| PR_procesado_traducido_09_cpx.fasta | | | I | | | | | | | | | | V | | | | | | | | |

## 3.- Summary Table:

In the Summary Table output format, the default screening is set to >75%, so mutations with a frequency ≤ 75% will not appear in this format.

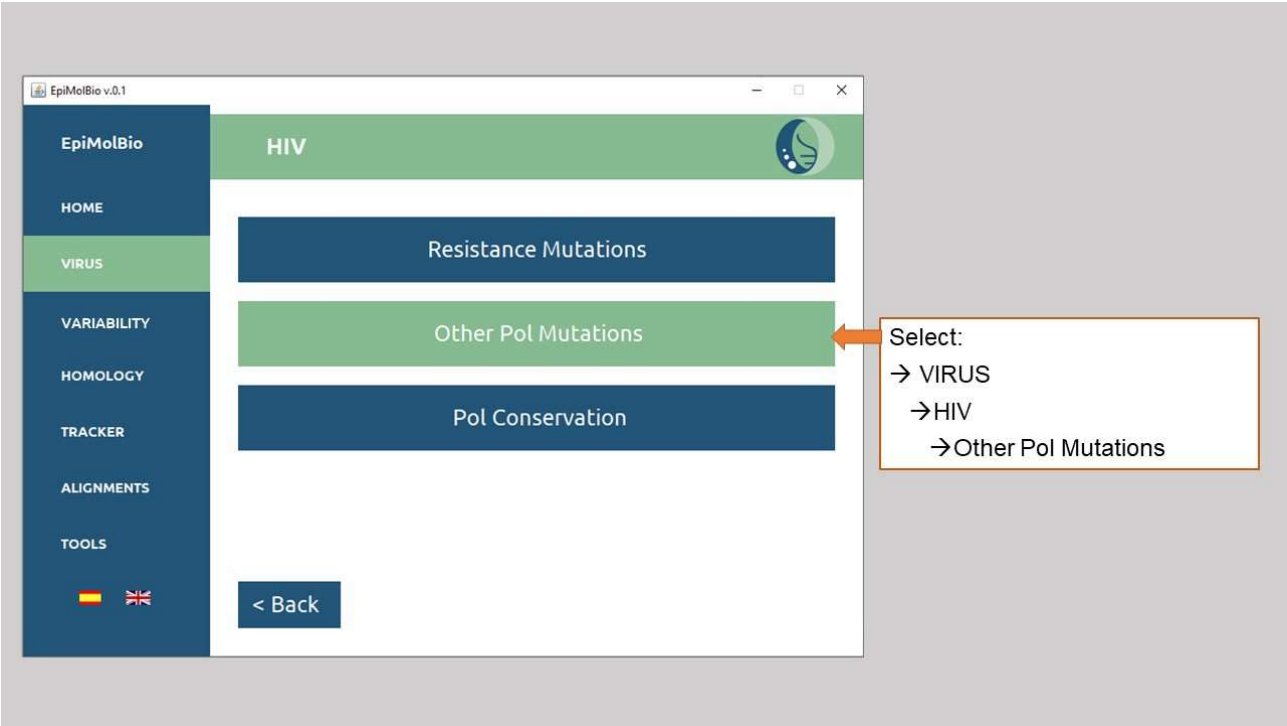
At the top of the table, you will find the title of the analysis. In the first column, 'File,' the names of the input files used to generate the table are listed. In the 'Residues' column, each position is displayed with its reference amino acid and the mutated residue, colored according to the color code, indicating the occurrence percentage for that position. In the 'Total Sequences' column, the total number of sequences from each input file is shown.

Example of Summary Table output format for the analysis of Other Pol Mutations:

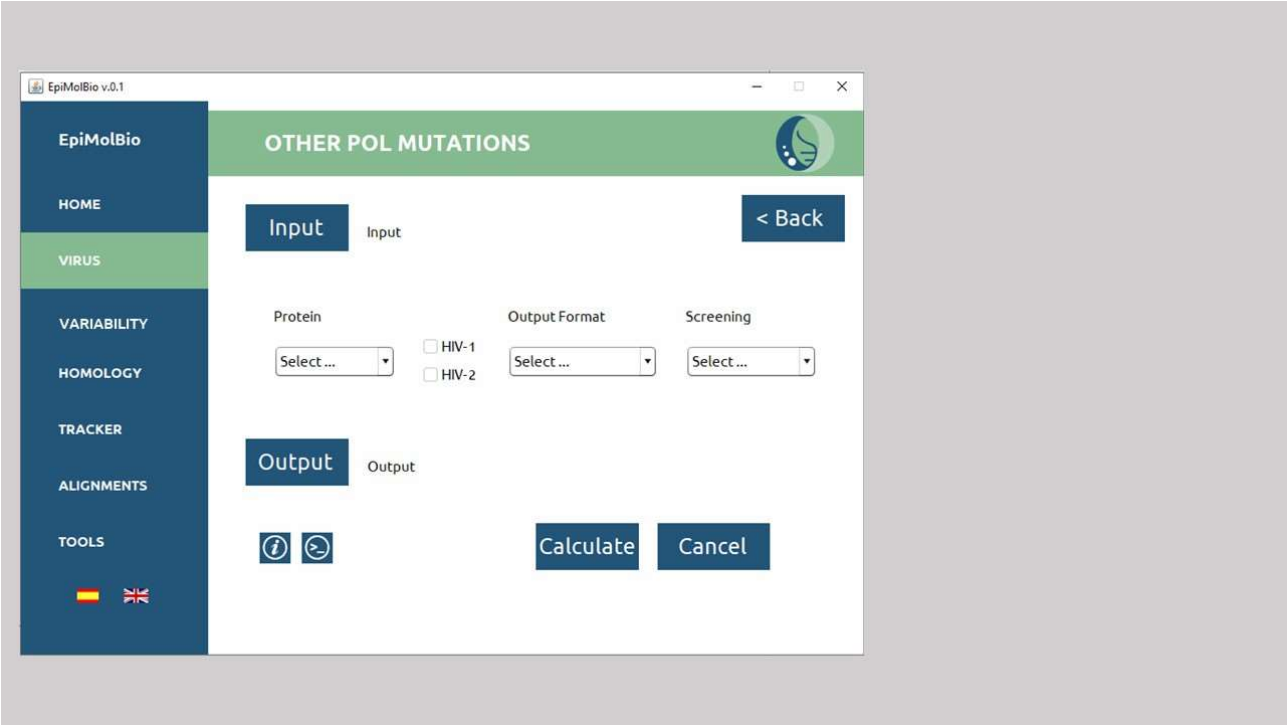
| Summary Table Other Pol Mutations Protease HIV-1 > 75% | | |
|---|---|---|
| File | Residues | Total Sequences |
| PR_procesado_traducido_01_AE.fasta | V3I, E35D, M36I, S37N, R41K, H69K, L89M | 26849 |
| PR_procesado_traducido_02_AG.fasta | V3I, I13V, K20I, M36I, R41K, H69K, L89M | 9577 |
| PR_procesado_traducido_03_A6B.fasta | V3I, E35D, M36I, S37N, R41K, H69K, L89M | 310 |
| PR_procesado_traducido_04_cpx.fasta | V3I, M36I, R41K, H69K | 15 |
| PR_procesado_traducido_05_DF.fasta | V3I, S37N, R41K | 24 |

# Step-by-step:

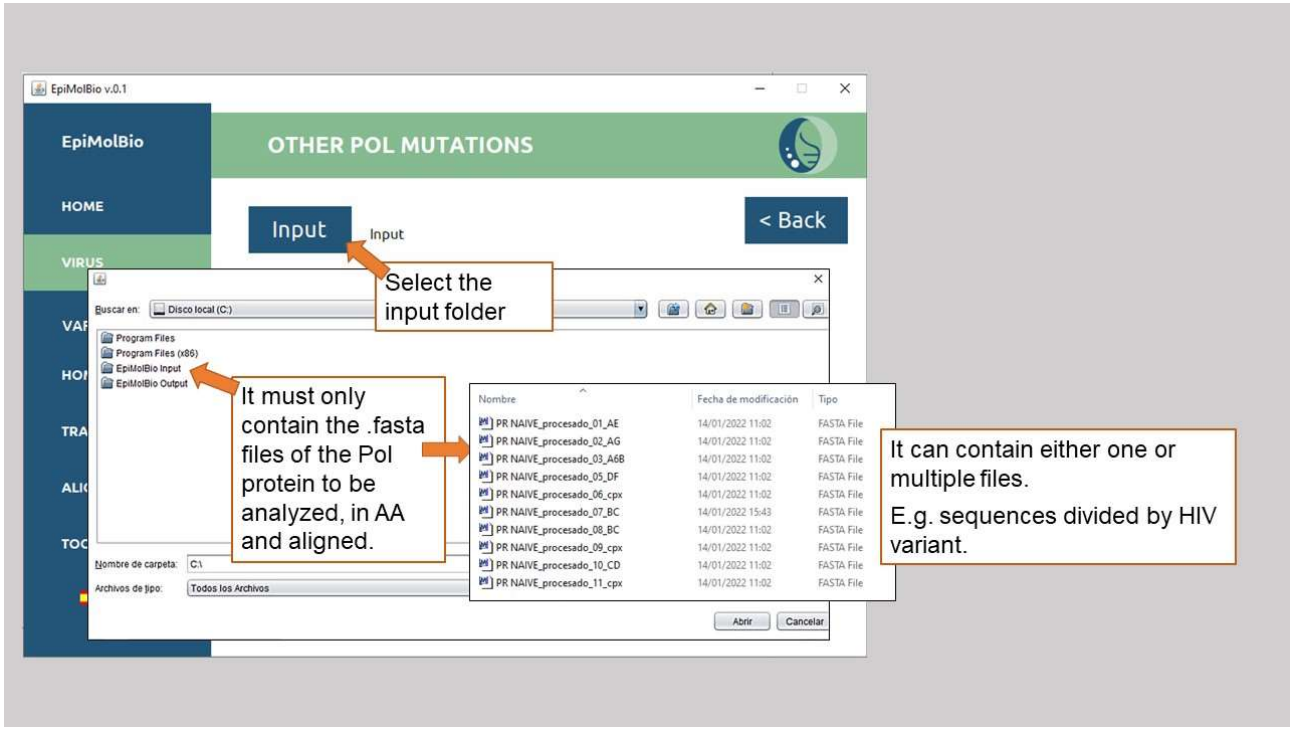
## 1)



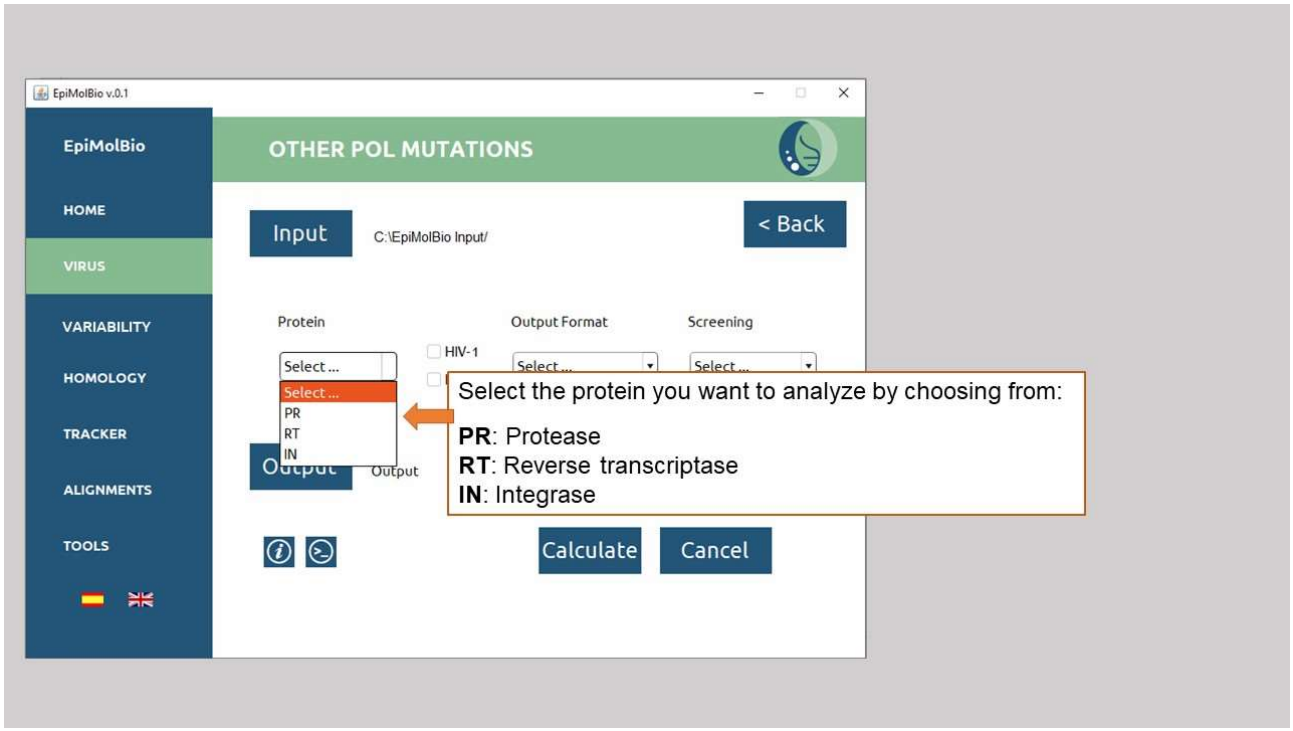

## 2)



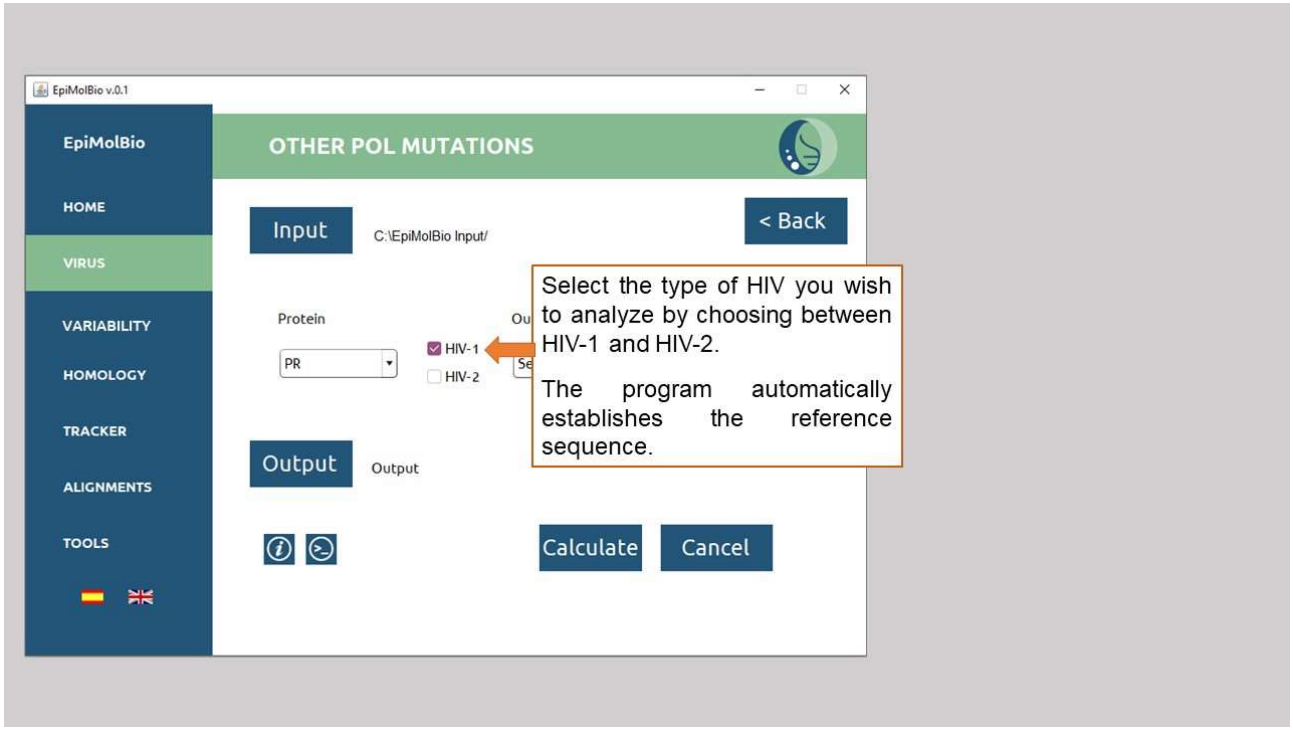
## 3)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
OTHER POL MUTATIONS
Input
Input
< Back
Select the input folder
Buscar en:
Disco local (C:)
Program Files
Program Files (x86)
EpiMolBio Input
EpiMolBio Output
It must only contain the .fasta files of the Pol protein to be analyzed, in AA and aligned.
Nombre
Fecha de modificación
Tipo
PR NAIVE_procesado_01_AE
PR NAIVE_procesado_02_AG
PR NAIVE_procesado_03_A6B
PR NAIVE_procesado_05_DF
PR NAIVE_procesado_06_cpx
PR NAIVE_procesado_07_BC
PR NAIVE_procesado_08_BC
PR NAIVE_procesado_09_cpx
PR NAIVE_procesado_10_CD
PR NAIVE_procesado_11_cpx
14/01/2022 11:02
14/01/2022 15:43
FASTA File
Nombre de carpeta:
Archivos de tipo:
Todos los Archivos
Abrir
Cancelar
It can contain either one or multiple files.
E.g. sequences divided by HIV variant.

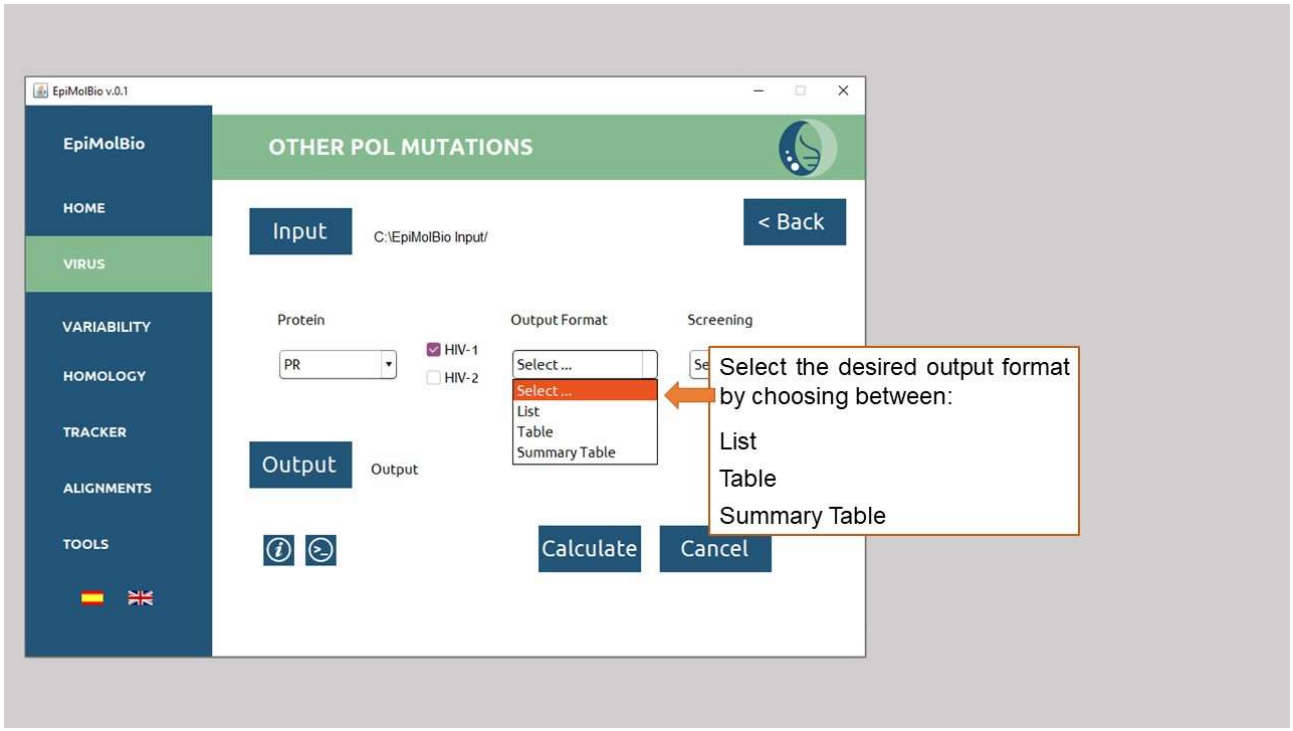
## 4)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
OTHER POL MUTATIONS
Input
C:\EpiMolBio Input/
< Back
Protein
Output Format
Screening
HIV-1
Select ...
Select ...
PR
RT
IN
Output
Calculate
Cancel
Select the protein you want to analyze by choosing from:
PR: Protease
RT: Reverse transcriptase
IN: Integrase

## 5)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
OTHER POL MUTATIONS
Input
C:\EpiMolBio Input/
< Back
Protein
PR
HIV-1
HIV-2
Output
Output
Calculate
Cancel
Select the type of HIV you wish to analyze by choosing between HIV-1 and HIV-2.
The program automatically establishes the reference sequence.

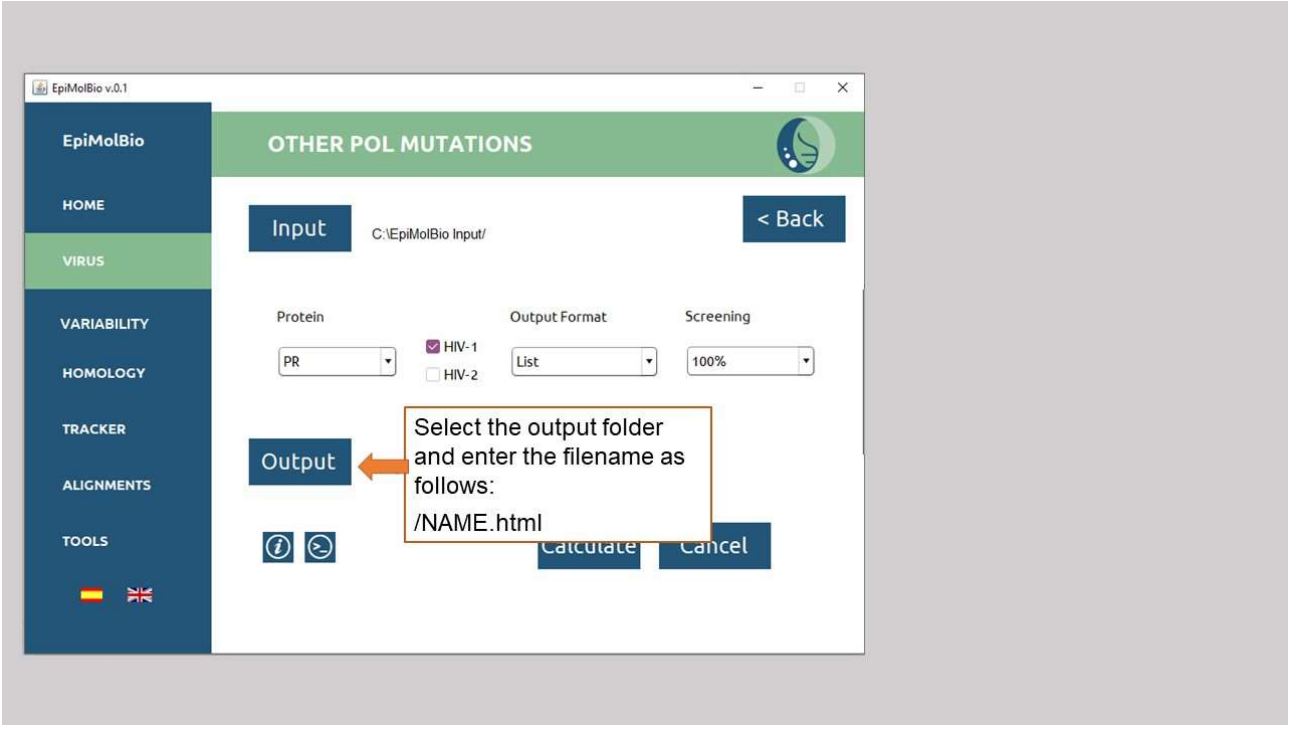
## 6)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
OTHER POL MUTATIONS
Input
C:\EpiMolBio Input/
< Back
Protein
PR
HIV-1
HIV-2
Output Format
Select ...
Select ...
List
Table
Summary Table
Screening
Output
Output
Calculate
Cancel
Select the desired output format by choosing between:
List
Table
Summary Table

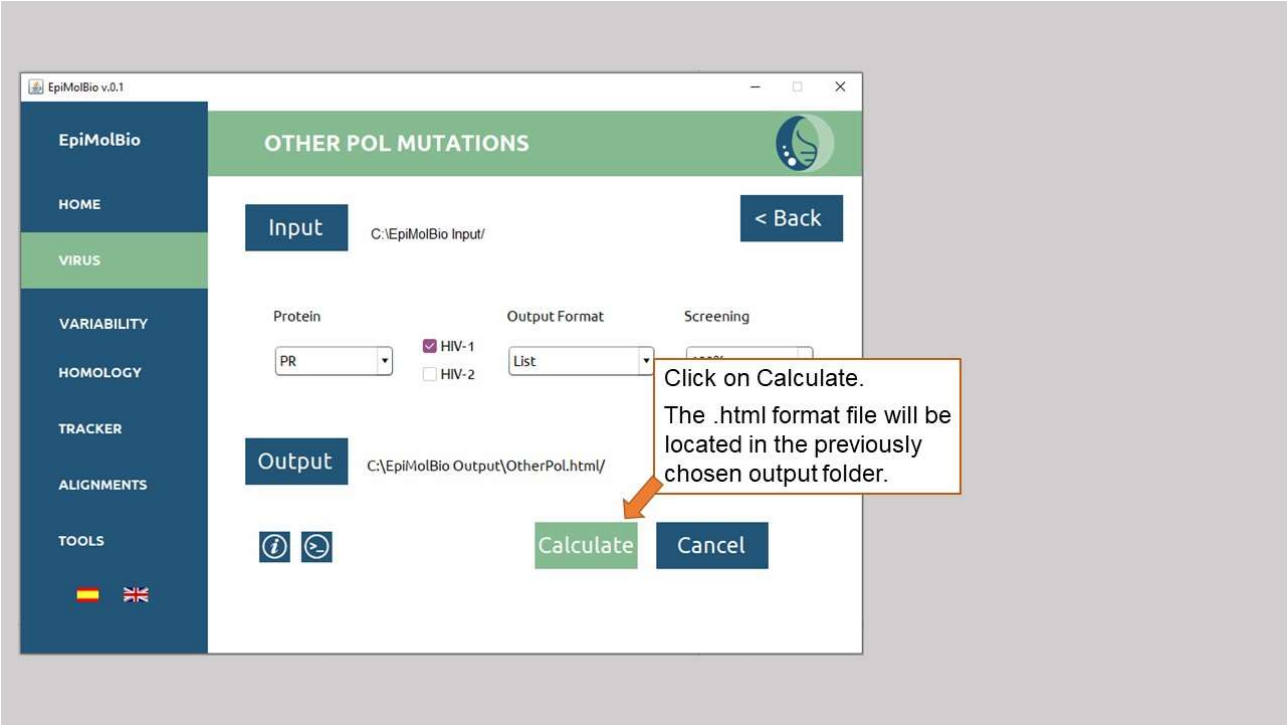
## 7)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
OTHER POL MUTATIONS
Input
C:\EpiMolBio Input/
< Back
Protein
PR
HIV-1
HIV-2
Output Format
List
Screening
Select ...
Select ...
100%
> 75%
Output
Output
Calculate
Cancel
If you choose the List Output Format:
Select the screening percentage you want to apply: 100% or >75%.

## 8)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
OTHER POL MUTATIONS
Input
C:\EpiMolBio Input/
< Back
Protein
PR
HIV-1
HIV-2
Output Format
List
Screening
100%
Output
Select the output folder and enter the filename as follows:
/NAME.html
Calculate
Cancel

**9)**


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
OTHER POL MUTATIONS
Input
C:\EpiMolBio Input/
< Back
Protein
PR
HIV-1
HIV-2
Output Format
List
Screening
Output
C:\EpiMolBio Output\OtherPol.html/
Calculate
Cancel
Click on Calculate.
The .html format file will be located in the previously chosen output folder.

# I.1.C) POL COONSERVATION

This function generates an .html table that shows the consensus of the input sequences with the most prevalent amino acid at each position in the selected Pol protein's amino acid sequence. It allows you to identify the most conserved residue of the protein for each position. Additionally, it displays a table with residues that have a conservation level above 75% for each position. Gaps (-) and question marks (?) are excluded from the analysis. This function is suitable for the *pol* gene proteins of both HIV-1 and HIV-2.

The **input file** should be a folder containing exclusively aligned sequences of the Pol protein that you want to analyze (protease, reverse transcriptase, or integrase) in amino acids and in .fasta format. This folder may contain a single file or multiple .fasta files if you want to analyze the same in different sequence groups (e.g., files divided by HIV variants).

In the '**Protein**' field, you will need to select the protein you want to analyze:

**PR**: protease

**RT**: reverse transcriptase

**IN**: integrase

Select the **type of HIV** to be analyzed: HIV-1 (HXB2, NCBI K03455.1) or HIV-2 (ALI, NCBI AF082339), to establish the length of each protein according to the reference sequence.

The **output file** will be an .html file. You need to select the output folder where you want the .html files to appear and name the files by writing .html at the end.

In the output file, at the top, you will find the title of the analysis, followed by the name of the input file. Below that, the consensus sequence for each file will be displayed, with residues colored according to their occurrence percentage, following the color code described in the Overview. If more than one residue has the same percentage at the same position, they will appear in parentheses. Below the consensus sequence, in the 'Position' column, each position with its reference amino acid will be listed. In the 'Residues' column, the most frequent amino acid for each position will be shown, followed by its occurrence percentage, colored according to the color code, provided that its frequency is higher than 75%. The 'Total Positions' column displays the total number of valid residues for that position.

Example of output format for Pol Conservation analysis:

**Pol Conservation Protease HIV-1 > 75%**

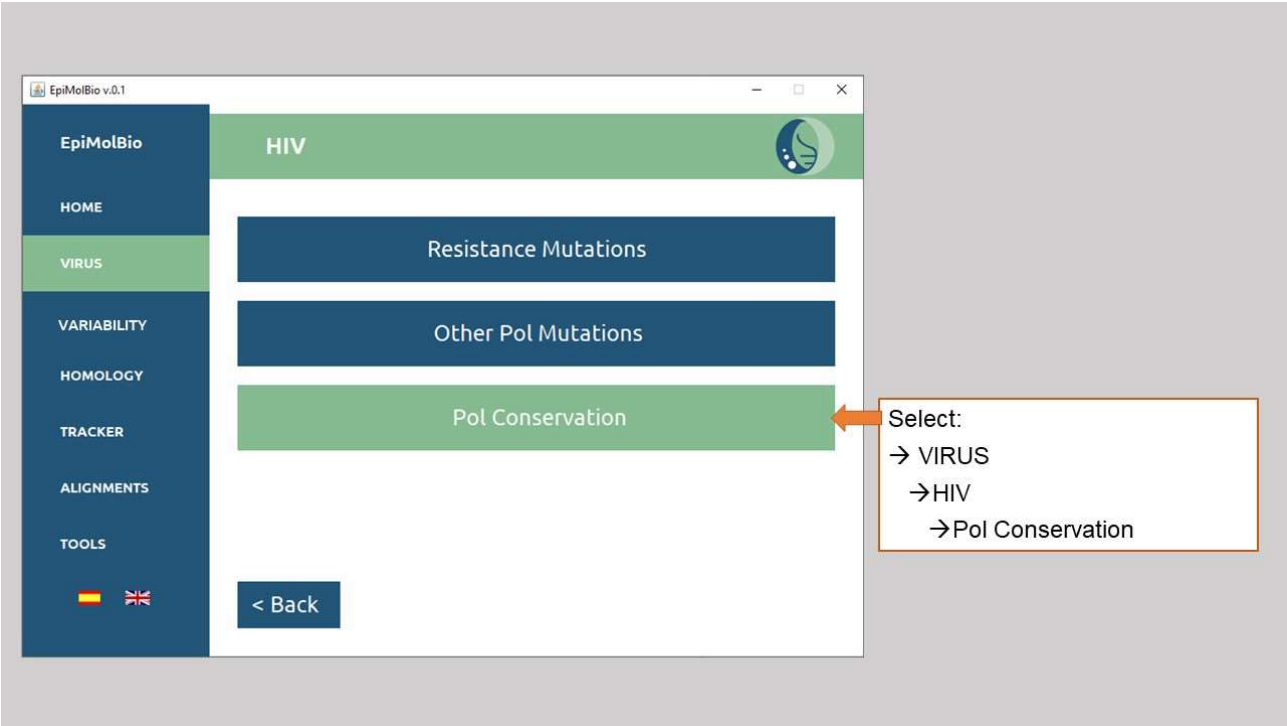
**PR_procesado_traducido_01_AE.fasta**

CONSENSUS PQITLWQRPLVTVKIGGQLKEALLDTGADDTVLEDINLPGKWKPKMIGGIGGFIKVRQYDQILIEICGKKAIGTVLVGPTPVNIIGRNMLTQIGCTLNF

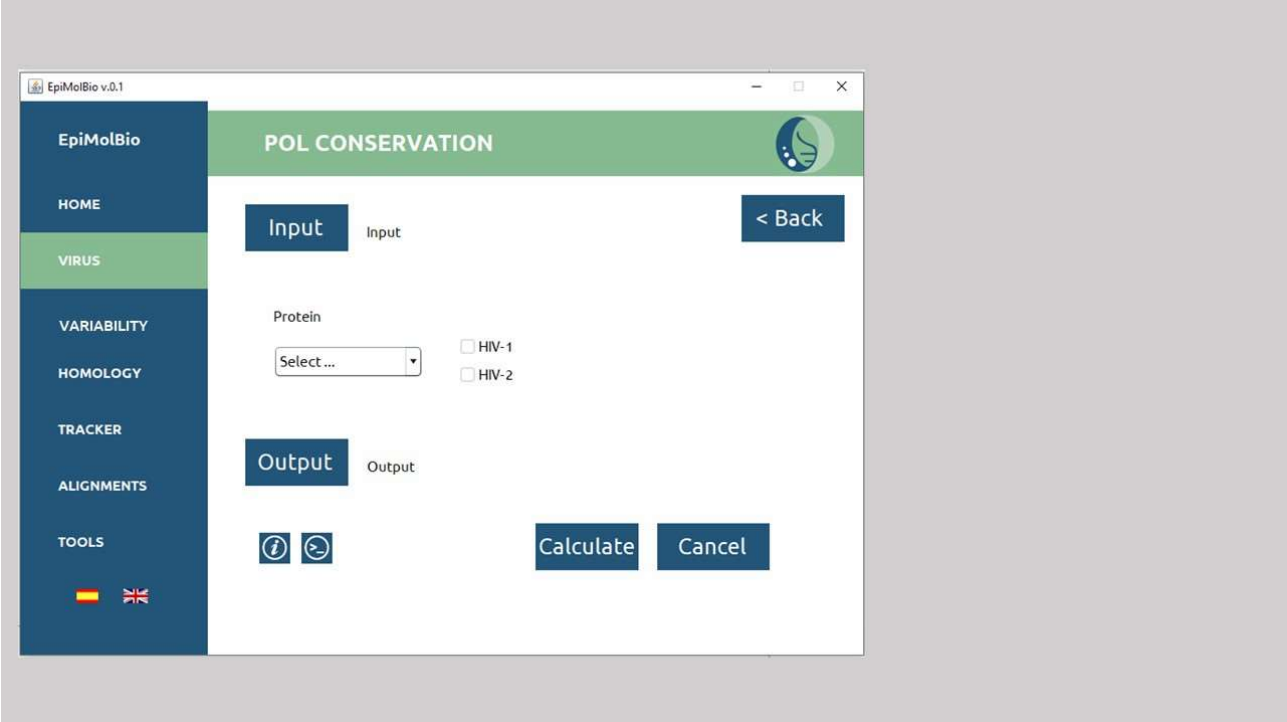
| Position | Residues | Total Positions |
|---|---|---|
| P1 | P(99.896%) | 26838 |
| Q2 | Q(99.782%) | 26649 |
| V3 | I(99.858%) | 26831 |
| T4 | T(99.858%) | 26816 |
| L5 | L(99.888%) | 26780 |
| W6 | W(99.929%) | 26836 |
| Q7 | Q(99.751%) | 26536 |
| R8 | R(99.922%) | 26792 |
| P9 | P(99.955%) | 26613 |

# Step-by-step:

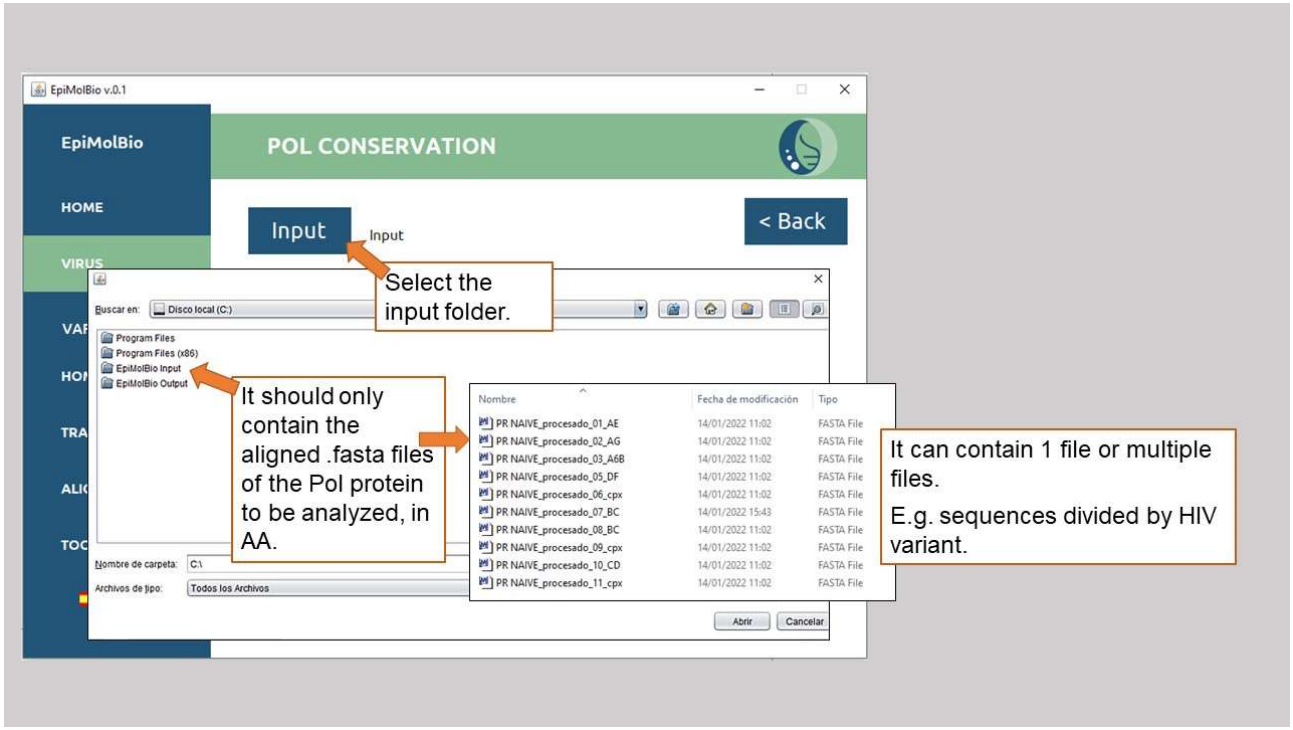
## 1)



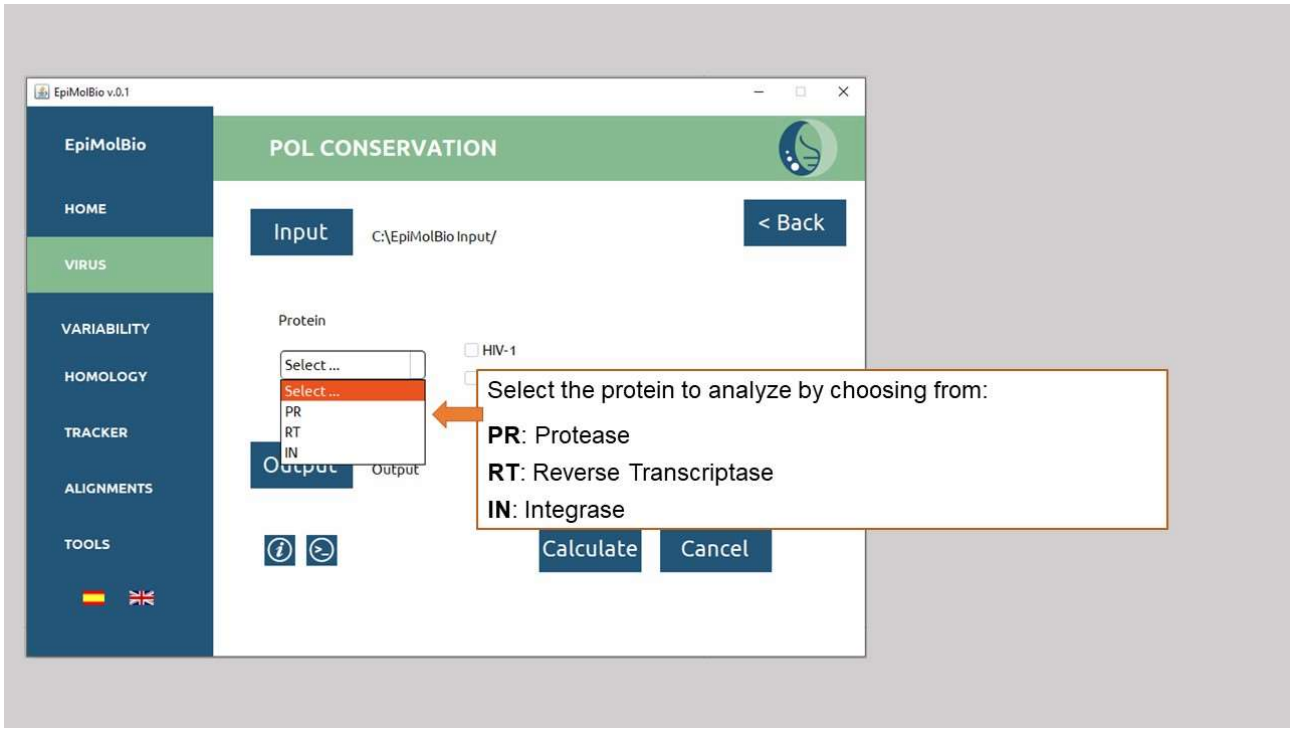

## 2)



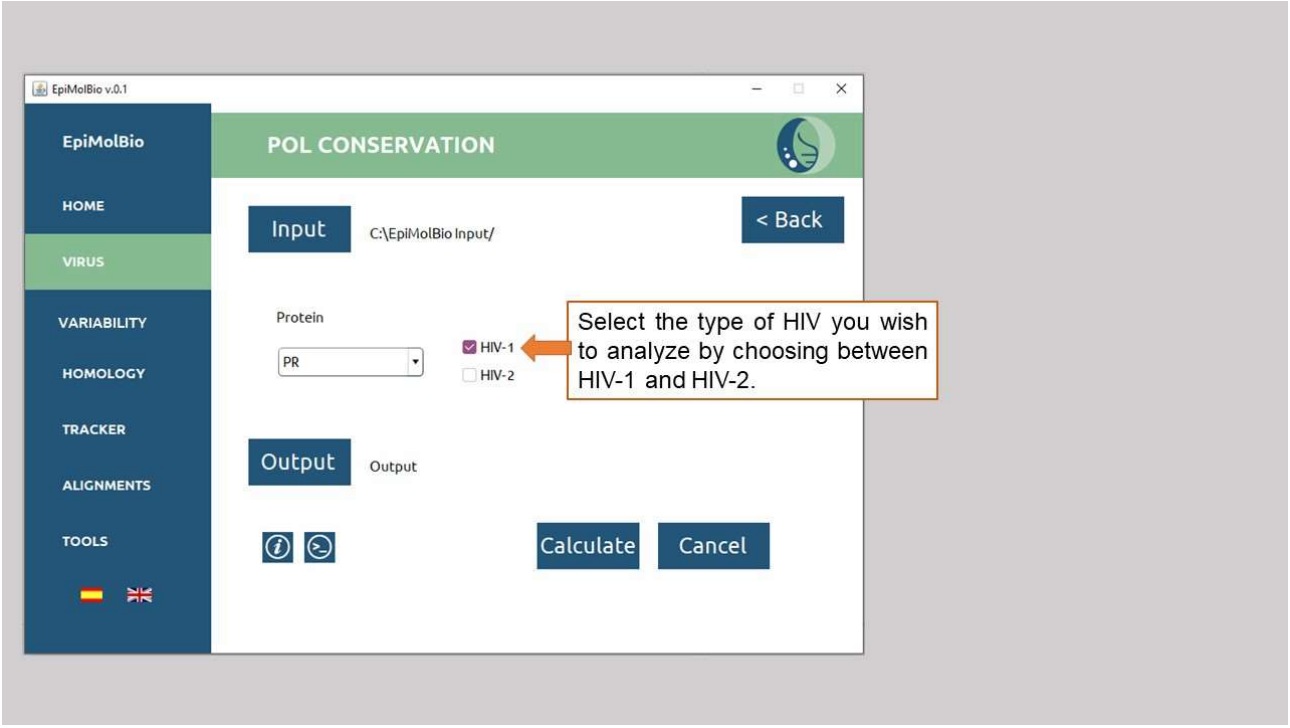
## 3)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
POL CONSERVATION
Input
Input
< Back
Select the input folder.
Buscar en:
Disco local (C:)
Program Files
Program Files (x86)
EpiMolBio Input
EpiMolBio Output
It should only contain the aligned .fasta files of the Pol protein to be analyzed, in AA.
Nombre
Fecha de modificación
Tipo
PR NAIVE_procesado_01_AE
PR NAIVE_procesado_02_AG
PR NAIVE_procesado_03_A6B
PR NAIVE_procesado_05_DF
PR NAIVE_procesado_06_cpx
PR NAIVE_procesado_07_BC
PR NAIVE_procesado_08_BC
PR NAIVE_procesado_09_cpx
PR NAIVE_procesado_10_CD
PR NAIVE_procesado_11_cpx
FASTA File
It can contain 1 file or multiple files.
E.g. sequences divided by HIV variant.
Nombre de carpeta:
Archivos de tipo:
Todos los Archivos
Abrir
Cancelar

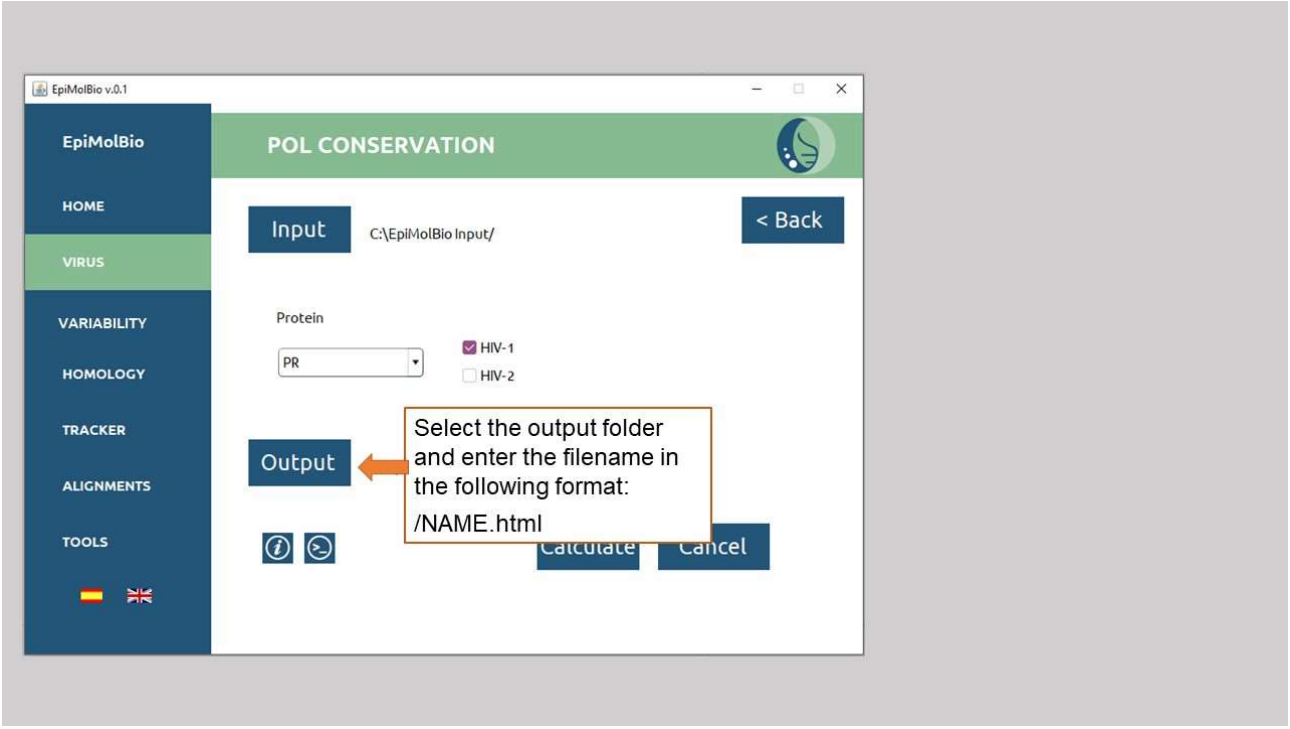
## 4)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
POL CONSERVATION
Input
C:\EpiMolBio Input/
< Back
Protein
Select ...
Select ...
PR
RT
IN
HIV-1
Output
Output
Select the protein to analyze by choosing from:
PR: Protease
RT: Reverse Transcriptase
IN: Integrase
Calculate
Cancel

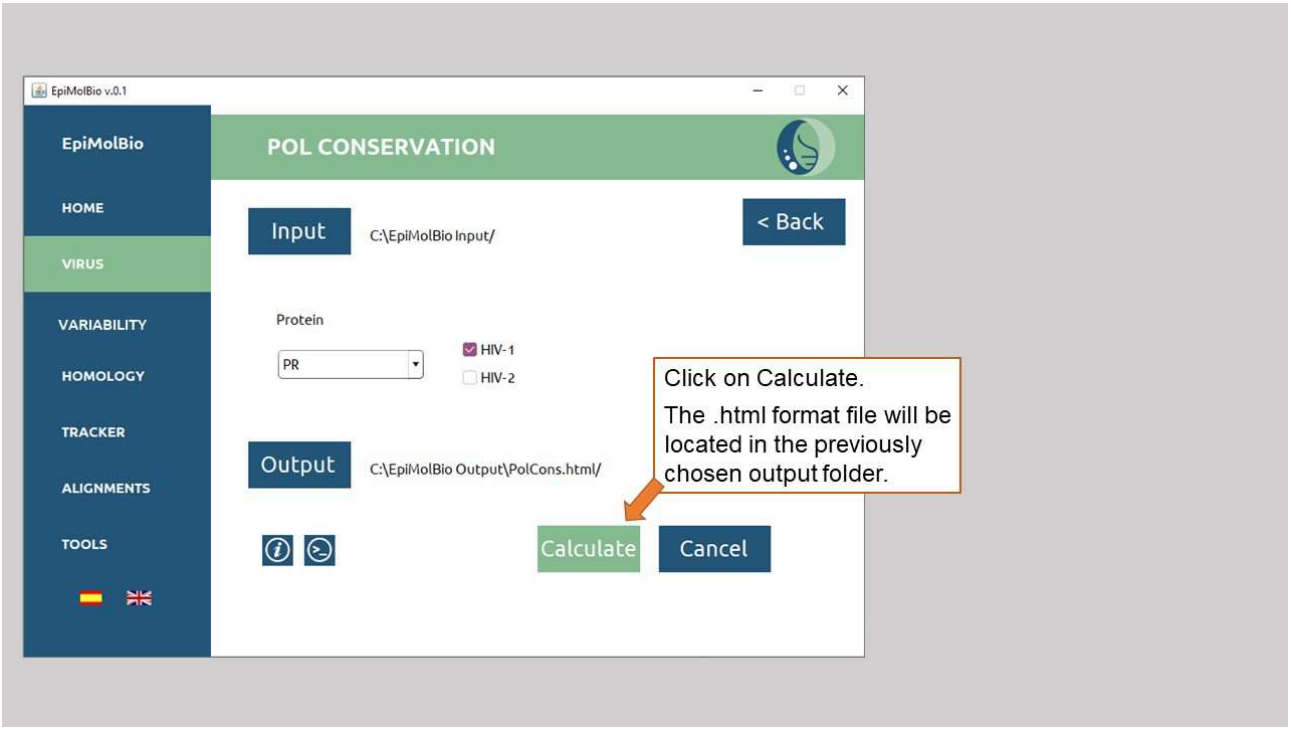
## 5)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
POL CONSERVATION
Input
C:\EpiMolBio Input/
< Back
Protein
PR
HIV-1
HIV-2
Select the type of HIV you wish to analyze by choosing between HIV-1 and HIV-2.
Output
Output
Calculate
Cancel

## 6)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
POL CONSERVATION
Input
C:\EpiMolBio Input/
< Back
Protein
PR
HIV-1
HIV-2
Output
Select the output folder and enter the filename in the following format:
/NAME.html
Calculate
Cancel

# 7)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
POL CONSERVATION
Input
C:\EpiMolBio Input/
< Back
Protein
PR
HIV-1
HIV-2
Output
C:\EpiMolBio Output\PolCons.html/
Calculate
Cancel
Click on Calculate.
The .html format file will be located in the previously chosen output folder.

# I.2. SARS-CoV-2

# PROTEIN TRACKER

This function generates .fasta files with nucleotide or amino acid sequences of the selected proteins from SARS-CoV-2 based on complete genome sequences.

The **input file** should be a folder containing exclusively .fasta files with complete genome sequences of SARS-CoV-2 in nucleotides. EpiMoBio tracks the proteins within the range of positions found in the reference sequence Wuhan (NC 045512.2).

In the '**Select Protein**' field, you will need to choose the SARS-CoV-2 protein you want to track, or select 'All' to track all proteins.

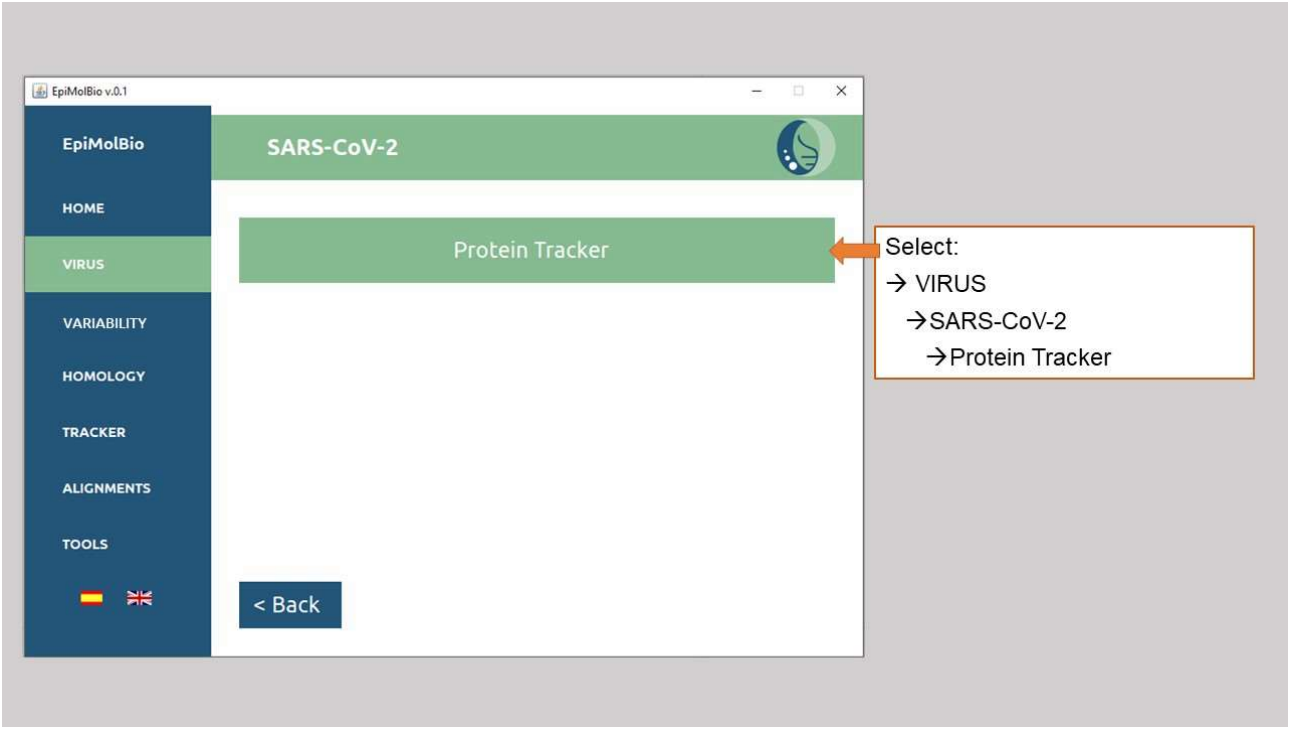
Next, choose whether you want the output .fasta file in nucleotides (select '**Not Translate**') or translated into amino acids (select '**Translate**').

The **output file** will be a .fasta file. For each file in the input folder, a corresponding output file will be generated with the found sequences in nucleotides or translated into amino acids. The number of input sequences may not match the number of tracked proteins. When an input sequence contains many mutations or is incomplete in the region that corresponds to the searched protein, the Protein Tracker cannot recognize it and therefore, the program will not retrieve it. The effectiveness of the Protein Tracker is summarized in **Appendix II**.

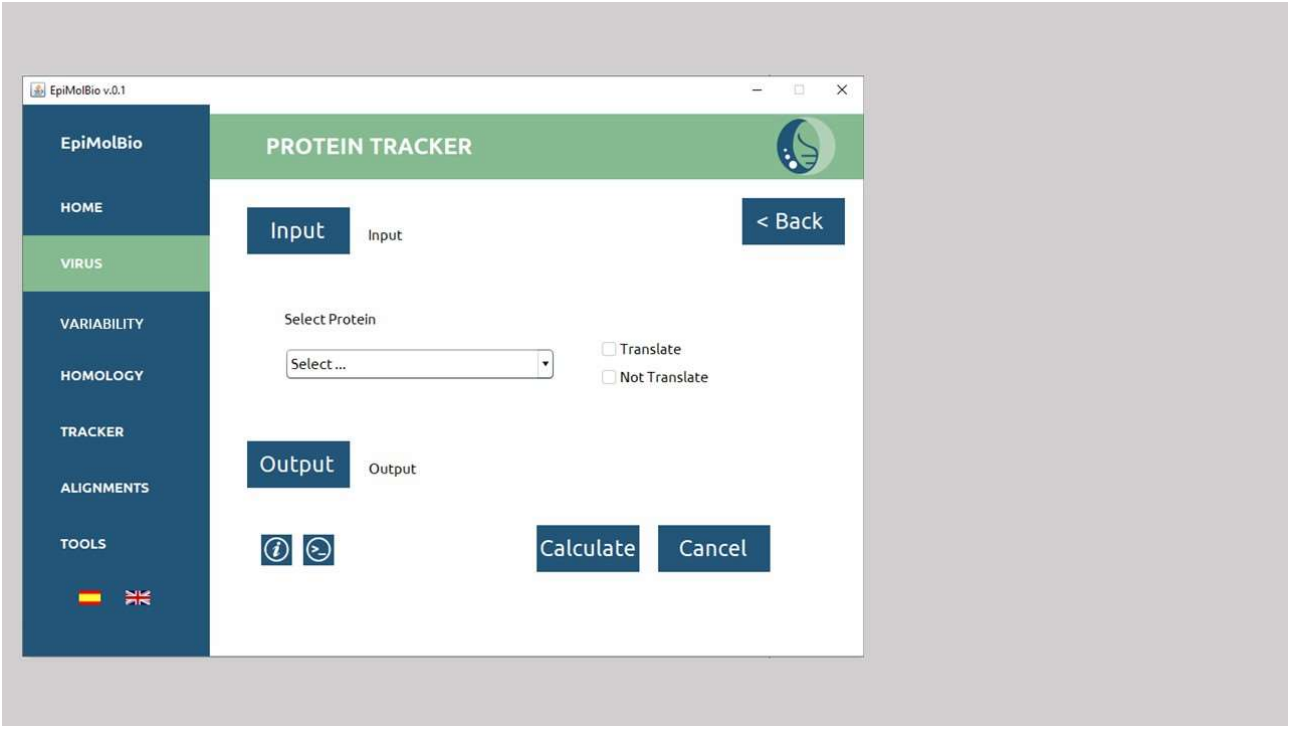
You will need to select the **output folder** where you want the .fasta files to appear. The file will be automatically named as follows: Selected Protein_Tracked_Input File Name.fasta (e.g., S (Spike)_Tracked_sequences.fasta). If 'All' is selected in 'Select Protein,' each protein will be separated into a .fasta file with the corresponding name.

# Step-by-step

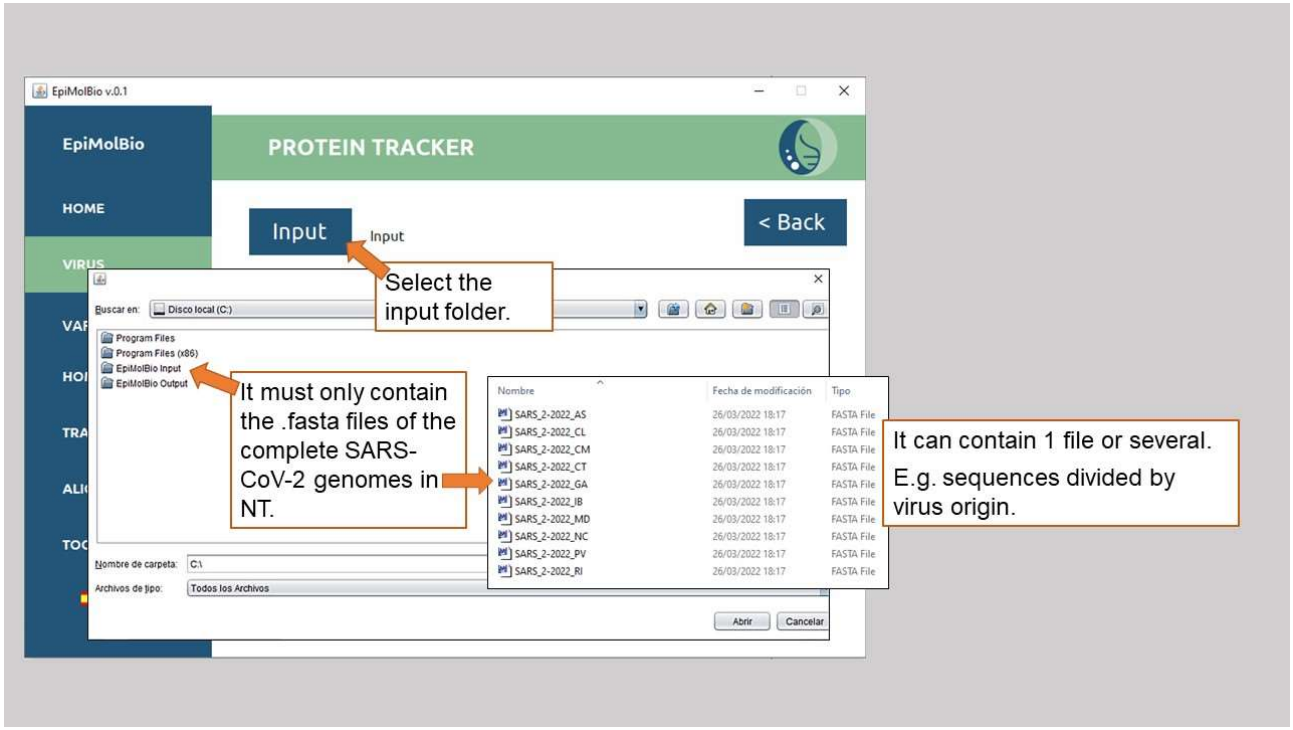
## 1)



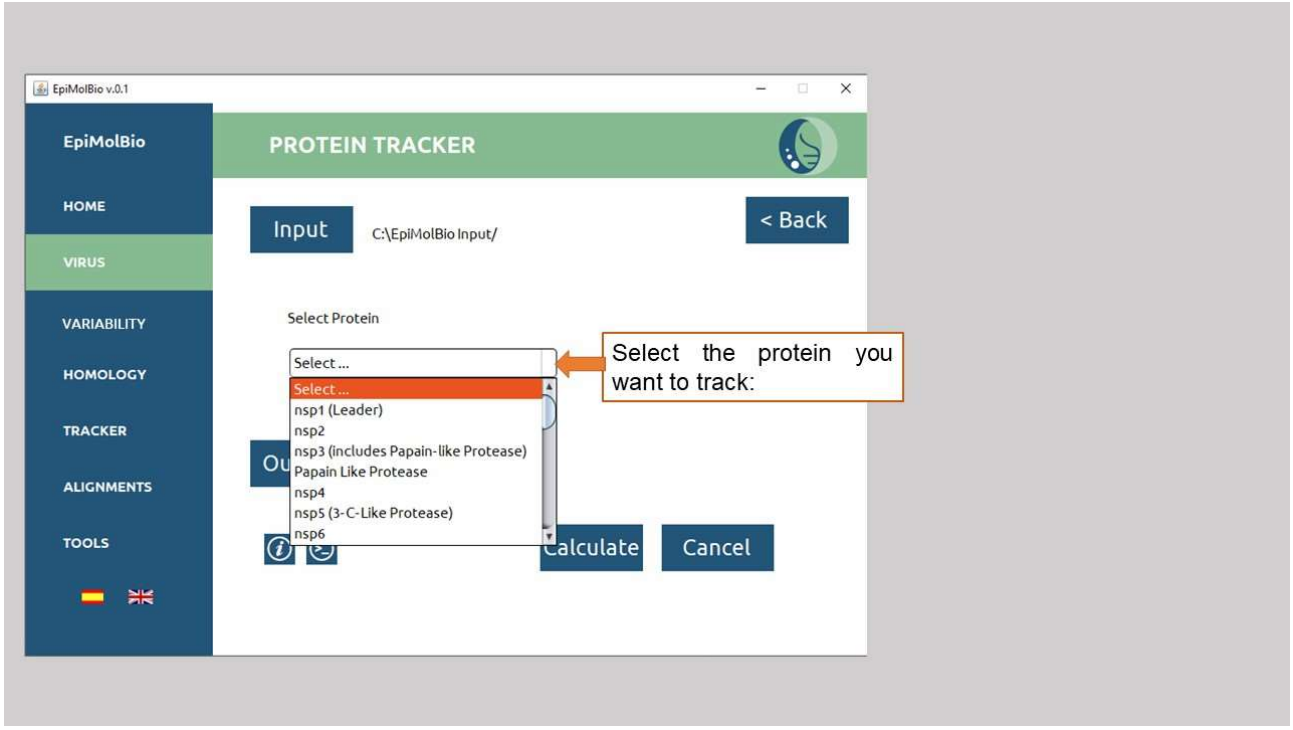

## 2)



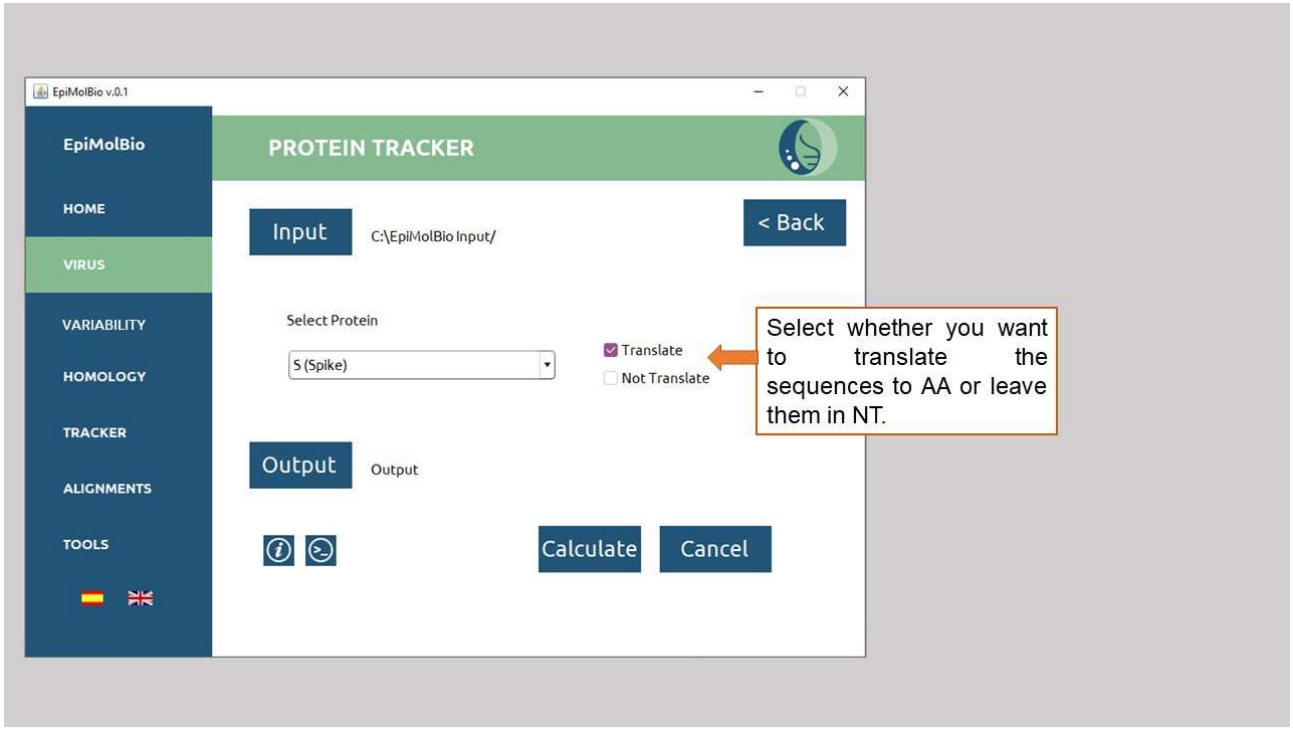
## 3)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
PROTEIN TRACKER
Input
Input
< Back
Select the input folder.
Buscar en:
Disco local (C:)
Program Files
Program Files (x86)
EpiMolBio Input
EpiMolBio Output
It must only contain the .fasta files of the complete SARS-CoV-2 genomes in NT.
Nombre
Fecha de modificación
Tipo
SARS_2-2022_AS
SARS_2-2022_CL
SARS_2-2022_CM
SARS_2-2022_CT
SARS_2-2022_GA
SARS_2-2022_IB
SARS_2-2022_MD
SARS_2-2022_NC
SARS_2-2022_PV
SARS_2-2022_RI
26/03/2022 18:17
FASTA File
It can contain 1 file or several.
E.g. sequences divided by virus origin.
Nombre de carpeta:
C:\
Archivos de tipo:
Todos los Archivos
Abrir
Cancelar

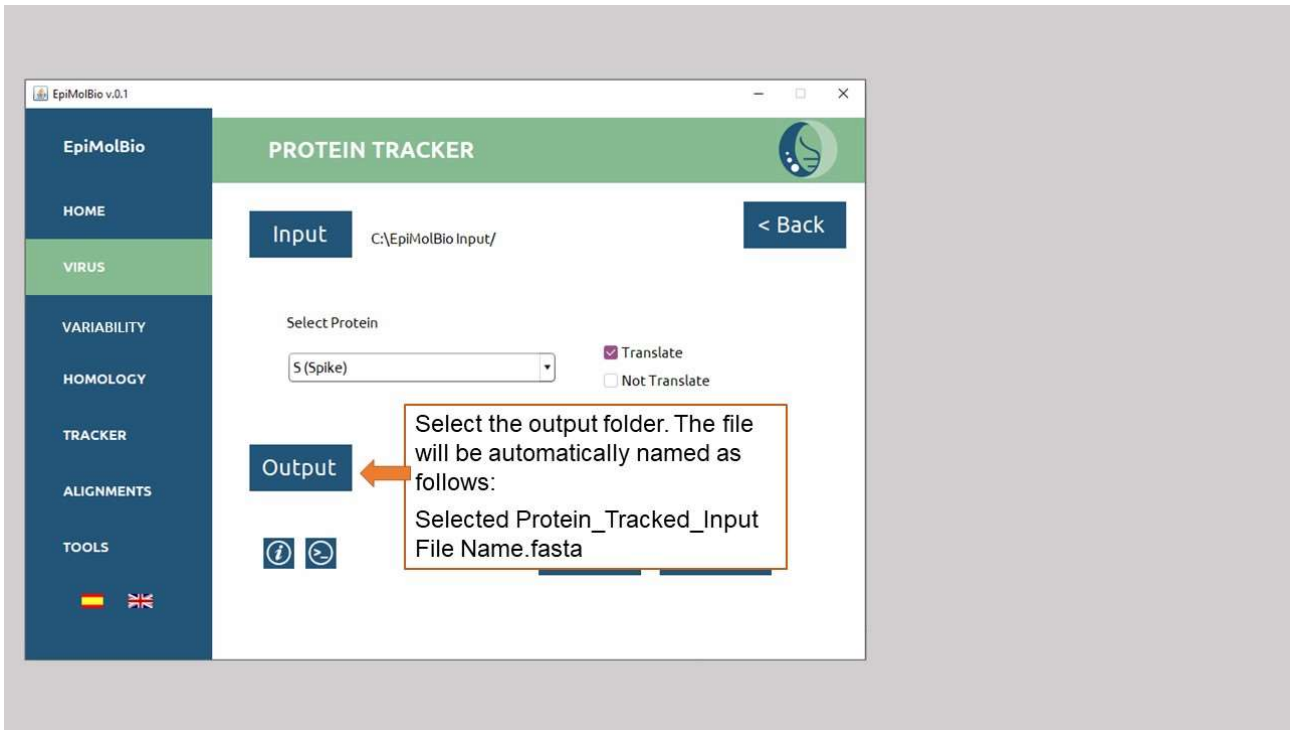
## 4)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
PROTEIN TRACKER
Input
C:\EpiMolBio Input/
< Back
Select Protein
Select ...
Select ...
nsp1 (Leader)
nsp2
nsp3 (includes Papain-like Protease)
Papain Like Protease
nsp4
nsp5 (3-C-Like Protease)
nsp6
Select the protein you want to track:
Calculate
Cancel

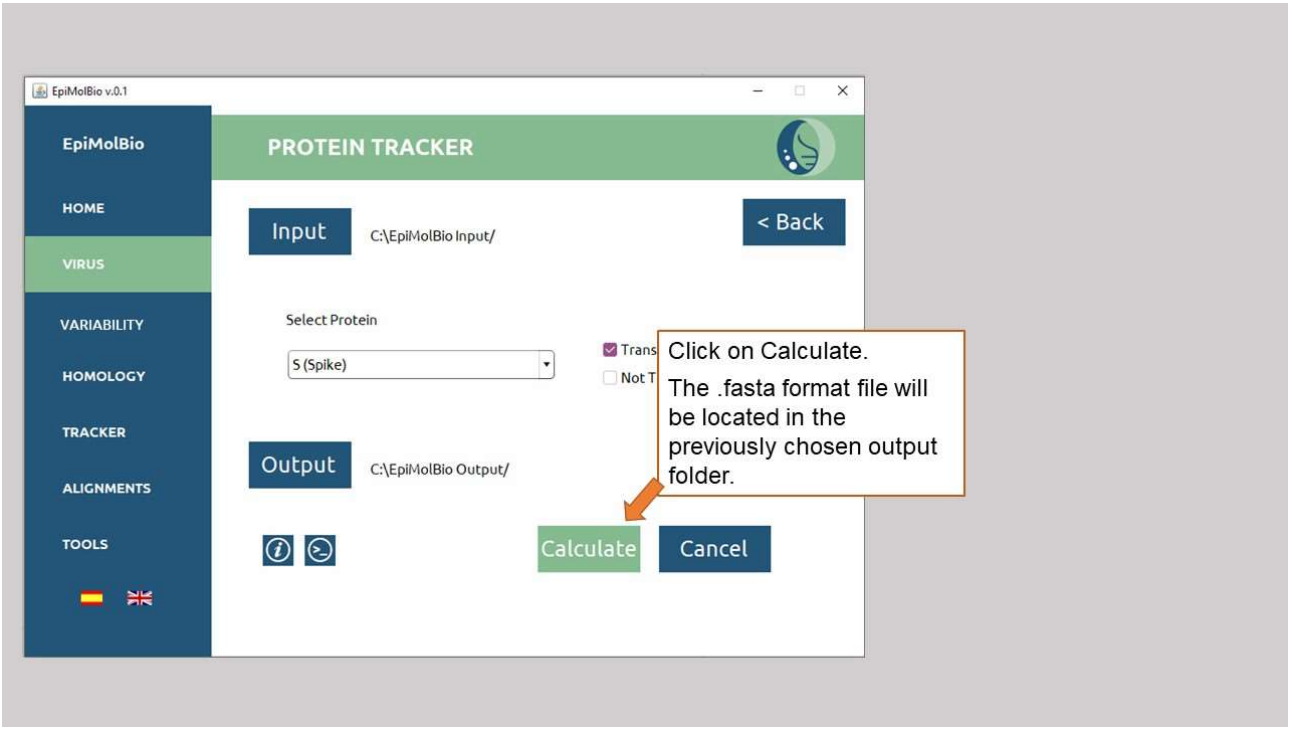
## 5)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
PROTEIN TRACKER
Input
C:\EpiMolBio Input/
< Back
Select Protein
S (Spike)
Translate
Not Translate
Select whether you want to translate the sequences to AA or leave them in NT.
Output
Output
Calculate
Cancel

## 6)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
PROTEIN TRACKER
Input
C:\EpiMolBio Input/
< Back
Select Protein
S (Spike)
Translate
Not Translate
Output
Select the output folder. The file will be automatically named as follows:
Selected Protein_Tracked_Input File Name.fasta

## 7)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
PROTEIN TRACKER
Input
C:\EpiMolBio Input/
< Back
Select Protein
S (Spike)
Trans
Not T
Output
C:\EpiMolBio Output/
Calculate
Cancel
Click on Calculate.
The .fasta format file will be located in the previously chosen output folder.

# II. VARIABILITY

In this section, you will find several specific tools for the analysis of genetic sequence variability.

# II.1. POLYMORPHISMS

It allows the detection of polymorphisms, providing their location and frequency of occurrence. The search can be performed by analyzing each residue individually (II.1.A) or by analyzing codons (II.1.B).

## II.1.A) INDIVIDUAL

This tool allows obtaining the frequencies of occurrence of mutations using any sequence provided by the user as a reference. These analyses can be performed for both amino acid and nucleotide sequences. To perform the analysis on nucleotide sequences, it will be necessary to use the **Find and Replace** tool in **File Editing** and replace 'N' with '?' so that they are excluded from the analysis.

**1.-Mutated Positions:**

This tool allows **detecting, locating, and determining the frequency of mutations at one or several positions** with respect to a user-provided reference sequence. With this tool, you can choose to search for specific positions or select the minimum frequency of mutations you want to display. Both gaps (-) and question marks (?) are excluded from the analysis.

In the output file, at the top, you will find the title of the analysis followed by the name of the input file. Below that, in the 'Position' column, each position is listed with its reference amino acid or nucleotide based on the user-provided reference sequence. In the 'Residues' column, the found residue is displayed along with its occurrence percentage, colored according to the color code described in the Overview, which can be consulted in the .html output file by clicking on the blue symbol. If the 'All Positions' field is selected, the percentage of the reference residue will also be shown. The 'Total Positions' column displays the total number of valid sequences for that position.

Example of Mutated Positions output format with minimum value of 0.0 selecting All Positions:

**Variability Polymorphisms Individual All Positions**

**PR_01_AE.fasta**

| Position | Residues | Total Positions |
|---|---|---|
| P1 | P(99.896%) S(0.078%) A(0.004%) L(0.007%) T(0.007%) H(0.004%) V(0.004%) | 26838 |
| Q2 | Q(99.782%) E(0.071%) S(0.019%) H(0.056%) D(0.004%) K(0.023%) L(0.015%) P(0.008%) R(0.011%) T(0.004%) *(0.008%) | 26649 |
| I3 | I(99.858%) V(0.078%) N(0.015%) L(0.041%) T(0.007%) | 26831 |
| T4 | T(99.858%) M(0.004%) I(0.048%) N(0.019%) P(0.022%) S(0.034%) F(0.004%) A(0.007%) H(0.004%) | 26816 |
| L5 | L(99.888%) F(0.075%) V(0.015%) S(0.004%) R(0.007%) I(0.007%) T(0.004%) | 26780 |
| W6 | W(99.929%) G(0.030%) R(0.022%) *(0.007%) C(0.011%) | 26836 |

To perform this analysis, the **input format** must be **folder** containing exclusively .fasta files of the sequences to be analyzed in nucleotides or amino acids.

Next, you should choose '**Mutated Positions**' in the dropdown menu for '**Output Format**'.

In the '**Reference**' field, you need to input the reference sequence in letters without line breaks, considering that the sequence should be entered in nucleotides or amino acids based on whether the input files are untranslated or translated.

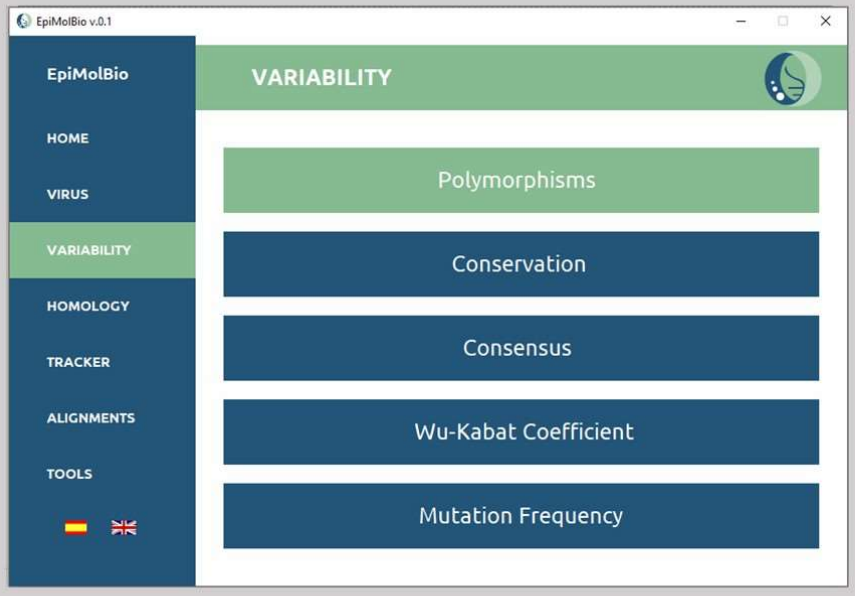
The '**Minimum Value**' field can be left blank if you want the result to show all detected mutations regardless of their frequency of occurrence. If you want to filter mutations based on a specific frequency, you need to input the minimum value in decimal numeric format (e.g., 90.0 to show only mutations that occur at a percentage greater than 90%).

The '**Mutations**' field can be left blank if you want the result to show all detected mutations. If you want to search for specific mutations, you need to input the position of the residue where the mutation is sought (e.g., 2). When the input files and the reference are in nucleotides, enter the position that corresponds to the desired mutation in nucleotides (e.g., 6). If you want to search for multiple mutations, separate them with a comma ',' without spaces (e.g., 6,8,11).

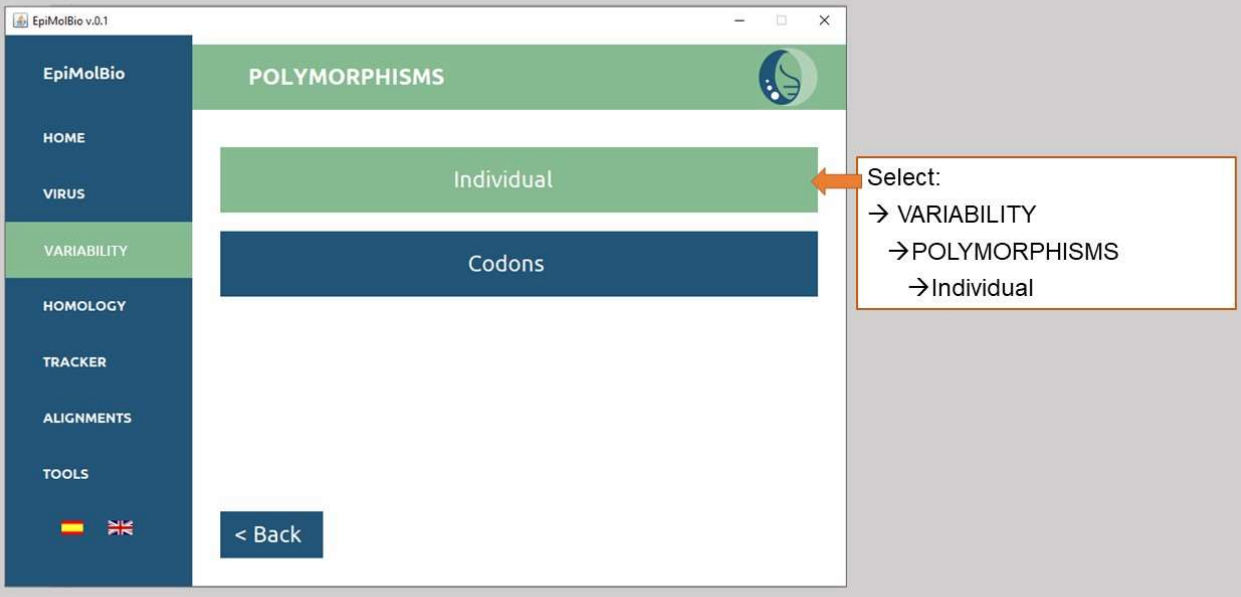
The result is displayed in an .html file. In the '**Output**' field, you should select the folder where you want to save the result and name the file with the .html extension.

# Step-by-step:

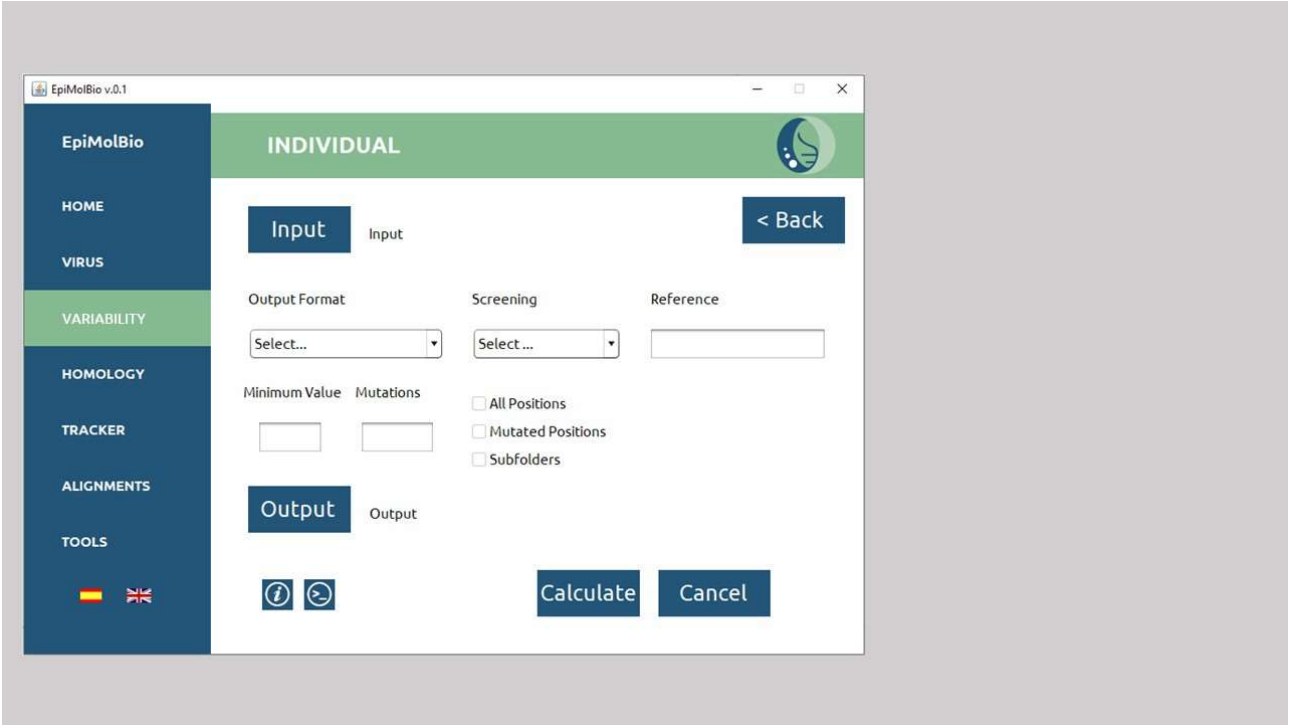
## 1)



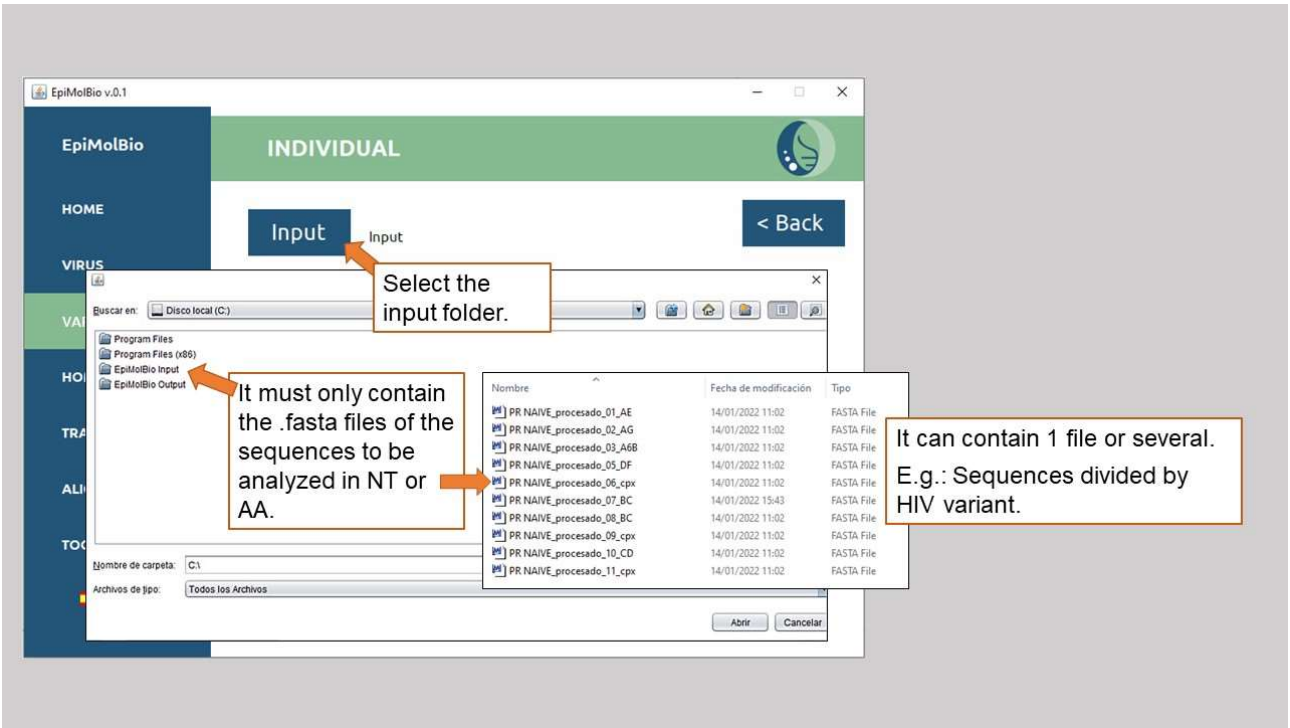

## 2)



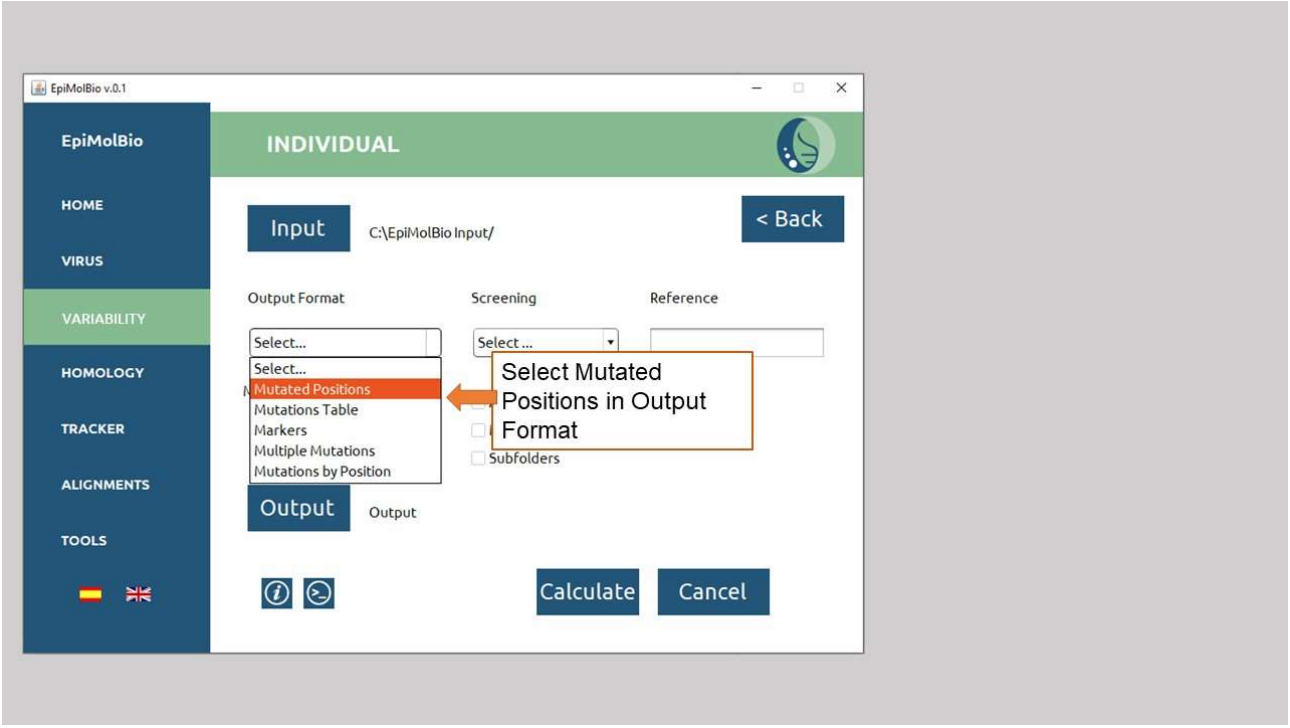
## 3)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
INDIVIDUAL
Input
Input
< Back
Output Format
Screening
Reference
Select...
Select ...
Minimum Value
Mutations
All Positions
Mutated Positions
Subfolders
Output
Output
Calculate
Cancel

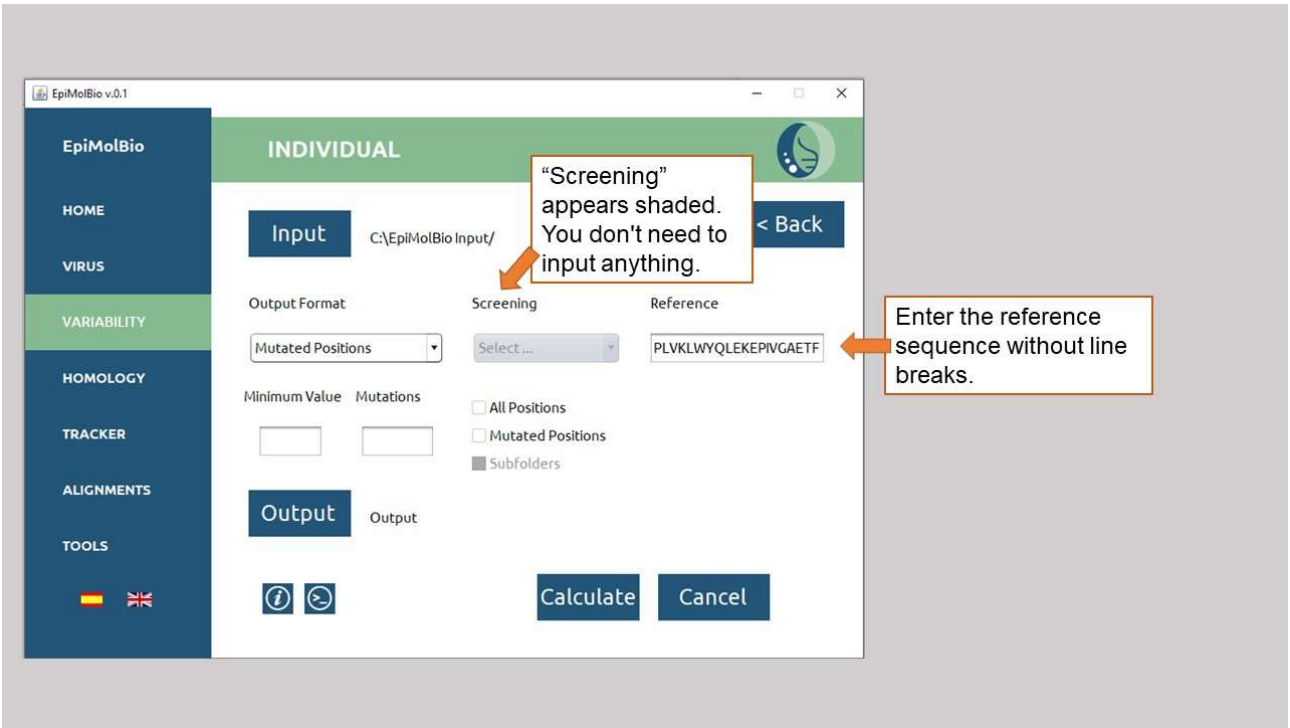
## 4)


EpiMolBio v.0.1
EpiMolBio
HOME
INDIVIDUAL
Input
Input
< Back
Select the input folder.
Program Files
Program Files (x86)
EpiMolBio Input
EpiMolBio Output
It must only contain the .fasta files of the sequences to be analyzed in NT or AA.
It can contain 1 file or several.
E.g.: Sequences divided by HIV variant.
Abrir
Cancelar

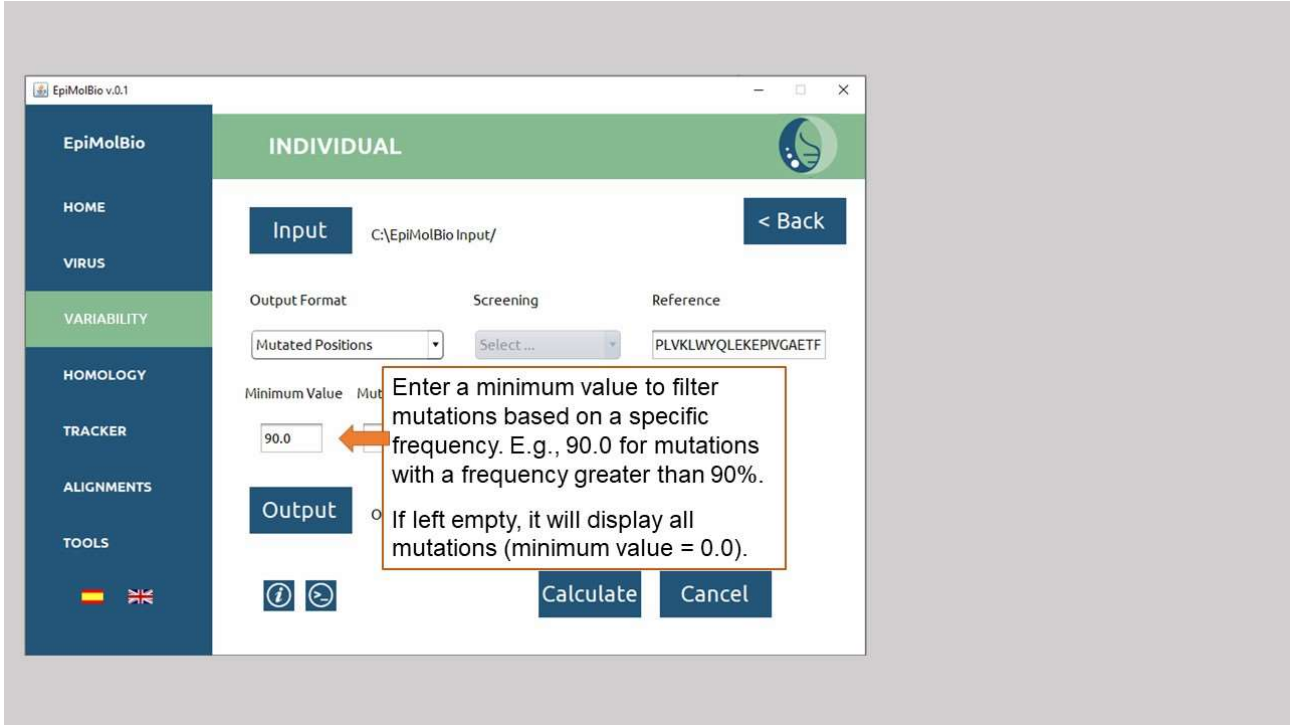
## 5)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
INDIVIDUAL
Input
C:\EpiMolBio Input/
< Back
Output Format
Screening
Reference
Select...
Select...
Mutated Positions
Mutations Table
Markers
Multiple Mutations
Mutations by Position
Select ...
Subfolders
Select Mutated Positions in Output Format
Output
Output
Calculate
Cancel

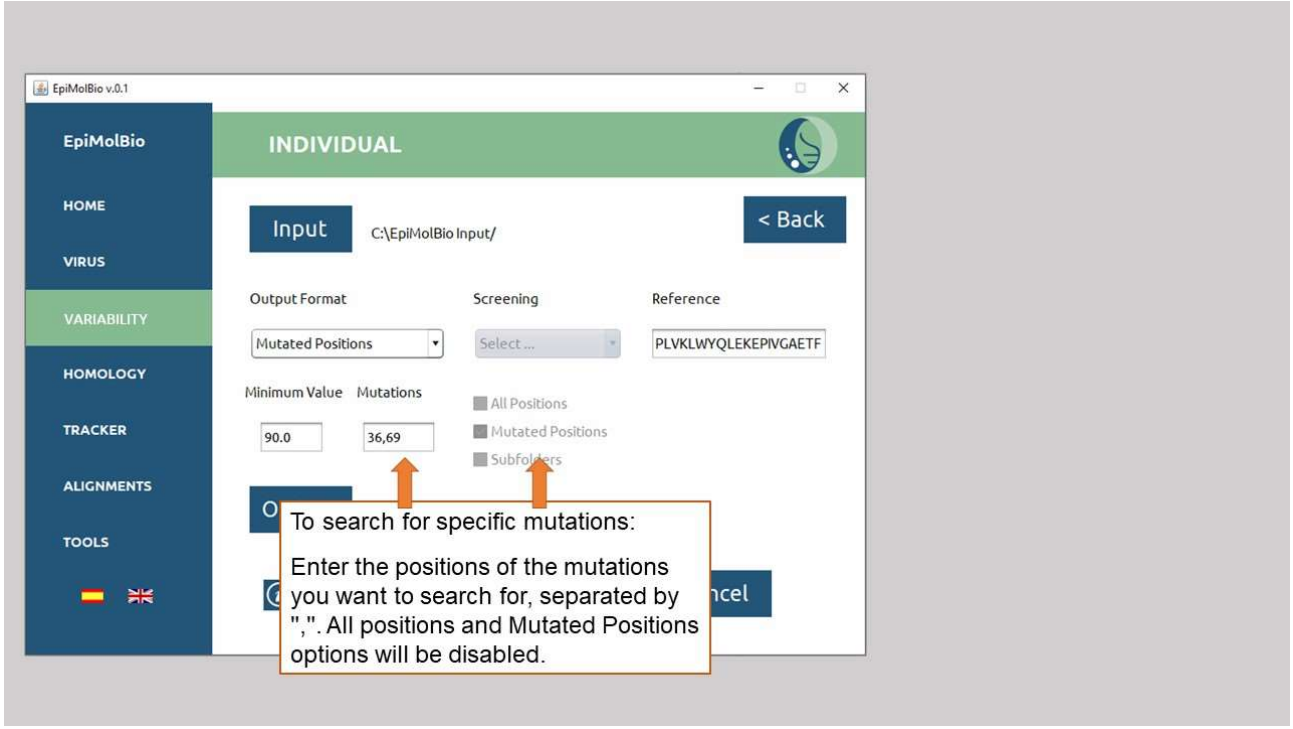
## 6)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
INDIVIDUAL
Input
C:\EpiMolBio Input/
< Back
"Screening" appears shaded. You don't need to input anything.
Output Format
Screening
Reference
Mutated Positions
Select ...
PLVKLWYQLEKEPIVGAETF
Enter the reference sequence without line breaks.
Minimum Value
Mutations
All Positions
Mutated Positions
Subfolders
Output
Output
Calculate
Cancel

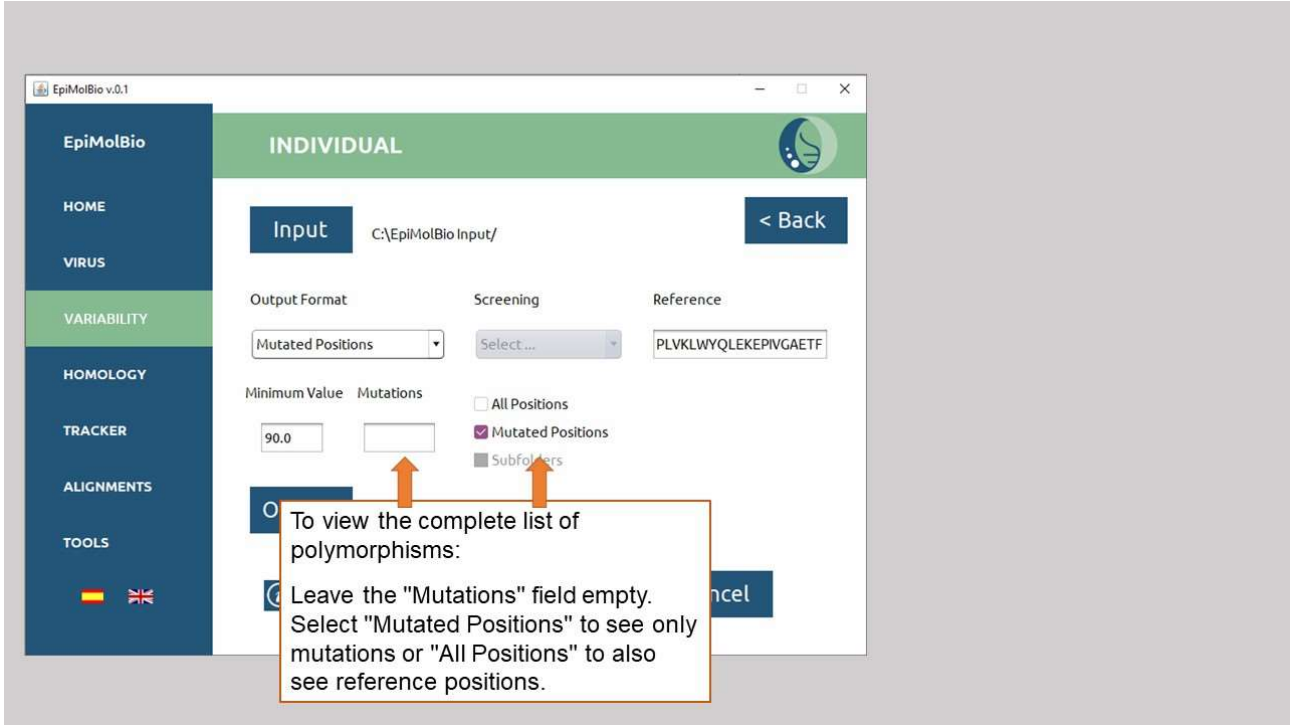
## 7)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
INDIVIDUAL
Input
C:\EpiMolBio Input/
< Back
Output Format
Screening
Reference
Mutated Positions
Select ...
PLVKLWYQLEKEPIVGAETF
Minimum Value
90.0
Output
Calculate
Cancel
Enter a minimum value to filter mutations based on a specific frequency. E.g., 90.0 for mutations with a frequency greater than 90%.
If left empty, it will display all mutations (minimum value = 0.0).

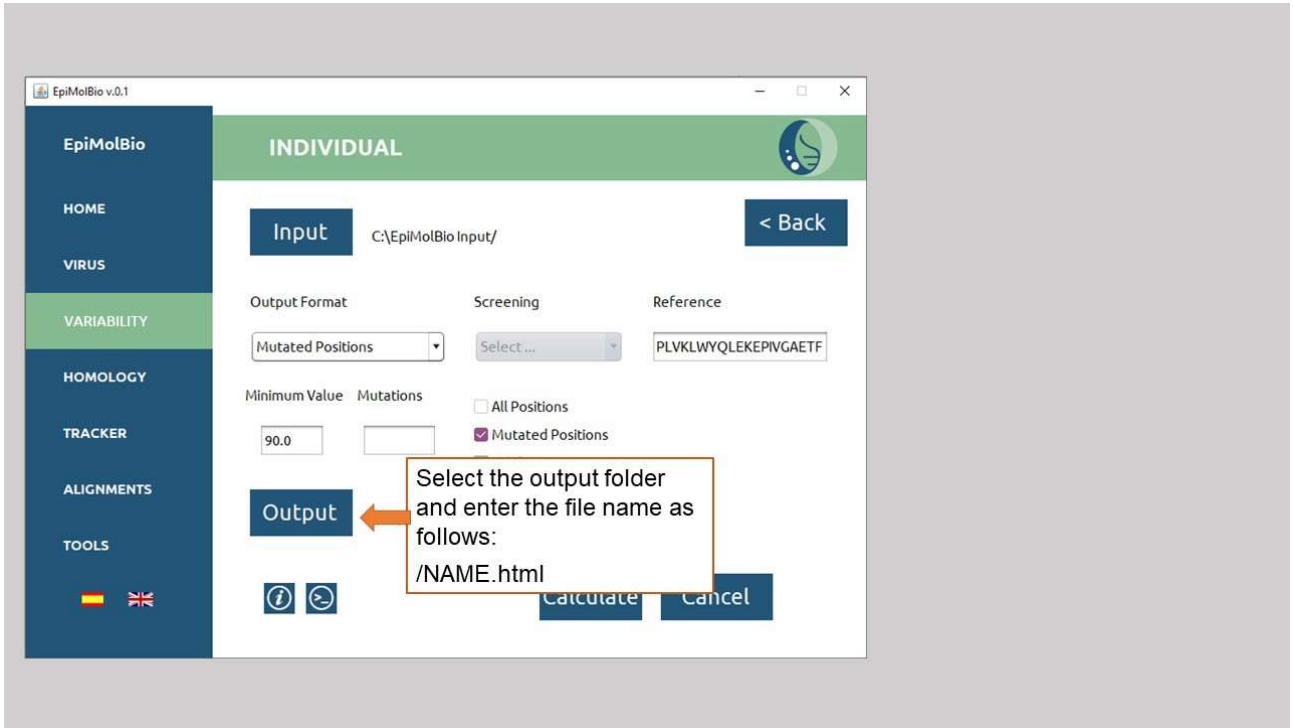
## 8)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
INDIVIDUAL
Input
C:\EpiMolBio Input/
< Back
Output Format
Screening
Reference
Mutated Positions
Select ...
PLVKLWYQLEKEPIVGAETF
Minimum Value
Mutations
90.0
36,69
All Positions
Mutated Positions
To search for specific mutations:
Enter the positions of the mutations you want to search for, separated by ",". All positions and Mutated Positions options will be disabled.

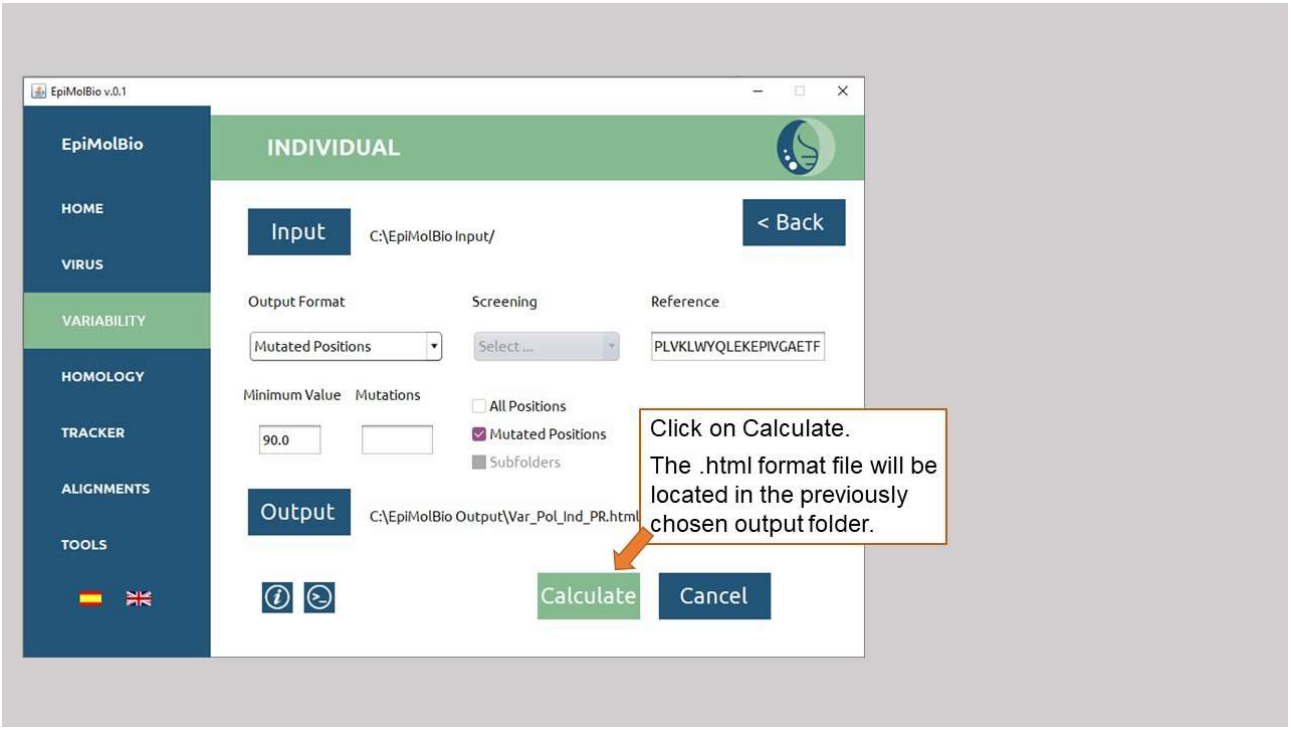
## 9)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
INDIVIDUAL
Input
C:\EpiMolBio Input/
< Back
Output Format
Screening
Reference
Mutated Positions
Select ...
PLVKLWYQLEKEPIVGAETF
Minimum Value
Mutations
90.0
All Positions
Mutated Positions
To view the complete list of polymorphisms:
Leave the "Mutations" field empty. Select "Mutated Positions" to see only mutations or "All Positions" to also see reference positions.

## 10)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
INDIVIDUAL
Input
C:\EpiMolBio Input/
< Back
Output Format
Screening
Reference
Mutated Positions
Select ...
PLVKLWYQLEKEPIVGAETF
Minimum Value
Mutations
90.0
All Positions
Mutated Positions
Output
Select the output folder and enter the file name as follows:
/NAME.html
Cancel

## 11)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
INDIVIDUAL
Input
C:\EpiMolBio Input/
< Back
Output Format
Screening
Reference
Mutated Positions
Select ...
PLVKLWYQLEKEPIVGAETF
Minimum Value
Mutations
90.0
All Positions
Mutated Positions
Subfolders
Output
C:\EpiMolBio Output\Var_Pol_Ind_PR.html
Click on Calculate.
The .html format file will be located in the previously chosen output folder.
Calculate
Cancel

## 2.-Mutations Table:

In a similar manner to 'Mutated Positions,' with this output format, you can **detect, locate, and determine the frequency of occurrence of mutations** with respect to a reference sequence provided by the user. You can also choose to search for specific individual mutations or select the minimum frequency of occurrence for the mutations you want to display. The difference with the previous format is that it allows you to input files in **subfolders**, and the output file is a **.csv** file that can be easily modified using Excel.

Example of Mutations Table output format:

| | A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|---|
| 1 | File | Number of Sequences | Reference | Position | Change | Percentage | Mutated Sequences |
| 2 | PR_01_AE.fasta | 26838 | P | 1 | P | 99.90% | 26810 |
| 3 | PR_01_AE.fasta | 26838 | P | 1 | S | 0.08% | 21 |
| 4 | PR_01_AE.fasta | 26838 | P | 1 | A | 0.00% | 1 |
| 5 | PR_01_AE.fasta | 26838 | P | 1 | L | 0.01% | 2 |
| 6 | PR_01_AE.fasta | 26838 | P | 1 | T | 0.01% | 2 |
| 7 | PR_01_AE.fasta | 26838 | P | 1 | H | 0.00% | 1 |
| 8 | PR_01_AE.fasta | 26838 | P | 1 | V | 0.00% | 1 |
| 9 | PR_01_AE.fasta | 26649 | Q | 2 | Q | 99.78% | 26591 |
| 10 | PR_01_AE.fasta | 26649 | Q | 2 | E | 0.07% | 19 |
| 11 | PR_01_AE.fasta | 26649 | Q | 2 | S | 0.02% | 5 |
| 12 | PR_01_AE.fasta | 26649 | Q | 2 | H | 0.06% | 15 |
| 13 | PR_01_AE.fasta | 26649 | Q | 2 | D | 0.00% | 1 |
| 14 | PR_01_AE.fasta | 26649 | Q | 2 | K | 0.02% | 6 |
| 15 | PR_01_AE.fasta | 26649 | Q | 2 | L | 0.02% | 4 |
| 16 | PR_01_AE.fasta | 26649 | Q | 2 | P | 0.01% | 2 |
| 17 | PR_01_AE.fasta | 26649 | Q | 2 | R | 0.01% | 3 |
| 18 | PR_01_AE.fasta | 26649 | Q | 2 | T | 0.00% | 1 |
| 19 | PR_01_AE.fasta | 26649 | Q | 2 | * | 0.01% | 2 |
| 20 | PR_01_AE.fasta | 26831 | V | 3 | I | 99.86% | 26793 |

The output .csv file is a table that contains the following information: in the first column, the names of the input files are listed. In the second column, the total number of sequences for each input file is shown. In the third column, the reference amino acid or nucleotide is displayed based on the user-provided reference sequence. In the next column, the positions where mutations have been detected are listed. In the following column, the detected amino acid or nucleotide change is shown. In the sixth column, the frequency of occurrence of the detected change is displayed. In the last column, the total number of mutated sequences is shown.

Since it is a .csv file, you can easily manipulate it using Excel to apply filters or merge columns as needed. If the 'All Positions' field is selected, all positions will be shown, whether they are mutated or not, and the percentage of the reference amino acid will also be displayed.

To perform this analysis, in the **input field** select a folder containing exclusively .fasta files or a folder containing subfolders with .fasta files (mark the 'Subfolders' option if using subfolders). The input files can be in nucleotides or amino acids. If the input files are in nucleotides, use the Find and Replace tool in File Editing to replace 'N' with '?' to exclude them from the analysis.

Choose '**Mutation Table**' from the dropdown menu for '**Output Format**.'

In the '**Reference**' field, input the reference sequence in letters without line breaks. Ensure that the sequence is entered in nucleotides or amino acids based on whether the input files are untranslated or translated.

The '**Minimum Value**' field can be left blank if you want the result to show all detected mutations regardless of their frequency of occurrence. To filter mutations based on a specific frequency, input the minimum value in decimal numeric format (e.g., 75.0 to show only mutations that occur at a percentage greater than 75%).

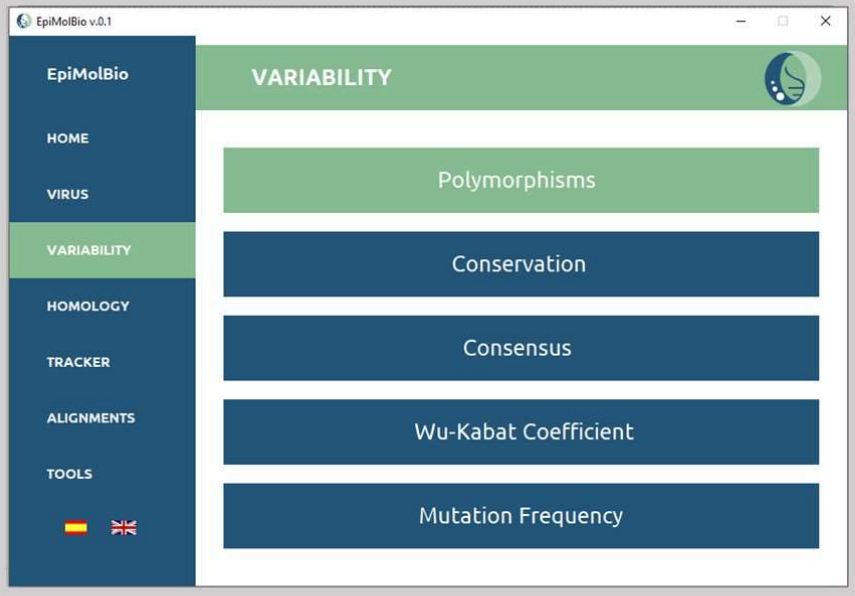
The '**Mutations**' field can be left blank if you want the result to show all detected mutations. To search for specific mutations, input the mutation(s) by typing the reference, position, and mutated residue (e.g., Q2H). If the input files and the reference are in nucleotides, input the mutation with the reference and the mutation in nucleotides (e.g., A6C). To search for multiple mutations, separate them with a comma ',' without spaces.

You can choose to view the information for all positions or only for mutated positions by selecting the '**All Positions**' or '**Mutated Positions**' checkbox, respectively. If you have input something in the 'Mutations' field, these options will be disabled.

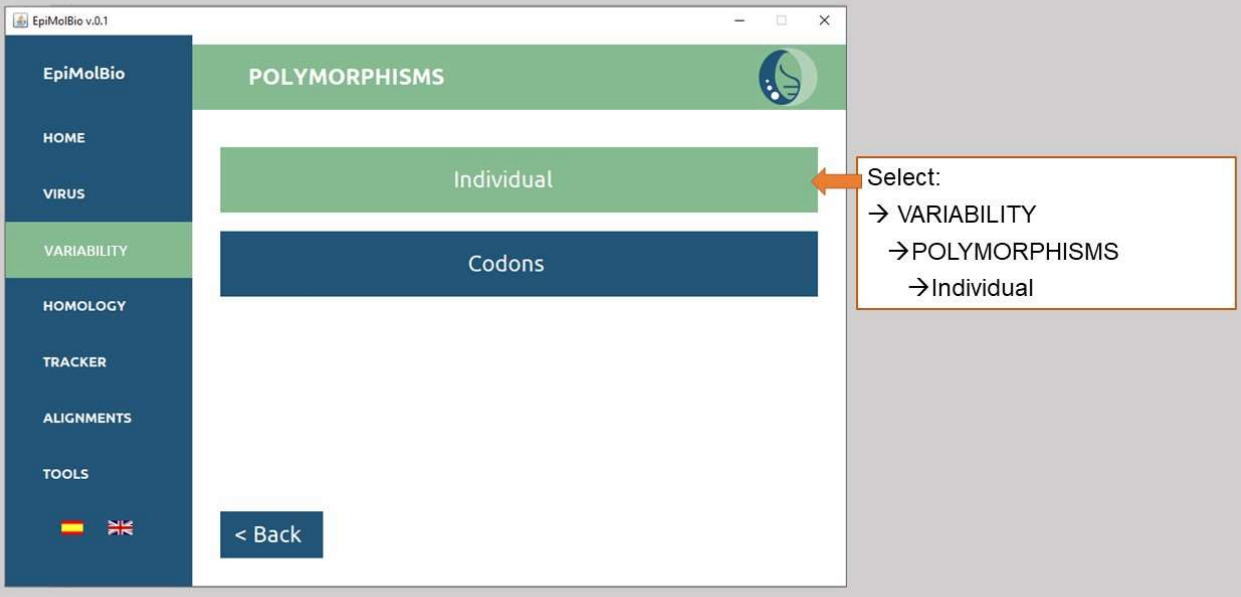
The result is displayed in a .csv file. In the '**Output**' field, select the folder where you want to save the result and name the file with the .csv extension.

# Step-by-step:

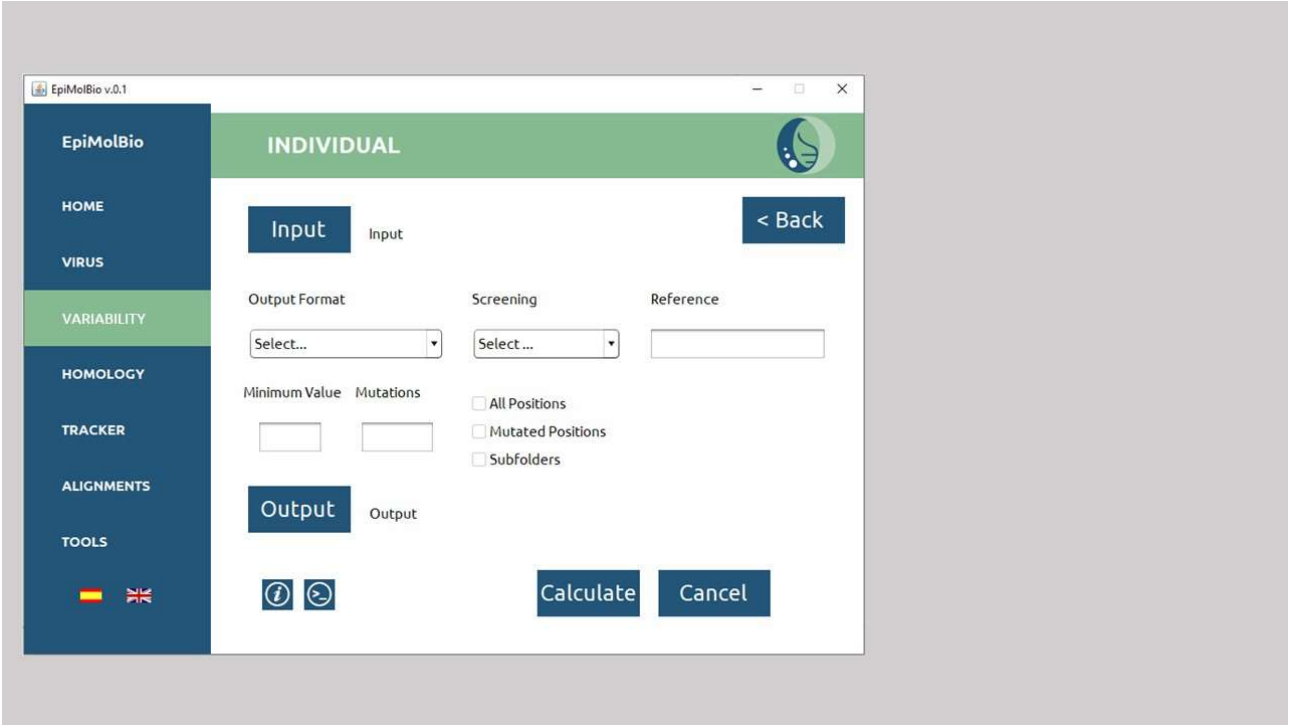
## 1)



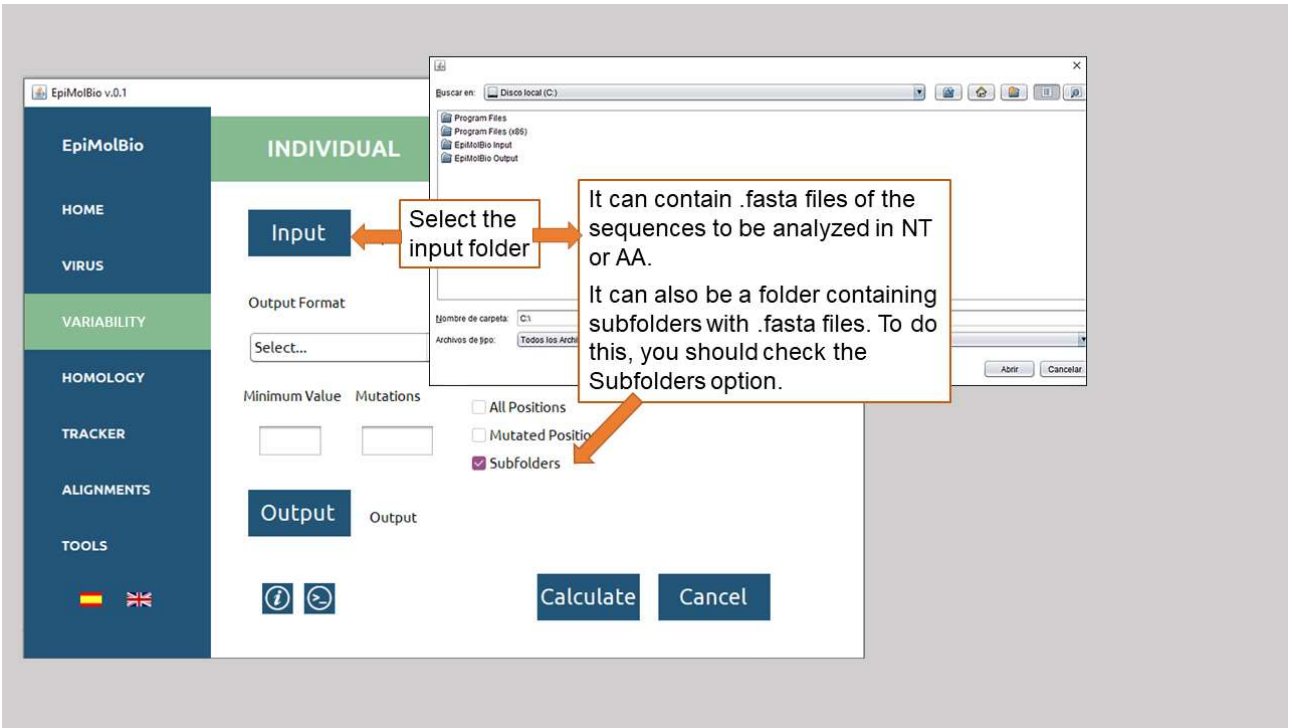

## 2)



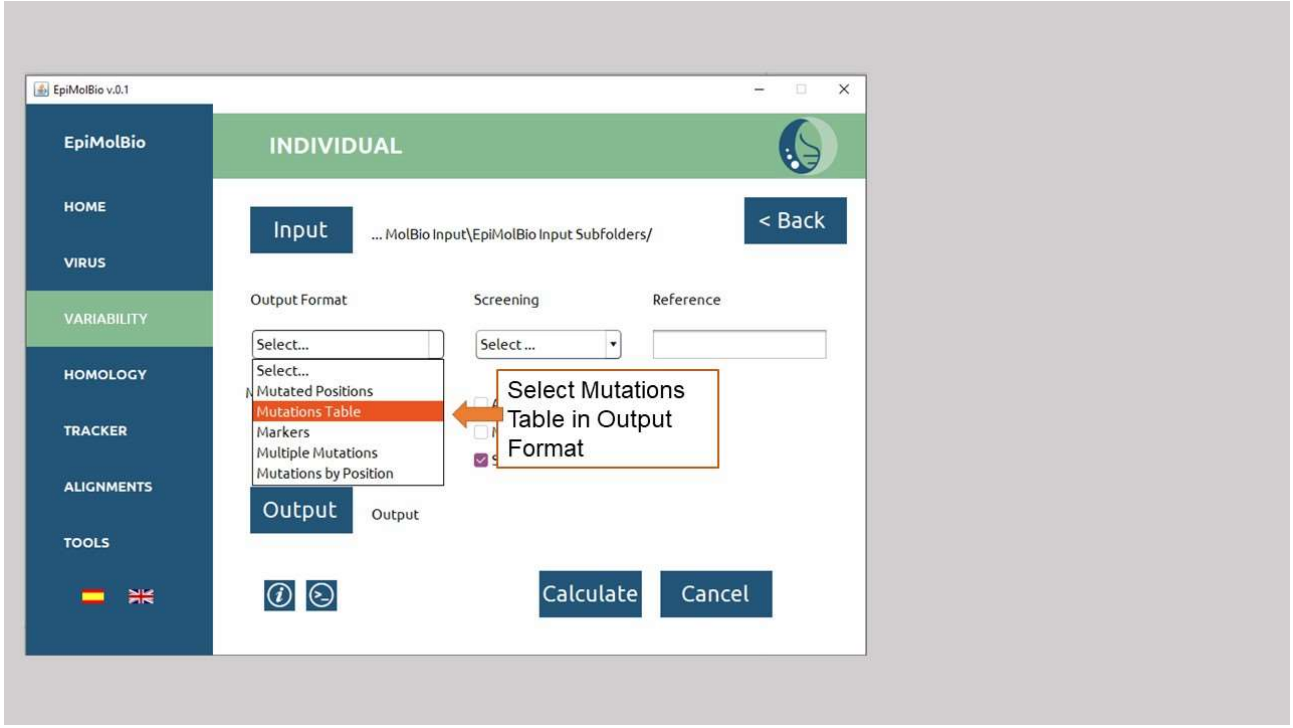
## 3)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
INDIVIDUAL
Input
Input
< Back
Output Format
Screening
Reference
Select...
Select ...
Minimum Value
Mutations
All Positions
Mutated Positions
Subfolders
Output
Output
Calculate
Cancel

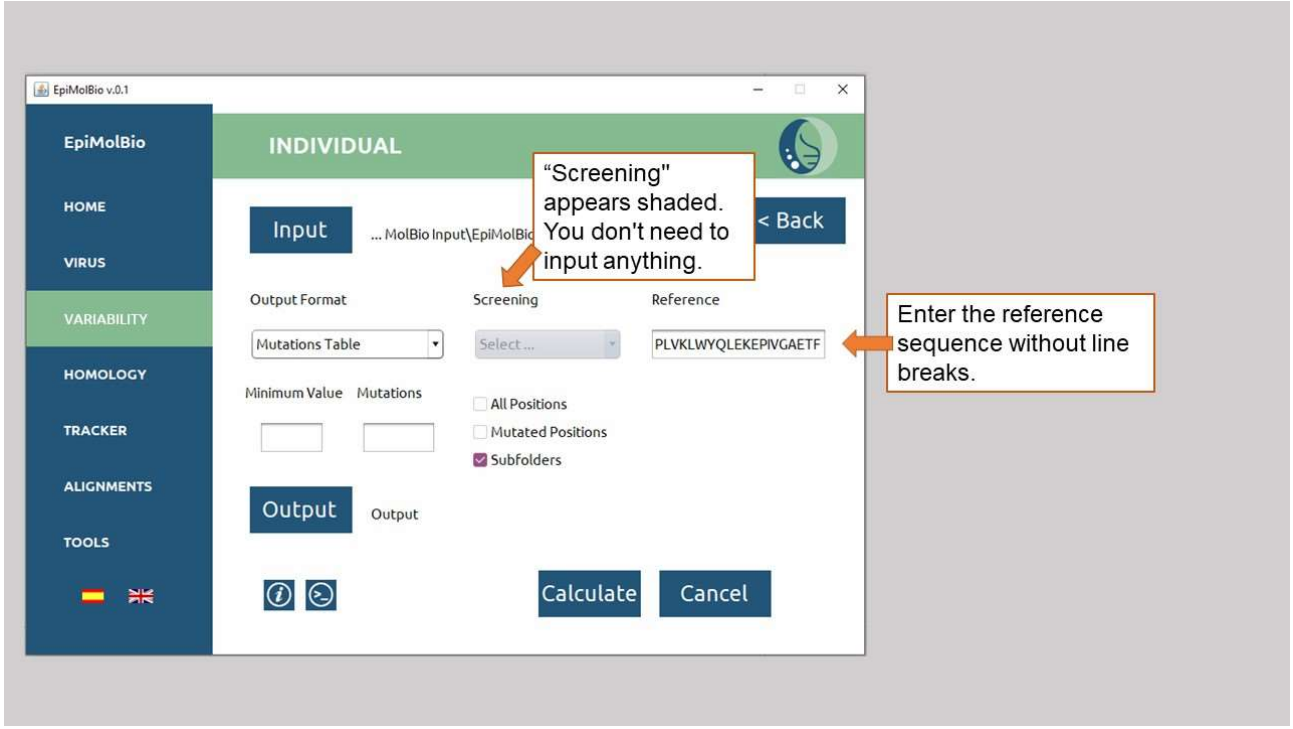
## 4)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
INDIVIDUAL
Input
Select the input folder
It can contain .fasta files of the sequences to be analyzed in NT or AA.
It can also be a folder containing subfolders with .fasta files. To do this, you should check the Subfolders option.
Output Format
Select...
Minimum Value
Mutations
All Positions
Subfolders
Output
Output
Calculate
Cancel

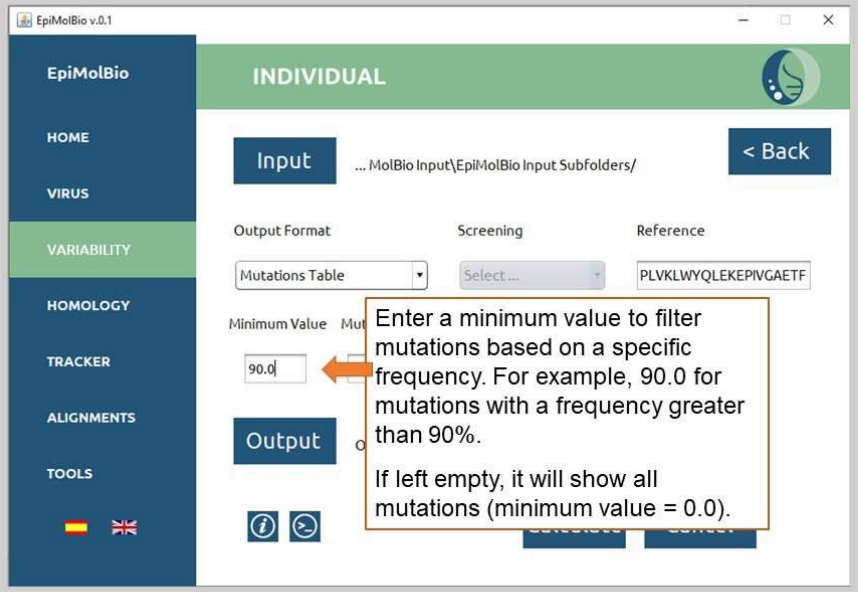
## 5)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
INDIVIDUAL
Input
... MolBio Input\EpiMolBio Input Subfolders/
< Back
Output Format
Screening
Reference
Select...
Select ...
Select...
Mutated Positions
Mutations Table
Markers
Multiple Mutations
Mutations by Position
Select Mutations Table in Output Format
Output
Output
Calculate
Cancel

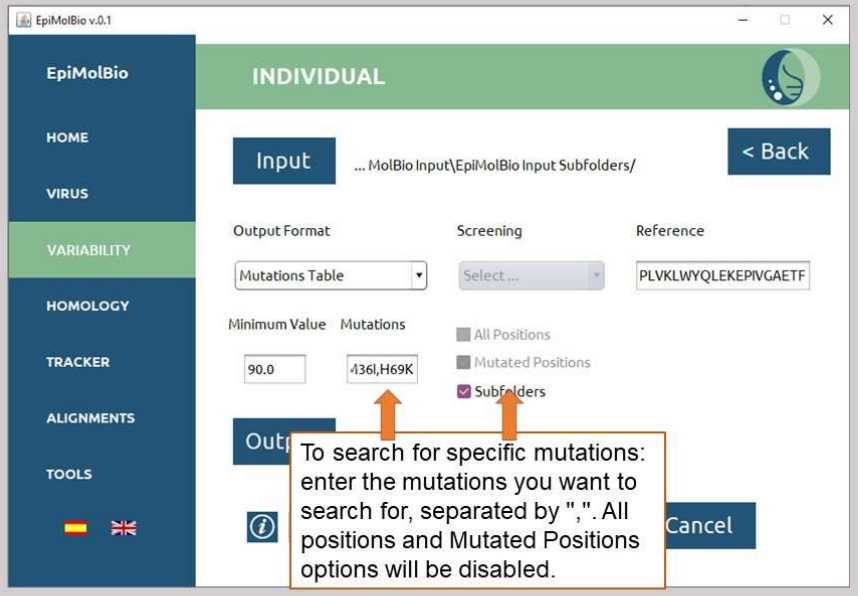
## 6)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
INDIVIDUAL
Input
... MolBio Input\EpiMolBio
< Back
"Screening" appears shaded. You don't need to input anything.
Output Format
Screening
Reference
Mutations Table
Select ...
PLVKLWYQLEKEPIVGAETF
Enter the reference sequence without line breaks.
Minimum Value
Mutations
All Positions
Mutated Positions
Subfolders
Output
Output
Calculate
Cancel

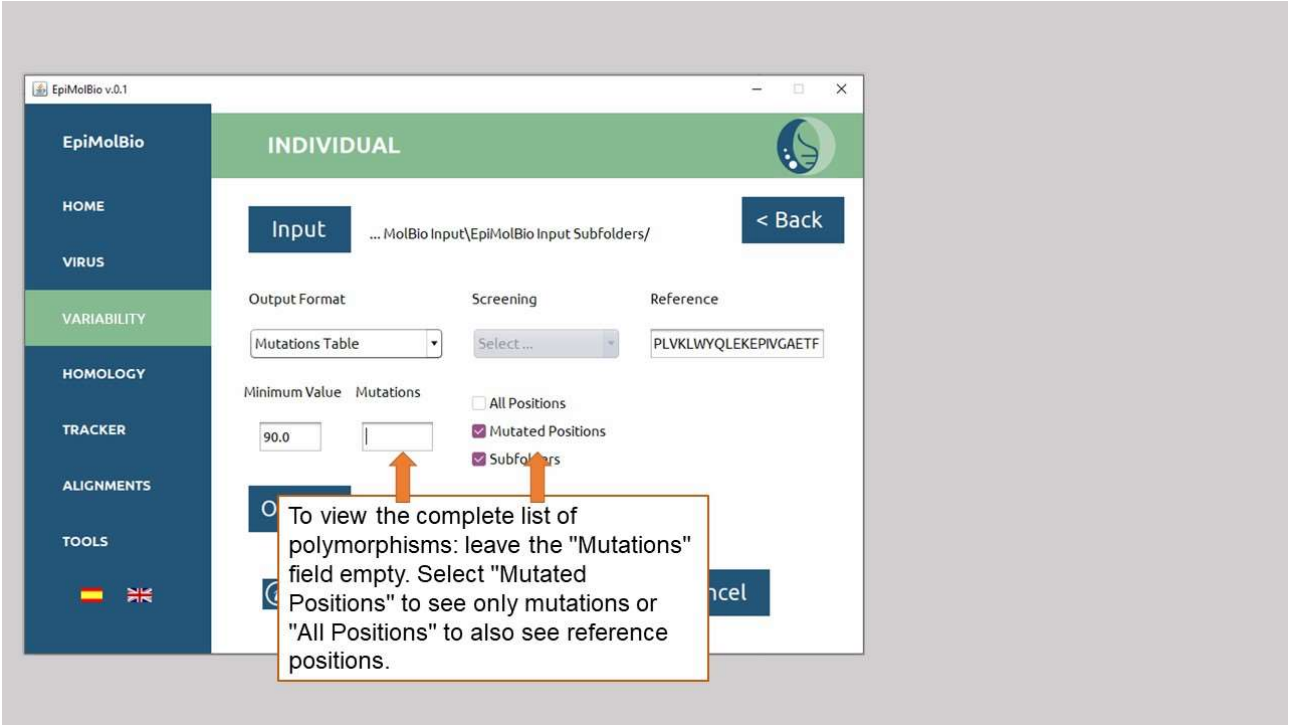
## 7)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
INDIVIDUAL
Input
... MolBio Input\EpiMolBio Input Subfolders/
< Back
Output Format
Screening
Reference
Mutations Table
Select ...
PLVKLWYQLEKEPIVGAETF
Minimum Value
90.0
Output
Enter a minimum value to filter mutations based on a specific frequency. For example, 90.0 for mutations with a frequency greater than 90%.
If left empty, it will show all mutations (minimum value = 0.0).

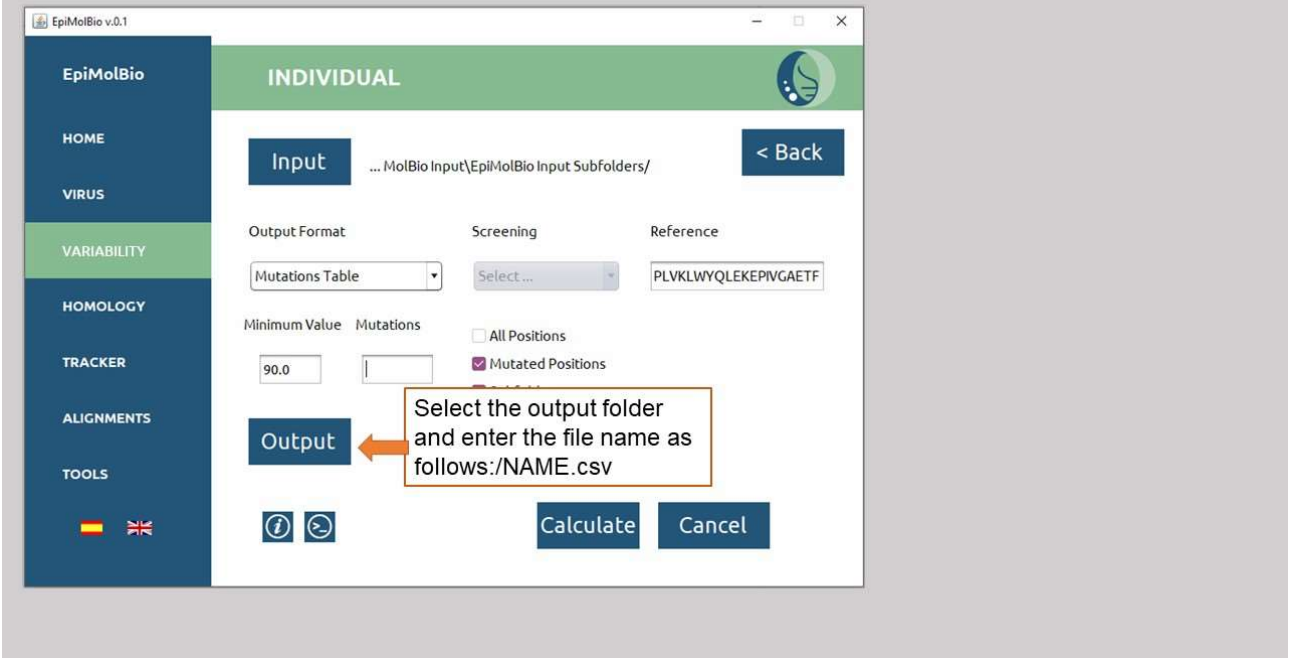
## 8)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
INDIVIDUAL
Input
... MolBio Input\EpiMolBio Input Subfolders/
< Back
Output Format
Screening
Reference
Mutations Table
Select ...
PLVKLWYQLEKEPIVGAETF
Minimum Value
Mutations
90.0
All Positions
Mutated Positions
Subfolders
Cancel
To search for specific mutations: enter the mutations you want to search for, separated by ",". All positions and Mutated Positions options will be disabled.

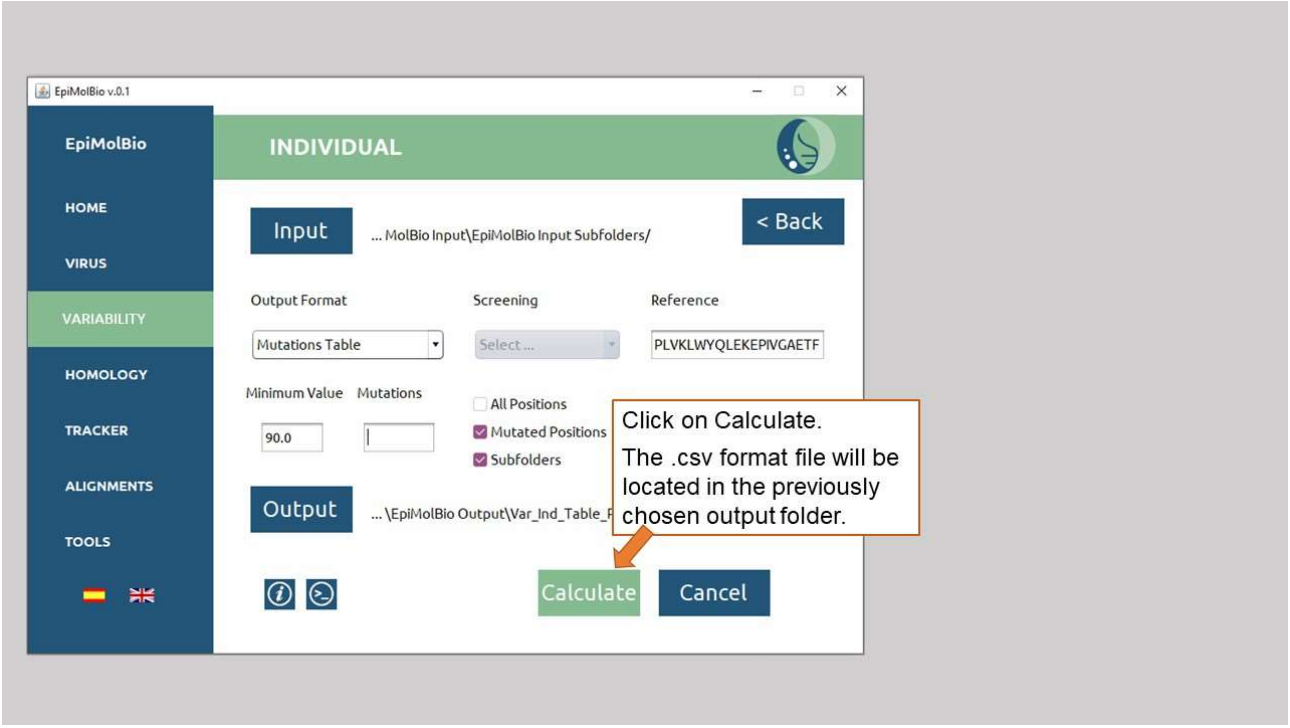
**9)**


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
INDIVIDUAL
Input
... MolBio Input\EpiMolBio Input Subfolders/
< Back
Output Format
Screening
Reference
Mutations Table
Select ...
PLVKLWYQLEKEPIVGAETF
Minimum Value
Mutations
90.0
All Positions
Mutated Positions
To view the complete list of polymorphisms: leave the "Mutations" field empty. Select "Mutated Positions" to see only mutations or "All Positions" to also see reference positions.

**10)**


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
INDIVIDUAL
Input
... MolBio Input\EpiMolBio Input Subfolders/
< Back
Output Format
Screening
Reference
Mutations Table
Select ...
PLVKLWYQLEKEPIVGAETF
Minimum Value
Mutations
90.0
All Positions
Mutated Positions
Output
Select the output folder and enter the file name as follows:/NAME.csv
Calculate
Cancel

## 11)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
INDIVIDUAL
Input
... MolBio Input\EpiMolBio Input Subfolders/
< Back
Output Format
Screening
Reference
Mutations Table
Select ...
PLVKLWYQLEKEPIVGAETF
Minimum Value
Mutations
90.0
All Positions
Mutated Positions
Subfolders
Output
... \EpiMolBio Output\Var_Ind_Table_F
Click on Calculate.
The .csv format file will be located in the previously chosen output folder.
Calculate
Cancel

## 3.- Markers:

This function allows **detecting exclusive mutations** present in each input file compared to the rest of the files provided as input. You can set whether you want the mutations to be present at a frequency greater than 75% or 90%. For instance, if the input files are divided by viral variants, you can identify mutations characteristic of each variant using this function. These characteristic mutations are referred to as 'markers.' Both gaps (-) and question marks (?) are excluded from the analysis.

In the output file, at the top, you will find the title of the analysis. Below that, in the 'File' column, the name of the analyzed input file is shown. In the second column, 'Markers,' the markers detected according to the selected screening percentage are displayed. In the third column, 'Total Sequences,' the total number of sequences analyzed is shown. If markers are found for a file, the reference amino acid is indicated, followed by the position, and the mutation is colored according to the color code described in the Overview, which can be consulted in the .html output file by clicking on the blue symbol.

Example of Markers output format:

| Variability Polymorphisms Individual Markers >= 90% | | |
|---|---|---|
| File | Markers | Total Sequences |
| PR_107_01B.fasta | Q92K | 4 |
| PR_108_BC.fasta | T74S | 15 |
| PR_112_01B.fasta | L63M | 5 |
| PR_118_BC.fasta | K70T | 13 |
| PR_11_cpx.fasta | G16A | 380 |
| PR_129_56G.fasta | K14D, E65K | 1 |

To perform this analysis, select a **folder** containing exclusively .fasta files of the sequences to be analyzed in amino acids as **input**.

Choose '**Markers**' from the dropdown menu for '**Output Format**.'

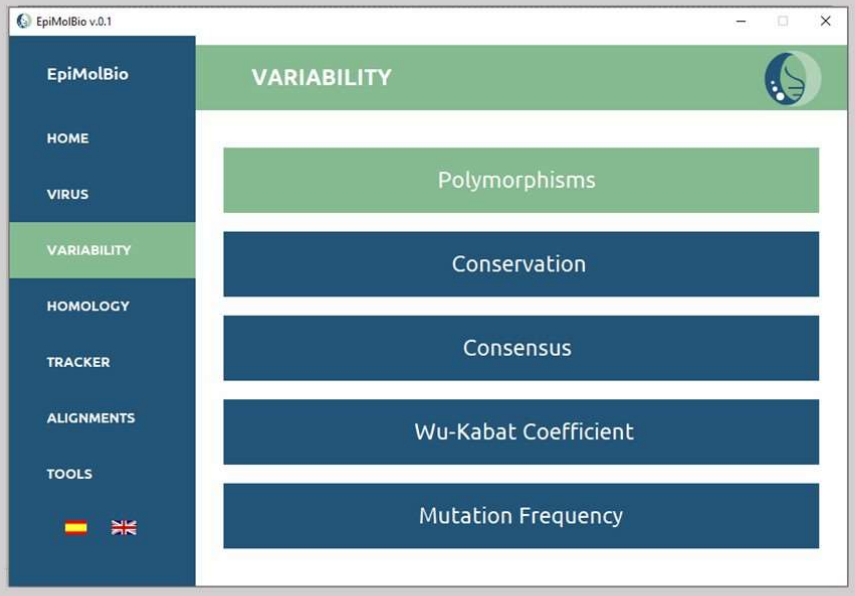
In the '**Screening**' field, select whether you want the detected markers to be present at a frequency greater than 75% (Show > 75%) or 90% (Show >= 90%).

In the '**Reference**' field, input the reference sequence in letters without line breaks.

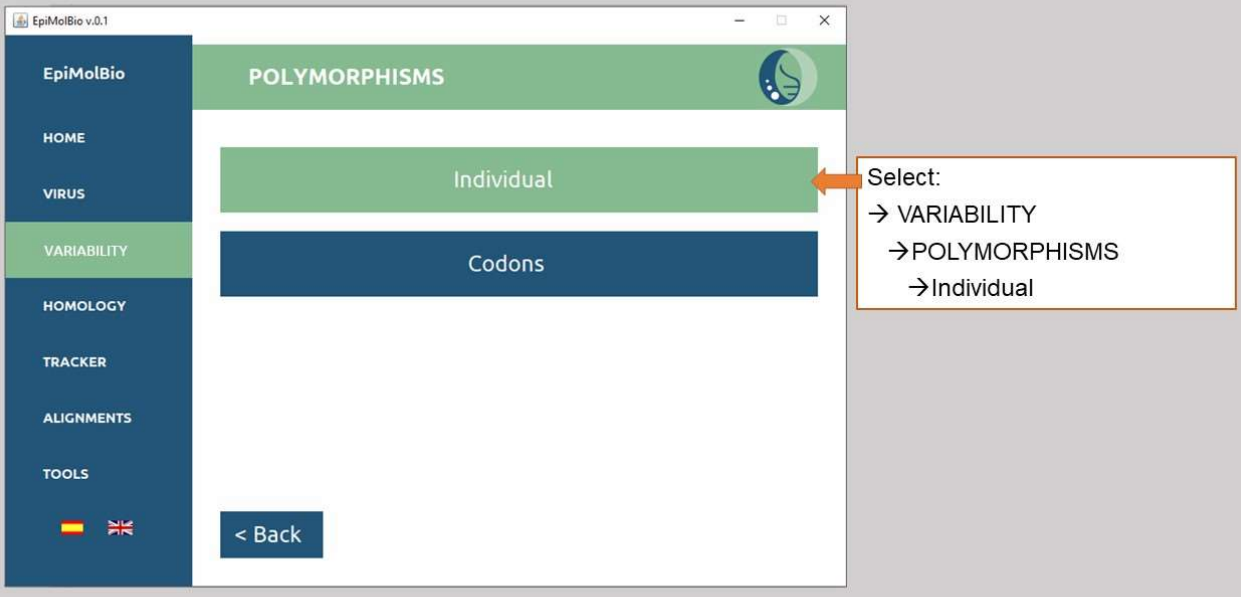
The result is displayed in an .html file. In the '**Output**' field, select the folder where you want to save the result and name the file with the .html extension.

# Step-by-step:

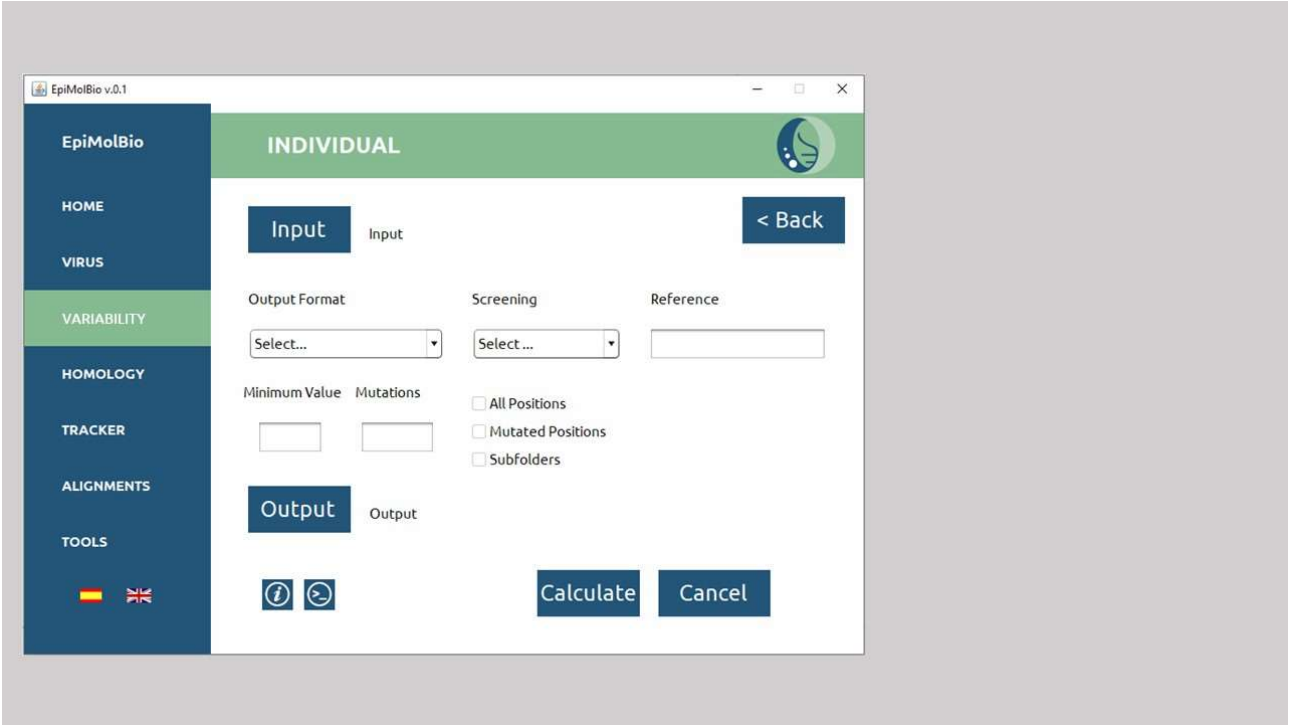
## 1)



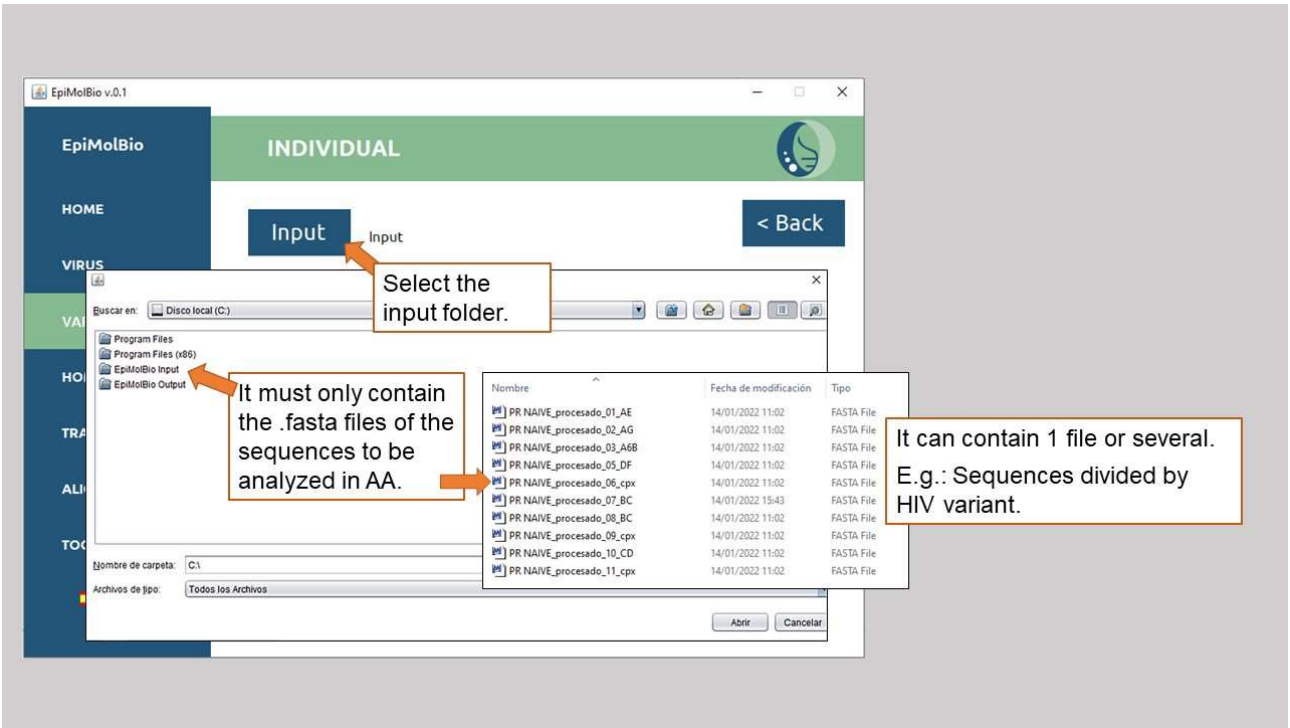

## 2)



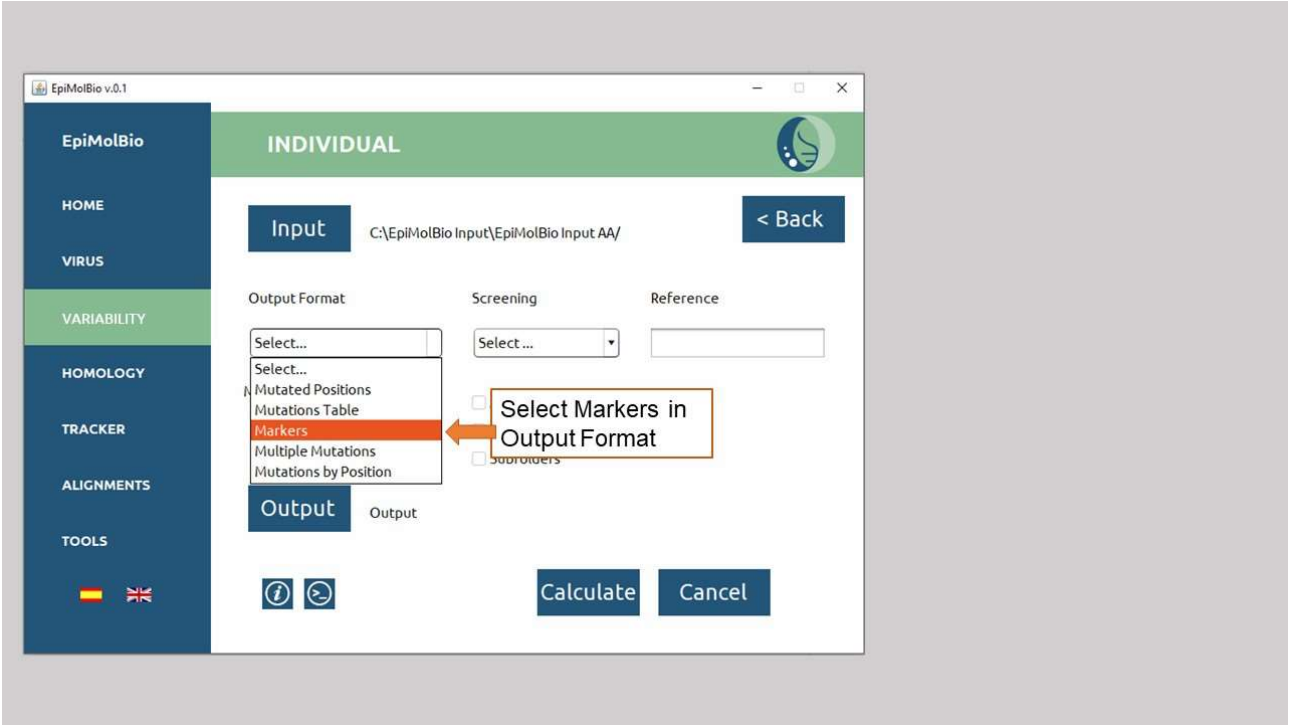
## 3)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
INDIVIDUAL
Input
Input
< Back
Output Format
Screening
Reference
Select...
Select ...
Minimum Value
Mutations
All Positions
Mutated Positions
Subfolders
Output
Output
Calculate
Cancel

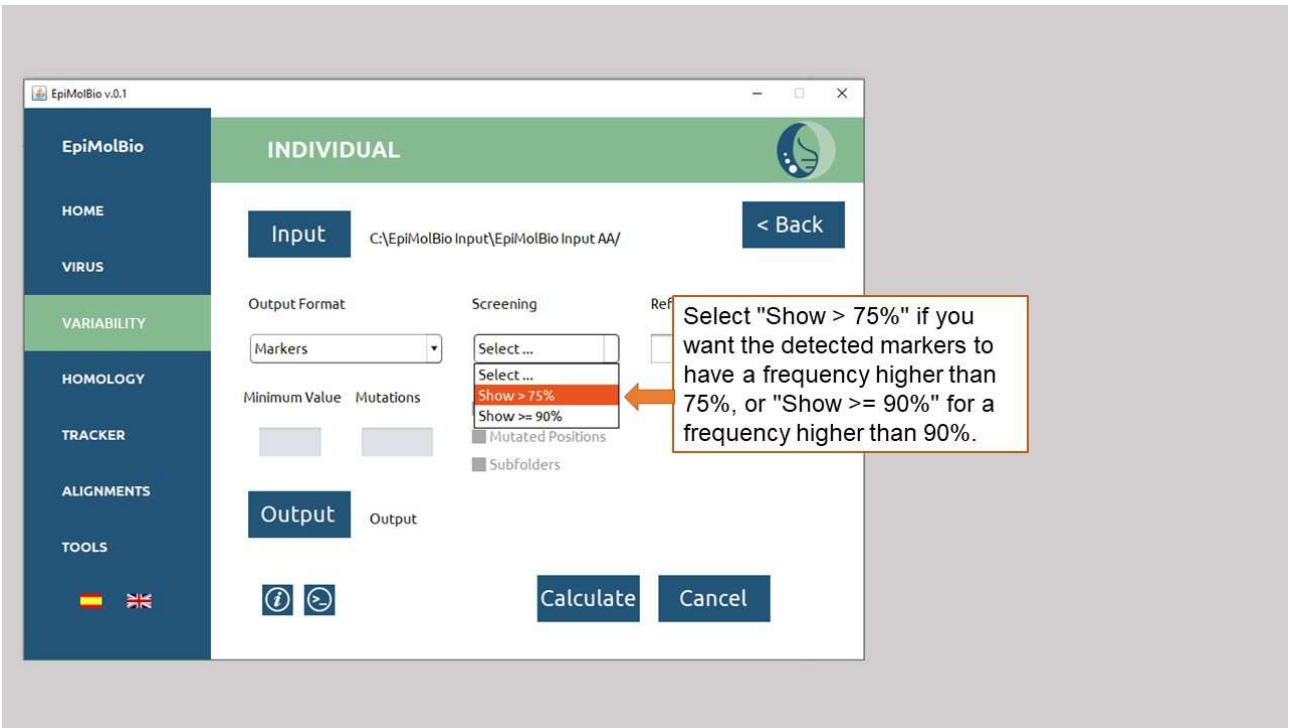
## 4)


EpiMolBio v.0.1
EpiMolBio
HOME
INDIVIDUAL
Input
Input
< Back
Select the input folder.
Buscar en:
Disco local (C:)
Program Files
Program Files (x86)
EpiMolBio Input
EpiMolBio Output
It must only contain the .fasta files of the sequences to be analyzed in AA.
Nombre
Fecha de modificación
Tipo
PR NAIVE_procesado_01_AE
PR NAIVE_procesado_02_AG
PR NAIVE_procesado_03_A6B
PR NAIVE_procesado_05_DF
PR NAIVE_procesado_06_cpx
PR NAIVE_procesado_07_BC
PR NAIVE_procesado_08_BC
PR NAIVE_procesado_09_cpx
PR NAIVE_procesado_10_CD
PR NAIVE_procesado_11_cpx
FASTA File
It can contain 1 file or several.
E.g.: Sequences divided by HIV variant.
Nombre de carpeta:
C:\
Archivos de tipo:
Todos los Archivos
Abrir
Cancelar

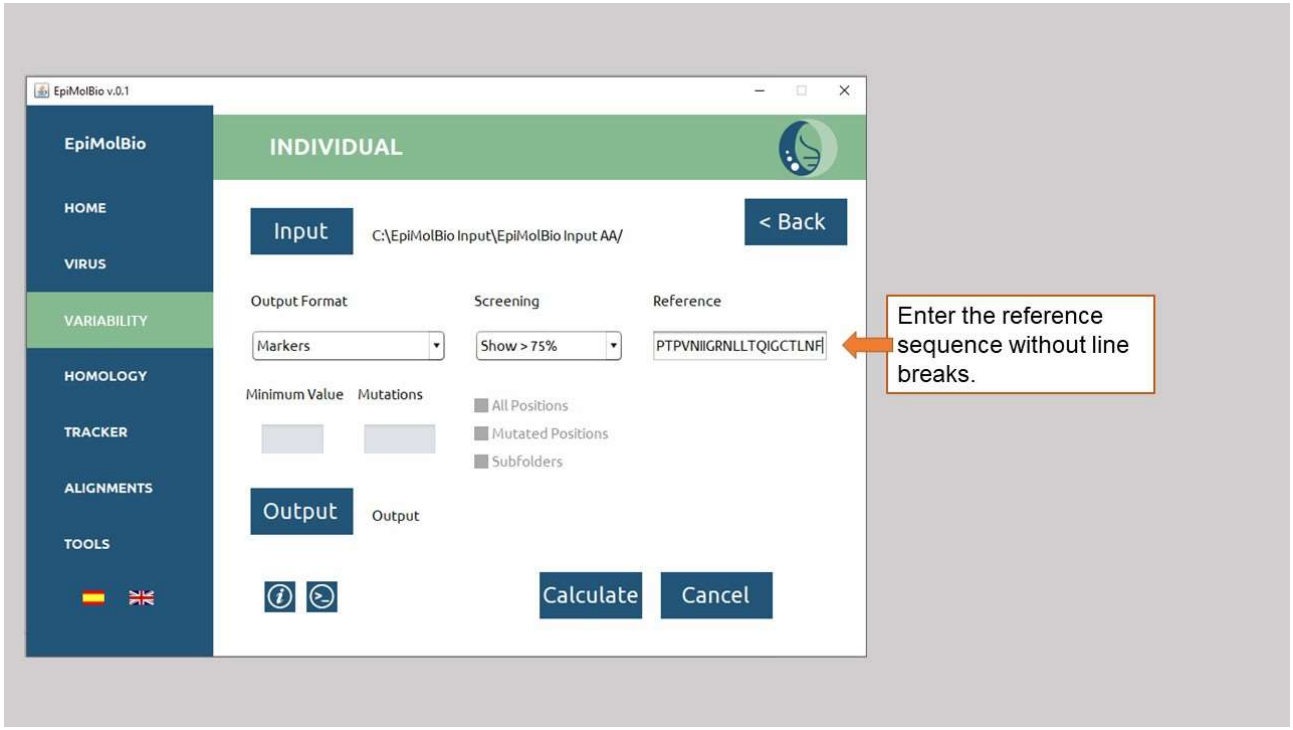
## 5)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
INDIVIDUAL
Input
C:\EpiMolBio Input\EpiMolBio Input AA/
< Back
Output Format
Screening
Reference
Select...
Select ...
Select...
Mutated Positions
Mutations Table
Markers
Multiple Mutations
Mutations by Position
Select Markers in Output Format
Output
Output
Calculate
Cancel

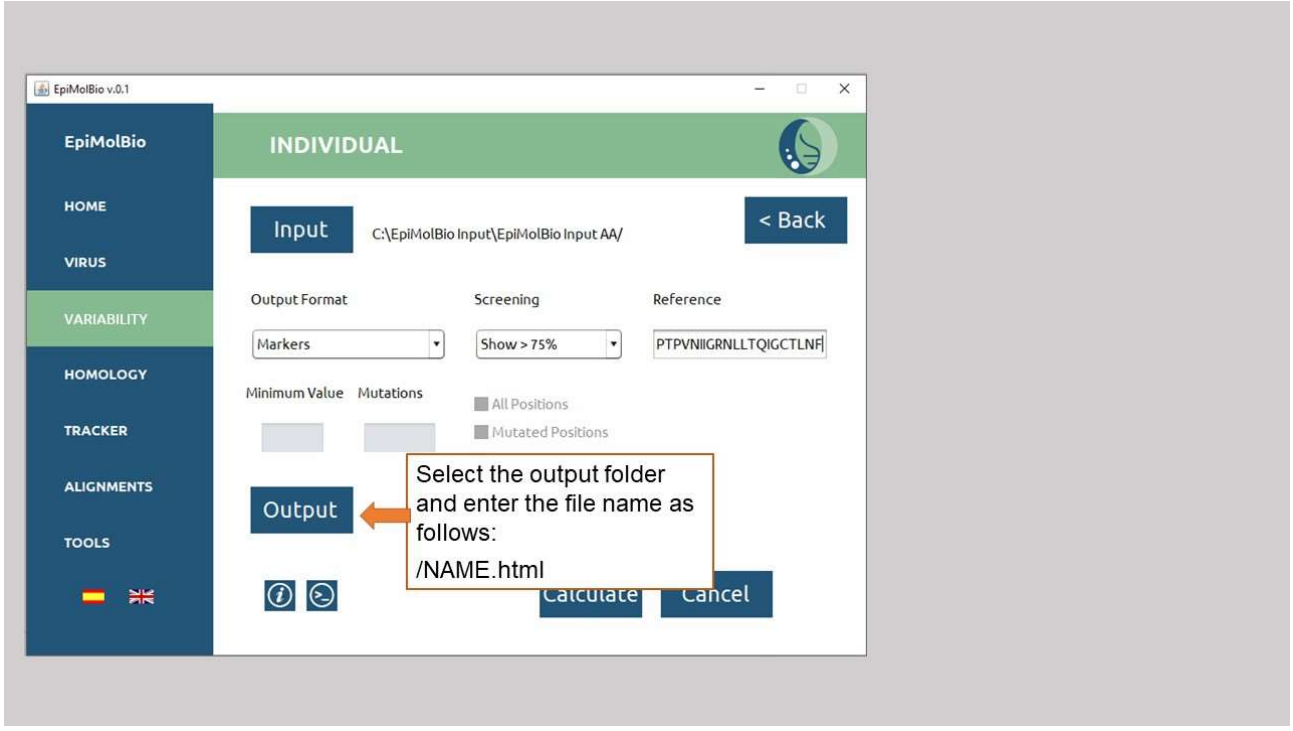
## 6)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
INDIVIDUAL
Input
C:\EpiMolBio Input\EpiMolBio Input AA/
< Back
Output Format
Screening
Markers
Select ...
Select ...
Show > 75%
Show >= 90%
Minimum Value
Mutations
Mutated Positions
Subfolders
Select "Show > 75%" if you want the detected markers to have a frequency higher than 75%, or "Show >= 90%" for a frequency higher than 90%.
Output
Output
Calculate
Cancel

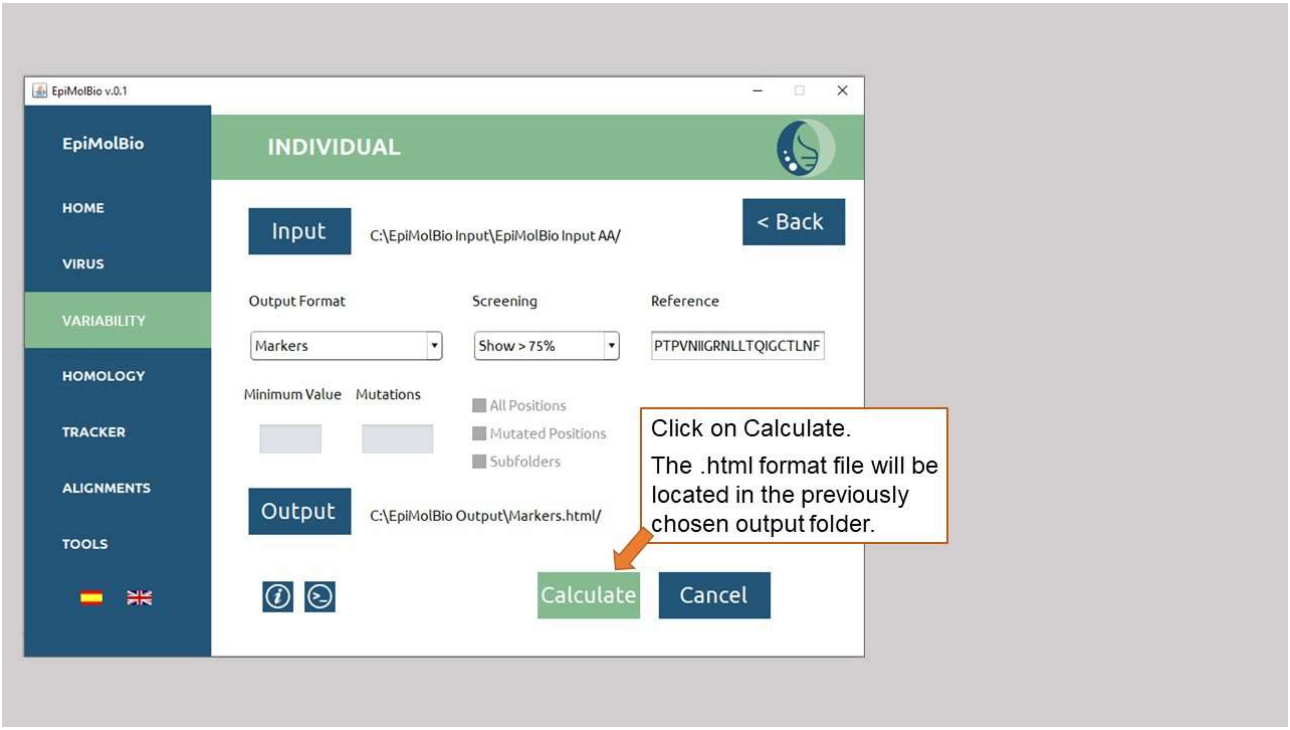
## 7)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
INDIVIDUAL
Input
C:\EpiMolBio Input\EpiMolBio Input AA/
< Back
Output Format
Screening
Reference
Markers
Show > 75%
PTPVNIIGRNLLTQIGCTLNF
Minimum Value
Mutations
All Positions
Mutated Positions
Subfolders
Output
Output
Calculate
Cancel
Enter the reference sequence without line breaks.

## 8)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
INDIVIDUAL
Input
C:\EpiMolBio Input\EpiMolBio Input AA/
< Back
Output Format
Screening
Reference
Markers
Show > 75%
PTPVNIIGRNLLTQIGCTLNF
Minimum Value
Mutations
All Positions
Mutated Positions
Output
Select the output folder and enter the file name as follows:
/NAME.html
Calculate
Cancel

**9)**


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
INDIVIDUAL
Input
C:\EpiMolBio Input\EpiMolBio Input AA/
< Back
Output Format
Screening
Reference
Markers
Show > 75%
PTPVNIIGRNLLTQIGCTLNF
Minimum Value
Mutations
All Positions
Mutated Positions
Subfolders
Output
C:\EpiMolBio Output\Markers.html/
Click on Calculate.
The .html format file will be located in the previously chosen output folder.
Calculate
Cancel

## 4.-Multiple Mutations:

This tool enables the **detection and analysis of the frequency of occurrence of mutation combinations**, which need to be entered in the 'Mutations' field. It also allows for the input of files organized in subfolders.

The output is a .csv file with the following columns: the first column contains the names of the input files. The second column displays the total number of valid sequences per file. In the third column, the number of times the combined mutations appear is recorded. The fourth column shows the frequency of occurrence of that specific combination of mutations.

This analysis supports both amino acid and nucleotide combinations. Gaps (-) and question marks (?) are excluded from the analysis.

Example of Multiple Mutations output format:

| | A | B | C | D |
|---|---|---|---|---|
| 1 | File | Number of Sequences | M46I/V32I | Frequency |
| 2 | PR_01_AE.fasta | 26504 | 2 | 0.008 |
| 3 | PR_02_AG.fasta | 9418 | 0 | 0 |
| 4 | PR_03_A6B.fasta | 300 | 1 | 0.333 |
| 5 | PR_04_cpx.fasta | 15 | 0 | 0 |
| 6 | PR_05_DF.fasta | 24 | 0 | 0 |
| 7 | PR_06_cpx.fasta | 732 | 0 | 0 |
| 8 | PR_07_BC.fasta | 10819 | 0 | 0 |
| 9 | PR_08_BC.fasta | 2326 | 0 | 0 |
| 10 | PR_09_cpx.fasta | 91 | 0 | 0 |
| 11 | PR_100_01C.fasta | 5 | 0 | 0 |
| 12 | PR_101_01B.fasta | 4 | 0 | 0 |

To perform this analysis, select as **input** a folder containing exclusively .fasta files in amino acids or nucleotides, or a folder containing subfolders with .fasta files (for this, mark the 'Subfolders' option). Both gaps (-) and question marks (?) will be excluded from the analysis.

If you want to conduct the analysis in nucleotides, you can use the Find and Replace function in File Editing within Tools to change 'N' to '?', as the Multiple Mutations tool does not automatically exclude 'N' from the analysis.

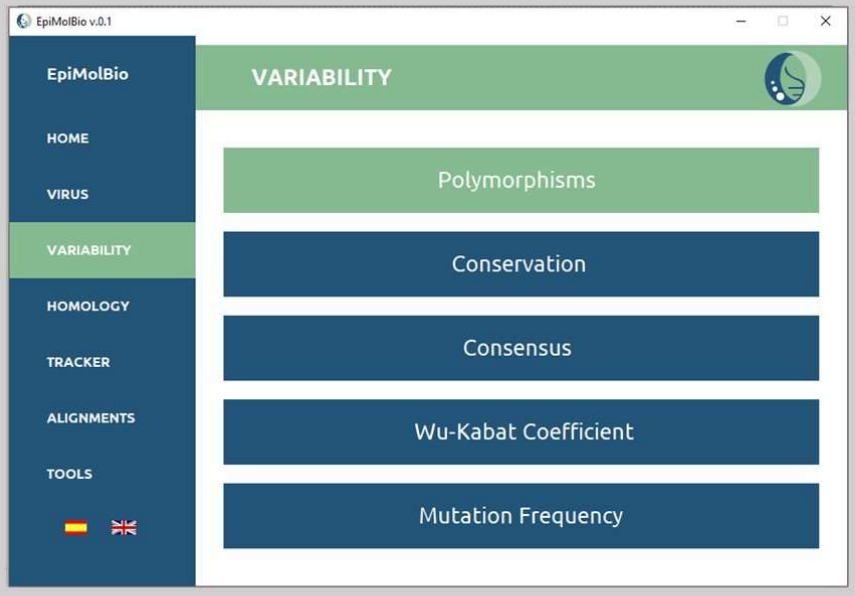
Choose '**Multiple Mutations**' from the dropdown menu for '**Output Format**.'

In the '**Mutations**' field, enter the combination of mutations you want to search for by typing the reference, position, and mutated residue of each mutation, separated by a comma ',' without spaces (e.g., D614G,A222V). If the input files and the reference are in nucleotides, input the mutation with the reference and the mutation in nucleotides.

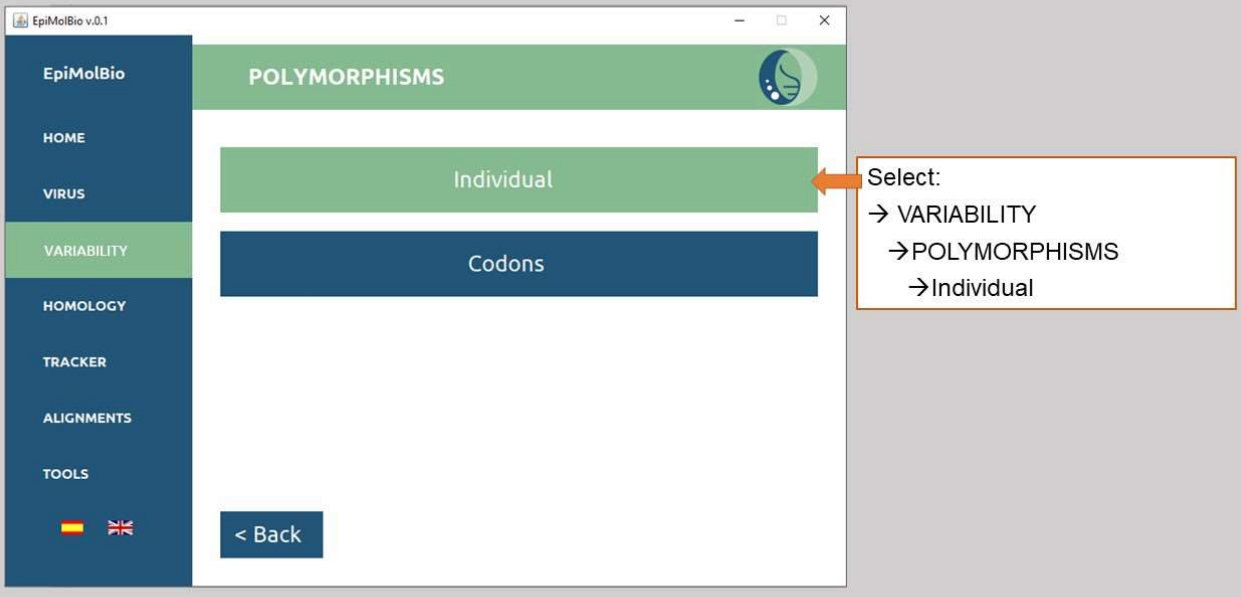
The result will be displayed in a .csv file. In the '**Output**' field, select the folder where you want to save the result and name the file with the .csv extension.

# Step-by-step:

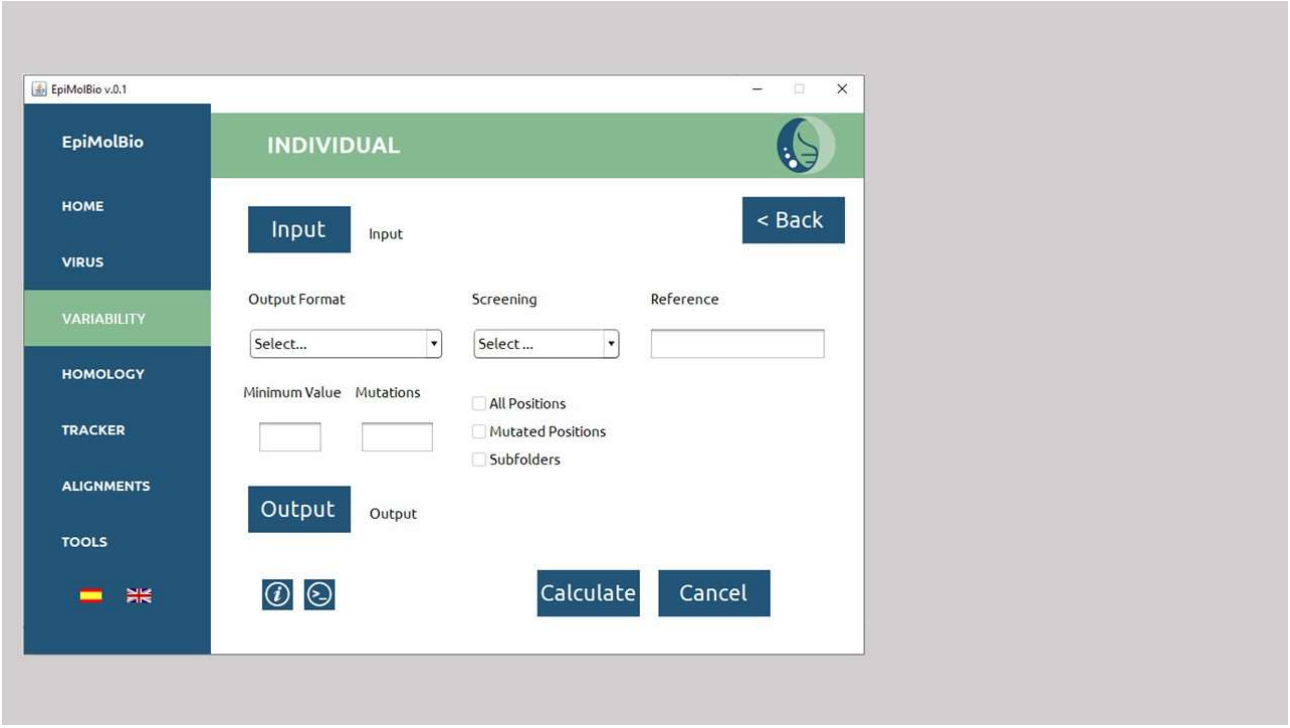
## 1)



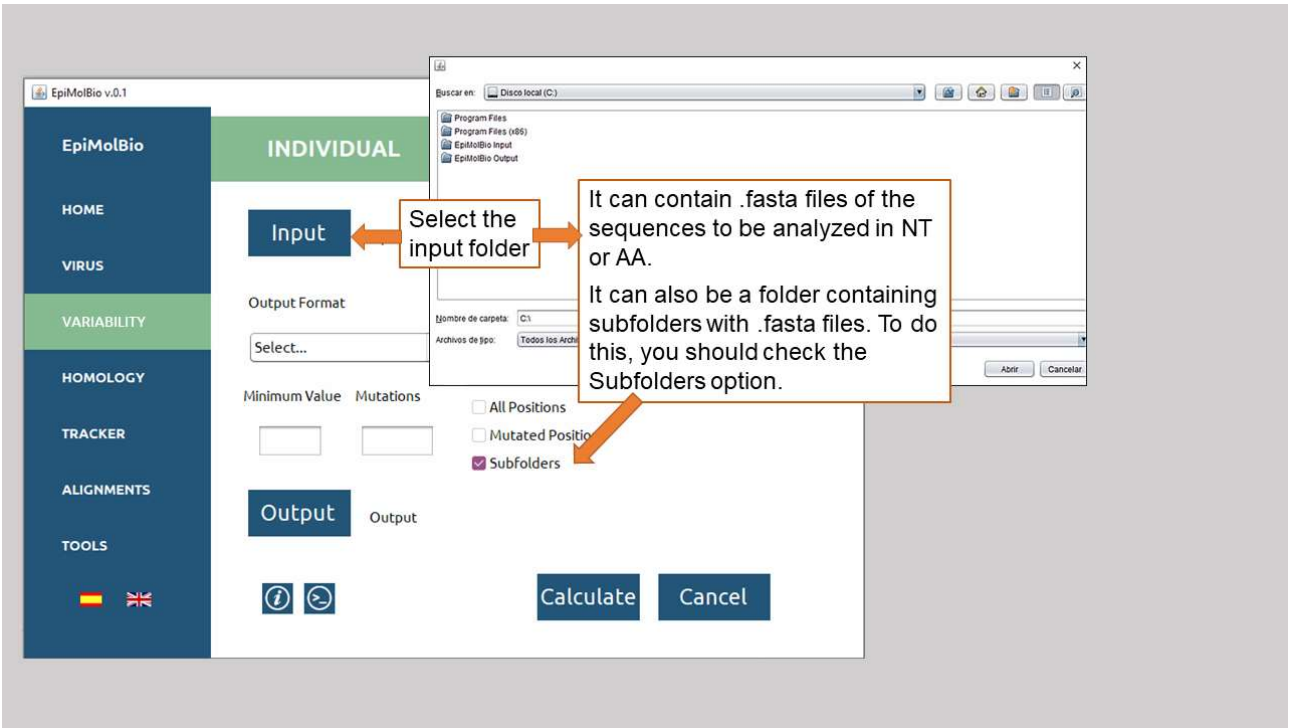

## 2)



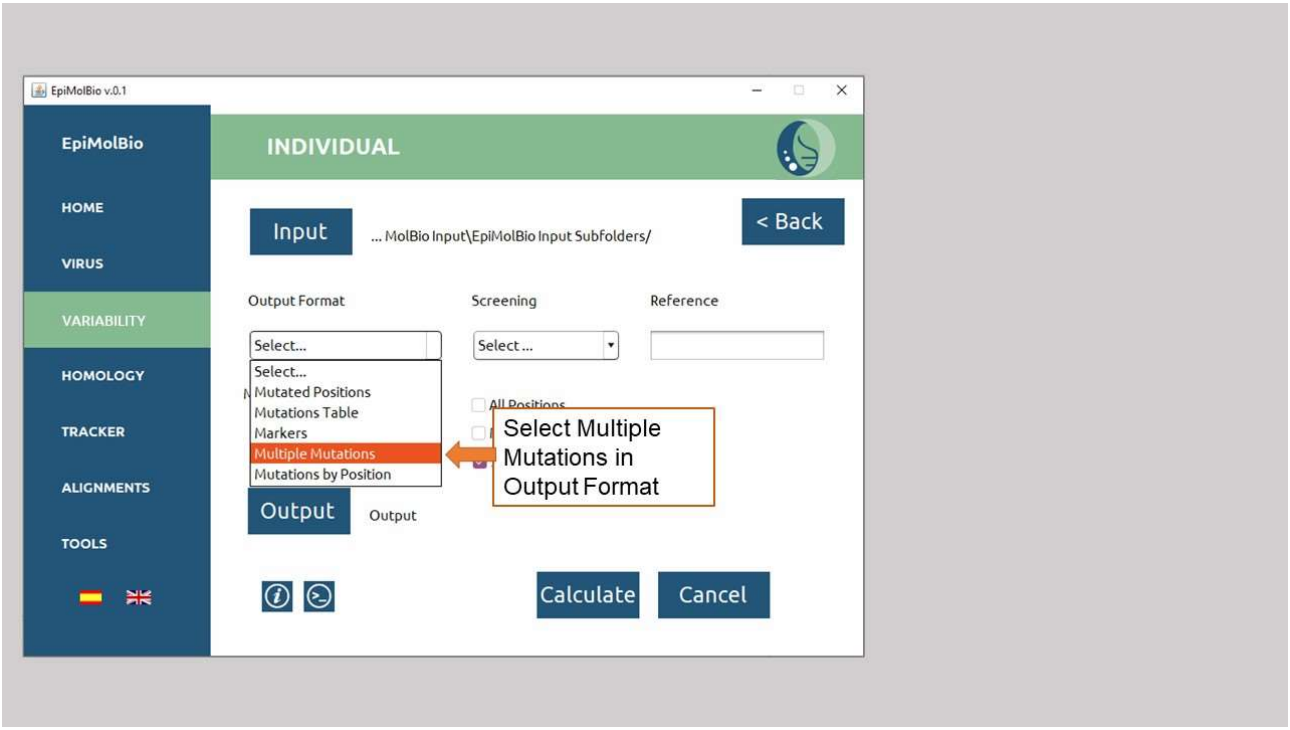
## 3)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
INDIVIDUAL
Input
Input
< Back
Output Format
Screening
Reference
Select...
Select ...
Minimum Value
Mutations
All Positions
Mutated Positions
Subfolders
Output
Output
Calculate
Cancel

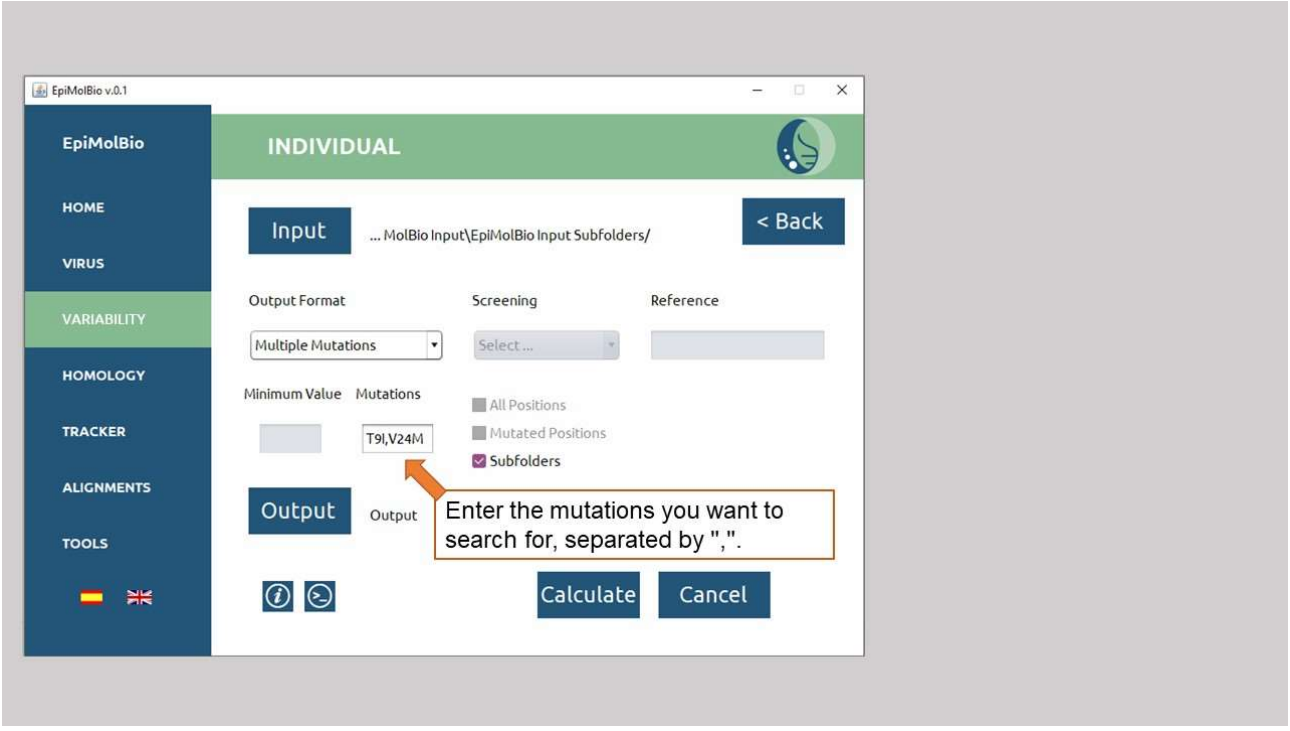
## 4)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
INDIVIDUAL
Input
Select the input folder
It can contain .fasta files of the sequences to be analyzed in NT or AA.
It can also be a folder containing subfolders with .fasta files. To do this, you should check the Subfolders option.
Output Format
Select...
Minimum Value
Mutations
All Positions
Subfolders
Output
Output
Calculate
Cancel

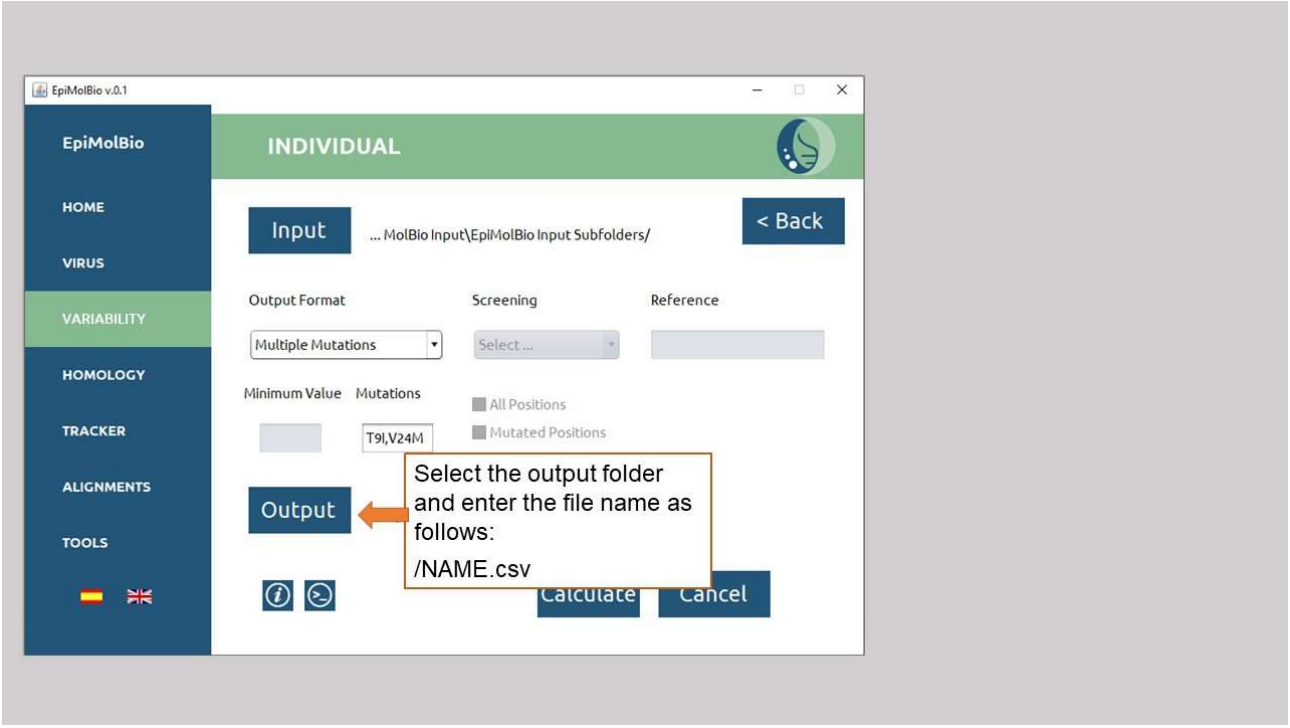
## 5)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
INDIVIDUAL
Input
... MolBio Input\EpiMolBio Input Subfolders/
< Back
Output Format
Screening
Reference
Select...
Select ...
Select...
Mutated Positions
Mutations Table
Markers
Multiple Mutations
Mutations by Position
Select Multiple Mutations in Output Format
Output
Output
Calculate
Cancel

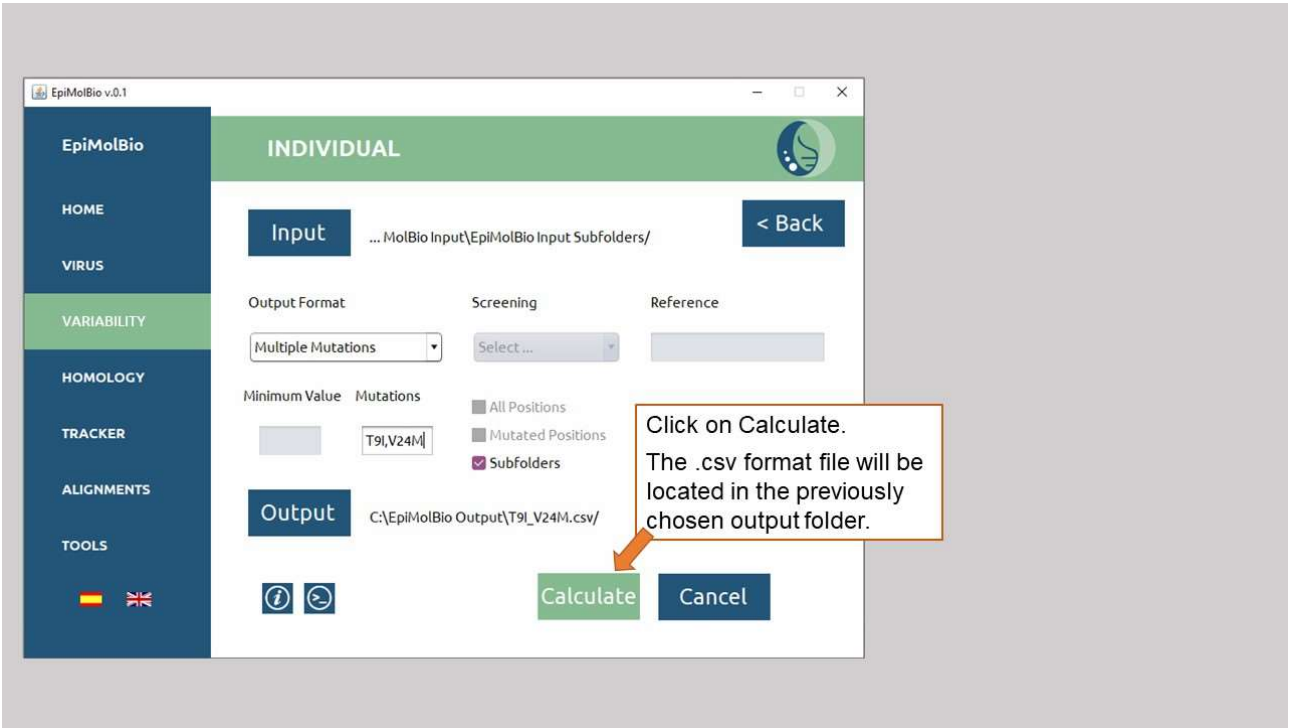
## 6)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
INDIVIDUAL
Input
... MolBio Input\EpiMolBio Input Subfolders/
< Back
Output Format
Screening
Reference
Multiple Mutations
Select ...
Minimum Value
Mutations
T9I,V24M
All Positions
Mutated Positions
Subfolders
Output
Output
Enter the mutations you want to search for, separated by ",".
Calculate
Cancel

## 7)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
INDIVIDUAL
Input
... MolBio Input\EpiMolBio Input Subfolders/
< Back
Output Format
Screening
Reference
Multiple Mutations
Select ...
Minimum Value
Mutations
T9I,V24M
All Positions
Mutated Positions
Output
Select the output folder and enter the file name as follows:
/NAME.csv
Calculate
Cancel

## 8)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
INDIVIDUAL
Input
... MolBio Input\EpiMolBio Input Subfolders/
< Back
Output Format
Screening
Reference
Multiple Mutations
Select ...
Minimum Value
Mutations
T9I,V24M
All Positions
Mutated Positions
Subfolders
Output
C:\EpiMolBio Output\T9I_V24M.csv/
Click on Calculate.
The .csv format file will be located in the previously chosen output folder.
Calculate
Cancel

## 5.- Mutations by Position:

This function allows for the **detection and calculation of the frequency of occurrence of residues** at one position or several combined positions, which should be previously entered in the 'Mutations' field. This way, you can identify which residues are present at the positions of interest in your sequences and in what combinations (e.g., searching for residues present at all three binding sites of a protein).

The output .csv file is a table with the following columns: In the first column, the names of the input files are listed. The second column displays the combination of residues detected at the positions specified in the 'Mutations' field. In the third column, the number of times each combination appears is shown. The fourth column indicates the frequency of occurrence of that specific combination of residues. The last column provides the number of valid sequences for those positions.

This analysis supports both amino acid and nucleotide combinations. Gaps (-) and question marks (?) are excluded from the analysis. If you want to conduct the analysis in nucleotides, you can use the Find and Replace function in File Editing within Tools to change 'N' to '?', as the function does not automatically exclude 'N' from the analysis.

Example of Mutations by Position output format entering 3 positions:

| | A | B | C | D | E |
|---|---|---|---|---|---|
| 1 | File | Residues (12,15,17) | Number of Mutations | Frequency | Number of Sequences |
| 2 | PR_01_AE.fasta | AIG | 289 | 1.134 | 25494 |
| 3 | PR_01_AE.fasta | AVG | 48 | 0.188 | 25494 |
| 4 | PR_01_AE.fasta | HIG | 1 | 0.004 | 25494 |
| 5 | PR_01_AE.fasta | IIG | 70 | 0.275 | 25494 |
| 6 | PR_01_AE.fasta | ILG | 1 | 0.004 | 25494 |
| 7 | PR_01_AE.fasta | IVG | 12 | 0.047 | 25494 |
| 8 | PR_01_AE.fasta | KIG | 13 | 0.051 | 25494 |

To perform this analysis, select as **input** a folder containing exclusively .fasta files in amino acids or nucleotides, or a folder containing subfolders with .fasta files (for this, mark the 'Subfolders' option).

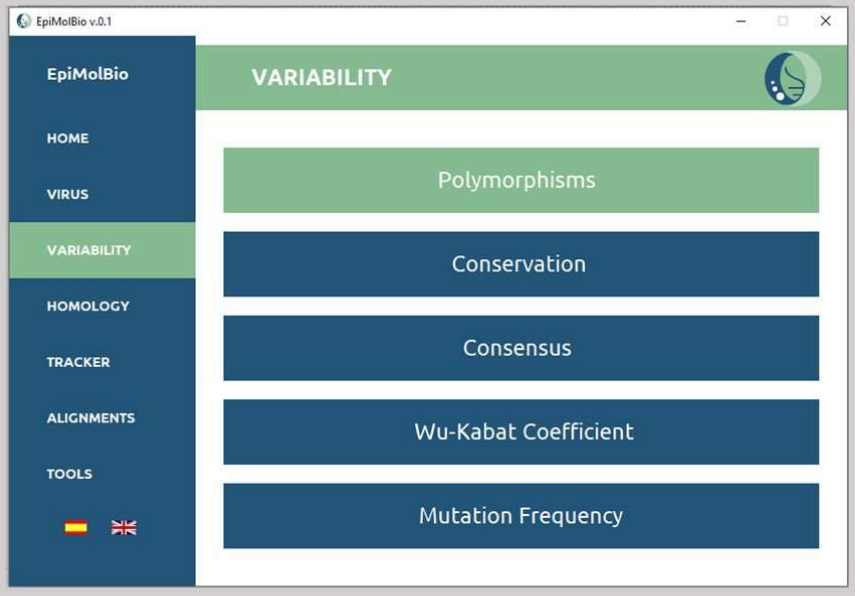
Choose '**Mutations by Position**' from the dropdown menu for '**Output Format**.'

In the '**Mutations**' field, enter the positions you want to analyze by typing the number of each position separated by a comma ',' without spaces (e.g., 9,22,30).

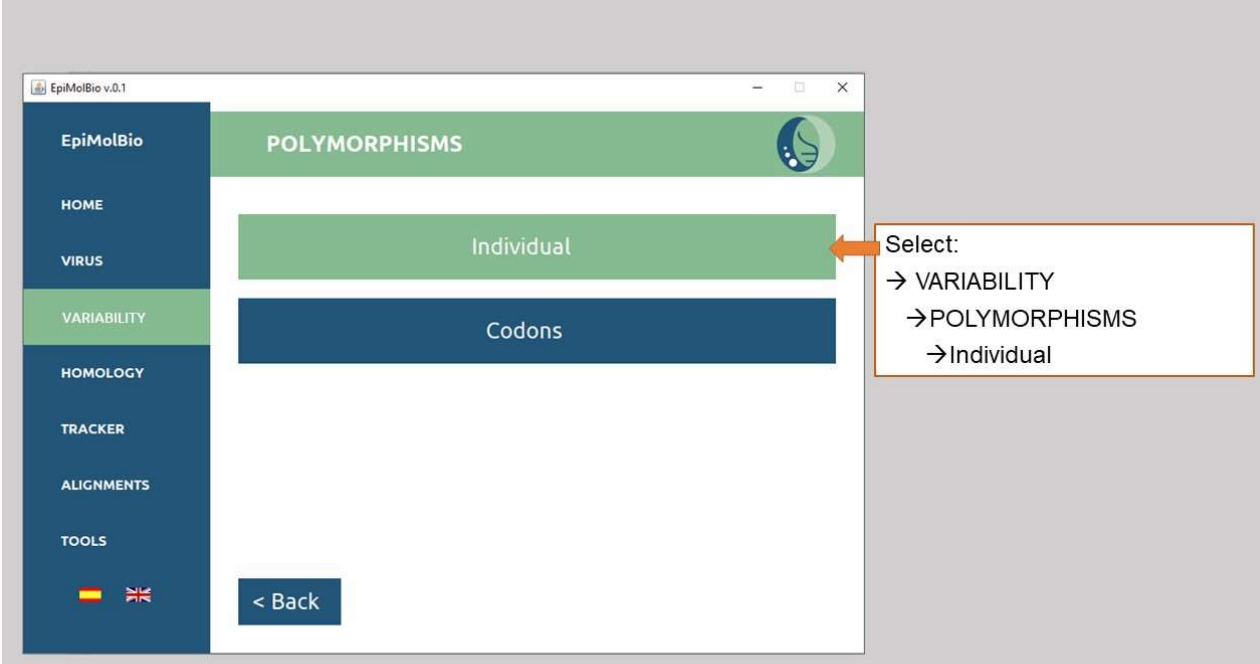
The result will be displayed in a .csv file. In the '**Output**' field, select the folder where you want to save the result and name the file with the .csv extension.

# Step-by-Step:

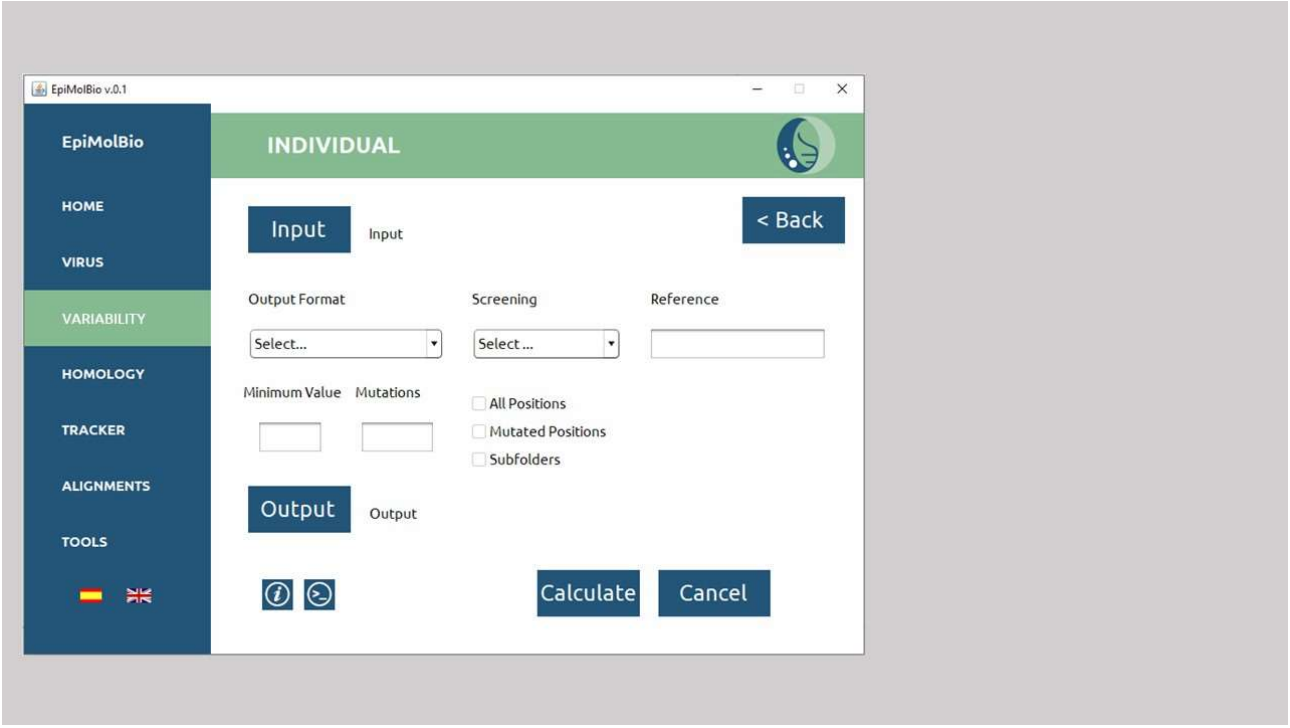
## 1)



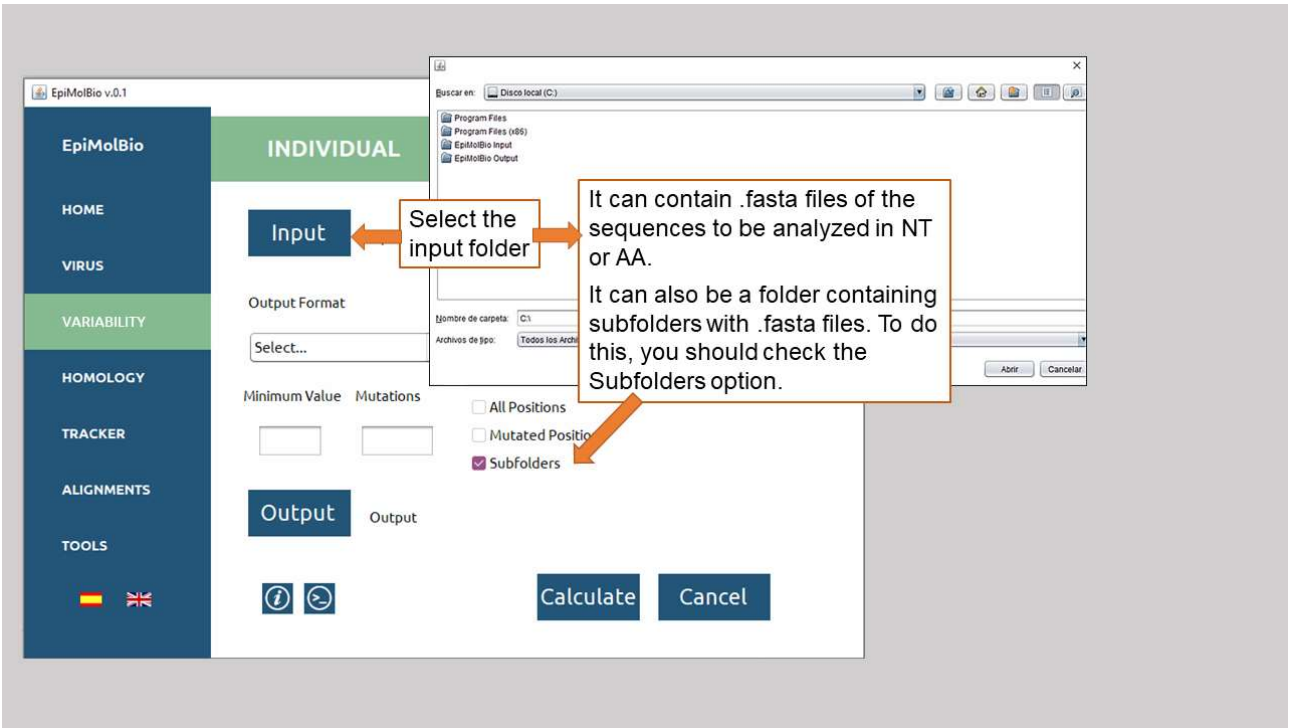

## 2)



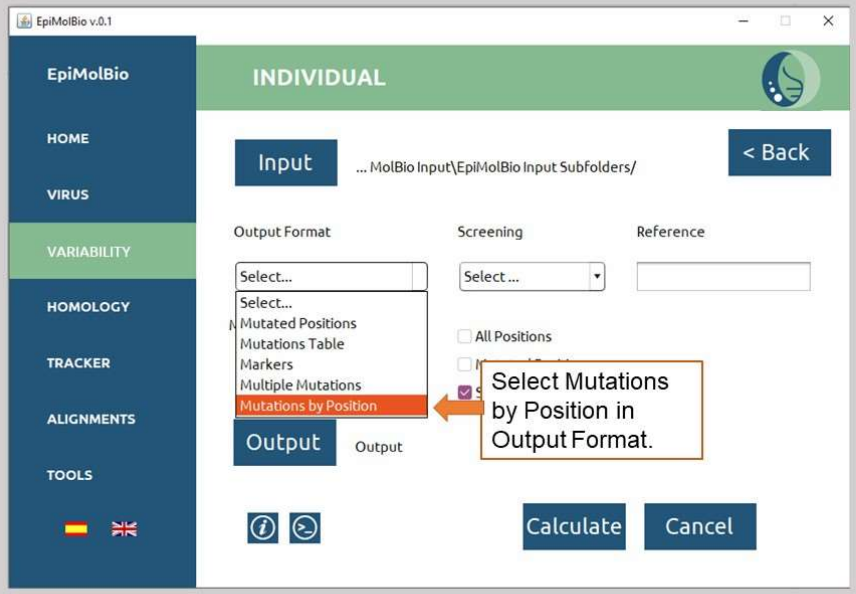
## 3)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
INDIVIDUAL
Input
Input
< Back
Output Format
Screening
Reference
Select...
Select ...
Minimum Value
Mutations
All Positions
Mutated Positions
Subfolders
Output
Output
Calculate
Cancel

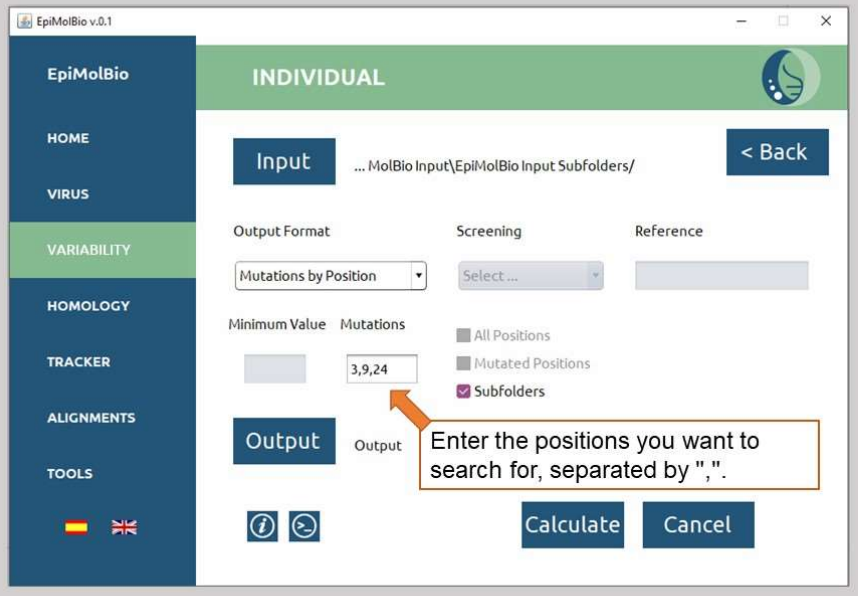
## 4)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
INDIVIDUAL
Input
Select the input folder
It can contain .fasta files of the sequences to be analyzed in NT or AA.
It can also be a folder containing subfolders with .fasta files. To do this, you should check the Subfolders option.
Output Format
Select...
Minimum Value
Mutations
All Positions
Subfolders
Output
Output
Calculate
Cancel

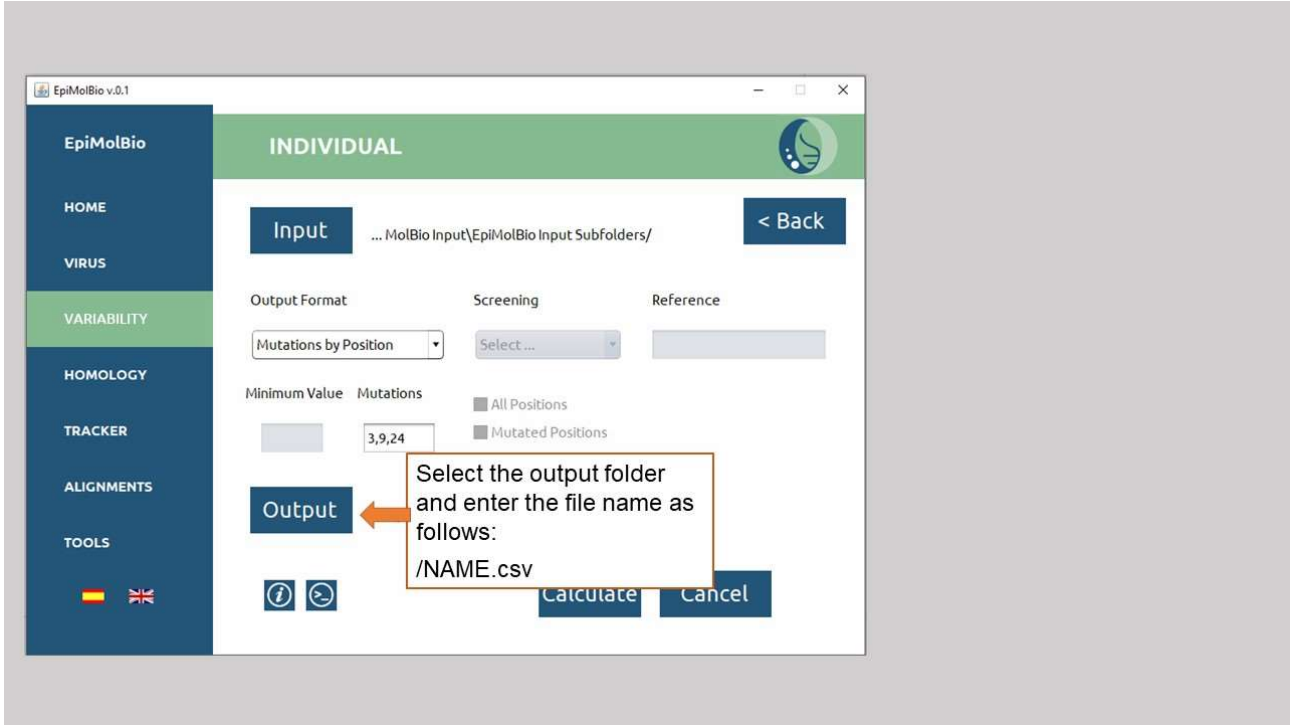
## 5)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
INDIVIDUAL
Input
... MolBio Input\EpiMolBio Input Subfolders/
< Back
Output Format
Screening
Reference
Select...
Select ...
Select...
Mutated Positions
Mutations Table
Markers
Multiple Mutations
Mutations by Position
All Positions
Select Mutations by Position in Output Format.
Output
Output
Calculate
Cancel

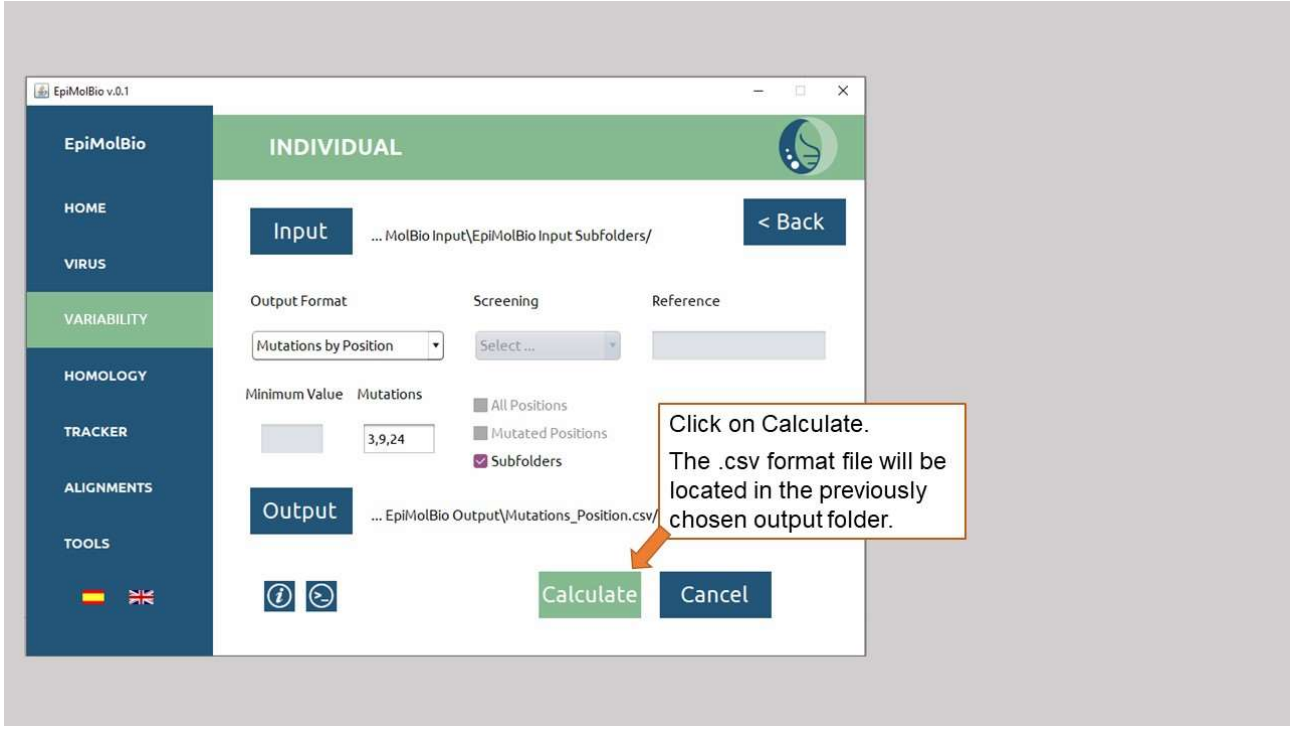
## 6)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
INDIVIDUAL
Input
... MolBio Input\EpiMolBio Input Subfolders/
< Back
Output Format
Screening
Reference
Mutations by Position
Select ...
Minimum Value
Mutations
3,9,24
All Positions
Mutated Positions
Subfolders
Output
Output
Enter the positions you want to search for, separated by ",".
Calculate
Cancel

## 7)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
INDIVIDUAL
Input
... MolBio Input\EpiMolBio Input Subfolders/
< Back
Output Format
Screening
Reference
Mutations by Position
Select ...
Minimum Value
Mutations
3,9,24
All Positions
Mutated Positions
Output
Select the output folder and enter the file name as follows:
/NAME.csv
Calculate
Cancel

## 8)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
INDIVIDUAL
Input
... MolBio Input\EpiMolBio Input Subfolders/
< Back
Output Format
Screening
Reference
Mutations by Position
Select ...
Minimum Value
Mutations
3,9,24
All Positions
Mutated Positions
Subfolders
Output
... EpiMolBio Output\Mutations_Position.csv/
Click on Calculate.
The .csv format file will be located in the previously chosen output folder.
Calculate
Cancel

# II.1.B) CODONS

This function allows for the detection of all **codons that differ from those in the user-provided reference sequence and their frequency of occurrence**. It uses nucleotide sequences in .fasta format as input and excludes gaps (-) and '?' from the analysis.

The **input** file format should be a folder containing exclusively .fasta files with nucleotide sequences to be analyzed.

In the '**Screening**' field, you can select the minimum frequency of occurrence for codon detection. Choose between 100% to detect all codons different from the reference sequence or >75% to detect only those codons that appear with a frequency higher than 75%.

In the '**Reference**' field, input the reference sequence in letters without line breaks, ensuring that the sequence is in nucleotides.

The **output** format will be a .html file. Select the output folder where you want the .html files to appear and name the files with .html extension.

In the output file, at the top, you will find the title of the analysis followed by the name of the input file. Below that, in the 'Position' column, you will see the analyzed codon corresponding to the reference sequence, with the amino acid it encodes and the nucleotides inside parentheses. In the 'Residues' column, all detected variations will be displayed in the following format: encoded amino acid [detected codon] (frequency of occurrence colored according to the color code described in the Overview section, which can be found in the .html output file by clicking on the blue symbol). In the 'Total Positions' column, the total number of valid sequences for that codon will be shown. If a codon does not have any variations, it will not appear in the output file.

Example of Polymorphisms Codons output:

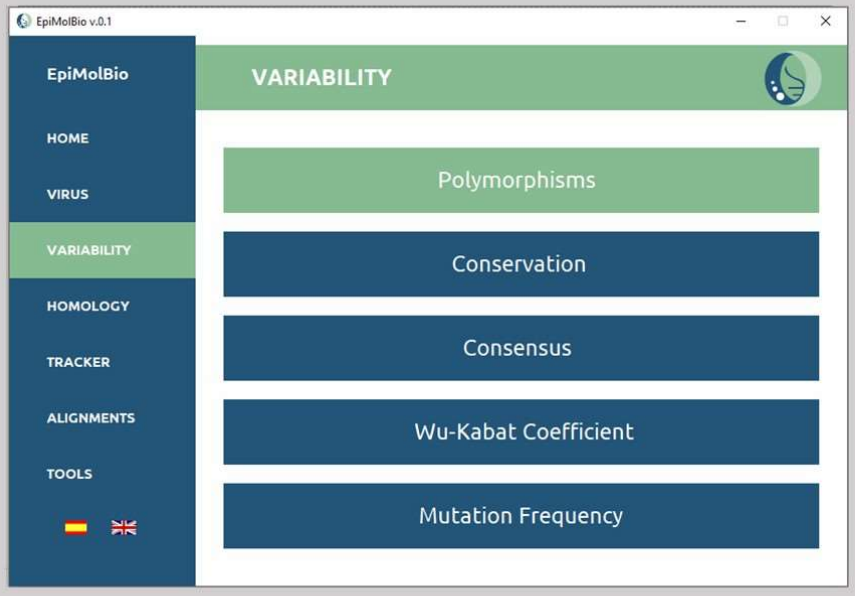
**Variability Polymorphisms Codons > 75%**

**PR_01_AE.fasta**

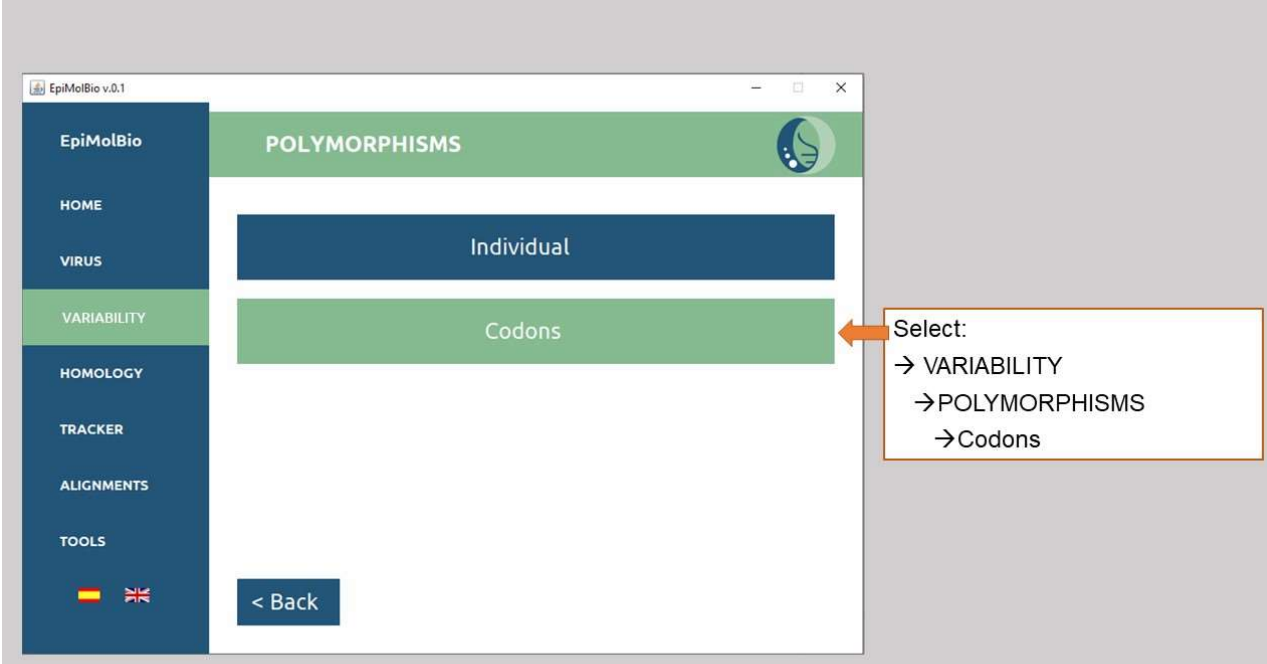
| Position | Residues | Total Positions |
|---|---|---|
| 2 Q(CAG) | Q[CAA(96.554%)] | 26844 |
| 3 V(GTC) | I[ATC(99.765%)] | 26847 |
| 10 L(CTC) | L[CTT(78.294%)] | 26840 |
| 14 K(AAG) | K[AAA(91.358%)] | 26845 |
| 17 G(GGG) | G[GGA(92.468%)] | 26845 |
| 18 Q(CAA) | Q[CAG(89.655%)] | 26845 |
| 35 E(GAA) | D[GAT(80.933%)] | 26847 |
| 36 M(ATG) | I[ATA(98.678%)] | 26846 |
| 37 S(AGT) | N[AAT(90.951%)] | 26844 |

# Step-by-step:

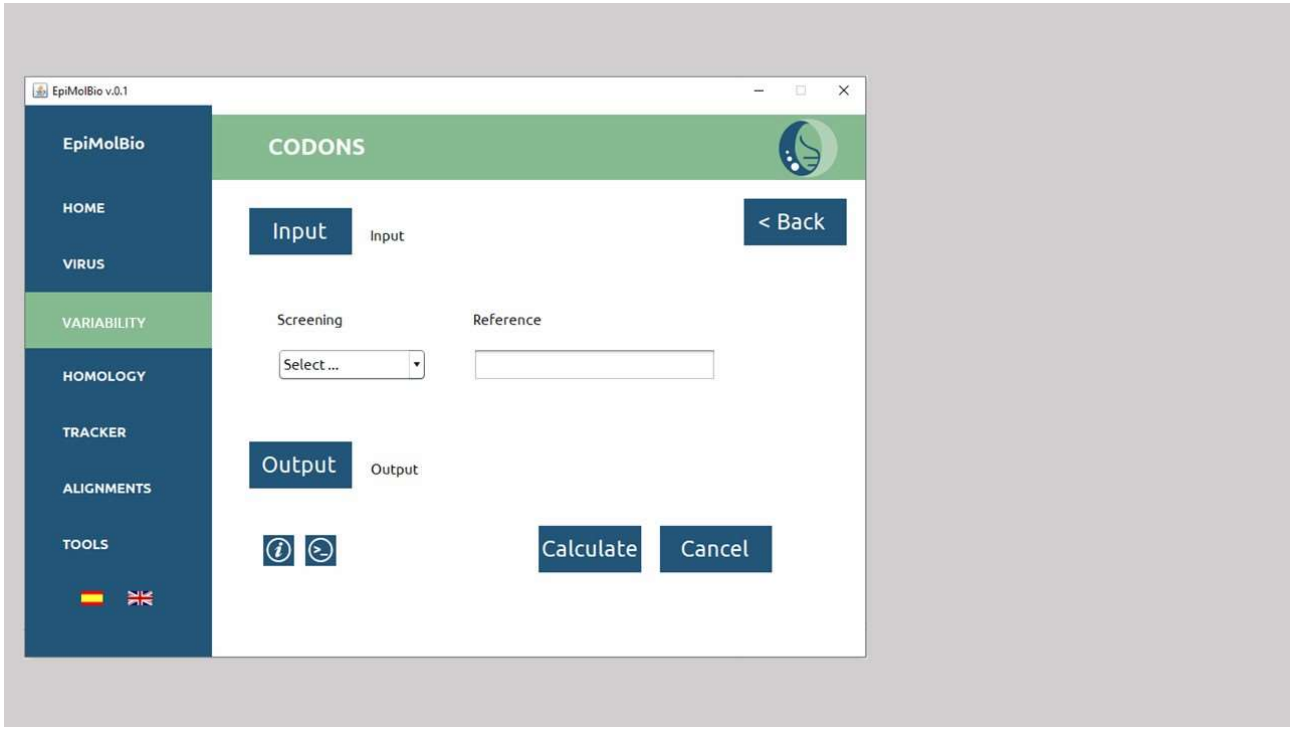
## 1)



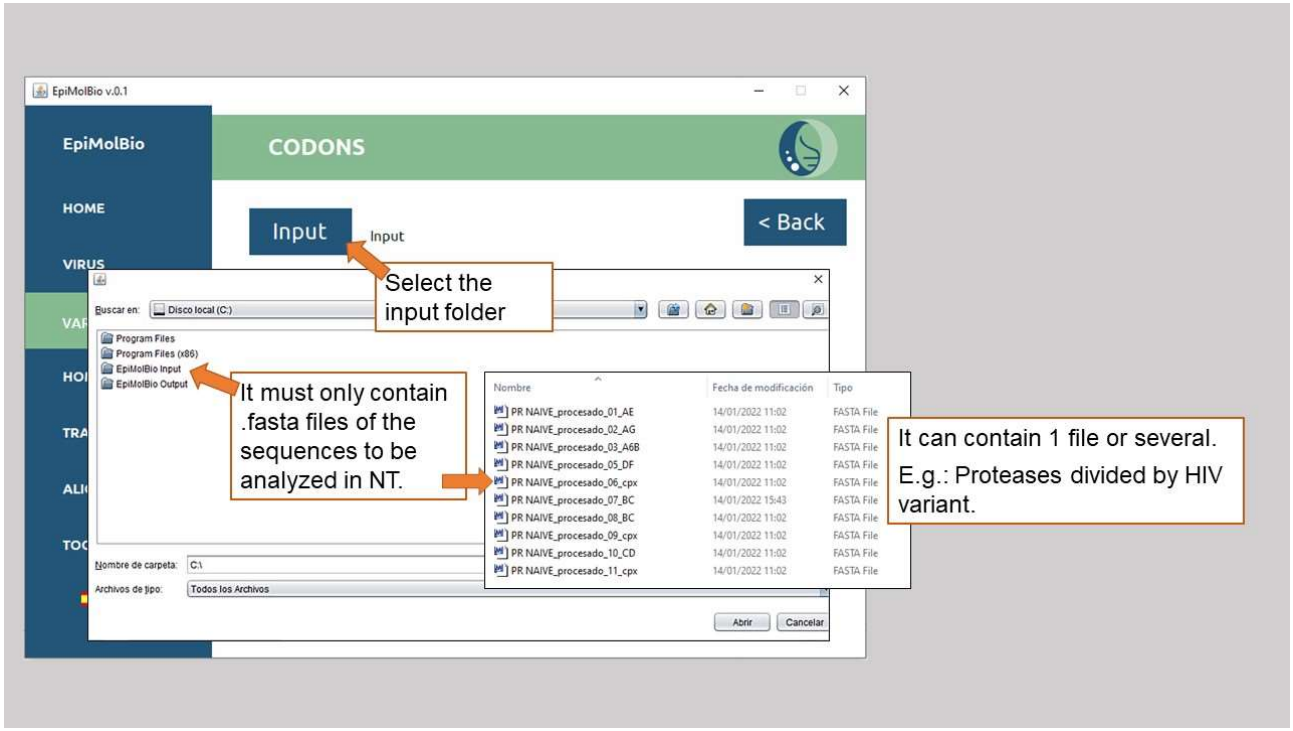

## 2)



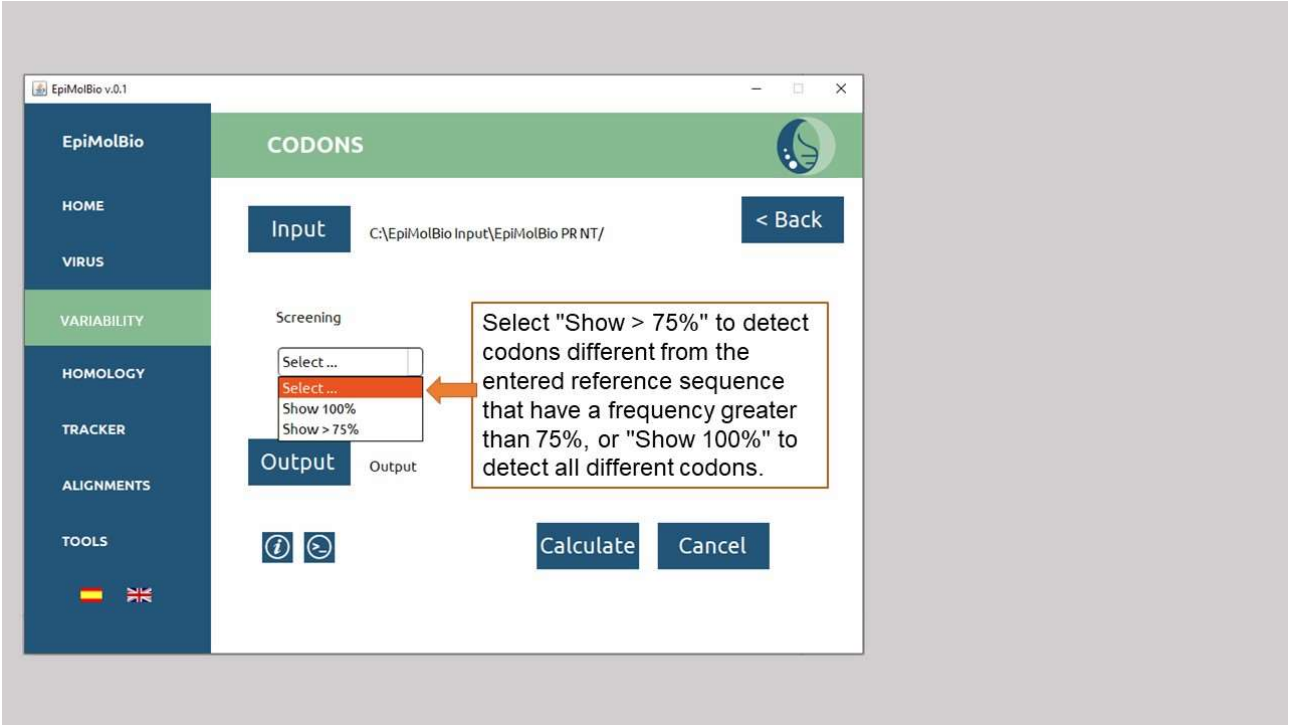
## 3)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
CODONS
Input
Input
< Back
Screening
Reference
Select ...
Output
Output
Calculate
Cancel

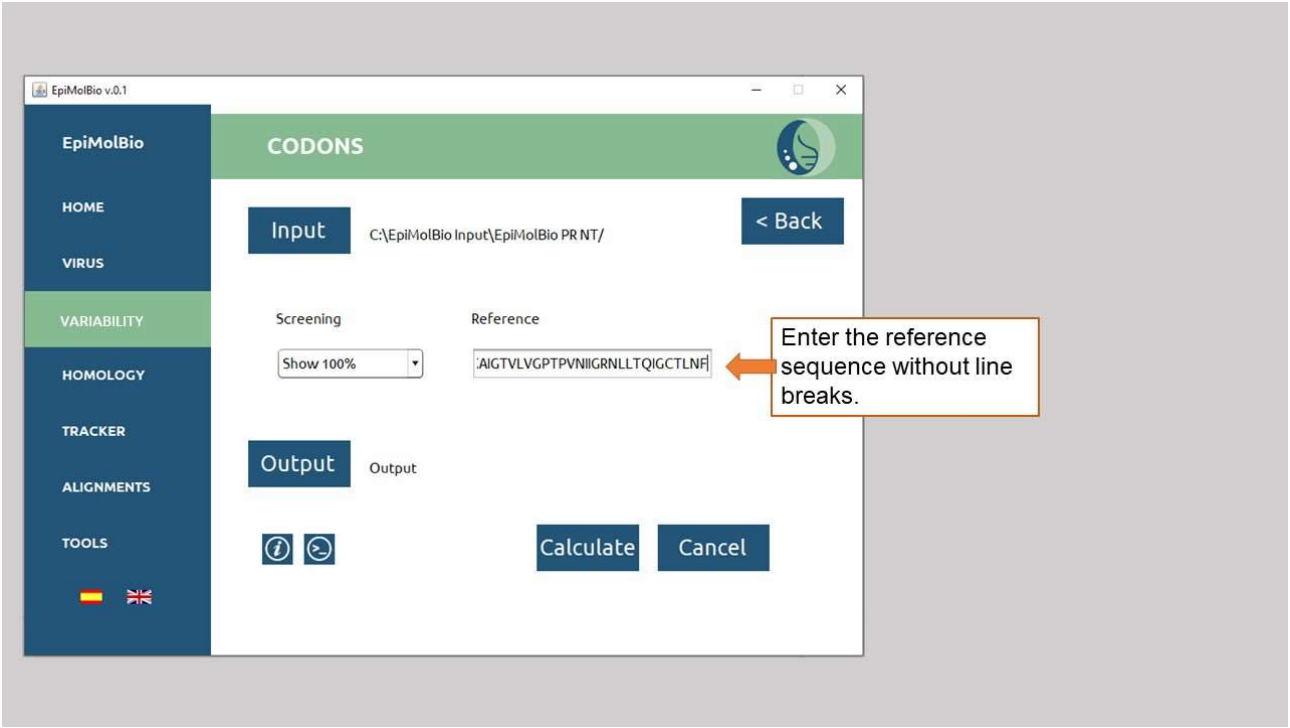
## 4)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
CODONS
Input
Input
< Back
Select the input folder
Buscar en:
Disco local (C:)
Program Files
Program Files (x86)
EpiMolBio Input
EpiMolBio Output
It must only contain .fasta files of the sequences to be analyzed in NT.
Nombre
Fecha de modificación
Tipo
PR.NAIVE_procesado_01_AE
PR.NAIVE_procesado_02_AG
PR.NAIVE_procesado_03_A6B
PR.NAIVE_procesado_05_DF
PR.NAIVE_procesado_06_cpx
PR.NAIVE_procesado_07_BC
PR.NAIVE_procesado_08_BC
PR.NAIVE_procesado_09_cpx
PR.NAIVE_procesado_10_CD
PR.NAIVE_procesado_11_cpx
14/01/2022 11:02
14/01/2022 15:43
FASTA File
It can contain 1 file or several.
E.g.: Proteases divided by HIV variant.
Nombre de carpeta:
C:\
Archivos de tipo:
Todos los Archivos
Abrir
Cancelar

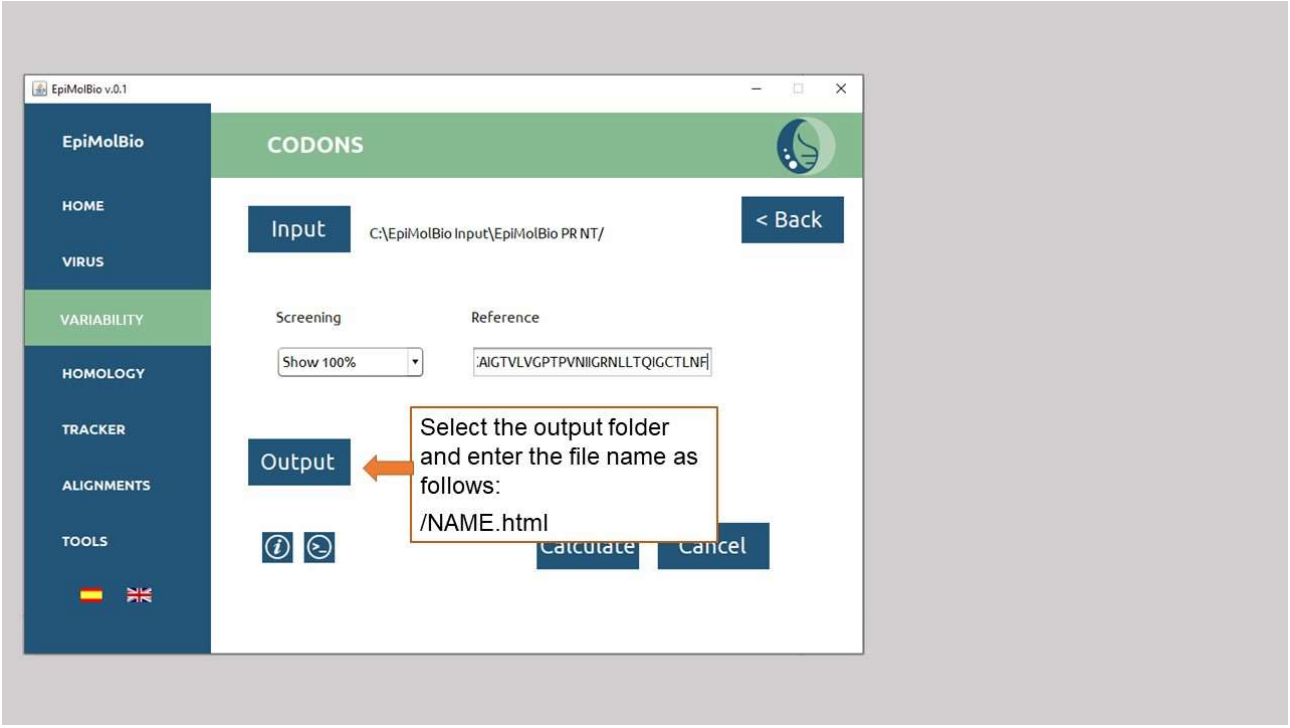
## 5)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
CODONS
Input
C:\EpiMolBio Input\EpiMolBio PR NT/
< Back
Screening
Select ...
Select ...
Show 100%
Show > 75%
Output
Output
Select "Show > 75%" to detect codons different from the entered reference sequence that have a frequency greater than 75%, or "Show 100%" to detect all different codons.
Calculate
Cancel

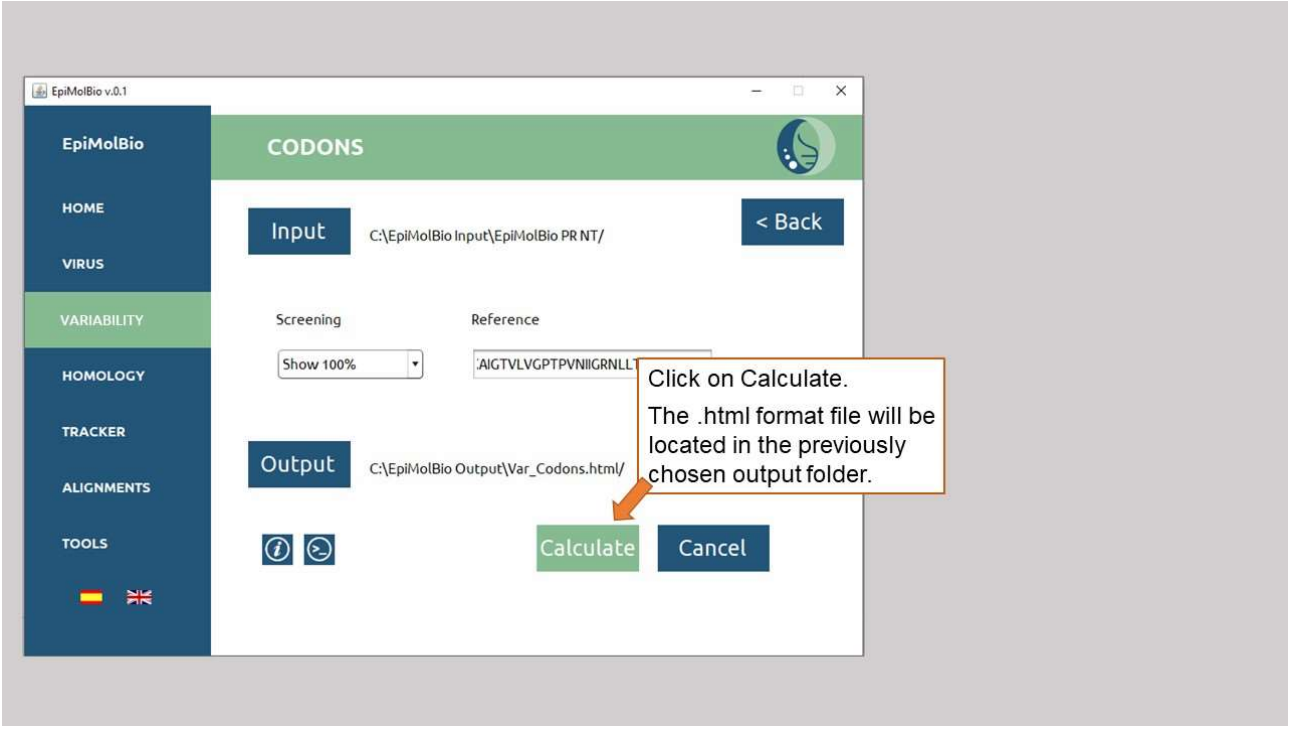
## 6)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
CODONS
Input
C:\EpiMolBio Input\EpiMolBio PR NT/
< Back
Screening
Reference
Show 100%
AIGTVLVGPTPVNIIGRNLLTQIGCTLNF
Enter the reference sequence without line breaks.
Output
Output
Calculate
Cancel

## 7)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
CODONS
Input
C:\EpiMolBio Input\EpiMolBio PR NT/
< Back
Screening
Reference
Show 100%
AIGTVLVGPTPVNIIGRNLLTQIGCTLNF
Output
Select the output folder and enter the file name as follows:
/NAME.html
Calculate
Cancel

## 8)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
CODONS
Input
C:\EpiMolBio Input\EpiMolBio PR NT/
< Back
Screening
Reference
Show 100%
Output
C:\EpiMolBio Output\Var_Codons.html/
Click on Calculate.
The .html format file will be located in the previously chosen output folder.
Calculate
Cancel

# II.2. CONSERVATION

This function allows you to determine the degree of conservation of sequences of interest by providing the most prevalent residue or codon and its percentage of occurrence. It also enables you to apply filters to narrow down the frequencies and generate consensus sequences from the input files.

## II.2.A) INDIVIDUAL

This function allows you to obtain the most conserved amino acid or nucleotide for each position in the analyzed sequence. Gaps (-) and question marks (?) are excluded from the analysis. In some output formats, depending on the chosen filtering criteria, two residues may appear as the most conserved if they have the same frequency of occurrence. In such cases, both residues will be shown in the result.

To use this function, you should provide as **input** a folder containing only .fasta files with aligned sequences to be analyzed. The sequences can be either nucleotides or amino acids. If you are analyzing nucleotide sequences, you will need to use the Find and Replace tool in the File Editing to replace 'N' with '?' to exclude them from the analysis.

There are two different **output formats** to choose from: **List** and **Table**. Both formats are in .html files. The **List** format uses a default filtering of **>75%**, showing only the most conserved residues with a frequency greater than 75%. The **Table** format uses a default filtering of **100%**, showing all completely conserved residues.

In the '**Reference**' field, you should enter a reference sequence without line breaks in NT or AA depending on the input file.

The **output** file will be in .html format. You need to select the output folder where you want the .html files to appear and name the files with .html at the end.

**1.-List:**

In the List output format, the analysis title followed by the input file name is displayed at the top. Below, the consensus sequence for each file is shown with residues colored according to their percentage using the color code described in Overview, which can be accessed in the .html output file by clicking on the blue symbol. In the 'Position' column, the positions that contain a conserved residue with a frequency greater than 75% are listed, along with their reference amino acid. The 'Residues' column shows the most frequent amino acid for each position, followed by its percentage colored according to the color code. The 'Total Positions' column displays the total number of valid sequences for that position.

Example of List output format for Individual Conservation analysis:

**Variability Conservation Individual List**

**PR_01_AE.fasta**

CONSENSUS PQITLWQRPLVTVKIGGQLKEALLDTGADDTVLEDINLPGKWKPKMIGGIGGFIKVRQYDQILIEICGKKAIGTVLVGPTPVNIIGRNMLTQIGCTLNF

| Position | Residues | Total Positions |
|---|---|---|
| P1 | P(99.896%) | 26838 |
| Q2 | Q(99.782%) | 26649 |
| V3 | I(99.858%) | 26831 |
| T4 | T(99.858%) | 26816 |
| L5 | L(99.888%) | 26780 |
| W6 | W(99.929%) | 26836 |
| Q7 | Q(99.751%) | 26536 |

## 2.-Table:

In the Table output format, the analysis title is displayed at the top. Below that, the first two rows show the reference sequence and the positions of each of its residues. The analysis results are displayed in the following rows, showing the input file name and three rows corresponding to the most frequent residue (nucleotide or amino acid) with the cell colored according to the previously described color code, the conservation frequency per position also with the cell colored, and the number of valid sequences for each position.

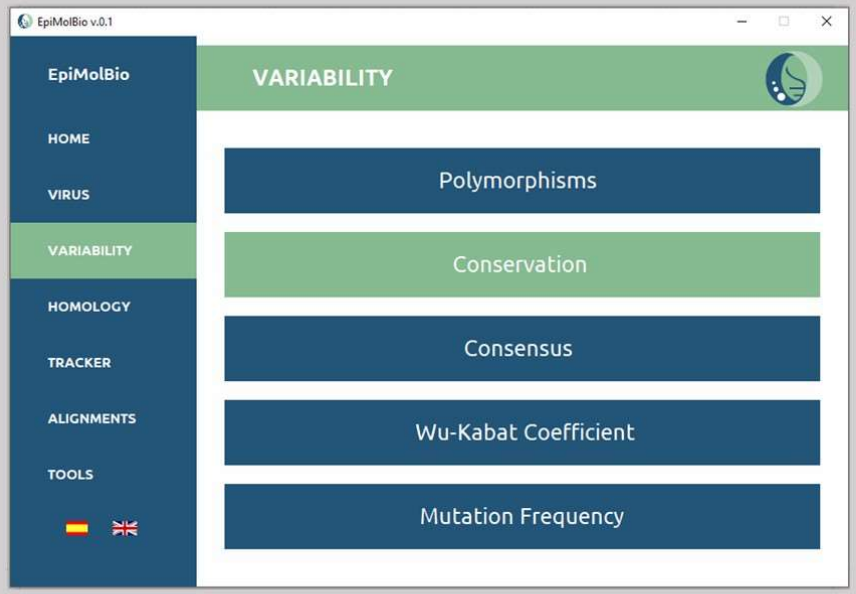
Example of Table output format for Individual Conservation analysis:

**Variability Conservation Individual Table**

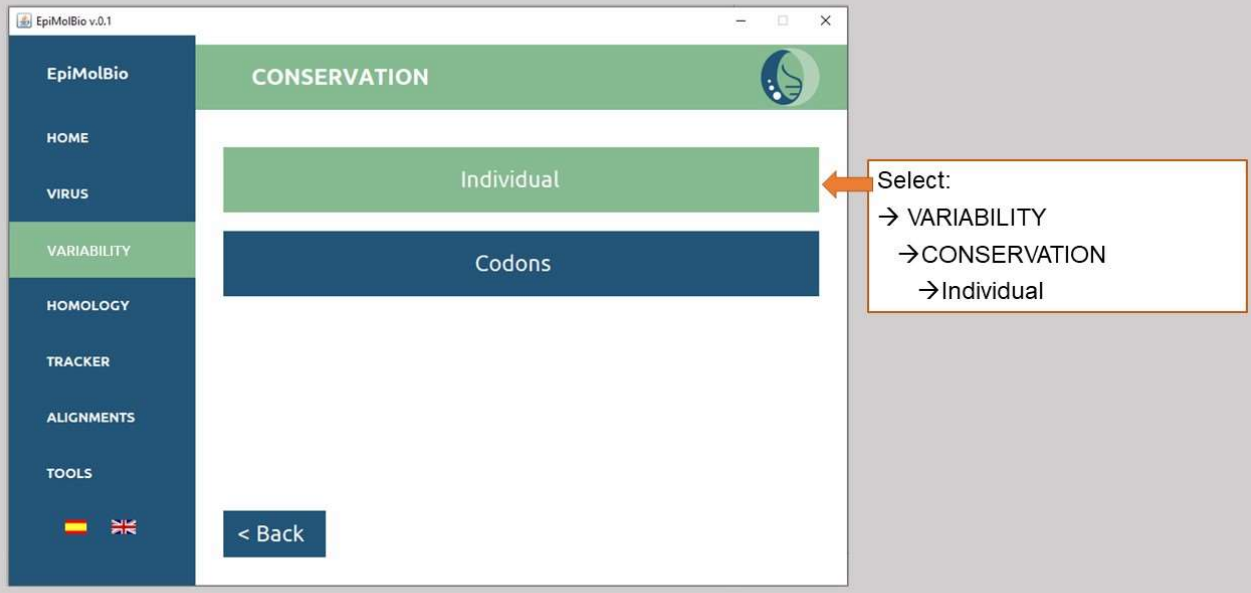
| File | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Reference | P | Q | V | T | L | W | Q | R | P | L | V | T | I | K | I | G |
| | Position | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 |
| PR_01_AE.fasta | Residue | P | Q | I | T | L | W | Q | R | P | L | V | T | V | K | I | G |
| | Conservation | 99.896 | 99.782 | 99.858 | 99.858 | 99.888 | 99.929 | 99.751 | 99.922 | 99.955 | 87.446 | 99.512 | 95.028 | 56.099 | 95.188 | 88.392 | 72.550 |
| | Number of Sequences | 26838 | 26649 | 26831 | 26816 | 26780 | 26836 | 26536 | 26792 | 26613 | 25952 | 26416 | 26469 | 26150 | 26270 | 26206 | 25636 |
| PR_02_AG.fasta | Residue | P | Q | I | T | L | W | Q | R | P | L | V | T | V | R | I | G |
| | Conservation | 99.948 | 99.819 | 99.529 | 99.728 | 99.958 | 99.875 | 99.838 | 99.895 | 99.947 | 83.181 | 96.532 | 89.274 | 91.362 | 59.803 | 85.337 | 70.800 |
| | Number of Sequences | 9571 | 9416 | 9557 | 9561 | 9560 | 9575 | 9248 | 9557 | 9351 | 9186 | 9313 | 9295 | 9308 | 8916 | 9289 | 9233 |

# Step-by-step:

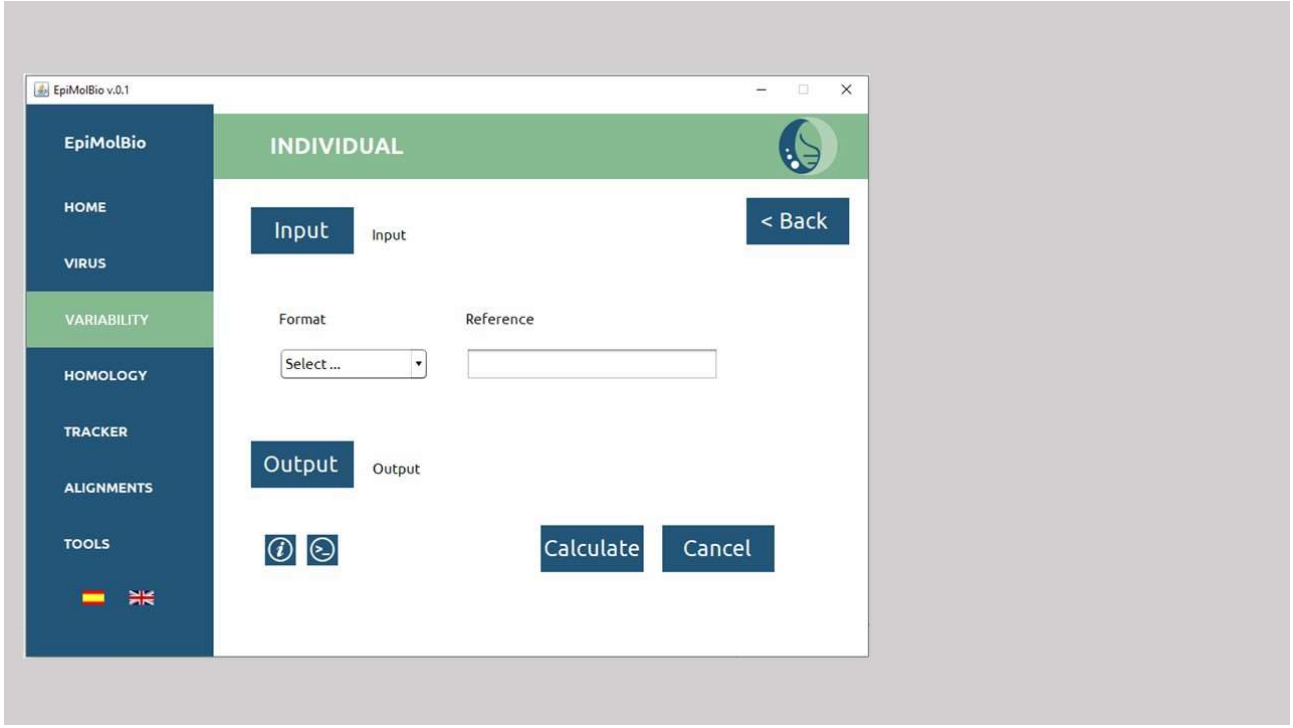
## 1)



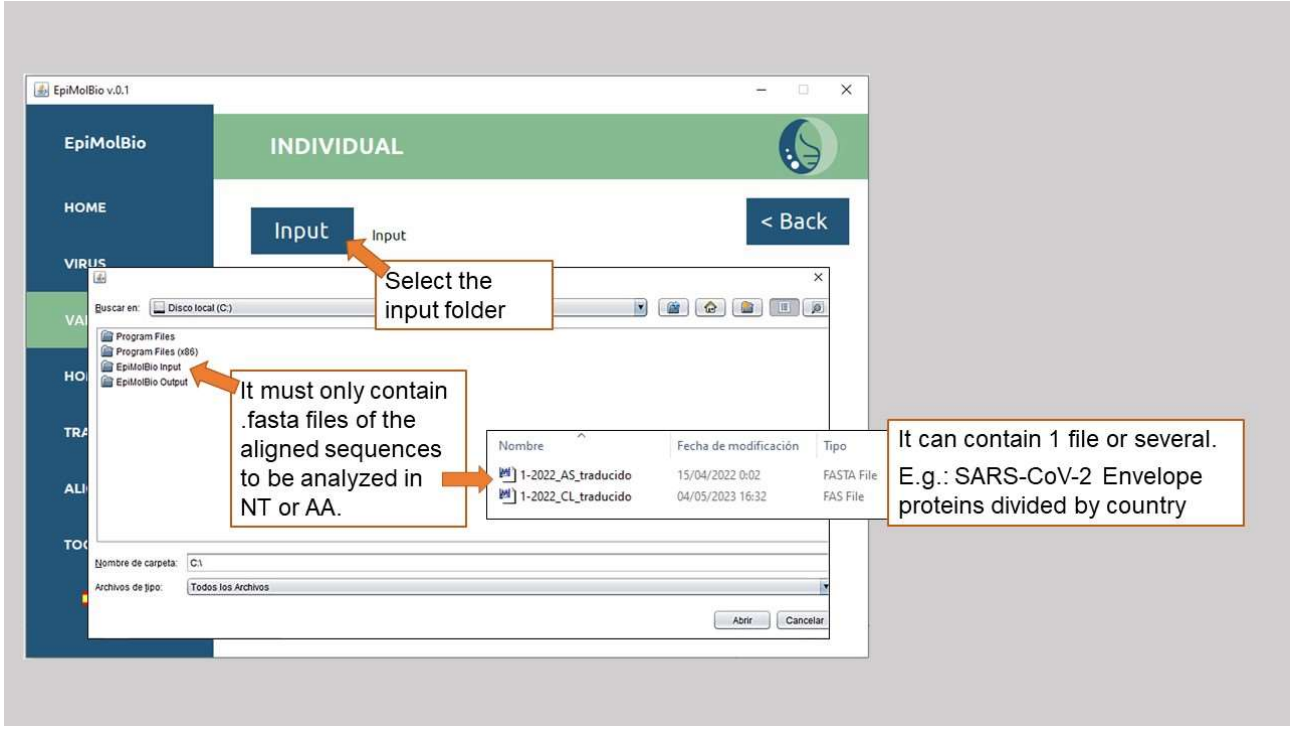

## 2)



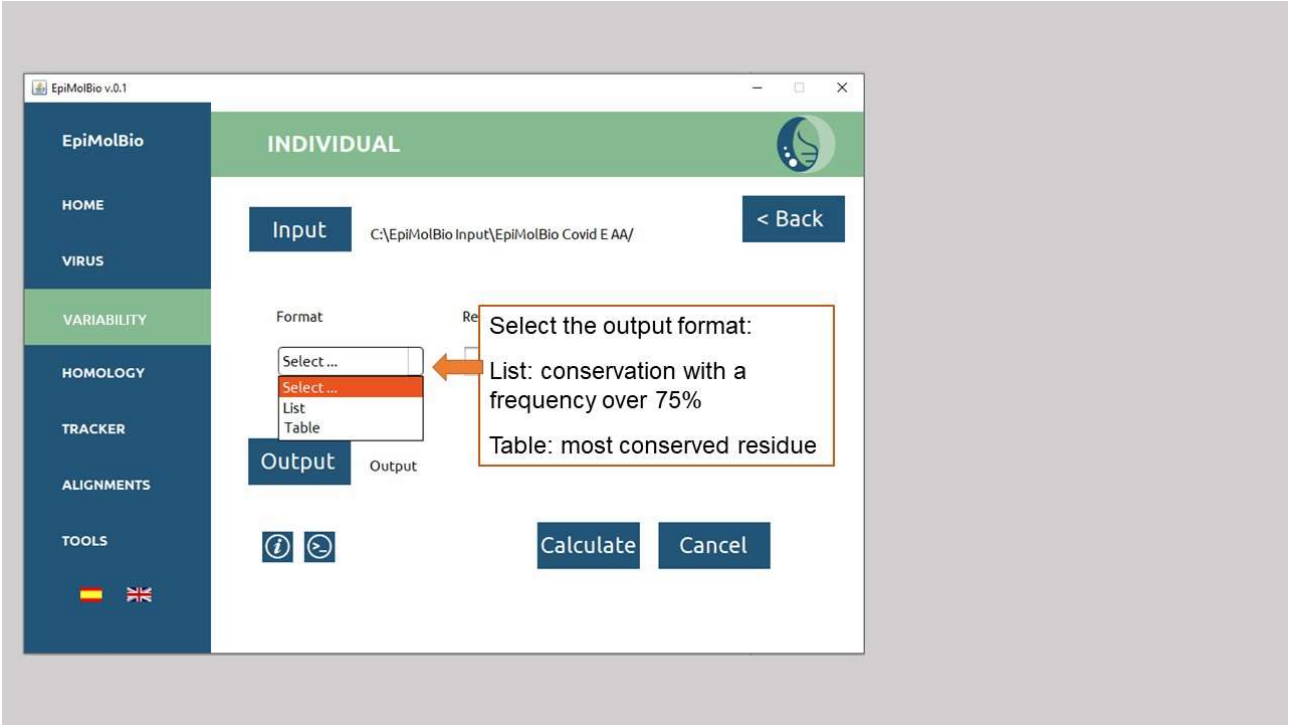
## 3)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
INDIVIDUAL
Input
Input
< Back
Format
Reference
Select ...
Output
Output
Calculate
Cancel

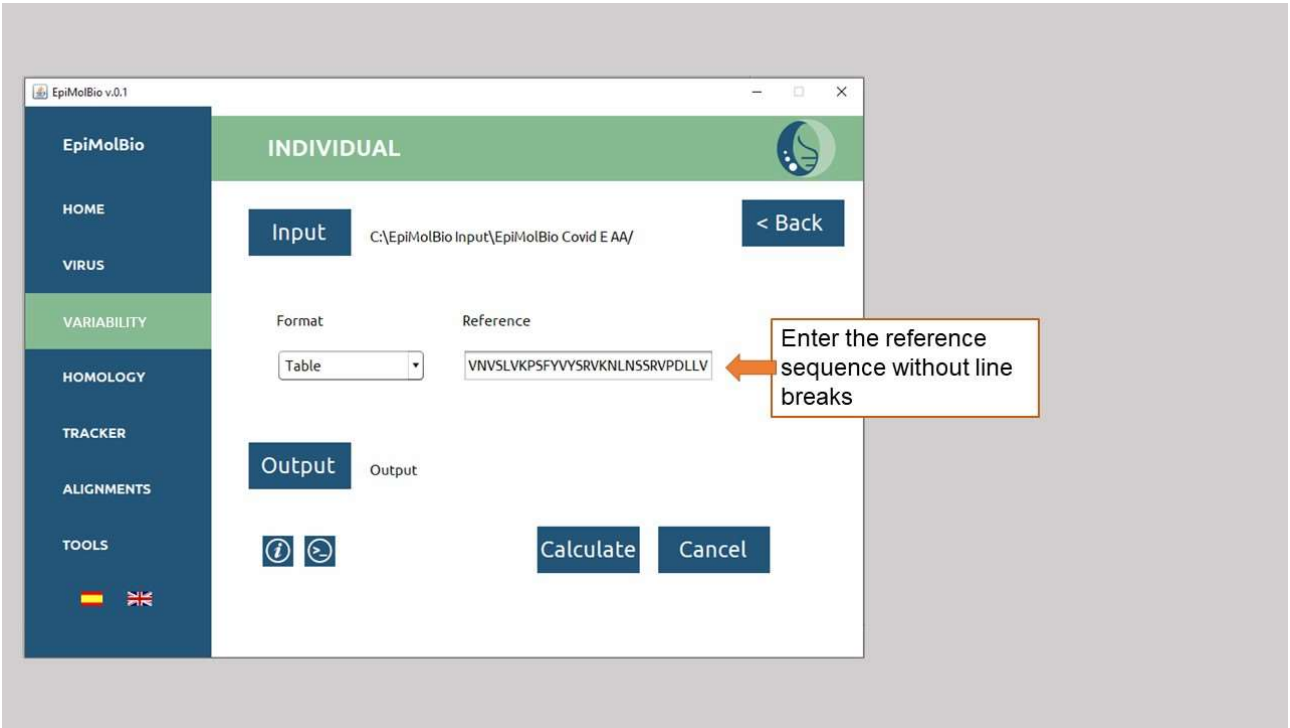
## 4)


EpiMolBio v.0.1
EpiMolBio
HOME
INDIVIDUAL
Input
Input
< Back
Select the input folder
Buscar en:
Disco local (C:)
Program Files
Program Files (x86)
EpiMolBio Input
EpiMolBio Output
It must only contain .fasta files of the aligned sequences to be analyzed in NT or AA.
Nombre
Fecha de modificación
Tipo
1-2022_AS_traducido
15/04/2022 0:02
FASTA File
1-2022_CL_traducido
04/05/2023 16:32
FAS File
It can contain 1 file or several.
E.g.: SARS-CoV-2 Envelope proteins divided by country
Nombre de carpeta:
C:\
Archivos de tipo:
Todos los Archivos
Abrir
Cancelar

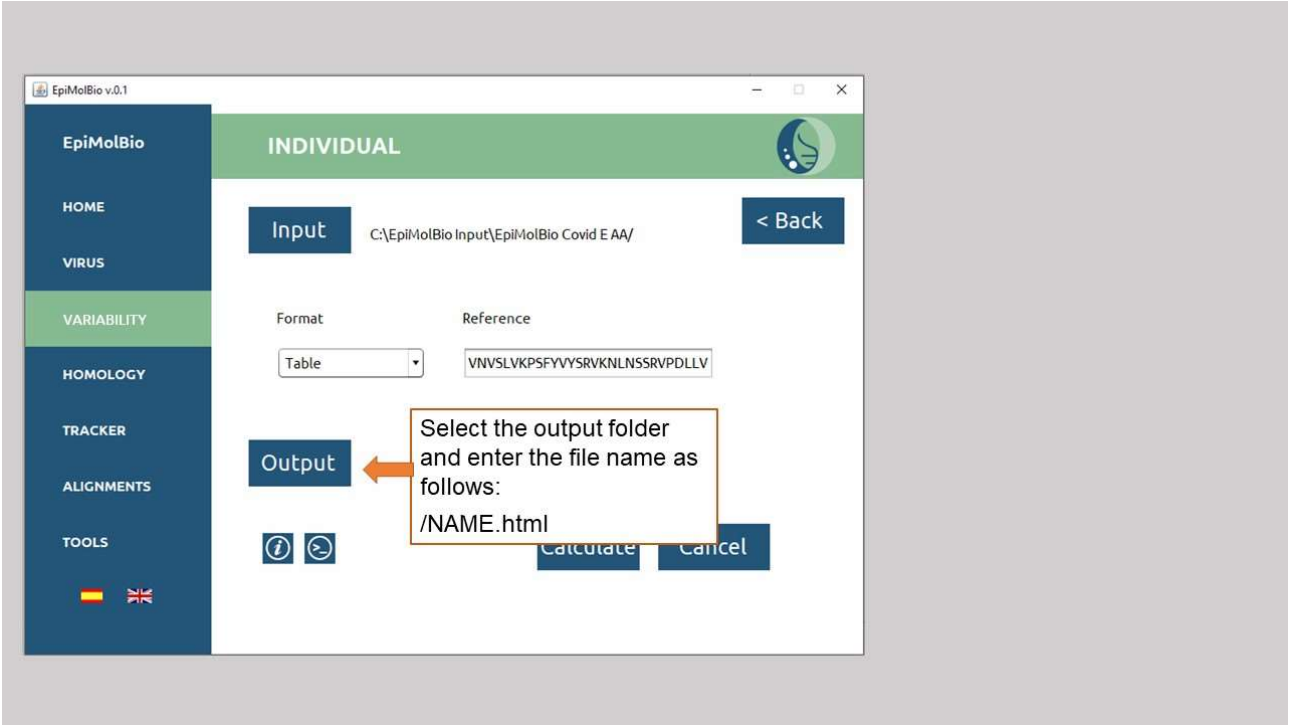
## 5)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
INDIVIDUAL
Input
C:\EpiMolBio Input\EpiMolBio Covid E AA/
< Back
Format
Select ...
Select ...
List
Table
Output
Output
Select the output format:
List: conservation with a frequency over 75%
Table: most conserved residue
Calculate
Cancel

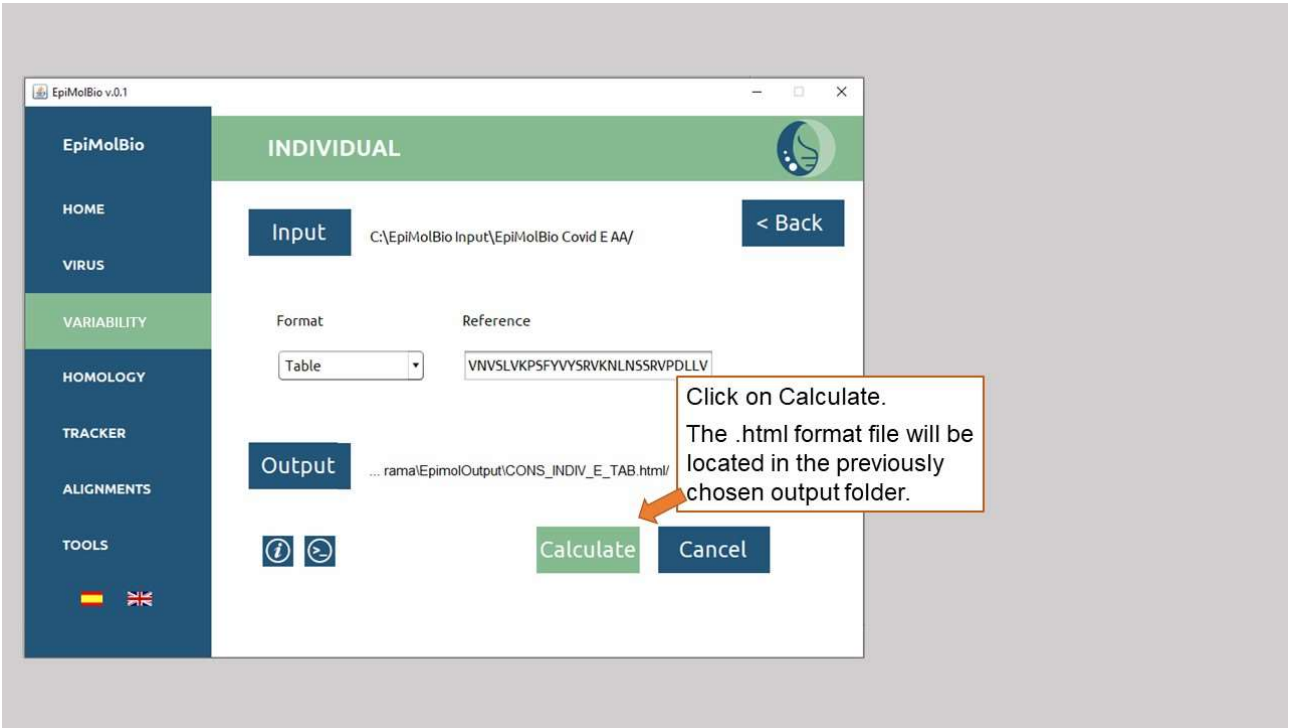
## 6)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
INDIVIDUAL
Input
C:\EpiMolBio Input\EpiMolBio Covid E AA/
< Back
Format
Reference
Table
VNVSLVKPSFYVYSRVKNLNSSRVPDLLV
Enter the reference sequence without line breaks
Output
Output
Calculate
Cancel

## 7)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
INDIVIDUAL
Input
C:\EpiMolBio Input\EpiMolBio Covid E AA/
< Back
Format
Reference
Table
VNVSLVKPSFYVYSRVKNLNSSRVPDLLV
Output
Select the output folder and enter the file name as follows:
/NAME.html
Calculate
Cancel

## 8)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
INDIVIDUAL
Input
C:\EpiMolBio Input\EpiMolBio Covid E AA/
< Back
Format
Reference
Table
VNVSLVKPSFYVYSRVKNLNSSRVPDLLV
Output
... rama\EpimolOutput\CONS_INDIV_E_TAB.html/
Click on Calculate.
The .html format file will be located in the previously chosen output folder.
Calculate
Cancel

## II.2.B) CODONS

This function allows you to obtain **the most conserved codon for each triplet** in an analyzed nucleotide sequence. Gaps (-) and 'N' are excluded from the analysis.

The **input** file format must be a folder containing exclusively .fasta files with aligned nucleotide sequences.

In the '**Screening**' field, choose the screening percentage: **100%** to display all detected codons or **>75%** to show only those with a frequency of occurrence higher than 75%.

In the '**Reference**' field, enter a reference sequence without line breaks and in nucleotides.

The **output** format will be a file with the extension .html. You will need to select the output folder where you want the .html files to appear and name the files with .html at the end.

In the output file, the analysis title is displayed at the top, followed by the input file name. The 'Position' column shows the position of the amino acid encoded by each analyzed codon, with the codon in parentheses. The 'Residues' column displays all the detected residues [the codon and its percentage] if the chosen screening is 100%. If you choose the >75% screening, only the most conserved codon that appears with a frequency > 75% in the analyzed sequences will be displayed. The percentages will be colored according to the **color code** described in Overview, which can be consulted in the output .html file by clicking on the blue symbol. The 'Total Positions' column shows the total number of valid sequences for that position.

Example of output format with 100% screening:

**Variability Conservation Codons 100%**

**PR_01_AE.fasta**

| Position | Residues | Total Positions |
|---|---|---|
| 1 P(CCT) | P[CCT(99.810%)] S[TCC(0.011%)] P[CCA(0.011%)] ?[CCY(0.022%)] P[CCC(0.026%)] S[TCT(0.067%)] A[GCT(0.004%)] L[CTT(0.004%)] ?[SCT(0.007%)] ?[CCW(0.004%)] P[CCG(0.007%)] T[ACT(0.007%)] H[CAT(0.004%)] ?[YCT(0.004%)] ?[CMT(0.004%)] V[GTC(0.004%)] L[CTC(0.004%)] | 26849 |
| 2 Q(CAG) | Q[CAA(96.554%)] Q[CAG(2.503%)] E[GAA(0.071%)] ?[CAR(0.596%)] S[TCA(0.019%)] H[CAT(0.034%)] D[GAC(0.004%)] K[AAA(0.022%)] L[CTG(0.004%)] H[CAC(0.022%)] ?[CAM(0.034%)] ?[CWW(0.004%)] ?[YAA(0.004%)] ?[CAW(0.026%)] ?[CMM(0.015%)] ?[SAA(0.007%)] P[CCT(0.004%)] P[CCC(0.004%)] L[CTC(0.011%)] ?[CWM(0.004%)] R[CGA(0.004%)] R[CGC(0.004%)] T[ACA(0.004%)] ?[CRA(0.019%)] ?[CMA(0.004%)] *[TAG(0.007%)] ?[CAV(0.004%)] ?[MAR(0.004%)] ?[CWA(0.007%)] R[AGA(0.004%)] | 26844 |

Example of output format with 75% screening:

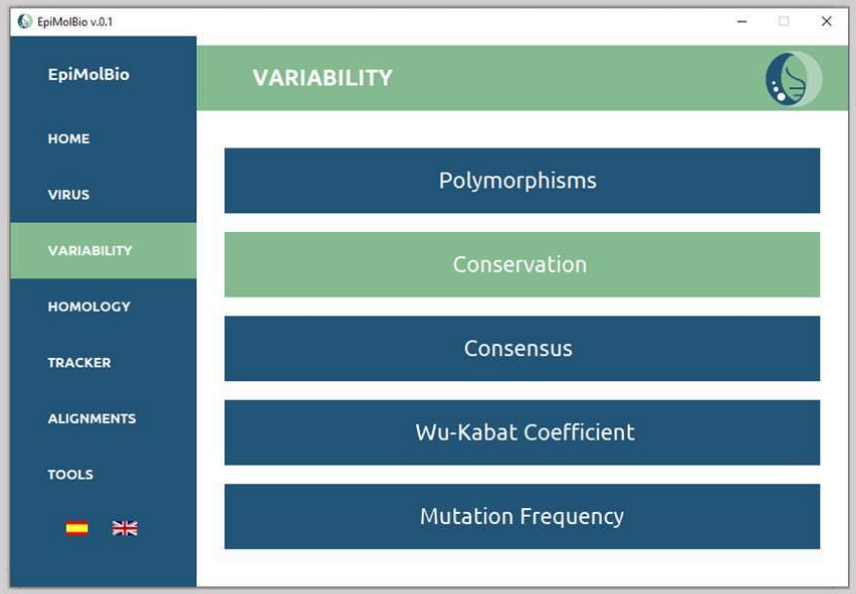
**Variability Conservation Codons > 75%**

**PR_01_AE.fasta**

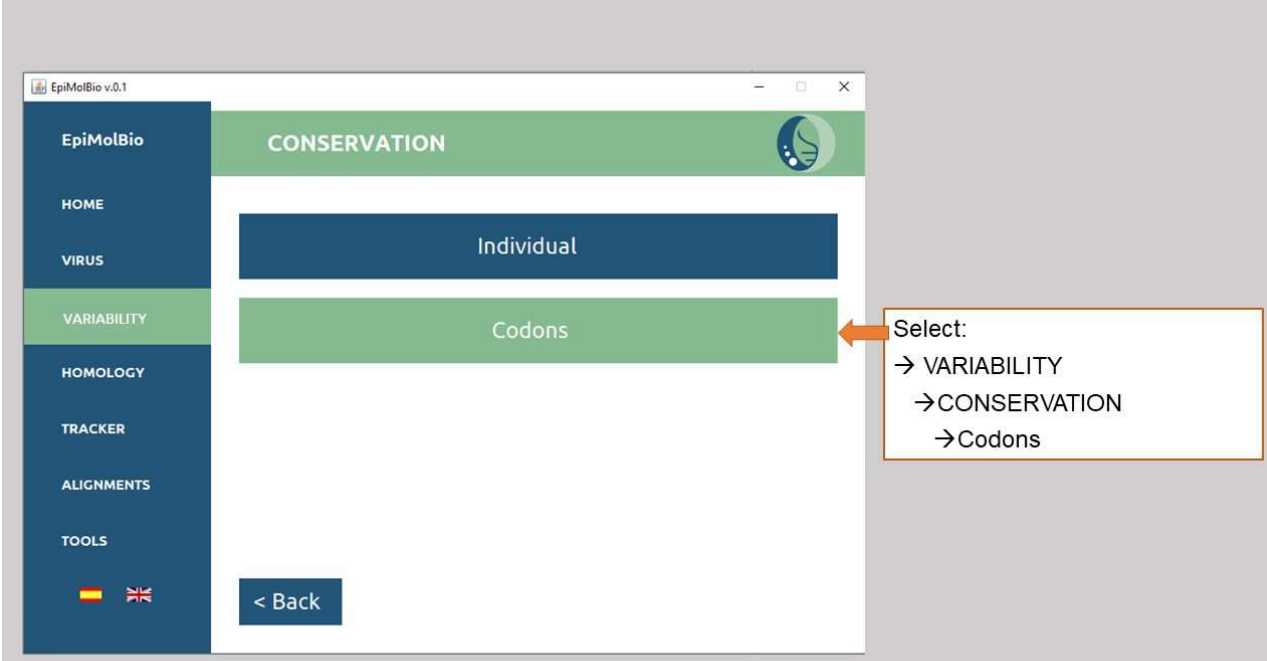
| Position | Residues | Total Positions |
|---|---|---|
| 1 P(CCT) | P[CCT(99.810%)] | 26849 |
| 2 Q(CAG) | Q[CAA(96.554%)] | 26844 |
| 3 V(GTC) | I[ATC(99.765%)] | 26847 |

# Step-by-step:

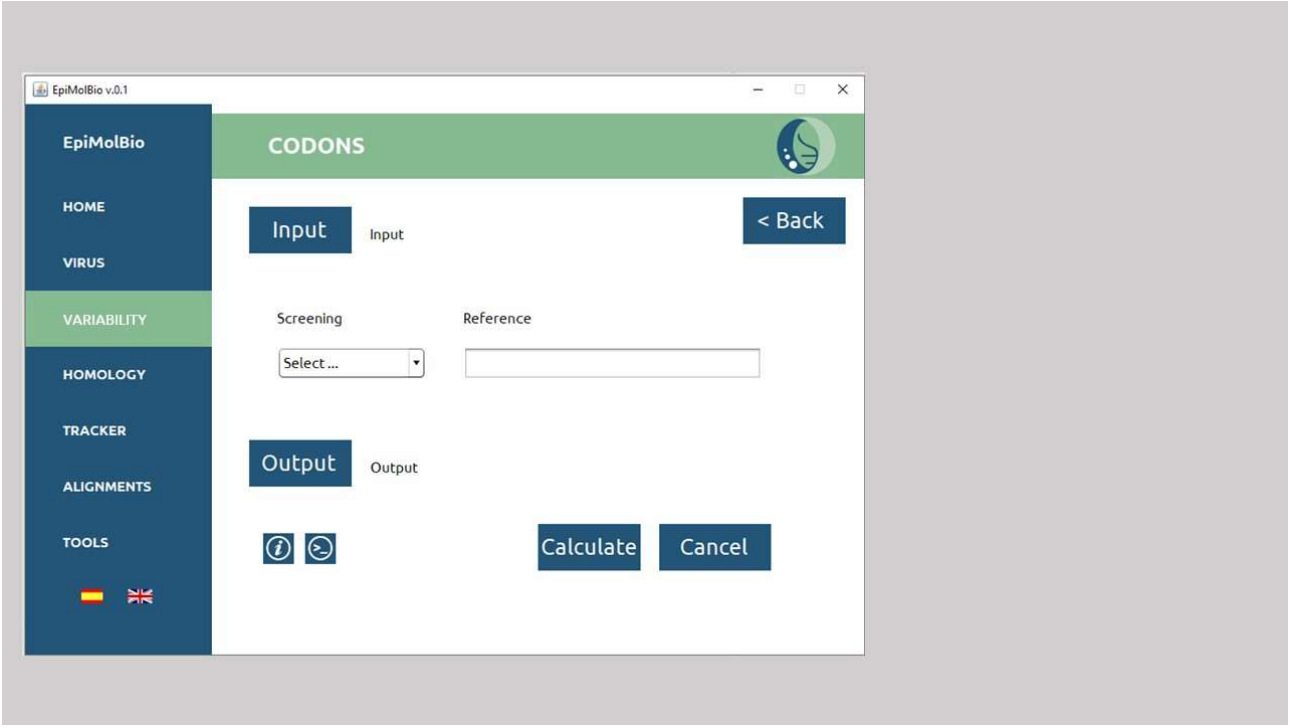
## 1)



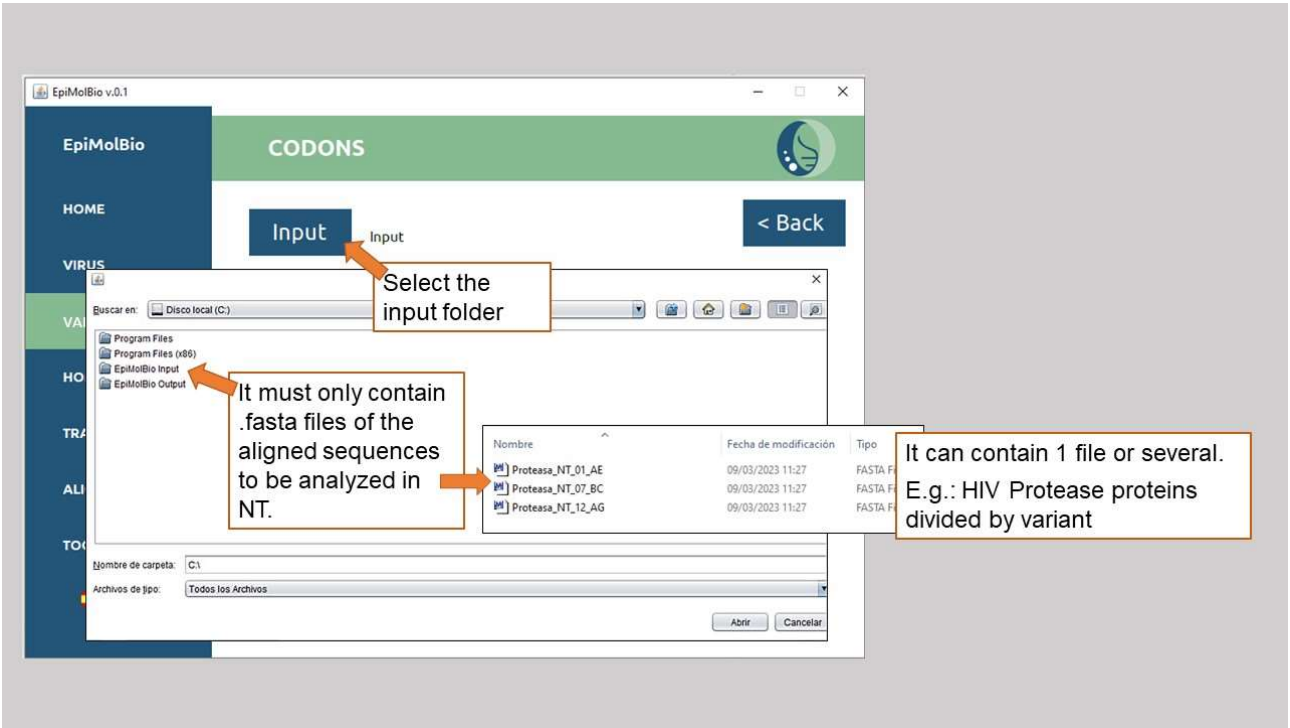

## 2)



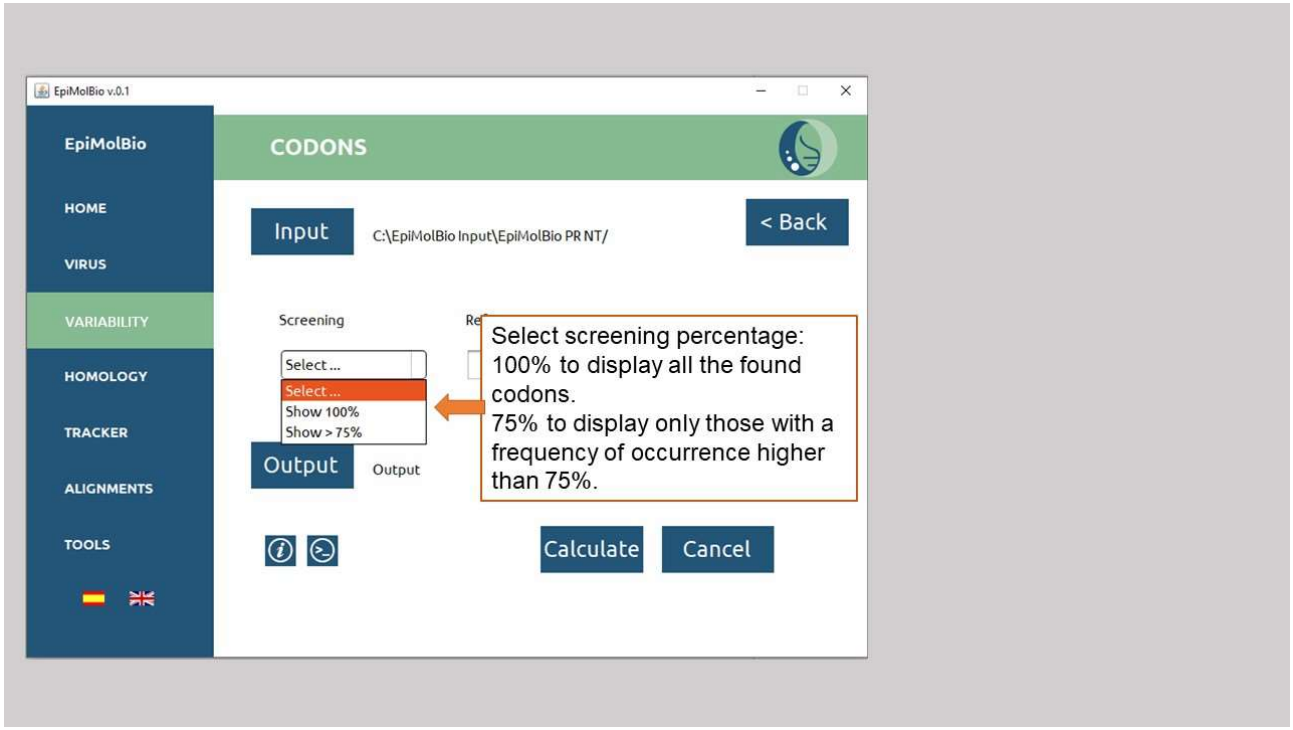
## 3)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
CODONS
Input
Input
< Back
Screening
Reference
Select ...
Output
Output
Calculate
Cancel

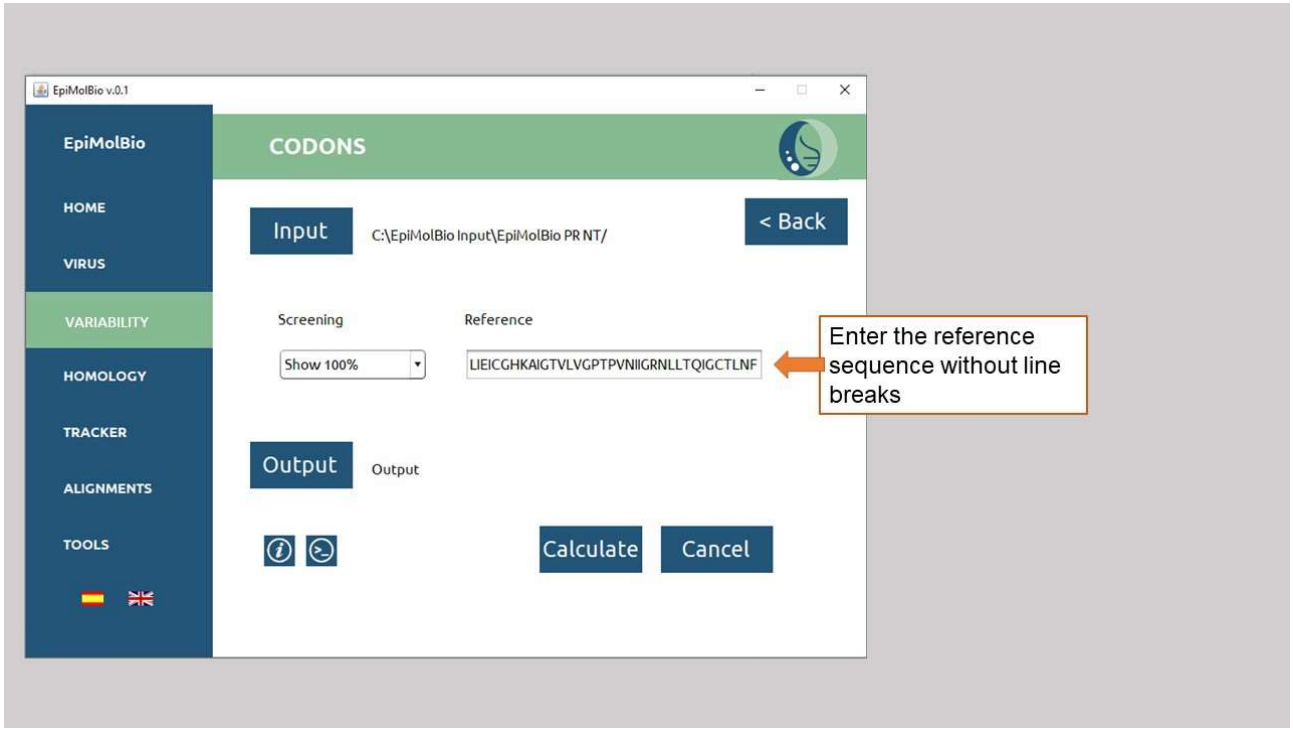
## 4)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
CODONS
Input
Input
< Back
Select the input folder
Buscar en:
Disco local (C:)
Program Files
Program Files (x86)
EpiMolBio Input
EpiMolBio Output
It must only contain .fasta files of the aligned sequences to be analyzed in NT.
Nombre
Fecha de modificación
Tipo
Proteasa_NT_01_AE
Proteasa_NT_07_BC
Proteasa_NT_12_AG
09/03/2023 11:27
It can contain 1 file or several.
E.g.: HIV Protease proteins divided by variant
Nombre de carpeta:
C:\
Archivos de tipo:
Todos los Archivos
Abrir
Cancelar

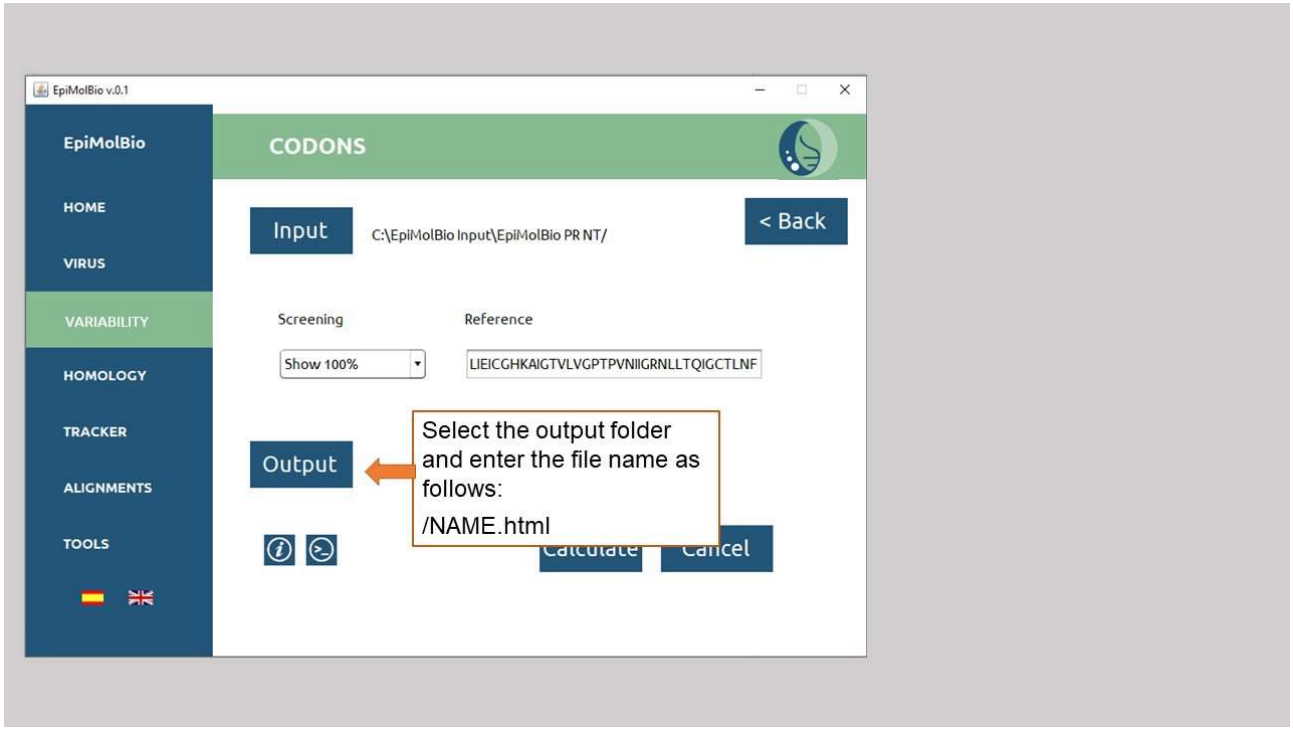
## 5)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
CODONS
Input
C:\EpiMolBio Input\EpiMolBio PR NT/
< Back
Screening
Select ...
Select ...
Show 100%
Show > 75%
Output
Output
Select screening percentage:
100% to display all the found codons.
75% to display only those with a frequency of occurrence higher than 75%.
Calculate
Cancel

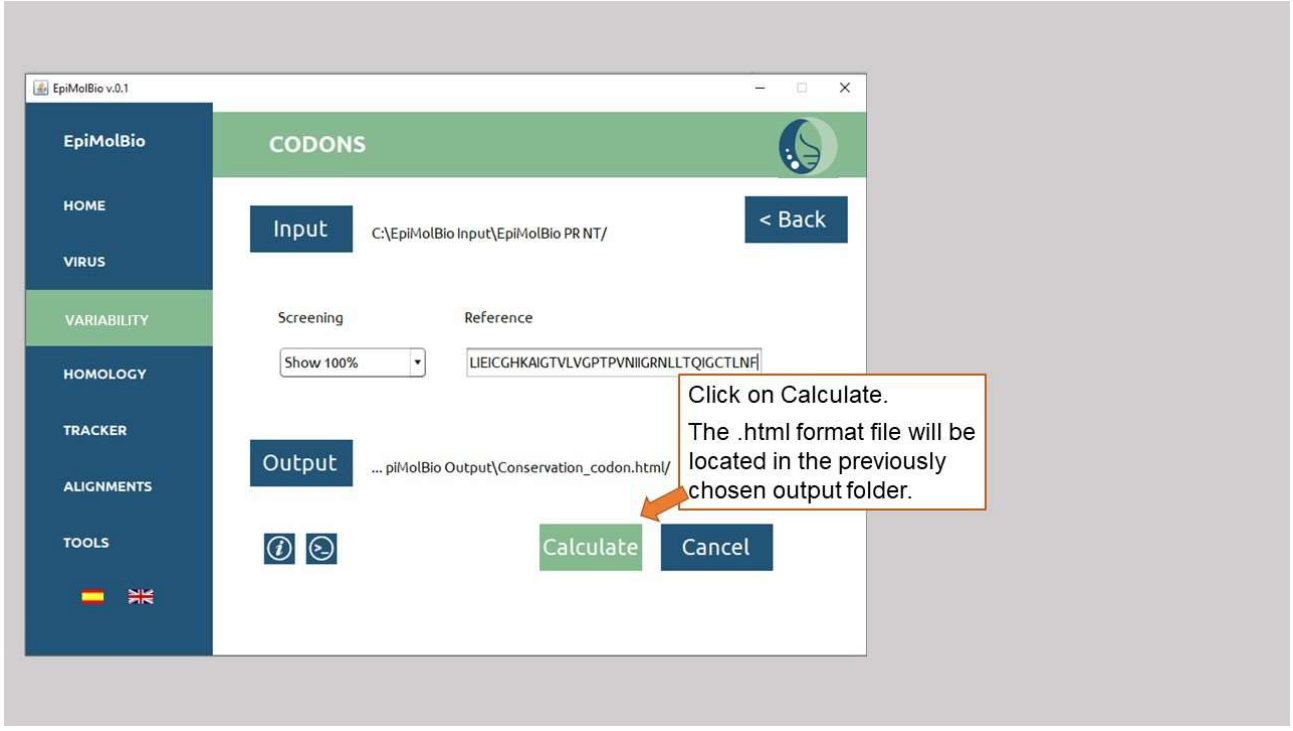
## 6)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
CODONS
Input
C:\EpiMolBio Input\EpiMolBio PR NT/
< Back
Screening
Reference
Show 100%
LIEICGHKAIGTVLVGPTPVNIIGRNLLTQIGCTLNF
Enter the reference sequence without line breaks
Output
Output
Calculate
Cancel

## 7)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
CODONS
Input
C:\EpiMolBio Input\EpiMolBio PR NT/
< Back
Screening
Reference
Show 100%
LIEICGHKAIGTVLVGPTPVNIIGRNLLTQIGCTLNF
Output
Select the output folder and enter the file name as follows:
/NAME.html
Calculate
Cancel

## 8)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
CODONS
Input
C:\EpiMolBio Input\EpiMolBio PR NT/
< Back
Screening
Reference
Show 100%
LIEICGHKAIGTVLVGPTPVNIIGRNLLTQIGCTLNF
Output
... piMolBio Output\Conservation_codon.html/
Click on Calculate.
The .html format file will be located in the previously chosen output folder.
Calculate
Cancel

# II.3. CONSENSUS

This function allows you to obtain consensus sequences from other consensus sequences generated from .fasta sequences entered as input files. Multiple rounds of successive analysis can be performed to obtain consensuses of consensuses. For example, in the first round, you can obtain consensus sequences of different variants of a virus, and in subsequent rounds, you can generate consensus sequences that encompass the previously processed variant consensuses to obtain the consensus of consensuses of the virus.

To obtain consensus sequences derived from other consensus sequences, you will need to perform multiple analyses using different rounds. Each round will involve generating a consensus sequence from the consensus sequences obtained in the previous round, and this process can be repeated iteratively.

Round 1:

The **input** format should be a folder containing only .fasta files with aligned sequences. The sequences can be either nucleotides or amino acids. To perform the analysis on nucleotide sequences, you need to use the 'Find and Replace' tool in the File Editing option to replace 'N' with '?' to exclude them from the analysis.

In the '**Select Round**' field, choose '**Round 1**'.

In the '**Reference**' field, you need to input a reference sequence without line breaks in nucleotides or amino acids, depending on the input files.

The **output** format for Round 1 will be a text file. You need to select the output folder where you want the file to appear without the need to name it. This text file will serve as input for subsequent rounds and is automatically named 'Consensus'. It is recommended to change the folder location and rename it before repeating this analysis to avoid overwriting it.

Successive Rounds:

In the **input**, we need to select a folder containing **exclusively** the **.txt file** from the previous round.

In the '**Select Round**' field, choose '**Successive Rounds**'.

In the '**Reference**' field, you need to input a reference sequence without line breaks in nucleotides or amino acids, depending on the input files.

The **output** format will be an .html file and another .txt file. You need to select the output folder where you want the files to appear, without the need to name them. Both files will be automatically named 'Consensus,' so remember to rename the files before repeating this analysis to avoid overwriting them.

If you want to create more levels of consensus, all you need to do is merge the '.txt' files from the successive rounds by copying and pasting them into a single .txt file while respecting the line breaks they contain, and then repeat the successive rounds step.

The .html file obtained in successive rounds is a table with the consensus obtained after the analysis. At the top, you will see the title of the analysis. The 'Reference' row displays the introduced reference sequence. In the next row, 'Position,' you can see the position of the

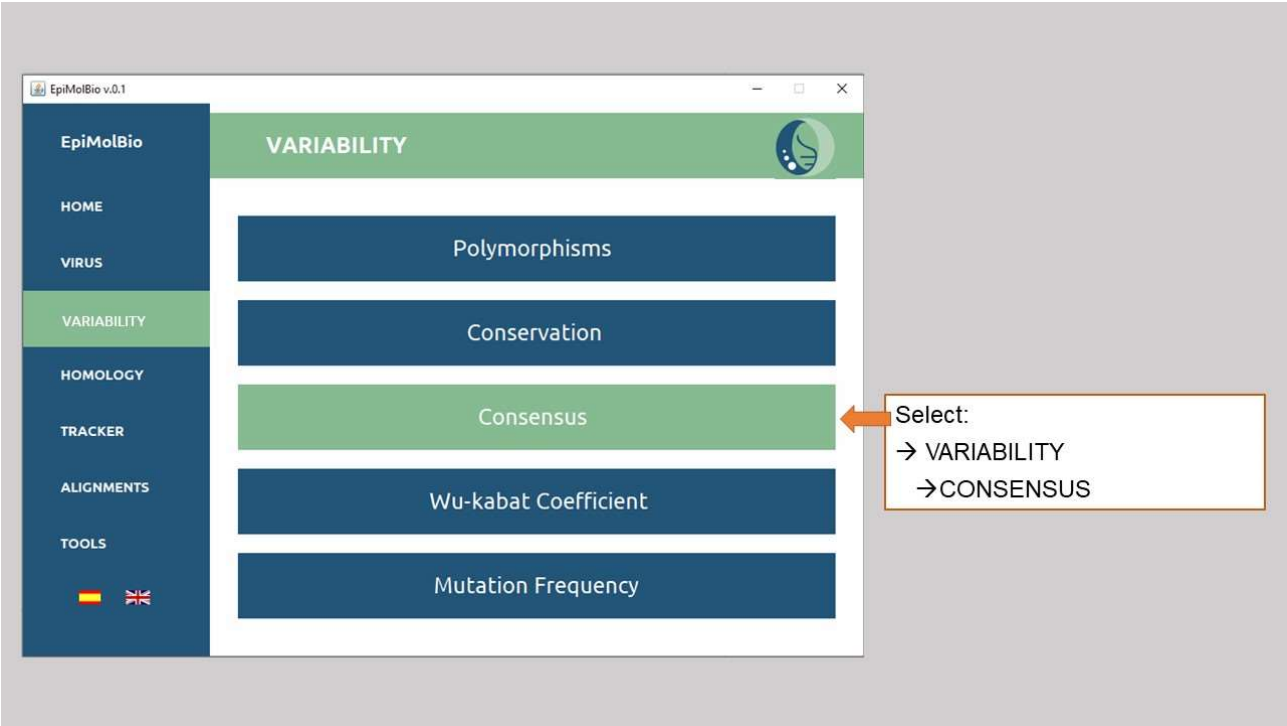
analyzed residue. In the ‘Residue’ row, the most frequent nucleotide or amino acid for each position is displayed. The ‘Conservation’ row shows the percentage of conservation. In the last row, ‘Number of Sequences,’ you can see the number of valid sequences for each position. Rows 3 and 4 are displayed with cells colored according to the color code described in the Overview section, which can be accessed in the .html output file by clicking on the blue symbol. It's worth mentioning that the loaded file is not displayed in the table of successive rounds.

Example of output format for Successive Consensus Rounds:

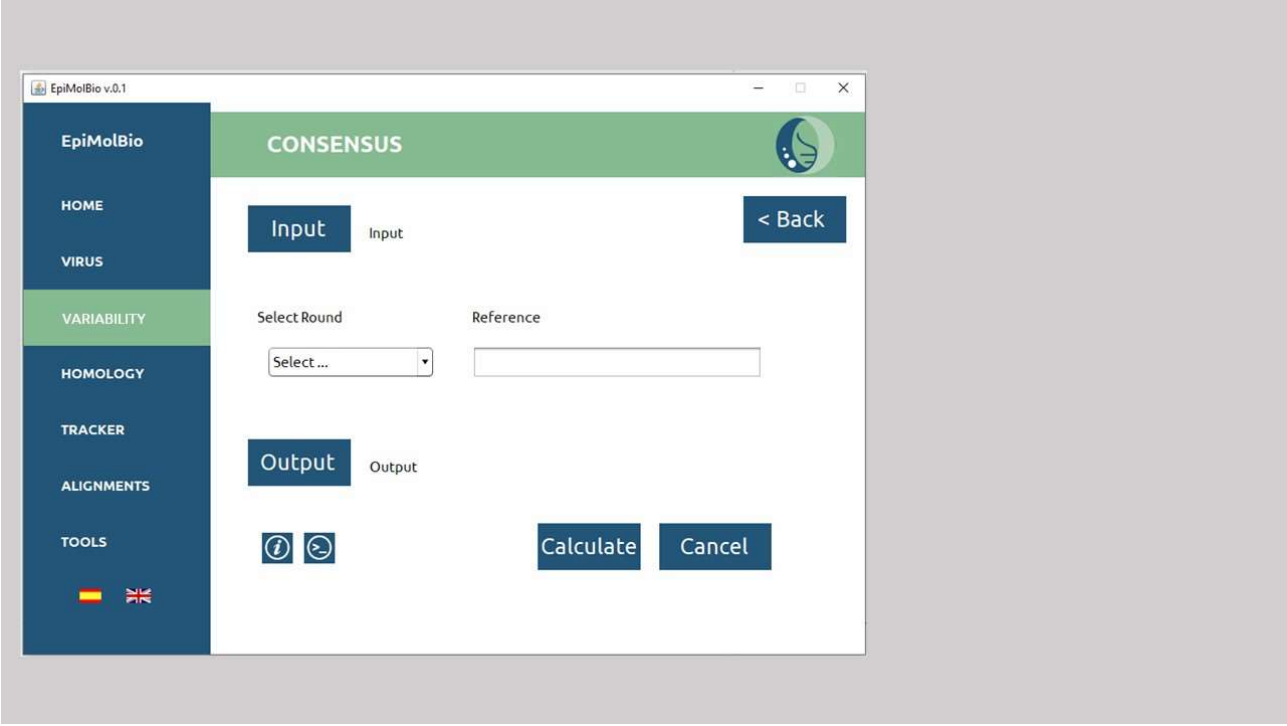
| | Variability Consensus | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Reference | P | Q | V | T | L | W | Q | R | P | L | V | T | I | K | I |
| Position | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
| Residue | **P** | **Q** | **I** | **T** | **L** | **W** | **Q** | **R** | **P** | **L** | **V** | **T** | **I** | **K** | **I** |
| Conservation | **99.867** | **99.880** | **99.037** | **98.725** | **99.810** | **99.946** | **98.497** | **99.882** | **99.987** | **76.744** | **97.137** | **78.090** | **50.465** | **74.586** | **61.991** |
| Number of Sequences | 146 | 146 | 146 | 146 | 146 | 146 | 146 | 146 | 146 | 146 | 146 | 146 | 146 | 146 | 146 |

# Step-by-step:

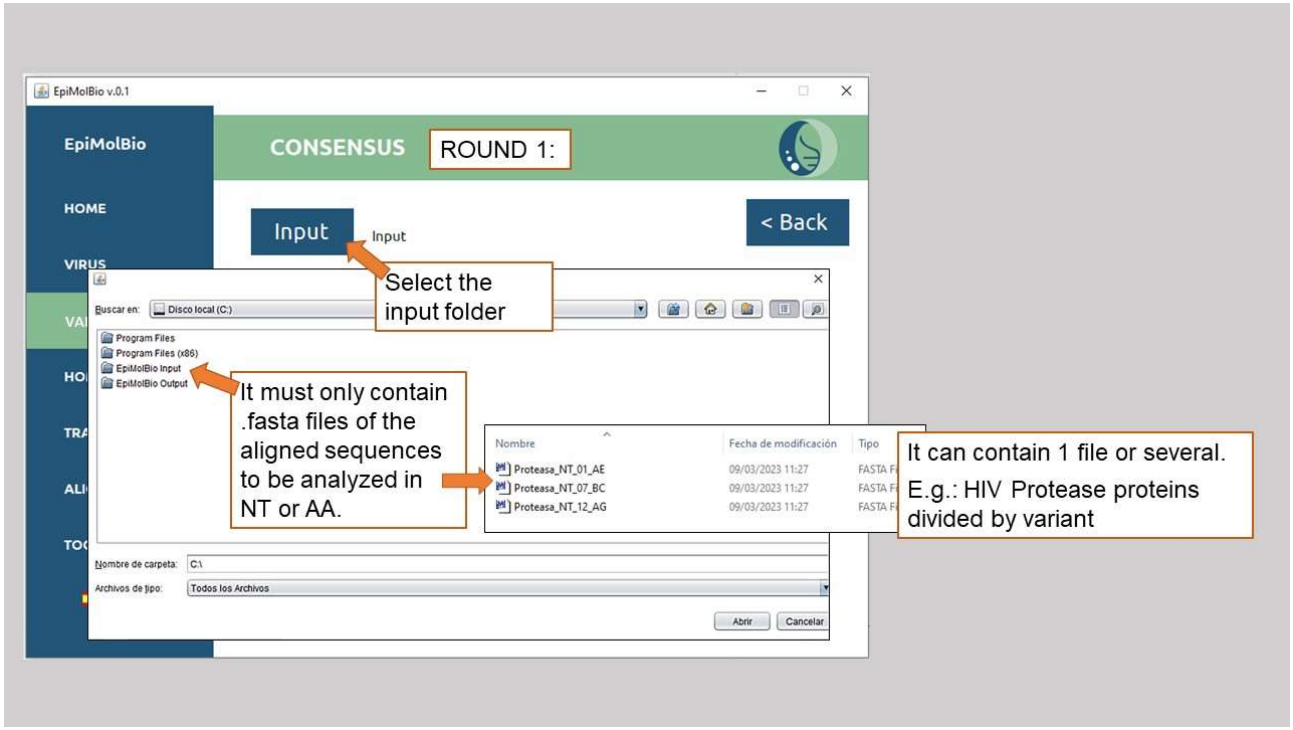
## 1)



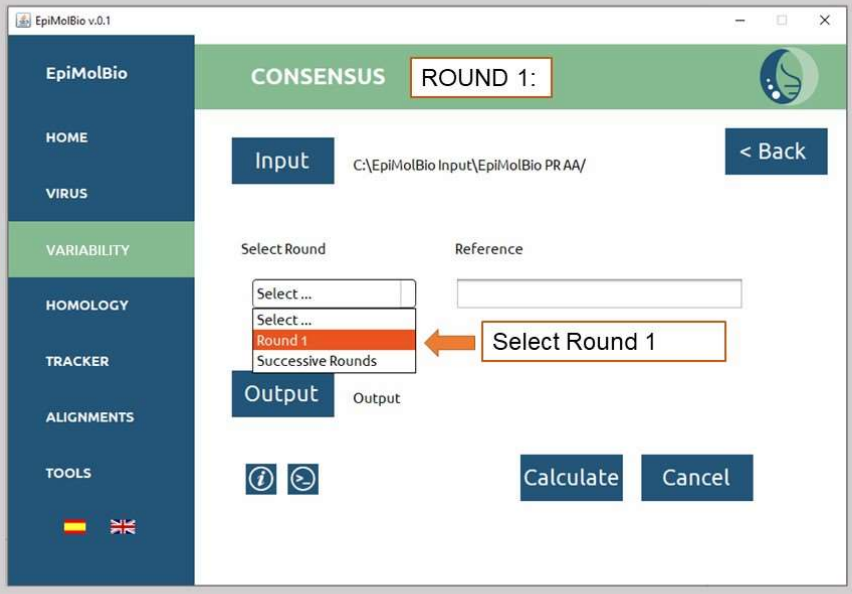

## 2)



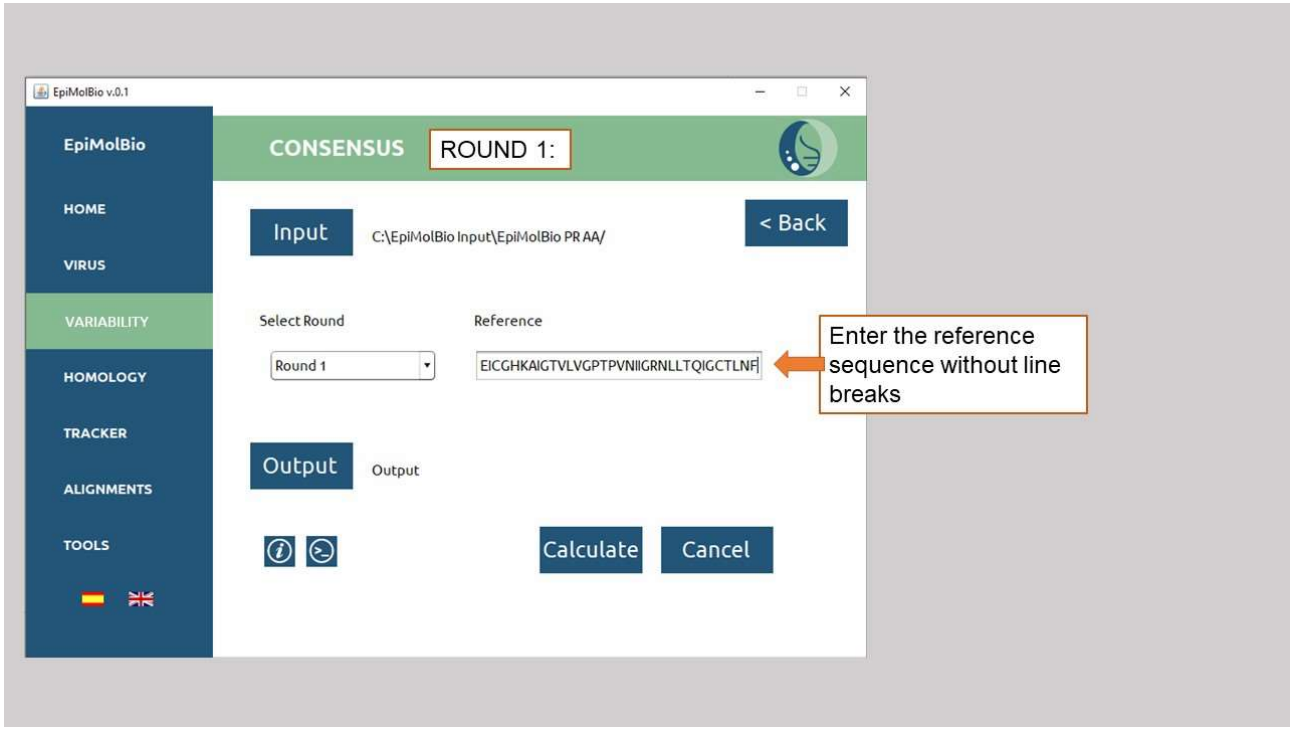
## 3)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
CONSENSUS
ROUND 1:
Input
Input
< Back
Select the input folder
Buscar en:
Disco local (C:)
Program Files
Program Files (x86)
EpiMolBio Input
EpiMolBio Output
It must only contain .fasta files of the aligned sequences to be analyzed in NT or AA.
Nombre
Fecha de modificación
Tipo
Proteasa_NT_01_AE
Proteasa_NT_07_BC
Proteasa_NT_12_AG
09/03/2023 11:27
It can contain 1 file or several.
E.g.: HIV Protease proteins divided by variant
Nombre de carpeta:
C:\
Archivos de tipo:
Todos los Archivos
Abrir
Cancelar

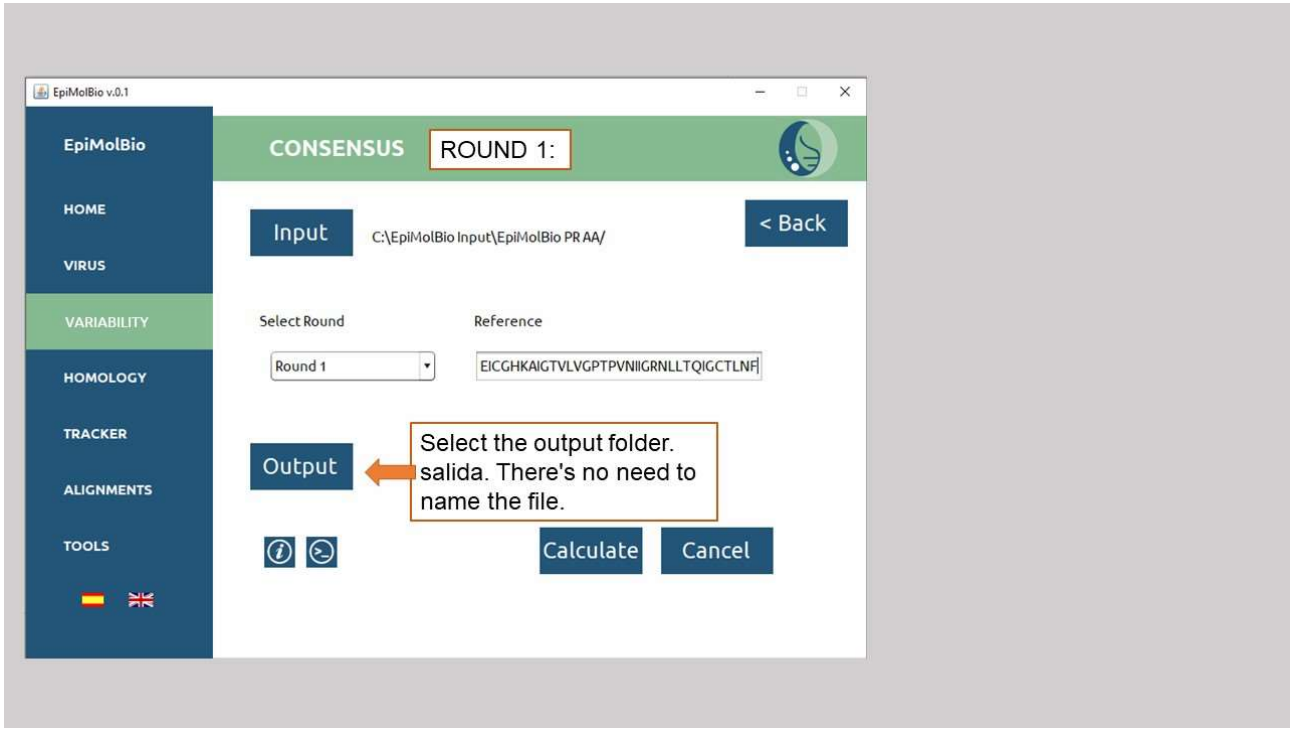
## 4)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
CONSENSUS
ROUND 1:
Input
C:\EpiMolBio Input\EpiMolBio PR AA/
< Back
Select Round
Reference
Select ...
Select ...
Round 1
Successive Rounds
Select Round 1
Output
Output
Calculate
Cancel

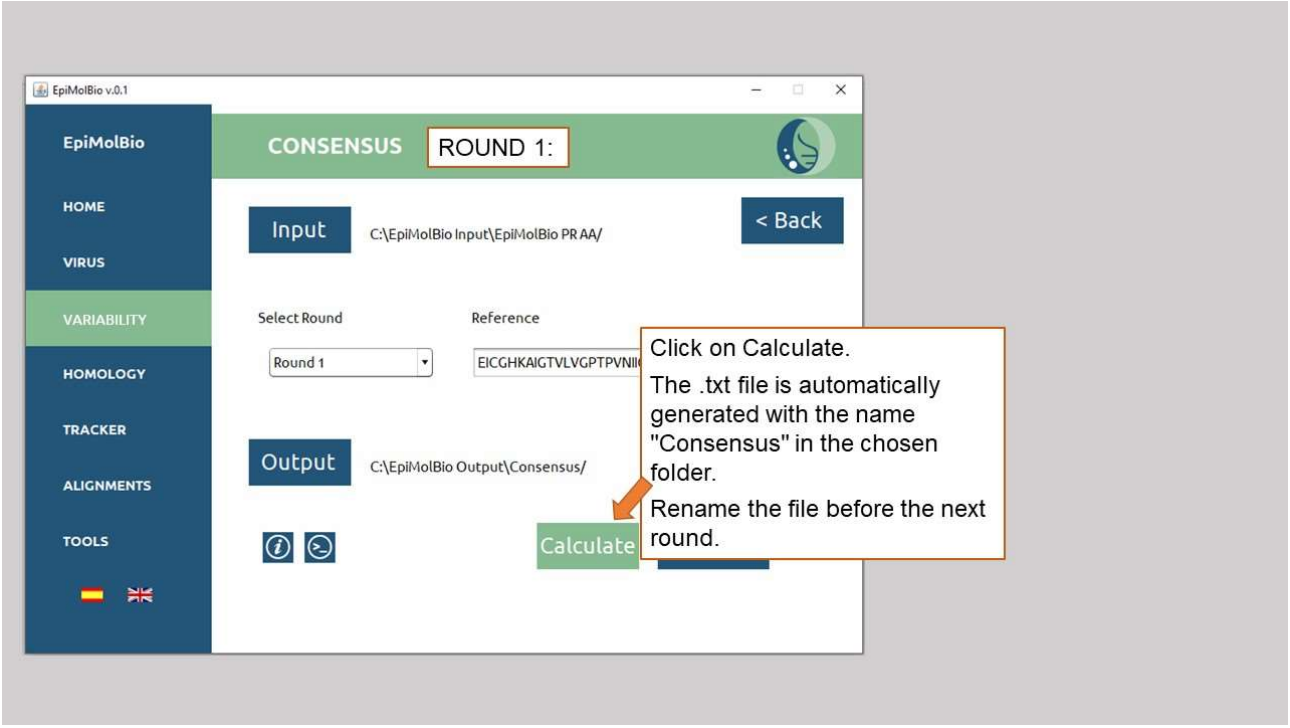
## 5)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
CONSENSUS
ROUND 1:
Input
C:\EpiMolBio Input\EpiMolBio PR AA/
< Back
Select Round
Reference
Round 1
EICGHKAIGTVLVGPTPVNIIGRNLLTQIGCTLNF
Enter the reference sequence without line breaks
Output
Output
Calculate
Cancel

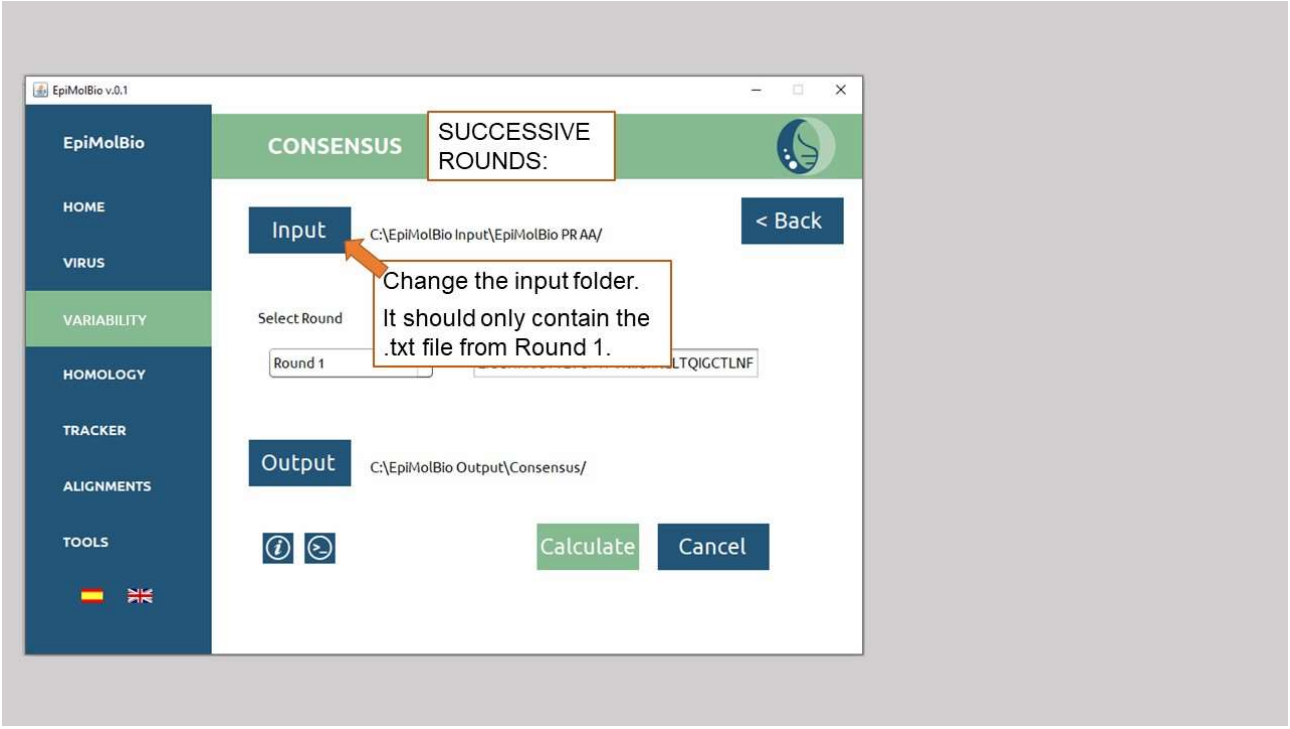
## 6)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
CONSENSUS
ROUND 1:
Input
C:\EpiMolBio Input\EpiMolBio PR AA/
< Back
Select Round
Reference
Round 1
EICGHKAIGTVLVGPTPVNIIGRNLLTQIGCTLNF
Output
Select the output folder. salida. There's no need to name the file.
Calculate
Cancel

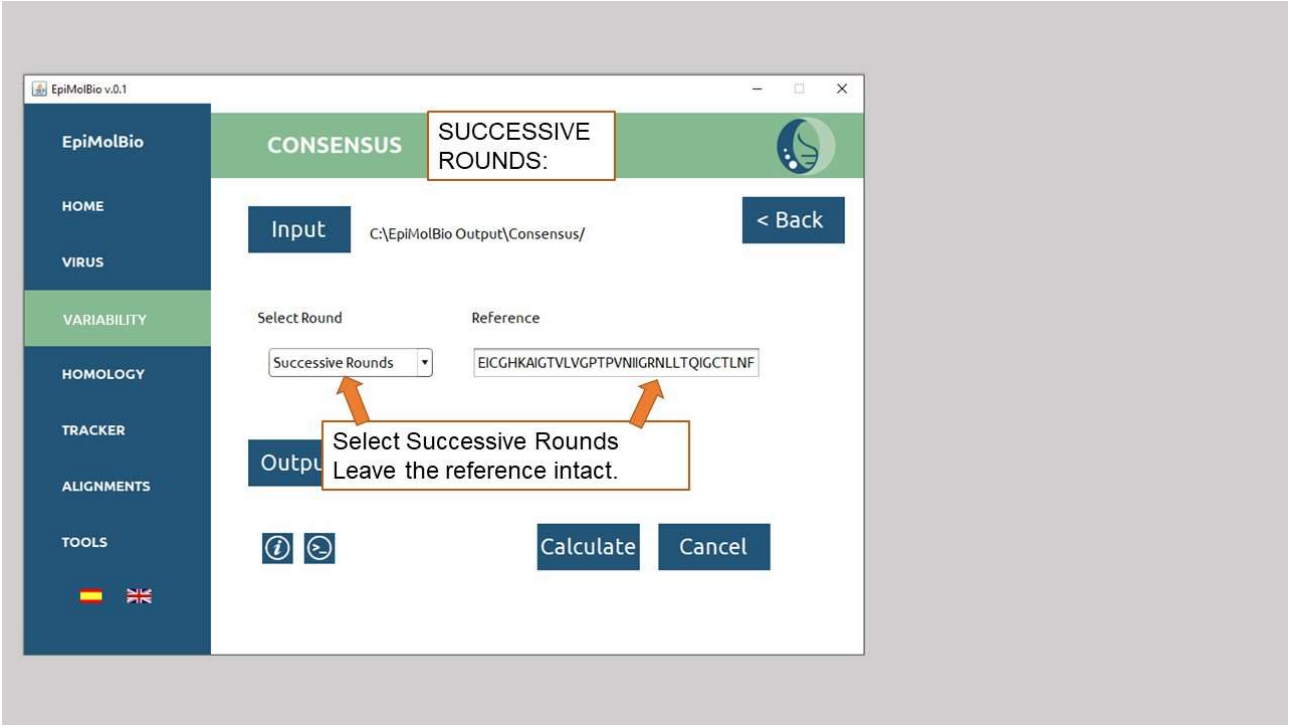
## 7)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
CONSENSUS
ROUND 1:
Input
C:\EpiMolBio Input\EpiMolBio PR AA/
< Back
Select Round
Reference
Round 1
Output
C:\EpiMolBio Output\Consensus/
Calculate
Click on Calculate.
The .txt file is automatically generated with the name "Consensus" in the chosen folder.
Rename the file before the next round.

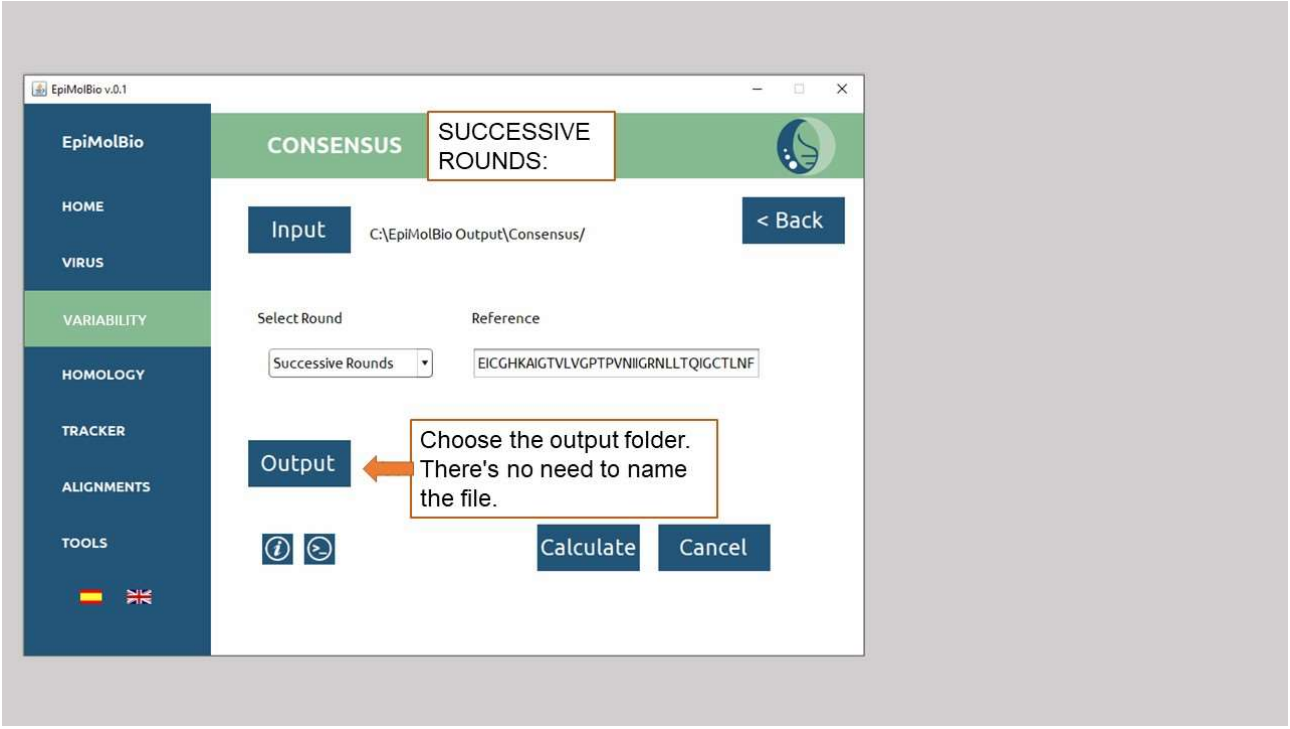
## 8)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
CONSENSUS
SUCCESSIVE ROUNDS:
Input
C:\EpiMolBio Input\EpiMolBio PR AA/
< Back
Change the input folder.
It should only contain the .txt file from Round 1.
Select Round
Round 1
Output
C:\EpiMolBio Output\Consensus/
Calculate
Cancel

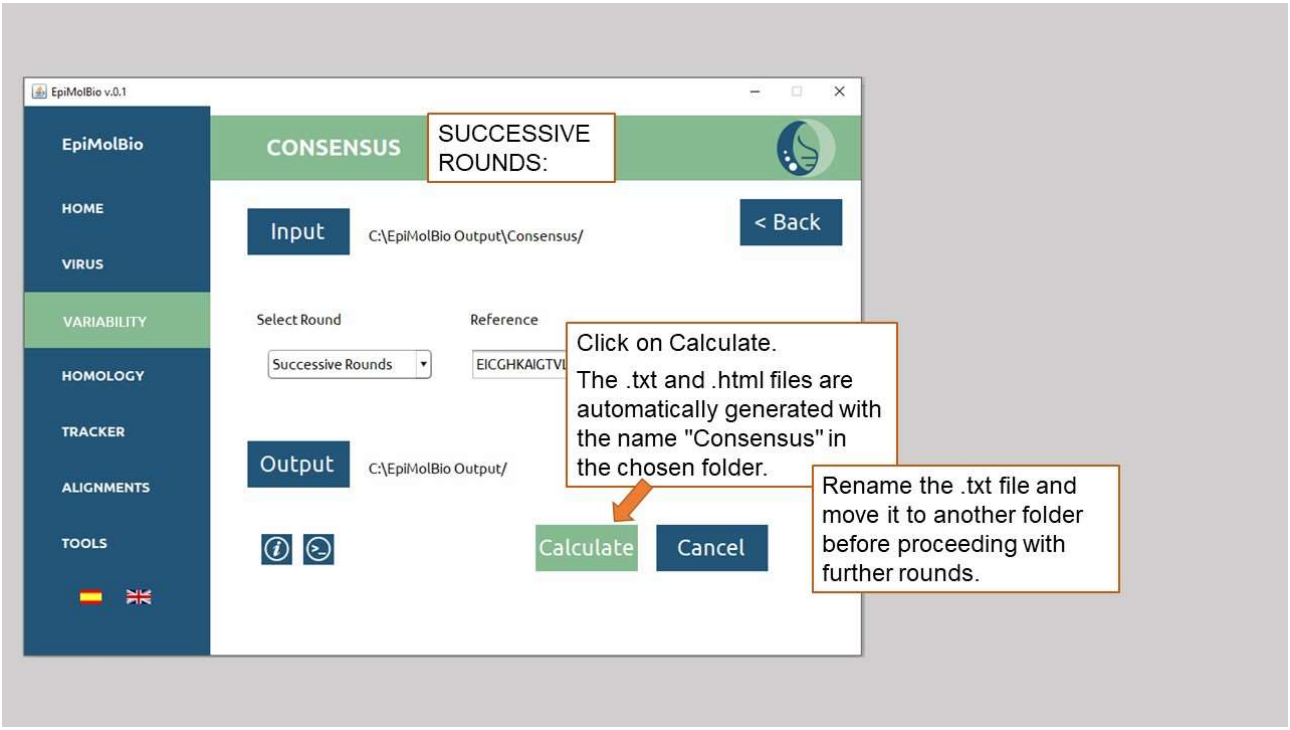
**9)**


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
CONSENSUS
SUCCESSIVE ROUNDS:
Input
C:\EpiMolBio Output\Consensus/
< Back
Select Round
Reference
Successive Rounds
EICGHKAIGTVLVGPTPVNIIGRNLLTQIGCTLNF
Select Successive Rounds
Leave the reference intact.
Outpu
Calculate
Cancel

**10)**


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
CONSENSUS
SUCCESSIVE ROUNDS:
Input
C:\EpiMolBio Output\Consensus/
< Back
Select Round
Reference
Successive Rounds
EICGHKAIGTVLVGPTPVNIIGRNLLTQIGCTLNF
Output
Choose the output folder.
There's no need to name
the file.
Calculate
Cancel

## 11)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
CONSENSUS
SUCCESSIVE ROUNDS:
Input
C:\EpiMolBio Output\Consensus/
< Back
Select Round
Reference
Successive Rounds
Output
C:\EpiMolBio Output/
Calculate
Cancel
Click on Calculate.
The .txt and .html files are automatically generated with the name "Consensus" in the chosen folder.
Rename the .txt file and move it to another folder before proceeding with further rounds.

# II.4.WU-KABAT COEFFICIENT

With this function, you can obtain the Wu-Kabat variability coefficient (WK) of protein sequences. The WK coefficient allows studying the susceptibility of an amino acid position to evolutionary replacements (Kabat et al., 1977). It is calculated using the following formula:

$$Variability = Nk/n$$

Where N is the number of sequences in the alignment, k is the number of different amino acids at a specific position, and n is the number of times the most frequent amino acid is repeated at that position.

Therefore, a WK of 1 indicates that the same amino acid was found at that position in the entire set of sequences, while a WK >1 indicates relative variability at the respective site, with greater diversity as the WK value increases.

The **input** file must be a folder containing exclusively .fasta files with aligned amino acid sequences. The sequences should be in amino acids.

In the '**Length**' field, you should enter the length in amino acids of the protein to be analyzed.

The **output** format will be a .csv file. You should select the output folder where you want the result to appear and name the file by adding .csv at the end.

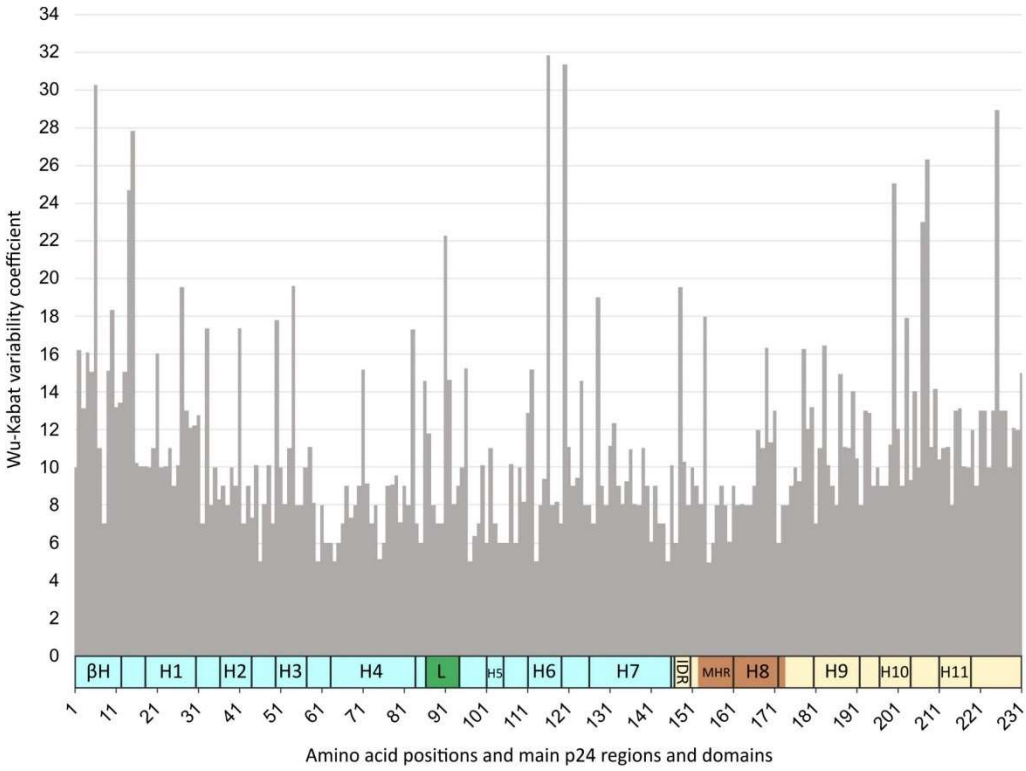
The output file is a table that can be opened in Excel. The table shows, in the first column, the names of the input files, in the second column, the position of each analyzed residue, in the third column, the Wu-Kabat index, in the fourth column, the number of valid sequences for that position, in the fifth column, the number of different amino acids at that position, and in the sixth column, the absolute frequency of the most frequent amino acid for that position.

Example of Wu-Kabat Coefficient output:

| | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| 1 | File | Position | Wu-Kabat | Number of Sequences | Number of Amino Acids | Frequency |
| 2 | PR_01_AE.fa | 1 | 7.007 | 26838 | 7 | 26810 |
| 3 | PR_01_AE.fa | 2 | 11.024 | 26649 | 11 | 26591 |
| 4 | PR_01_AE.fa | 3 | 5.007 | 26831 | 5 | 26793 |
| 5 | PR_01_AE.fa | 4 | 9.013 | 26816 | 9 | 26778 |
| 6 | PR_01_AE.fa | 5 | 7.008 | 26780 | 7 | 26750 |
| 7 | PR_01_AE.fa | 6 | 5.004 | 26836 | 5 | 26817 |
| 8 | PR_01_AE.fa | 7 | 10.025 | 26536 | 10 | 26470 |
| 9 | PR_01_AE.fa | 8 | 5.004 | 26792 | 5 | 26771 |
| 10 | PR_01_AE.fa | 9 | 6.003 | 26613 | 6 | 26601 |
| 11 | PR_01_AE.fa | 10 | 14.866 | 25952 | 13 | 22694 |
| 12 | PR_01_AE.fa | 11 | 9.044 | 26416 | 9 | 26287 |
| 13 | PR_01_AE.fa | 12 | 12.628 | 26469 | 12 | 25153 |
| 14 | PR_01_AE.fa | 13 | 17.825 | 26150 | 10 | 14670 |
| 15 | PR_01_AE.fa | 14 | 13.657 | 26270 | 13 | 25006 |

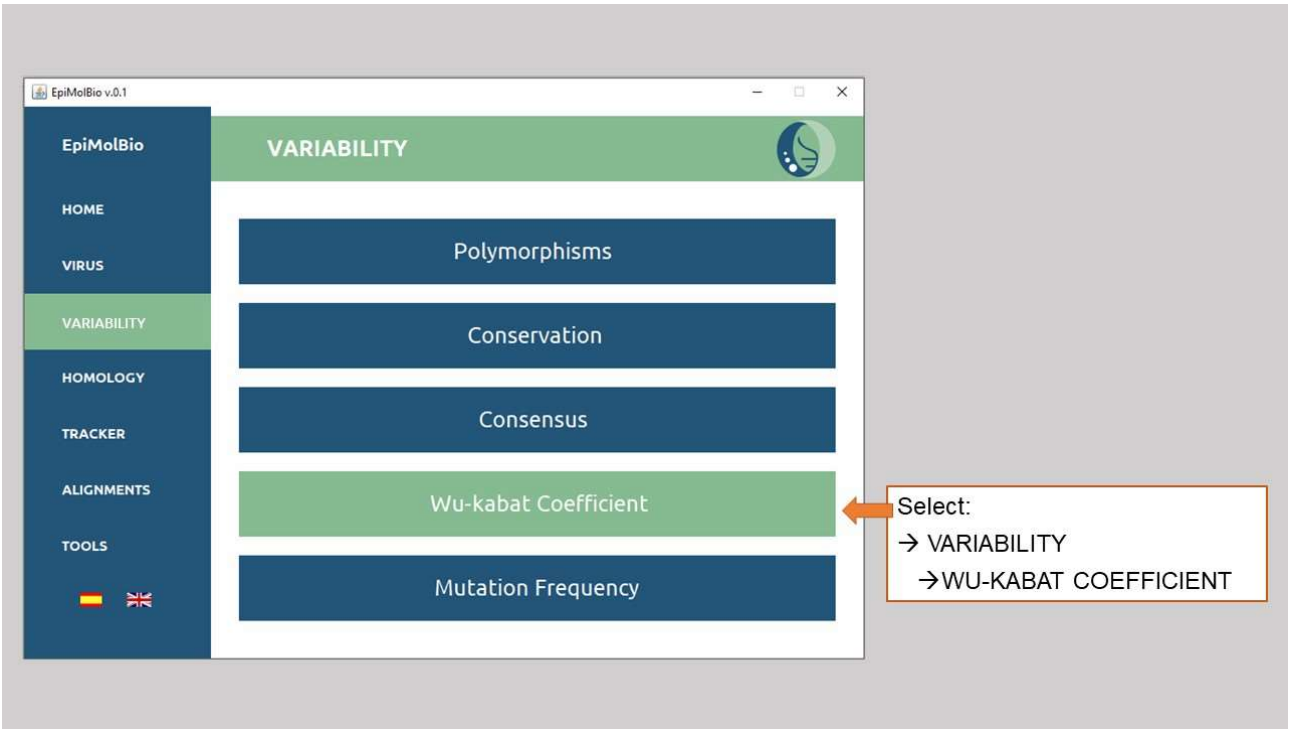
With columns 2 and 3 (Position and Wu-Kabat) of the output table, you can create a graph to visualize the Wu-Kabat variability coefficient of a protein.

Example: Diagram of the Wu-Kabat variability coefficient in HIV Capsid protein p24 sequences of HIV-1 group M (Troyano-Hernáez P, Reinosa R, Holguín Á. HIV Capsid Protein Genetic Diversity Across HIV-1 Variants and Impact on New Capsid-Inhibitor Lenacapavir. Front Microbiol. 2022 Apr 12;13:854974. doi: 10.3389/fmicb.2022.854974. PMID: 35495642; PMCID: PMC9039614)

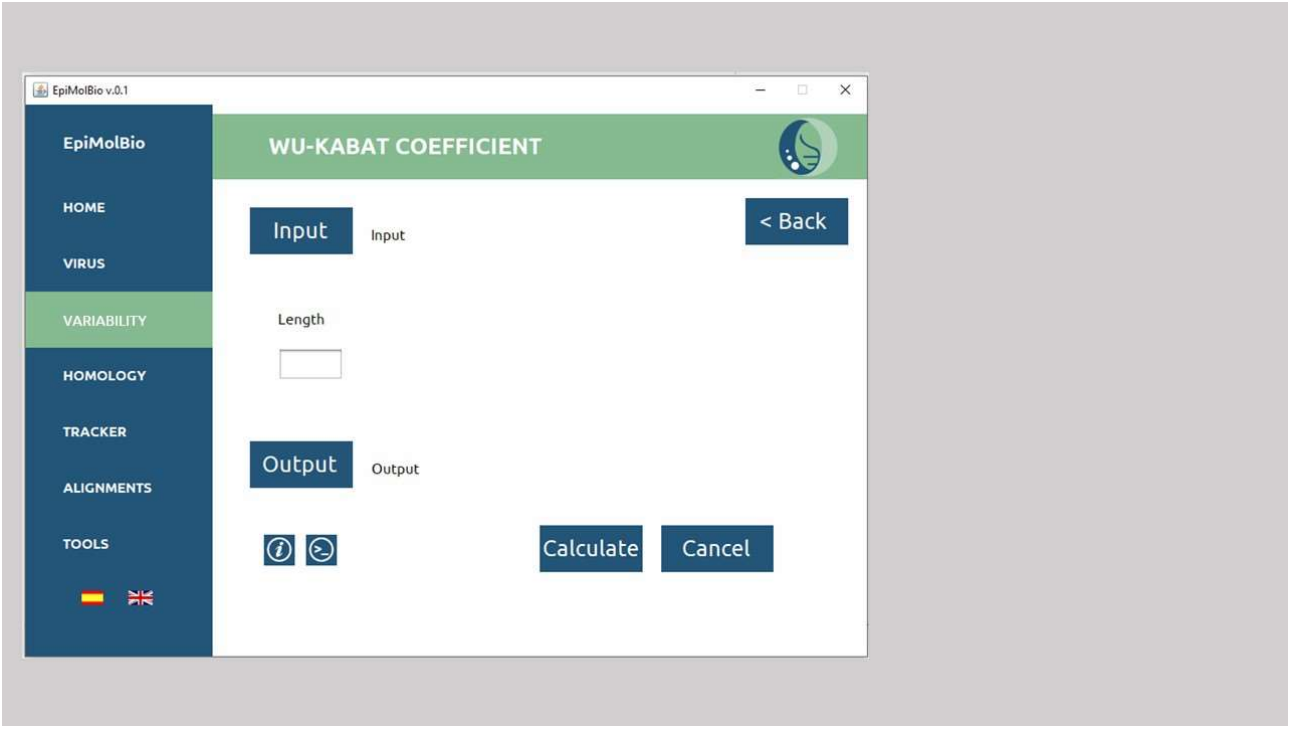


# Step-by-step:

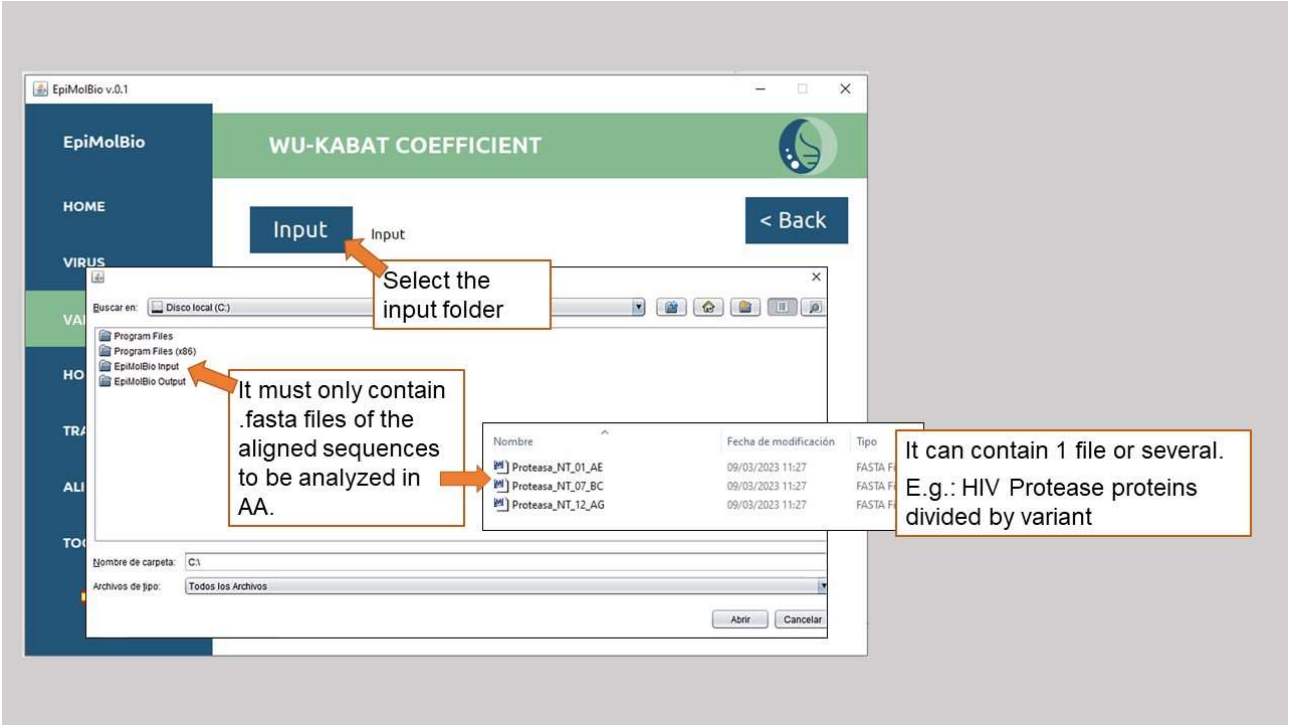
## 1)



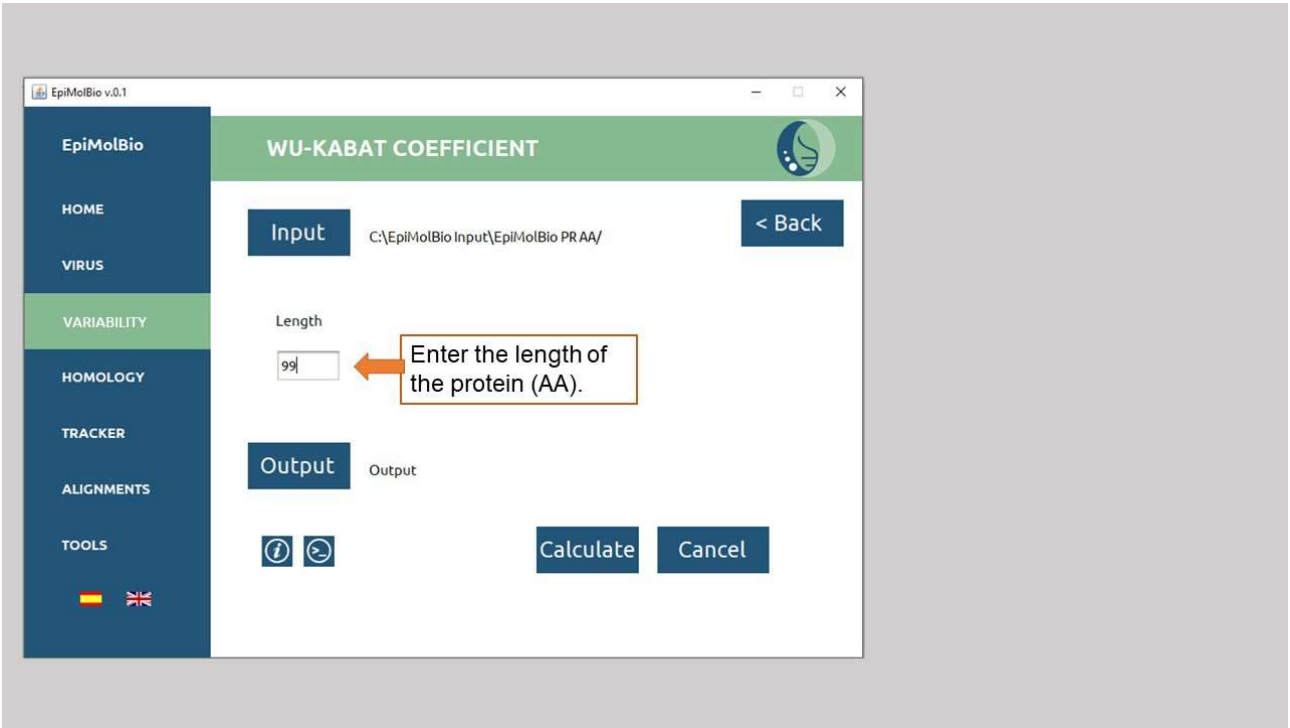

## 2)



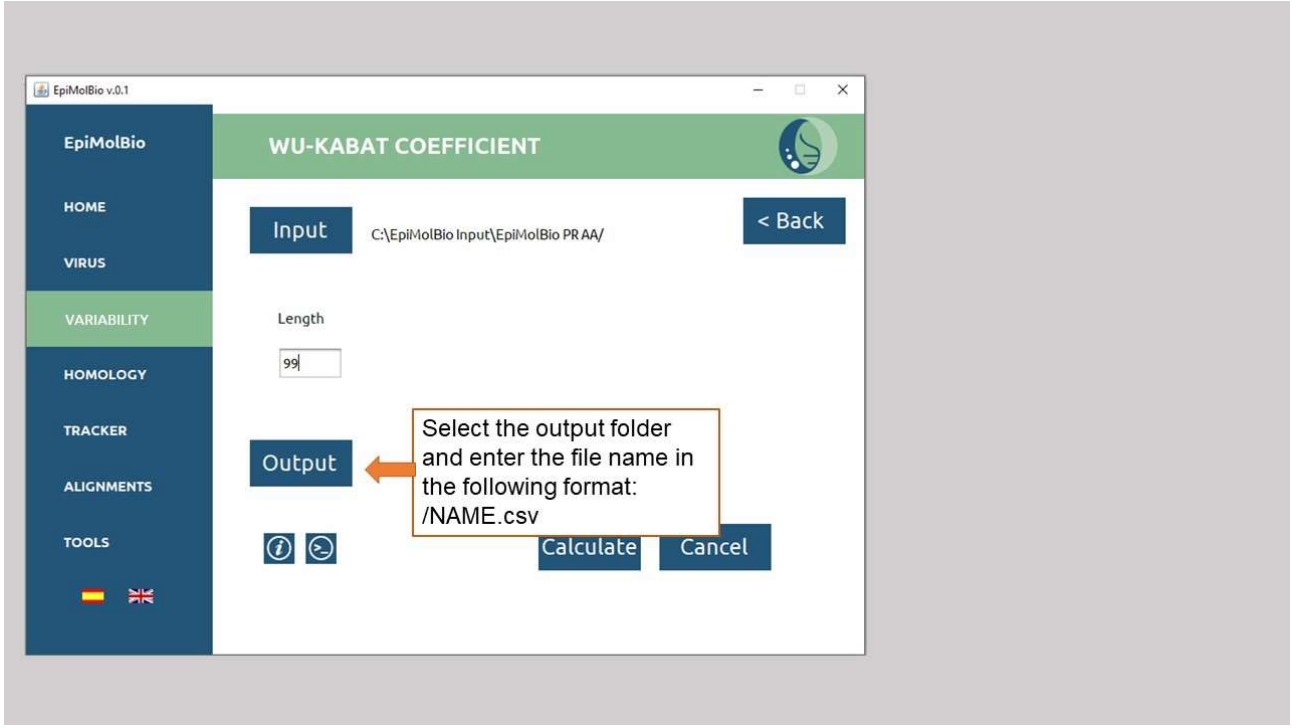
## 3)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
WU-KABAT COEFFICIENT
Input
Input
< Back
Select the input folder
Buscar en:
Disco local (C:)
Program Files
Program Files (x86)
EpiMolBio Input
EpiMolBio Output
It must only contain .fasta files of the aligned sequences to be analyzed in AA.
Nombre
Fecha de modificación
Tipo
Proteasa_NT_01_AE
Proteasa_NT_07_BC
Proteasa_NT_12_AG
09/03/2023 11:27
It can contain 1 file or several.
E.g.: HIV Protease proteins divided by variant
Nombre de carpeta:
C:\
Archivos de tipo:
Todos los Archivos
Abrir
Cancelar

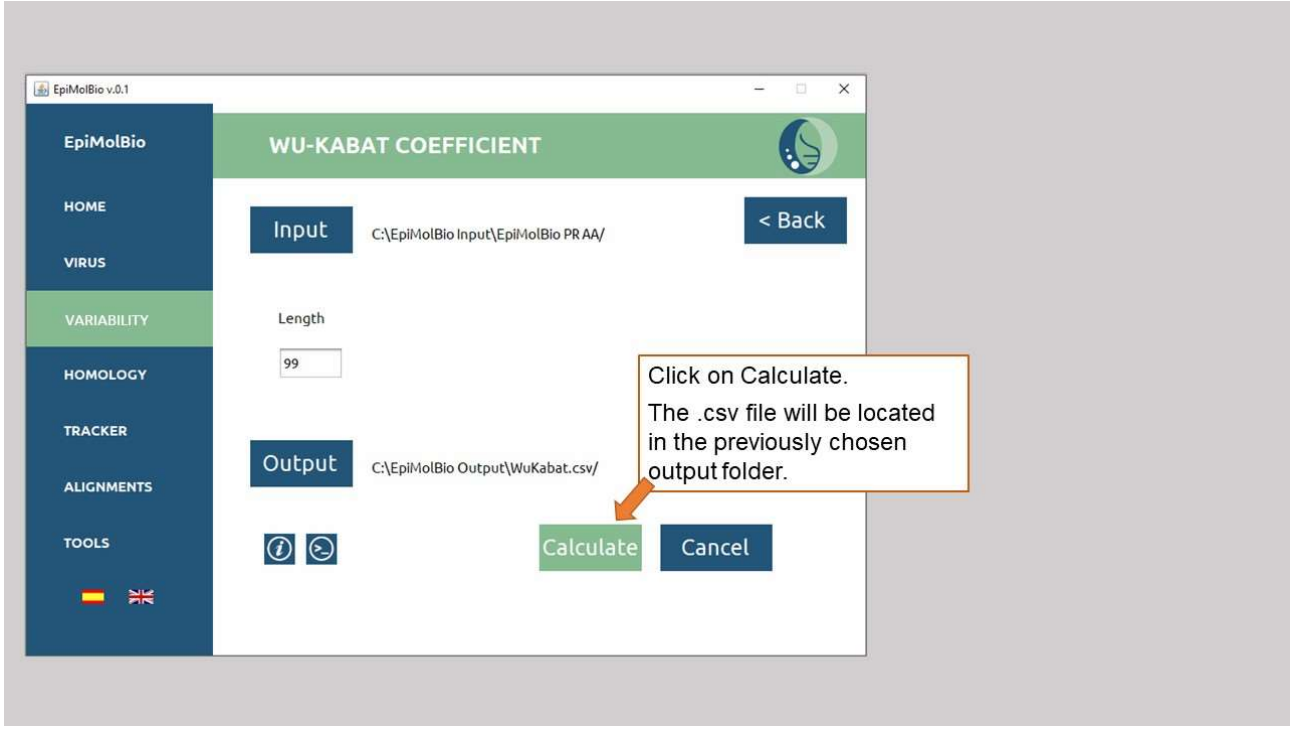
## 4)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
WU-KABAT COEFFICIENT
Input
C:\EpiMolBio Input\EpiMolBio PR AA/
< Back
Length
99
Enter the length of the protein (AA).
Output
Output
Calculate
Cancel

## 5)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
WU-KABAT COEFFICIENT
Input
C:\EpiMolBio Input\EpiMolBio PR AA/
< Back
Length
99
Output
Select the output folder and enter the file name in the following format: /NAME.csv
Calculate
Cancel

## 6)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
WU-KABAT COEFFICIENT
Input
C:\EpiMolBio Input\EpiMolBio PR AA/
< Back
Length
99
Output
C:\EpiMolBio Output\WuKabat.csv/
Click on Calculate.
The .csv file will be located in the previously chosen output folder.
Calculate
Cancel

# II.5. MUTATION FREQUENCY

With this function, you can obtain a series of parameters related to the frequency of mutations in a group of nucleotide or amino acid sequences. This analysis ignores gaps and missing residues 'N' when the input file is in nucleotides and ignores gaps, question marks '?', and stop codons '*' when the input file is in amino acids.

The analyzed parameters are as follows:

**Mutation frequency**: Number of mutated residues / Total valid positions.

**Mutation frequency percentage**: Mutation frequency x 100.

**Conservation percentage**: 100 - mutation frequency percentage.

**Average mutations per sequence**: Number of mutated residues / Total sequences in the input file.

The format of the **input** file should be a folder containing exclusively the .fasta files with the sequences to be analyzed. The sequences can be aligned or unaligned, in case you want to detect insertions and/or deletions.

Check the box '**Nucleotides**' or '**Amino Acids**' depending on whether the input sequences are translated or not.

Check the '**Align**' field when the input sequences are not aligned. The program will perform an automatic alignment with respect to the reference sequence to perform the calculations correctly.

In the '**Reference**' field, you should enter a reference sequence without line breaks, in nucleotides or amino acids, according to the input files.

The **output** format will be a .csv file. You should select the output folder where you want the .csv file to appear and name the file by writing .csv at the end.

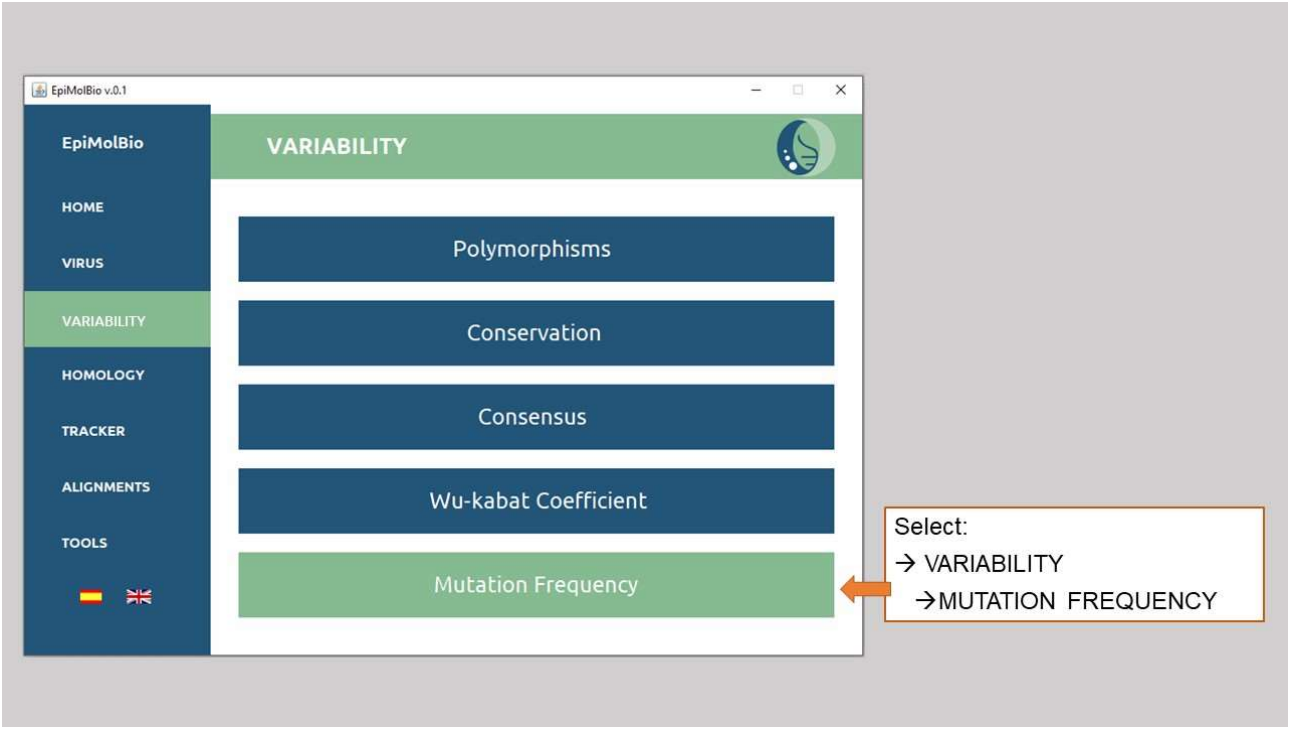
The output .csv file consists of a table that can be opened in Excel. The table contains the following columns: file name of the input sequence, mutation frequency, mutation frequency percentage, conservation percentage, value of average mutations per sequence, and total number of sequences analyzed per file.

Example of Mutation Frequency output format:

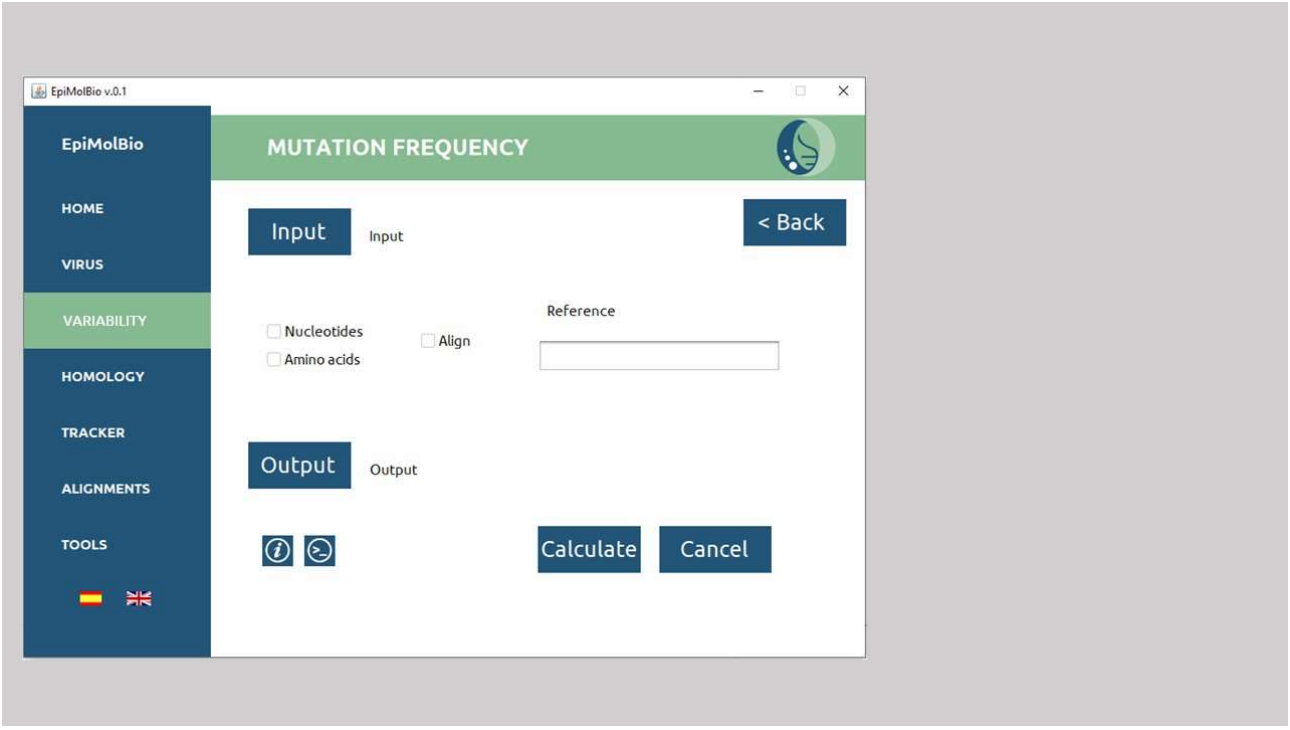
| | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| 1 | File | Mut. Frequency | Mut. Frequency % | Conservation % | Average Mut. Sec. | Total Sec. |
| 2 | 1-2022-AS.fasta | 0.10456 | 10.46% | 89.54% | 10.138 | 94 |
| 3 | 1-2022_CL.fasta | 0.01139 | 1.14% | 98.86% | 0.84 | 287 |

# Step-by-step:

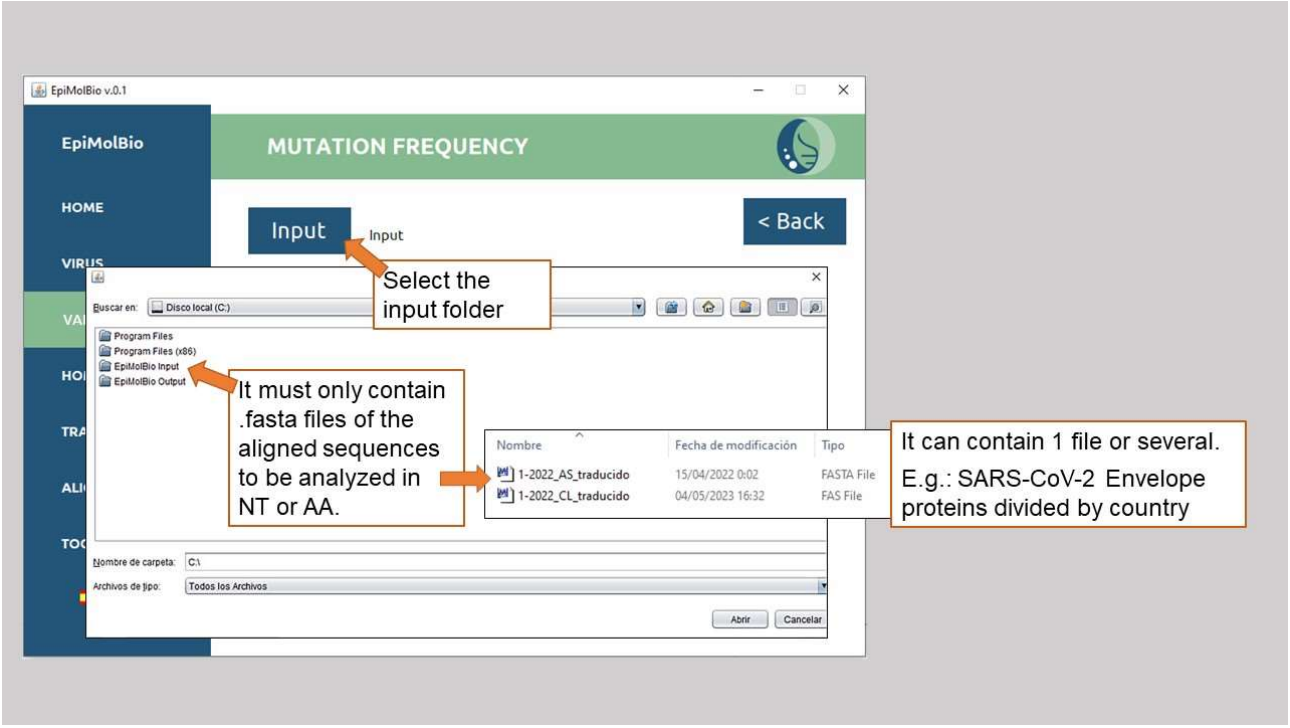
## 1)



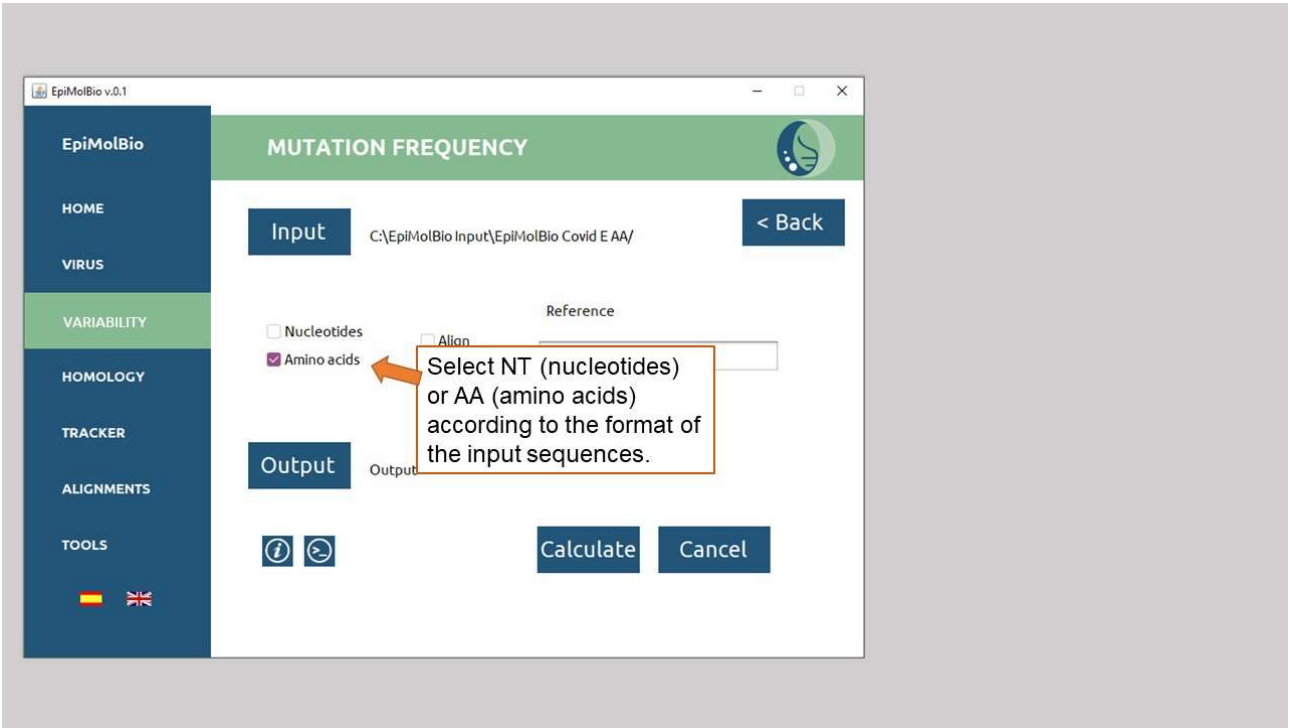

## 2)



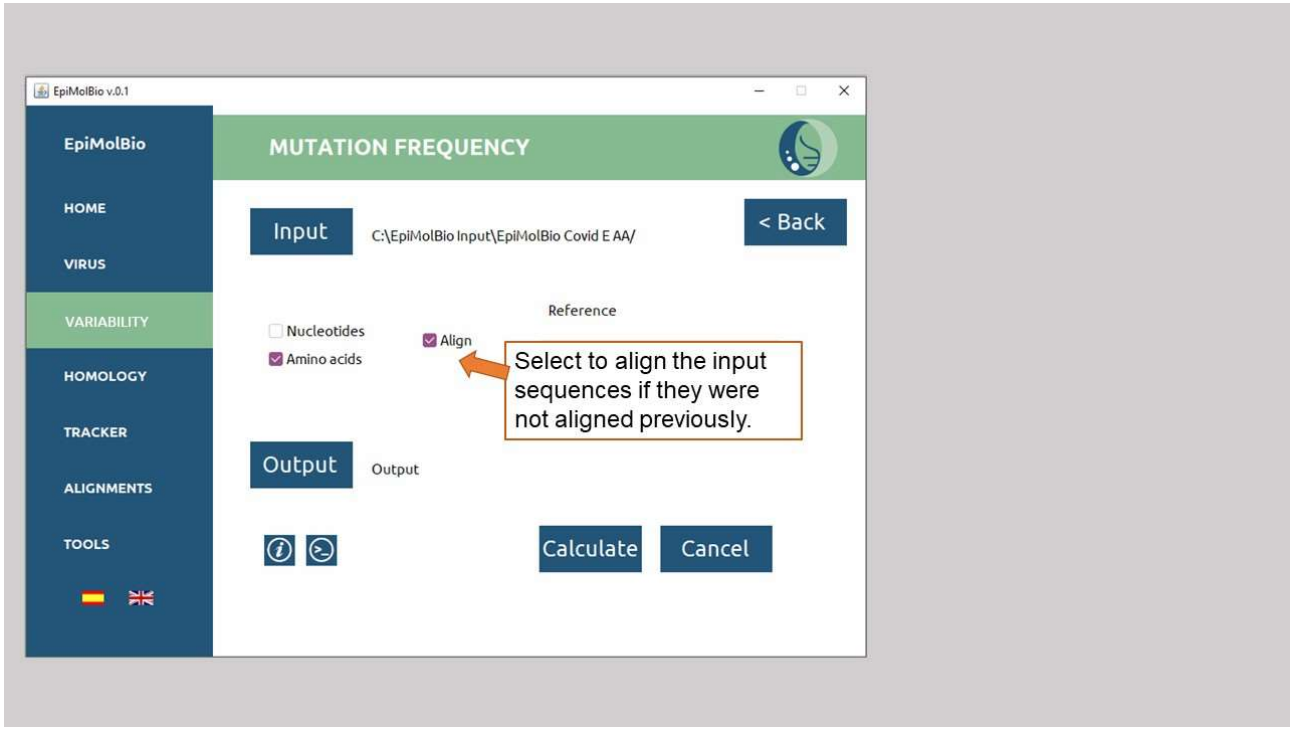
## 3)


EpiMolBio v.0.1
EpiMolBio
HOME
MUTATION FREQUENCY
Input
Input
< Back
Select the input folder
Buscar en:
Disco local (C:)
Program Files
Program Files (x86)
EpiMolBio Input
EpiMolBio Output
It must only contain .fasta files of the aligned sequences to be analyzed in NT or AA.
Nombre
Fecha de modificación
Tipo
1-2022_AS_traducido
15/04/2022 0:02
FASTA File
1-2022_CL_traducido
04/05/2023 16:32
FAS File
It can contain 1 file or several.
E.g.: SARS-CoV-2 Envelope proteins divided by country
Nombre de carpeta:
C:\
Archivos de tipo:
Todos los Archivos
Abrir
Cancelar

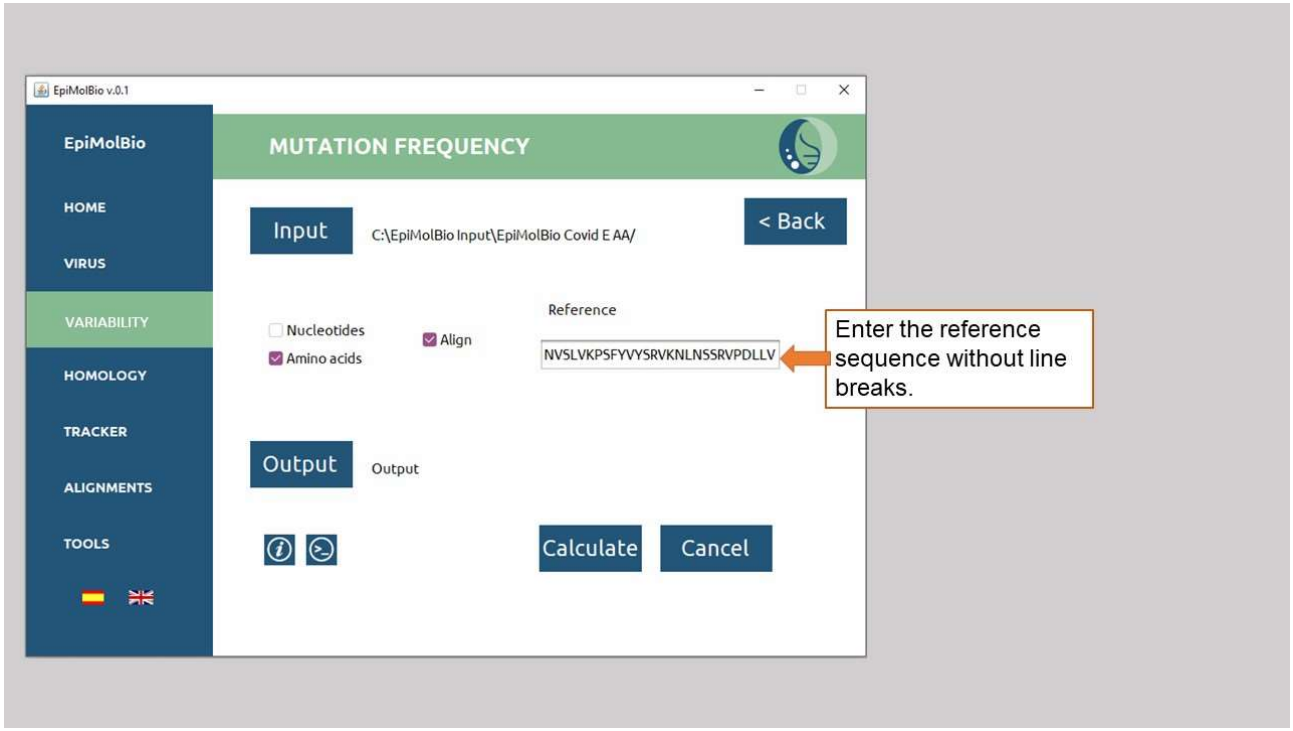
## 4)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
MUTATION FREQUENCY
Input
C:\EpiMolBio Input\EpiMolBio Covid E AA/
< Back
Reference
Nucleotides
Align
Amino acids
Select NT (nucleotides) or AA (amino acids) according to the format of the input sequences.
Output
Output
Calculate
Cancel

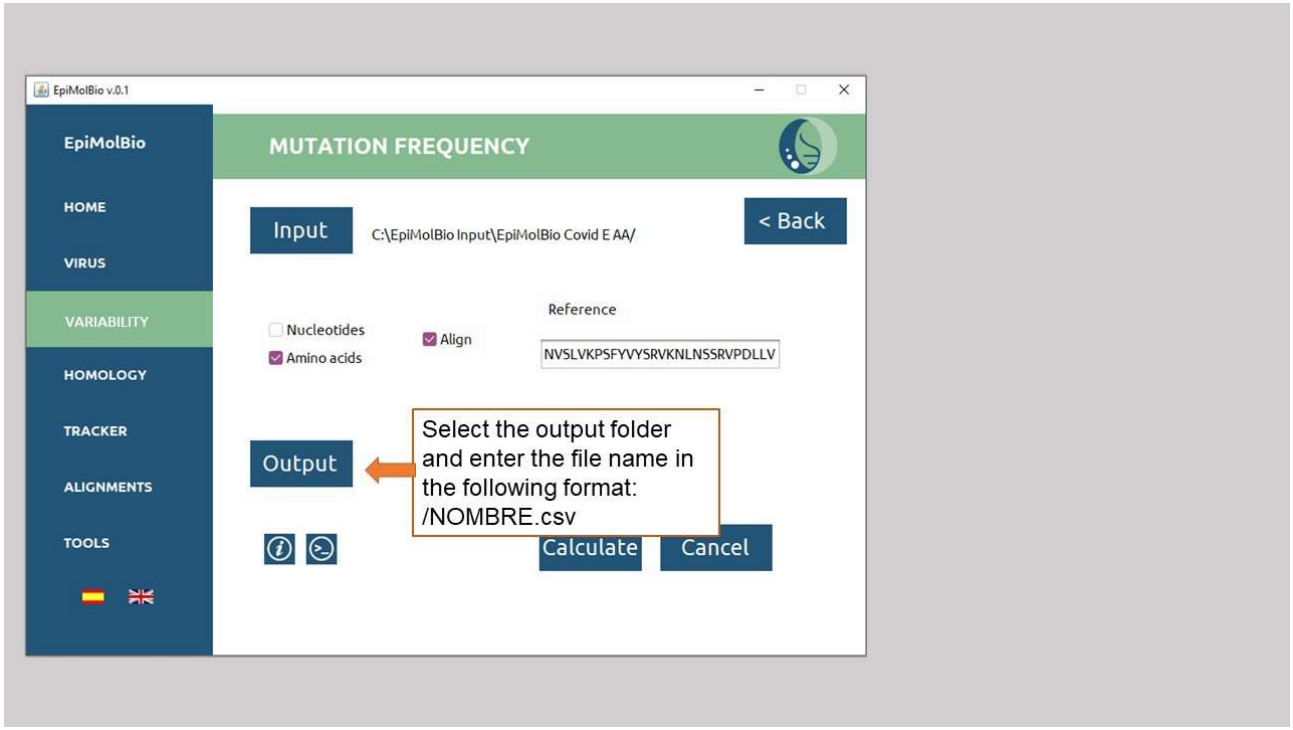
## 5)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
MUTATION FREQUENCY
Input
C:\EpiMolBio Input\EpiMolBio Covid E AA/
< Back
Reference
Nucleotides
Amino acids
Align
Select to align the input sequences if they were not aligned previously.
Output
Output
Calculate
Cancel

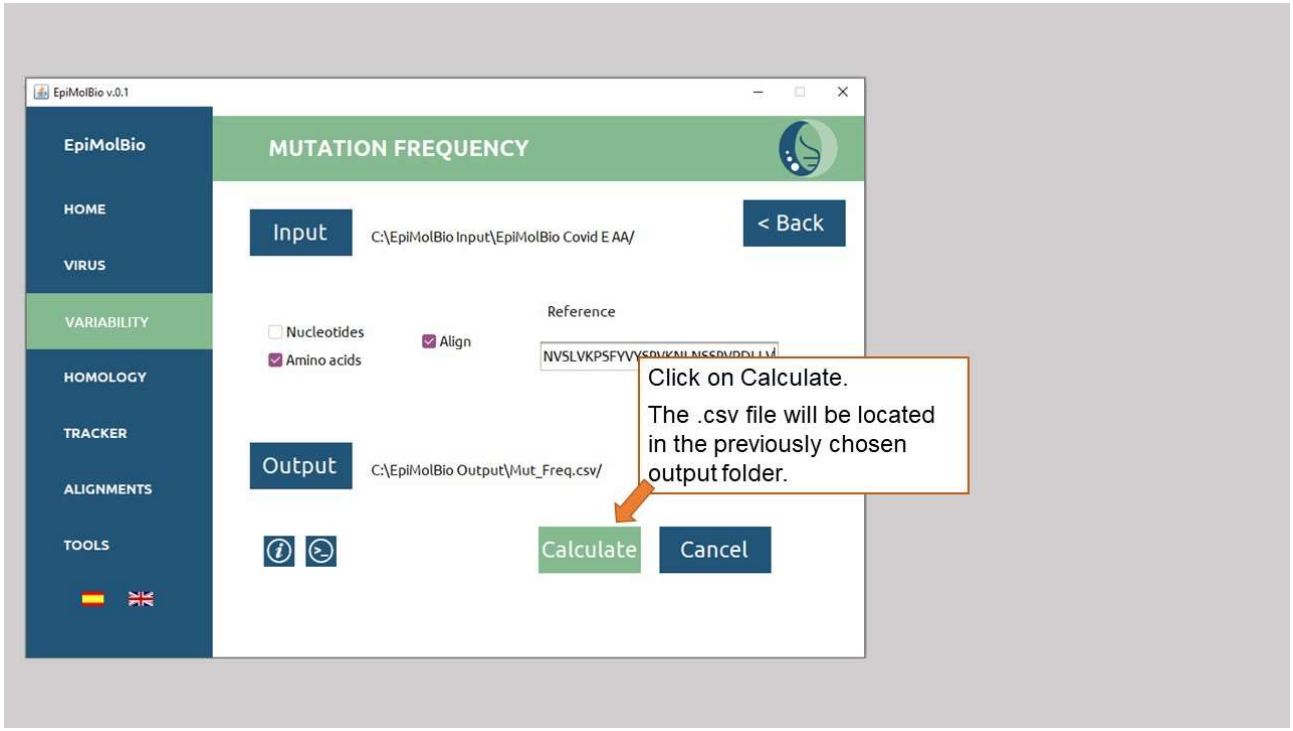
## 6)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
MUTATION FREQUENCY
Input
C:\EpiMolBio Input\EpiMolBio Covid E AA/
< Back
Reference
Nucleotides
Amino acids
Align
NVSLVKPSFYVYSRVKNLNSSRVPDLLV
Enter the reference sequence without line breaks.
Output
Output
Calculate
Cancel

## 7)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
MUTATION FREQUENCY
Input
C:\EpiMolBio Input\EpiMolBio Covid E AA/
< Back
Reference
Nucleotides
Amino acids
Align
NVSLVKPSFYVYSRVKNLNSSRVPDLLV
Output
Select the output folder and enter the file name in the following format: /NOMBRE.csv
Calculate
Cancel

## 8)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
MUTATION FREQUENCY
Input
C:\EpiMolBio Input\EpiMolBio Covid E AA/
< Back
Reference
Nucleotides
Amino acids
Align
Output
C:\EpiMolBio Output\Mut_Freq.csv/
Click on Calculate.
The .csv file will be located in the previously chosen output folder.
Calculate
Cancel

# III.HOMOLOGY

## III.1.SIMILARITY

This function allows **searching for a specific problem sequence entered by the user within the sequences of the input file**, generating an .html table indicating the proportion of sequences per file that contain the problem sequence. It enables searching throughout the entire length of the sequence or within a specific region. For example, it can be used to determine the proportion of sequences containing a specific peptide.

In the output file, at the top, the analysis title is displayed followed by the problem sequence being searched for. Below, in the 'File' column, the name of each analyzed file is shown. In the 'Frequency' column, the occurrence frequency of the problem sequence in that file is displayed, colored according to the color code described in the Overview section, which can be accessed in the .html output file by clicking on the blue symbol. Additionally, in the 'Total Sequences' column, the total number of sequences in the analyzed file is shown.

Example of Homology Similarity analysis output:

**Homology Similarity Range 5 - 20**

**Problem Sequence: WQRPLVT**

| File | Frequency | Total Sequences |
|---|---|---|
| PR_01_AE.fasta | 76.643% | 26849 |
| PR_02_AG.fasta | 64.728% | 9577 |
| PR_03_A6B.fasta | 69.355% | 310 |
| PR_04_cpx.fasta | 53.333% | 15 |
| PR_05_DF.fasta | 4.167% | 24 |
| PR_06_cpx.fasta | 61.126% | 746 |
| PR_07_BC.fasta | 76.695% | 10916 |

The **input** file should be the folder containing exclusively .fasta files with aligned sequences of the protein to be analyzed in nucleotides or amino acids. The input sequences may not be aligned, but in that case, the search region cannot be specified in the 'Range' field.

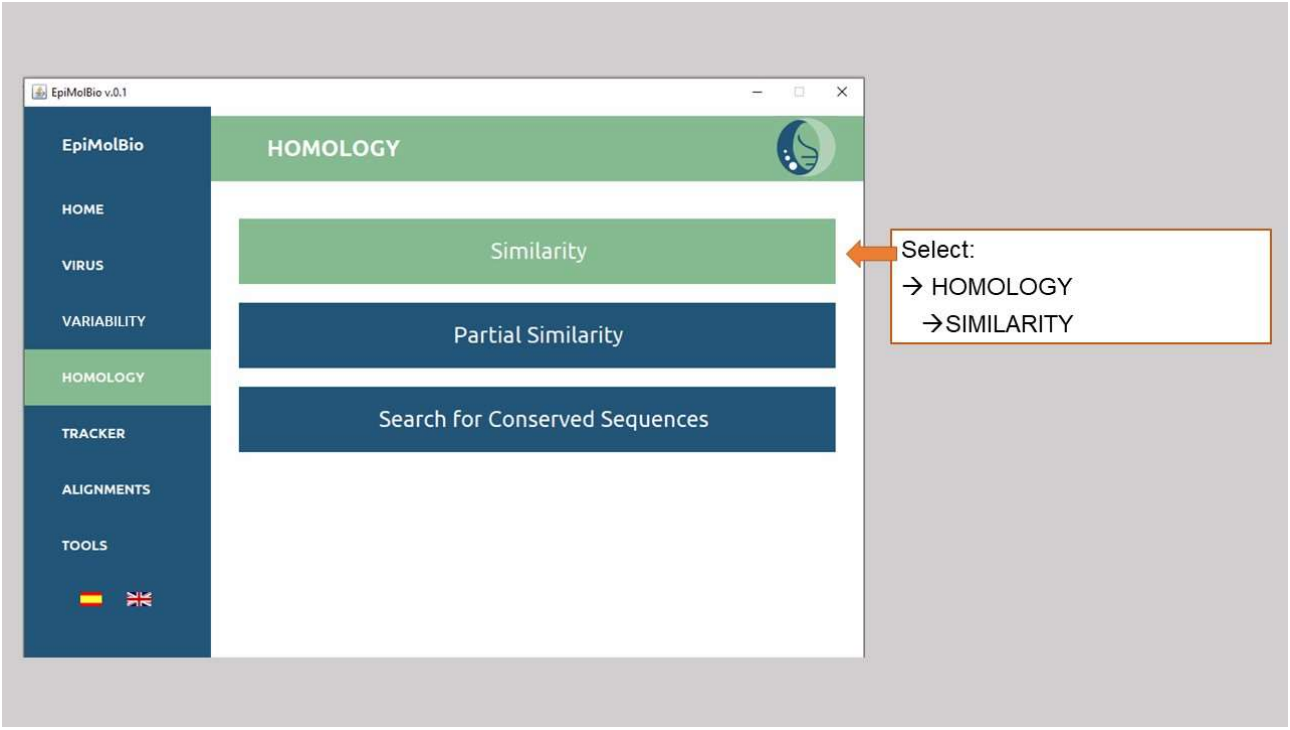
In the '**Problem Sequence**' field, enter the sequence to be searched in letters without spaces, in nucleotides or amino acids depending on whether the input files are translated or not (e.g., KLKPGMDGPKVK).

In the '**Range**' field, input the positions of nucleotides or amino acids that encompass the region of the input protein where the problem sequence is to be searched (e.g., to search between amino acid 10 and 30 inclusive, enter '10' in the first box and '30' in the second box). If you leave this field blank, it will search throughout the sequence.

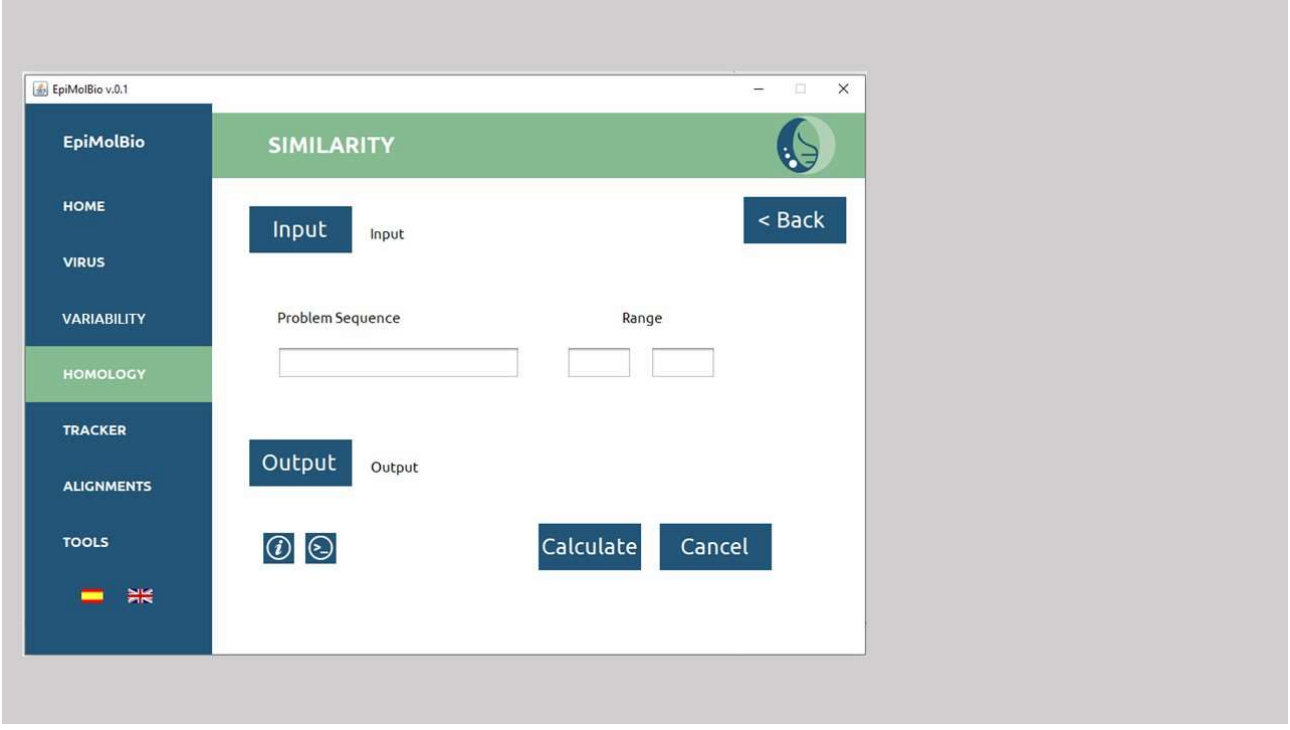
For the **output**, select the output folder where the .html files should appear and name the files by adding '.html' at the end.

# Step-by-step:

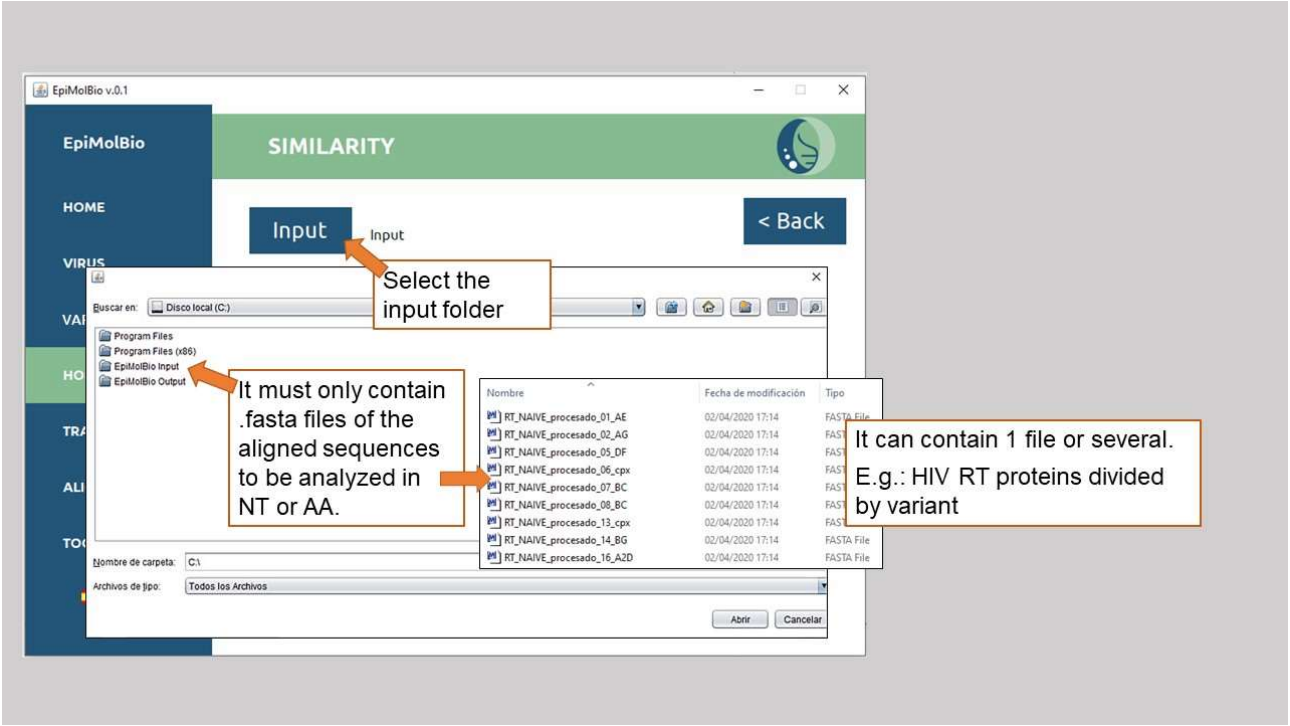
## 1)



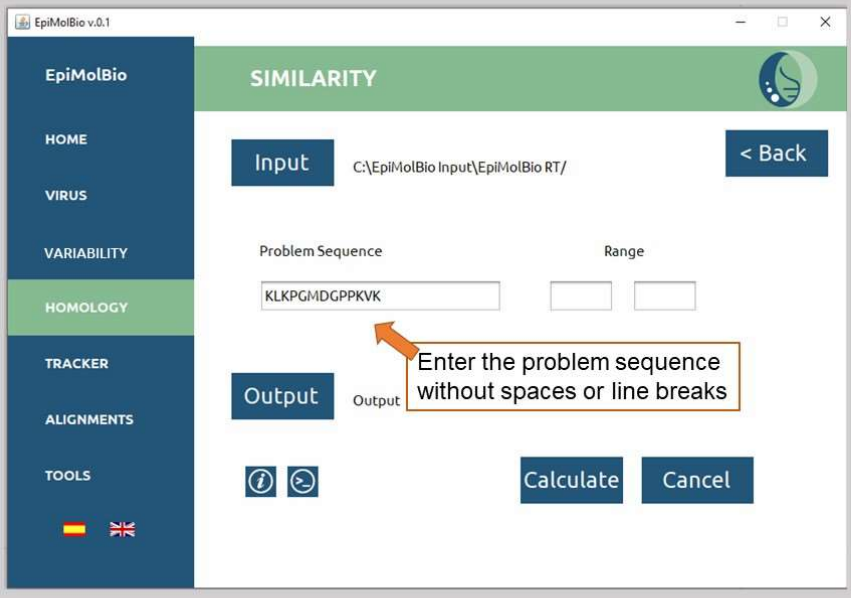

## 2)



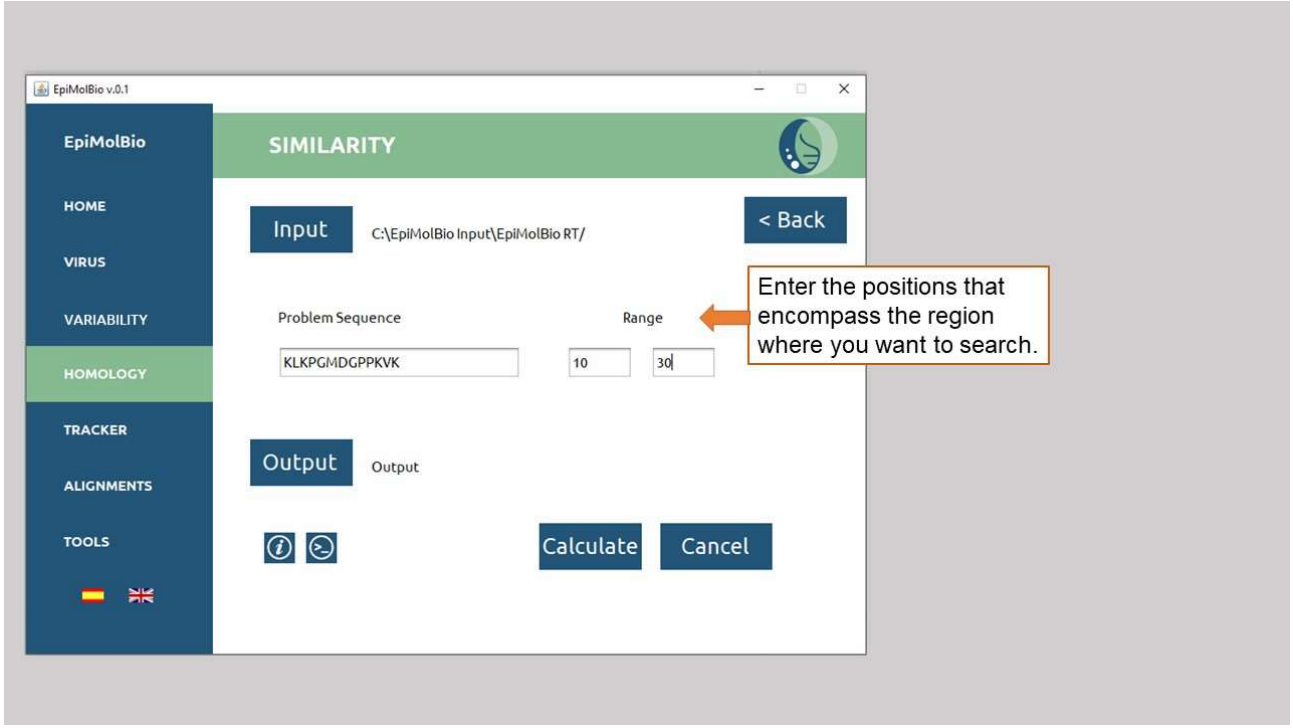
## 3)


EpiMolBio v.0.1
EpiMolBio
HOME
SIMILARITY
Input
Input
< Back
Select the input folder
Buscar en:
Disco local (C:)
Program Files
Program Files (x86)
EpiMolBio Input
EpiMolBio Output
It must only contain .fasta files of the aligned sequences to be analyzed in NT or AA.
Nombre
Fecha de modificación
Tipo
RT_NAIVE_procesado_01_AE
RT_NAIVE_procesado_02_AG
RT_NAIVE_procesado_05_DF
RT_NAIVE_procesado_06_cpx
RT_NAIVE_procesado_07_BC
RT_NAIVE_procesado_08_BC
RT_NAIVE_procesado_13_cpx
RT_NAIVE_procesado_14_BG
RT_NAIVE_procesado_16_A2D
02/04/2020 17:14
FASTA File
It can contain 1 file or several.
E.g.: HIV RT proteins divided by variant
Nombre de carpeta:
C:\
Archivos de tipo:
Todos los Archivos
Abrir
Cancelar

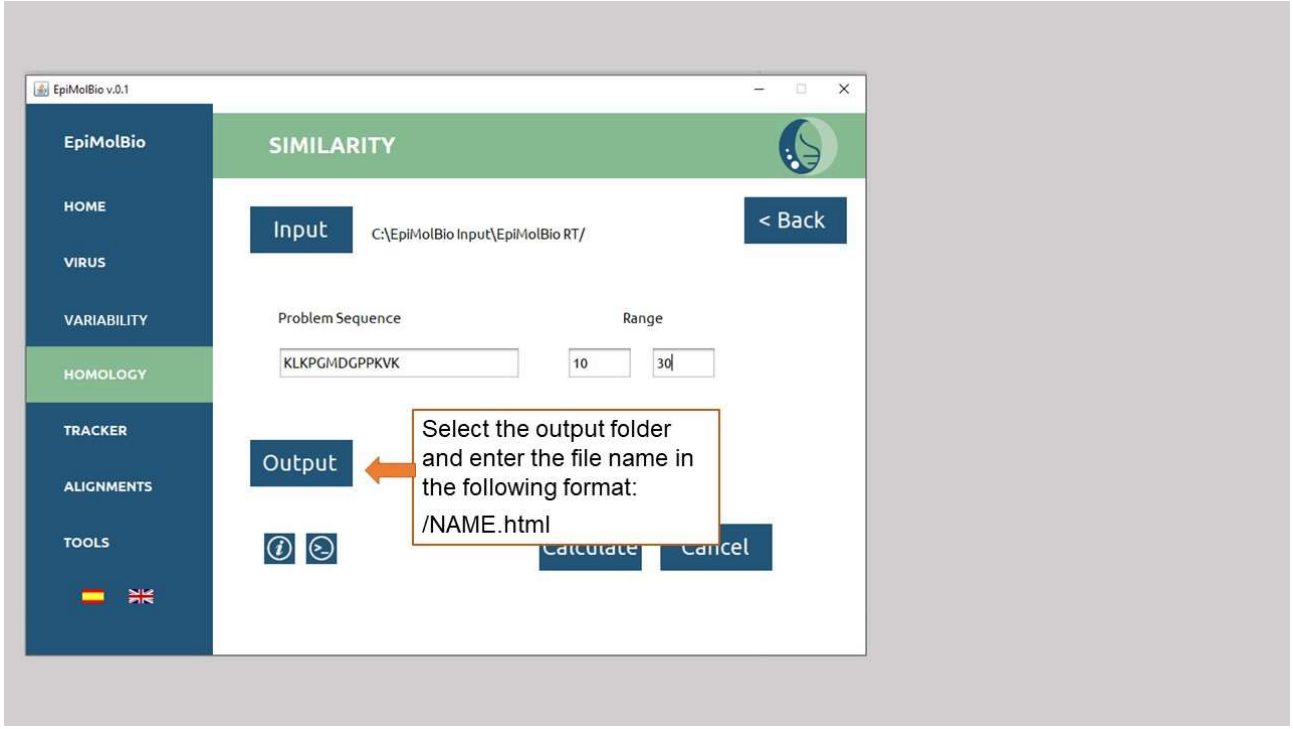
## 4)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
SIMILARITY
Input
C:\EpiMolBio Input\EpiMolBio RT/
< Back
Problem Sequence
Range
KLKPGMDGPPKVK
Enter the problem sequence without spaces or line breaks
Output
Output
Calculate
Cancel

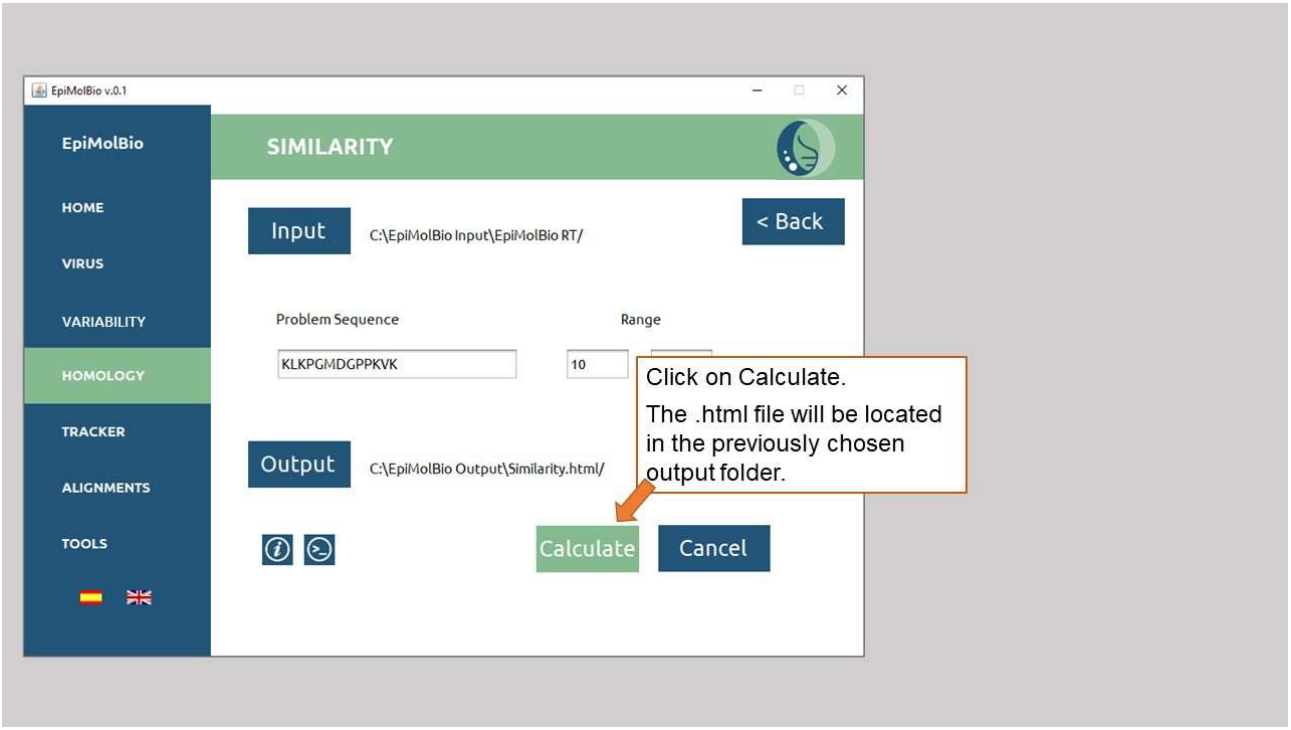
## 5)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
SIMILARITY
Input
C:\EpiMolBio Input\EpiMolBio RT/
< Back
Problem Sequence
Range
KLKPGMDGPPKVK
10
30
Enter the positions that encompass the region where you want to search.
Output
Output
Calculate
Cancel

## 6)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
SIMILARITY
Input
C:\EpiMolBio Input\EpiMolBio RT/
< Back
Problem Sequence
Range
KLKPGMDGPPKVK
10
30
Output
Select the output folder and enter the file name in the following format:
/NAME.html
Calculate
Cancel

# 7)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
SIMILARITY
Input
C:\EpiMolBio Input\EpiMolBio RT/
< Back
Problem Sequence
Range
KLKPGMDGPPKVK
10
Output
C:\EpiMolBio Output\Similarity.html/
Calculate
Cancel
Click on Calculate.
The .html file will be located in the previously chosen output folder.

# III.2.PARTIAL SIMILARITY

This function allows **comparing a user-entered sequence with sequences from the input file to search for similar regions between them, with the ability to define the percentage of similarity**. The program cuts both sequences into multiple fragments for comparison, and the length of these fragments can be specified. After the analysis, a .html table is generated containing the found sequence fragments and their percentage of similarity with the input sequence. For example, comparing the SARS-CoV-2 3C-like sequence (input sequence) with HIV PR sequence (input file) to locate regions of 10 similar amino acids at 50% similarity.

The **output** format is an .html table where, at the top, the analysis title is displayed. Below that, in the 'File' column, the name of the analyzed file is shown. Next, under the 'Sequence' column, the headers of the sequences from the input file are displayed. In the 'Input Sequence' column, a fragment of the input sequence for the search is shown, and in the 'Found Sequence' column, the result is displayed: the identified sequence fragment from the input sequences based on the chosen analysis parameters. Next to it, in the 'Similarity' column, the similarity percentage is displayed, colored according to the color code described in the Overview section, accessible in the .html output file by clicking on the blue symbol.

Example of Homology Partial Similarity analisys output with a 50% of similarity:

**Homology Partial Similarity 50.0%**

| File | Sequence | Input Sequence | Found Sequence | Similarity |
|---|---|---|---|---|
| PR_D.fasta | >D.ET.2003.ETH_G_230.AB285830 | GHRATGTVLV | GTDTTITVNV | 50.000% |
| PR_D.fasta | >D.ET.2003.ETH_G_230.AB285830 | IGRNLLTQLG | NGMNGRTILG | 50.000% |
| PR_D.fasta | >D.ET.2003.ETH_G_230.AB285830 | GRNLLTQLGC | GMNGRTILGS | 50.000% |
| PR_D.fasta | >D.ET.2003.ETH_G_230.AB285830 | NLLTQLGCTL | NGRTILGSAL | 50.000% |
| PR_D.fasta | >D.JP.-.patient_88.AB356098 | YDQIHVEICG | LTQDHVDILG | 50.000% |
| PR_D.fasta | >D.JP.-.patient_88.AB356098 | DQIHVEICGH | TQDHVDILGP | 50.000% |

The **input** file should be the folder containing exclusively .fasta files with sequences of the protein to be analyzed in nucleotides or amino acids. It is recommended to adjust the number of sequences based on their length to avoid excessively slowing down the processing speed. If the sequences are very long (e.g., complete genomes), it's preferable to perform searches in batches.

In the '**Length**' field, input the length of the fragments into which the input sequence and the entry sequence will be divided for comparison (e.g., 10, the sequences will be divided into segments of 10 amino acids or nucleotides for comparison).

In the '**% Similarity**' field, define the minimum percentage of similarity that the compared fragments should have to appear in the results. Enter the value as a number with one decimal place without the '%' symbol (e.g., 50.0 for 50% similarity).

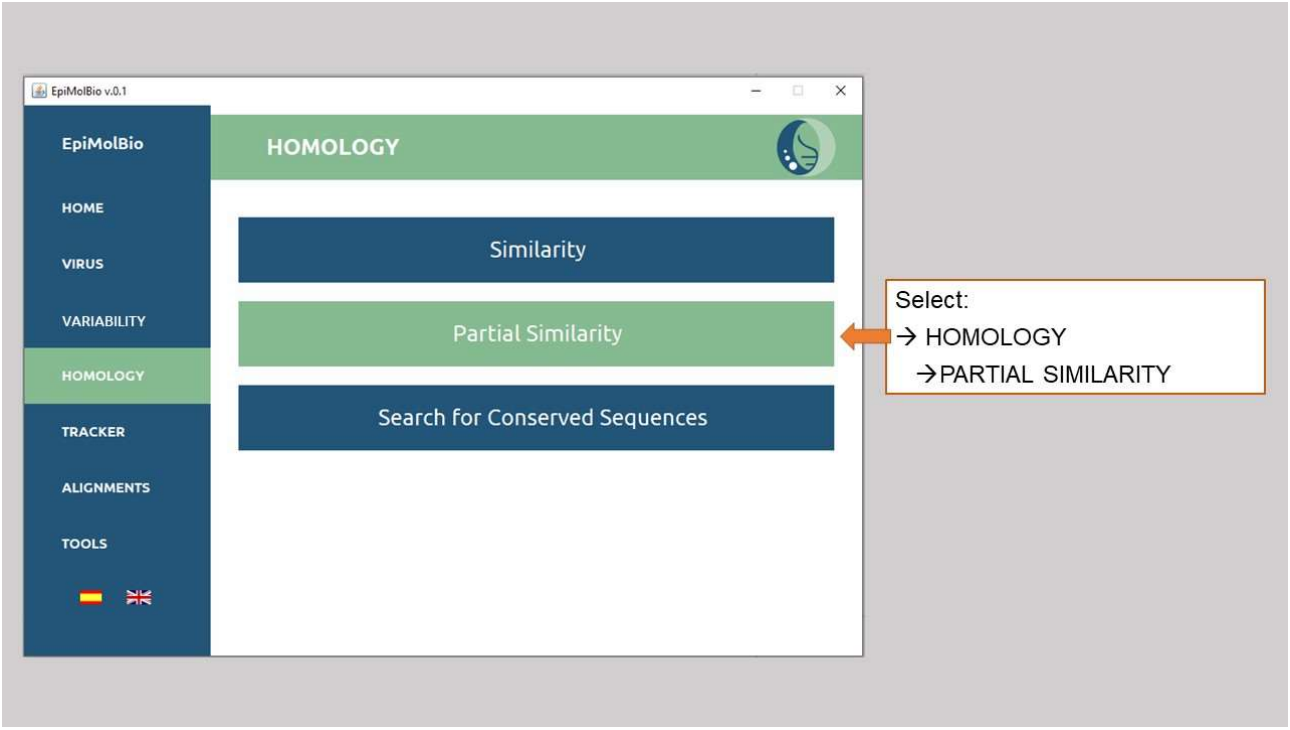
In the '**Align**' field, if you choose 'Align,' the fragments will be aligned with each other, resulting in a more exhaustive but slower search. You can choose 'Not Align' for a faster but less comprehensive search. It is recommended to use the second option for an initial overview and the first option for a thorough analysis if relevant results are found.

In the '**Sequence**' field, enter the sequence you want to search for in nucleotides or amino acids, depending on the input file. The sequence must not contain line breaks or spaces (continuing with the previous example: if the input consists of sequences of the HIV PR, in the Sequence field, enter the complete sequence of the SARS-CoV-2 3C-like protein).

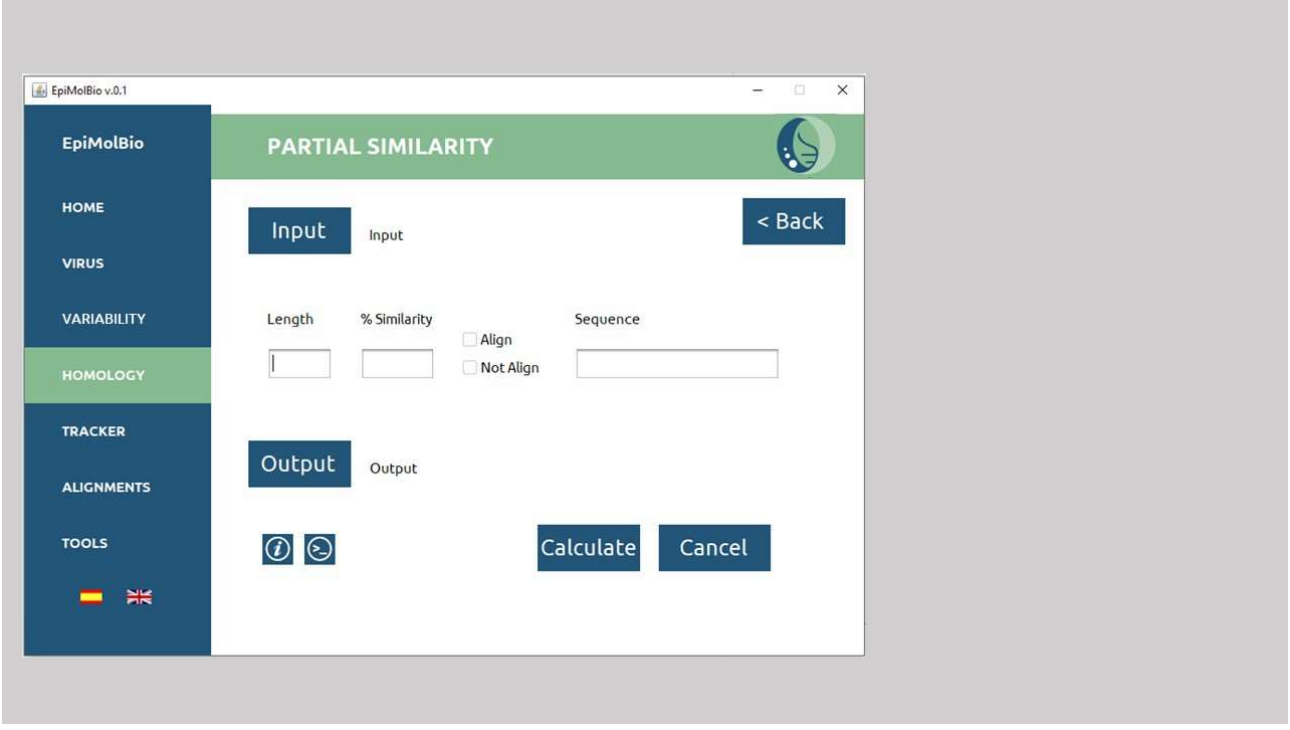
For the **output**, select the output folder where you want the .html files to appear and name the files by adding '.html' at the end.

# Step-by-step:

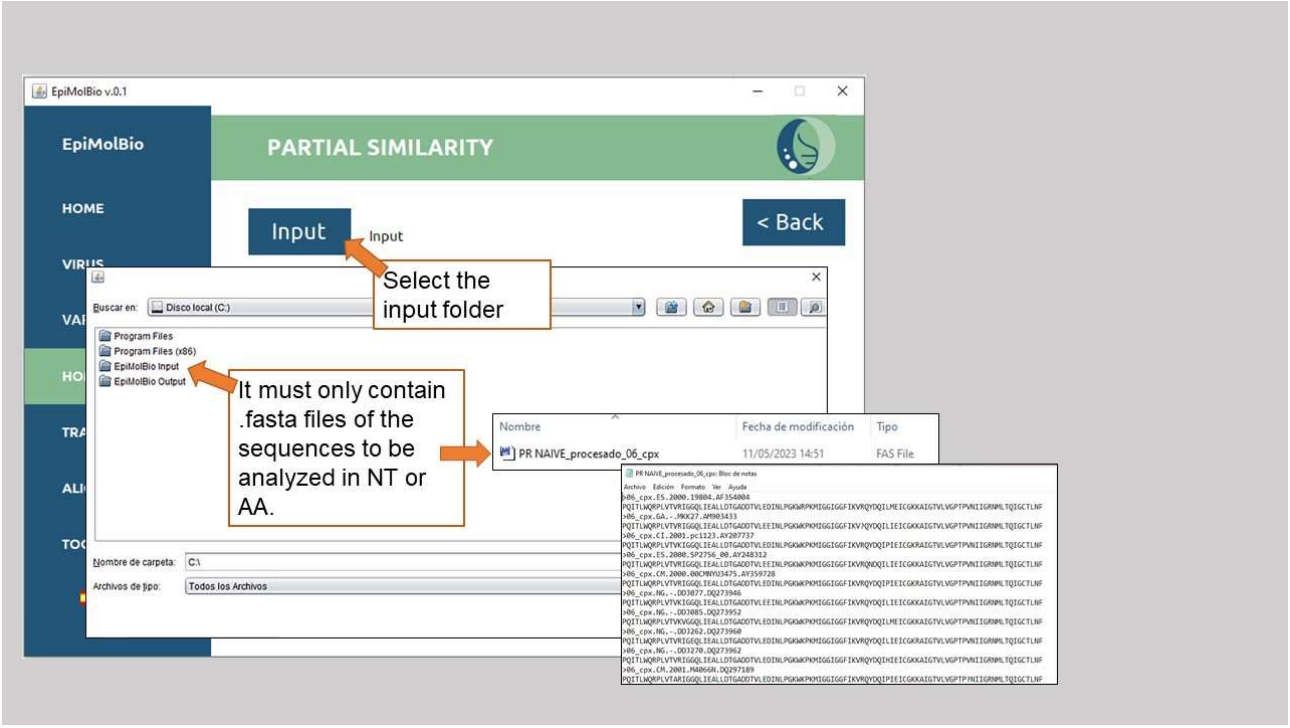
## 1)



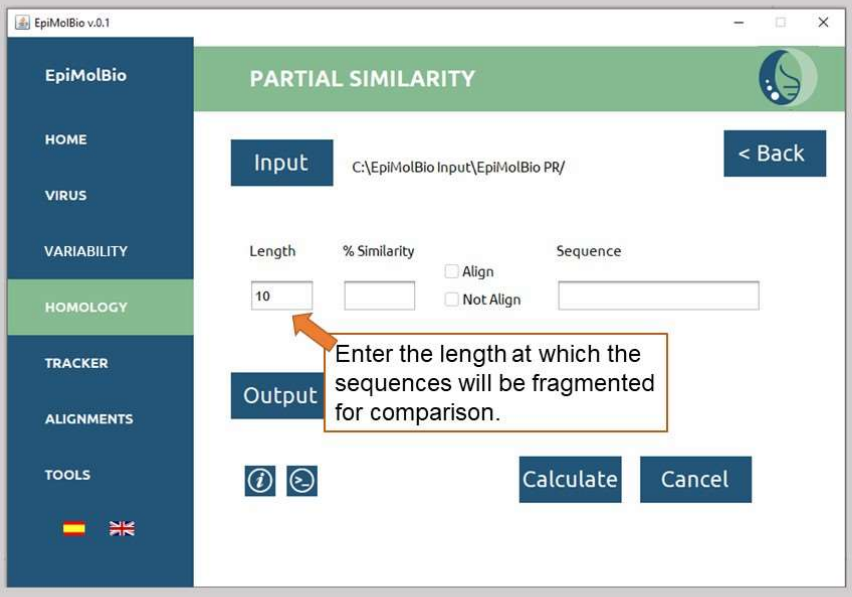

## 2)



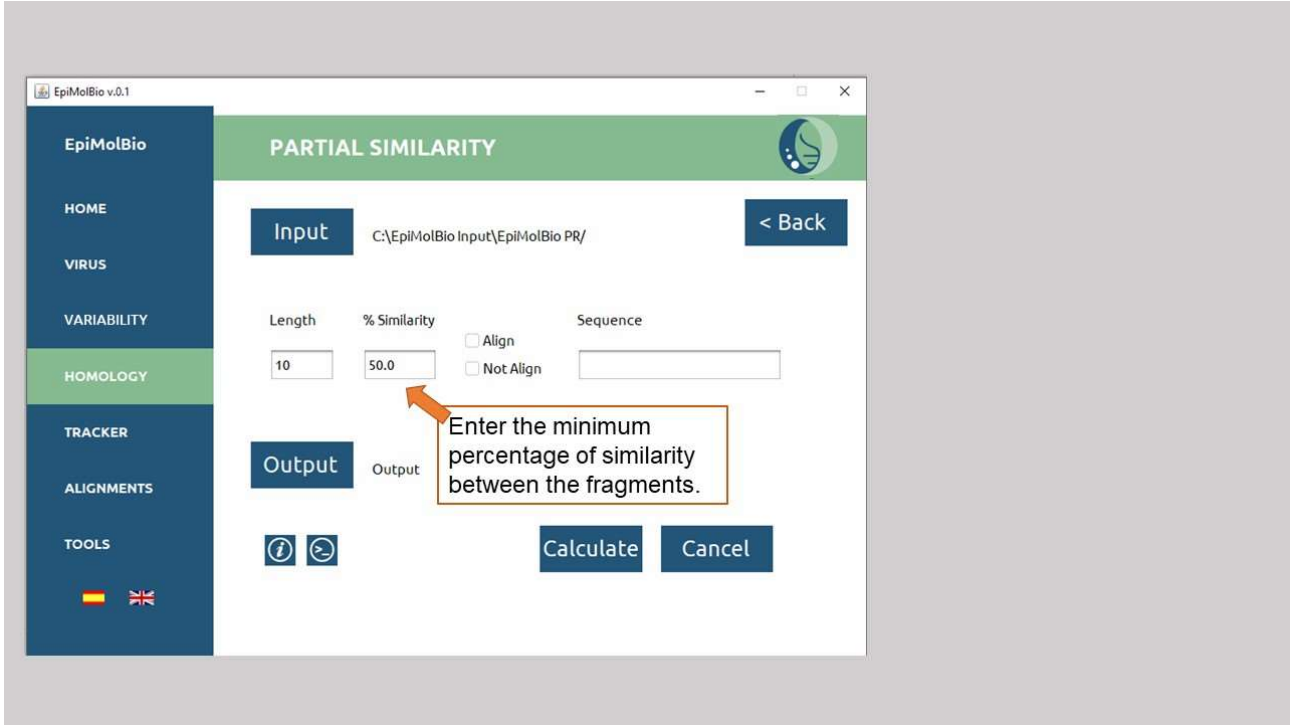
## 3)


EpiMolBio v.0.1
EpiMolBio
HOME
PARTIAL SIMILARITY
Input
Input
< Back
Select the input folder
Buscar en:
Disco local (C:)
Program Files
Program Files (x86)
EpiMolBio Input
EpiMolBio Output
It must only contain .fasta files of the sequences to be analyzed in NT or AA.
Nombre
Fecha de modificación
Tipo
PR NAIVE_procesado_06_cpx
11/05/2023 14:51
FAS File
Nombre de carpeta:
C:\
Archivos de tipo:
Todos los Archivos

## 4)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
PARTIAL SIMILARITY
Input
C:\EpiMolBio Input\EpiMolBio PR/
< Back
Length
% Similarity
Align
Not Align
Sequence
10
Enter the length at which the sequences will be fragmented for comparison.
Output
Calculate
Cancel

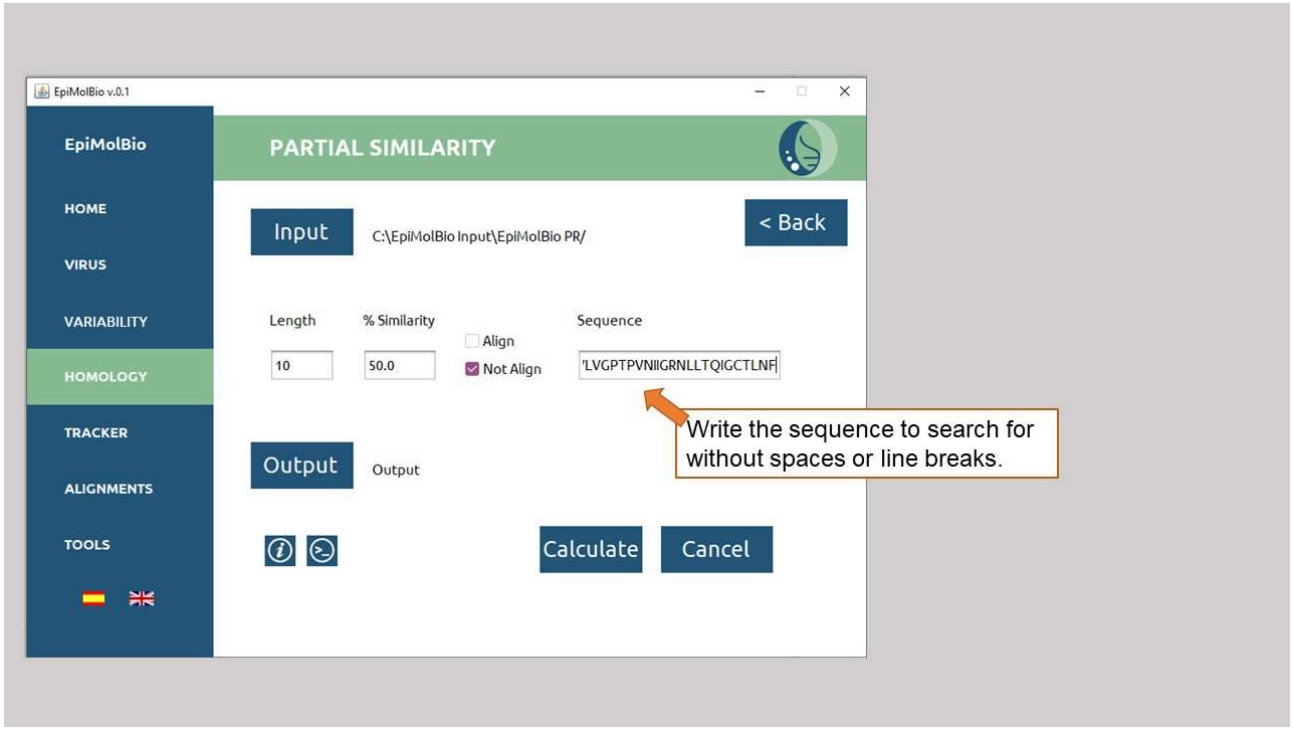
## 5)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
PARTIAL SIMILARITY
Input
C:\EpiMolBio Input\EpiMolBio PR/
< Back
Length
% Similarity
Align
Not Align
Sequence
10
50.0
Enter the minimum percentage of similarity between the fragments.
Output
Output
Calculate
Cancel

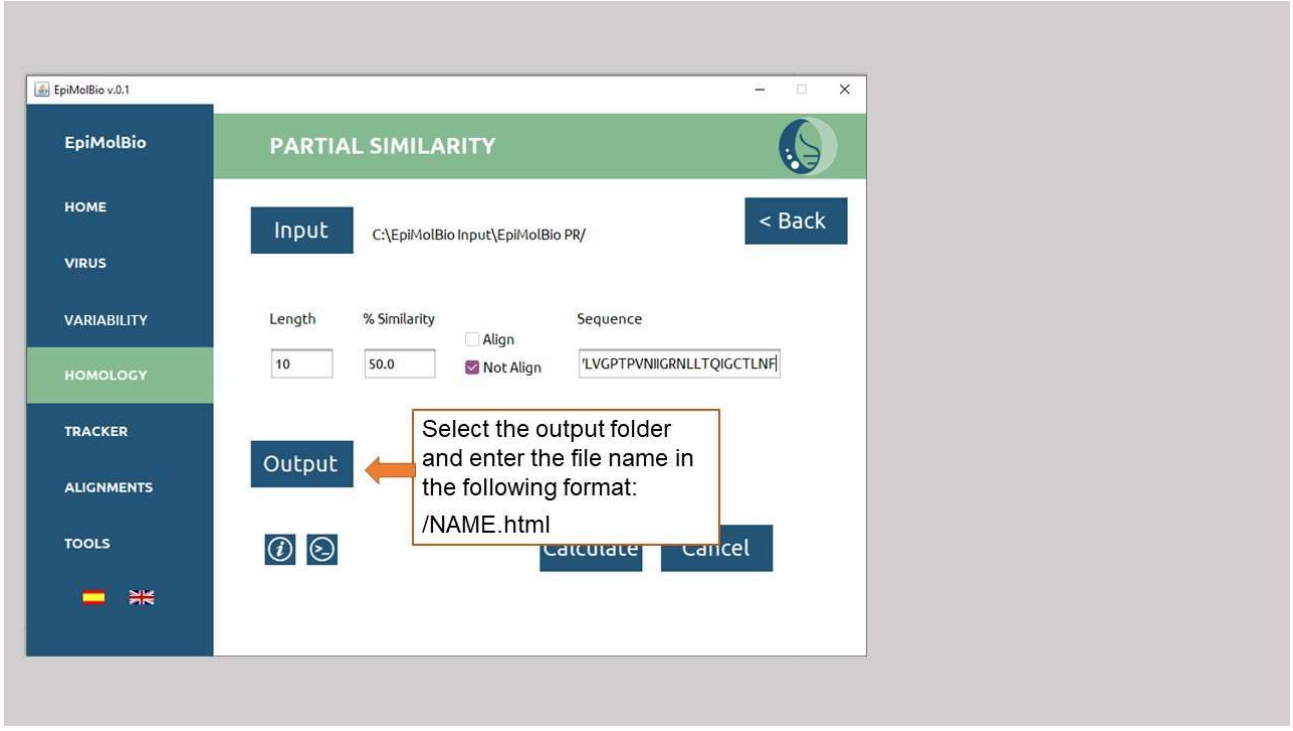
## 6)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
PARTIAL SIMILARITY
Input
C:\EpiMolBio Input\EpiMolBio PR/
< Back
Length
% Similarity
Align
Not Align
Sequence
10
50.0
Choose Align for exhaustive searches where the fragments are aligned, or Not Align for quick searches.
Output
Output
Calculate
Cancel

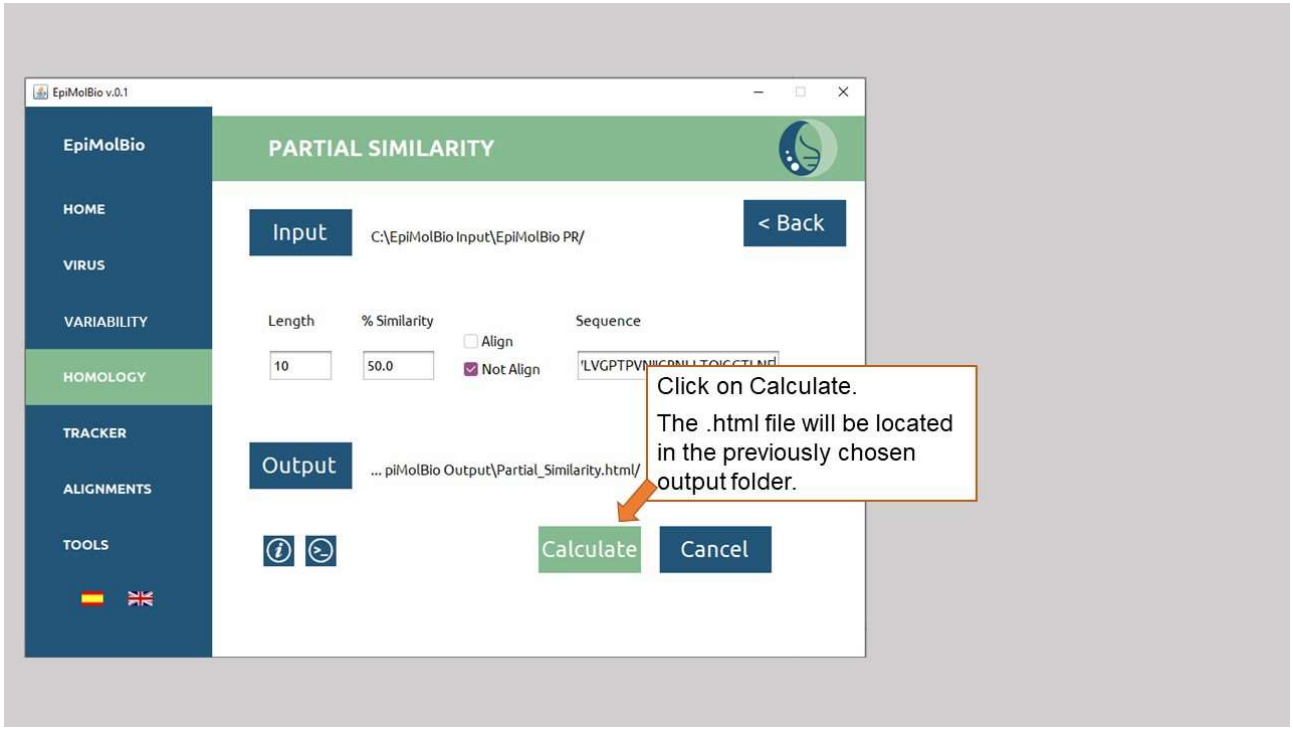
## 7)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
PARTIAL SIMILARITY
Input
C:\EpiMolBio Input\EpiMolBio PR/
< Back
Length
% Similarity
Align
Not Align
Sequence
10
50.0
'LVGPTPVNIIGRNLLTQIGCTLNF
Write the sequence to search for without spaces or line breaks.
Output
Output
Calculate
Cancel

## 8)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
PARTIAL SIMILARITY
Input
C:\EpiMolBio Input\EpiMolBio PR/
< Back
Length
% Similarity
Align
Not Align
Sequence
10
50.0
'LVGPTPVNIIGRNLLTQIGCTLNF
Output
Select the output folder and enter the file name in the following format:
/NAME.html
Cancel

**9)**


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
PARTIAL SIMILARITY
Input
C:\EpiMolBio Input\EpiMolBio PR/
< Back
Length
% Similarity
Align
Not Align
Sequence
10
50.0
Output
... piMolBio Output\Partial_Similarity.html/
Calculate
Cancel
Click on Calculate.
The .html file will be located in the previously chosen output folder.

# III.3.SEARCH FOR CONSERVED SEQUENCES

This function **enables the extraction of conserved sequence fragments from a set of input sequences, allowing you to search within a specific region, choose the fragment length, and define the conservation percentage**. For example, it can be used to search for conserved peptides for the subsequent design of diagnostic or therapeutic aptamers.

The **output** format is an .html table where, at the top, the analysis title is displayed. In the 'File' column, the name of the analyzed file appears, followed by the 'Length' column displaying the length of the fragment. The 'Region' column shows the region of the input sequence where the fragment has been found. In the 'Fragment' column, the result is displayed with the obtained sequence, followed by the 'Frequency' column, which shows the conservation percentage, colored according to the color code described in the Overview section. This color code can be referenced in the .html output file by clicking on the blue symbol.

Example of Homology Search fro Conserved Sequences analysis output:

**Homology Search for Conserved Sequences Length 10 - 10 95.0%**

| File | Length | Region | Fragment | Frequency |
|---|---|---|---|---|
| PR_71_BF1.fasta | 10 | 22 - 31 | ALLDTGADDT | 100.000% |
| PR_71_BF1.fasta | 10 | 23 - 32 | LLDTGADDTV | 100.000% |
| PR_71_BF1.fasta | 10 | 24 - 33 | LDTGADDTVL | 100.000% |
| PR_71_BF1.fasta | 10 | 25 - 34 | DTGADDTVLE | 100.000% |
| PR_130_A1B.fasta | 10 | 1 - 10 | PQITLWQRPL | 100.000% |
| PR_130_A1B.fasta | 10 | 2 - 11 | QITLWQRPLV | 100.000% |
| PR_130_A1B.fasta | 10 | 3 - 12 | ITLWQRPLVT | 100.000% |

The **input** file should be the folder containing exclusively .fasta files with sequences to be analyzed in nucleotides or amino acids. It is advisable for the input sequences to be aligned and without insertions if you intend to search for conserved fragments within a specific region, or if the input sequences are not significantly different from each other.

In the '**Range**' field, select 'Full Range' when you want to search across the entire length of the input sequences, or choose 'Select Range' if you want to search within a specific region of the input sequences.

If you have chosen 'Select Range' in the previous field, in the '**Select Range**' field, input the positions of nucleotides or amino acids that encompass the region where you want to search for the conserved fragment. For example, to search between amino acid 10 and 30 inclusive, enter '10' in the first box and '30' in the second box.

In the **'Conservation %**' field, define the minimum conservation percentage that the fragments must have to appear in the results. Enter the value as a number with one decimal place without the '%' symbol. For instance, use '80.0' for 80% conservation.

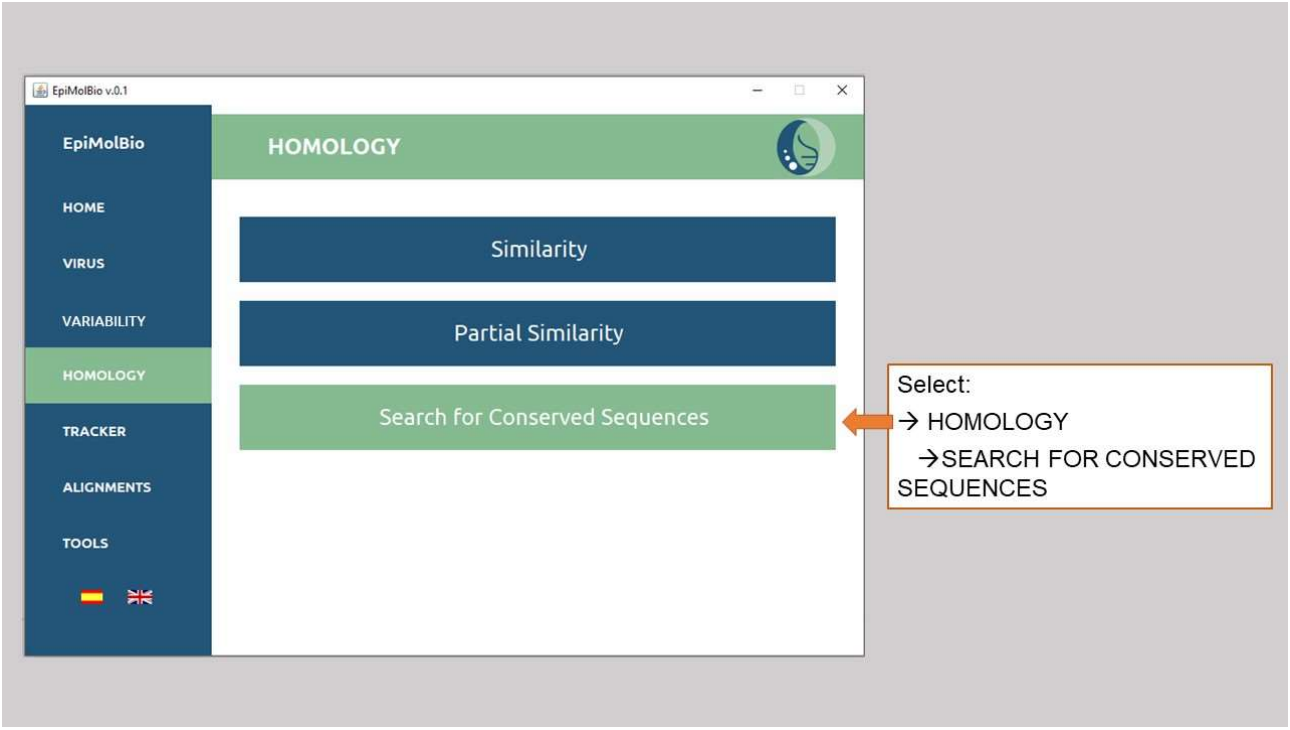
In the '**Sequence Length**' field, input the desired length for the resulting conserved fragments. If you want a length between 20 and 25 residues, enter '20' in the first box and '25' in the second. If you want them to have exactly 20 residues, enter '20' in both boxes.

In the '**Reference**' field, you can optionally input a reference sequence to expedite the calculation process. This reference sequence should be without spaces or line breaks and should be in nucleotides or amino acids, depending on the input file. If you leave this field blank, an automatic consensus sequence will be generated as a reference.

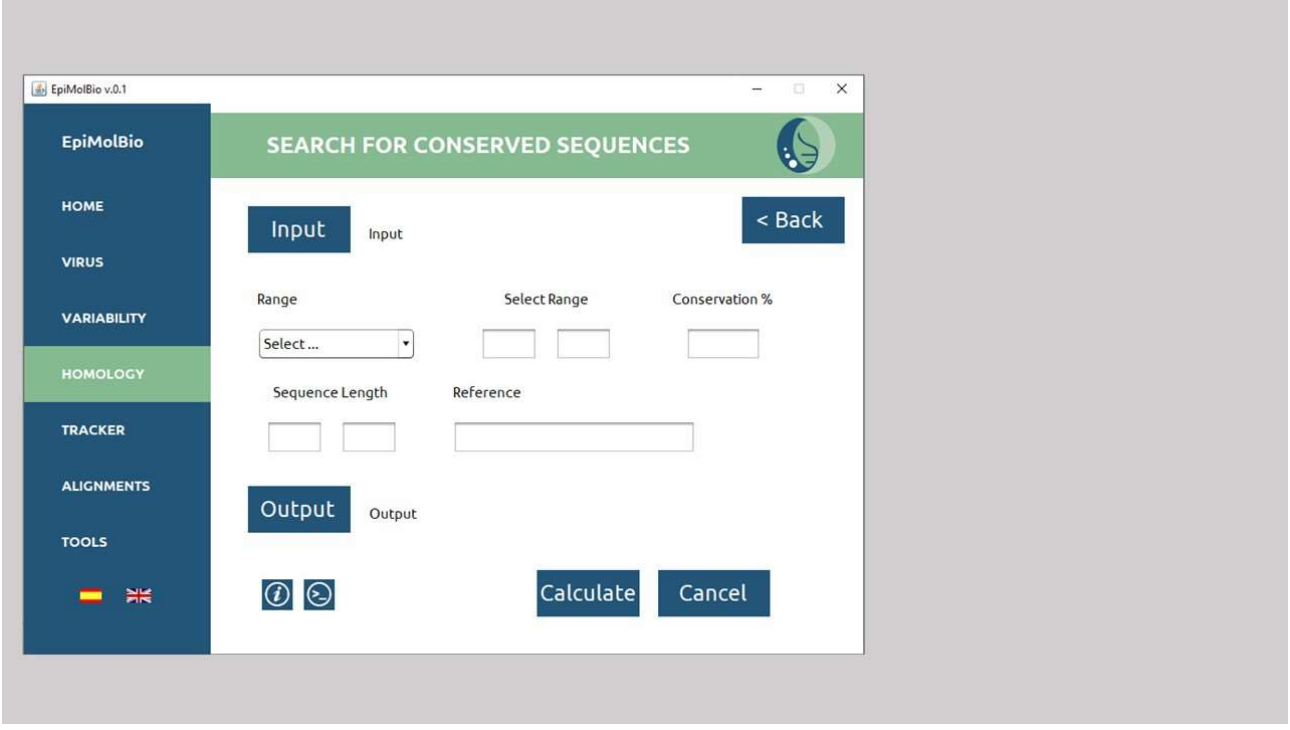
For the **output**, select the output folder where you want the .html files to appear and name the files by adding '.html' at the end.

# Step-by-step:

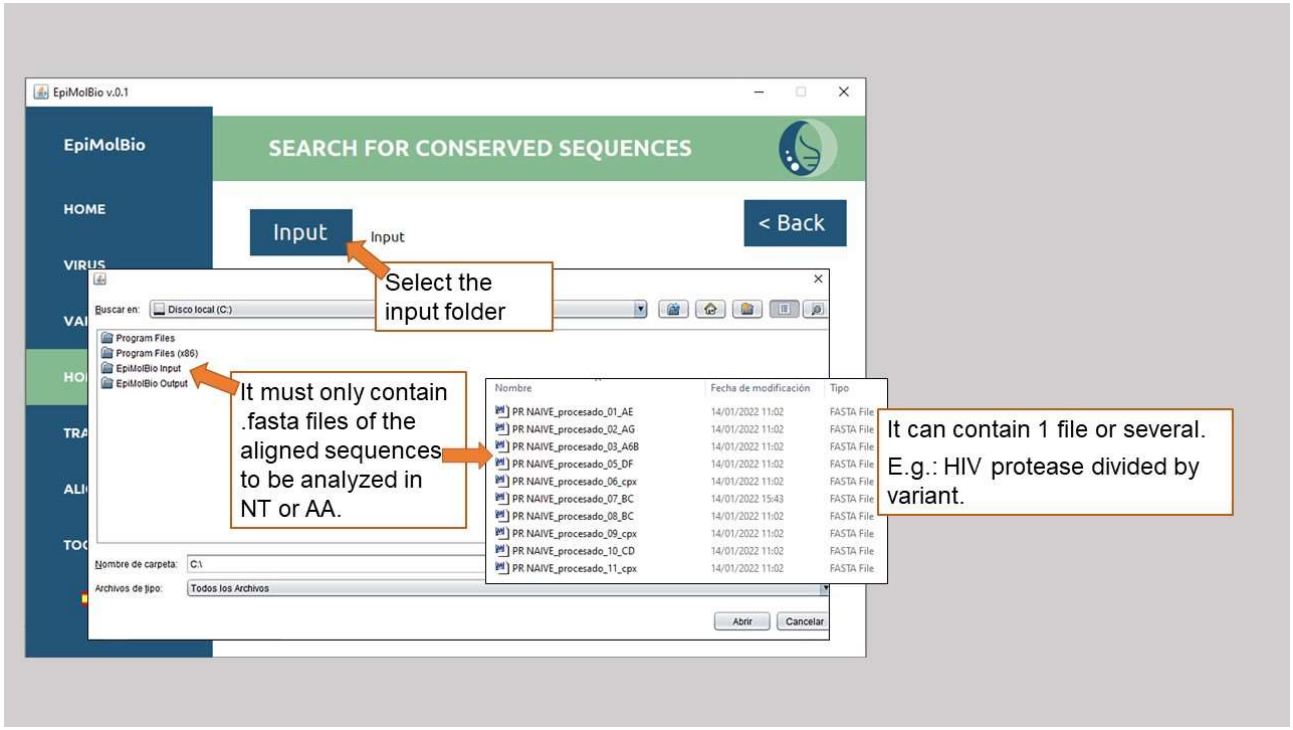
## 1)



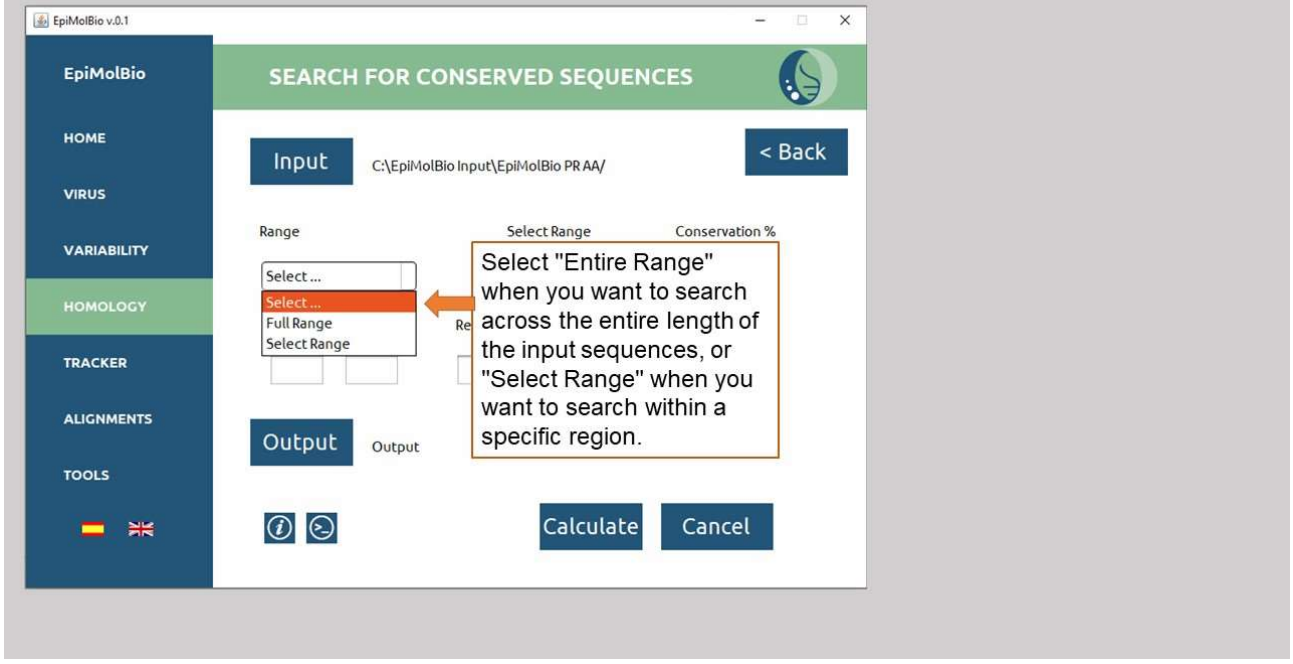

## 2)



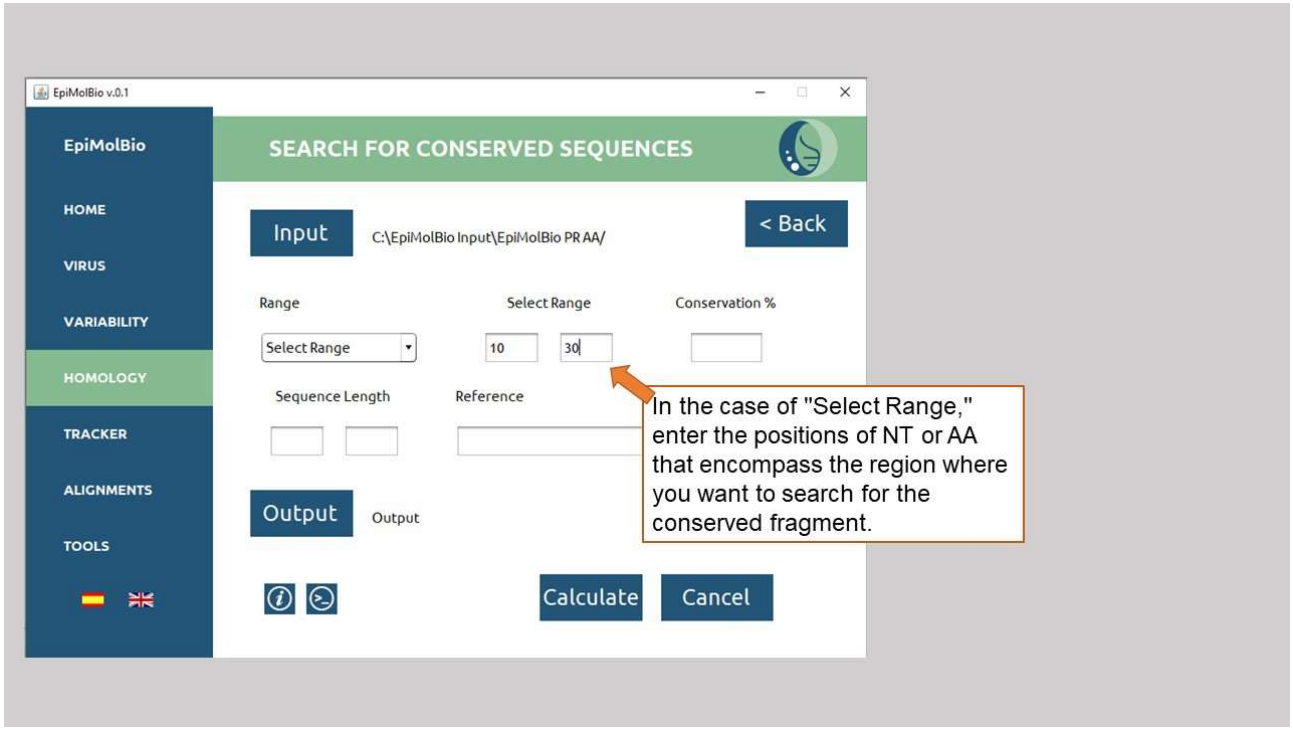
## 3)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
SEARCH FOR CONSERVED SEQUENCES
Input
Input
< Back
Select the input folder
Buscar en:
Disco local (C:)
Program Files
Program Files (x86)
EpiMolBio Input
EpiMolBio Output
It must only contain .fasta files of the aligned sequences to be analyzed in NT or AA.
Nombre
Fecha de modificación
Tipo
PR NAIVE_procesado_01_AE
PR NAIVE_procesado_02_AG
PR NAIVE_procesado_03_A6B
PR NAIVE_procesado_05_DF
PR NAIVE_procesado_06_cpx
PR NAIVE_procesado_07_BC
PR NAIVE_procesado_08_BC
PR NAIVE_procesado_09_cpx
PR NAIVE_procesado_10_CD
PR NAIVE_procesado_11_cpx
FASTA File
It can contain 1 file or several.
E.g.: HIV protease divided by variant.
Nombre de carpeta:
C:\
Archivos de tipo:
Todos los Archivos
Abrir
Cancelar

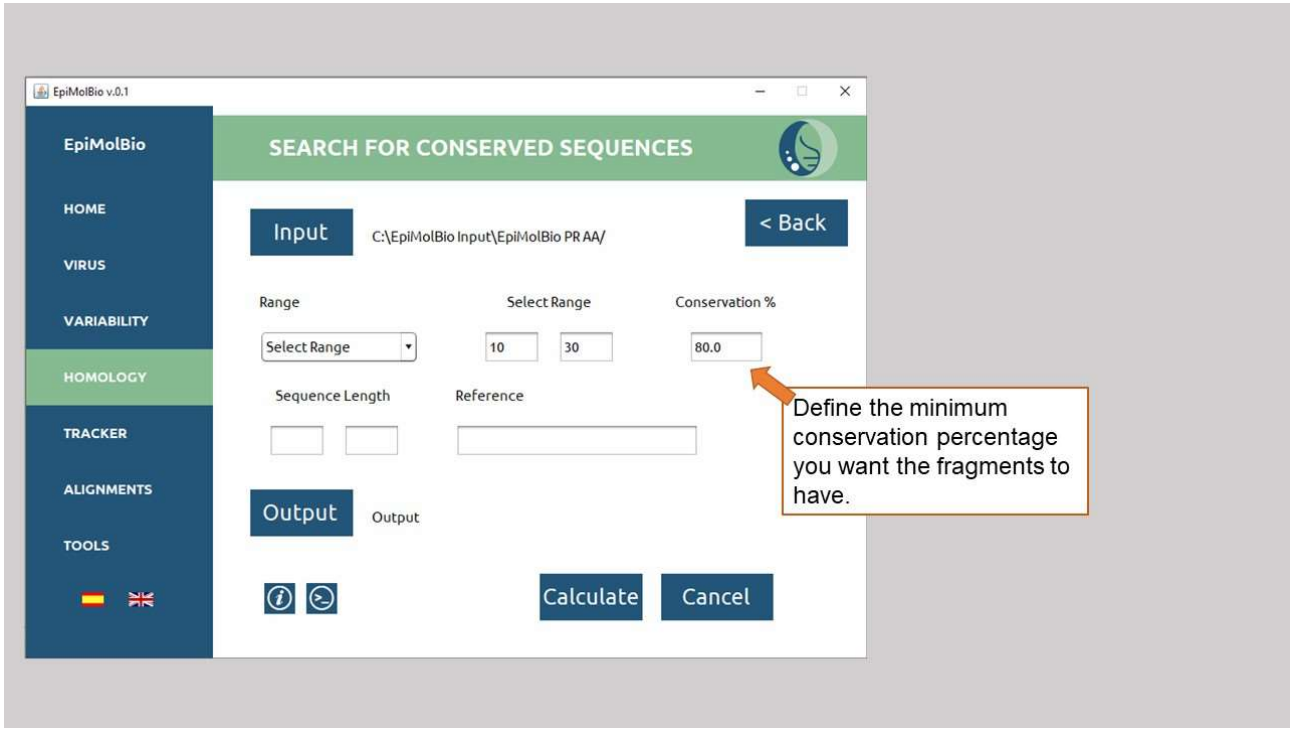
## 4)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
SEARCH FOR CONSERVED SEQUENCES
Input
C:\EpiMolBio Input\EpiMolBio PR AA/
< Back
Range
Select Range
Conservation %
Select ...
Select ...
Full Range
Select Range
Select "Entire Range" when you want to search across the entire length of the input sequences, or "Select Range" when you want to search within a specific region.
Output
Output
Calculate
Cancel

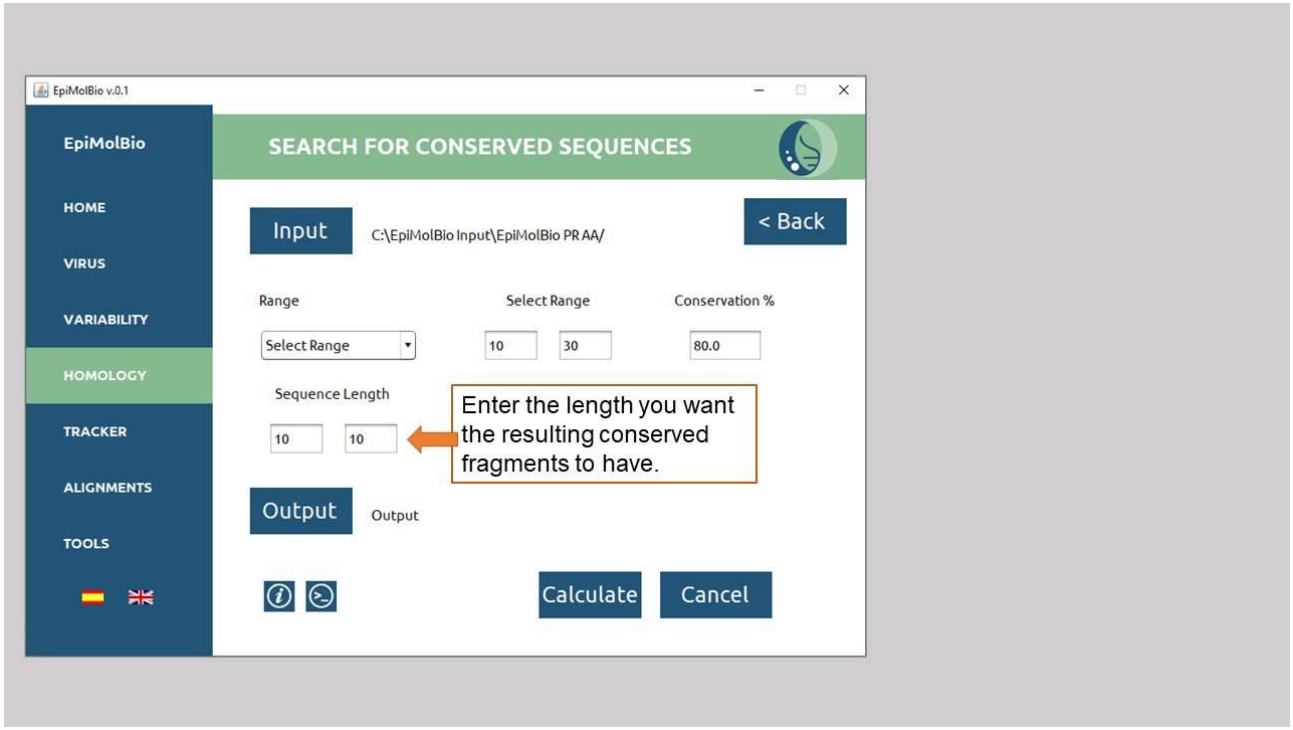
## 5)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
SEARCH FOR CONSERVED SEQUENCES
Input
C:\EpiMolBio Input\EpiMolBio PR AA/
< Back
Range
Select Range
Conservation %
Select Range
10
30
Sequence Length
Reference
Output
Output
Calculate
Cancel
In the case of "Select Range," enter the positions of NT or AA that encompass the region where you want to search for the conserved fragment.

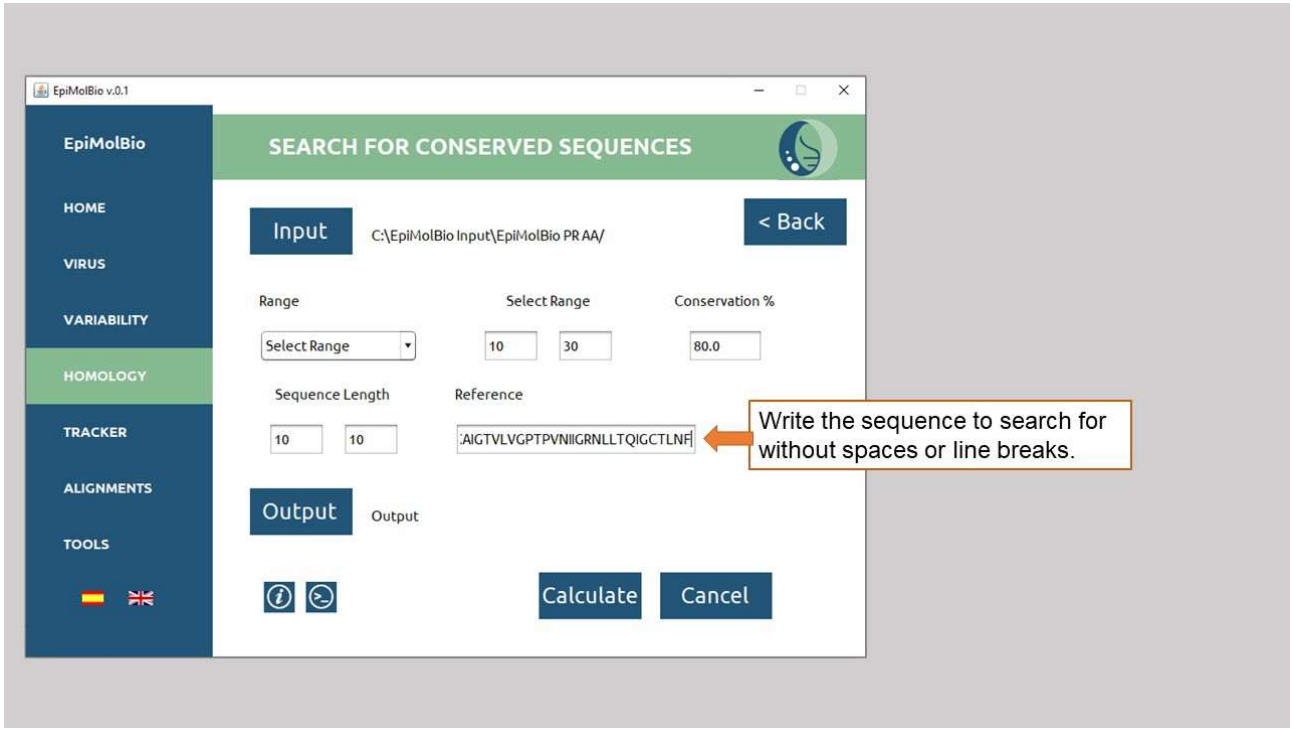
## 6)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
SEARCH FOR CONSERVED SEQUENCES
Input
C:\EpiMolBio Input\EpiMolBio PR AA/
< Back
Range
Select Range
Conservation %
Select Range
10
30
80.0
Sequence Length
Reference
Output
Output
Calculate
Cancel
Define the minimum conservation percentage you want the fragments to have.

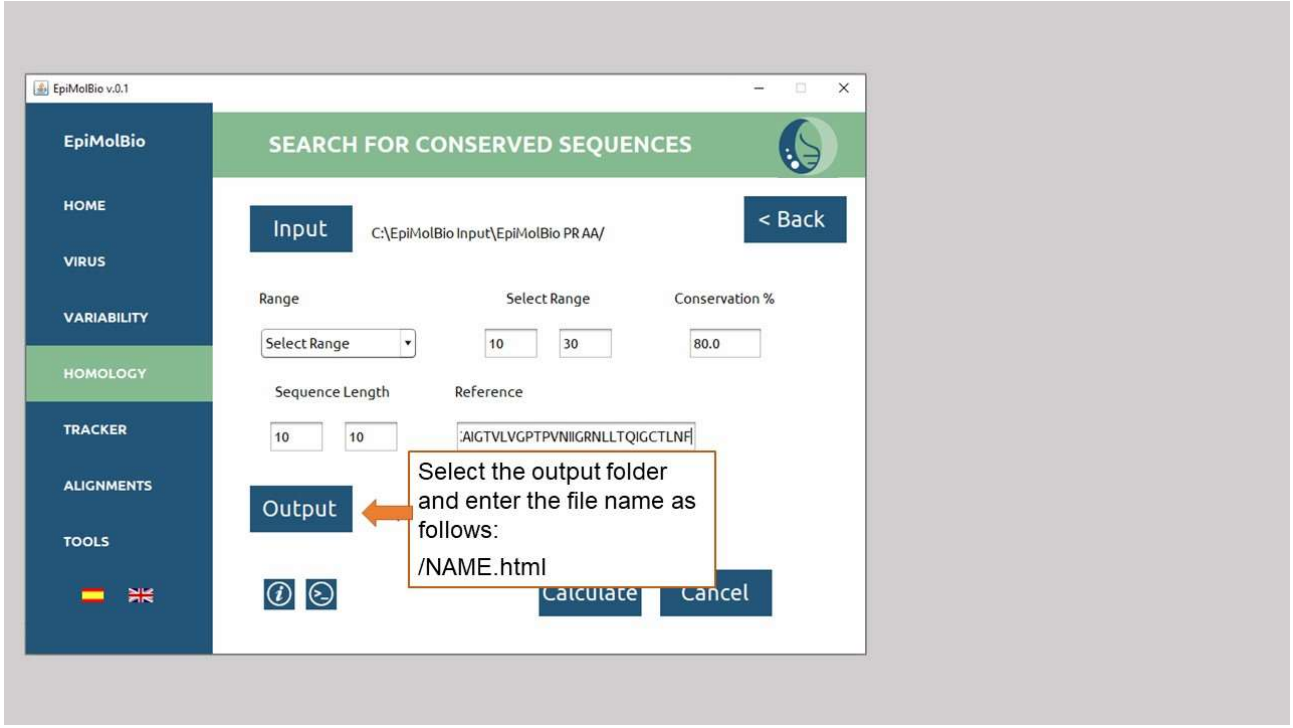
## 7)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
SEARCH FOR CONSERVED SEQUENCES
Input
C:\EpiMolBio Input\EpiMolBio PR AA/
< Back
Range
Select Range
Conservation %
Select Range
10
30
80.0
Sequence Length
10
10
Enter the length you want the resulting conserved fragments to have.
Output
Output
Calculate
Cancel

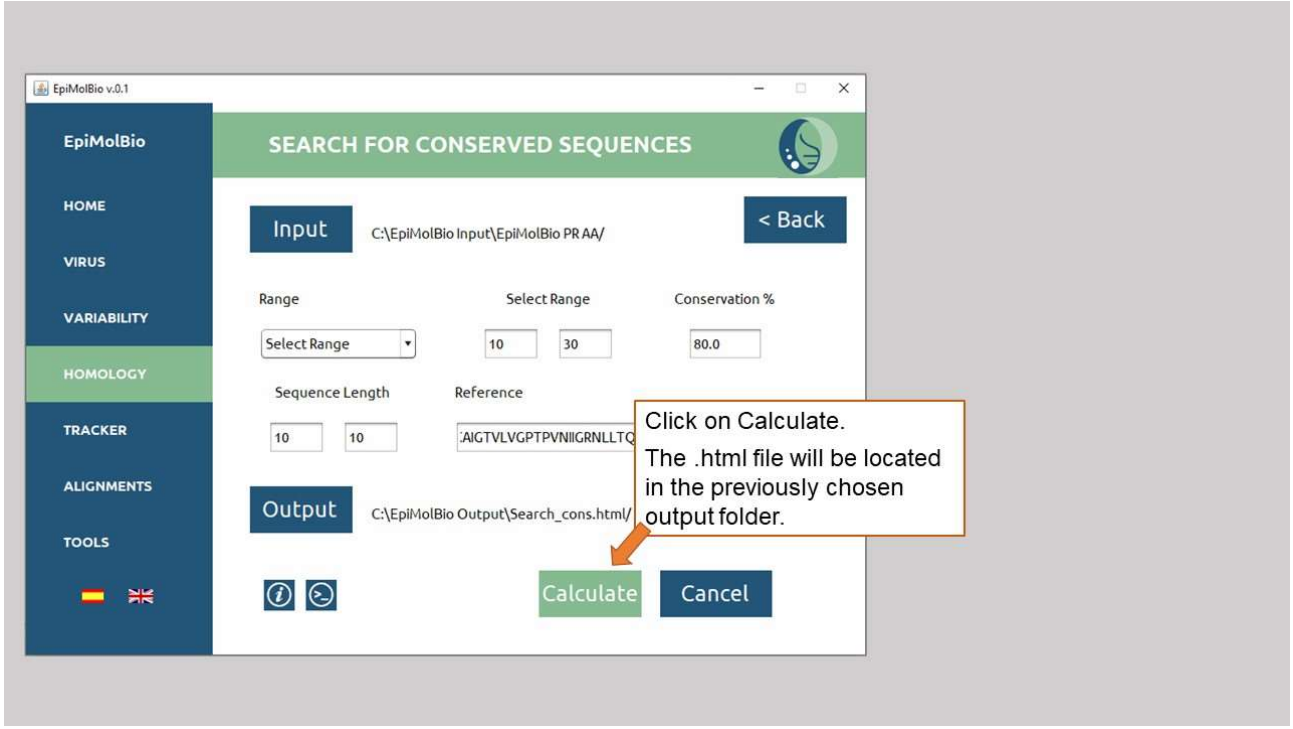
## 8)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
SEARCH FOR CONSERVED SEQUENCES
Input
C:\EpiMolBio Input\EpiMolBio PR AA/
< Back
Range
Select Range
Conservation %
Select Range
10
30
80.0
Sequence Length
Reference
10
10
AIGTVLVGPTPVNIIGRNLLTQIGCTLNF
Write the sequence to search for without spaces or line breaks.
Output
Output
Calculate
Cancel

## 9)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
SEARCH FOR CONSERVED SEQUENCES
Input
C:\EpiMolBio Input\EpiMolBio PR AA/
< Back
Range
Select Range
Conservation %
Select Range
10
30
80.0
Sequence Length
Reference
10
10
:AIGTVLVGPTPVNIIGRNLLTQIGCTLNF
Output
Select the output folder and enter the file name as follows:
/NAME.html
Calculate
Cancel

## 10)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
SEARCH FOR CONSERVED SEQUENCES
Input
C:\EpiMolBio Input\EpiMolBio PR AA/
< Back
Range
Select Range
Conservation %
Select Range
10
30
80.0
Sequence Length
Reference
10
10
:AIGTVLVGPTPVNIIGRNLLTQ
Output
C:\EpiMolBio Output\Search_cons.html/
Click on Calculate.
The .html file will be located in the previously chosen output folder.
Calculate
Cancel

# IV.TRACKER

## IV.1.SIMILARITY

This function **allows searching for target sequences in nucleotides or amino acids within a set of longer sequences that may be incomplete or of varying lengths**. The search is based on a user-entered reference sequence. For example, you can search for the HIV-2 protease within a group of HIV-2 genome sequences or the Spike protein sequence within the complete SARS-CoV-2 genome. You can choose the similarity percentage between what you're searching for and the reference sequence to avoid discarding valid sequences due to mutations. The **output** format is a .fasta file containing the found sequences.

The **input** file should be the folder containing exclusively .fasta files with preferably complete sequences where you want to perform the search. These sequences can be in nucleotides (NT) or amino acids (AA), they may not be aligned and can have different lengths.

Select whether you want to **translate** the resulting sequence or not. If the input is in NT and you want the output in NT, select 'Not Translate' (the reference sequence should be in NT). If the input is in NT and you want the output in AA, select 'Translate' (the reference sequence should be in AA). If the input is in AA, select 'Not Translate,' and the output will also be in AA (the reference should also be in AA).

Select '**Full Range**' if you want to search across the entire length of the input sequences, or choose '**Select Range**' to search within a specific region. In the latter case, the input sequences must be aligned.

If you have chosen 'Select Range' in the previous field, input the positions of nucleotides or amino acids that encompass the region where you want to search for the fragment (e.g., to search between AA 10 and 30 inclusive, enter '10' in the first box and '30' in the second box) in the '**Range**' field.

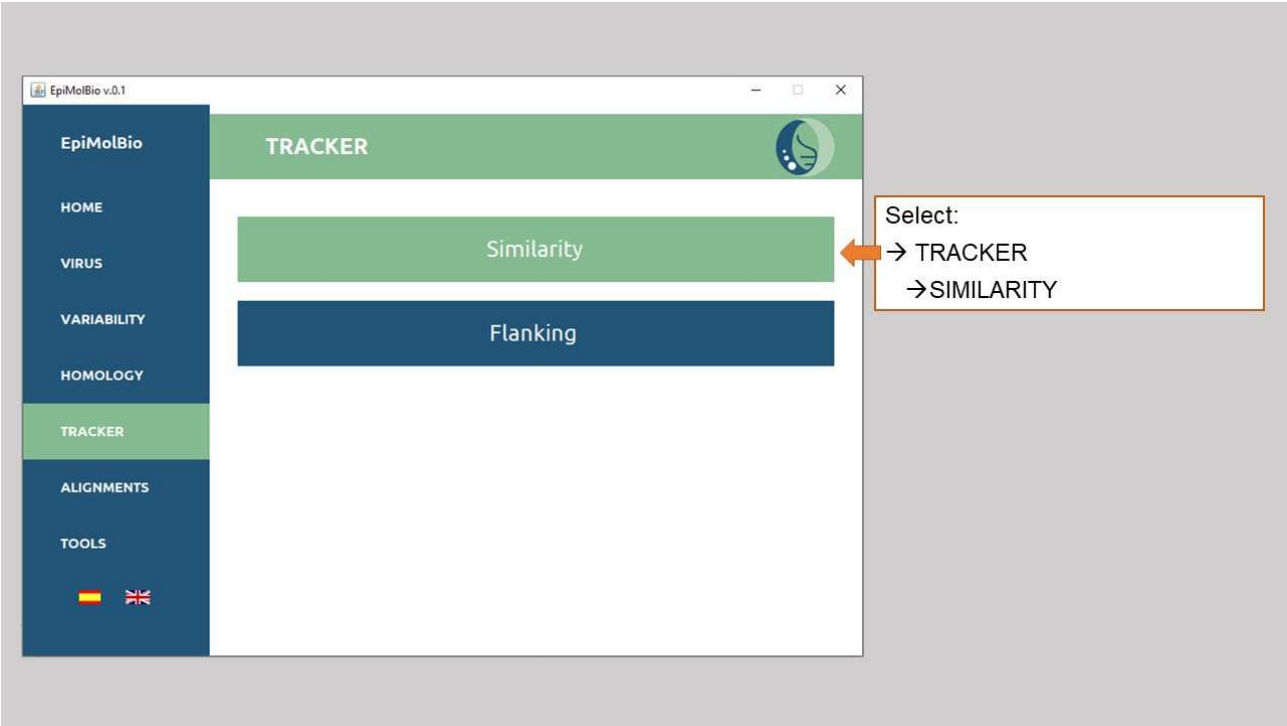
In the '**% Similarity**' field, define the minimum similarity percentage that you want the output sequences to have with respect to the reference sequence. Enter the value as a number with one decimal place without the '%' symbol. For instance, use '90.0' for 90% similarity.

In the '**Reference**' field, input the reference sequence you are searching for, without line breaks or spaces (for example the Spike reference sequence when searching SARS-CoV-2 complete genome).

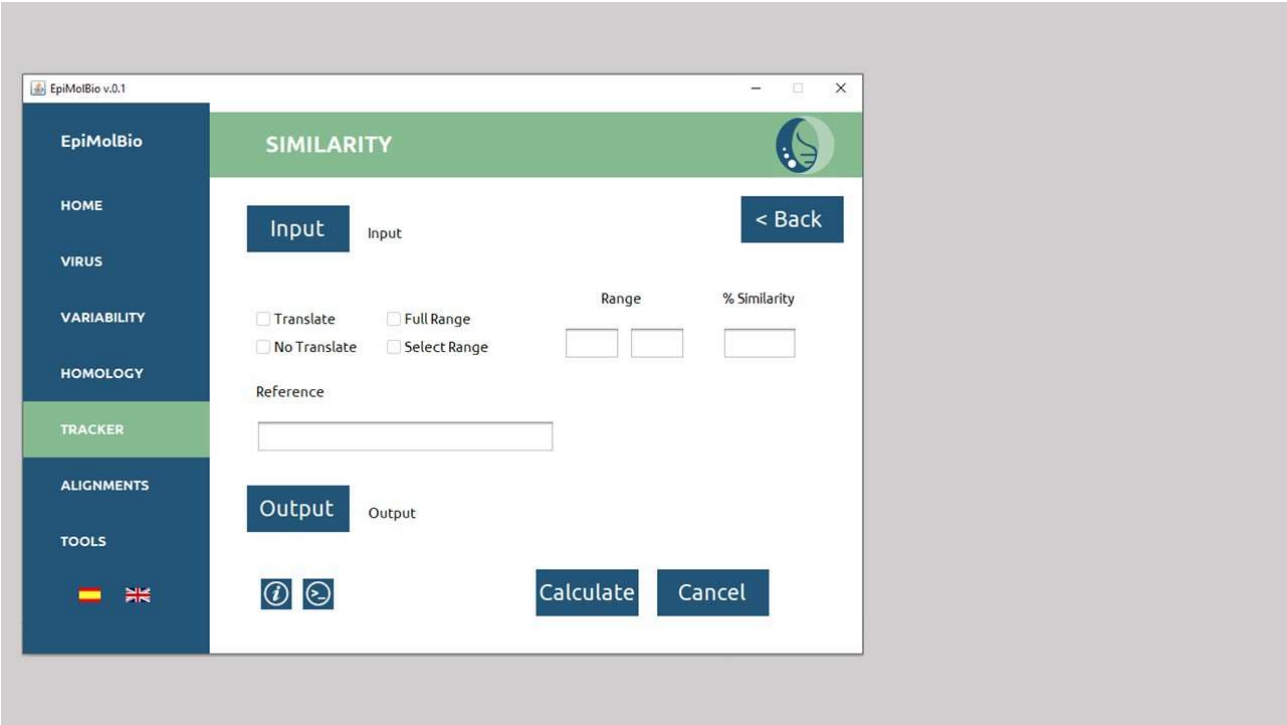
For the **output**, select the output folder where you want the .fasta files with the found sequences to appear. The files are automatically named as follows: Tracked_Similarity_InputFileName.fasta

# Step-by-step:

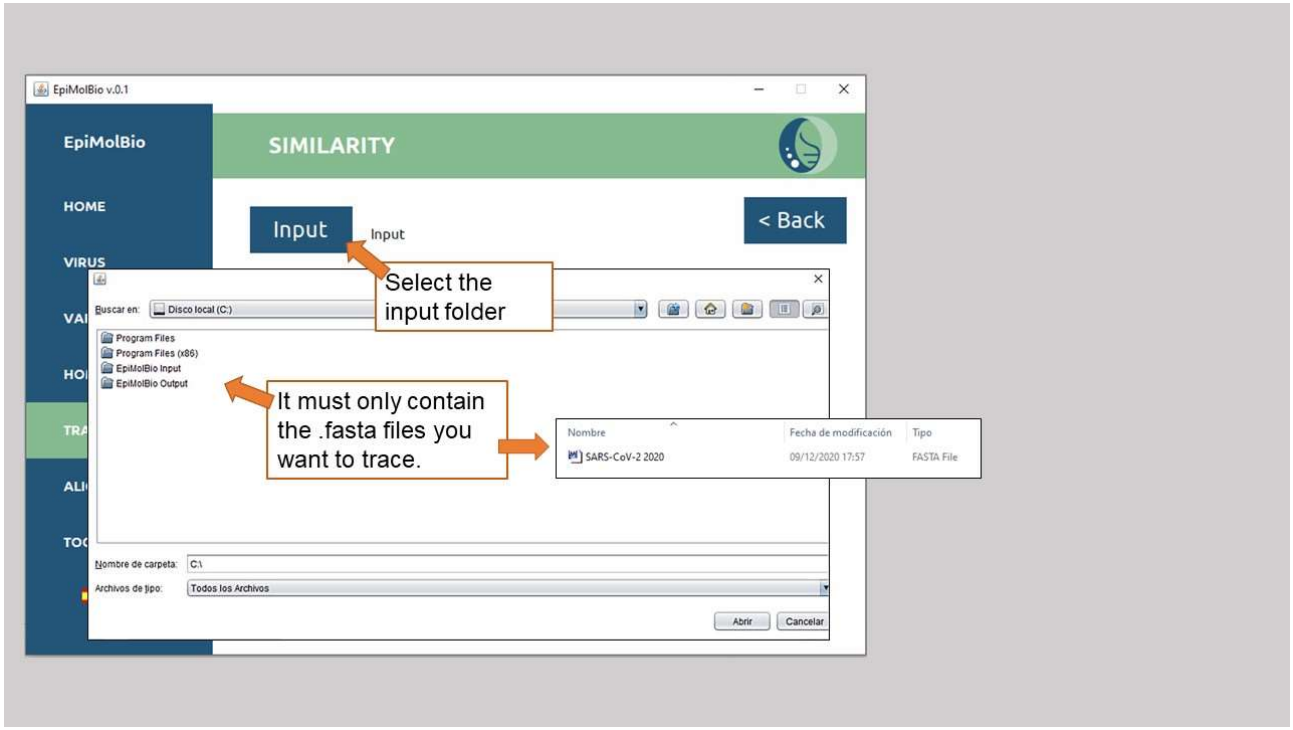
## 1)



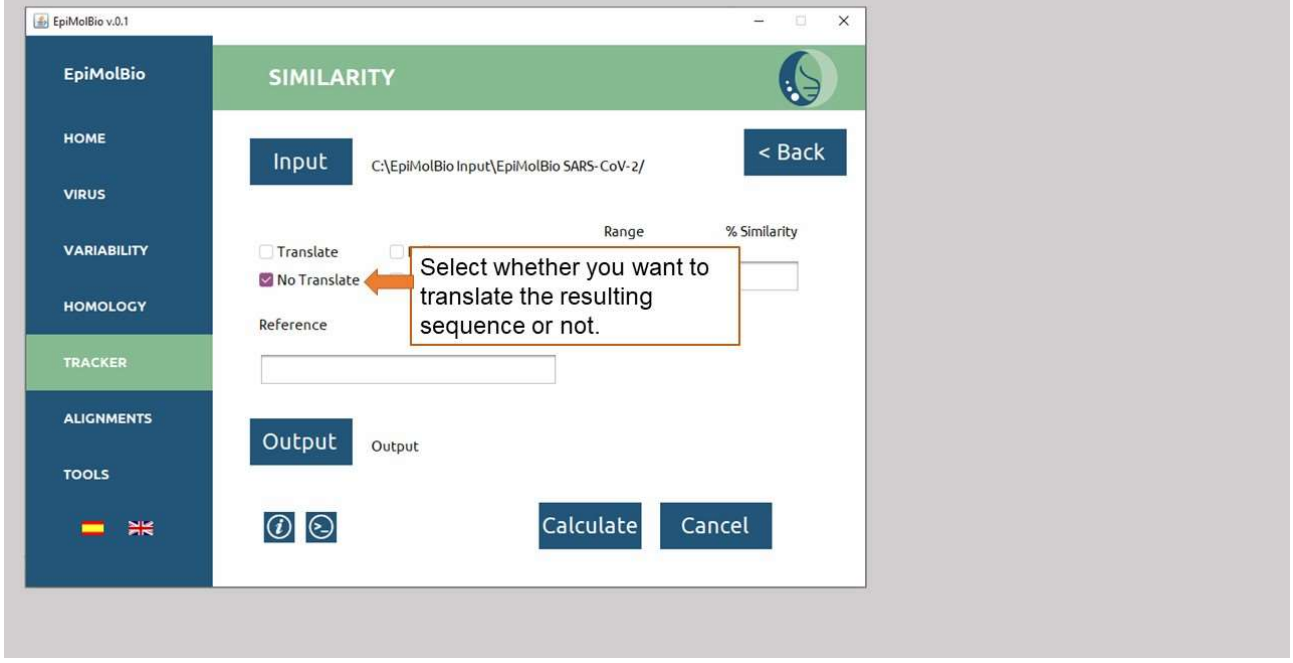

## 2)



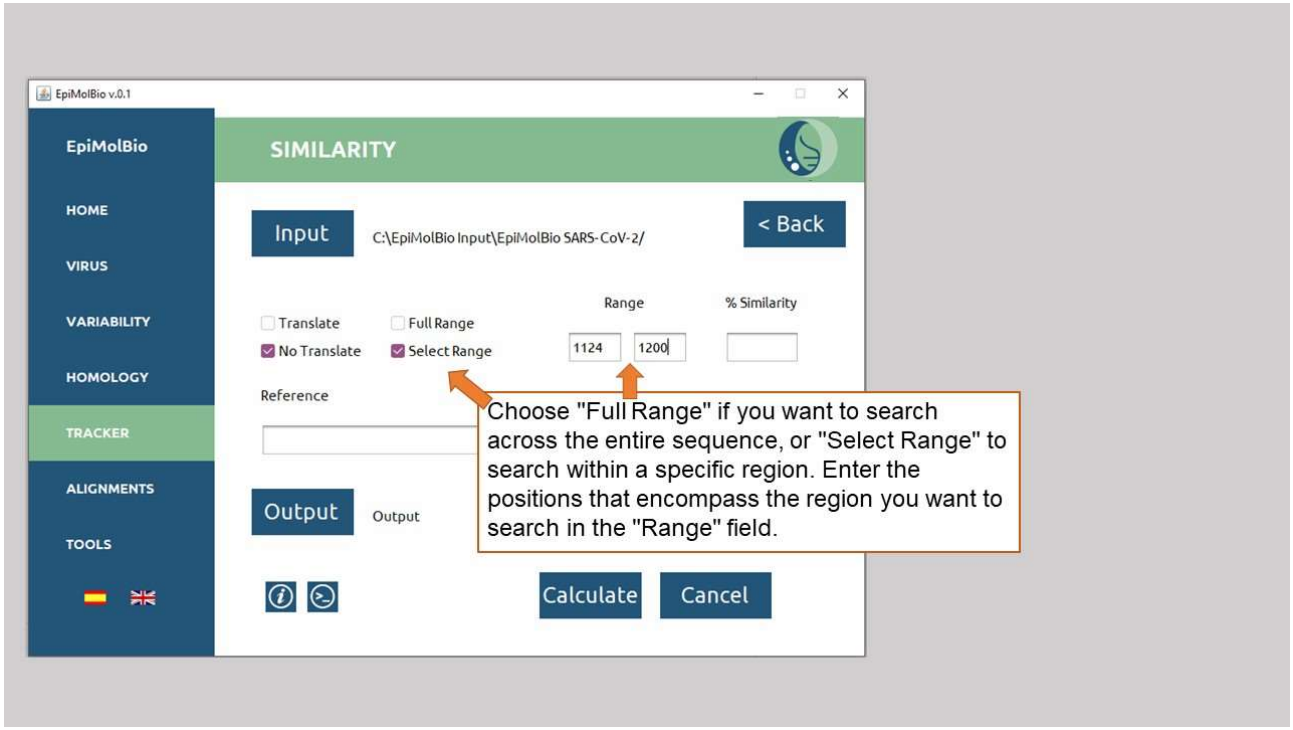
## 3)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
SIMILARITY
Input
Input
< Back
Select the input folder
Buscar en:
Disco local (C:)
Program Files
Program Files (x86)
EpiMolBio Input
EpiMolBio Output
It must only contain the .fasta files you want to trace.
Nombre
Fecha de modificación
Tipo
SARS-CoV-2 2020
09/12/2020 17:57
FASTA File
Nombre de carpeta:
C:\
Archivos de tipo:
Todos los Archivos
Abrir
Cancelar

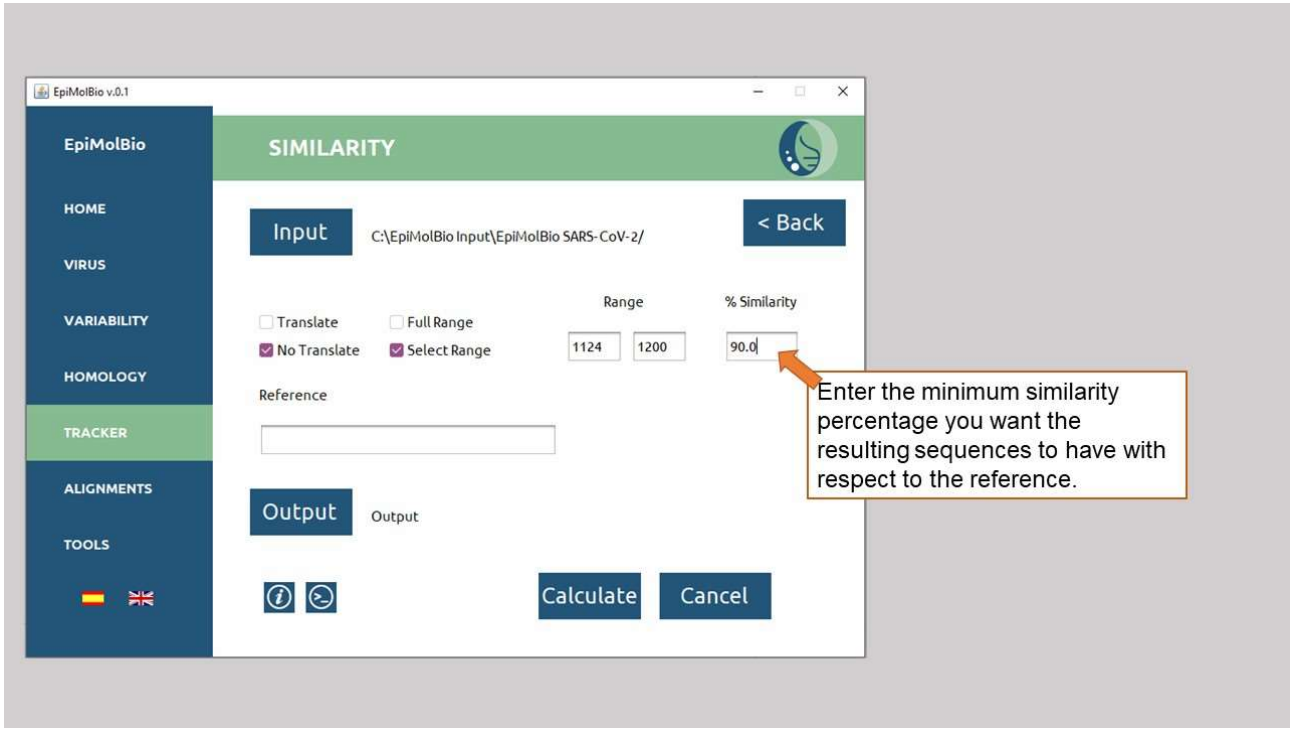
## 4)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
SIMILARITY
Input
C:\EpiMolBio Input\EpiMolBio SARS-CoV-2/
< Back
Range
% Similarity
Translate
No Translate
Select whether you want to translate the resulting sequence or not.
Reference
Output
Output
Calculate
Cancel

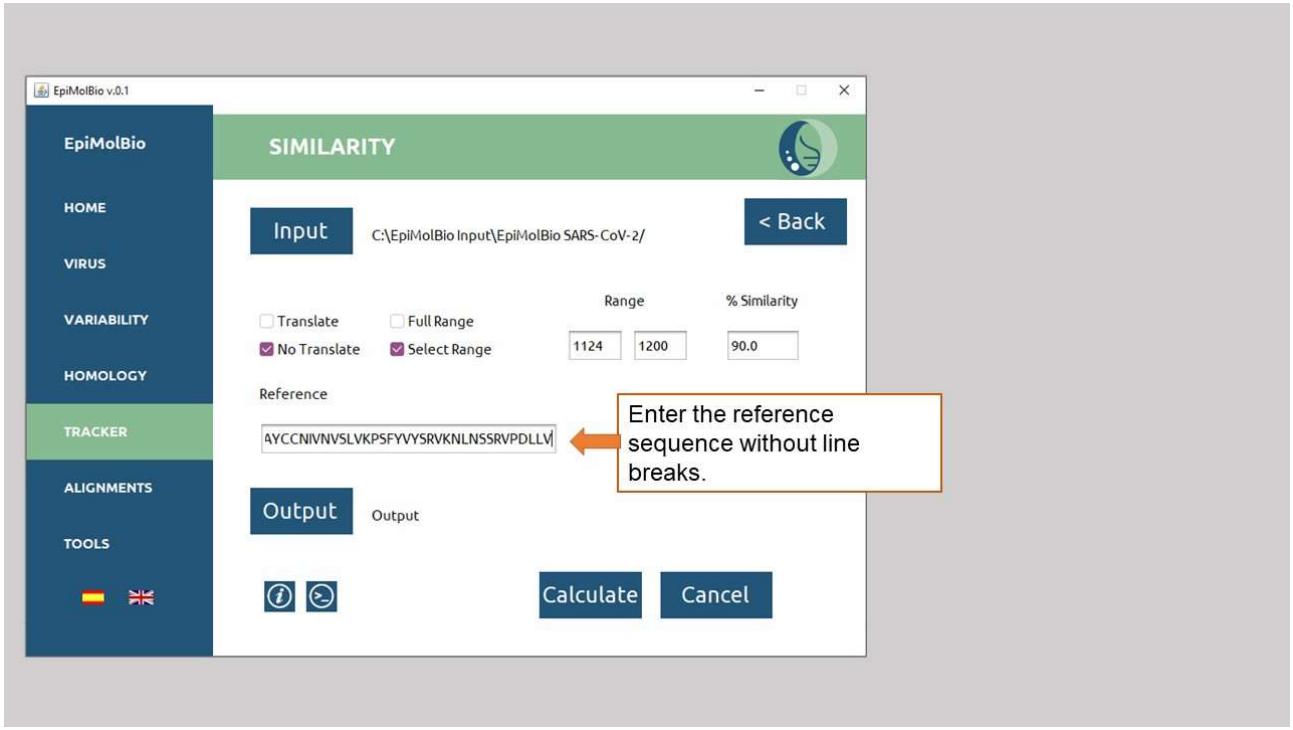
## 5)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
SIMILARITY
Input
C:\EpiMolBio Input\EpiMolBio SARS-CoV-2/
< Back
Range
% Similarity
Translate
Full Range
No Translate
Select Range
1124
1200
Reference
Output
Output
Calculate
Cancel
Choose "Full Range" if you want to search across the entire sequence, or "Select Range" to search within a specific region. Enter the positions that encompass the region you want to search in the "Range" field.

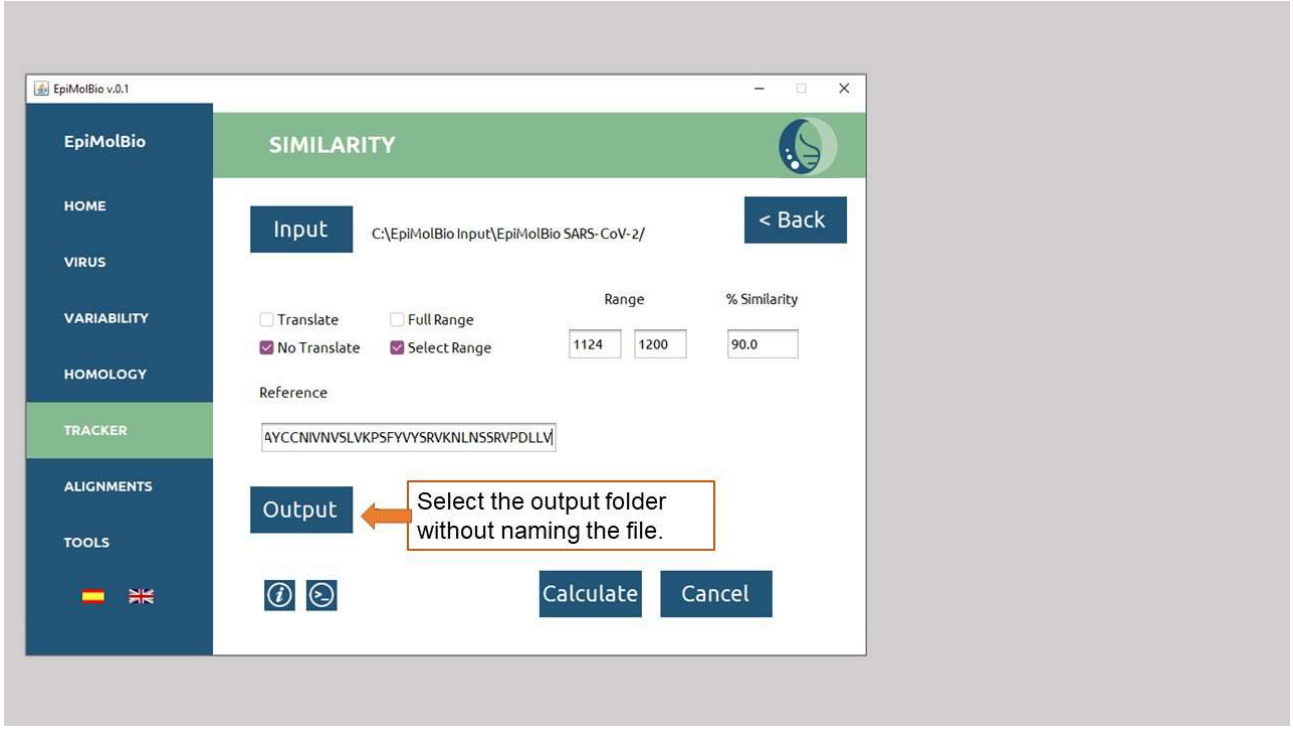
## 6)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
SIMILARITY
Input
C:\EpiMolBio Input\EpiMolBio SARS-CoV-2/
< Back
Range
% Similarity
Translate
Full Range
No Translate
Select Range
1124
1200
90.0
Reference
Output
Output
Calculate
Cancel
Enter the minimum similarity percentage you want the resulting sequences to have with respect to the reference.

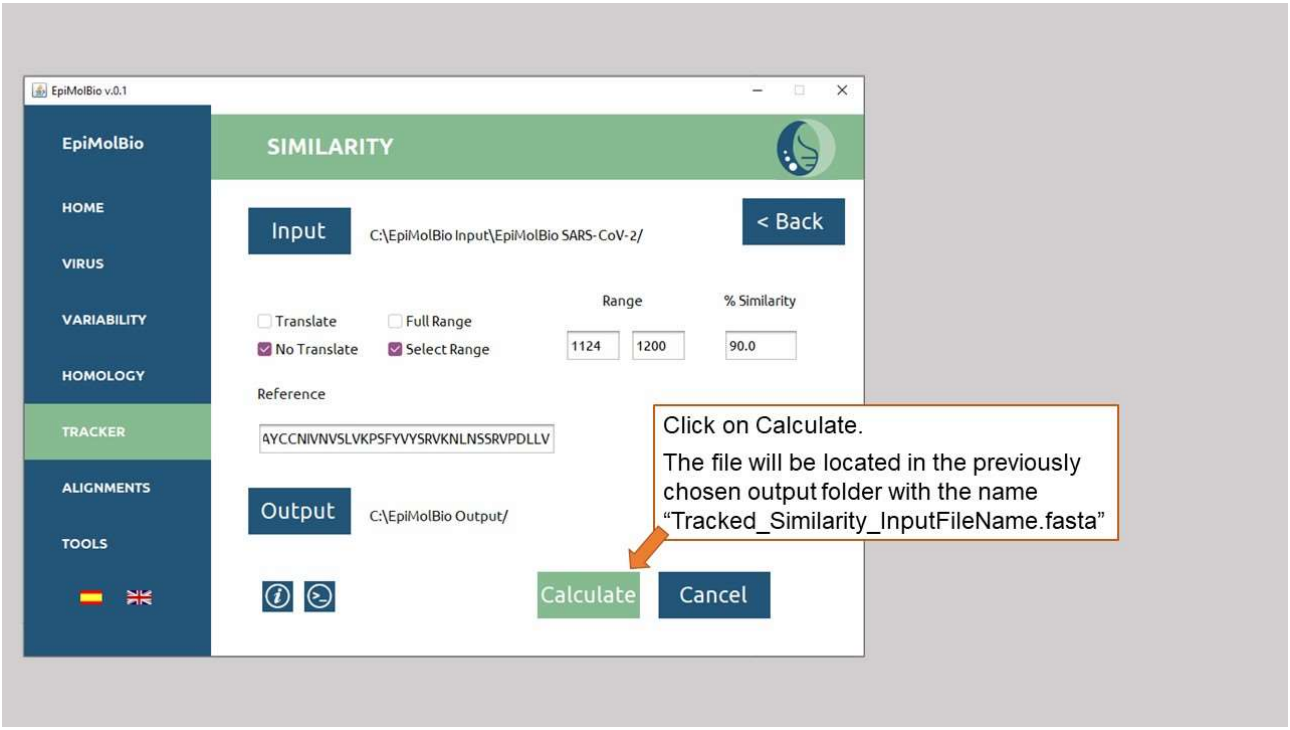
## 7)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
SIMILARITY
Input
C:\EpiMolBio Input\EpiMolBio SARS-CoV-2/
< Back
Range
% Similarity
Translate
Full Range
No Translate
Select Range
1124
1200
90.0
Reference
AYCCNIVNVSLVKPSFYVYSRVKNLNSSRVPDLLV
Enter the reference sequence without line breaks.
Output
Output
Calculate
Cancel

## 8)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
SIMILARITY
Input
C:\EpiMolBio Input\EpiMolBio SARS-CoV-2/
< Back
Range
% Similarity
Translate
Full Range
No Translate
Select Range
1124
1200
90.0
Reference
AYCCNIVNVSLVKPSFYVYSRVKNLNSSRVPDLLV
Output
Select the output folder without naming the file.
Calculate
Cancel

**9)**


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
SIMILARITY
Input
C:\EpiMolBio Input\EpiMolBio SARS-CoV-2/
< Back
Range
% Similarity
Translate
Full Range
No Translate
Select Range
1124
1200
90.0
Reference
AYCCNIVNVSLVKPSFYVYSRVKNLNSSRVPDLLV
Output
C:\EpiMolBio Output/
Calculate
Cancel
Click on Calculate.
The file will be located in the previously chosen output folder with the name "Tracked_Similarity_InputFileName.fasta"

# IV.2.FLANKING

This function **allows you to search for proteins within a set of complete genomic sequences using flanking sequences (preceding and succeeding sequences of around 15 residues) for the target protein**. These flanking sequences should be in amino acids and in the same reading frame as the protein sequence you're searching for. To locate the protein of interest, it must be complete within the input file. Incomplete sequences cannot be located. The output format is a .fasta file containing the found sequences.

The **input** file should be the folder containing exclusively .fasta files with complete nucleotide sequences where you want to perform the search. The input sequences may not be aligned or partially incomplete, but they must have similar lengths. The search will be conducted based on residue location ranges, therefore, if the sequences are not similar, the ranges won't correspond to the regions where the protein of interest is located, and the program will perform the search incorrectly.

In the '**Sequence Type**' field, select amino acids or nucleotides based on whether you want the output file to be translated or not.

In the '**Range**' field, enter the range of the genomic sequence where you want to search for the target protein. For instance, to search for the envelope protein in the SARS-CoV-2 genome, input 26050 in the first range box and 26650 in the second (numbers must be input without commas). The program will search between nucleotides 26,050 and 26,650 of the input genomes.

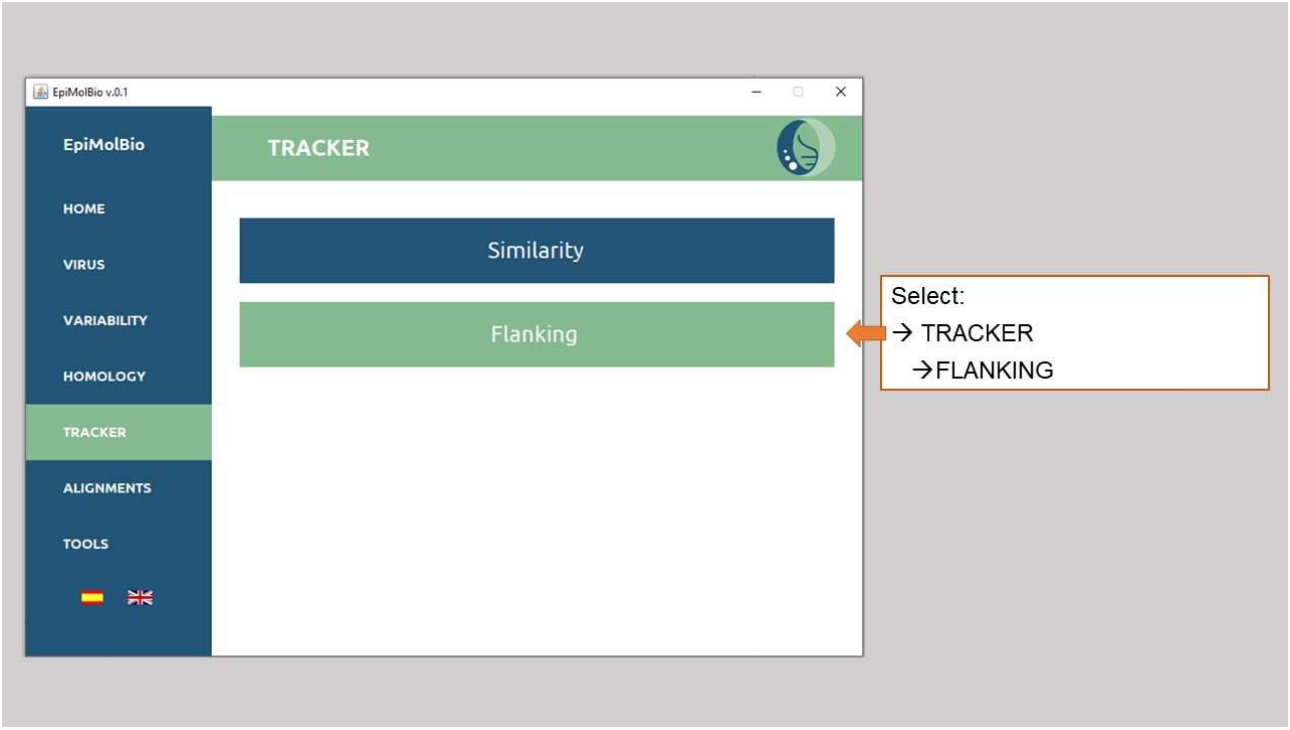
In the '**Size**' field, enter the length of the target protein in amino acids. For this example, enter 75 in this field.

In the '**Flanking Sequences**' field, input the 15 preceding and succeeding amino acids to the target protein you're searching for. For instance, enter TTSVPLAQADEYEL in the first flanking sequences box and TNILYFFCLELF in the second Flanking Sequences box.

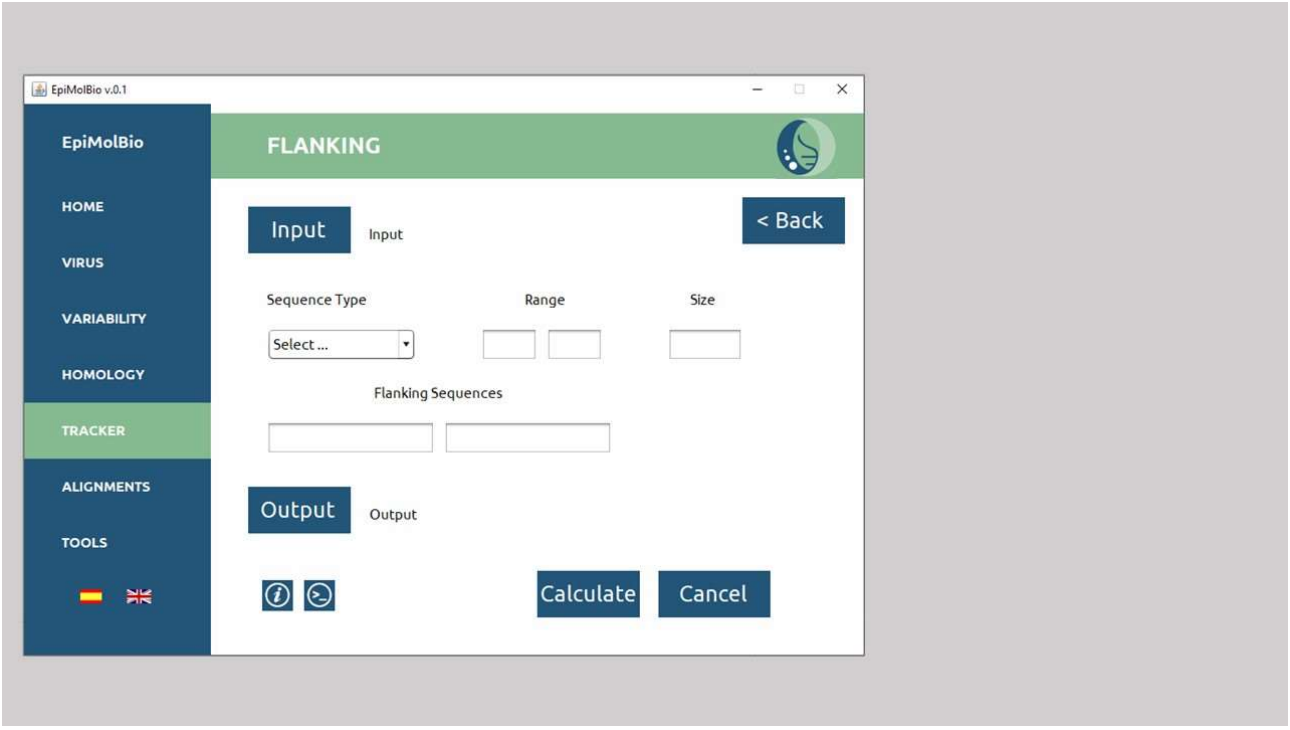
For the **output**, select the output folder where you want the **.fasta** files to appear. The files are automatically named as follows: 'Tracked_Flanking_InputFileName.fasta'.

# Step-by-step:

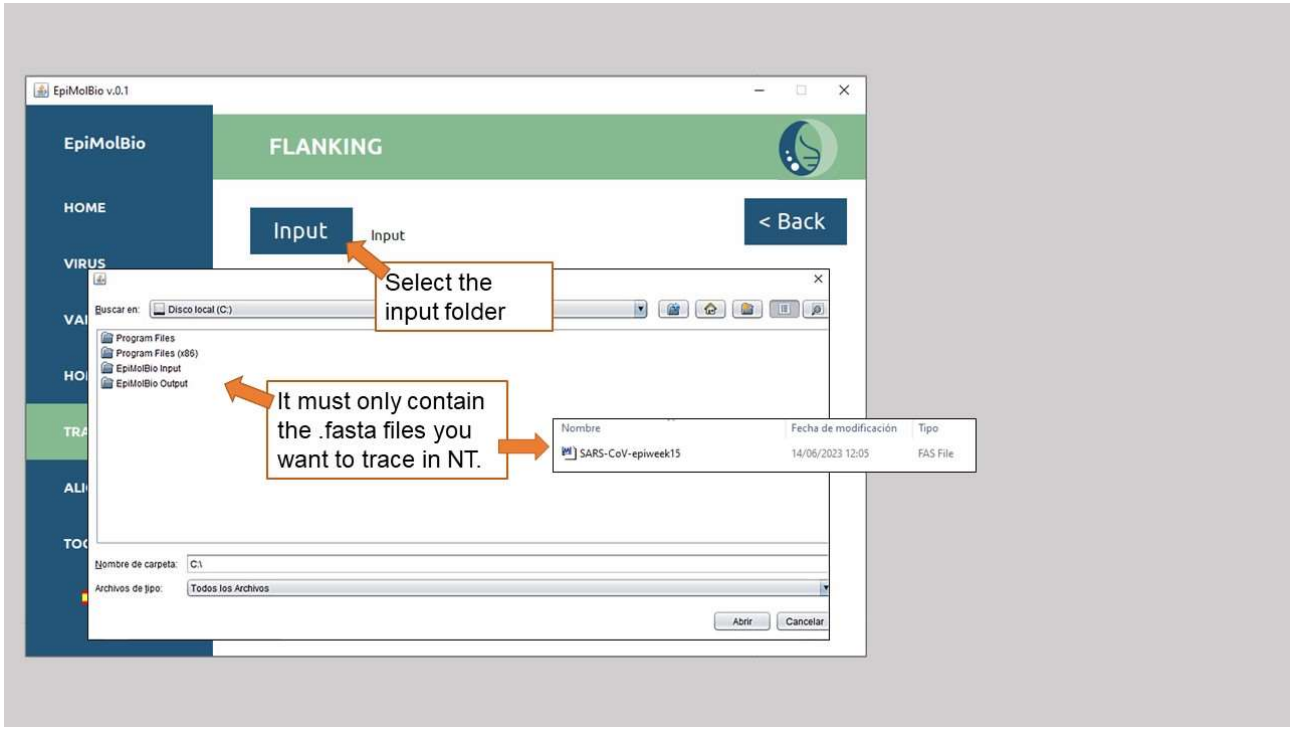
## 1)



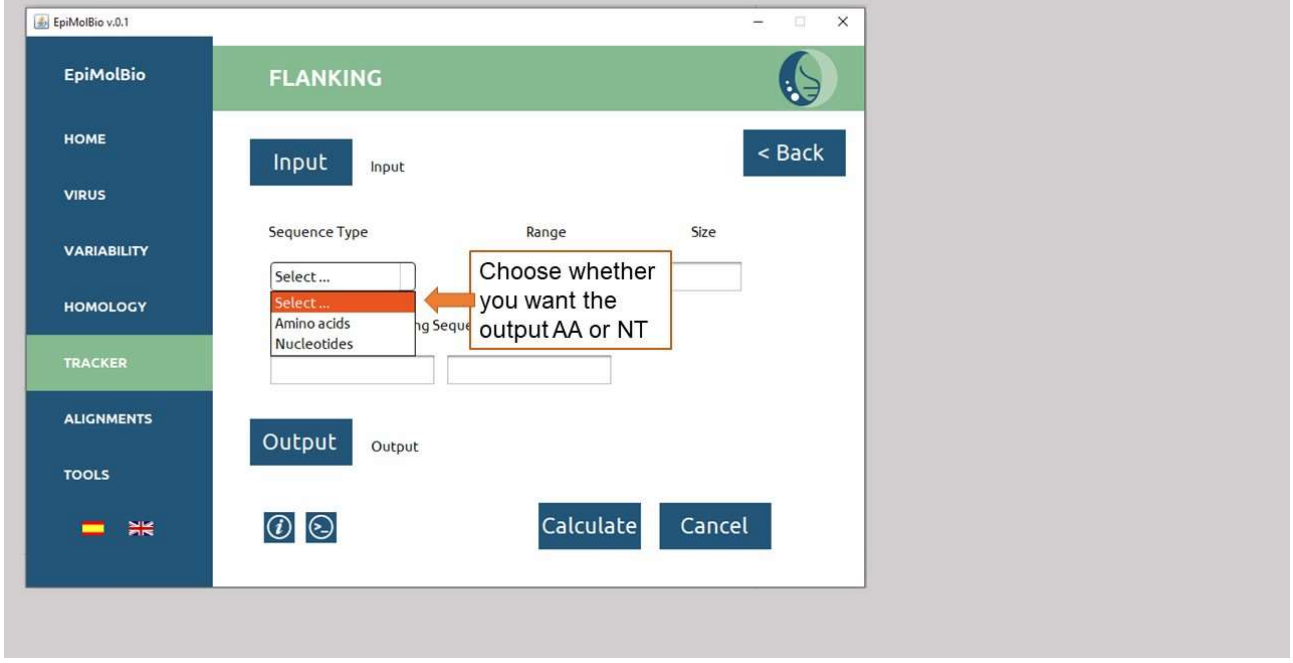

## 2)



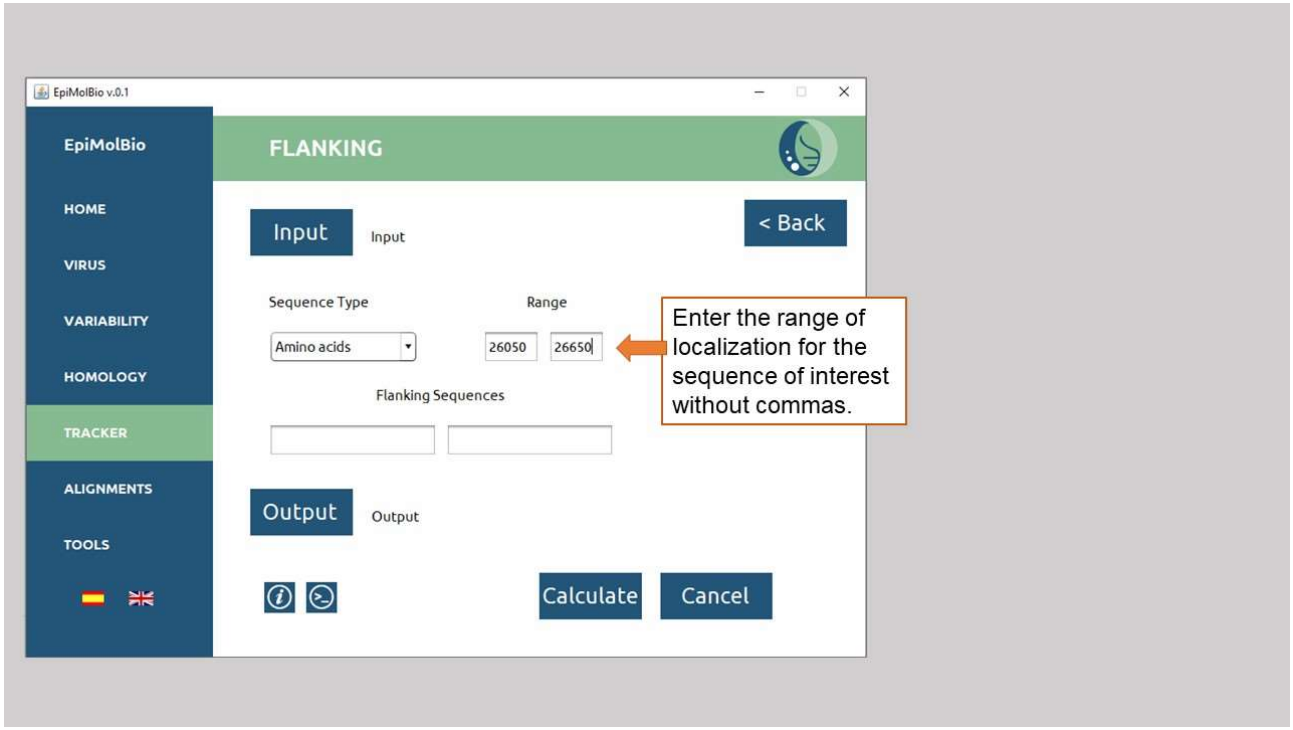
## 3)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
FLANKING
Input
Input
< Back
Select the input folder
Buscar en:
Disco local (C:)
Program Files
Program Files (x86)
EpiMolBio Input
EpiMolBio Output
It must only contain the .fasta files you want to trace in NT.
Nombre
Fecha de modificación
Tipo
SARS-CoV-epiweek15
14/06/2023 12:05
FAS File
Nombre de carpeta:
C:\
Archivos de tipo:
Todos los Archivos
Abrir
Cancelar

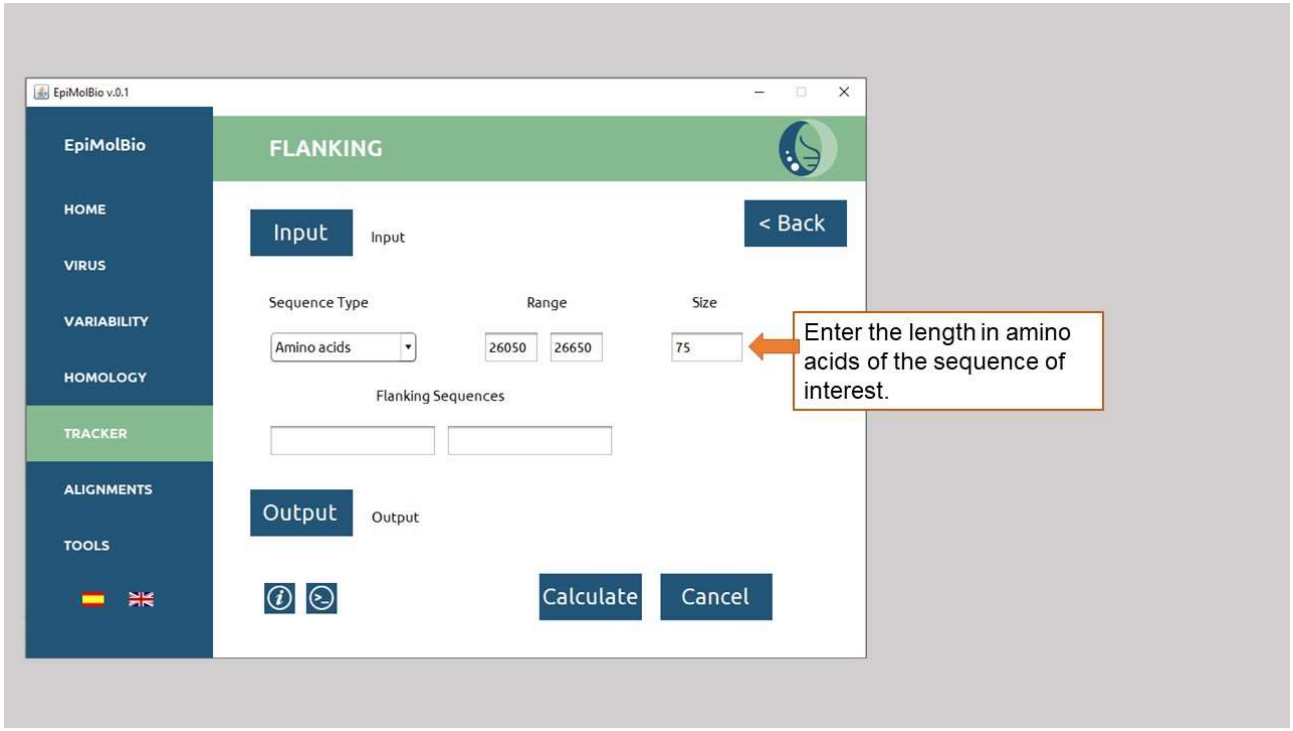
## 4)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
FLANKING
Input
Input
< Back
Sequence Type
Range
Size
Select ...
Select ...
Amino acids
Nucleotides
Choose whether you want the output AA or NT
Output
Output
Calculate
Cancel

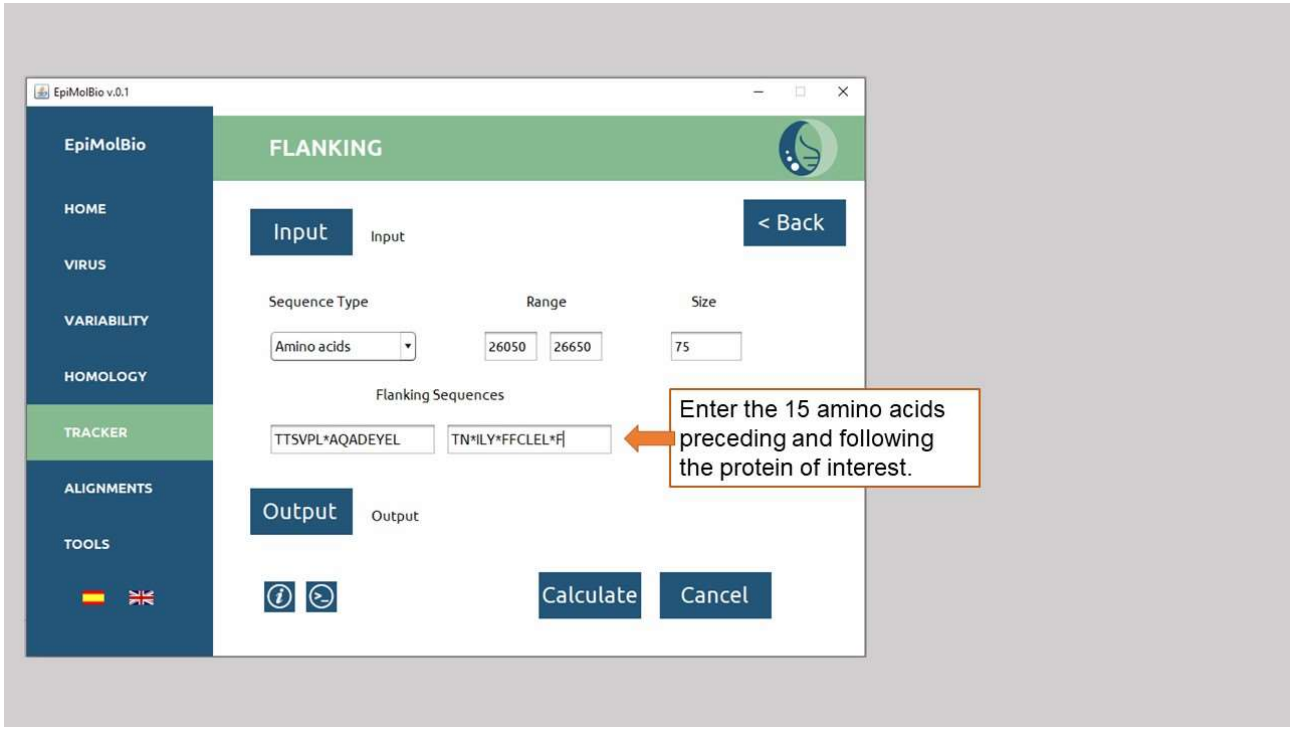
## 5)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
FLANKING
Input
Input
< Back
Sequence Type
Range
Amino acids
26050
26650
Flanking Sequences
Output
Output
Calculate
Cancel
Enter the range of localization for the sequence of interest without commas.

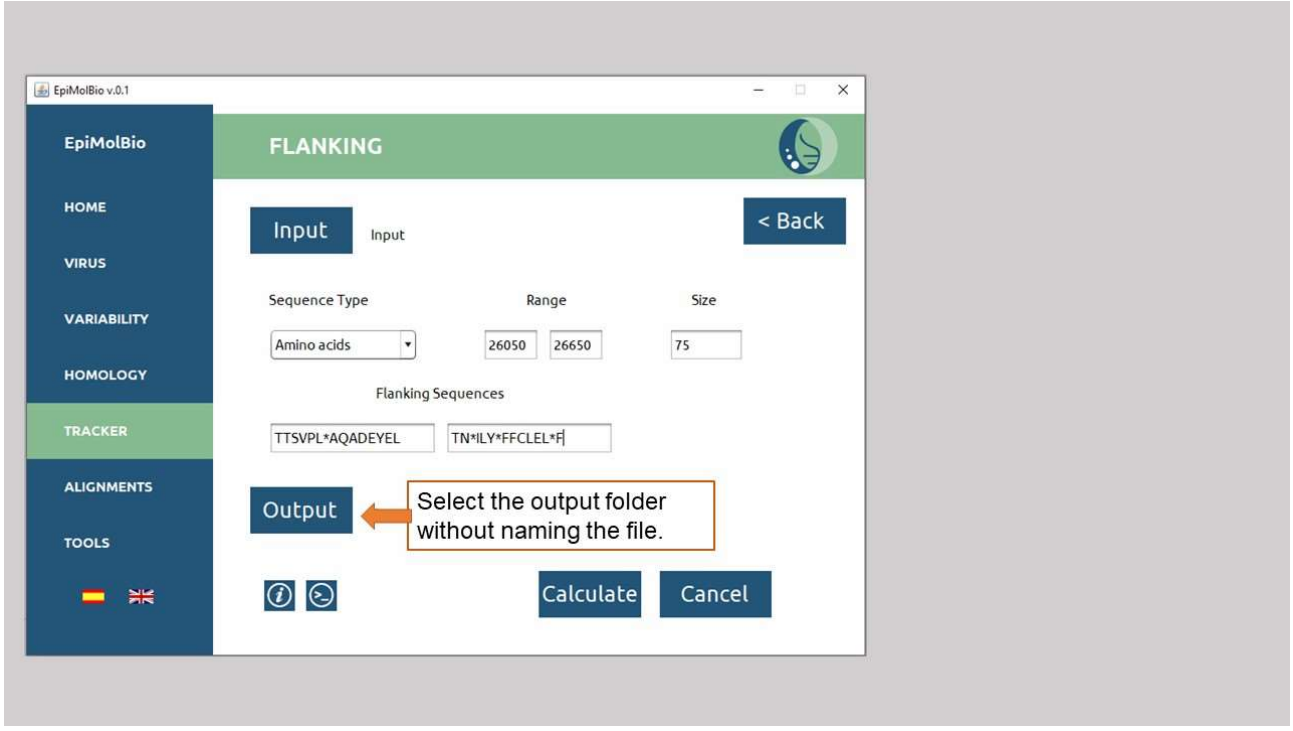
## 6)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
FLANKING
Input
Input
< Back
Sequence Type
Range
Size
Amino acids
26050
26650
75
Flanking Sequences
Output
Output
Calculate
Cancel
Enter the length in amino acids of the sequence of interest.

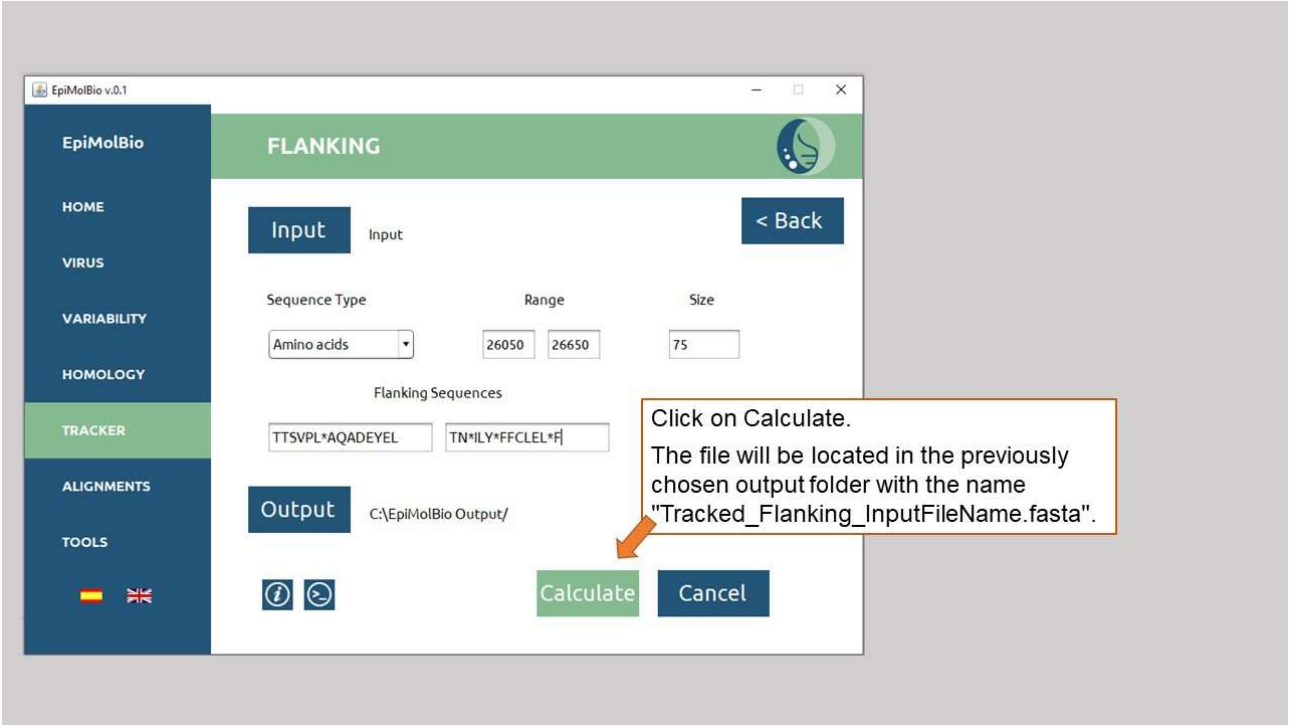
## 7)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
FLANKING
Input
Input
< Back
Sequence Type
Range
Size
Amino acids
26050
26650
75
Flanking Sequences
TTSVPL*AQADEYEL
TN*ILY*FFCLEL*F
Enter the 15 amino acids preceding and following the protein of interest.
Output
Output
Calculate
Cancel

## 8)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
FLANKING
Input
Input
< Back
Sequence Type
Range
Size
Amino acids
26050
26650
75
Flanking Sequences
TTSVPL*AQADEYEL
TN*ILY*FFCLEL*F
Output
Select the output folder without naming the file.
Calculate
Cancel

**9)**


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
FLANKING
Input
Input
< Back
Sequence Type
Range
Size
Amino acids
26050
26650
75
Flanking Sequences
TTSVPL*AQADEYEL
TN*ILY*FFCLEL*F
Output
C:\EpiMolBio Output/
Calculate
Cancel
Click on Calculate.
The file will be located in the previously chosen output folder with the name "Tracked_Flanking_InputFileName.fasta".

# V.ALINGNMENTS

## V.1.MULTIPLE ALIGNMENTS

This function allows aligning amino acid and nucleotide sequences using the publicly available MUSCLE v3.8.31 program by Robert C. Edgar: Edgar, R.C. Nucleic Acids Res 32(5), 1792-97. The alignment is carried out with respect to a reference sequence intruduced by the user, unless 'Preserve Insertions' is marked, in which case a multiple alignment will be performed among the input sequences.

MUSCLE has the limitation of being unable to align more than a few thousand not-too-long sequences. However, with this function, EpiMolBio can align the required sequences, speeding up the process and enabling the simultaneous alignment of thousands of sequences. For this purpose, you can choose to align the sequences in batches or one by one, in both cases with respect to the reference sequence.

The **input** file should be the folder containing exclusively .fasta files in amino acids or nucleotides with the sequences you want to align.

Select the '**Preserve Insertions**' box if you want to retain insertions. Otherwise, the program will automatically remove them. If this option is chosen, no further input is needed except for the output location.

In the '**Sequences per File**' field, enter a number so that the program divides the input file into smaller files with fewer sequences, allowing MUSCLE to align them. It is recommended to enter the **value of 1** in this field if you want to align **incomplete sequences** with respect to the reference or **when the number of sequences is high and their length is relatively large.**

In cases where the sequences have **many mutations** (insertions, deletions, and residue changes) relative to the reference, and the **number of sequences and their length are not excessively large**, it is recommended to enter the **total number of sequences of the input file in the 'Sequences per File'** field.

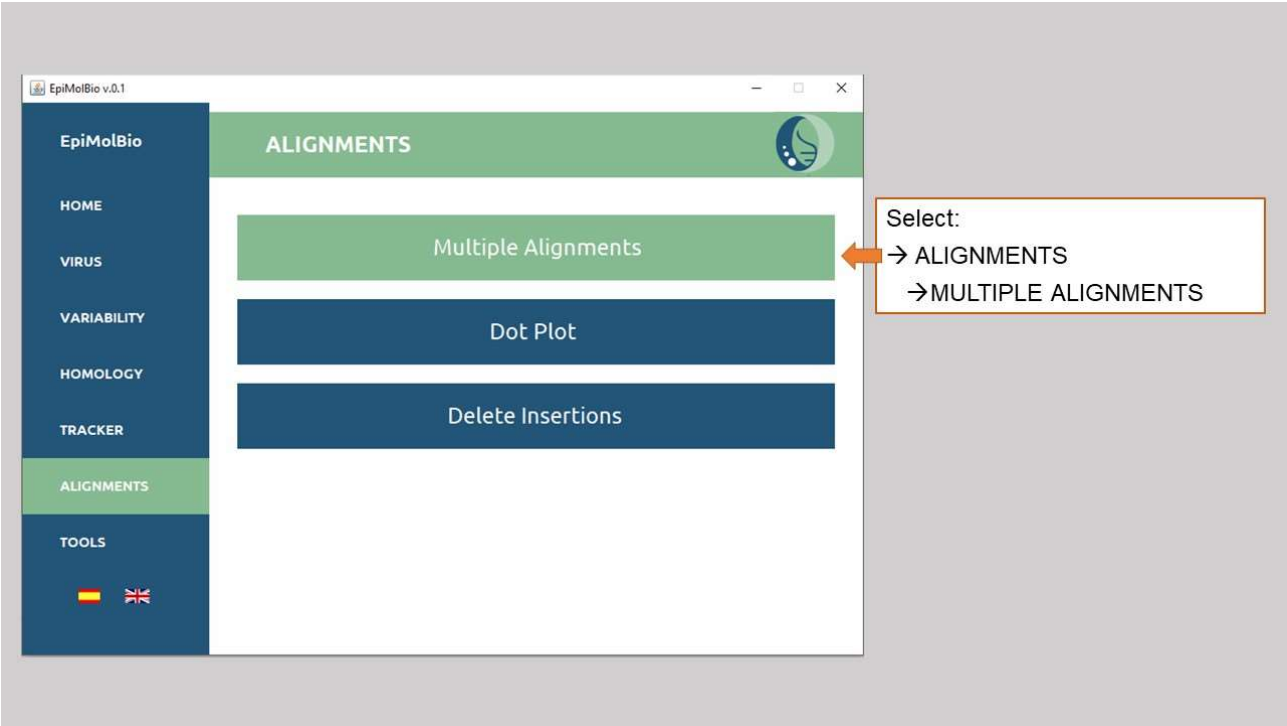
If the sequences have **many mutations** and the file to be aligned is **heavy** due to the sequences being lengthy or a high number of sequences (either a few long sequences or many short sequences), it's preferable to perform a **multiple alignment** by splitting the original file into smaller packages. To do this, input a value between **100-500** in the '**Sequences per File**' field. This way, the input file will be divided into packages with the specified number of sequences in each package. These packages will be aligned individually with respect to the reference sequence, gaps will be removed, and a unified file with the aligned sequences will be generated.

In the '**Reference**' field, input the reference sequence without line breaks in amino acids or nucleotides, depending on the format of the input file.

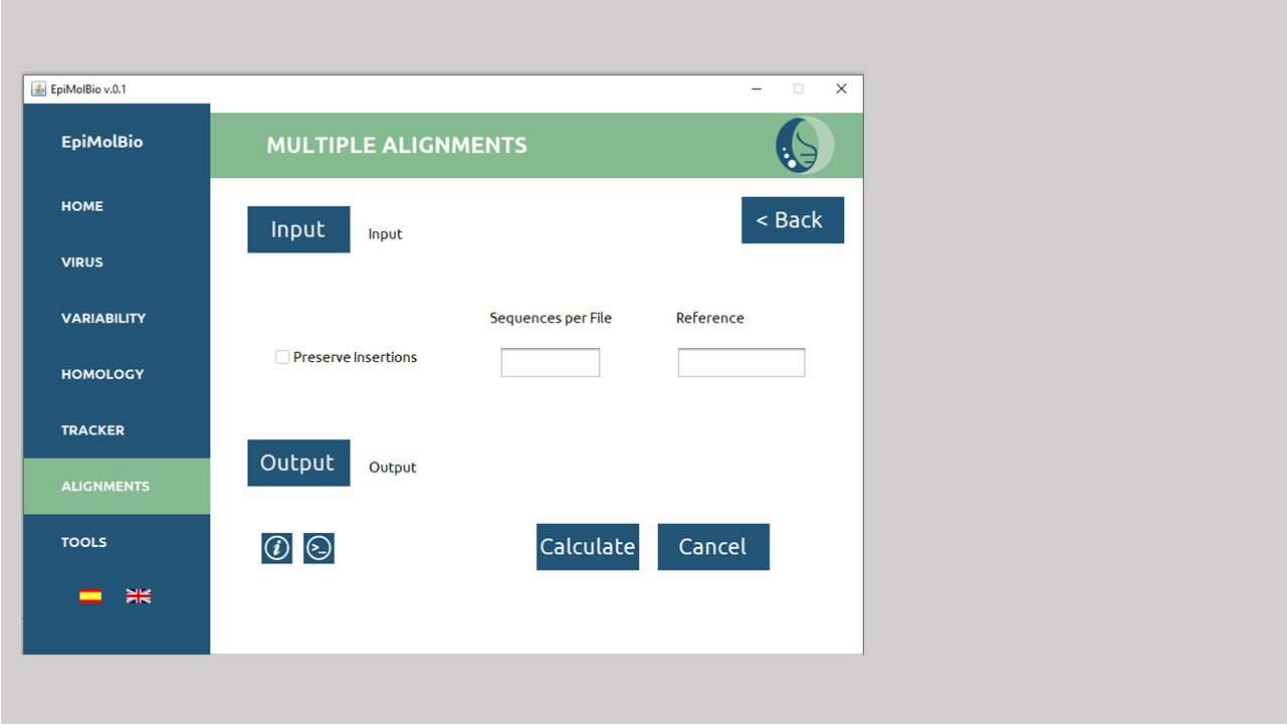
For the **output**, select the output folder where you want the aligned sequence .fasta files to appear. The files will be automatically named as follows: Aligned_InputFileName.fasta.

# Step-by-step:

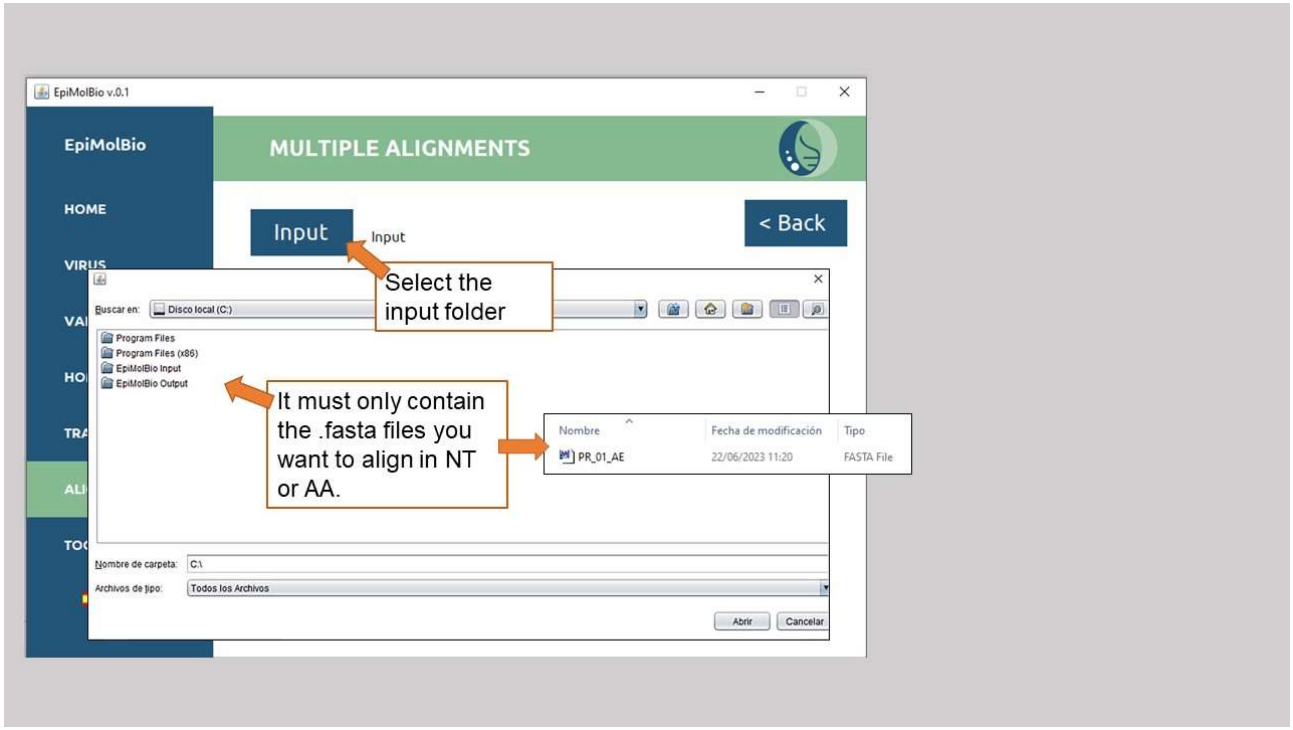
## 1)



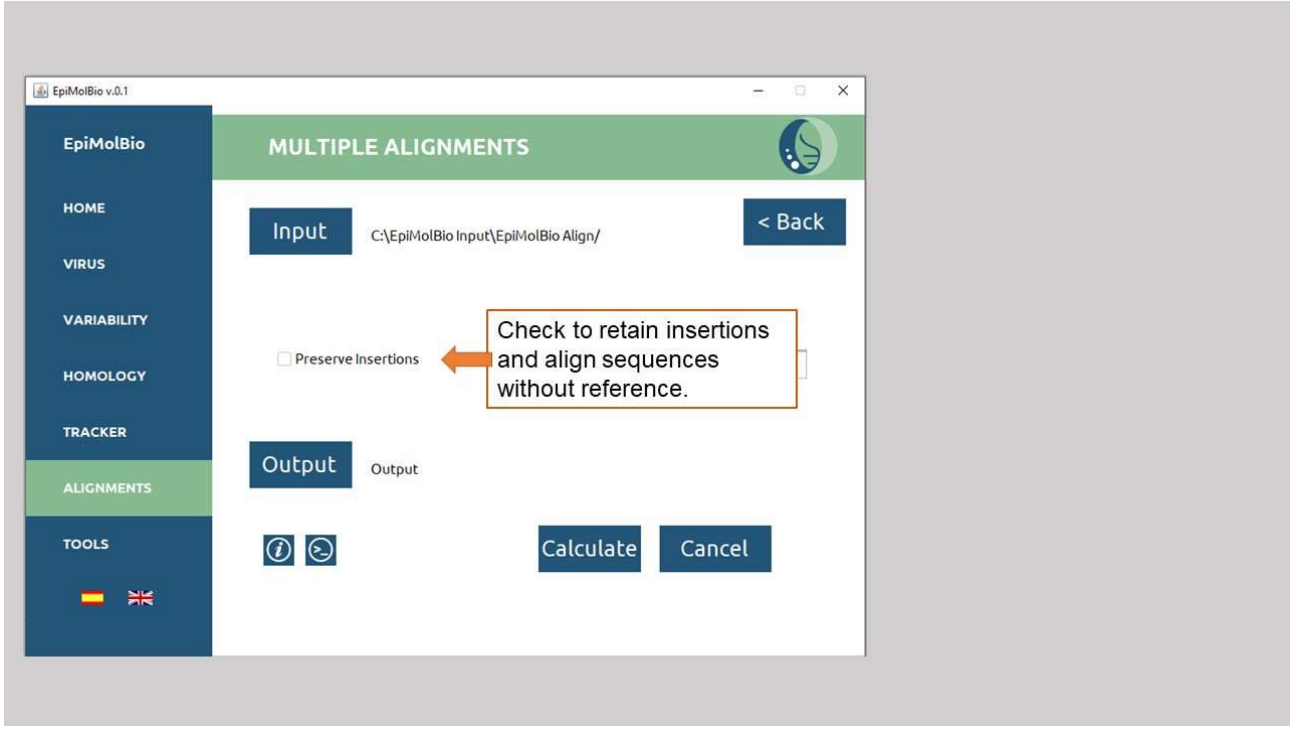

## 2)



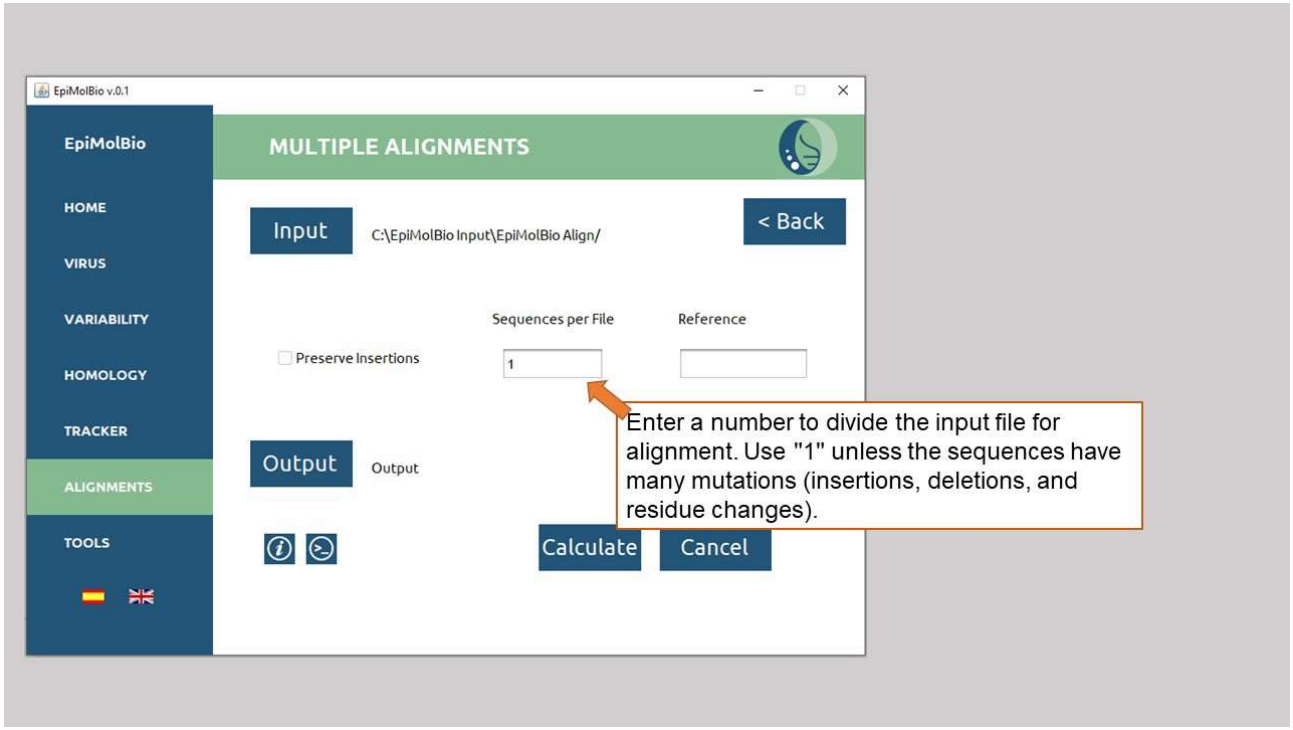
## 3)


EpiMolBio v.0.1
EpiMolBio
HOME
MULTIPLE ALIGNMENTS
Input
Input
< Back
Select the input folder
Buscar en:
Disco local (C:)
Program Files
Program Files (x86)
EpiMolBio Input
EpiMolBio Output
It must only contain the .fasta files you want to align in NT or AA.
Nombre
Fecha de modificación
Tipo
PR_01_AE
22/06/2023 11:20
FASTA File
Nombre de carpeta:
C:\
Archivos de tipo:
Todos los Archivos
Abrir
Cancelar

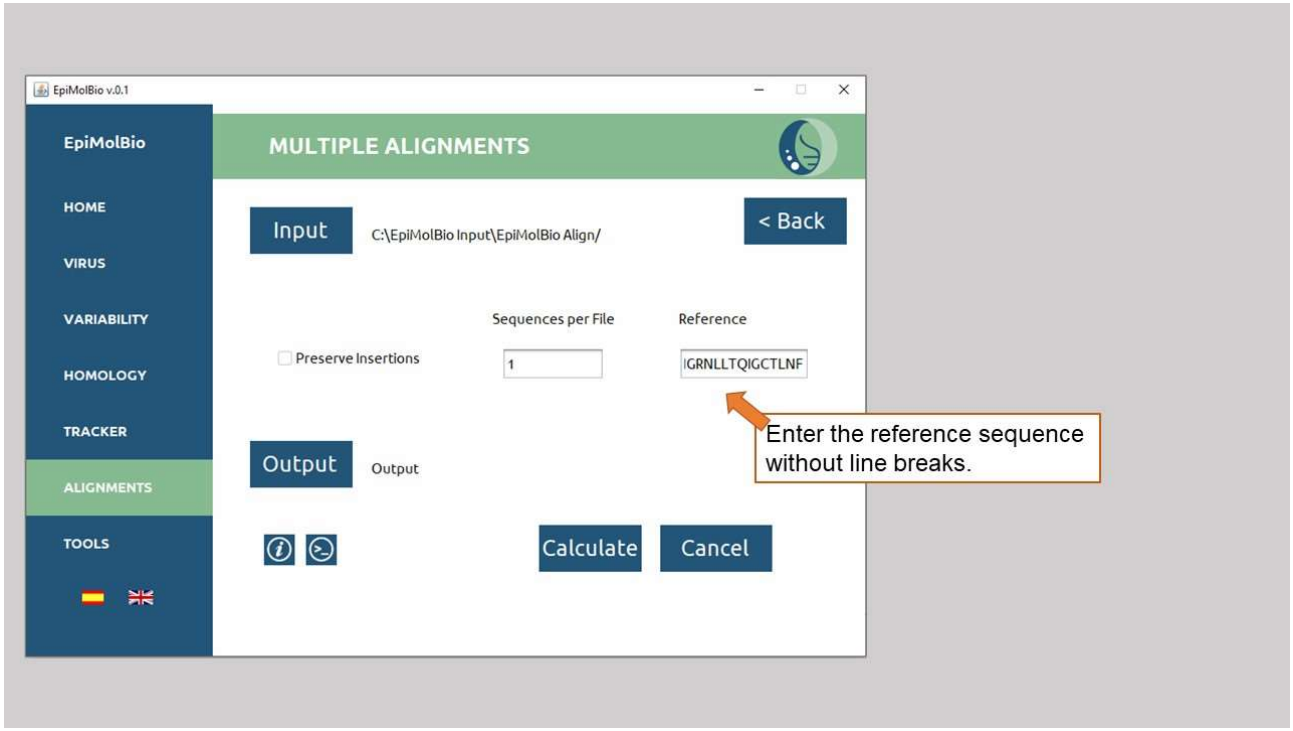
## 4)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
MULTIPLE ALIGNMENTS
Input
C:\EpiMolBio Input\EpiMolBio Align/
< Back
Preserve Insertions
Check to retain insertions and align sequences without reference.
Output
Output
Calculate
Cancel

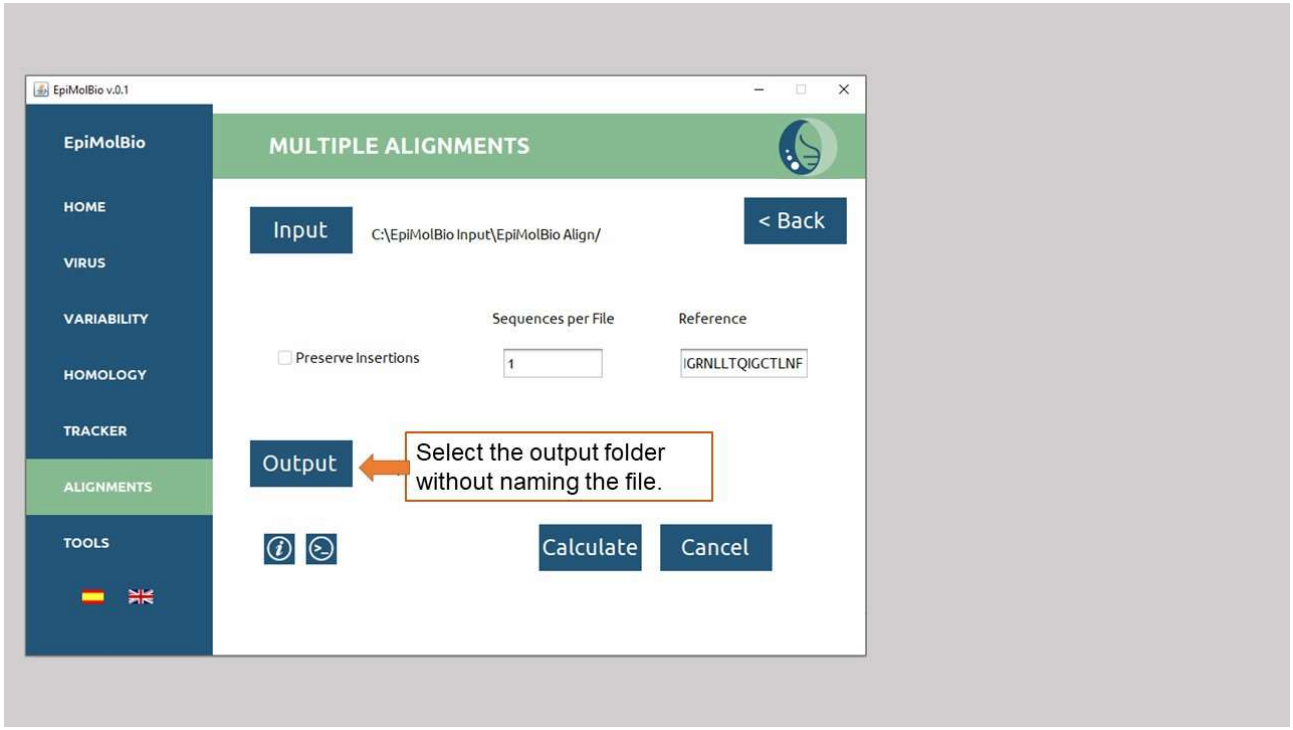
## 5)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
MULTIPLE ALIGNMENTS
Input
C:\EpiMolBio Input\EpiMolBio Align/
< Back
Sequences per File
Reference
Preserve Insertions
1
Output
Output
Calculate
Cancel
Enter a number to divide the input file for alignment. Use "1" unless the sequences have many mutations (insertions, deletions, and residue changes).

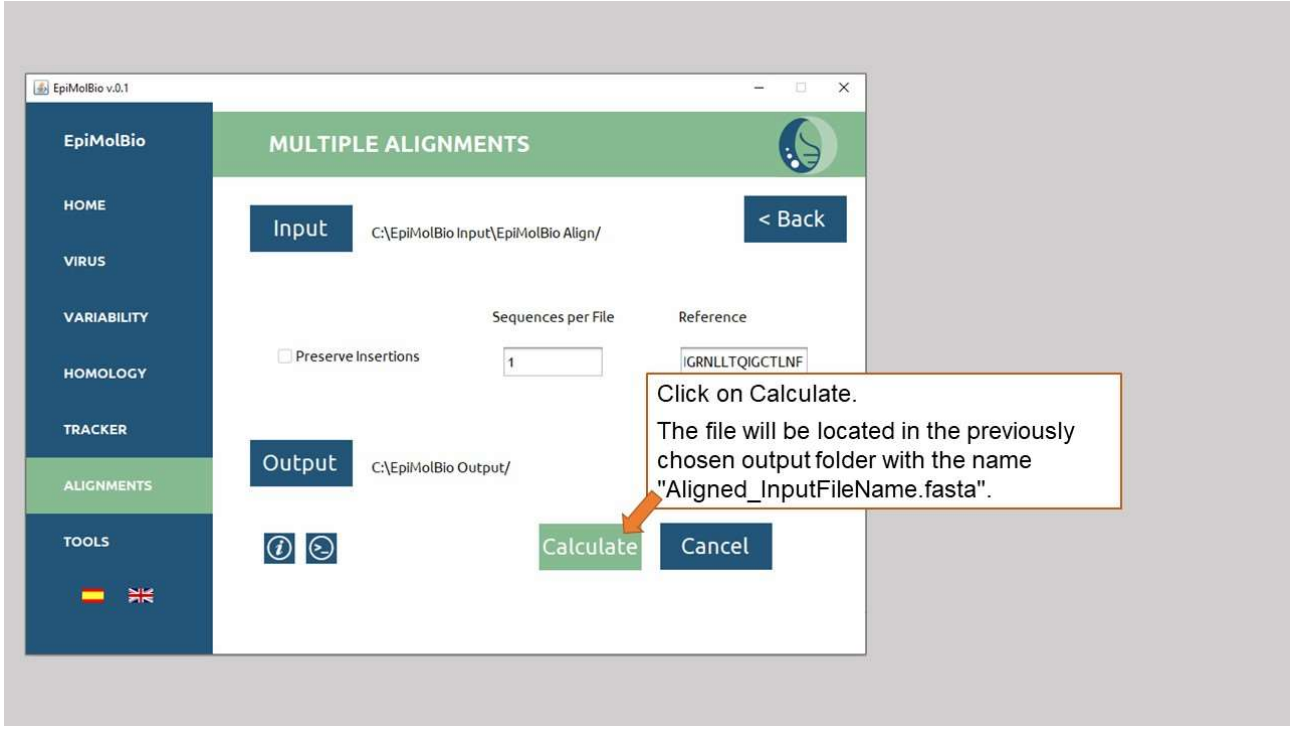
## 6)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
MULTIPLE ALIGNMENTS
Input
C:\EpiMolBio Input\EpiMolBio Align/
< Back
Sequences per File
Reference
Preserve Insertions
1
IGRNLLTQIGCTLNF
Output
Output
Calculate
Cancel
Enter the reference sequence without line breaks.

## 7)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
MULTIPLE ALIGNMENTS
Input
C:\EpiMolBio Input\EpiMolBio Align/
< Back
Sequences per File
Reference
Preserve Insertions
1
IGRNLLTQIGCTLNF
Output
Select the output folder without naming the file.
Calculate
Cancel

## 8)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
MULTIPLE ALIGNMENTS
Input
C:\EpiMolBio Input\EpiMolBio Align/
< Back
Sequences per File
Reference
Preserve Insertions
1
IGRNLLTQIGCTLNF
Output
C:\EpiMolBio Output/
Click on Calculate.
The file will be located in the previously chosen output folder with the name "Aligned_InputFileName.fasta".
Calculate
Cancel

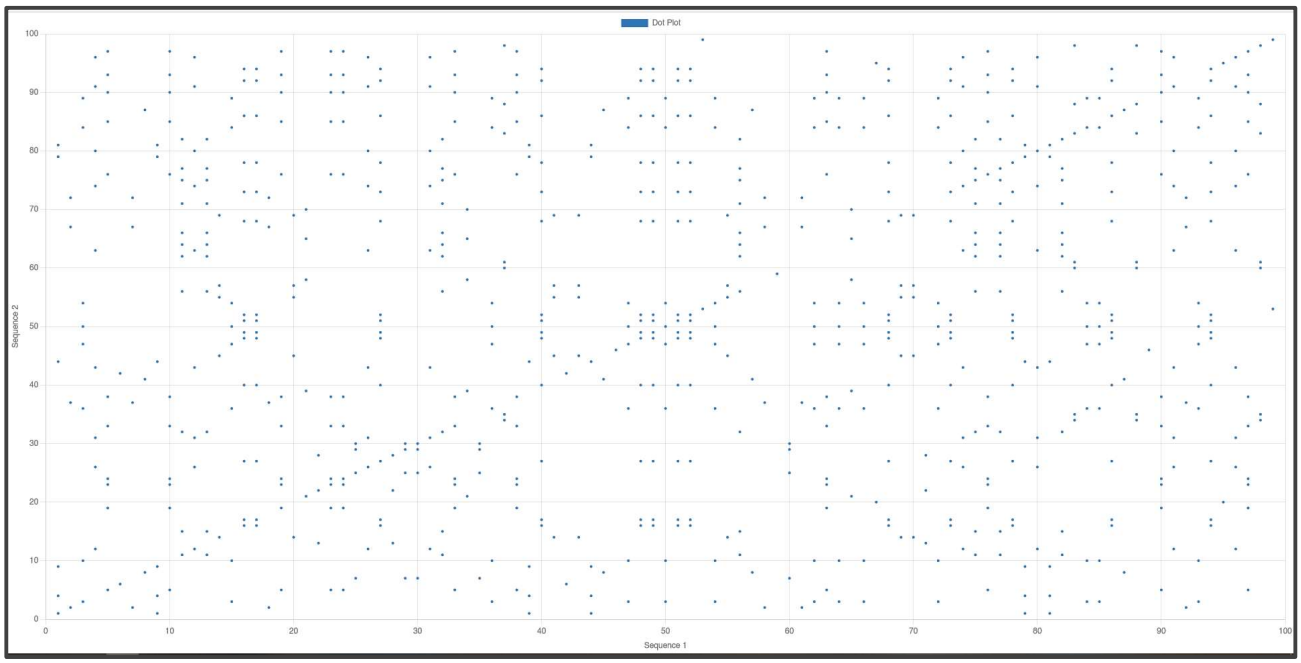
# V.2.DOT PLOT

This function allows you to generate a graphical representation for comparing DNA or protein sequences. The dot plot is constructed by plotting points on a two-dimensional matrix, where each axis represents a sequence. Each point in the dot plot is placed at position (x, y) when corresponding residues in the sequences match at those positions. A point is placed if there is a match; otherwise, the position is left empty.

By observing the dot plot, patterns such as regions of high similarity or repetitions in the sequences can be identified. Dot plots can also help detect inversions, deletions, or insertions in the sequences.

The **output** format is a graphical representation in .html format showing the compared sequences and their positions every 10.

Example of the output format for a Dot Plot analysis:




In the '**Sequence 1**' field, input the first sequence to be compared in nucleotides or amino acids without line breaks or spaces.

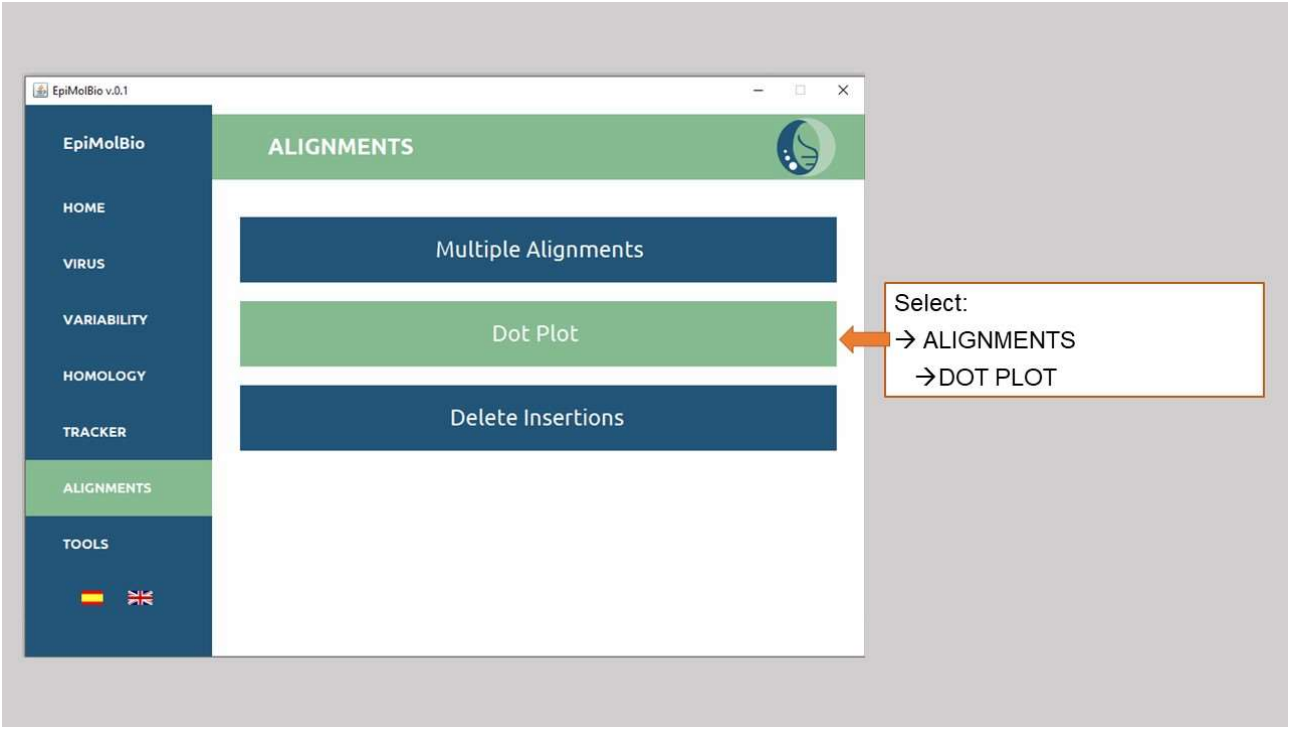
In the '**Sequence 2**' field, input the second sequence to be compared in nucleotides or amino acids without line breaks or spaces.

Both sequences should be in the same format: either both in nucleotides or both in amino acids.

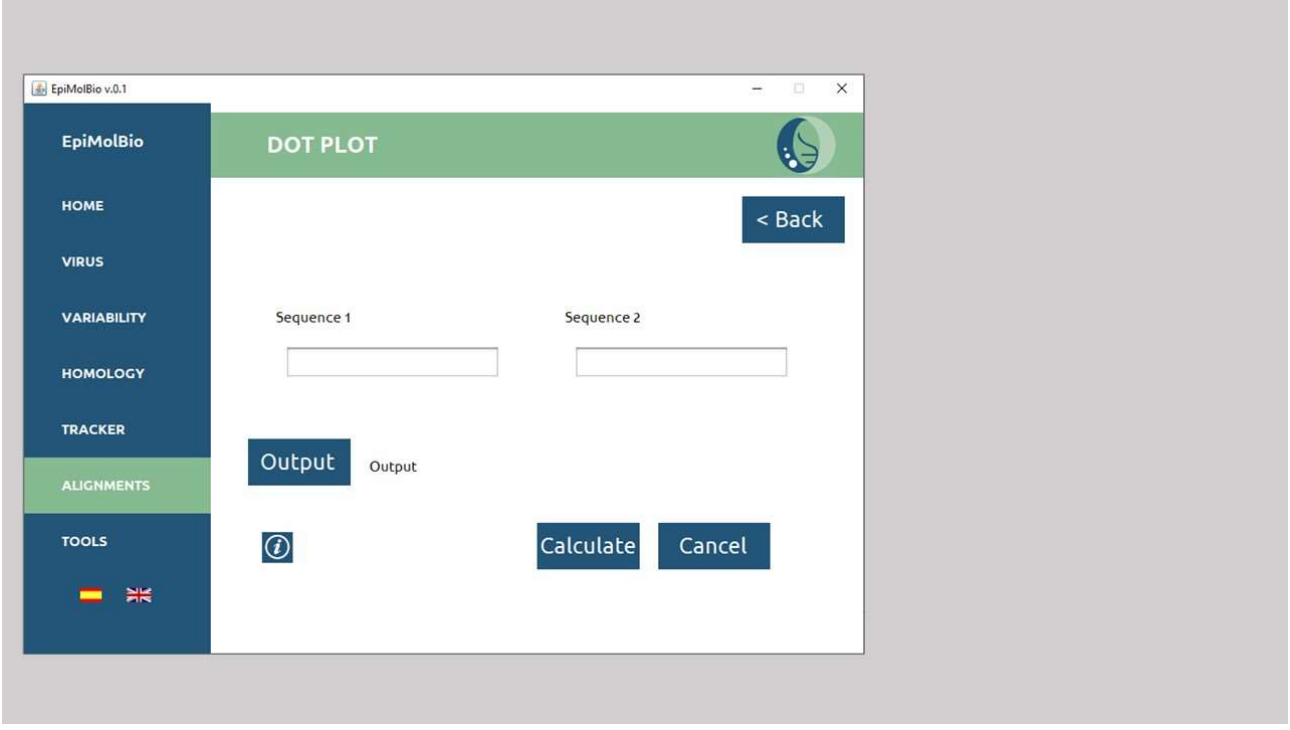
For the **output**, select the output folder where you want the .html file to appear and name the file by adding '.html' at the end.

# Step-by-step:

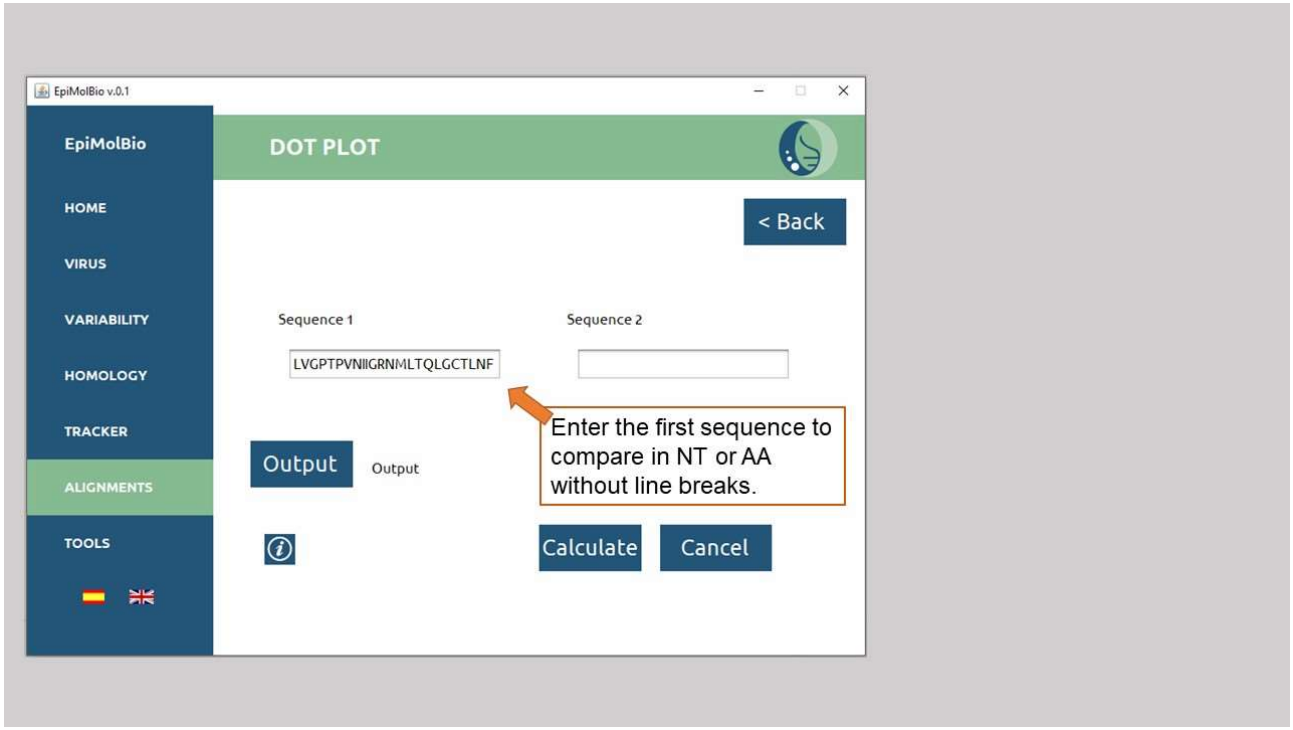
## 1)



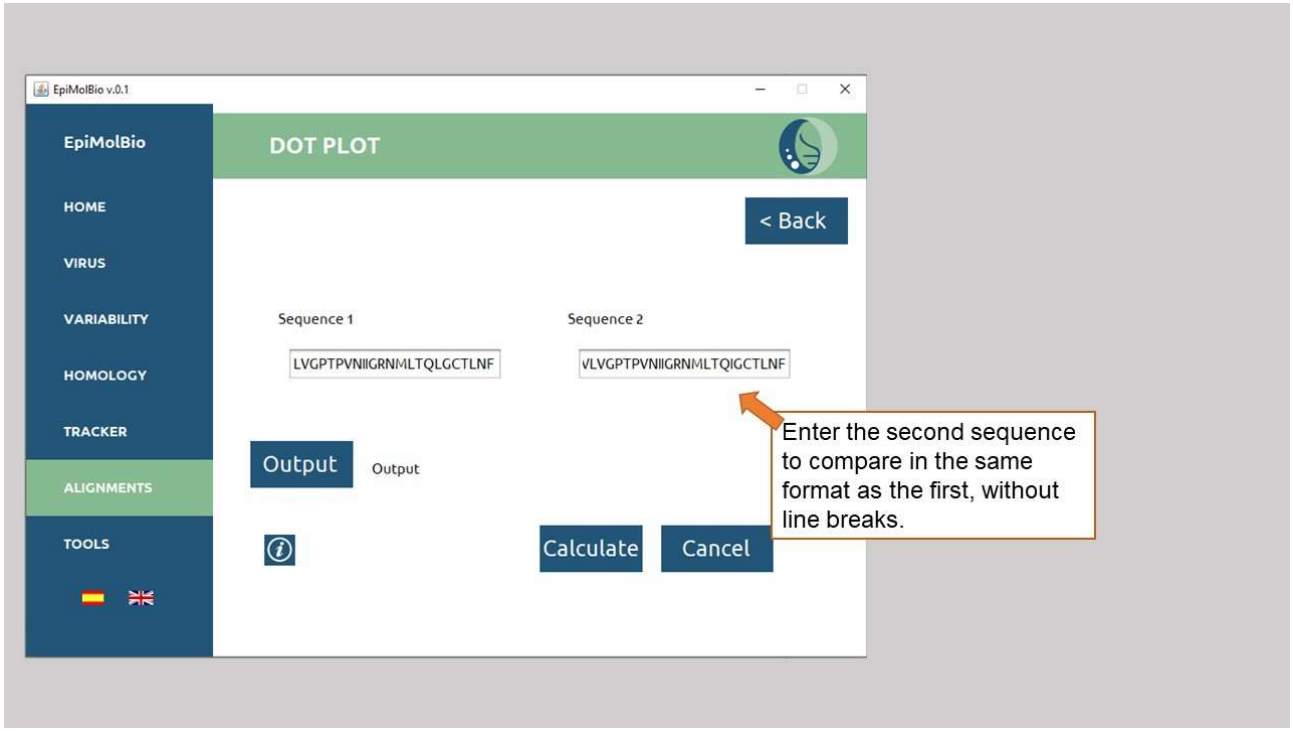

## 2)



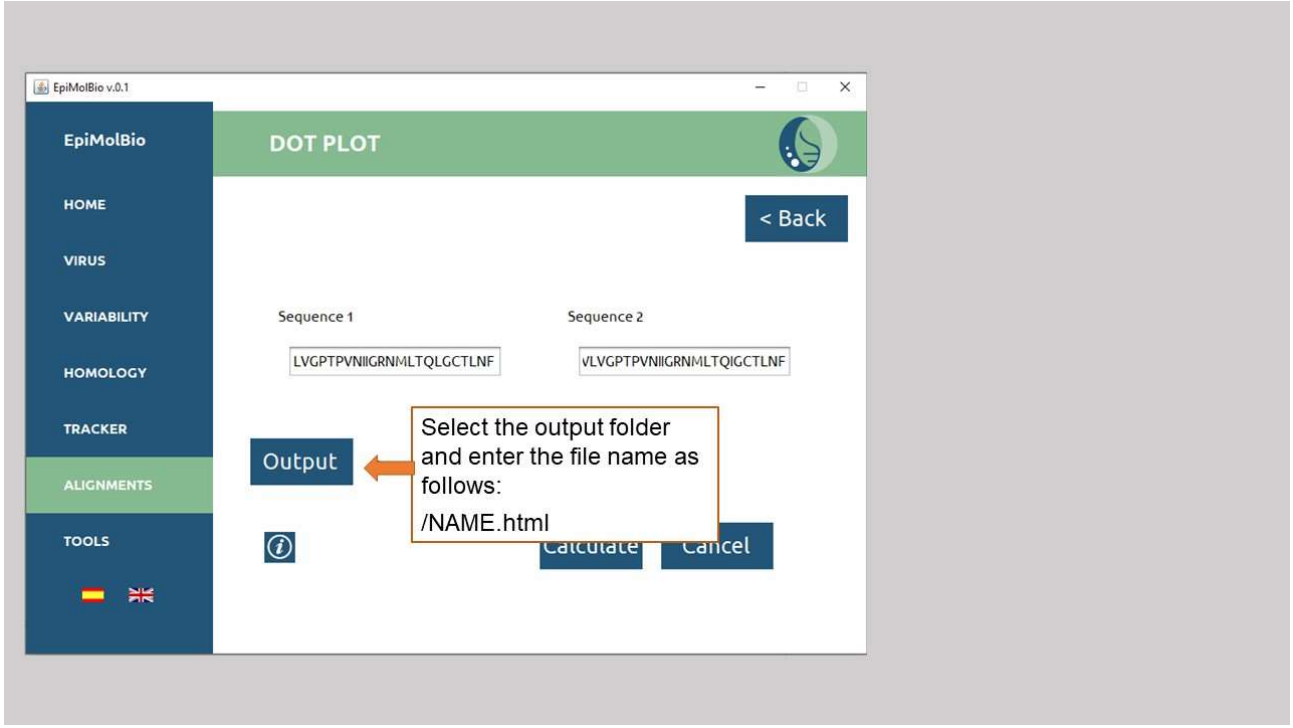
## 3)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
DOT PLOT
< Back
Sequence 1
Sequence 2
LVGPTPVNIIGRNMLTQLGCTLNF
Enter the first sequence to compare in NT or AA without line breaks.
Output
Output
Calculate
Cancel

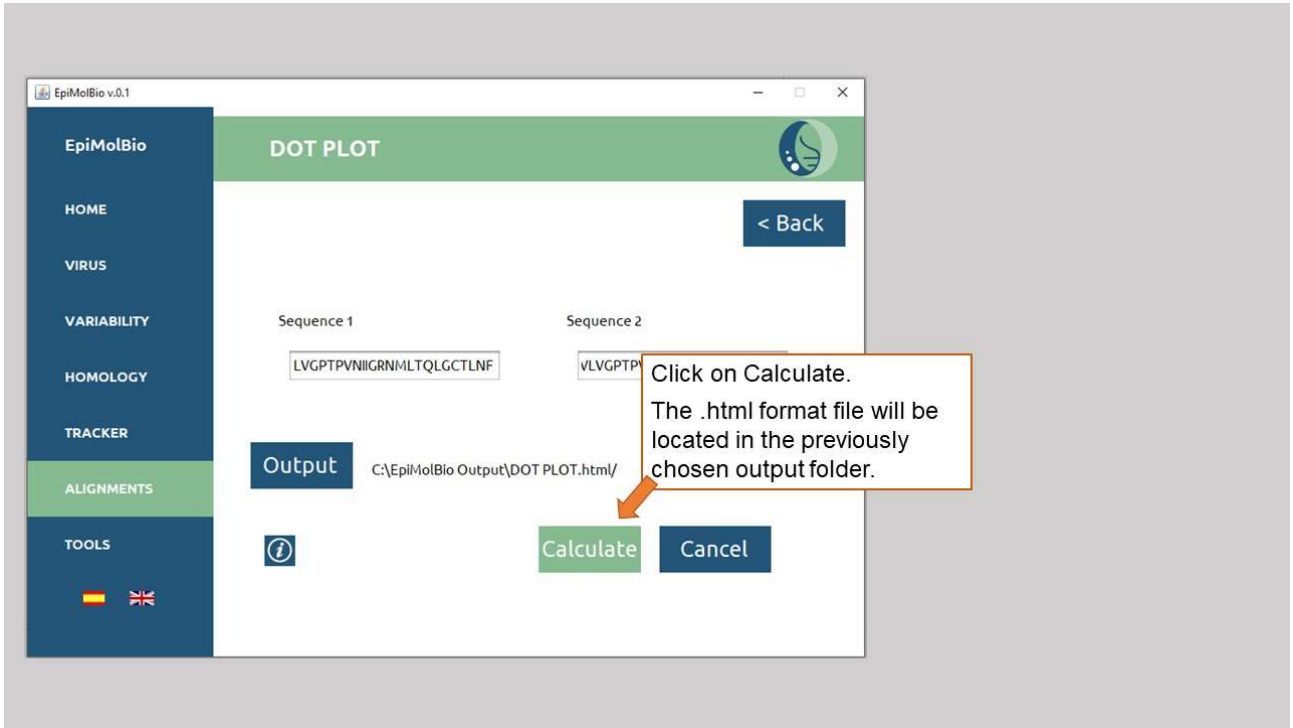
## 4)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
DOT PLOT
< Back
Sequence 1
Sequence 2
LVGPTPVNIIGRNMLTQLGCTLNF
VLVGPTPVNIIGRNMLTQIGCTLNF
Enter the second sequence to compare in the same format as the first, without line breaks.
Output
Output
Calculate
Cancel

## 5)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
DOT PLOT
< Back
Sequence 1
Sequence 2
LVGPTPVNIIGRNMLTQLGCTLNF
VLVGPTPVNIIGRNMLTQIGCTLNF
Output
Select the output folder and enter the file name as follows:
/NAME.html
Calculate
Cancel

## 6)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
DOT PLOT
< Back
Sequence 1
Sequence 2
LVGPTPVNIIGRNMLTQLGCTLNF
Output
C:\EpiMolBio Output\DOT PLOT.html/
Click on Calculate.
The .html format file will be located in the previously chosen output folder.
Calculate
Cancel

# V.3.DELETE INSERTIONS

This function **allows for the automatic removal of insertions from sequences in a file with respect to a reference that contains gaps**. When sequences have been previously aligned through multiple sequence alignment, or downloaded from a database that performs multiple sequence alignment with the reference (which must be present in the downloaded FASTA file), gaps are generated in the reference sequence corresponding to insertions. This function detects the gaps in the reference sequence and removes those positions from the input sequences, effectively eliminating the insertions. The output format is a .fasta file containing the resulting sequences without insertions.

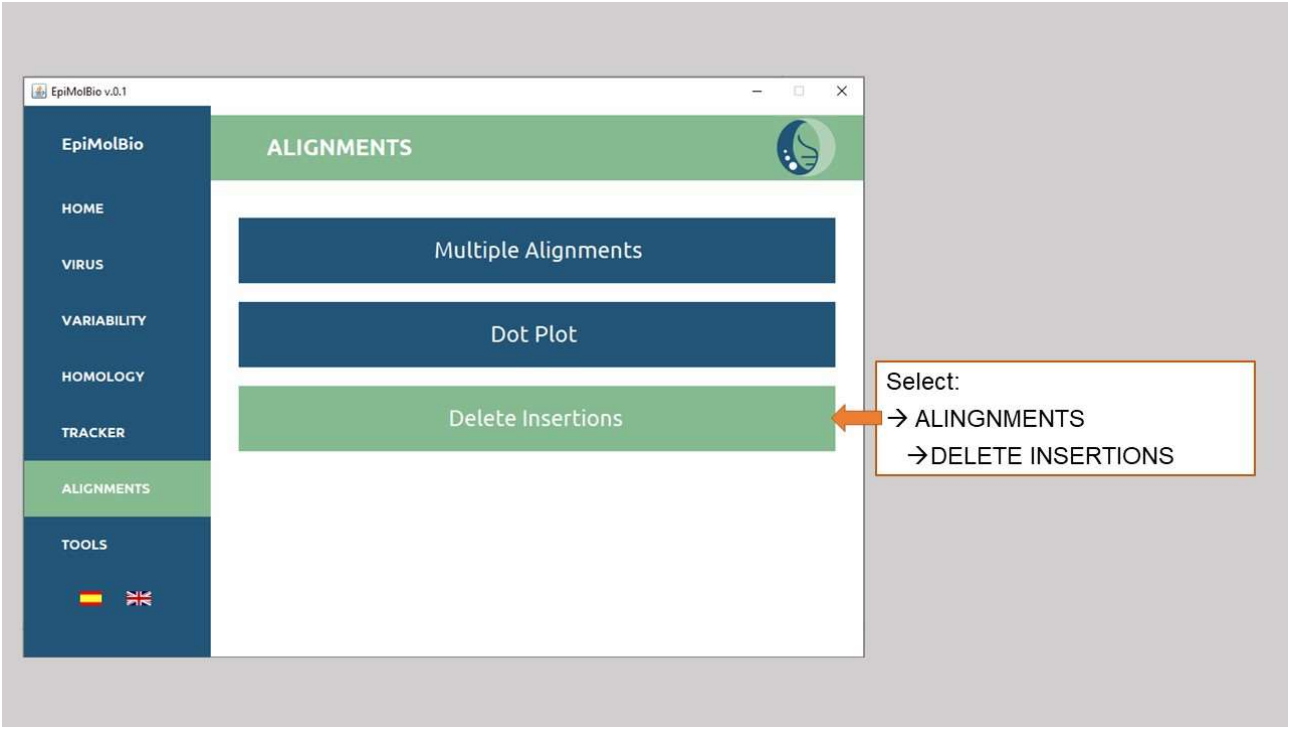
The **input** file should be the folder containing exclusively .fasta files with aligned sequences in nucleotides or amino acids.

In the '**Reference**' field, enter the reference sequence with gaps generated after multiple sequence alignment of the sequences, without line breaks or spaces in amino acids or nucleotides, depending on the input files.

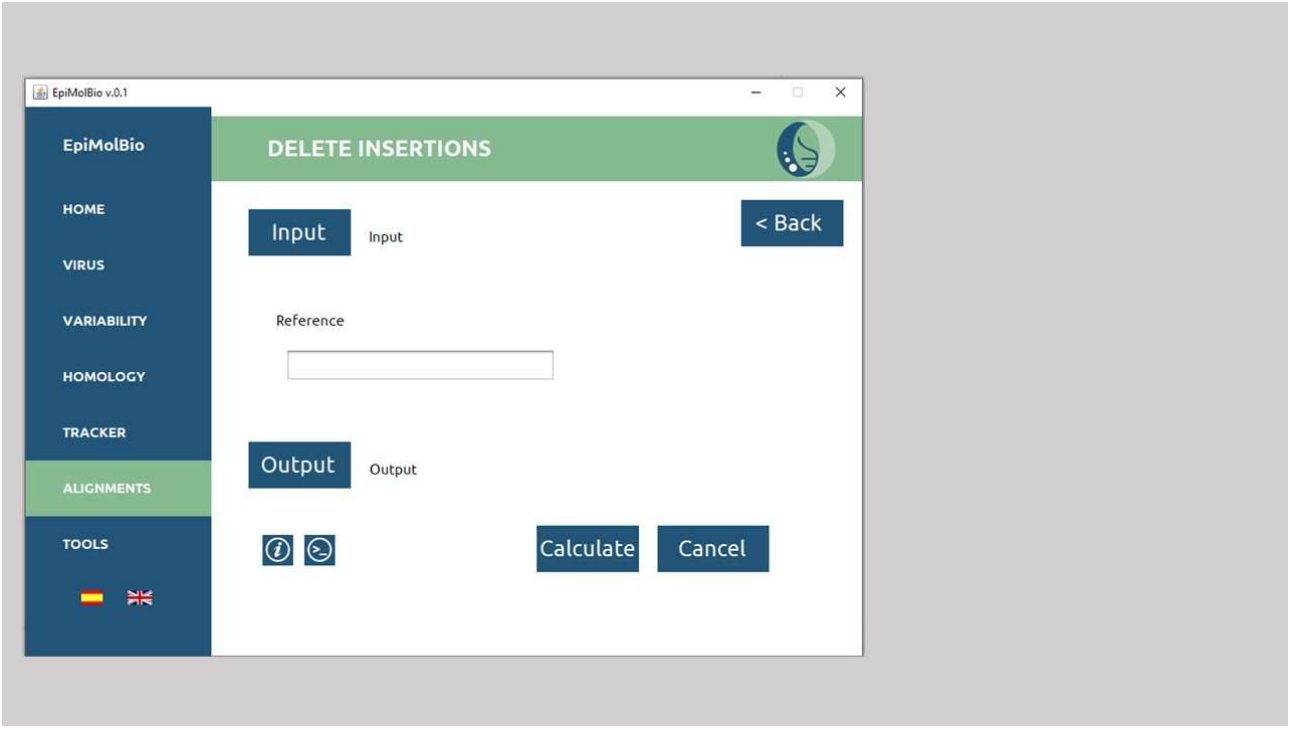
For the **output**, select the output folder where you want the .fasta files to appear. The files are automatically named as follows: Insertions_Deleted_InputFileName.fasta.

# Step-by-step:

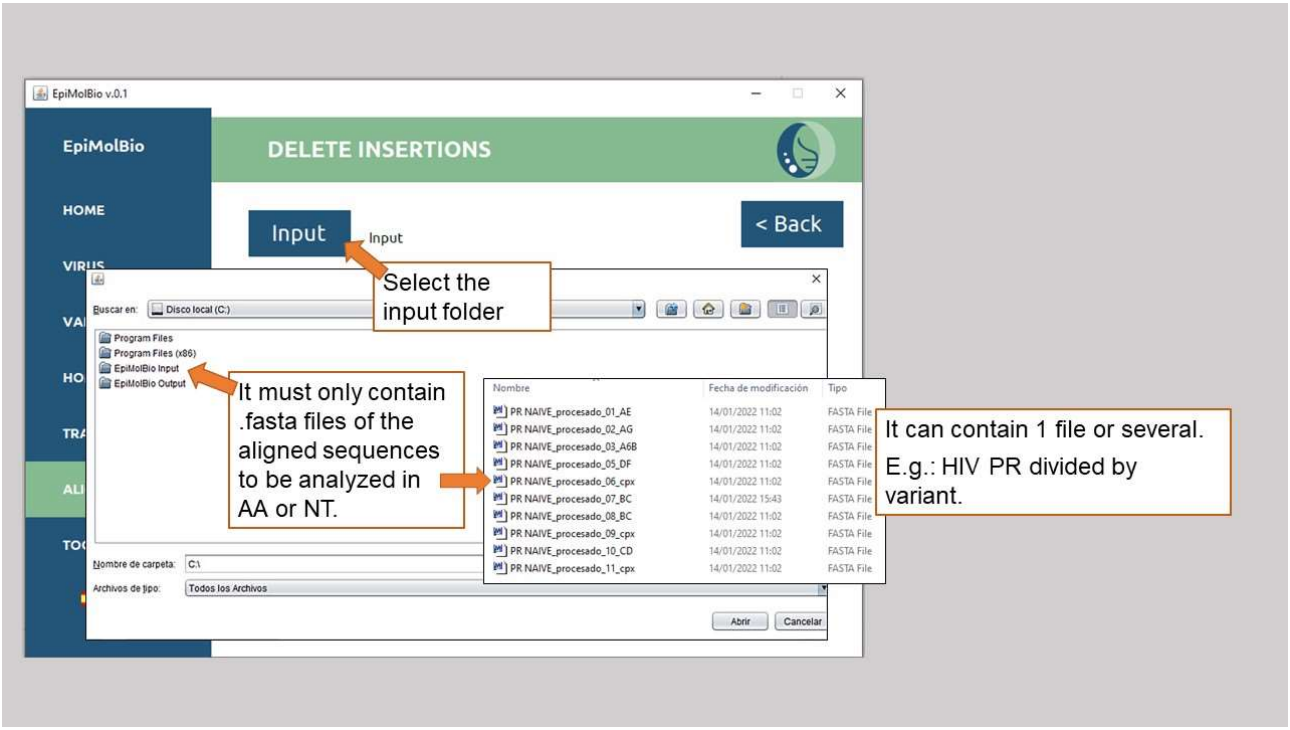
## 1)



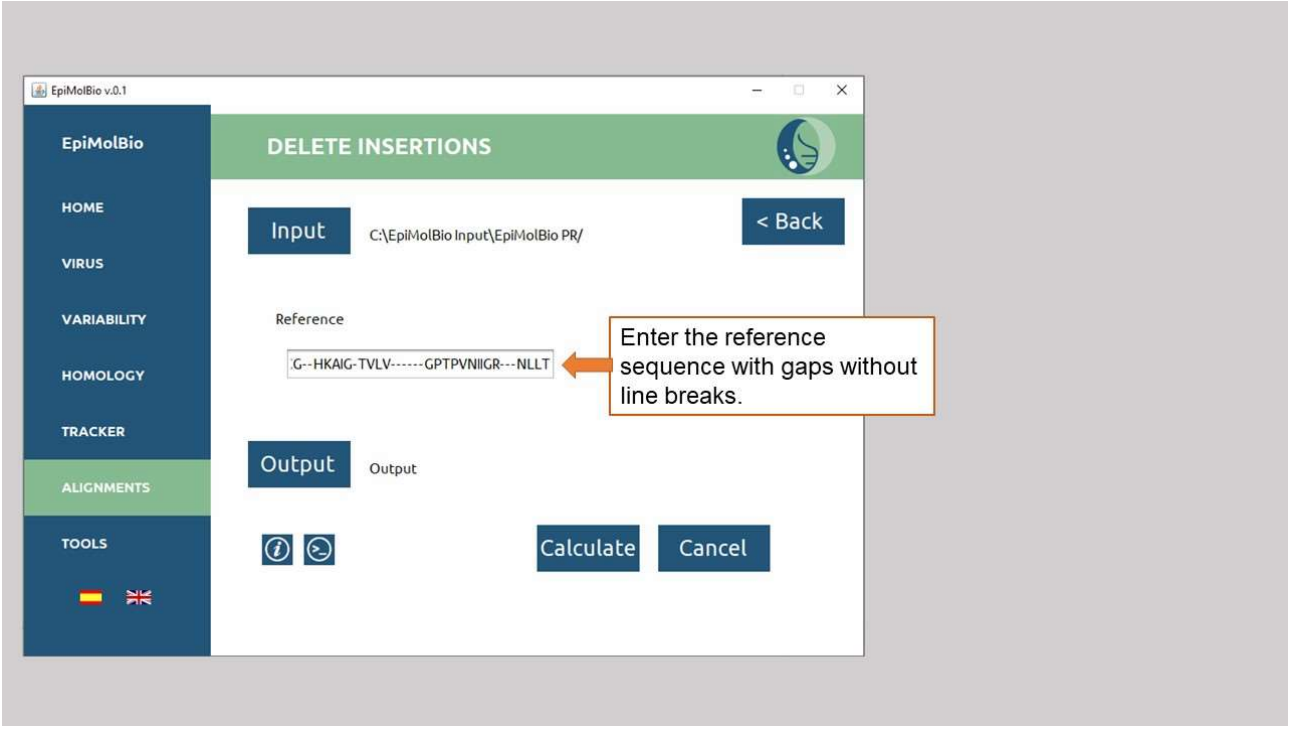

## 2)



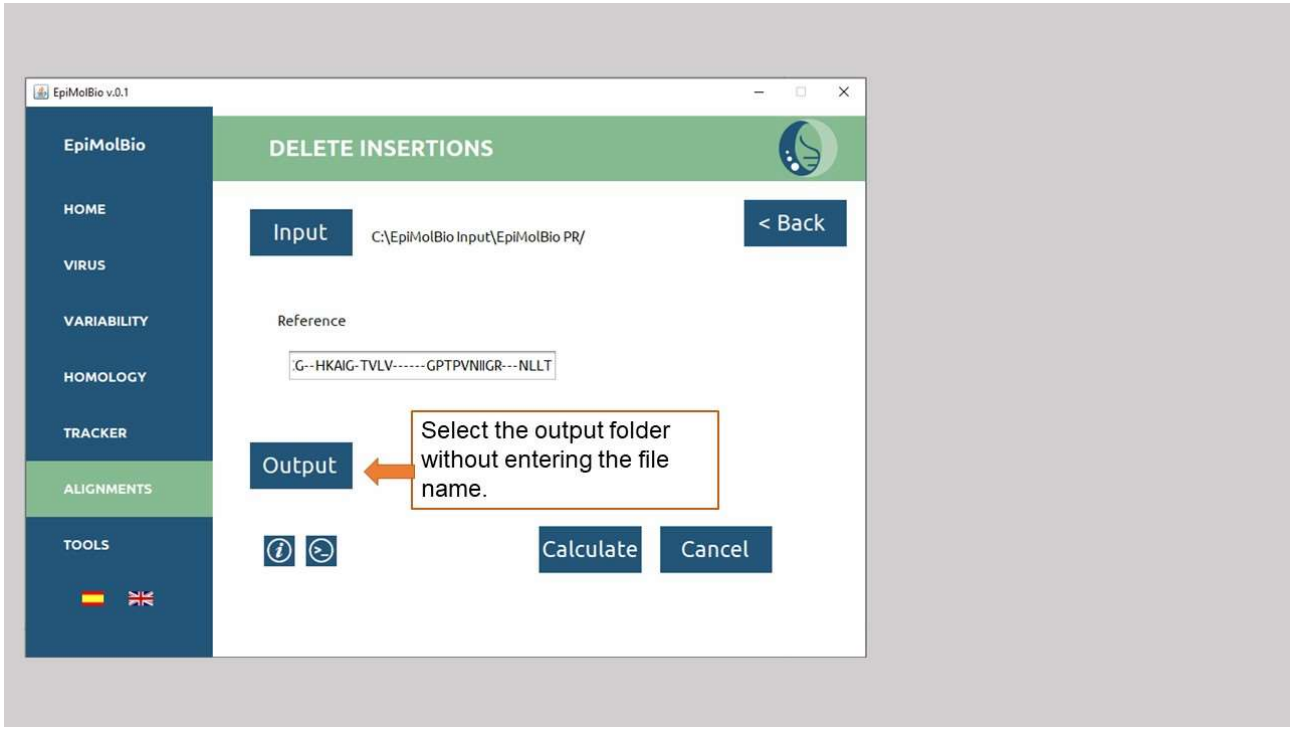
## 3)


EpiMolBio v.0.1
EpiMolBio
HOME
DELETE INSERTIONS
Input
Input
< Back
Select the input folder
Buscar en:
Disco local (C:)
Program Files
Program Files (x86)
EpiMolBio Input
EpiMolBio Output
It must only contain .fasta files of the aligned sequences to be analyzed in AA or NT.
Nombre
Fecha de modificación
Tipo
PR NAIVE_procesado_01_AE
PR NAIVE_procesado_02_AG
PR NAIVE_procesado_03_A6B
PR NAIVE_procesado_05_DF
PR NAIVE_procesado_06_cpx
PR NAIVE_procesado_07_BC
PR NAIVE_procesado_08_BC
PR NAIVE_procesado_09_cpx
PR NAIVE_procesado_10_CD
PR NAIVE_procesado_11_cpx
14/01/2022 11:02
14/01/2022 15:43
FASTA File
It can contain 1 file or several.
E.g.: HIV PR divided by variant.
Nombre de carpeta:
C:\
Archivos de tipo:
Todos los Archivos
Abrir
Cancelar

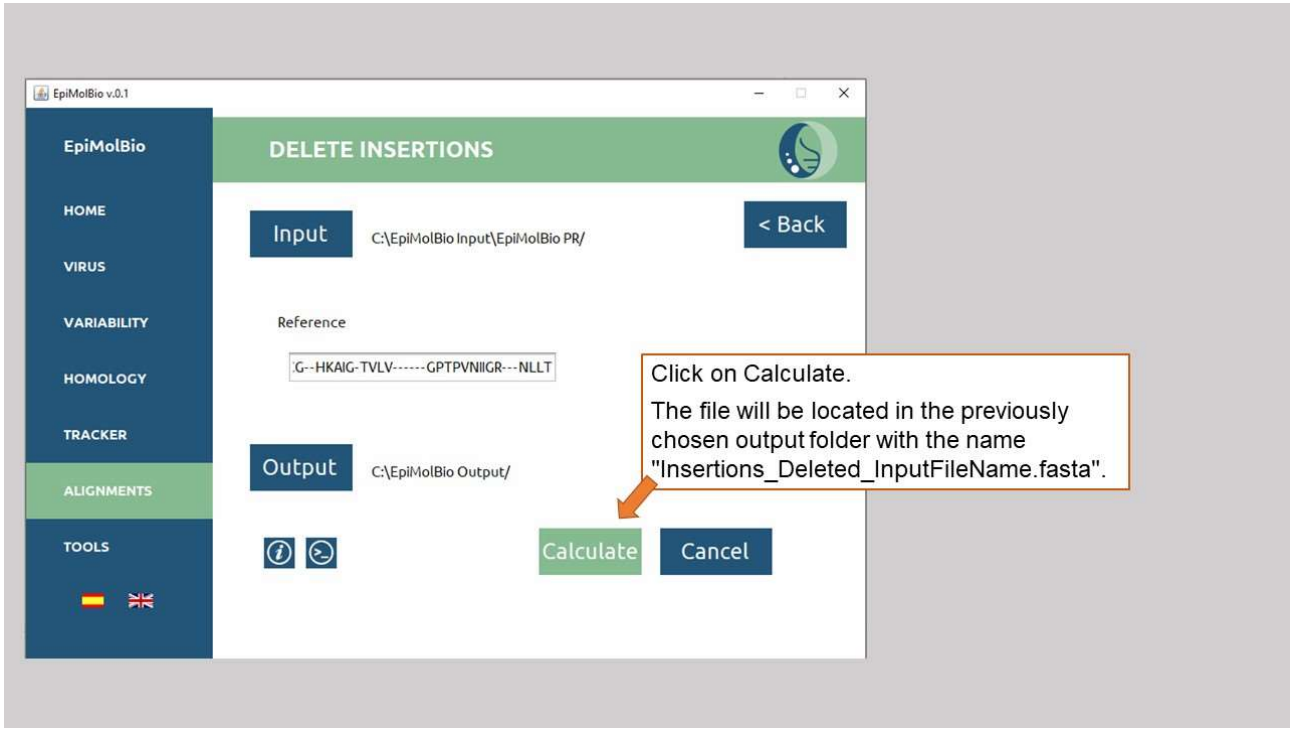
## 4)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
DELETE INSERTIONS
Input
C:\EpiMolBio Input\EpiMolBio PR/
< Back
Reference
:G--HKAIG-TVLV------GPTPVNIIGR---NLLT
Enter the reference sequence with gaps without line breaks.
Output
Output
Calculate
Cancel

## 5)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
DELETE INSERTIONS
Input
C:\EpiMolBio Input\EpiMolBio PR/
< Back
Reference
:G--HKAIG-TVLV------GPTPVNIIGR---NLLT
Output
Select the output folder without entering the file name.
Calculate
Cancel

## 6)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
DELETE INSERTIONS
Input
C:\EpiMolBio Input\EpiMolBio PR/
< Back
Reference
:G--HKAIG-TVLV------GPTPVNIIGR---NLLT
Output
C:\EpiMolBio Output/
Click on Calculate.
The file will be located in the previously chosen output folder with the name "Insertions_Deleted_InputFileName.fasta".
Calculate
Cancel

# VI. TOOLS

In this section, you will find various **tools for modifying files and sequences**, ranging from counting sequences, editing headers, or filtering sequences based on quality, to creating a semi-automated workflow using the ‘Function Programming’ tool.

## VI.1.FILE EDITING

### VI.1.A) MERGE FILES

This tool allows you to **merge multiple .fasta files into a single .fasta file**.

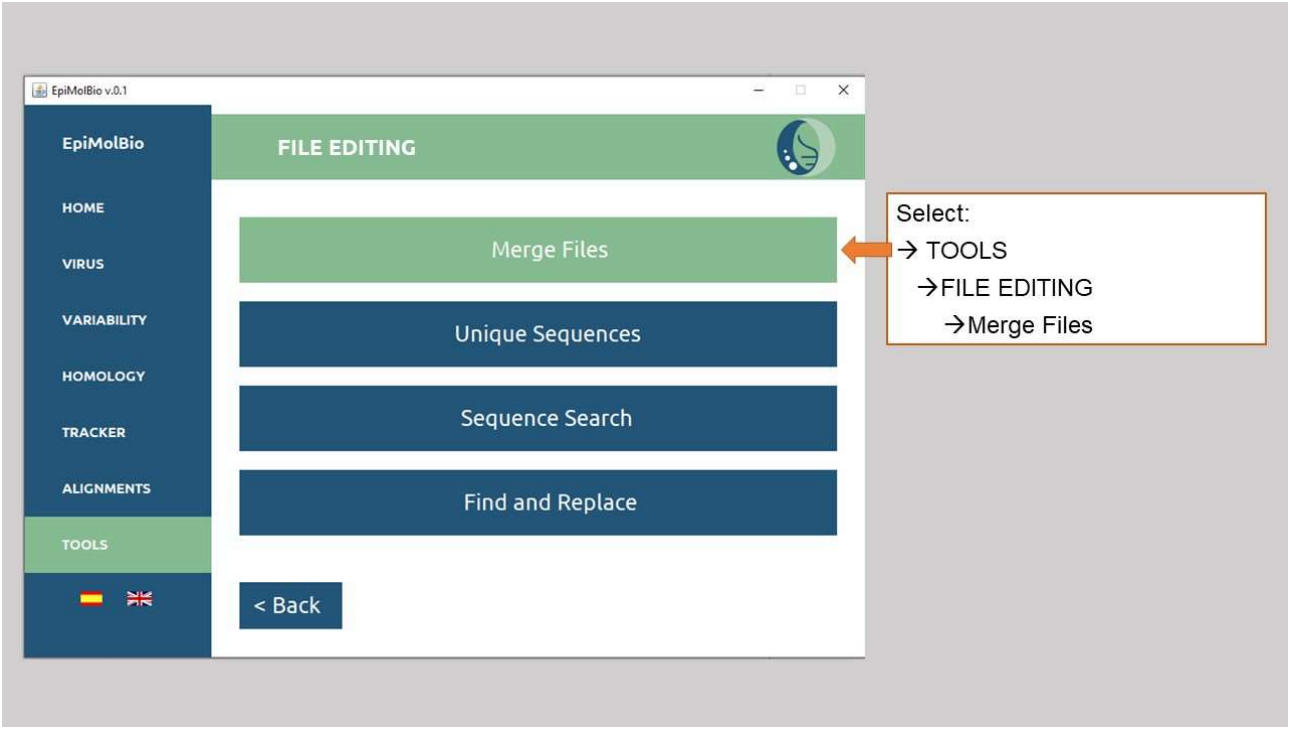
The **input** file should be the folder containing exclusively the .fasta files that you want to merge. These files can contain multiple sequences.

Mark the option ‘**Merge Files**.’

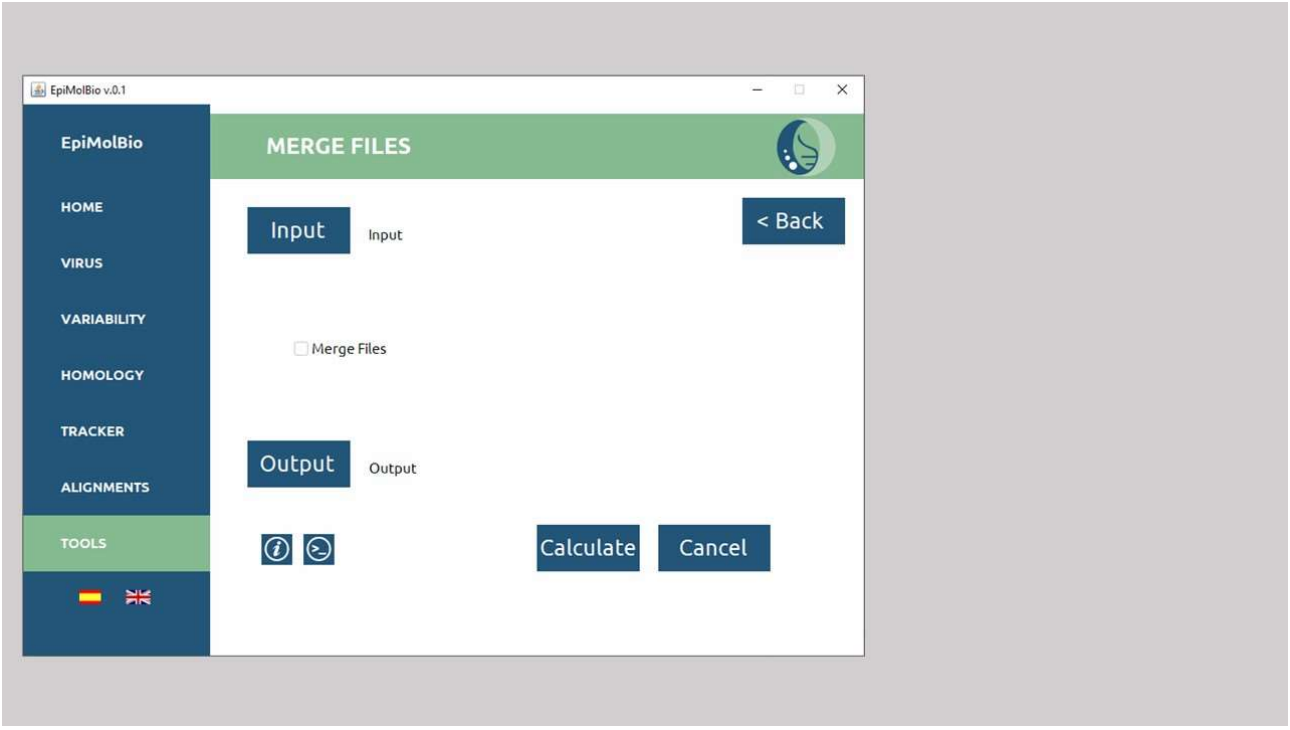
For the **output**, select the output folder where you want the .fasta file to appear and name it by adding ‘.fasta’ at the end. The output format is a .fasta file containing all the sequences from the input files.

# Step-by-step:

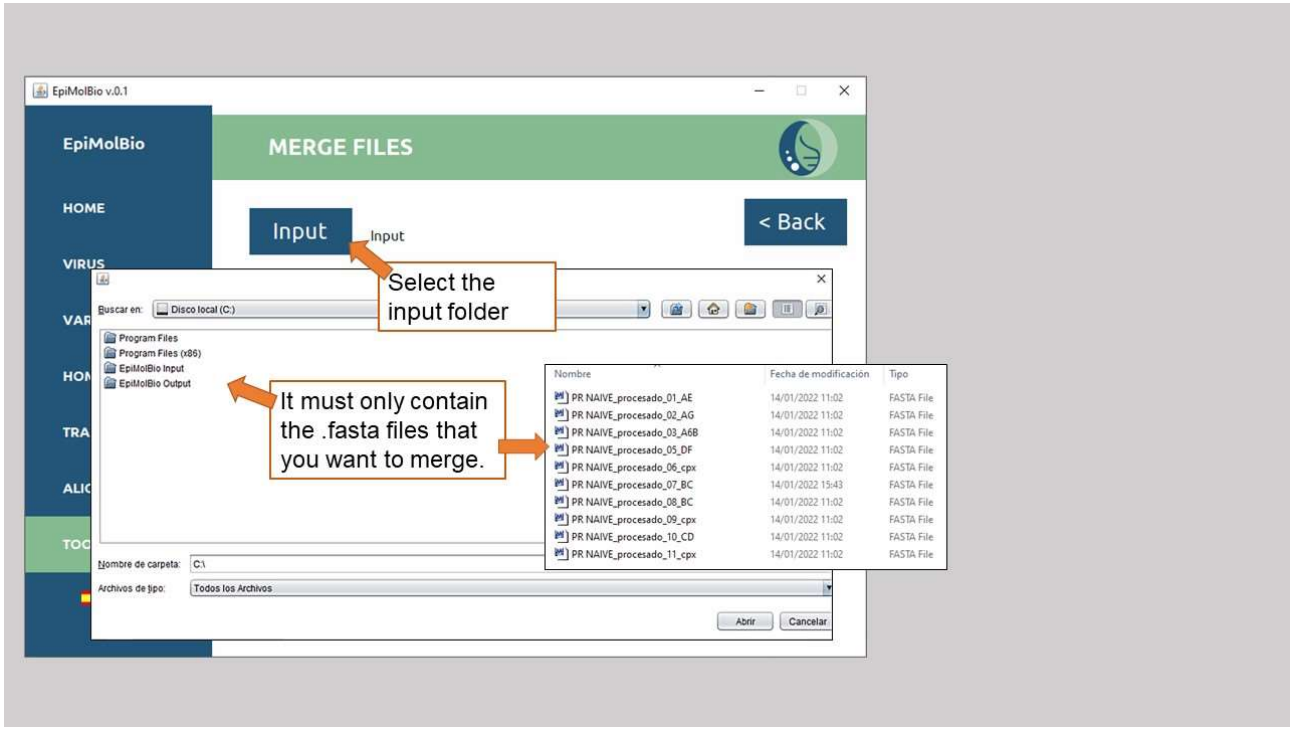
## 1)



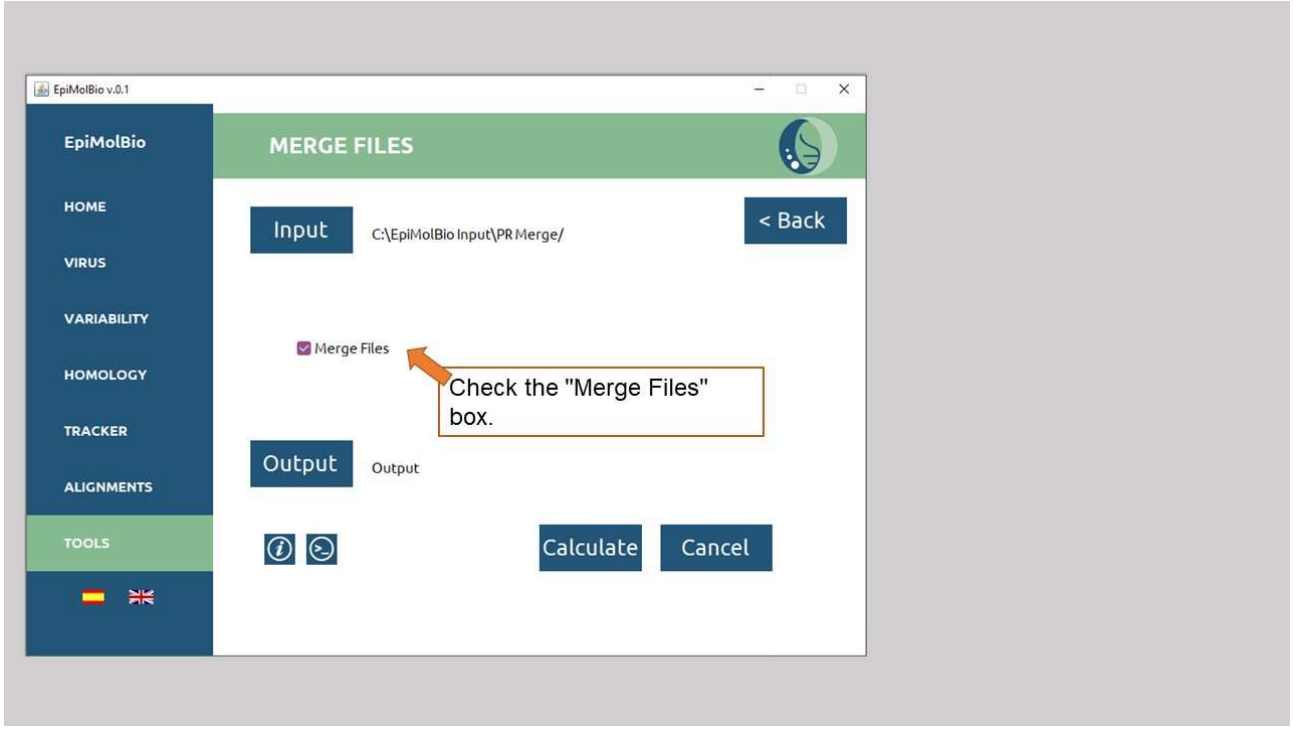

## 2)



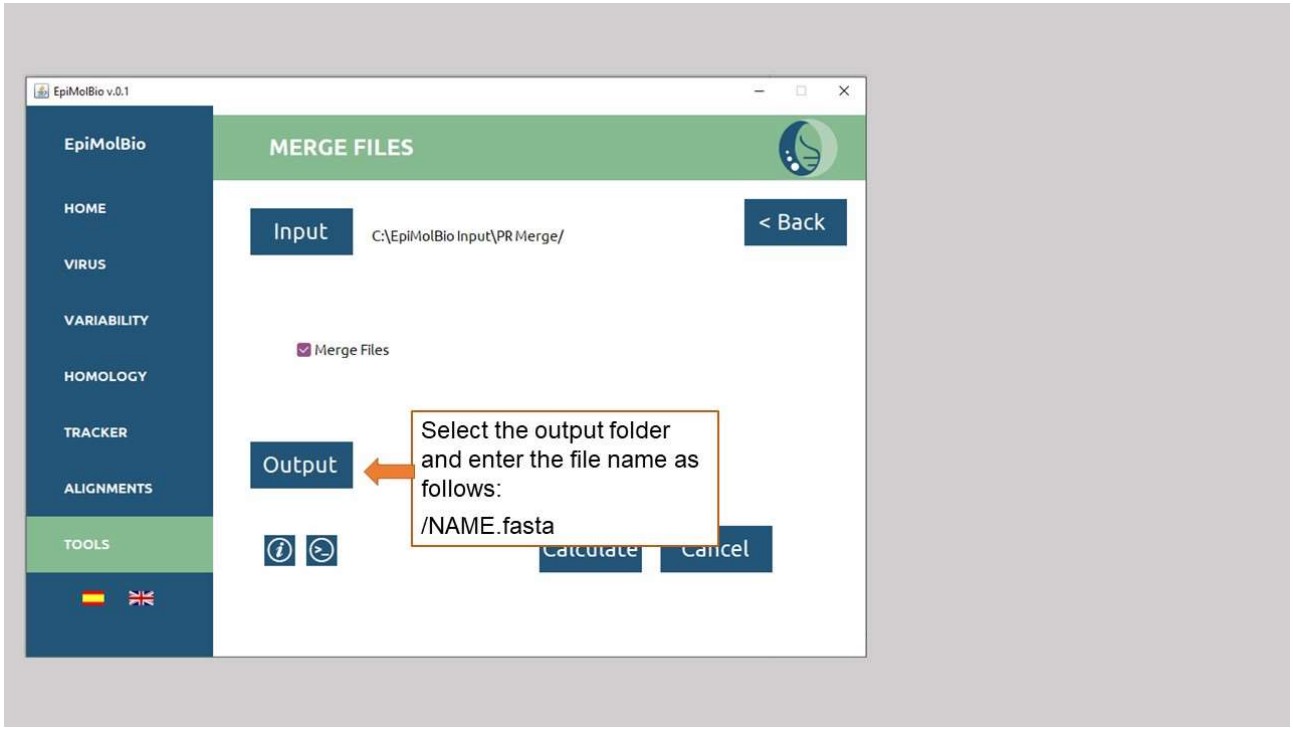
## 3)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
MERGE FILES
Input
Input
< Back
Select the input folder
Buscar en:
Disco local (C:)
Program Files
Program Files (x86)
EpiMolBio Input
EpiMolBio Output
It must only contain the .fasta files that you want to merge.
Nombre
Fecha de modificación
Tipo
PR NAIVE_procesado_01_AE
PR NAIVE_procesado_02_AG
PR NAIVE_procesado_03_A6B
PR NAIVE_procesado_05_DF
PR NAIVE_procesado_06_cpx
PR NAIVE_procesado_07_BC
PR NAIVE_procesado_08_BC
PR NAIVE_procesado_09_cpx
PR NAIVE_procesado_10_CD
PR NAIVE_procesado_11_cpx
14/01/2022 11:02
14/01/2022 15:43
FASTA File
Nombre de carpeta:
C:\
Archivos de tipo:
Todos los Archivos
Abrir
Cancelar

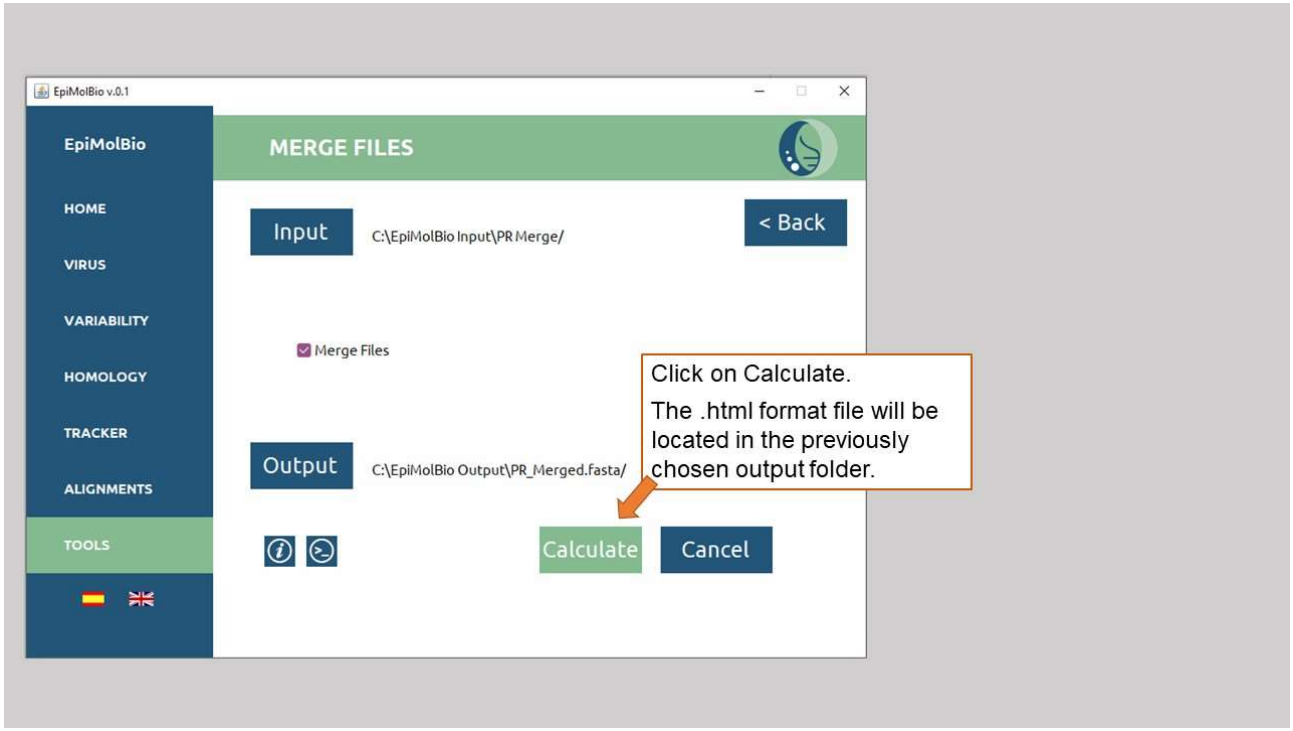
## 4)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
MERGE FILES
Input
C:\EpiMolBio Input\PR Merge/
< Back
Merge Files
Check the "Merge Files" box.
Output
Output
Calculate
Cancel

## 5)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
MERGE FILES
Input
C:\EpiMolBio Input\PR Merge/
< Back
Merge Files
Output
Select the output folder and enter the file name as follows:
/NAME.fasta
Calculate
Cancel

## 6)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
MERGE FILES
Input
C:\EpiMolBio Input\PR Merge/
< Back
Merge Files
Output
C:\EpiMolBio Output\PR_Merged.fasta/
Click on Calculate.
The .html format file will be located in the previously chosen output folder.
Calculate
Cancel

## VI.1.B) UNIQUE SEQUENCES

This tool allows you to **remove repeated sequences from one or multiple input .fasta files**. You can choose the minimum frequency with which you want to filter the sequences that will be kept in the output file. For example, if you have 15 sequences of the protease from HIV variant 06_cpx, with 6 of them being identical, applying a filtering at 0.0 will result in an output file containing 10 unique sequences, none repeated.

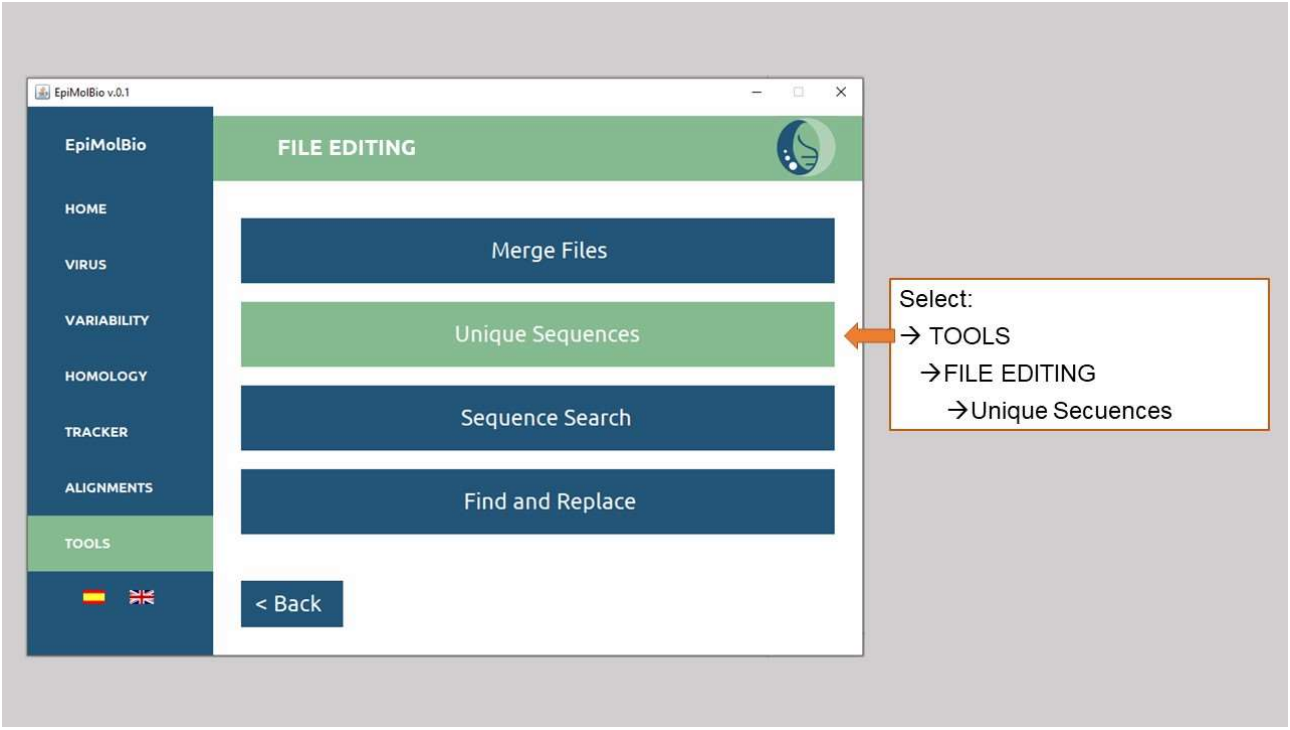
The **input** file should be the folder containing exclusively the .fasta files that you want to filter.

In the '**Minimum Frequency**' field, input the filtering frequency as a number with one decimal place. For example, input 90.0 to perform filtering at 90% and only retain sequences that are present at least 90% of the time, or input 0.0 to retain all unique sequences.

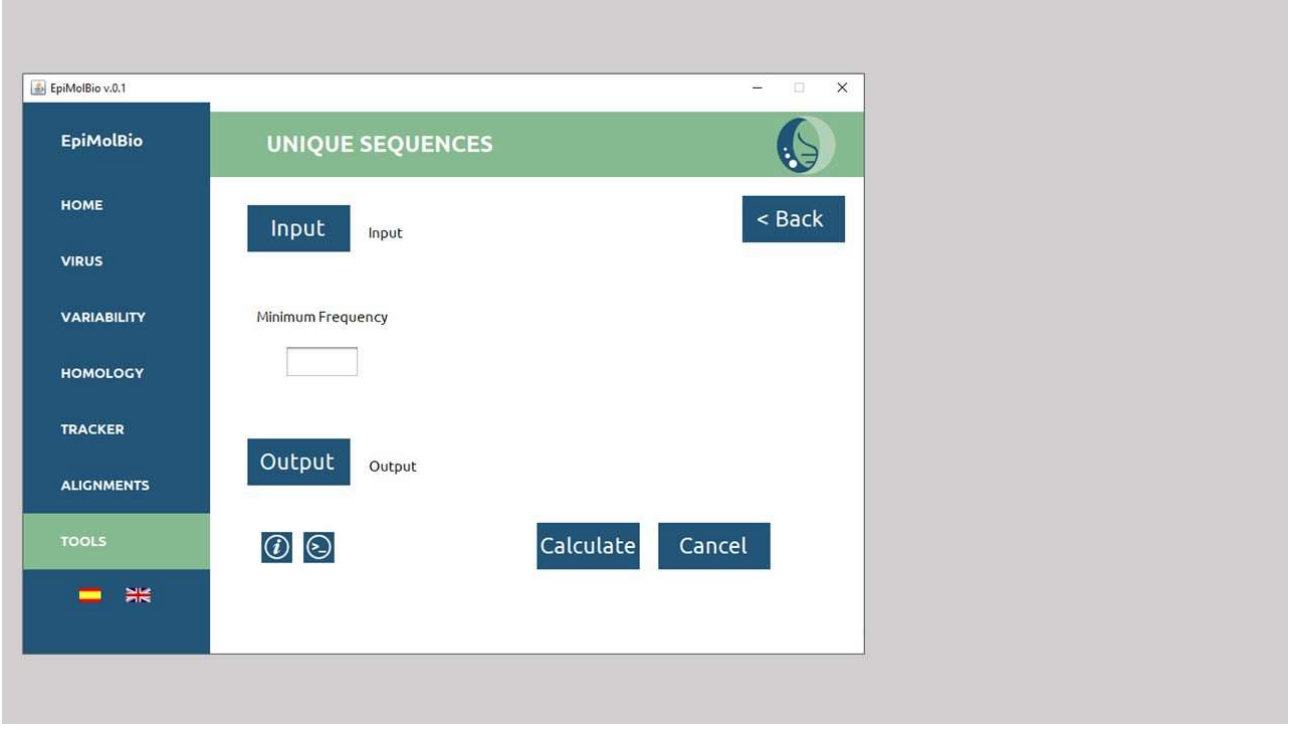
The **output** format is a .fasta file containing the filtered sequences. For the output, select the output folder where you want the .fasta file to appear, and it will be automatically named as follows: 'Unique_InputFileName.fasta'.

# Step-by-step:

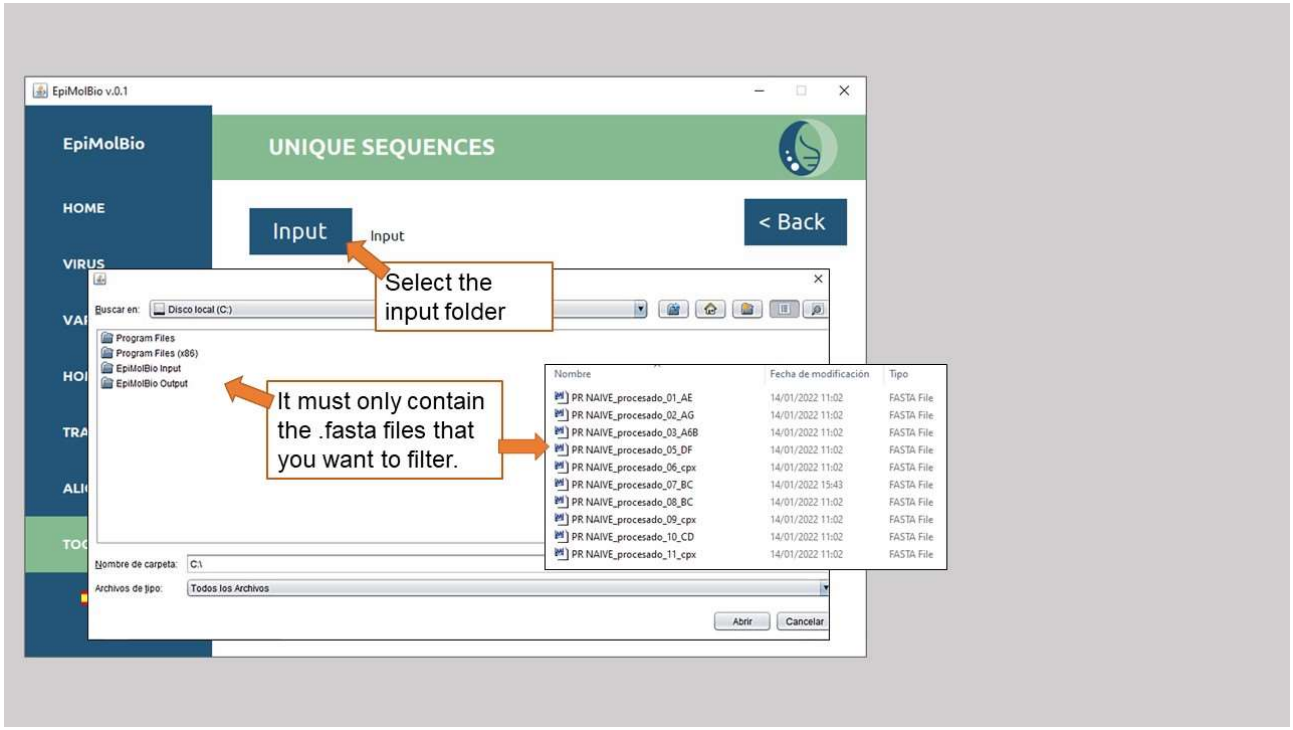
## 1)



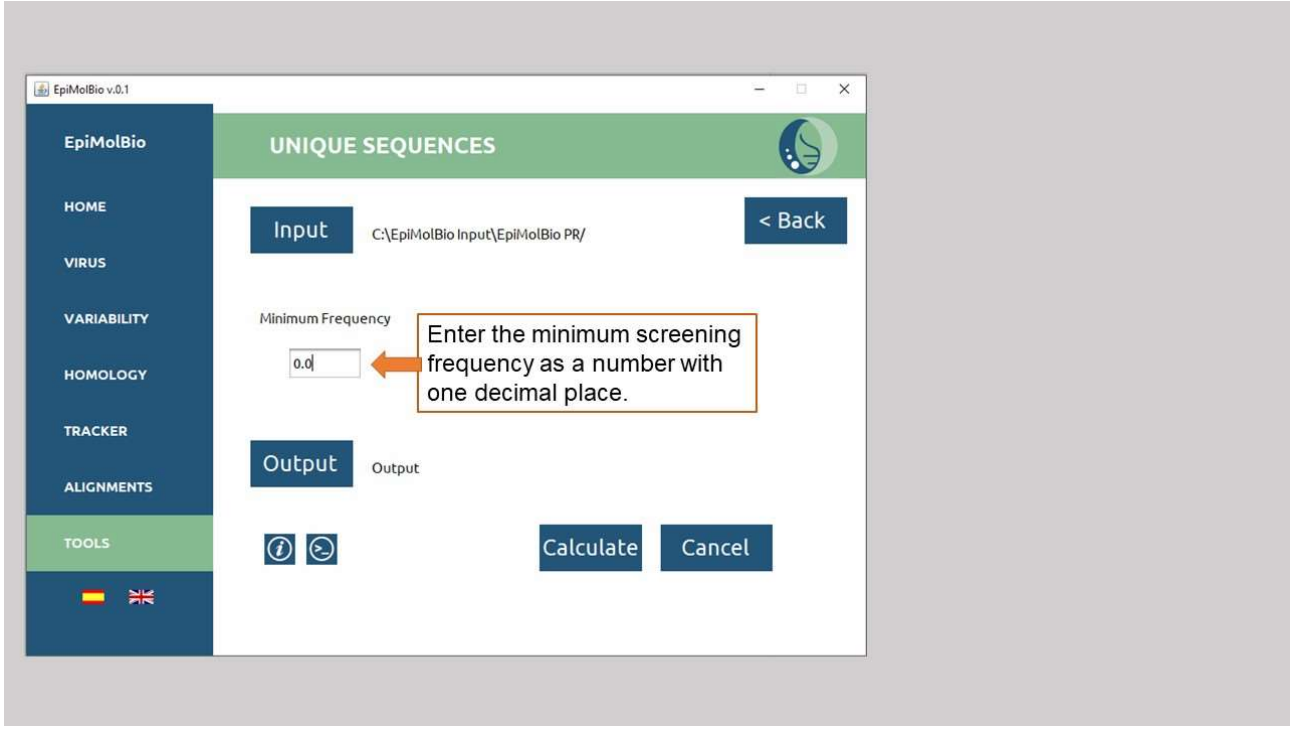

## 2)



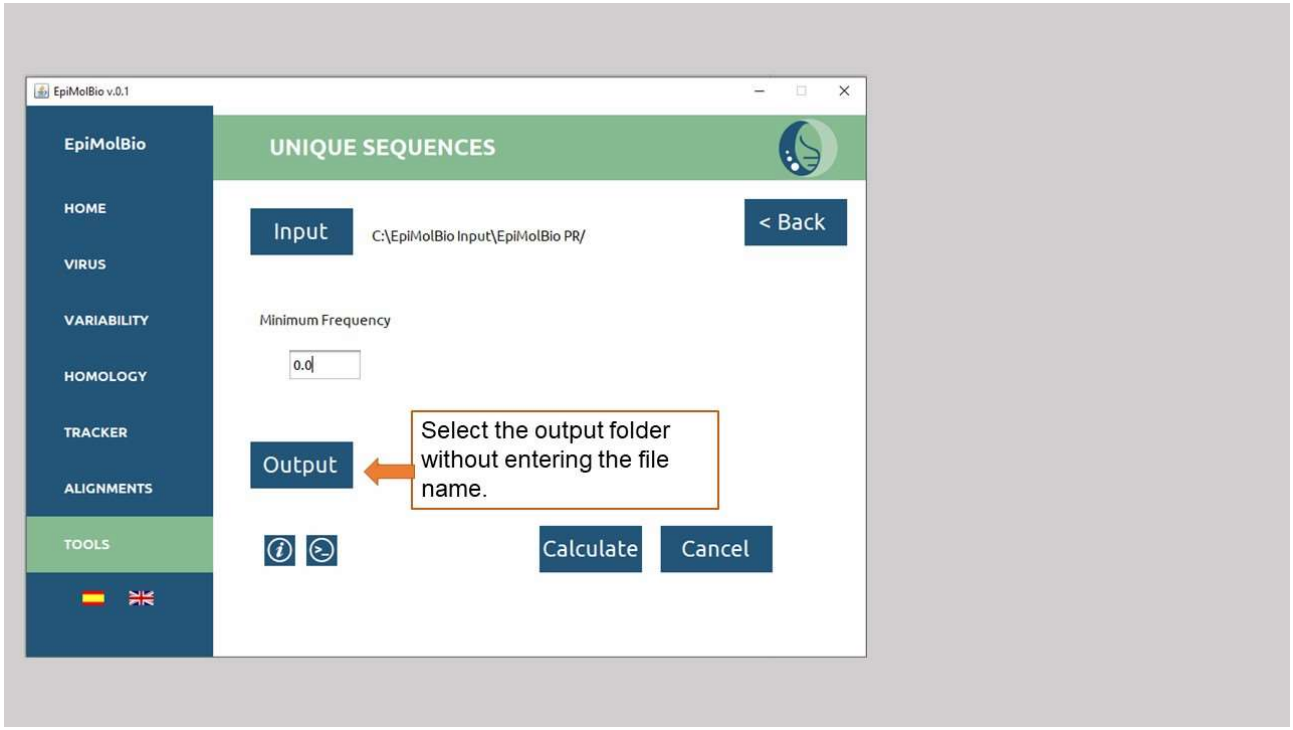
## 3)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
UNIQUE SEQUENCES
Input
Input
< Back
Select the input folder
Buscar en:
Disco local (C:)
Program Files
Program Files (x86)
EpiMolBio Input
EpiMolBio Output
It must only contain the .fasta files that you want to filter.
Nombre
Fecha de modificación
Tipo
PR NAIVE_procesado_01_AE
PR NAIVE_procesado_02_AG
PR NAIVE_procesado_03_A6B
PR NAIVE_procesado_05_DF
PR NAIVE_procesado_06_cpx
PR NAIVE_procesado_07_BC
PR NAIVE_procesado_08_BC
PR NAIVE_procesado_09_cpx
PR NAIVE_procesado_10_CD
PR NAIVE_procesado_11_cpx
14/01/2022 11:02
14/01/2022 15:43
FASTA File
Nombre de carpeta:
C:\
Archivos de tipo:
Todos los Archivos
Abrir
Cancelar

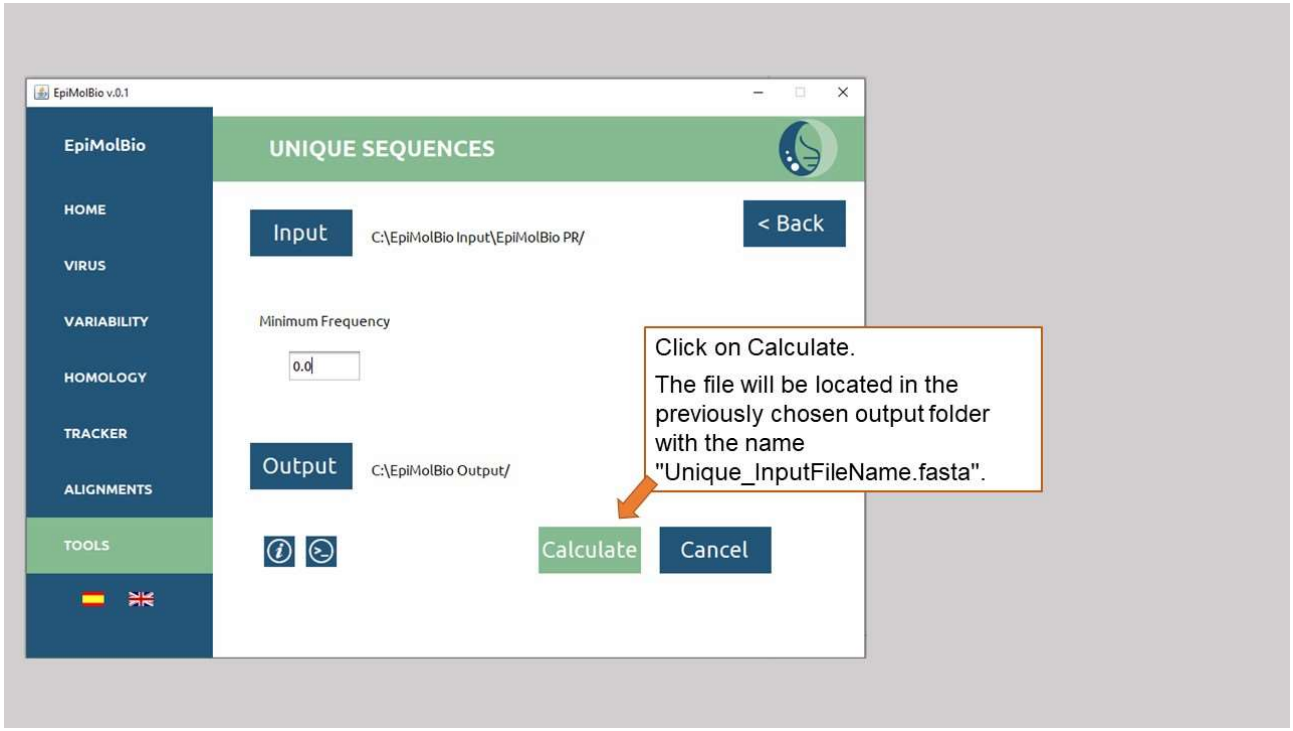
## 4)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
UNIQUE SEQUENCES
Input
C:\EpiMolBio Input\EpiMolBio PR/
< Back
Minimum Frequency
0.0
Enter the minimum screening frequency as a number with one decimal place.
Output
Output
Calculate
Cancel

## 5)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
UNIQUE SEQUENCES
Input
C:\EpiMolBio Input\EpiMolBio PR/
< Back
Minimum Frequency
0.0
Output
Select the output folder without entering the file name.
Calculate
Cancel

## 6)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
UNIQUE SEQUENCES
Input
C:\EpiMolBio Input\EpiMolBio PR/
< Back
Minimum Frequency
0.0
Output
C:\EpiMolBio Output/
Click on Calculate.
The file will be located in the previously chosen output folder with the name "Unique_InputFileName.fasta".
Calculate
Cancel

# VI.1.C) SEQUENCE SEARCH

This function is used to **filter sequences from .fasta files that contain one or several specific mutations**. The output can be either another .fasta file or a .csv table.

The **input** file should be the folder containing exclusively the .fasta files that you want to filter.

In the '**Format**' field, choose the type of output between FASTA (.fasta file) or CSV (.csv file).

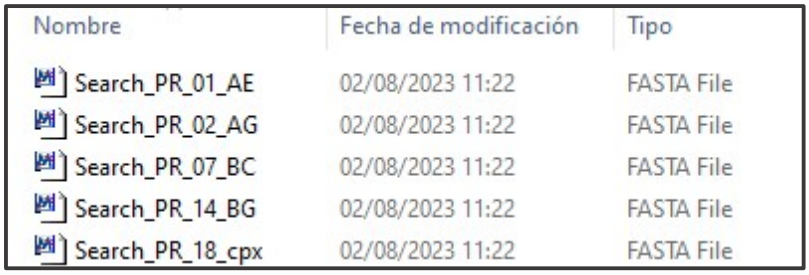
The .csv format consists of a table that can be opened in Excel. At the top, the analyzed input file is displayed. The first column shows the headers of the resulting sequences from the analysis, and the second column displays their sequences. Empty rows indicate that the chosen mutations were not found within any of the sequences in the input file indicated in the first column.

Example of .csv output format for Sequence Search Analysis:

| | A | B | C | D | E | F | G | H | I | J | K |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | PR_01_AE.fasta | | | | | | | | | | |
| 2 | >01_AE.VN.2 | PQITLWQRPLVTVKIGGQLKEALLDTGADDTVLEDINLPGKWKPKVIGGIGGFIKVRQYDQISIEICGKKAIGTVLVGPTPVNIIGRNMLTQIGCTLNF | | | | | | | | | |
| 3 | >01_AE.TH.2 | PQITLWQRPLVTVKIEGQLKEALLDTGADDTVLEDINLPGKWKPKVIGGIGGFIKVRQYDQILIEICGKRAIGTVLVGPTPVNIIGRNMLTQIGCTLNF | | | | | | | | | |
| 4 | >01_AE.CN.2 | PQITLWQRPLVTVKIGGQLKEALLDTGADDTVLEDINLPGKWKPKVIGGIGGFIKVRQYDQILIEICGKRAIGTVLVGPTPVNIIGRNMLTKIGCTLNF | | | | | | | | | |
| 5 | >01_AE.CN.2 | PQITLWQRPLVTIKIEGQLKEALLDTGADDTVLEDINLPGKLKPKVIGGIGGFIKVRQYDQILIELCGKKAIGTVLVGPTPVNIIGRNMLTQLGCTLNF | | | | | | | | | |
| 6 | >01_AE.CN.2 | PQITLWQRPLVTIKIGGQLKEALLDTGADDTVLEDINLPGKWKPKVIGGIGGFIKVRQYDQIPIEICGKKAIGTVLVGPTPVNIIGRNMLTQLGCTLNF | | | | | | | | | |
| 7 | >01_AE.CN.2 | PQITLWQRPCVTIKIGAELKEALLDTGADDPVLEDINLPGKWKPKVIGGIGGFIRVRHYDRVVIGICGRKAVRTVLVRPTPVNIIKRNMFSHLGFALNF | | | | | | | | | |
| 8 | >01_AE.CN.2 | PQITLWQRPLVTVKIGDQLREALLDTGADDTVLEEINLPGKWKPKVIGGIGGFIKVRQYDQISIEICGKKAIGTVLVGPTPVNIIGRNMLTQLGCTLNF | | | | | | | | | |
| 9 | >01_AE.CN.2 | PQITLWQRPLVTVKIGDQLREALLDTGADDTVLEEINLPGKWKPKVIGGIGGFIKVRQYDQISIEICGKKAIGTVLVGPTPVNIIGRNMLTQLGCTLNF | | | | | | | | | |

The .fasta format generates a .fasta file for each input file provided. Files in which the chosen mutations are not detected will not appear.

Example of .fasta output format for Sequence Search Analysis:

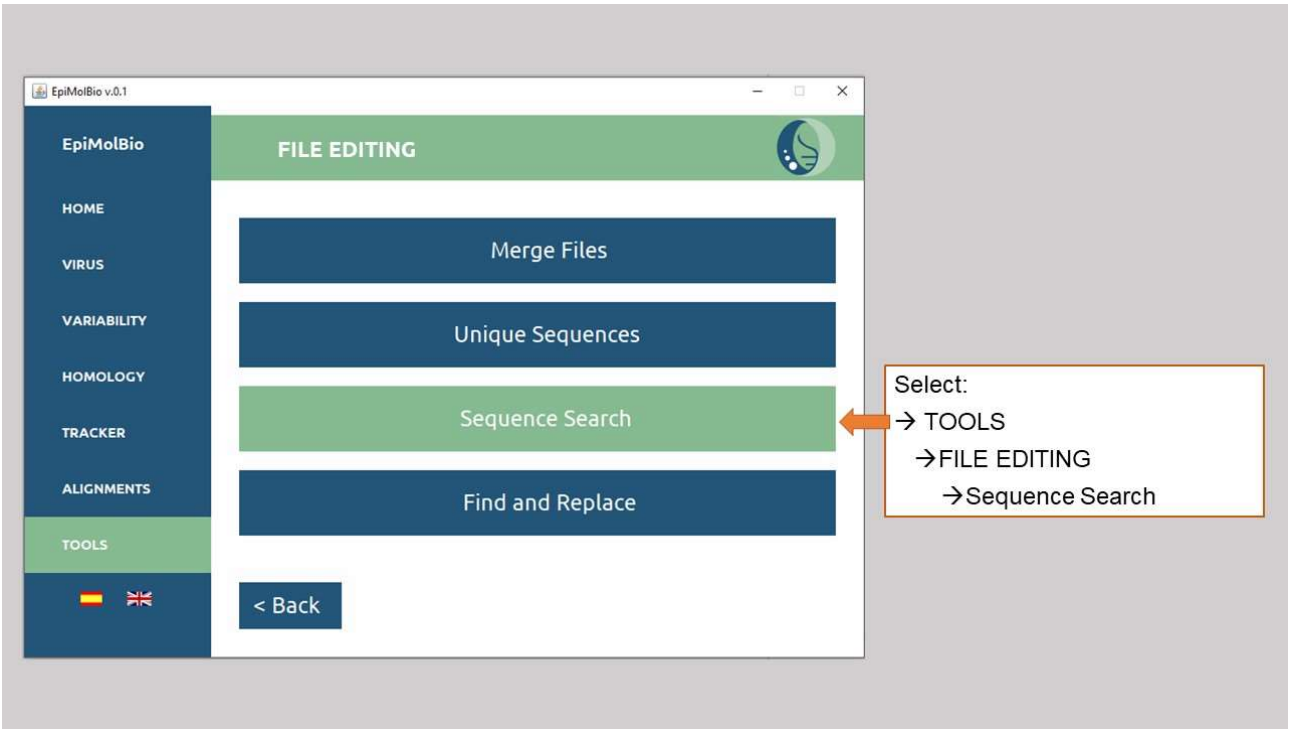



In the '**Mutations**' field, input one or multiple mutations separated by commas and without spaces (e.g., M50I,L74I).

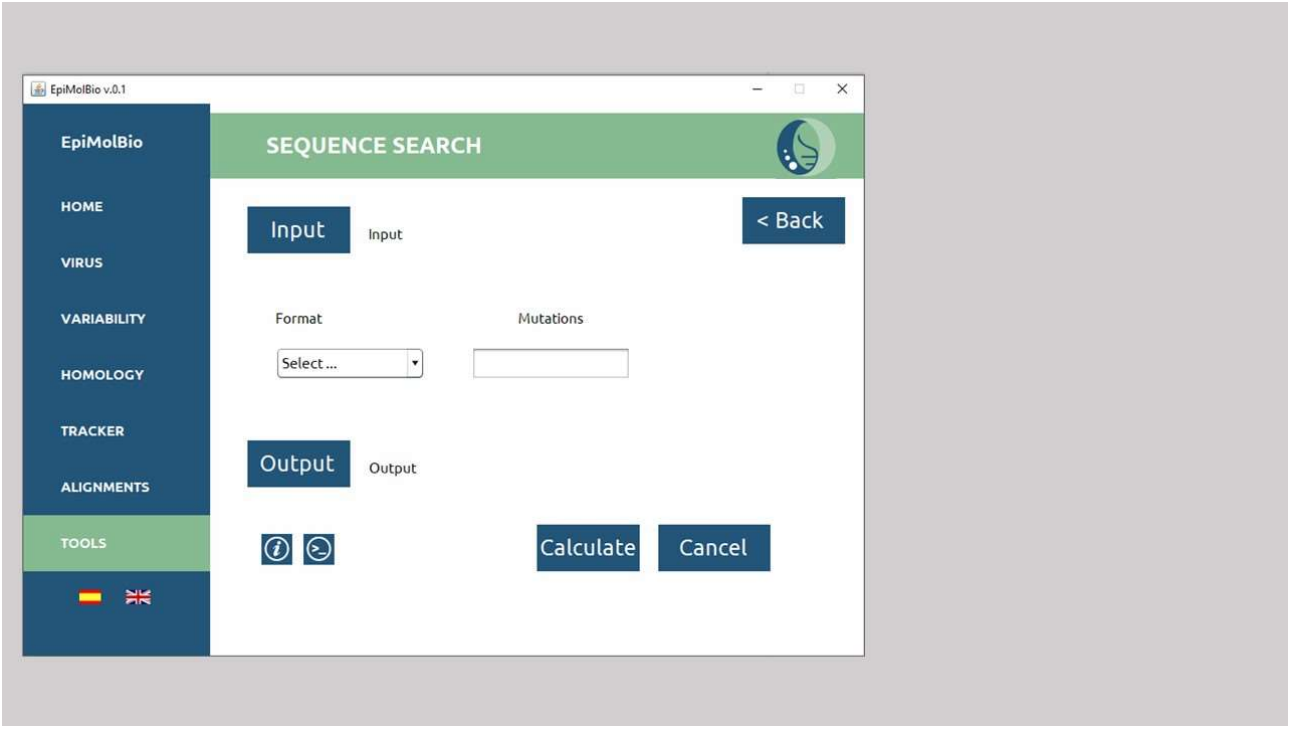
For the **output**, select the output folder where you want the results file to appear. If you choose the .csv output format, you need to name the file by adding '.csv' at the end. If you choose the .fasta output format, you don't need to name the file, as it will be automatically named as follows: Search_InputFileName.fasta.

# Step-by-step:

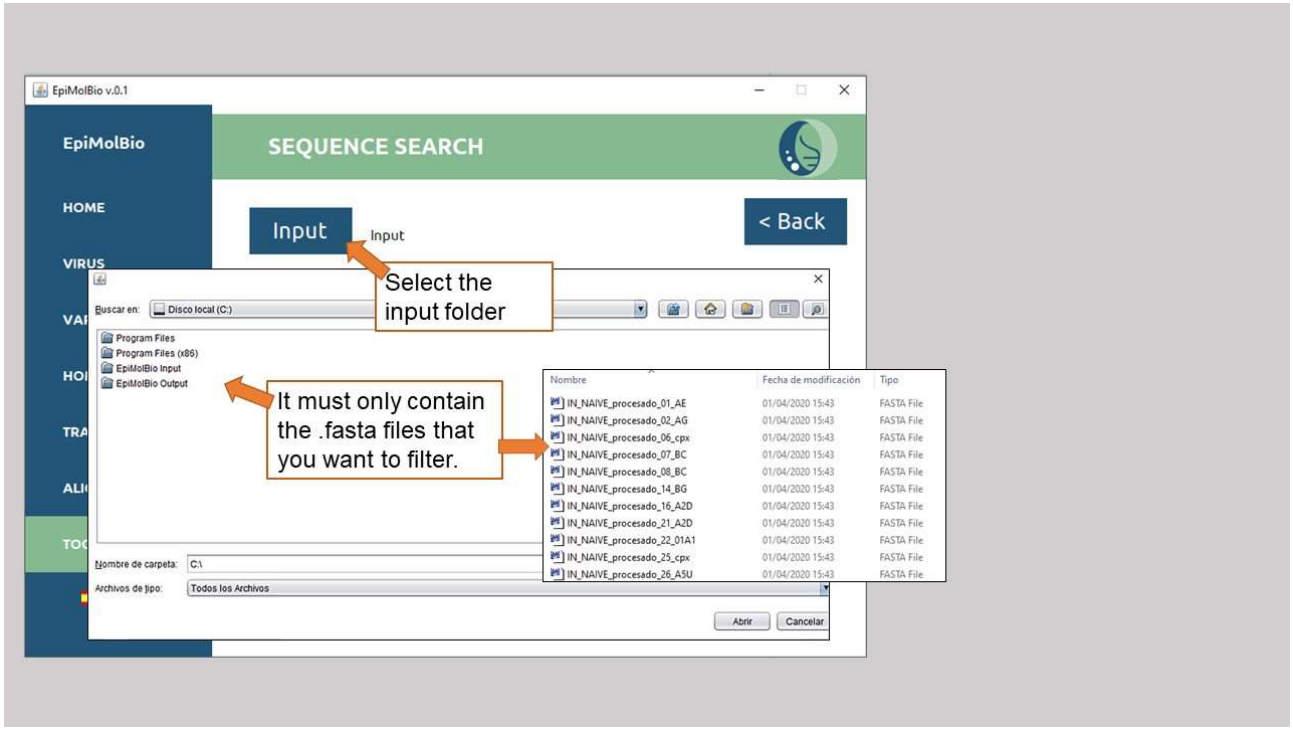
## 1)



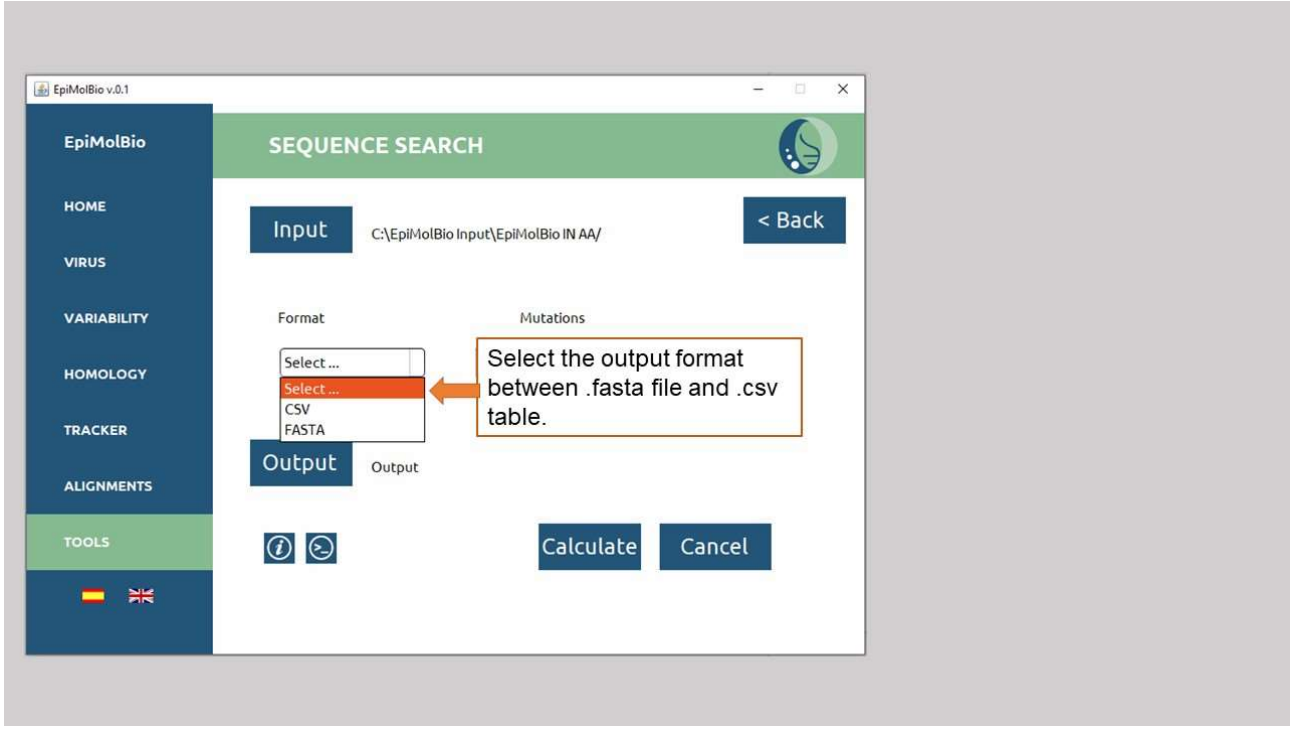

## 2)



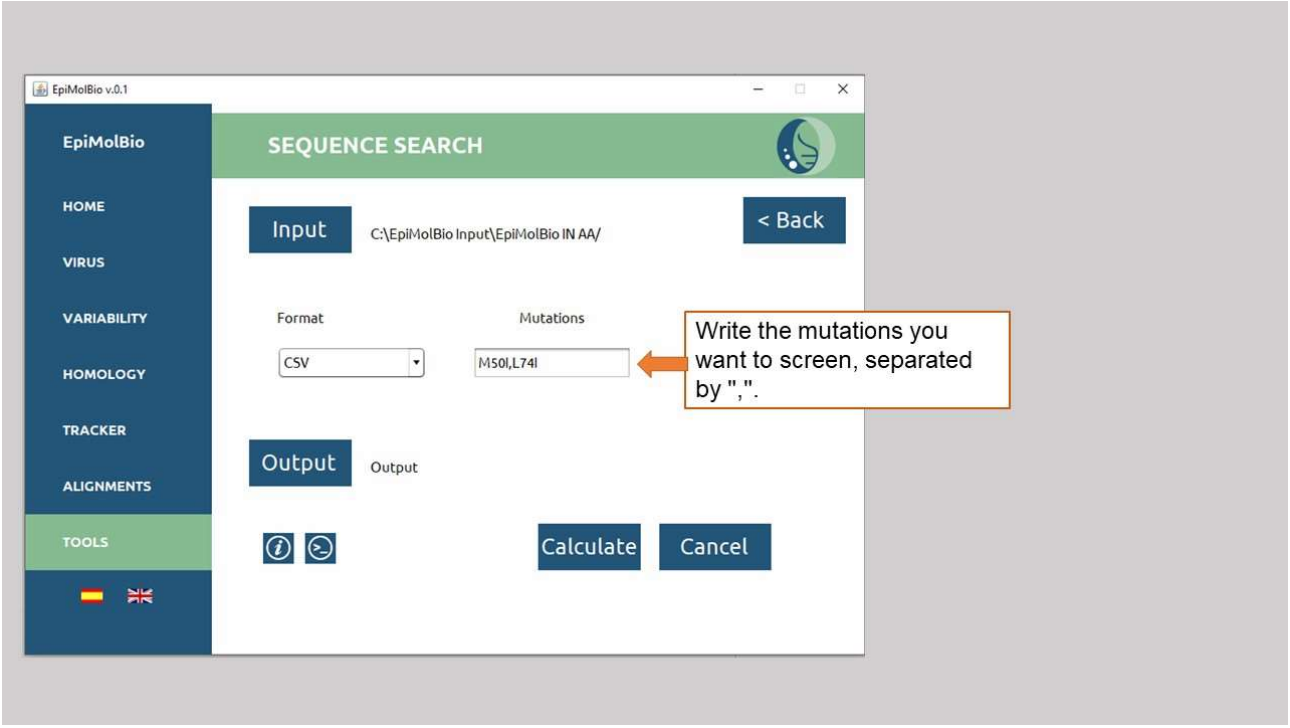
## 3)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
SEQUENCE SEARCH
Input
Input
< Back
Select the input folder
It must only contain the .fasta files that you want to filter.
Buscar en:
Disco local (C:)
Program Files
Program Files (x86)
EpiMolBio Input
EpiMolBio Output
Nombre de carpeta:
Archivos de tipo:
Todos los Archivos
Abrir
Cancelar

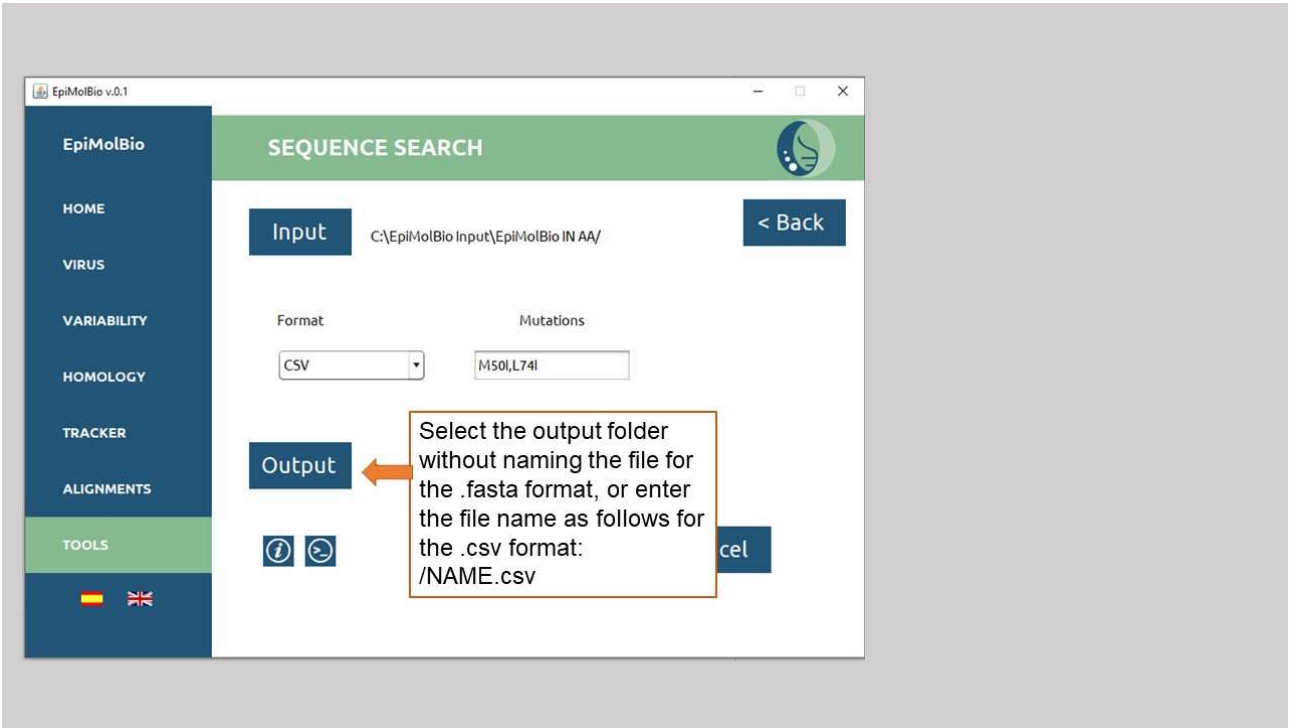
## 4)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
SEQUENCE SEARCH
Input
C:\EpiMolBio Input\EpiMolBio IN AA/
< Back
Format
Mutations
Select ...
Select ...
CSV
FASTA
Select the output format between .fasta file and .csv table.
Output
Output
Calculate
Cancel

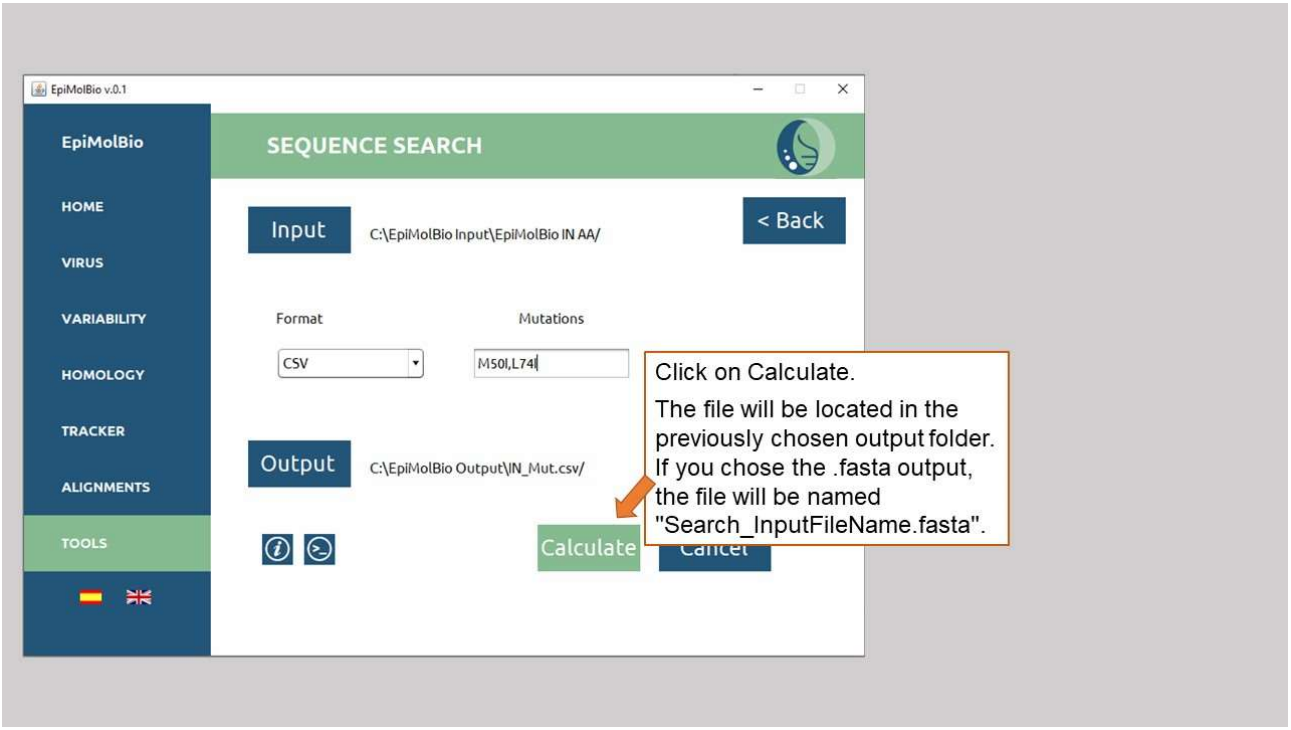
## 5)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
SEQUENCE SEARCH
Input
C:\EpiMolBio Input\EpiMolBio IN AA/
< Back
Format
Mutations
CSV
M50I,L74I
Write the mutations you want to screen, separated by ",".
Output
Output
Calculate
Cancel

## 6)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
SEQUENCE SEARCH
Input
C:\EpiMolBio Input\EpiMolBio IN AA/
< Back
Format
Mutations
CSV
M50I,L74I
Output
Select the output folder without naming the file for the .fasta format, or enter the file name as follows for the .csv format: /NAME.csv
cel

# 7)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
SEQUENCE SEARCH
Input
C:\EpiMolBio Input\EpiMolBio IN AA/
< Back
Format
Mutations
CSV
M50I,L74I
Output
C:\EpiMolBio Output\IN_Mut.csv/
Calculate
Cancel
Click on Calculate.
The file will be located in the previously chosen output folder. If you chose the .fasta output, the file will be named "Search_InputFileName.fasta".

# VI.1.D) FIND AND REPLACE

This tool allows you to **replace a series of characters with others, both in the header and in the genetic sequence, in one or multiple '.fasta' files**. For example, you can change '-' to '/' in the header of the .fasta files to use the header filtering function later, or replace 'N' with '?' to analyze nucleotide sequences using a function that only accepts amino acid sequences as input in EpiMolBio as in 'Variability, Polymorphisms, Individual'.

The **input** file should be the folder containing exclusively the '.fasta' files that you want to modify.

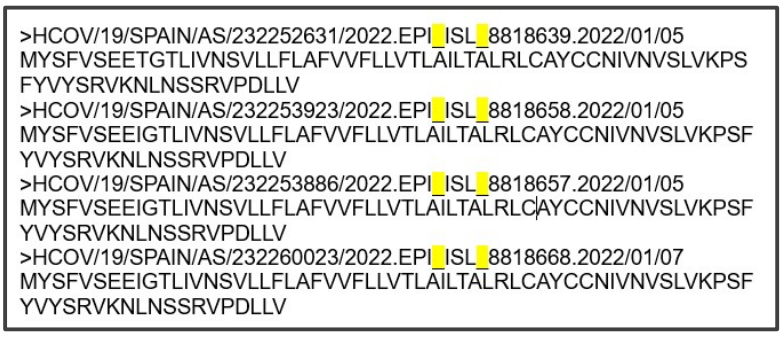
In the '**Search**' field, input the characters you want to replace (e.g., _).

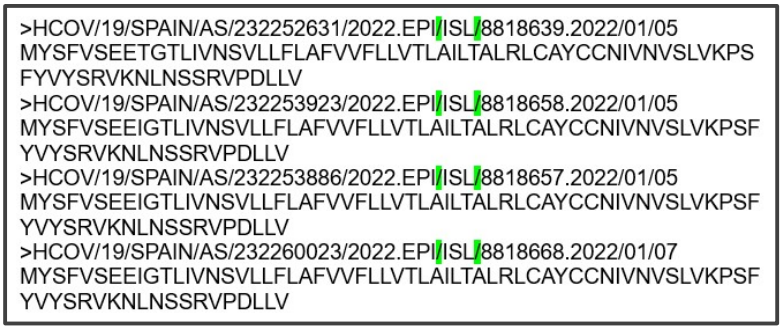
In the '**Replace**' field, input the new characters that will replace the previous ones (e.g., /).

Choose '**Sequence**' if you want to replace characters within the sequence, or '**Header**' if you want to replace characters within the header.

For the **output**, select the output folder where you want the modified '.fasta' file to appear without naming it. The files are automatically named as follows: 'Replace_InputFileName.fasta'.

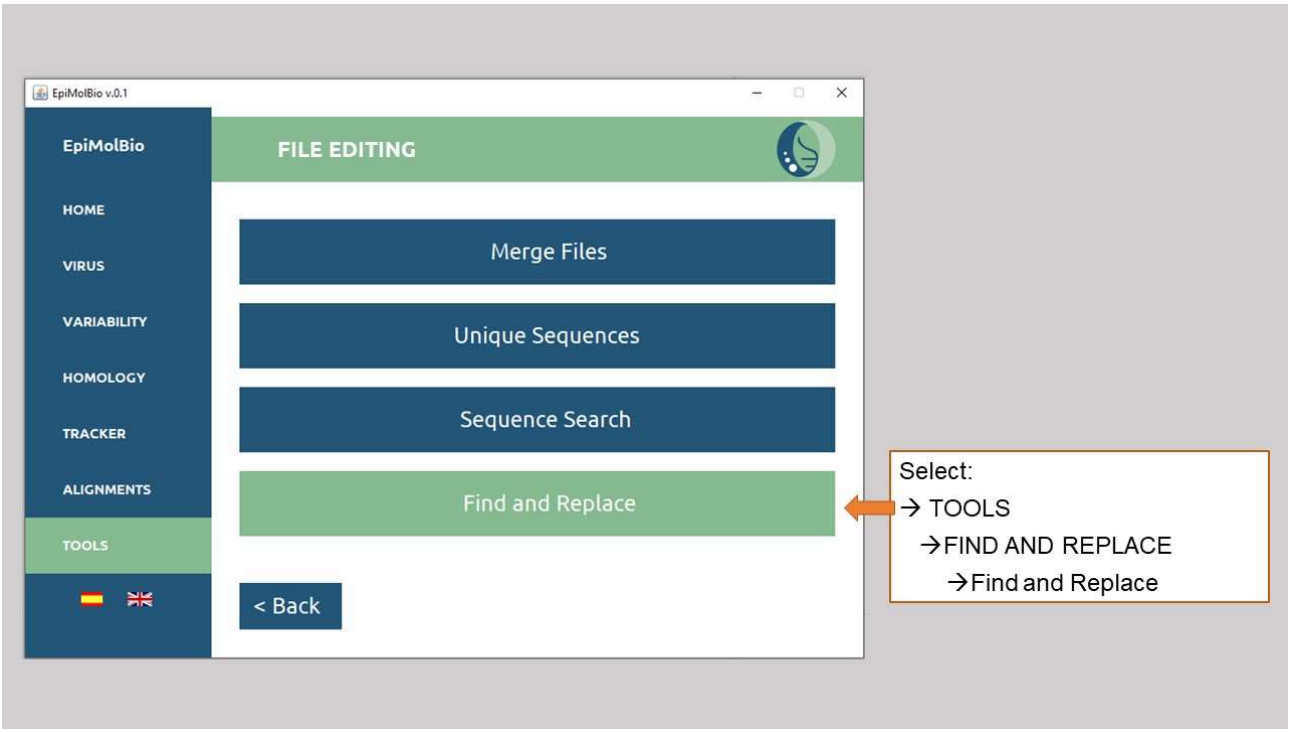
Example of original file and modified file from the Find and Replace tool:



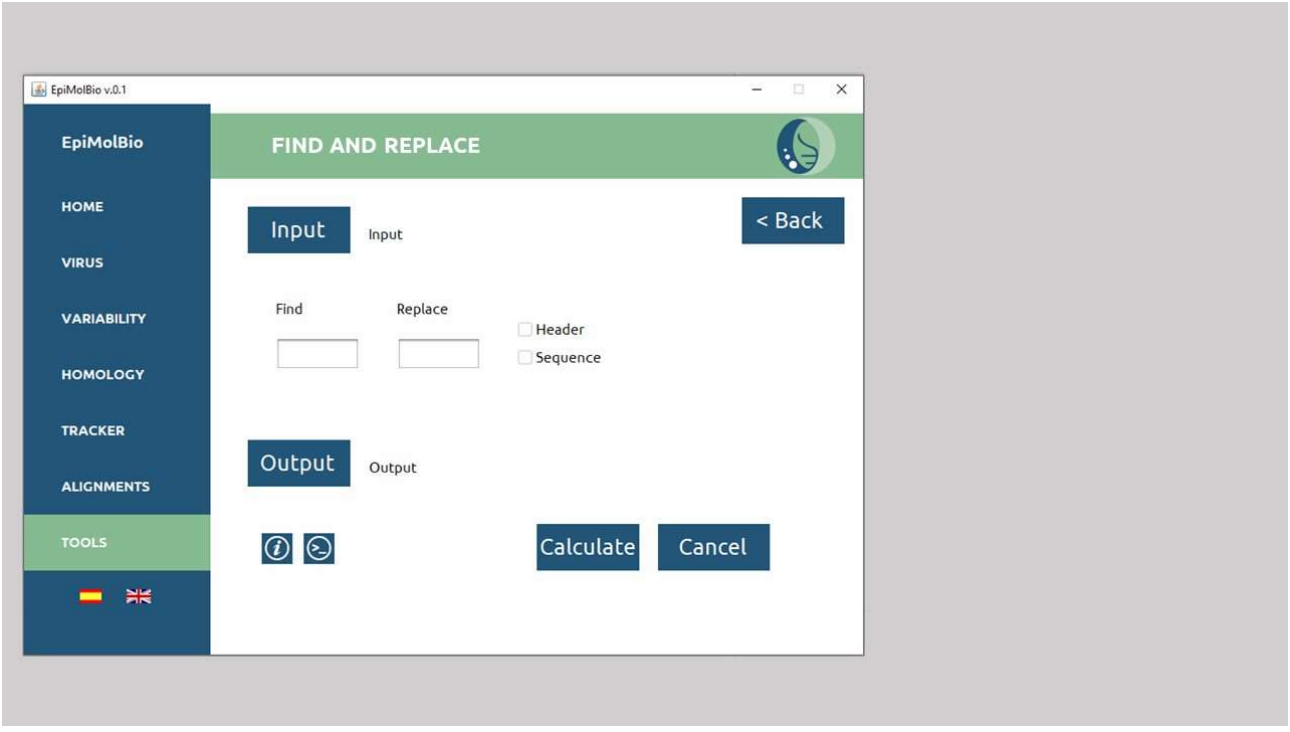



# Step-by-step:

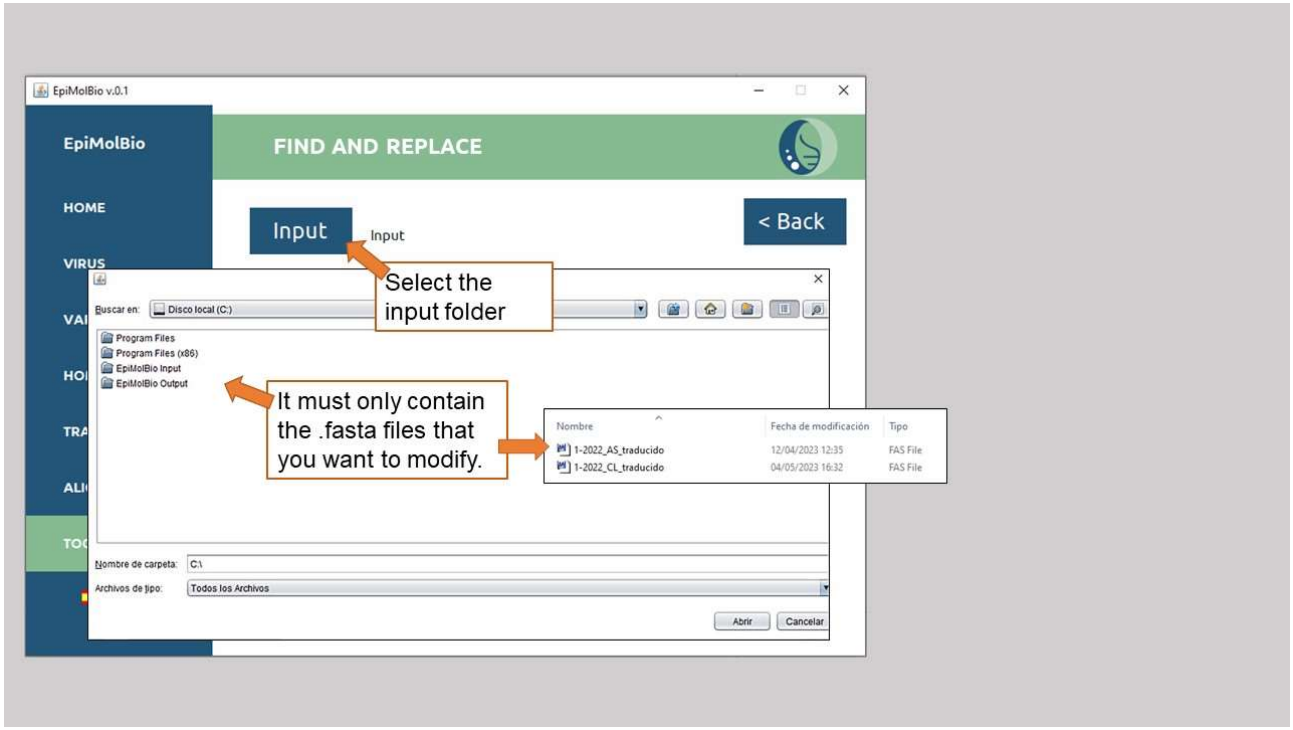
## 1)



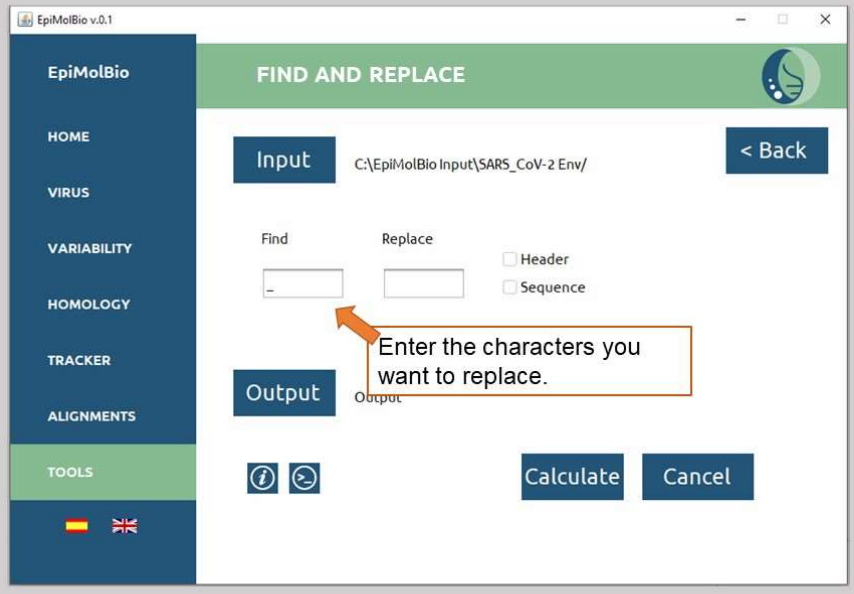

## 2)



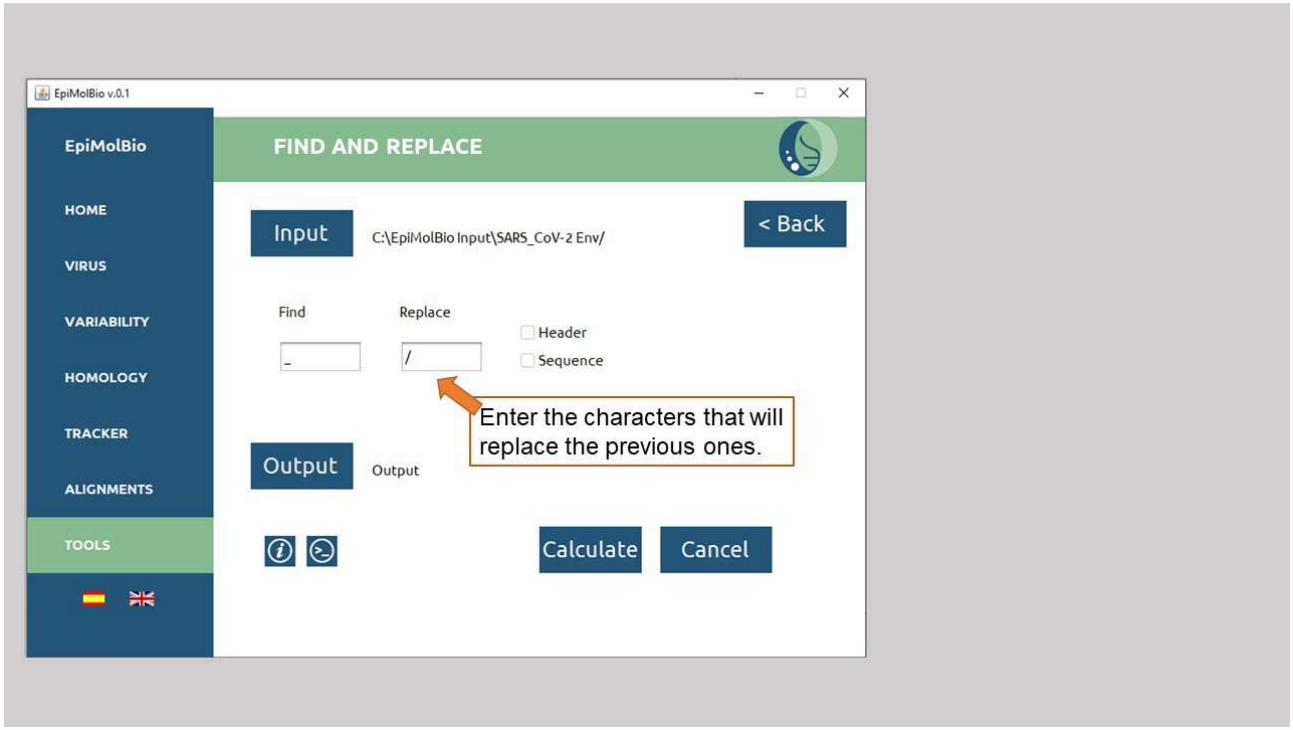
## 3)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
FIND AND REPLACE
Input
Input
< Back
Select the input folder
Buscar en:
Disco local (C:)
Program Files
Program Files (x86)
EpiMolBio Input
EpiMolBio Output
It must only contain the .fasta files that you want to modify.
Nombre
Fecha de modificación
Tipo
1-2022_AS_traducido
12/04/2023 12:35
FAS File
1-2022_CL_traducido
04/05/2023 16:32
FAS File
Nombre de carpeta:
C:\
Archivos de tipo:
Todos los Archivos
Abrir
Cancelar

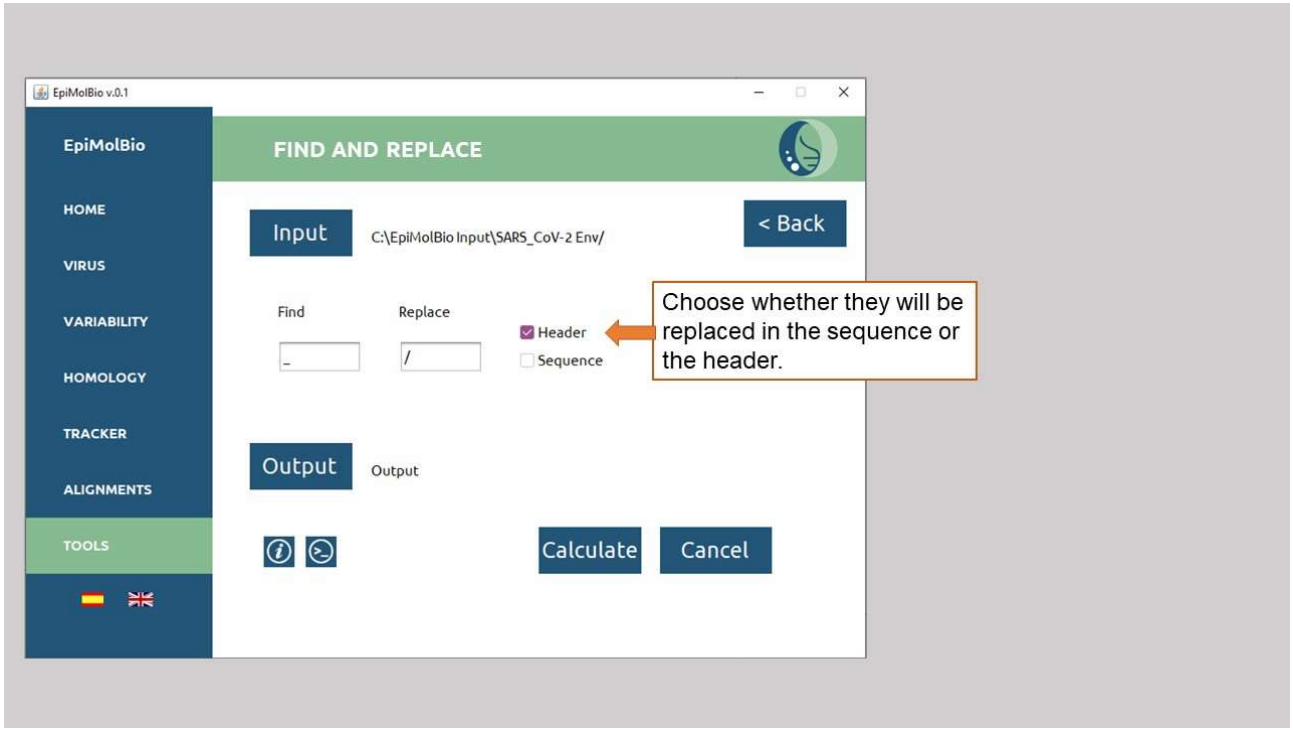
## 4)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
FIND AND REPLACE
Input
C:\EpiMolBio Input\SARS_CoV-2 Env/
< Back
Find
Replace
Header
Sequence
Enter the characters you want to replace.
Output
Output
Calculate
Cancel

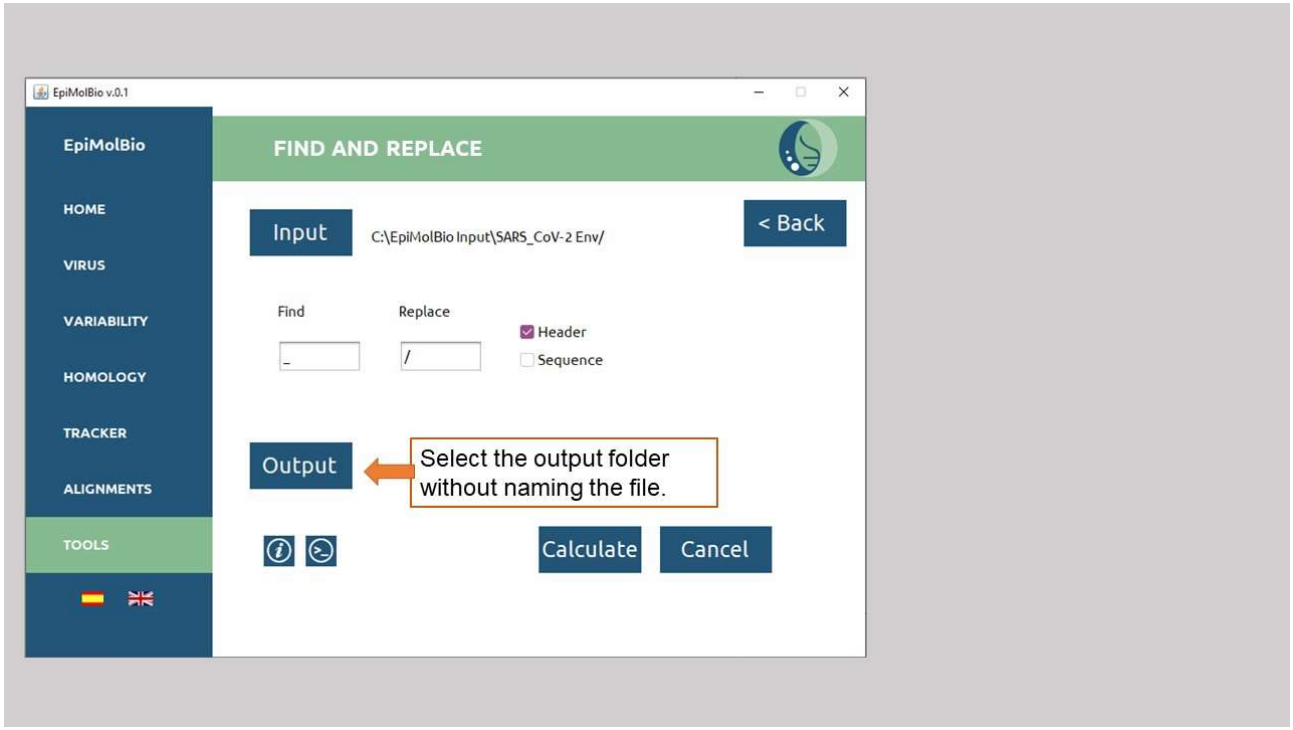
## 5)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
FIND AND REPLACE
Input
C:\EpiMolBio Input\SARS_CoV-2 Env/
< Back
Find
Replace
_
/
Header
Sequence
Enter the characters that will replace the previous ones.
Output
Output
Calculate
Cancel

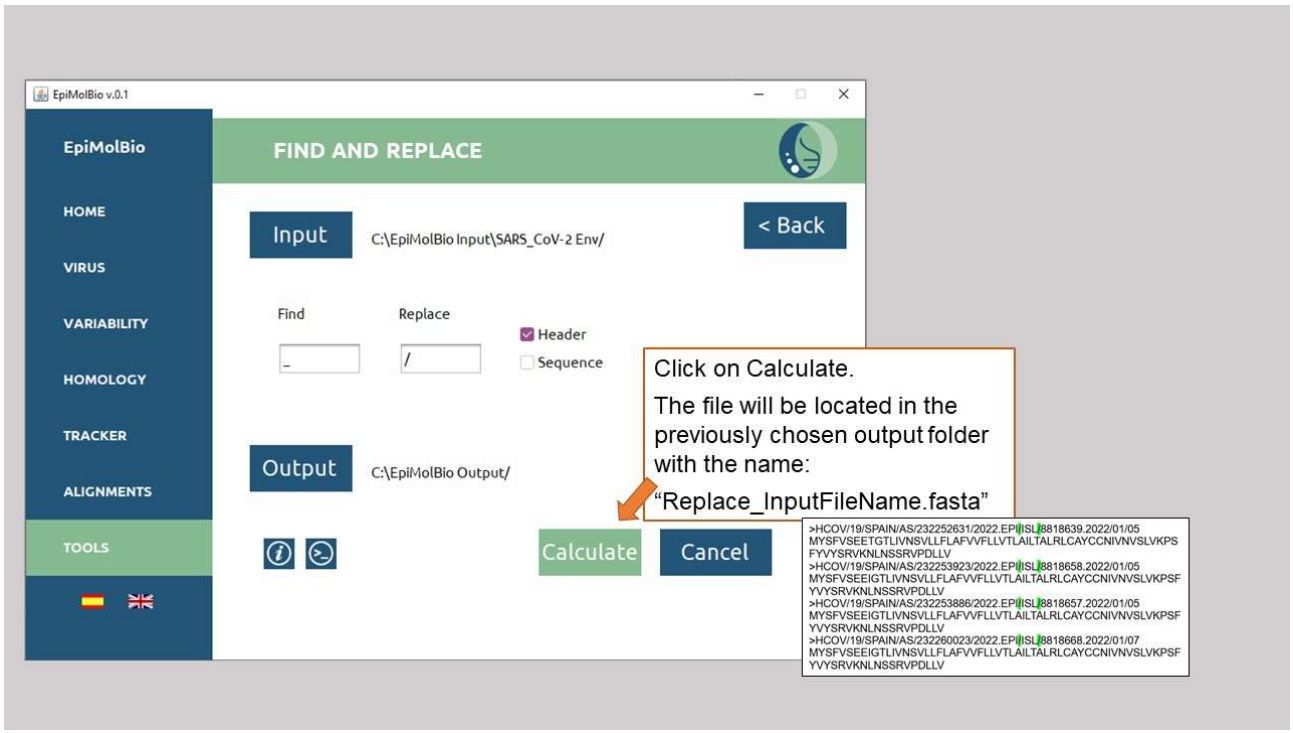
## 6)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
FIND AND REPLACE
Input
C:\EpiMolBio Input\SARS_CoV-2 Env/
< Back
Find
Replace
_
/
Header
Sequence
Choose whether they will be replaced in the sequence or the header.
Output
Output
Calculate
Cancel

## 7)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
FIND AND REPLACE
Input
C:\EpiMolBio Input\SARS_CoV-2 Env/
< Back
Find
Replace
_
/
Header
Sequence
Output
Select the output folder without naming the file.
Calculate
Cancel

## 8)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
FIND AND REPLACE
Input
C:\EpiMolBio Input\SARS_CoV-2 Env/
< Back
Find
Replace
_
/
Header
Sequence
Output
C:\EpiMolBio Output/
Click on Calculate.
The file will be located in the previously chosen output folder with the name:
"Replace_InputFileName.fasta"
Calculate
Cancel

# VI.2.FILTERS

## VI.2.A) HEADER FILTERING

This tool is used to **filter one or multiple '.fasta' format files based on parameters in their headers**, separating them into **different files according to the chosen parameter**. It also allows for **gap removal and simultaneous translation**.

For example, you can separate sequences by their variant, country of origin, year of sampling, name, or accession number based on the following header: 10_CD.ES.1998. AF000454.IC2258.

The **input** file should be the folder containing exclusively the '.fasta' files that you want to filter. All of them should have the same type of information in each parameter of the header and in the same order.

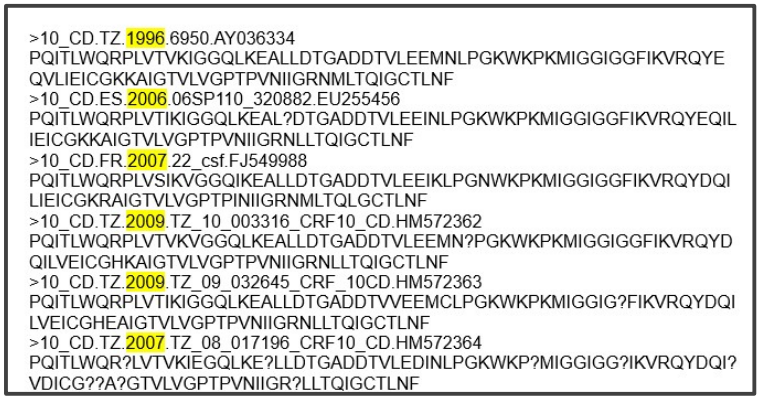
In the '**Header**' field, choose the item by which you want to perform the filtering. EpiMolBio includes up to 5 items. In the previous example, if you want to filter by year, you would need to select item 3.

In the '**Separator**' field, input the character that serves as a separator in the sequence headers. In the previous example, it would be a period '.'; in other cases, it could be another character such as '_' or '/'.

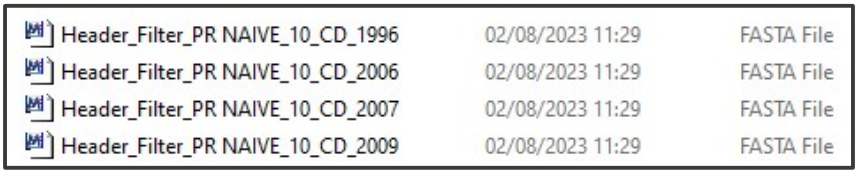
In the '**Options**' field, you can select '**Remove Gaps**' to eliminate all gaps and '**Translate**' to translate from nucleotides to amino acids.

For the **output**, select the output folder where you want the filtered file to appear without naming it. A separate file is generated for each filtered item, containing the '.fasta' sequences that share the same characters for that item. These files are automatically named as follows: 'Header_Filter_InputFileName.fasta'.

Example of input .fasta file for filtering by year:

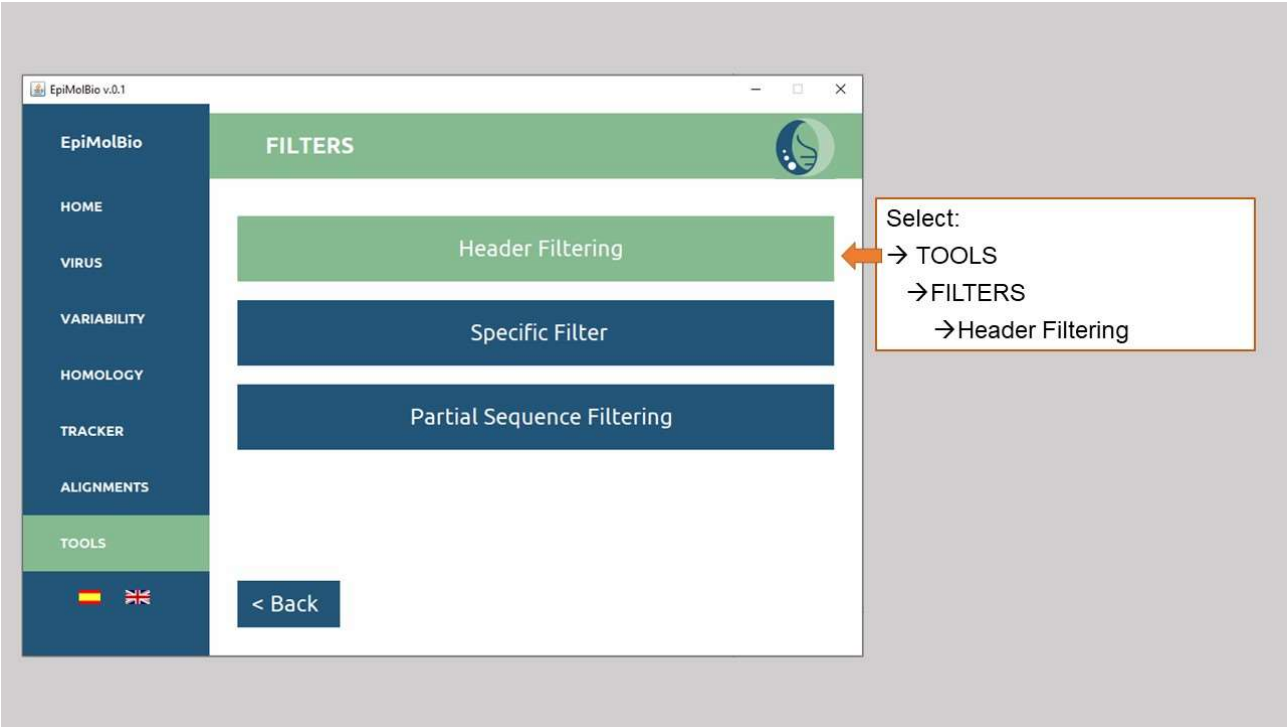



Example of output files with .fasta sequences grouped by year:

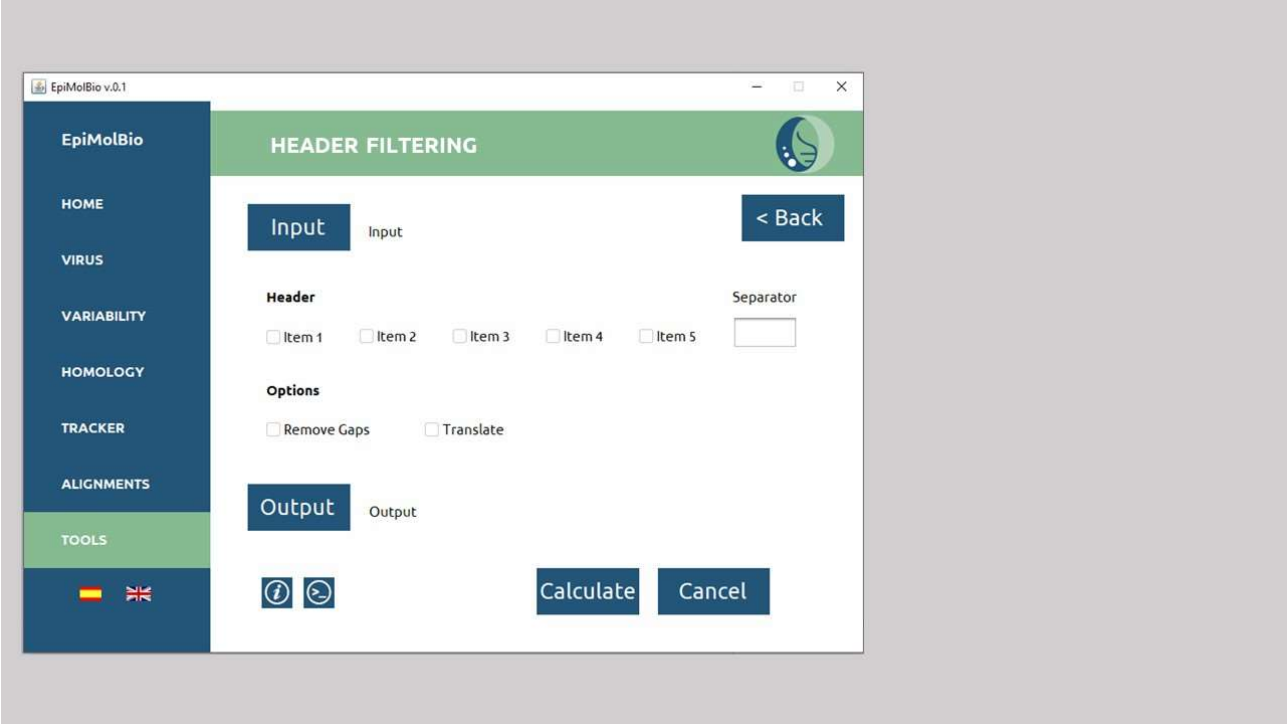


# Step-by-step:

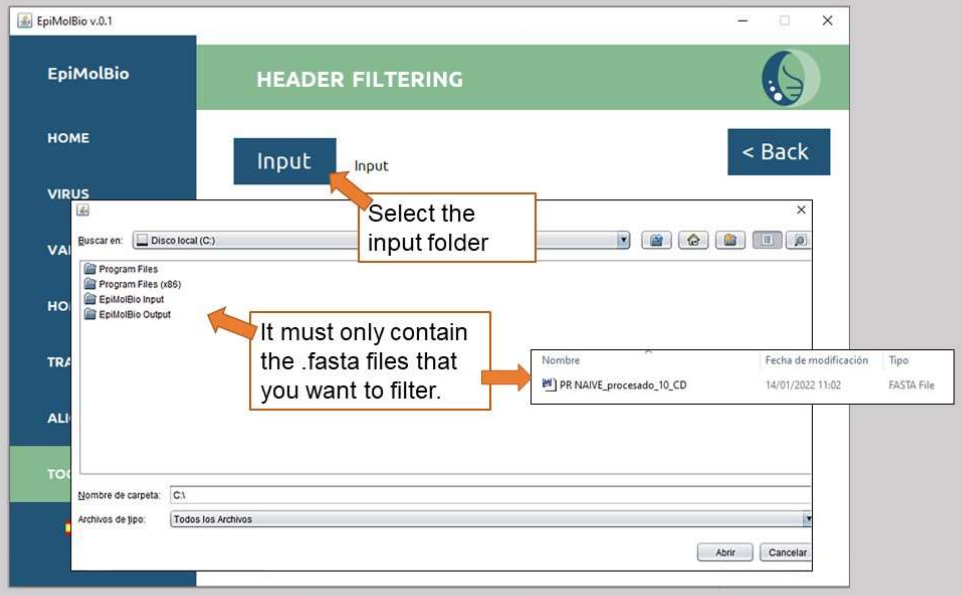
## 1)



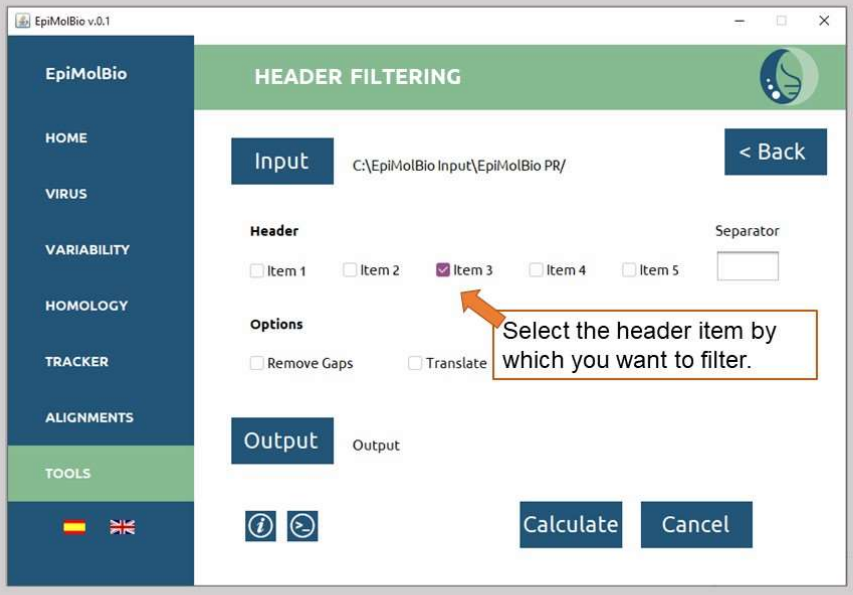

## 2)



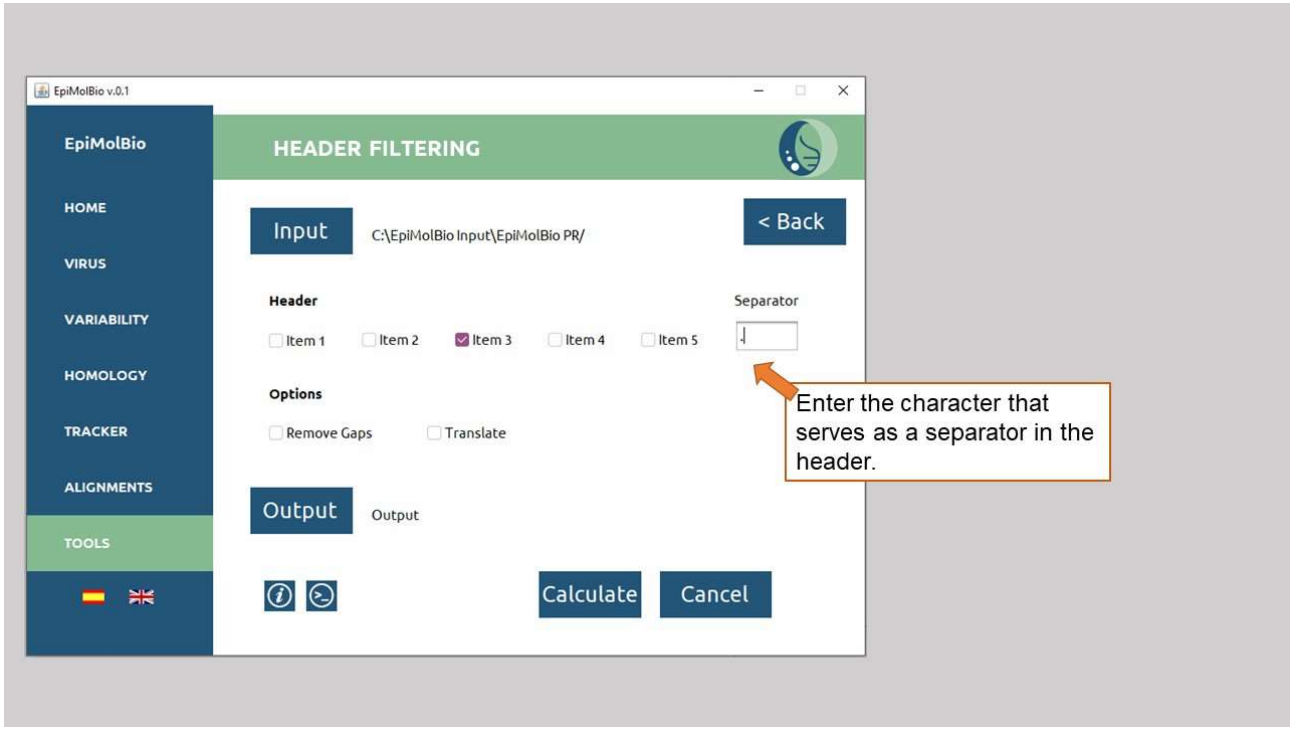
## 3)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
HEADER FILTERING
Input
Input
< Back
Select the input folder
Buscar en:
Disco local (C:)
Program Files
Program Files (x86)
EpiMolBio Input
EpiMolBio Output
It must only contain the .fasta files that you want to filter.
Nombre
Fecha de modificación
Tipo
PR NAIVE_procesado_10_CD
14/01/2022 11:02
FASTA File
Nombre de carpeta:
C:\
Archivos de tipo:
Todos los Archivos
Abrir
Cancelar

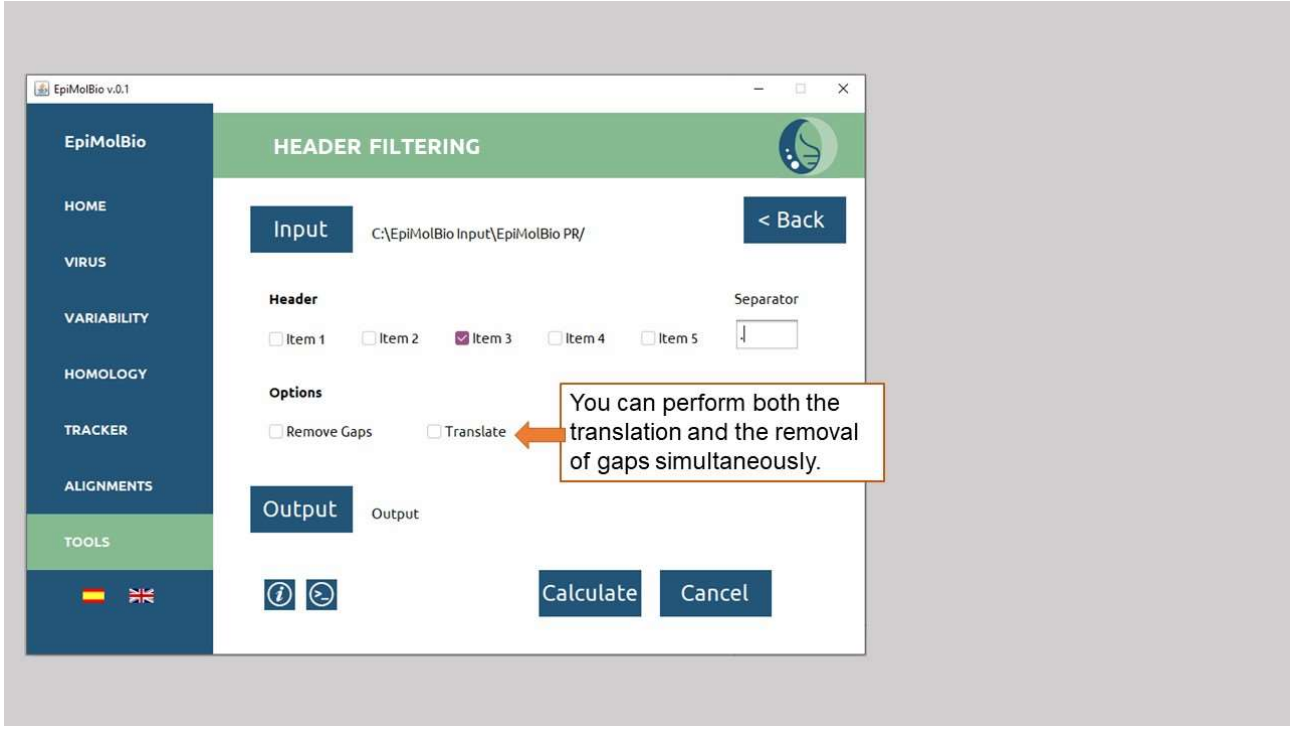
## 4)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
HEADER FILTERING
Input
C:\EpiMolBio Input\EpiMolBio PR/
< Back
Header
Separator
Item 1
Item 2
Item 3
Item 4
Item 5
Options
Remove Gaps
Translate
Select the header item by which you want to filter.
Output
Output
Calculate
Cancel

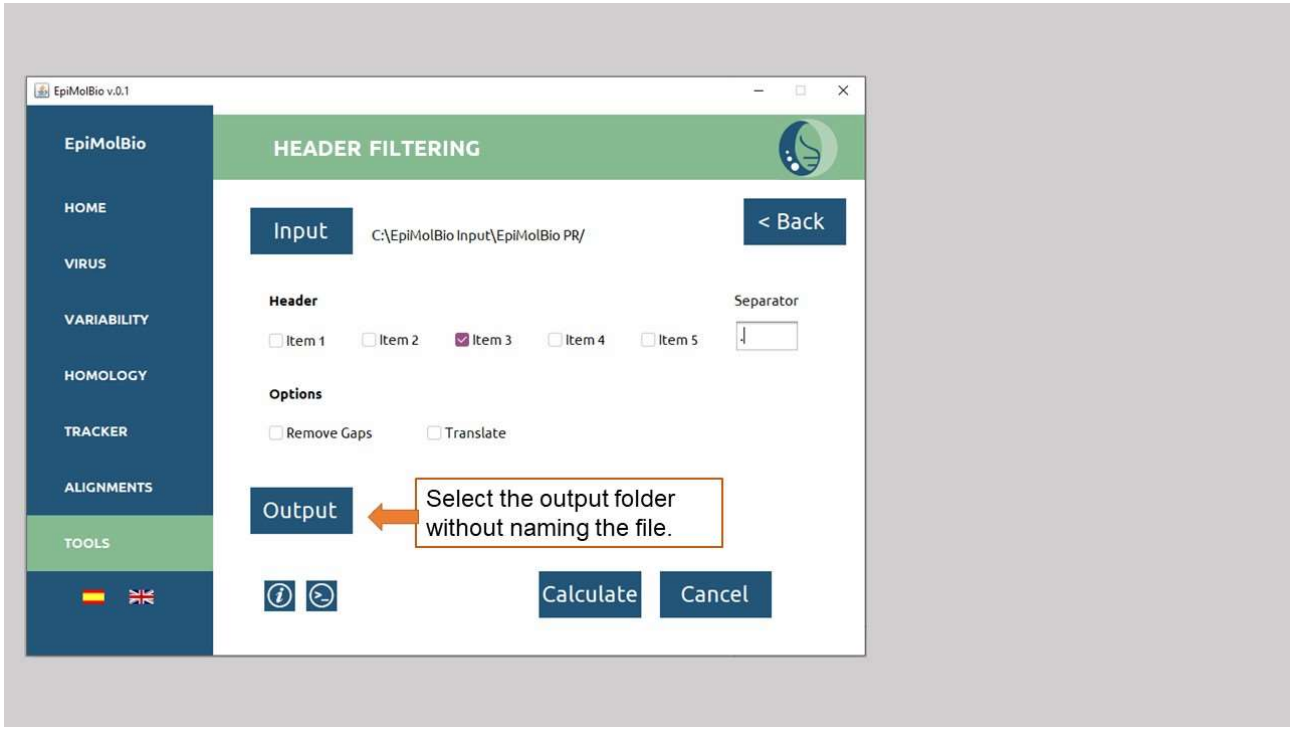
## 5)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
HEADER FILTERING
Input
C:\EpiMolBio Input\EpiMolBio PR/
< Back
Header
Separator
Item 1
Item 2
Item 3
Item 4
Item 5
Options
Remove Gaps
Translate
Output
Output
Calculate
Cancel
Enter the character that serves as a separator in the header.

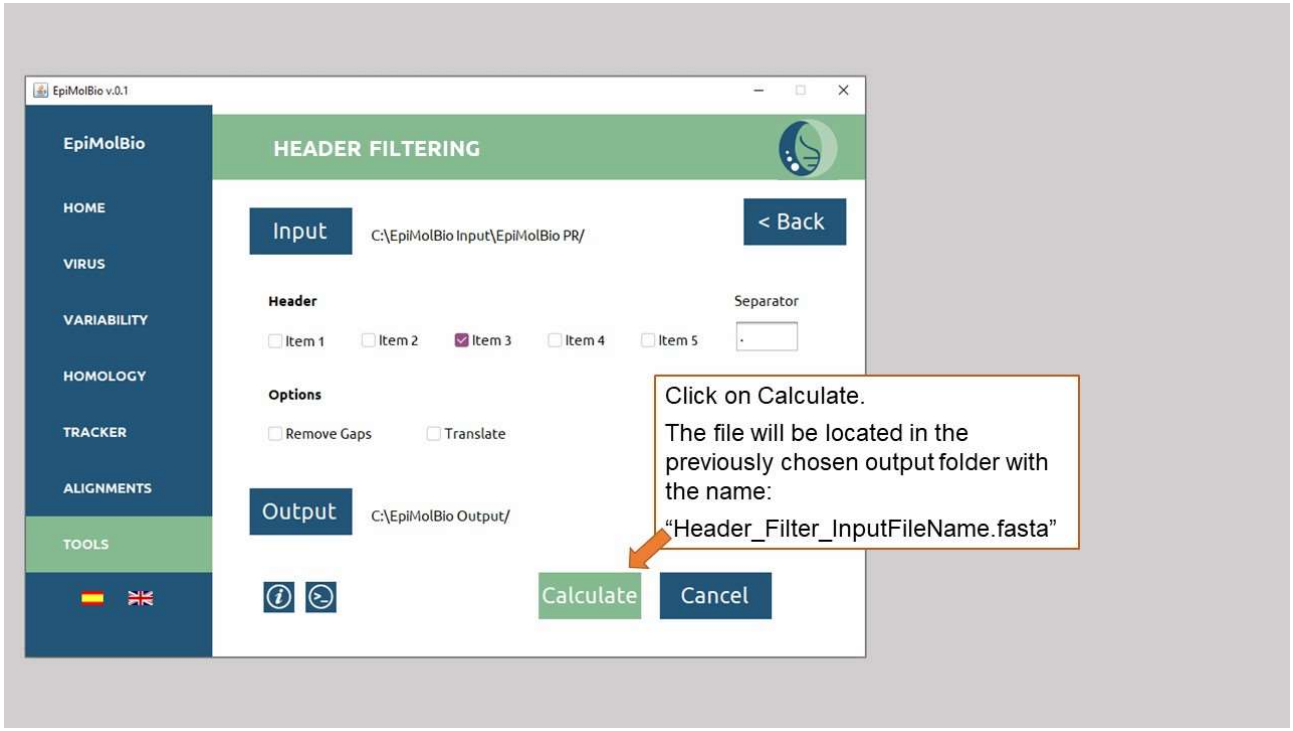
## 6)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
HEADER FILTERING
Input
C:\EpiMolBio Input\EpiMolBio PR/
< Back
Header
Separator
Item 1
Item 2
Item 3
Item 4
Item 5
Options
Remove Gaps
Translate
Output
Output
Calculate
Cancel
You can perform both the translation and the removal of gaps simultaneously.

## 7)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
HEADER FILTERING
Input
C:\EpiMolBio Input\EpiMolBio PR/
< Back
Header
Separator
Item 1
Item 2
Item 3
Item 4
Item 5
Options
Remove Gaps
Translate
Output
Select the output folder without naming the file.
Calculate
Cancel

## 8)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
HEADER FILTERING
Input
C:\EpiMolBio Input\EpiMolBio PR/
< Back
Header
Separator
Item 1
Item 2
Item 3
Item 4
Item 5
Options
Remove Gaps
Translate
Output
C:\EpiMolBio Output/
Click on Calculate.
The file will be located in the previously chosen output folder with the name:
“Header_Filter_InputFileName.fasta”
Calculate
Cancel

# VI.2.B) SPECIFIC FILTER

This tool is used to **filter sequences that have a specific set of characters in their headers** from files in .fasta format.

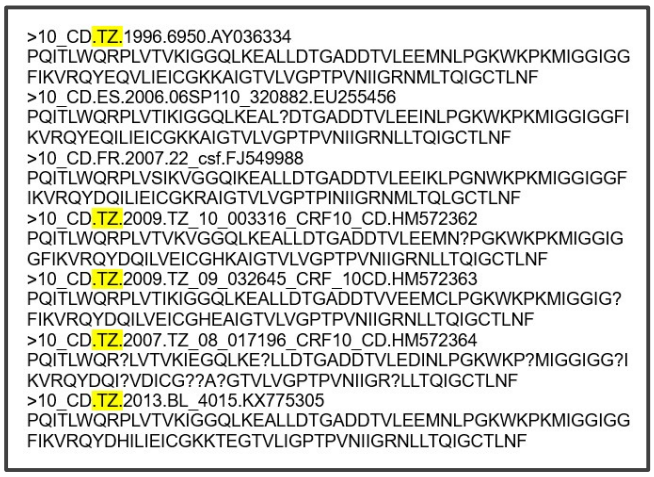
For example, filtering sequences by their variant, country of origin, year of sample collection, name, or accession number based on the following header: 10_CD.TZ.1996.AF000454.IC2258.

The **input** file should be the folder containing exclusively the .fasta files that you want to filter. All files should have the same type of information in each parameter of the header and in the same order.

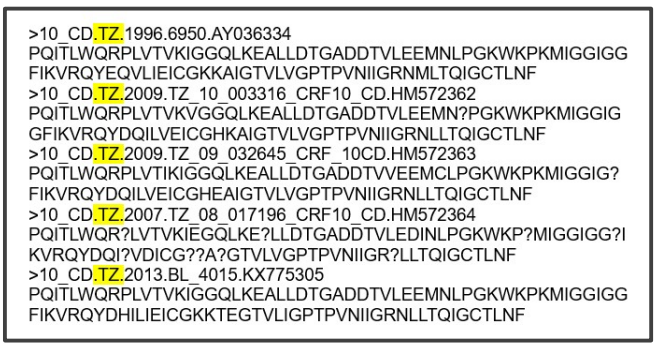
In the '**Filtering Sequence**' field, enter the series of characters by which you want to filter, along with the preceding and following separators. For example, if you want to filter by the country of origin, selecting Tanzania (TZ), you would enter '.TZ.'

For the **output**, select the output folder where you want the filtered file to appear without naming it. For each input file, an output file is generated containing the sequences that include the specified set of characters. These files are automatically named as follows: 'Specific_Filter_InputFileName.fasta'.

Example of input .fasta file for filtering by country of origin (Tanzania or TZ):

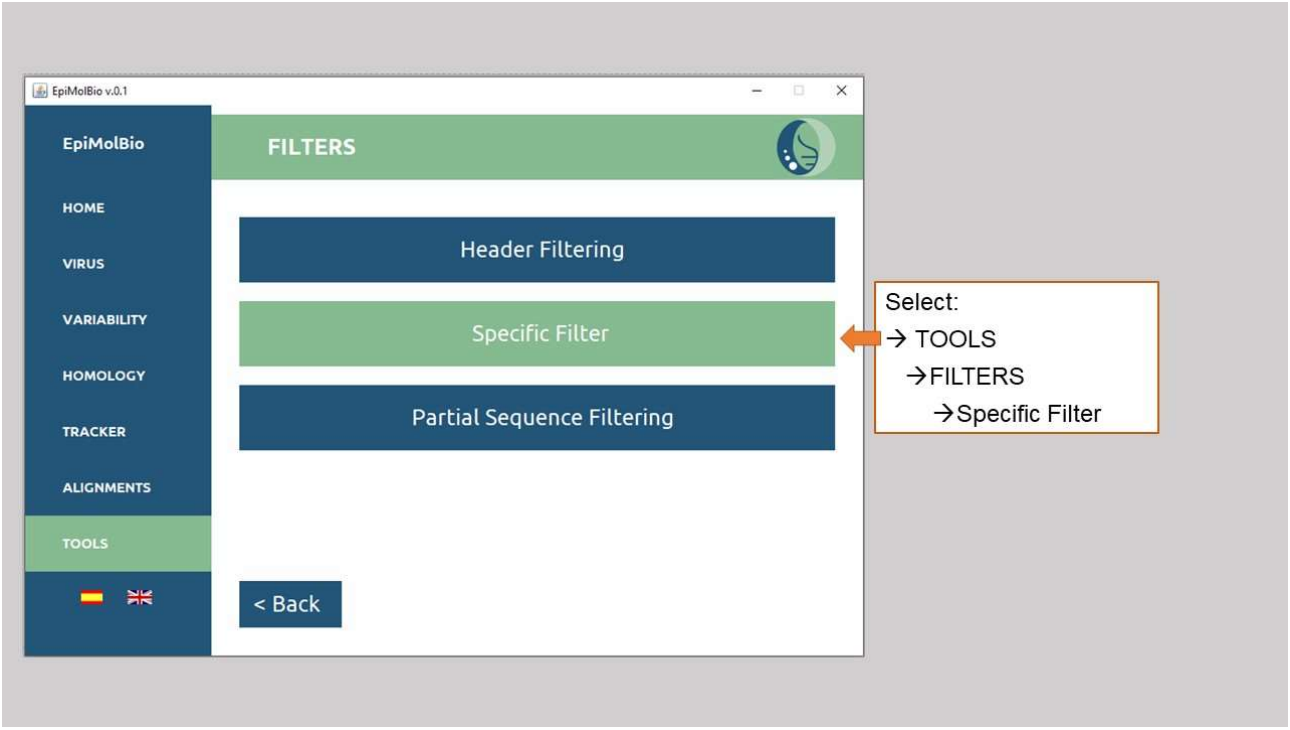



Example of output file with the .fasta sequences containing 'TZ' in their header:

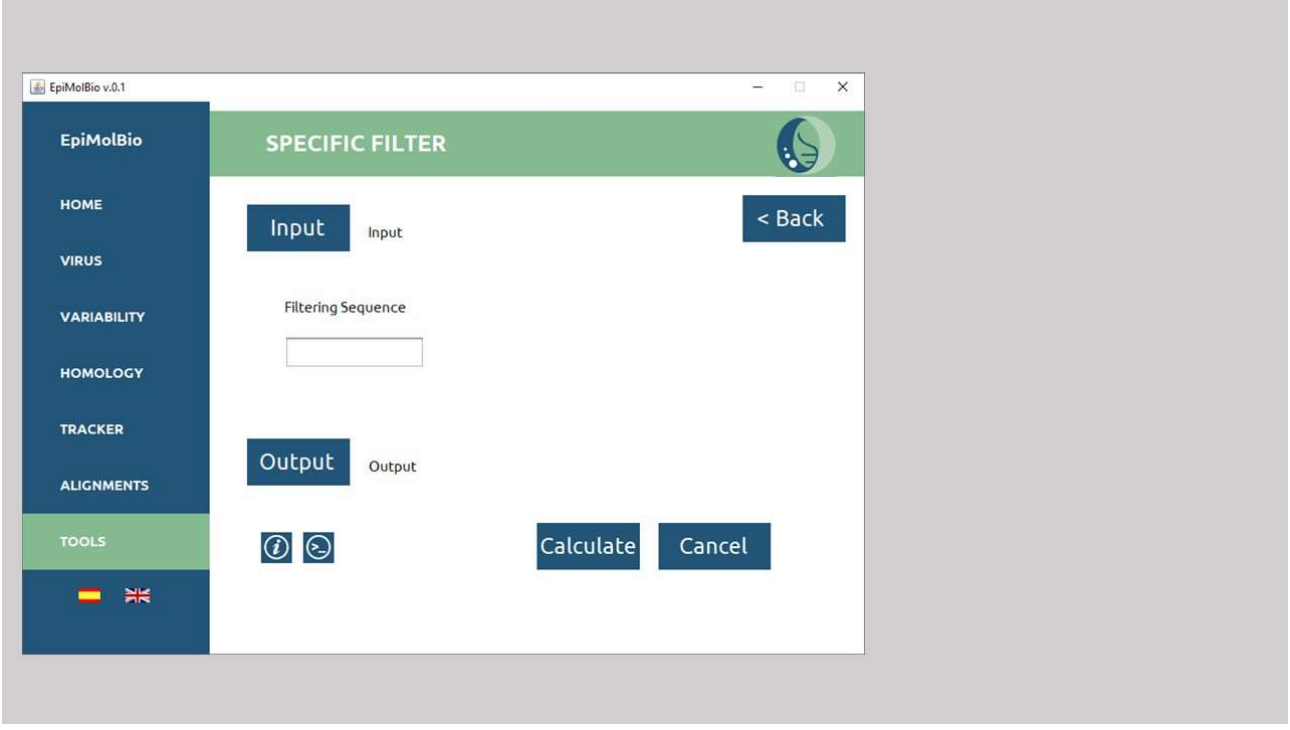


# Step-by-step:

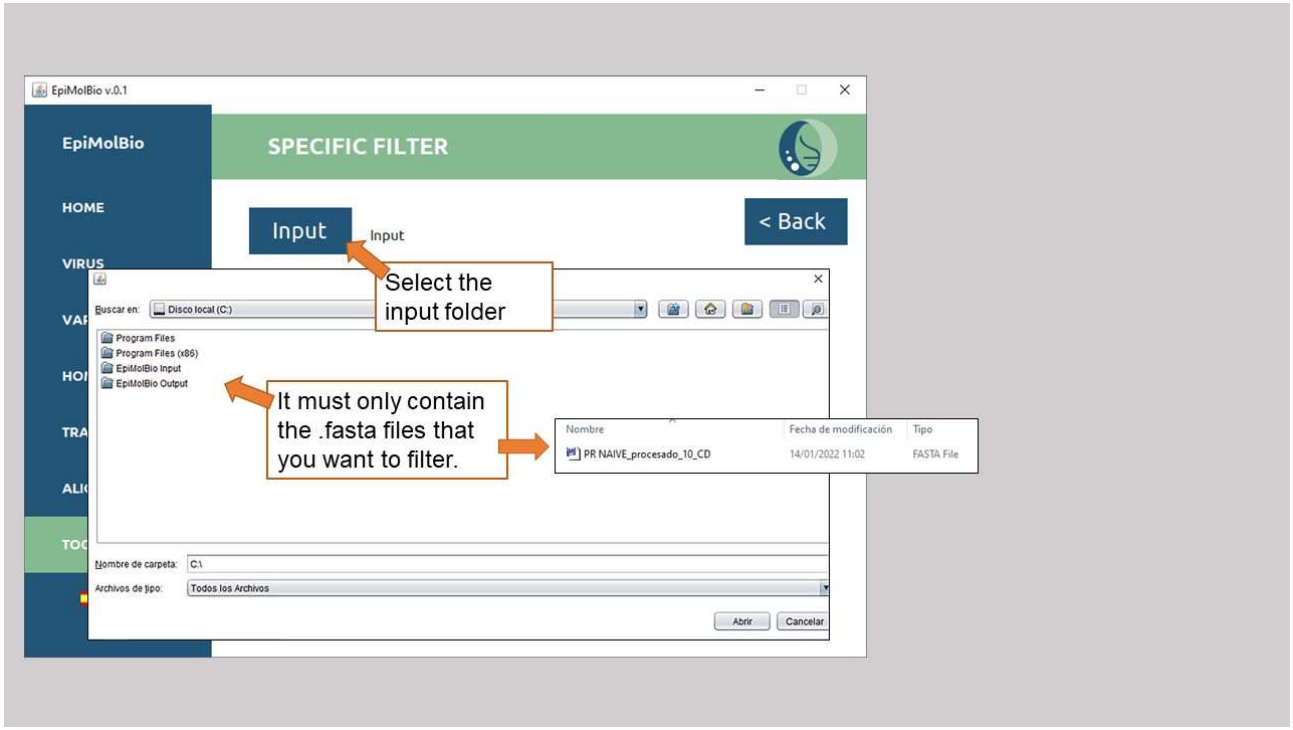
## 1)



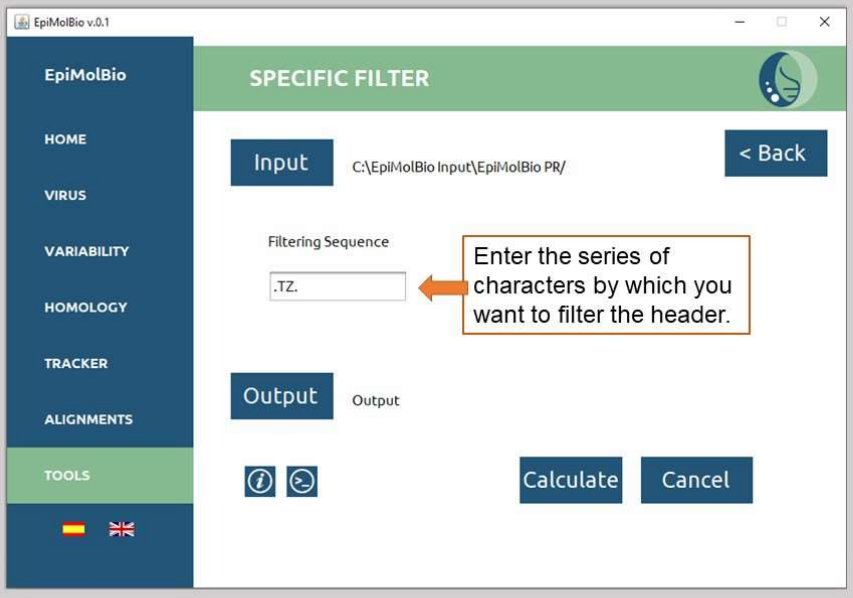

## 2)



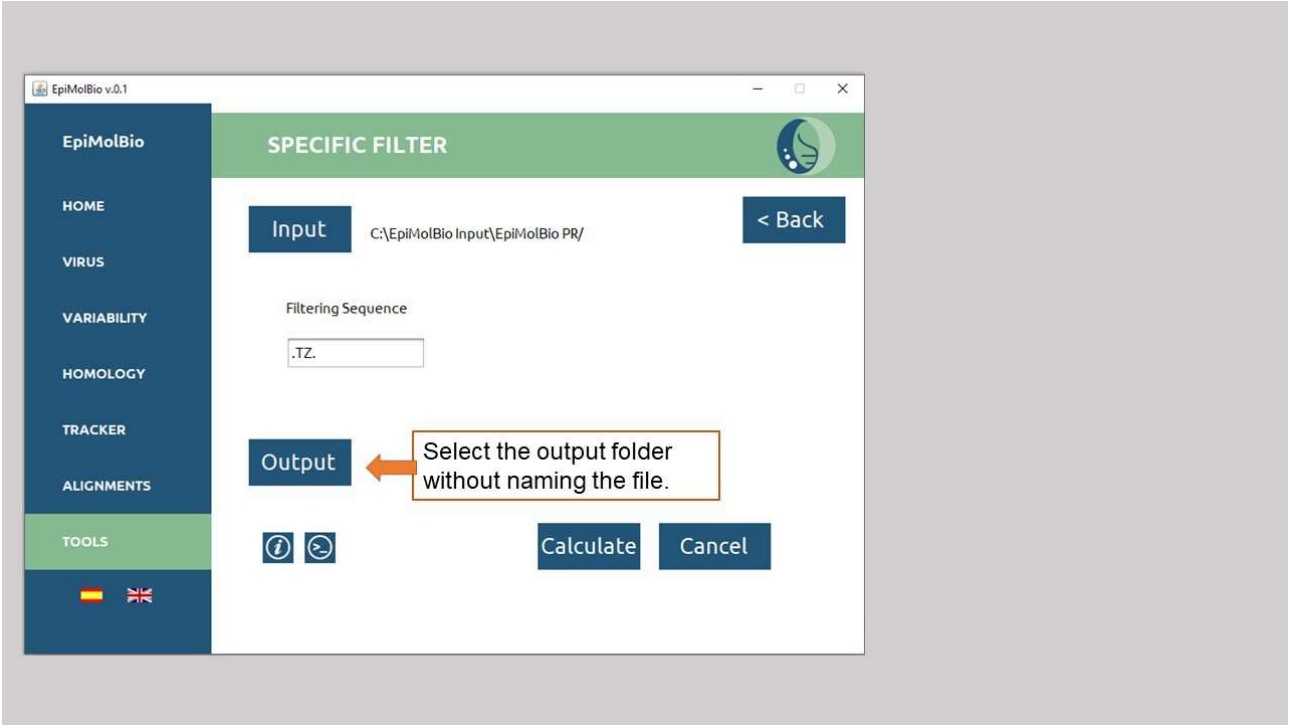
## 3)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
SPECIFIC FILTER
Input
Input
< Back
Select the input folder
Buscar en:
Disco local (C:)
Program Files
Program Files (x86)
EpiMolBio Input
EpiMolBio Output
It must only contain the .fasta files that you want to filter.
Nombre
Fecha de modificación
Tipo
PR NAIVE_procesado_10_CD
14/01/2022 11:02
FASTA File
Nombre de carpeta:
C:\
Archivos de tipo:
Todos los Archivos
Abrir
Cancelar

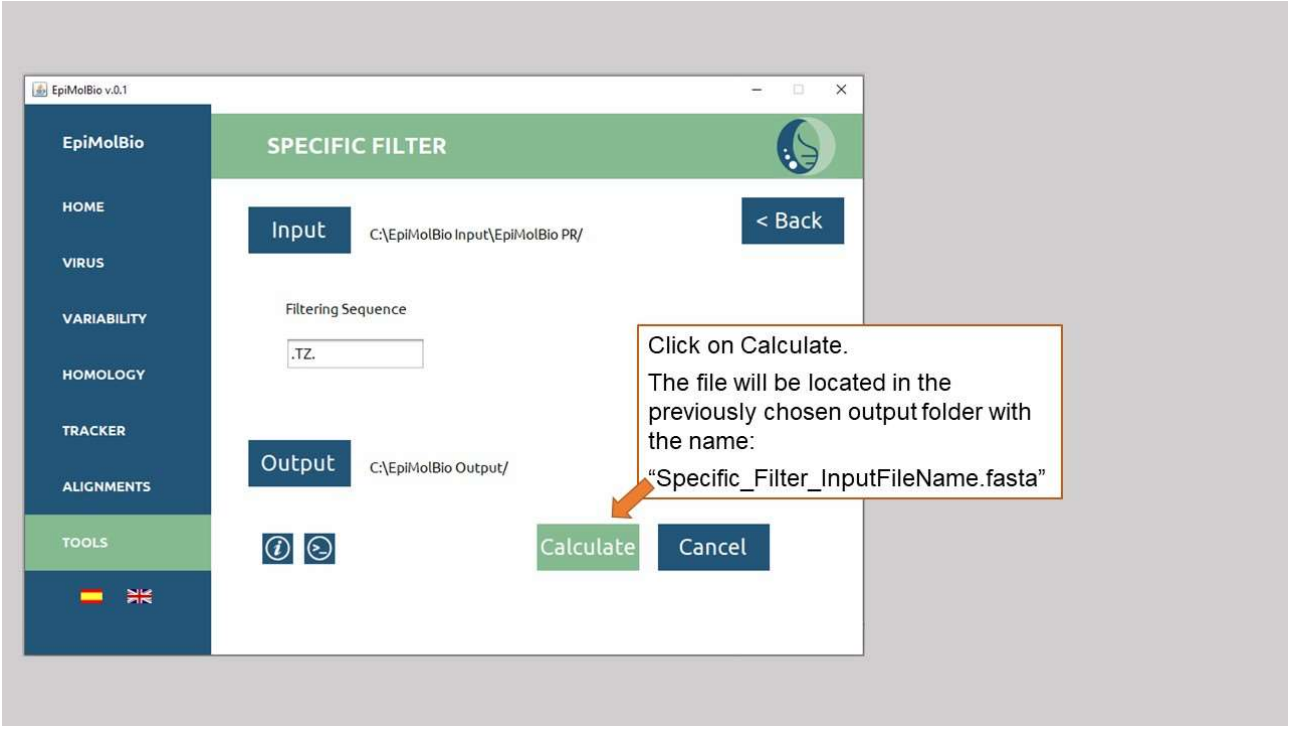
## 4)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
SPECIFIC FILTER
Input
C:\EpiMolBio Input\EpiMolBio PR/
< Back
Filtering Sequence
.TZ.
Enter the series of characters by which you want to filter the header.
Output
Output
Calculate
Cancel

## 5)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
SPECIFIC FILTER
Input
C:\EpiMolBio Input\EpiMolBio PR/
< Back
Filtering Sequence
.TZ.
Output
Select the output folder without naming the file.
Calculate
Cancel

## 6)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
SPECIFIC FILTER
Input
C:\EpiMolBio Input\EpiMolBio PR/
< Back
Filtering Sequence
.TZ.
Output
C:\EpiMolBio Output/
Click on Calculate.
The file will be located in the previously chosen output folder with the name:
“Specific_Filter_InputFileName.fasta”
Calculate
Cancel

# VI.2.C) PARTIAL SEQUENCE FILTERING

This tool allows you to **filter sequences based on their quality**. This quality depends on the amount of unknown residues contained in the sequences, appearing as '?' in amino acid sequences or as 'N' in nucleotide sequences. A higher quantity of these characters indicates lower quality. The tool enables you to set a quality threshold to obtain one or more .fasta files containing sequences that exceed this quality threshold. Additionally, an .html file is generated that displays the sequences that were lost when applying this filter.

The **input** file should be the folder containing exclusively the .fasta files that you want to filter.

In the '**Sequence Type**' field, select whether the input sequences are in nucleotides or amino acids.

In the '**Filter %**' field, input the percentage value for the filtering threshold with one decimal place. For example, entering 95.0 in the filtering percentage will remove sequences containing 5% or more 'N' from the output file, leaving those that contain none or <5% 'N'. If you enter 100.0, only sequences without 'N' will remain.

For the **output**, select the output folder where you want the filtered '.fasta' file to appear without naming it. The '.fasta' file will be automatically named as follows: Partial_Filter_InputFileName.fasta. The resulting '.html' file will be automatically named as Lost_Sequences.html.

Example of input file with 15 sequences, 7 of them with '?':

Example of the output .fasta file after applying the 100% filter, showing the 8 sequences without any '?':

```
>06_cpx.ES.2000.19804.AF354004
PQITLWQRPLVTVRIGGQLIEALLDTGADDTVLEDINLPGKWRPKMIGGIGGFIKVRQYDQI
LMEICGKKAIGTVLVGPTPVNIIGRNMLTQIGCTLNF
>06_cpx.CI.2001.pci123.AY207737
PQITLWQRPLVTVKIGGQLIEALLDTGADDTVLEDINLPGKWKPKMIGGIGGFIKVRQYDQI
PIEICGKRAIGTVLVGPTPVNIIGRNMLTQIGCTLNF
>06_cpx.ES.2000.SP2756_00.AY248312
PQITLWQRPLVTVRIGGQLIEALLDTGADDTVLEEINLPGKWKPKMIGGIGGFIKVRQNDQI
LIEICGKKAIGTVLVGPTPVNIIGRNMLTQIGCTLNF
>06_cpx.CM.2000.00CMNYU3475.AY359728
PQITLWQRPLVTVRIGGQLIEALLDTGADDTVLEDINLPGKWKPKMIGGIGGFIKVRQYDQI
PIEICGKRAIGTVLVGPTPVNIIGRNMLTQIGCTLNF
>06_cpx.NG.-.DDJ077.DQ273946
PQITLWQRPLVTVKIGGQLIEALLDTGADDTVLEEINLPGKWKPKMIGGIGGFIKVRQYDQIL
IEICGKKAIGTVLVGPTPVNIIGRNMLTQIGCTLNF
>06_cpx.NG.-.DDJ085.DQ273952
PQITLWQRPLVTVKVGGQLIEALLDTGADDTVLEDINLPGKWKPKMIGGIGGFIKVRQYDQI
LMEICGKKAIGTVLVGPTPVNIIGRNMLTQIGCTLNF
>06_cpx.NG.-.DDJ262.DQ273960
PQITLWQRPLVTVRIGEQLIEALLDTGADDTVLEDINLPGKWKPKMIGGIGGFIKVRQYDQIL
IEICGKRAIGTVLVGPTPVNIIGRNMLTQIGCTLNF
>06_cpx.NG.-.DDJ270.DQ273962
PQITLWQRPLVTVRIGGQLIEALLDTGADDTVLEDINLPGKWKPKMIGGIGGFIKVRQYDQI
HIEICGKKAIGTVLVGPTPVNIIGRNMLTQIGCTLNF
```

Example of the output .html file after applying the 100% filter, showing the number and percentage of lost sequences (the 7 sequences with '?'):

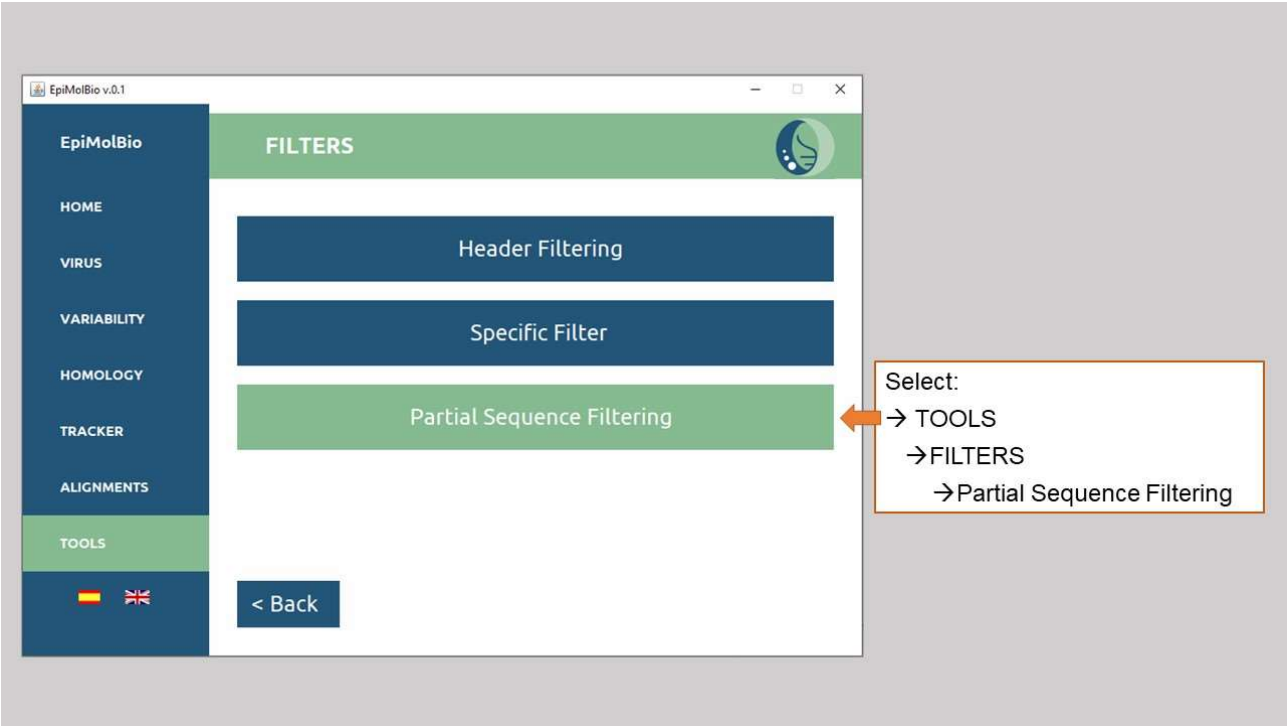
**Partial Sequence Filtering**

| File | Total Sequences | Recovered Sequences | Lost Sequences | Loss Percentage |
|---|---|---|---|---|
| 06_cpx.fasta | 15 | 8 | 7 | 46.666% |
| Total | 15 | 8 | 7 | 46.666% |

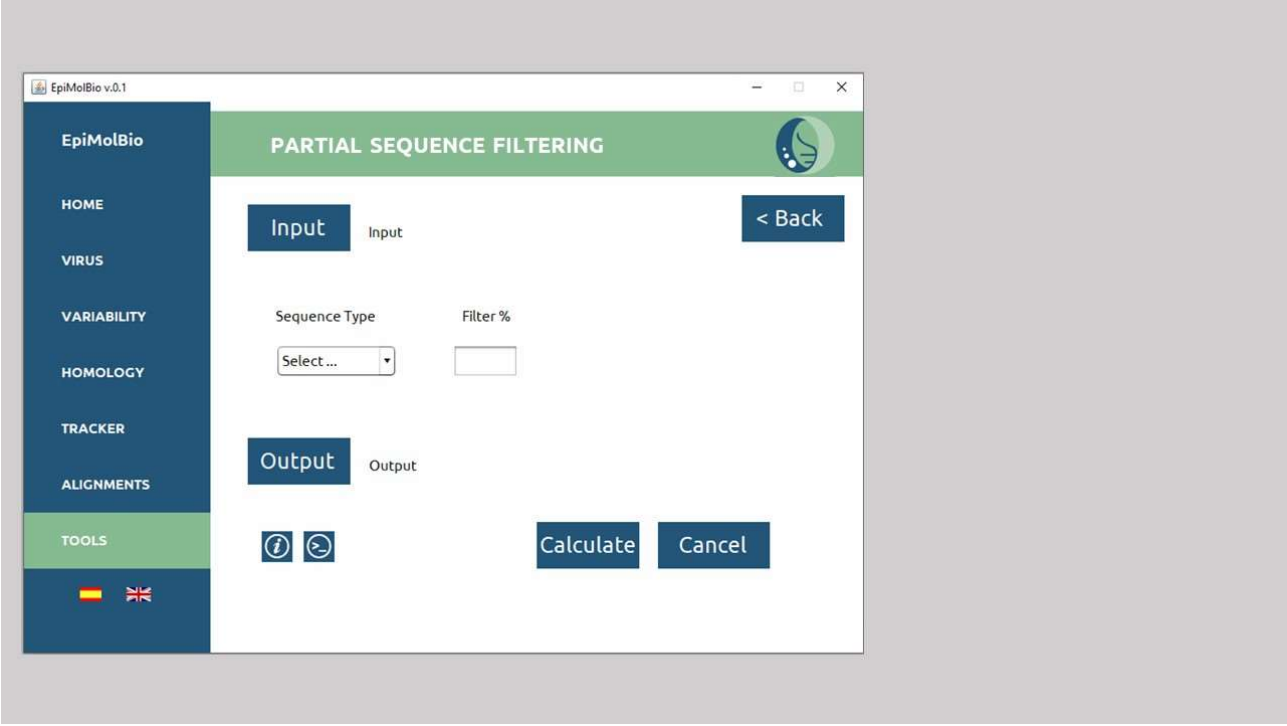
The resulting .html file shows at the top, the title of the analysis. In the ‘File’ column, the name of the input file will be displayed. ‘Total Sequences’ will show the total number of input sequences. The ‘Recovered Sequences’ column will display the number of sequences in the output file that have passed the quality criteria. The ‘Lost Sequences’ column will show the number of sequences that were removed because they did not meet the quality criteria. The ‘Loss Percentage’ column will display the percentage of lost sequences relative to the total, with color-coding based on the color code explained in the Overview section of the output ‘.html’ file. You can access this color code by clicking on the blue symbol in the output file.

# Step-by-step:

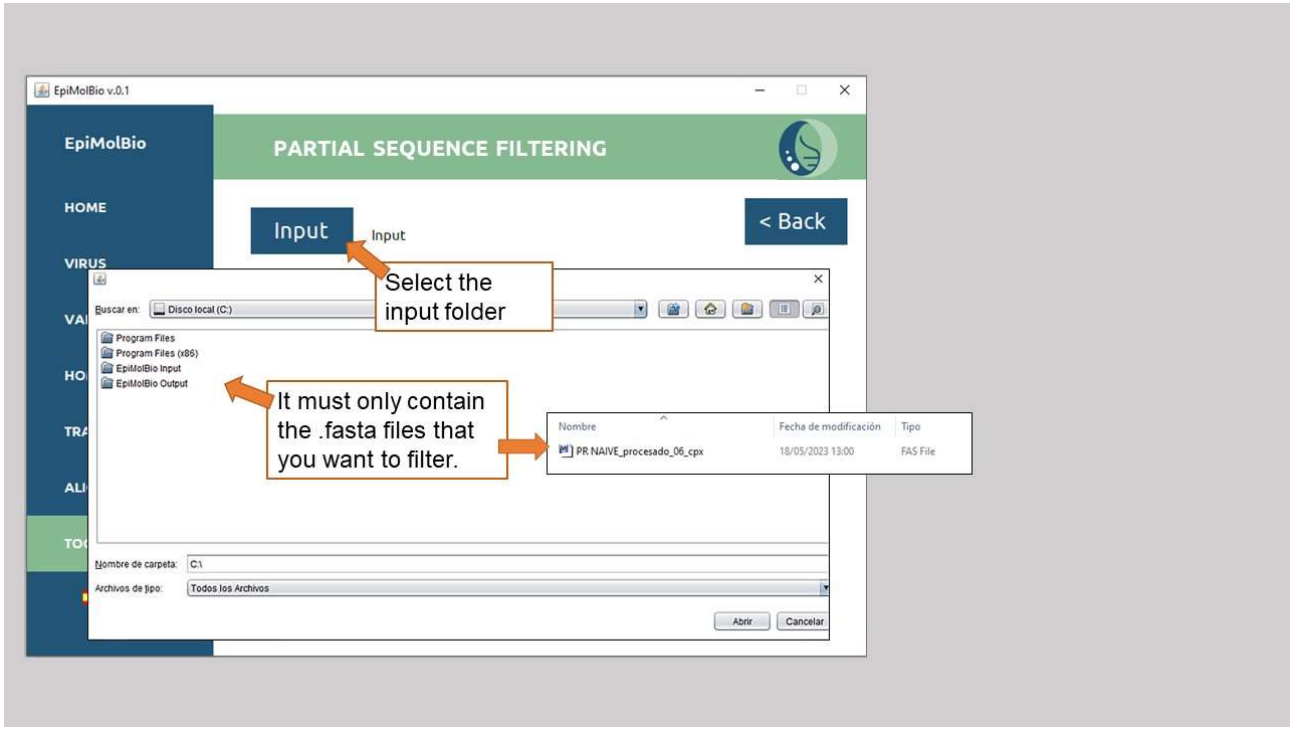
## 1)



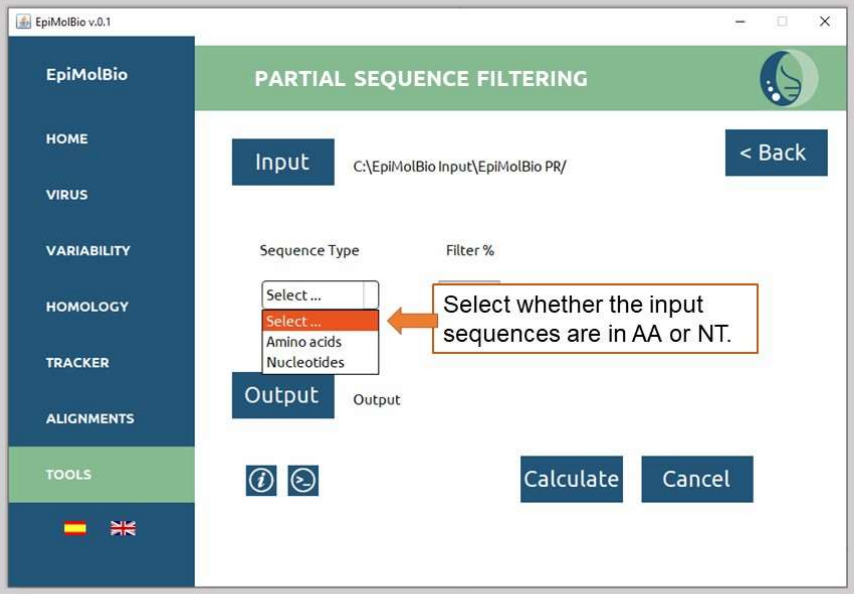

## 2)



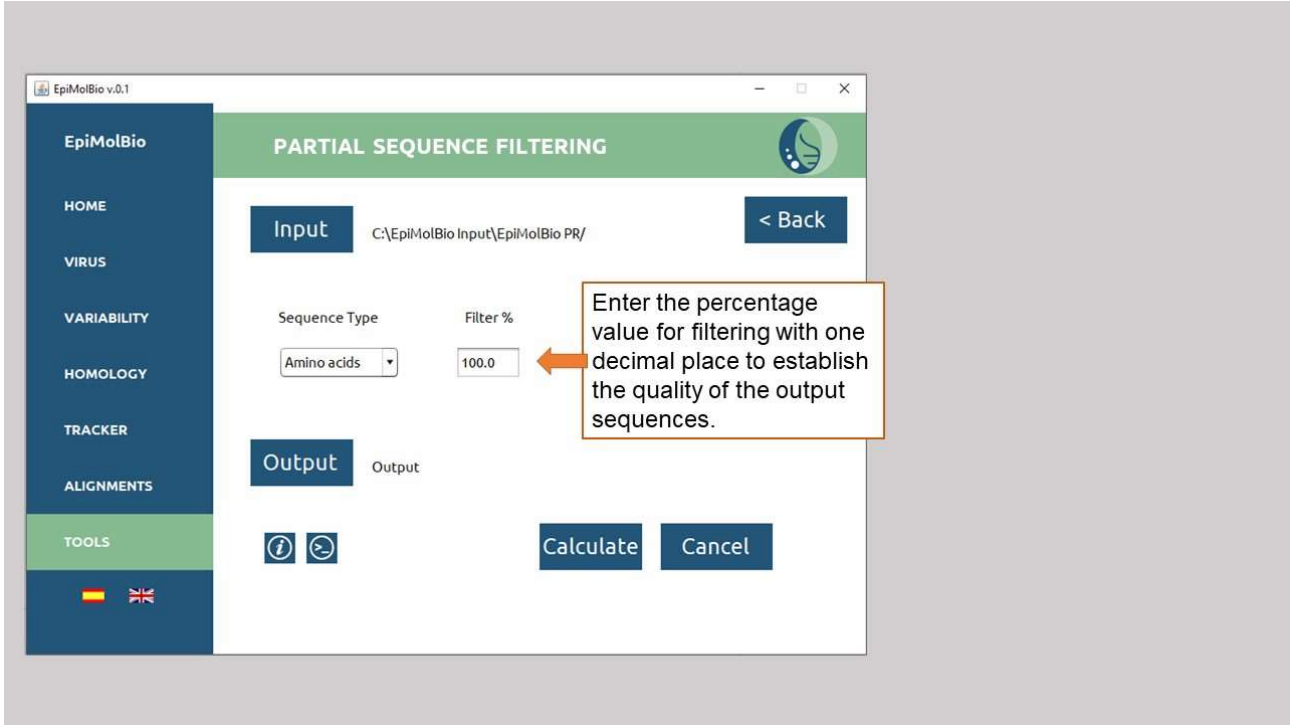
## 3)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
PARTIAL SEQUENCE FILTERING
Input
Input
< Back
Select the input folder
Buscar en:
Disco local (C:)
Program Files
Program Files (x86)
EpiMolBio Input
EpiMolBio Output
It must only contain the .fasta files that you want to filter.
Nombre
Fecha de modificación
Tipo
PR NAIVE_procesado_06_cpx
18/05/2023 13:00
FAS File
Nombre de carpeta:
C:\
Archivos de tipo:
Todos los Archivos
Abrir
Cancelar

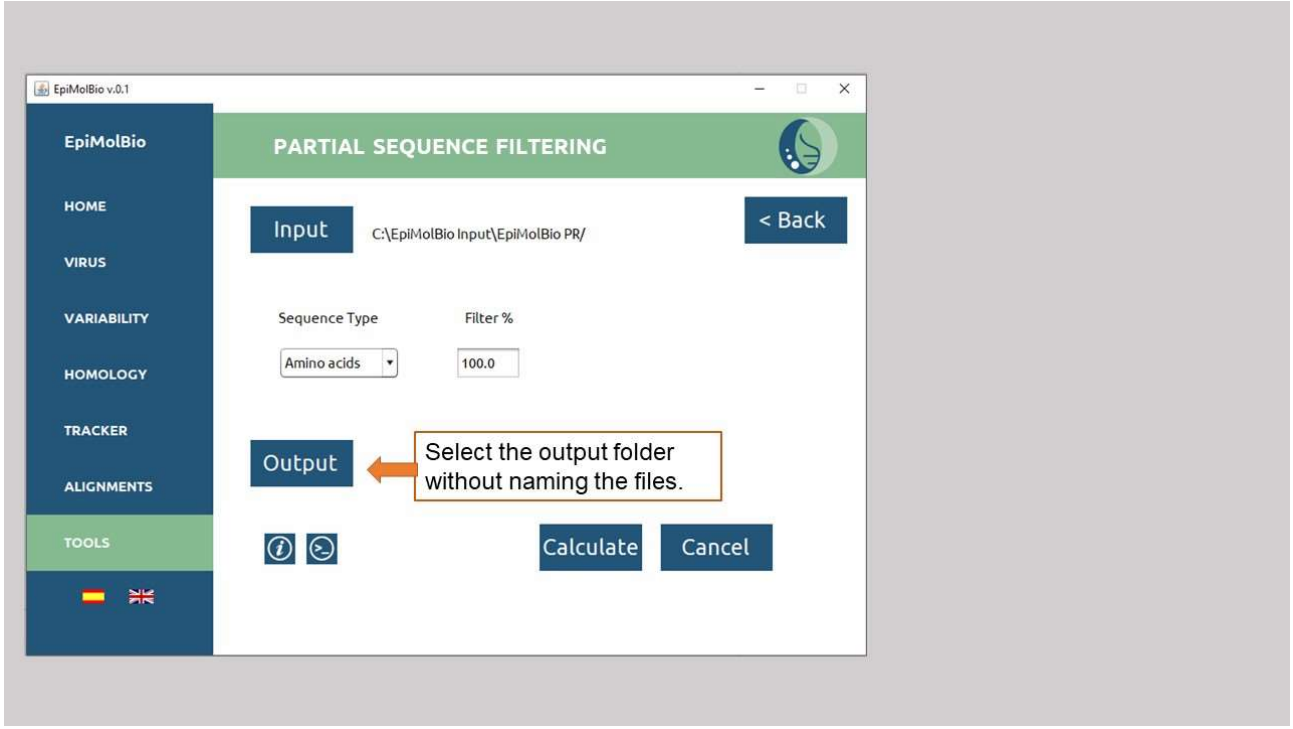
## 4)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
PARTIAL SEQUENCE FILTERING
Input
C:\EpiMolBio Input\EpiMolBio PR/
< Back
Sequence Type
Filter %
Select ...
Select ...
Amino acids
Nucleotides
Select whether the input sequences are in AA or NT.
Output
Output
Calculate
Cancel

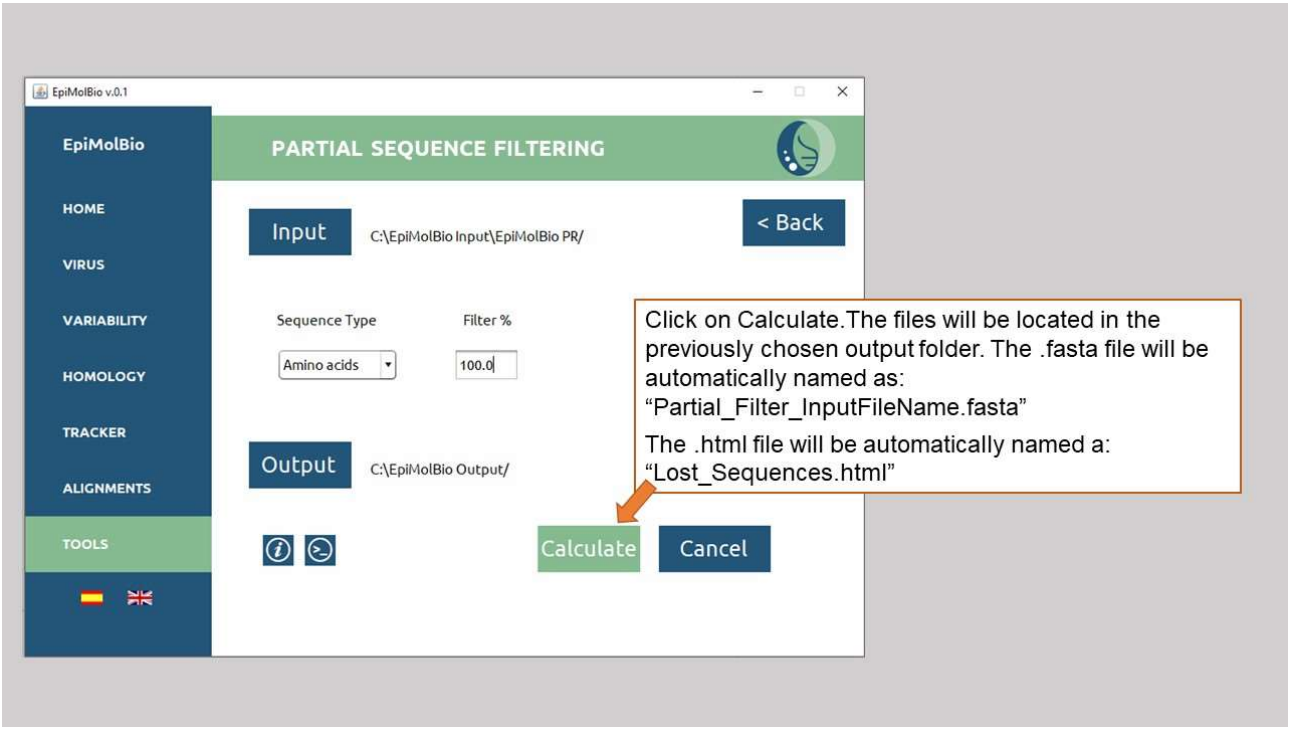
## 5)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
PARTIAL SEQUENCE FILTERING
Input
C:\EpiMolBio Input\EpiMolBio PR/
< Back
Sequence Type
Filter %
Amino acids
100.0
Enter the percentage value for filtering with one decimal place to establish the quality of the output sequences.
Output
Output
Calculate
Cancel

## 6)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
PARTIAL SEQUENCE FILTERING
Input
C:\EpiMolBio Input\EpiMolBio PR/
< Back
Sequence Type
Filter %
Amino acids
100.0
Output
Select the output folder without naming the files.
Calculate
Cancel

# 7)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
PARTIAL SEQUENCE FILTERING
Input
C:\EpiMolBio Input\EpiMolBio PR/
< Back
Sequence Type
Filter %
Amino acids
100.0
Output
C:\EpiMolBio Output/
Calculate
Cancel
Click on Calculate.The files will be located in the previously chosen output folder. The .fasta file will be automatically named as: "Partial_Filter_InputFileName.fasta"
The .html file will be automatically named a: "Lost_Sequences.html"

# VI.3.TRANSLATION

This tool allows you to translate nucleotide sequences from .fasta files into amino acids. Additionally, it offers the option to remove gaps from the sequences.

The **input** file should be a folder containing exclusively the .fasta files in NT you want to translate.

Select the '**Translate**' checkbox to perform the translation.

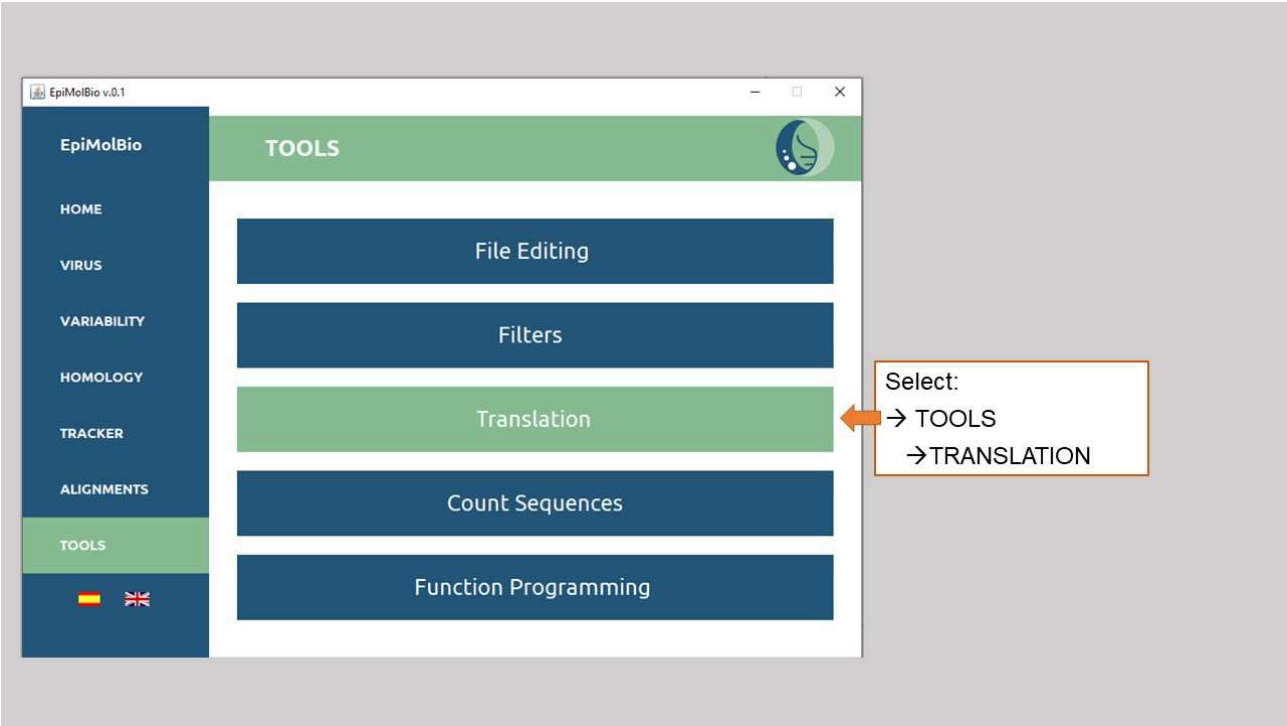
You can choose to check the '**Delete gaps**' checkbox to automatically eliminate all gaps from the sequences.

In the '**Frame**' field, choose the reading frame among frame 1, 2, or 3 to establish the starting nucleotide for counting codons. Generally, Frame 1 is used, if it is assumed that the sequences are in the correct reading frame.

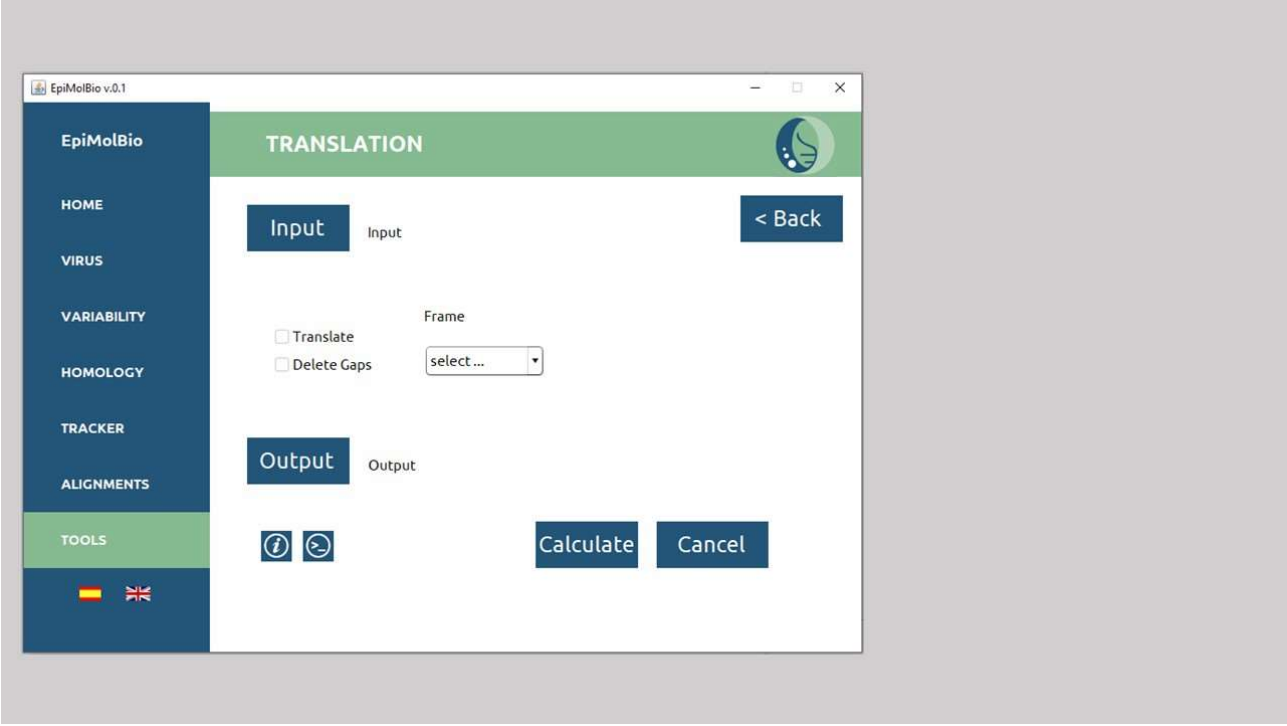
The **output** file will be a .fasta file with the sequences translated into amino acids. In the output, select the destination folder where you want the .fasta file to appear without naming it. This file will be automatically named as follows: 'Translated_InputFileName.fasta' or 'Translated_No_Gaps_InputFileName.fasta' if the option to remove gaps has been selected.

# Step-by-step:

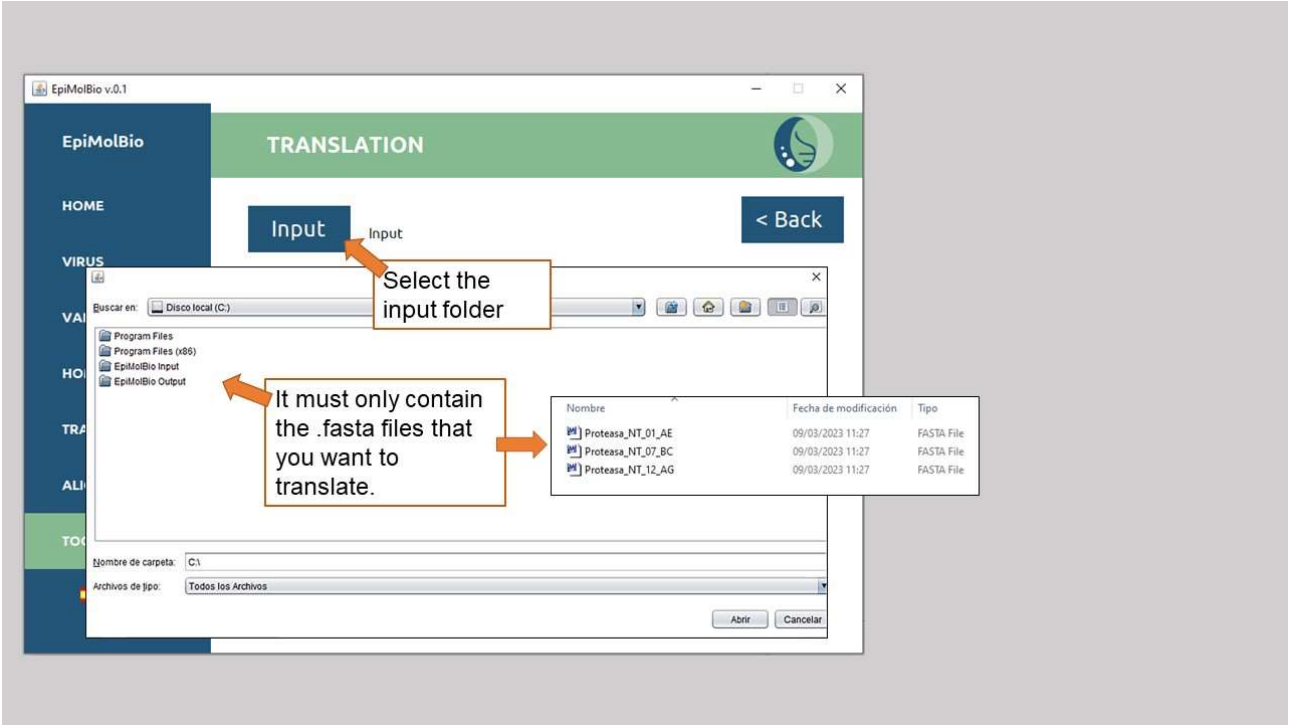
## 1)



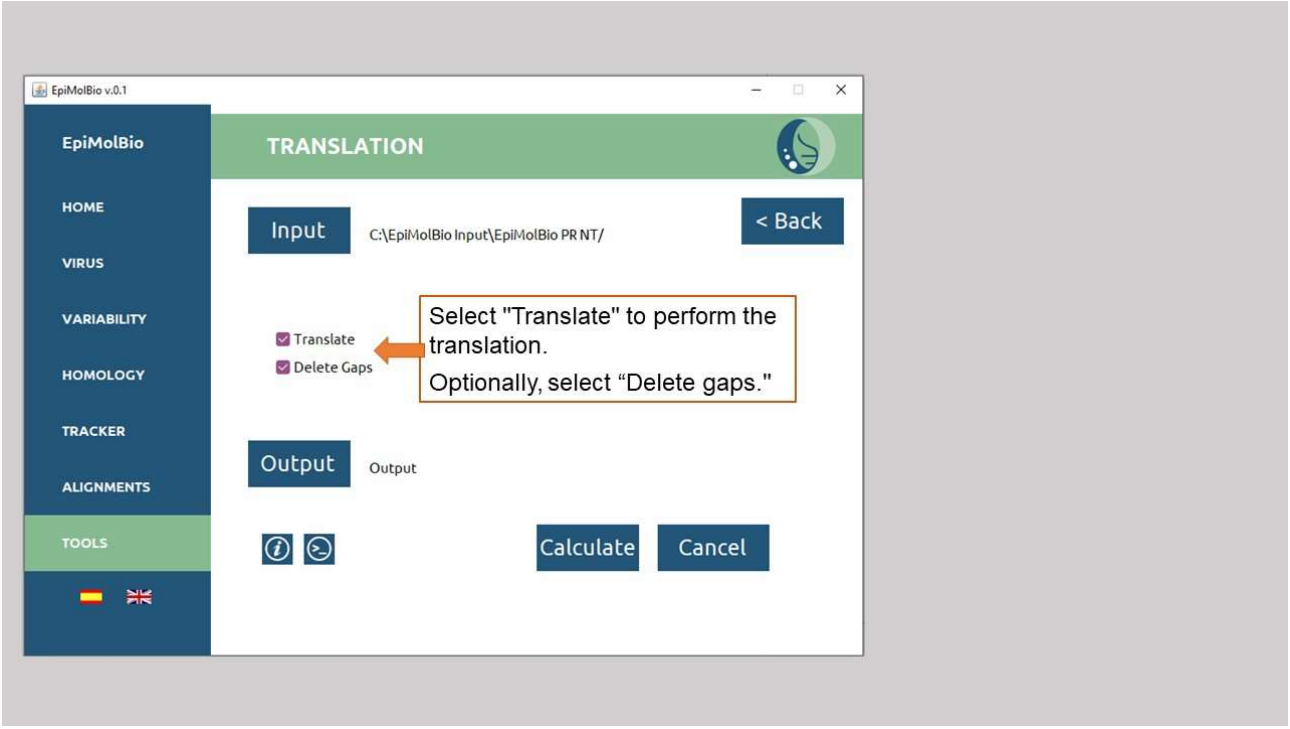

## 2)



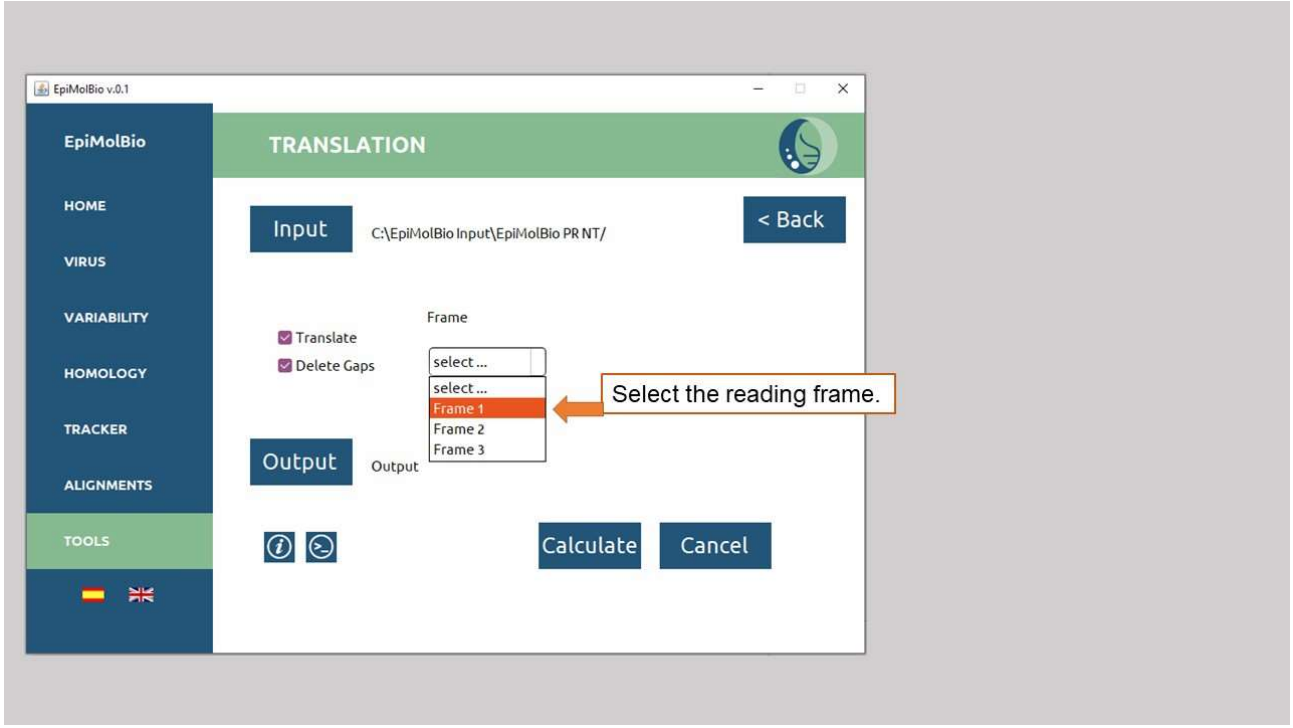
## 3)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
TRANSLATION
Input
Input
< Back
Select the input folder
Buscar en:
Disco local (C:)
Program Files
Program Files (x86)
EpiMolBio Input
EpiMolBio Output
It must only contain the .fasta files that you want to translate.
Nombre
Fecha de modificación
Tipo
Proteasa_NT_01_AE
Proteasa_NT_07_BC
Proteasa_NT_12_AG
09/03/2023 11:27
FASTA File
Nombre de carpeta:
C:\
Archivos de tipo:
Todos los Archivos
Abrir
Cancelar

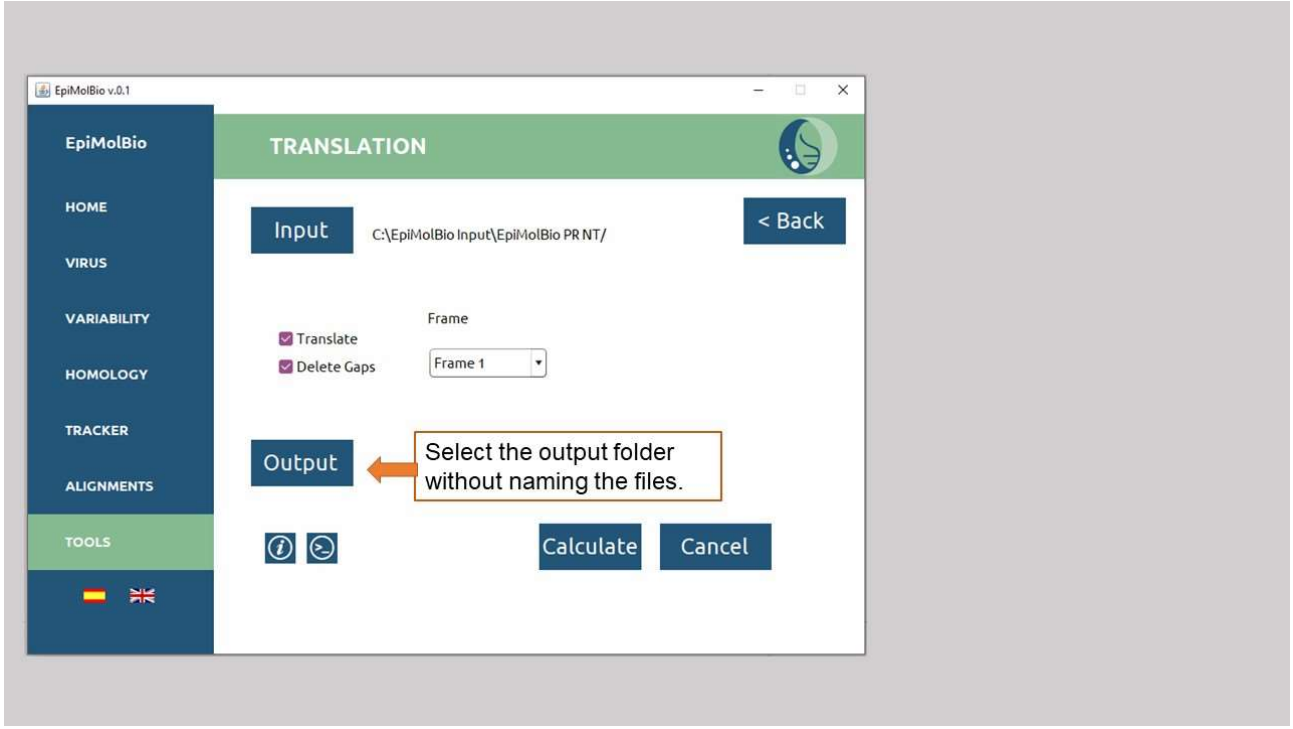
## 4)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
TRANSLATION
Input
C:\EpiMolBio Input\EpiMolBio PR NT/
< Back
Translate
Delete Gaps
Select "Translate" to perform the translation.
Optionally, select “Delete gaps."
Output
Output
Calculate
Cancel

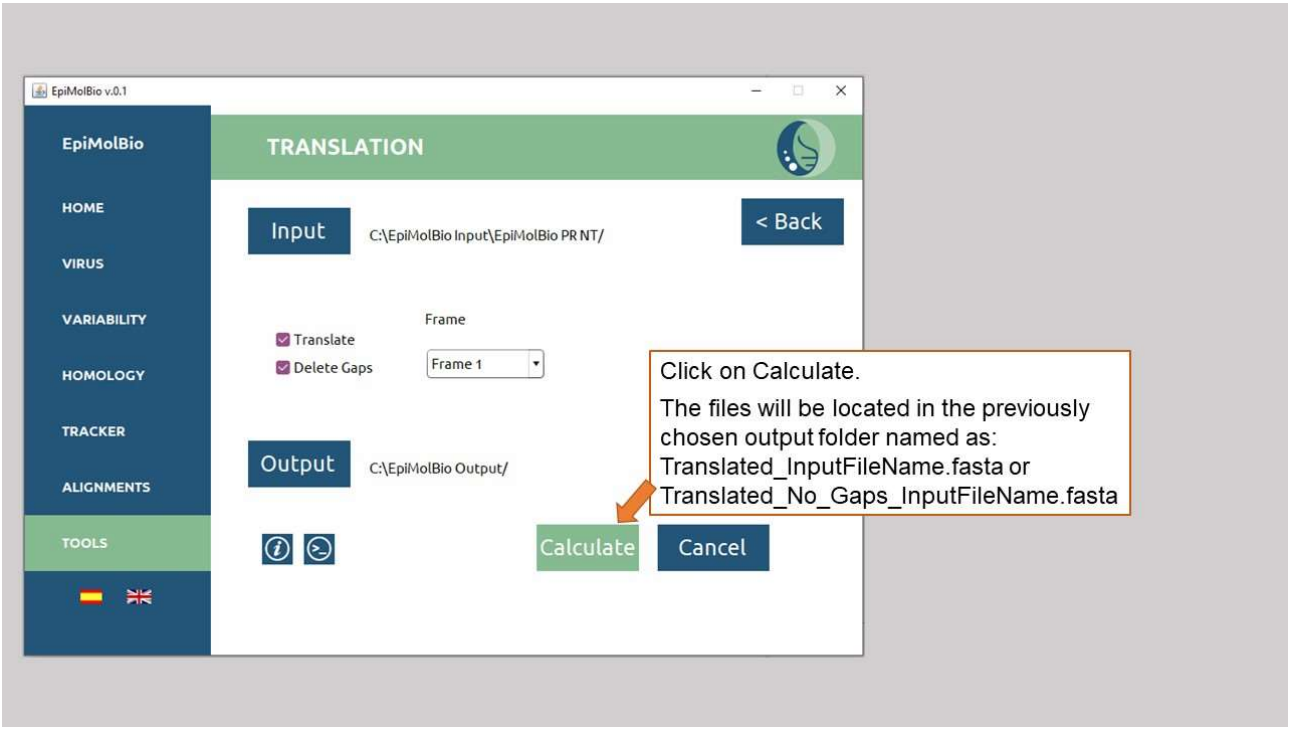
## 5)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
TRANSLATION
Input
C:\EpiMolBio Input\EpiMolBio PR NT/
< Back
Translate
Delete Gaps
Frame
select ...
select ...
Frame 1
Frame 2
Frame 3
Select the reading frame.
Output
Output
Calculate
Cancel

## 6)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
TRANSLATION
Input
C:\EpiMolBio Input\EpiMolBio PR NT/
< Back
Translate
Delete Gaps
Frame
Frame 1
Output
Select the output folder
without naming the files.
Calculate
Cancel

# 7)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
TRANSLATION
Input
C:\EpiMolBio Input\EpiMolBio PR NT/
< Back
Frame
Translate
Delete Gaps
Frame 1
Output
C:\EpiMolBio Output/
Calculate
Cancel
Click on Calculate.
The files will be located in the previously chosen output folder named as:
Translated_InputFileName.fasta or
Translated_No_Gaps_InputFileName.fasta

# VI.4. COUNT SEQUENCES

This tool is used to **count the total number of sequences in one or multiple .fasta files**, or to **determine how many of these sequences contain mutations** compared to a reference sequence.

The **input** file should be the folder containing exclusively the .fasta files that you want to count.

In the '**Format**' field, you can choose between the two functions. The option '**Table**,' generates a .csv table counting all the sequences in the input file. The option '**Mutated Sequences**,' generates a .csv table counting only the sequences that have mutations compared to the introduced reference sequence. If you choose 'Table,' the next step will be to set the output folder. If you choose 'Mutated Sequences,' you will need to fill out the following fields.

In the '**Reference**' field, enter the reference sequence without spaces or line breaks.

Choose between the boxes labeled '**AA**' and '**NT**' depending on whether the sequences in the input file are in nucleotides (NT) or amino acids (AA).

For the **output**, you'll need to select the output folder where you want the .csv file to appear and name the file by adding .csv at the end. The .csv output formats can be opened with Excel.

Example of the **Table** output format from the Count Sequences tool:

| | A | B |
|---|---|---|
| 1 | File | Number of Sequences |
| 2 | PR_01_AE.fasta | 26849 |
| 3 | PR_02_AG.fasta | 9577 |
| 4 | PR_03_A6B.fasta | 310 |
| 5 | PR_04_cpx.fasta | 15 |
| 6 | PR_05_DF.fasta | 24 |
| 7 | PR_06_cpx.fasta | 746 |
| 8 | PR_07_BC.fasta | 10916 |
| 9 | PR_08_BC.fasta | 2348 |
| 10 | PR_09_cpx.fasta | 94 |
| 11 | PR_100_01C.fasta | 5 |
| 12 | PR_101_01B.fasta | 4 |

The **Table** output format consists of a .csv table. In the table, the first column displays the names of the input files, and the second column shows the total number of sequences per file. At the end of the table, the total number of sequences across all files is indicated.

Example of the **Mutated Sequences** output format from the Sequence Counting tool:

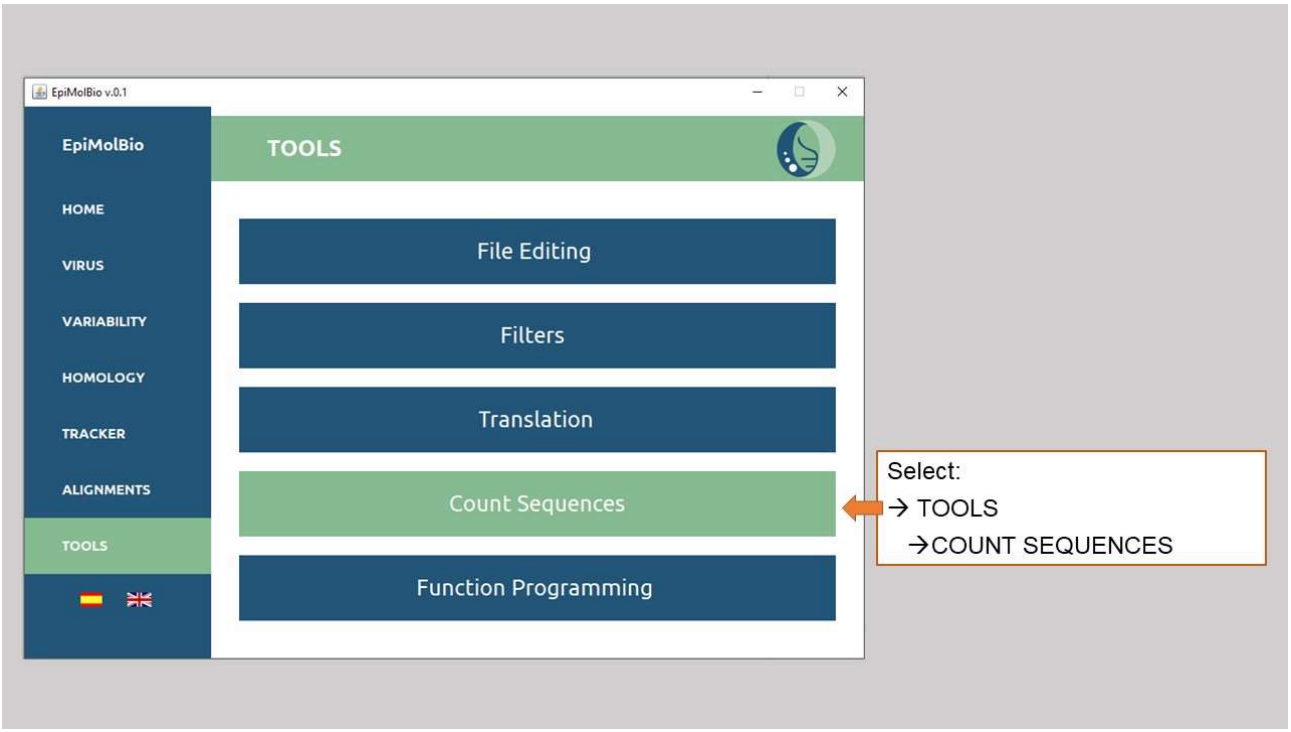
| | A | B | C | D |
|---|---|---|---|---|
| 1 | File | Mutated | Number of Sequences | Percentage |
| 2 | PR_01_AE.fasta | 26849 | 26849 | 100.00% |
| 3 | PR_02_AG.fasta | 9577 | 9577 | 100.00% |
| 4 | PR_03_A6B.fasta | 310 | 310 | 100.00% |
| 5 | PR_04_cpx.fasta | 15 | 15 | 100.00% |
| 6 | PR_05_DF.fasta | 24 | 24 | 100.00% |
| 7 | PR_06_cpx.fasta | 746 | 746 | 100.00% |
| 8 | PR_07_BC.fasta | 10916 | 10916 | 100.00% |
| 9 | PR_08_BC.fasta | 2348 | 2348 | 100.00% |
| 10 | PR_09_cpx.fasta | 94 | 94 | 100.00% |
| 11 | PR_100_01C.fasta | 5 | 5 | 100.00% |
| 12 | PR_101_01B.fasta | 4 | 4 | 100.00% |

The **Mutated Sequences** output format consists of a .csv table. In the table, the first column displays the names of the input files; the second column shows the total number of sequences that have mutations compared to the reference sequence; the third column displays the total number of sequences per input file; and the fourth column shows the frequency of the mutated sequences.

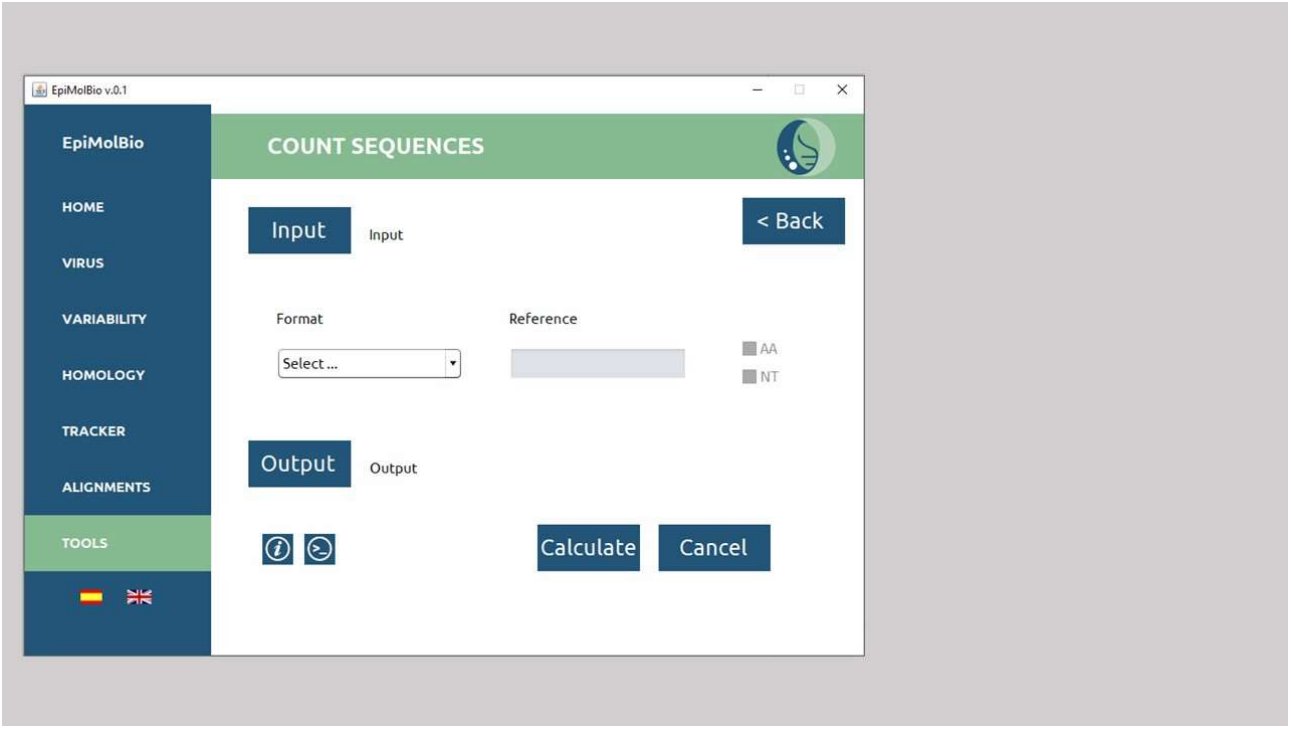
In this example, as sequences of an RNA virus (HIV) with high genetic variability and mutation frequency, all sequences contain at least 1 mutation compared to the reference sequence (HXB2 isolate of HIV).

# Step-by-step:

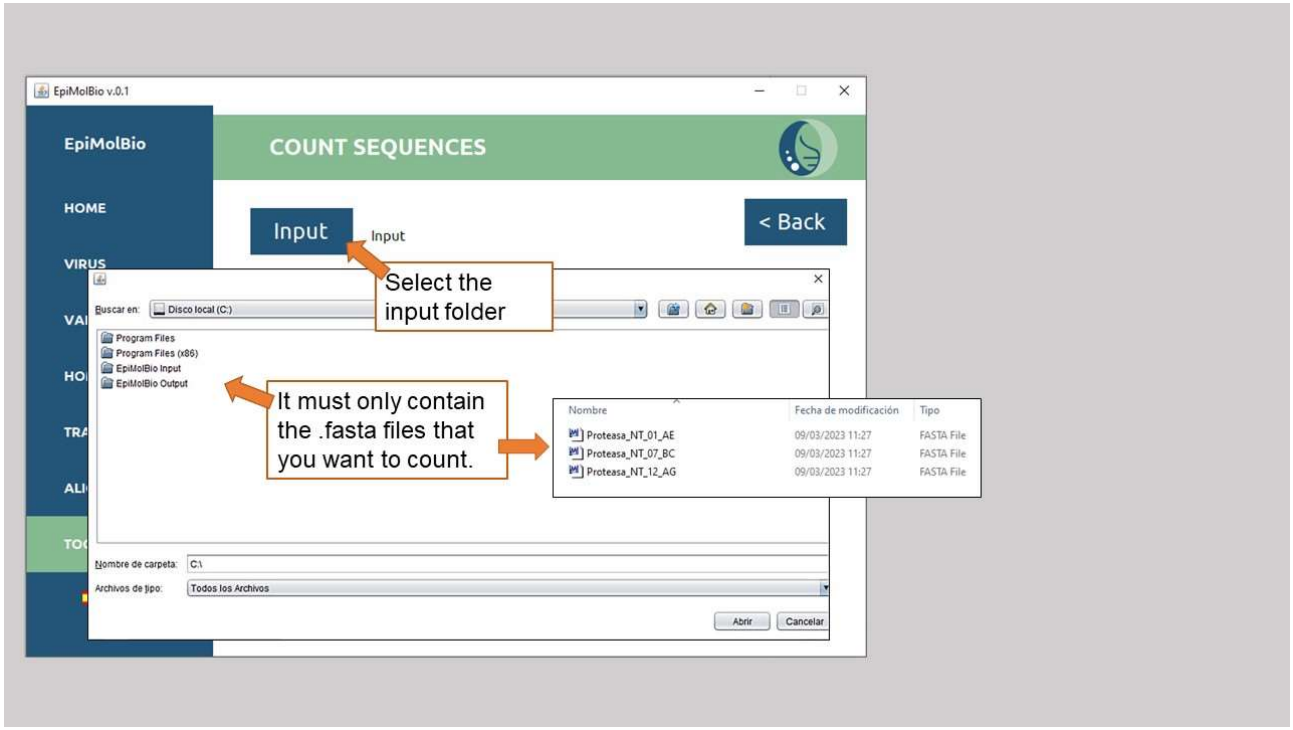
## 1)



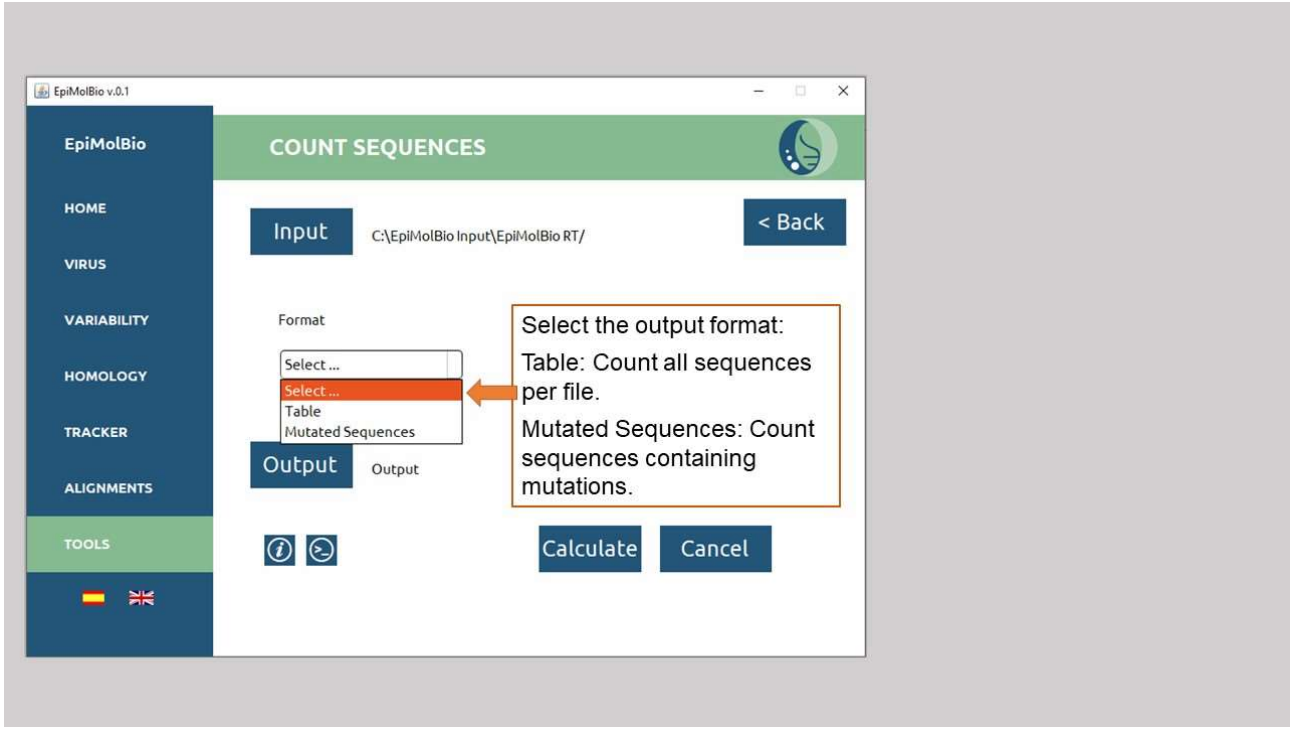

## 2)



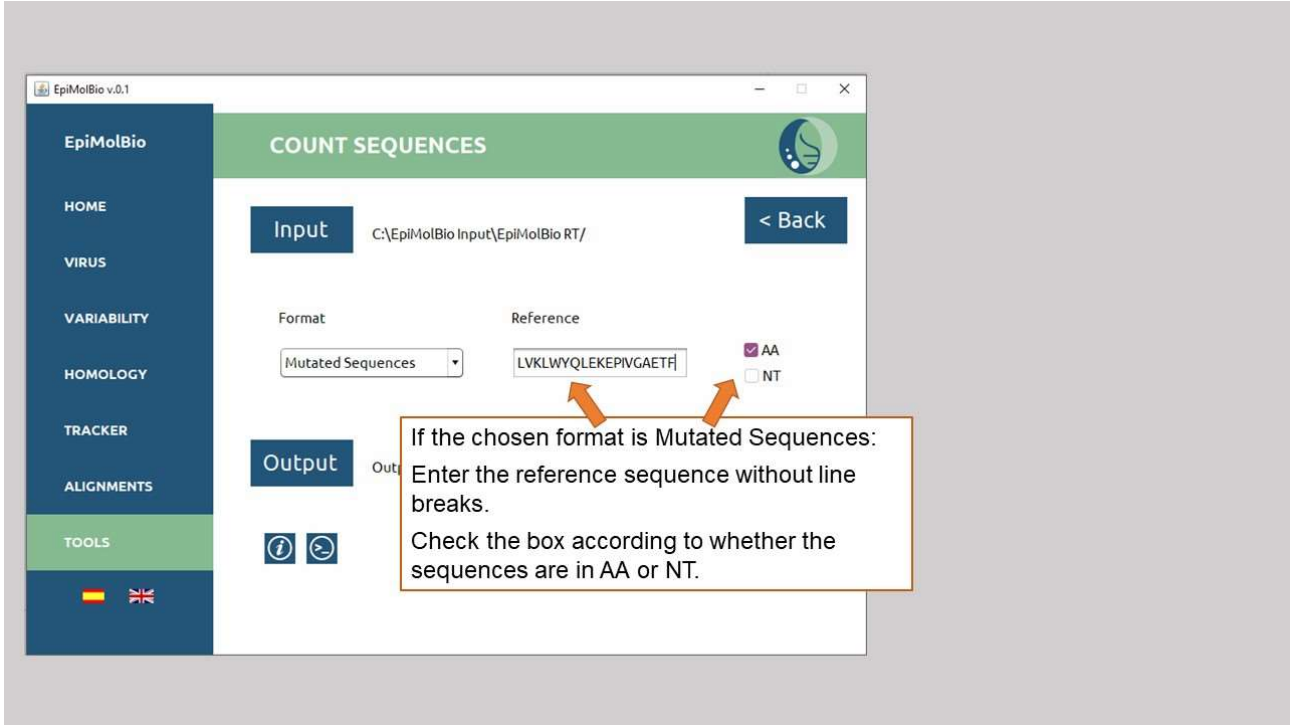
## 3)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
COUNT SEQUENCES
Input
Input
< Back
Select the input folder
Buscar en:
Disco local (C:)
Program Files
Program Files (x86)
EpiMolBio Input
EpiMolBio Output
It must only contain the .fasta files that you want to count.
Nombre
Fecha de modificación
Tipo
Proteasa_NT_01_AE
Proteasa_NT_07_BC
Proteasa_NT_12_AG
09/03/2023 11:27
FASTA File
Nombre de carpeta:
C:\
Archivos de tipo:
Todos los Archivos
Abrir
Cancelar

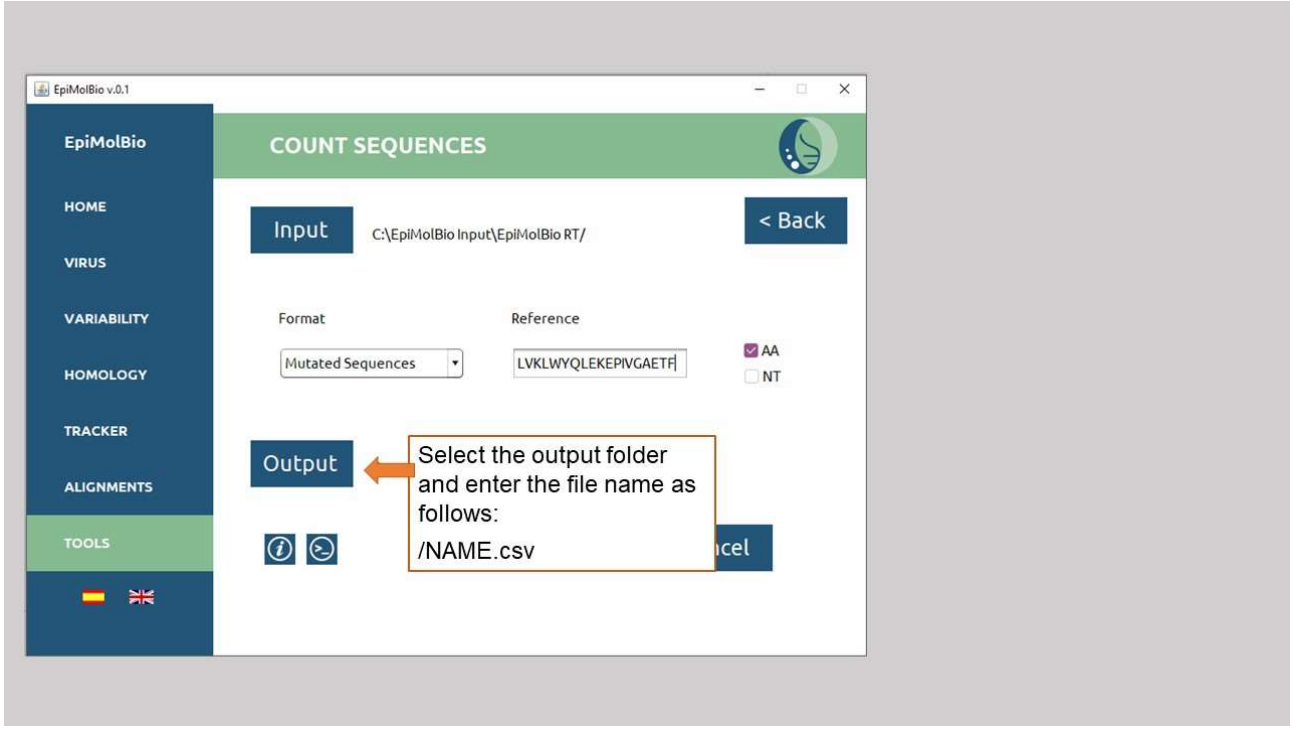
## 4)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
COUNT SEQUENCES
Input
C:\EpiMolBio Input\EpiMolBio RT/
< Back
Format
Select ...
Select ...
Table
Mutated Sequences
Output
Output
Select the output format:
Table: Count all sequences per file.
Mutated Sequences: Count sequences containing mutations.
Calculate
Cancel

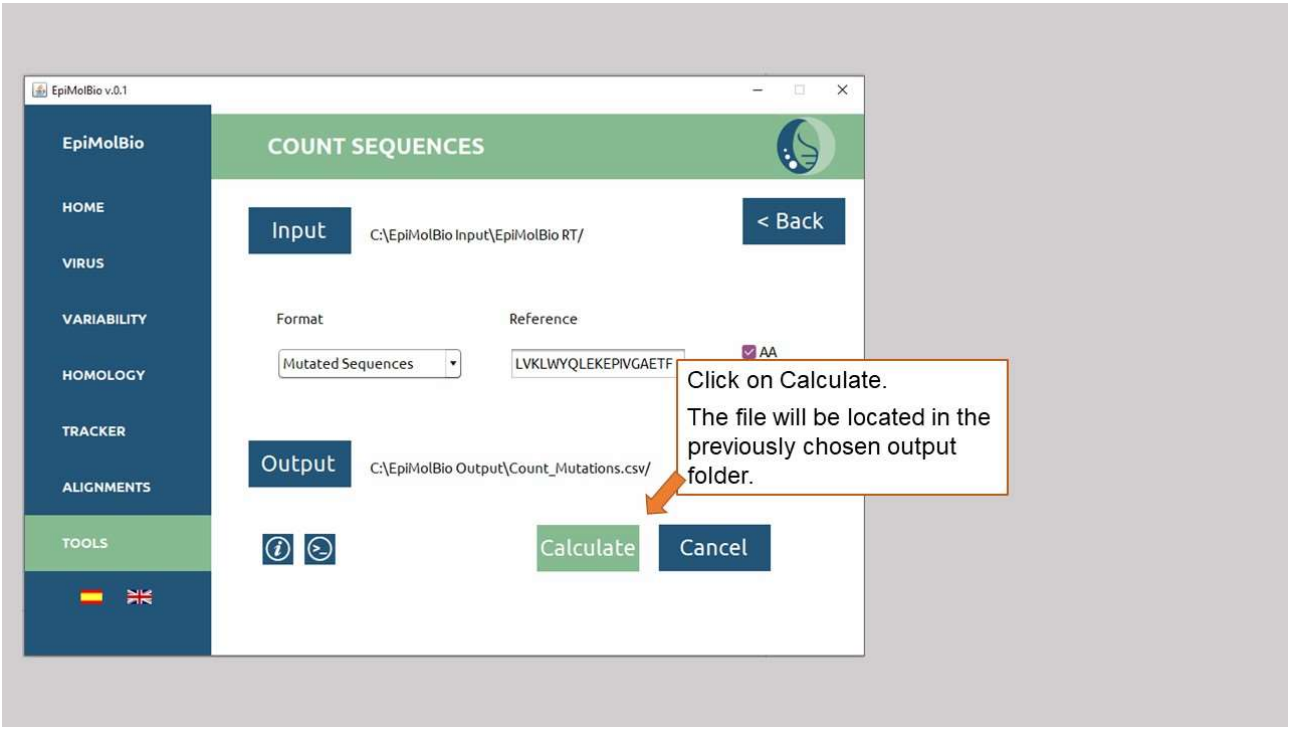
## 5)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
COUNT SEQUENCES
Input
C:\EpiMolBio Input\EpiMolBio RT/
< Back
Format
Reference
Mutated Sequences
LVKLWYQLEKEPIVGAETF
AA
NT
Output
If the chosen format is Mutated Sequences:
Enter the reference sequence without line breaks.
Check the box according to whether the sequences are in AA or NT.

## 6)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
COUNT SEQUENCES
Input
C:\EpiMolBio Input\EpiMolBio RT/
< Back
Format
Reference
Mutated Sequences
LVKLWYQLEKEPIVGAETF
AA
NT
Output
Select the output folder and enter the file name as follows:
/NAME.csv

# 7)


EpiMolBio v.0.1
EpiMolBio
HOME
VIRUS
VARIABILITY
HOMOLOGY
TRACKER
ALIGNMENTS
TOOLS
COUNT SEQUENCES
Input
C:\EpiMolBio Input\EpiMolBio RT/
< Back
Format
Reference
Mutated Sequences
LVKLWYQLEKEPIVGAETF
AA
Output
C:\EpiMolBio Output\Count_Mutations.csv/
Calculate
Cancel
Click on Calculate.
The file will be located in the previously chosen output folder.

# VI.5 FUNCTION PROGRAMMING

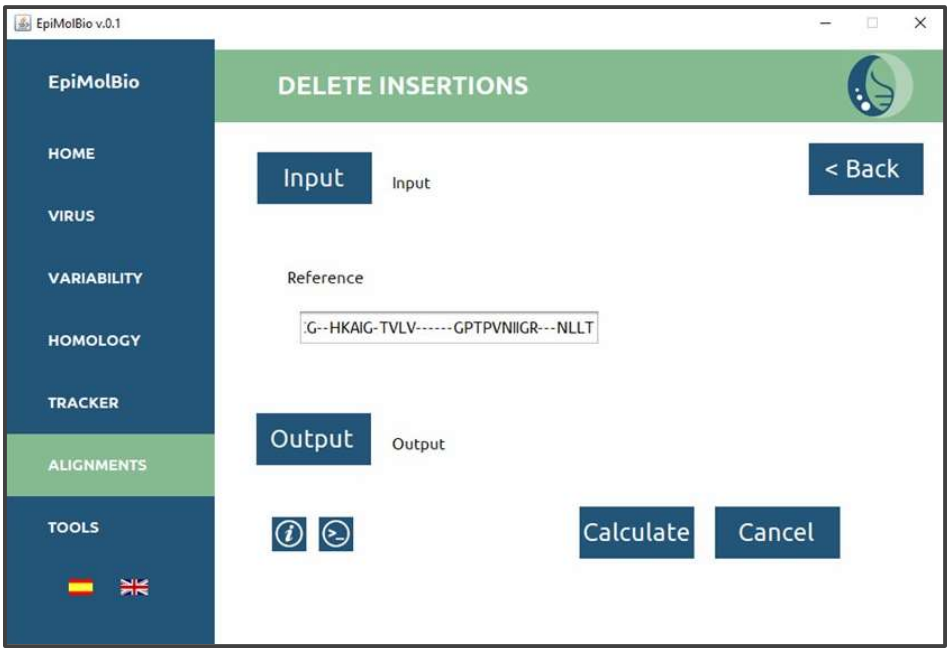
This tool **enables the automation of functions within the EpiMolBio program by chaining them together to be executed sequentially without manual intervention**. It is recommended for cases where the same process needs to be repeated or for executing functions that are time-consuming or involve multiple steps, such as processing a large volume of sequences.

To carry out this task, it is necessary to have a **.txt** file containing the **instructions** that will be automatically executed by this tool.

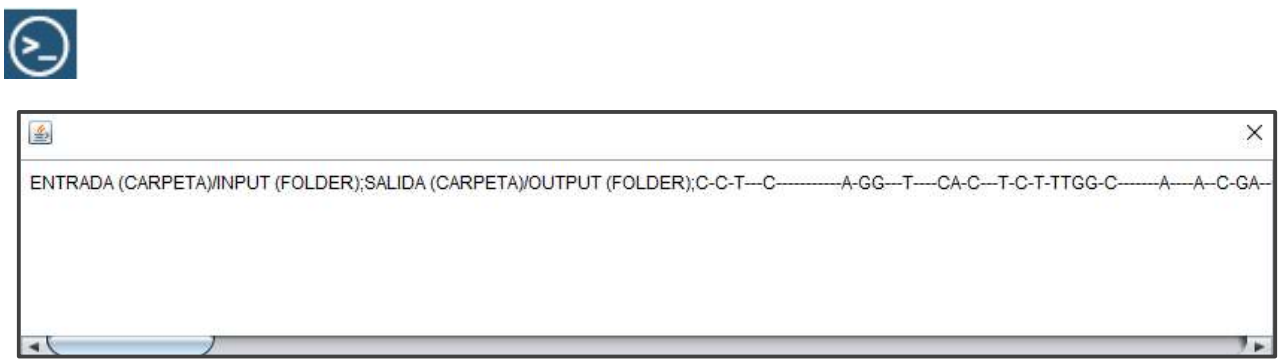
In the following example, we will automate the 'Delete Insertions' function (within Alignments) and the 'Translation' function (in Tools) to obtain translated sequences without insertions as output.

To create the first **text file**, select the first function that you want to automate without including input or output.

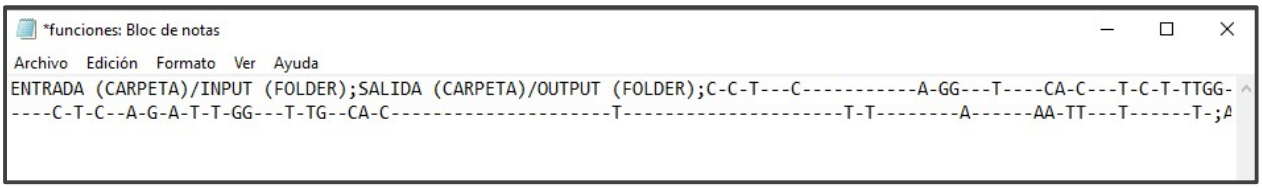
In our example, in the 'Delete Insertions' function, fill in the parameters of the function without including input or output. In this case, you only need to enter the reference sequence with gaps, without line breaks or spaces in AA or NT according to the input files.



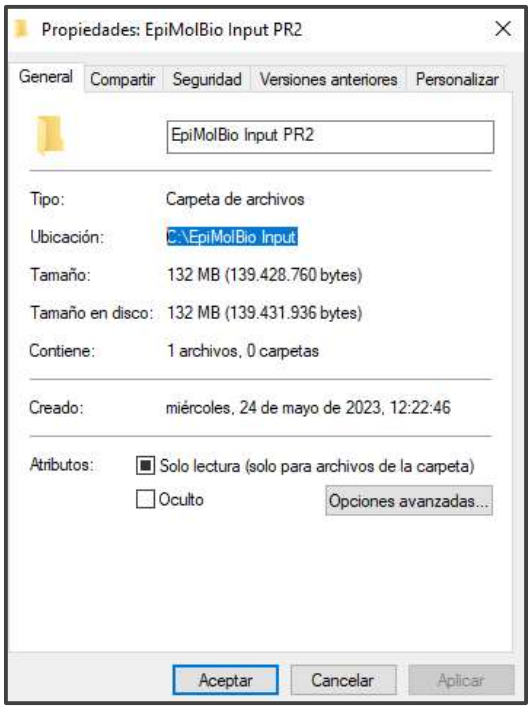

Once this is done, click on the Function Programming button. The following window will open:



Copy the content of the button and paste it into a text file as follows:




Replace **ENTRADA (CARPETA)/INPUT (FOLDER)** with the path of the folder containing the files to be analyzed. This can be obtained in Windows by right-clicking on the folder, then clicking on ‘Properties.’




You need to copy both the text written behind ‘Location’ and the name of the folder appearing at the top of the box, followed by ‘/’.

In this example, it would be as follows: **C:\EpiMolBio Input/EpiMolBio Input PR2/;OUTPUT...**

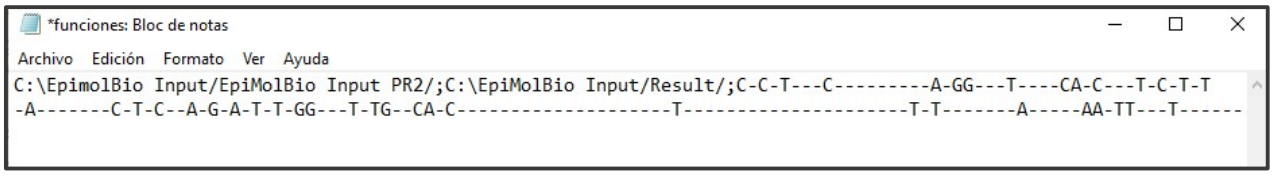
It's important to keep the ‘;’ without deleting it from the text file.

In Linux, you can directly copy the file and paste it into the text file, and it will paste the file path.

Next, you need to modify the output in the text file in the same way as the input, replacing **SALIDA (CARPETA)/OUTPUT (FOLDER)** with the path of the folder where you want to save the result of the first function. It's not necessary to create the folder beforehand; if you write the path directly, the program will create the folder automatically.

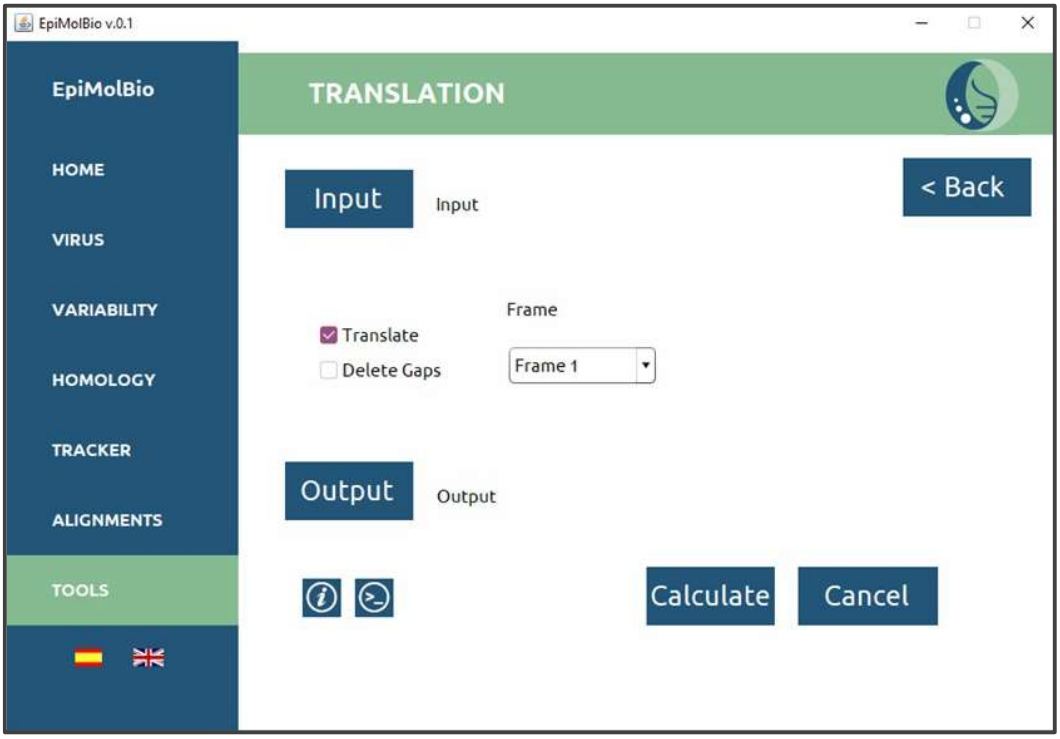
Creating a "Result" folder would look like this: **C:\EpiMolBio Input/Result/;**

Example of the text file with the input and output modified for the ‘Delete Insertions’ function:



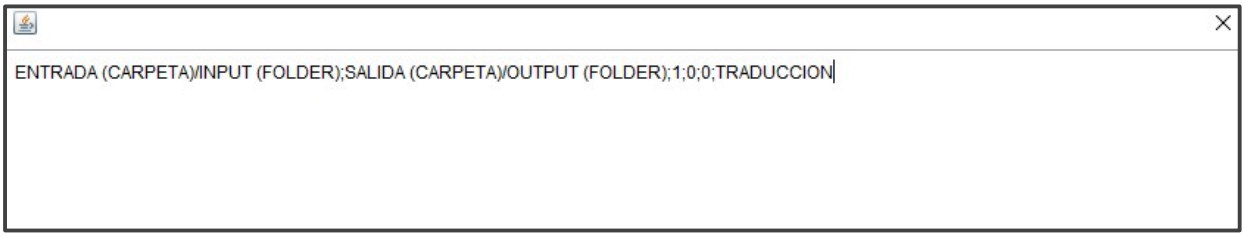

Once the first function has been automated, we proceed to automate a second one. To do this, the following instruction should start on a new line of text. Repeat the previous steps, replacing the input and output as explained earlier, always maintaining the ";" delimiter.

In our example, after removing the insertions, we proceed to translate:

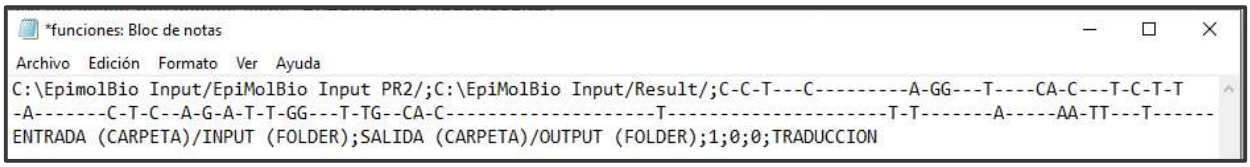



The information is completed in the function interface without entering the input or output. In this example, we click on the ‘Translate’ box and choose Frame 1 as the reading frame.

Next, click the **Function Programming** button to copy its content.

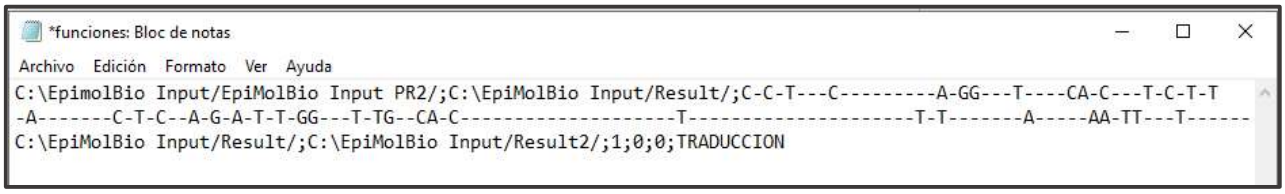


Paste the information on the next line of text in the previous text file:




To chain functions, as in this example, replace **ENTRADA (CARPETA)/INPUT (FOLDER)** with the output from the previous function (C:\EpiMolBio Input/Result/;).

Replace the output with another folder. In this example, we create another folder named ‘Result2’: **C:\EpiMolBio Input/Result2/;**




You can chain as many functions as needed by following this process: using the output of the previous process as input.

Once the text file creation is complete, save the file.

In the **Tools** function, choose **Function Programming**:

For the **input**, load the folder where the instructions file (the saved text file) is located.

In the **Function Programming** field, click the checkbox to activate the function.

Click **Calculate**.

In this example, we generate a fasta file without insertions (function 1) and translated (function 2) , which will be located in the ‘Result2’ folder.

It's not always necessary to chain functions; you can also automate independent functions. For example, you can automate the outputs of various HTML files. In this case, the input will not be the output from the previous line, but another separate input.

# EpiMolBio

## Analysis of Genetic Variability

License: Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 (registry number 2305114294344)

Developer: Roberto Reinosa

Collaborator: Paloma Troyano-Hernáez

Coordination and Supervision: África Holguín

Copyright: Roberto Reinosa Fernández and Biomedical Research Foundation of Ramón y Cajal University Hospital (FIBioHRC)