Protocol 3\_Using SampleMiner.

1. Launch Anaconda Navigator, click the “Environments” tag at the left panel of the screen, then open a terminal with the “SampleMiner” environment by clicking the “SampleMiner” tag at the “Environments” window.
2. Run SampleMiner in the terminal screen loaded with the “SampleMiner” environment by entering the entire path to SampleMiner at the command prompt and pressing the Enter key. For example, enter “Macintosh HD/Users/Current user/Applications/SampleMiner” on MacOS, or “C:\Program Files\SampleMiner” on Windows.
3. The above two steps will display SampleMiner’s main menu on the terminal screen topped by the program version and current working directory information.
4. Choose one of the several options offered by the “MAIN MENU” by inputting the option number at the “Please, enter your choice from above list:” prompt displayed just below the main menu, then pressing the enter button at the computer keyboard. This will execute one of the options currently available in SampleMiner v1.0 below:
   * + (0) Set working directory
     + (1) Import Patric tsv file
     + (2) Browse Patric with keywords
     + (3) Browse Patric with field specific keywords
     + (4) Fetch Patric isolates’ metadata with keywords
     + (5) Fetch Patric isolates’ metadata with field specific keywords
     + (6) Exclude / Reinclude Patric isolates’ from sample
     + (7) Describe Patrick sample with tables
     + (8) Describe Patrick sample with graphics
     + (9) Import Enterobase tsv file
     + (10) Browse Enterobase with keywords
     + (11) Browse Enterobase with field specific keywords
     + (12) Fetch Enterobase isolates’ metadata with keywords
     + (13) Fetch Enterobase isolates’ metadata with field specific keywords
     + (14) Exclude / Reinclude Enterobase isolates’ from sample
     + (15) Describe Enterobase sample with tables
     + (16) Describe Enterobase sample with graphics
     + (40) Fetch sample isolates’ DNA sequences from NCBI
     + (50) Quit
5. Setting a working directory to store data resulting from a sampling process employing SampleMiner.
   1. Input option “0” at the “Please, enter your choice from above list:” prompt to set a custom working directory. If a custom working directory is not set, the program will use the “default” directory instead. These directories are located within the “mined samples” subfolder of the “SampleMiner” folder, which were written to the local user’s “Documents” folder during program installation.
   2. Entering option “0” will allow the user to create a new working directory as well as to access that directory during subsequent steps of the sampling process.
6. Importing a dataset to be used for sampling with SampleMiner.
   1. Currently, SampleMiner v1.0 is prepared to explore Enterobase and Patric derived datasets.
   2. Importing Patric datasets:
      1. Access Patric at “<https://www.bv-brc.org/>”.
      2. Perform a general search to provide a Universe dataset from which to carry out the intended sampling process(es). For example, use the general term “Escherichia coli” to recover the largest possible amount of E. coli isolates and their metadata to the dataset, then use it as the Universe in downstream sampling processes intended to include isolates from certain types of E. coli, like “EPEC”, or from certain sources, like “urine”.
      3. Download the dataset as a comma-separated-values file (csv) to the user’s local computer, then convert it to tab-separated-values (tsv) file using a spreadsheet program. Naming the resulting file “Patric\_Metadata.tsv”.
      4. Locally transfer the tsv file to the “databases tsv files” subfolder within the “SampleMiner” folder at the current user’s “Documents” directory.
      5. At SampleMiner’s main menu choose option “(1) Import Patric tsv file”, entering “1” in the “Please, enter your choice from above list:” prompt.
      6. After SampleMiner automatically imports the Patric dataset, it will be ready to be used in the sampling process exploring the imported data.
      7. The imported dataset will consist of a local list of Python dictionaries, each dictionary describing an isolate present in the csv/tsv file downloaded from Patric.
   3. Importing Enterobase datasets:
      1. Access Enterobase at “<https://enterobase.warwick.ac.uk/>”.
      2. Perform a general search to provide a Universe dataset from which to carry out the intended sampling process(es). For example, use the general term “Escherichia coli” to recover the largest possible amount of E. coli isolates and their metadata to the dataset, then use it as the Universe in downstream sampling processes intended to include isolates from certain types of E. coli, like “EPEC”, or from certain sources, like “urine”.
      3. Download the four resulting datasets below as a comma-separated-values files (csv) to the user’s local computer, then convert them to tab-separated-values (tsv) files using a spreadsheet program. Name them “Enterobase\_Metadata\_1.tsv”, “Enterobase\_Metadata \_2.tsv”, “Enterobase\_Metadata \_3.tsv”, and “Enterobase\_Metadata \_4.tsv”. The four datasets to be downloaded consist of the Enterobase metadata pages:
         * + Assembly stats;
           + Achtman 7 Gene MLST;
           + Phylotypes;
           + Serotype Prediction.
      4. To avoid field conflicts involving the column header ‘ST’ in ‘Phylotypes’ tsv file during data searching with SampleMiner, the column headers ‘ST’ and ‘ST Complex’ in ‘Achtman 7 Gene MLST’ tsv file must be edited to read ‘MLST ST’ and ‘MLST ST Complex’.
      5. Locally transfer the tsv files to the “databases tsv files” subfolder within the “SampleMiner” folder at the current user’s “Documents” directory.
      6. At SampleMiner’s main menu choose option “(9) Import Enterobase tsv file”, entering “9” in the “Please, enter your choice from above list:” prompt.
      7. After SampleMiner automatically imports the Enterobase datasets, it will be ready to be used in the sampling process exploring the imported data.
      8. The imported datasets will be merged into a single local list of Python dictionaries, each dictionary describing an isolate present in the csv/tsv files downloaded from Enterobase.
7. Browsing Patric and Enterobase with keywords:
   1. Browsing the datasets will not result in retrieval of metadata. So, no files will be generated and stored in subfolders of the current working directory because of the process. Browsing options may be explored by the user to test the adequacy of chosen keywords before actual searches for data retrieval may be performed.
   2. Input option “2” or “10” at the “Please, enter your choice from above list:” prompt to begin a simple keyword search of the Patric or Enterobase datasets, respectively. This kind of keyword search does not limit the scanning process to specific fields of the datasets. So, it consists of a broad search.
   3. A screen informing the version of SampleMiner and the database in use will appear. This screen also presents instructions on how to enter keywords, conclude the entry of a set of keywords, and conclude the entry of all sets of keywords so the dataset search process may start.
   4. To enter the first set of keywords, type the keyword at the prompt “Please, enter keyword 1:” then press enter. Proceed adding the desired keywords in this manner for the current set. After the last keyword of the first set is added, type “-“ to finish the current set and start inputting the next set of keywords, or “--” to finish entering all sets and start scanning the dataset.
   5. Note that for each set of keywords, only dataset entries matching all keywords will be detected during browsing. On the other hand, the browsing process will compute the sum of unique dataset entries of all keyword sets. This means that dataset entries detected by more than one set of keywords will be counted only once.
   6. After the dataset scanning finishes, a screen summarizing the search process and its results will be displayed.
   7. Press any key to go back to the “Main Menu” screen.
8. Browsing Patric and Enterobase with field-specific keywords:
   1. Browsing the datasets will not result in retrieval of metadata. So, no files will be generated and stored in subfolders of the current working directory because of the process. Browsing options may be explored by the user to test the adequacy of chosen keywords before actual searches for data retrieval may be performed.
   2. Input option “3” or “11” at the “Please, enter your choice from above list:” prompt to begin a field-specific keyword search of the Patric or Enterobase datasets, respectively. This kind of keyword search limits the scanning process to specific fields of the datasets. So, instead of a broad search, it consists of a more precise search indicated to detect only dataset entries that fulfill the provided keywords at specific metadata fields.
   3. A screen informing the version of SampleMiner and the database in use will appear. This screen also presents the options of fields for the dataset, and instructions on how to enter keywords, conclude the entry of a set of keywords, and conclude the entry of all sets of keywords so the dataset search process may start.
   4. To enter the first field/keyword pair of the first set of keywords start by typing the number of the field to be scanned at the “Please, enter field 1:” prompt and pressing enter, then type the keyword at the prompt “Please, enter field 1's keyword:” and press enter. Proceed adding the desired pairs of fields/keywords in this manner for the current set. After the last pair of field/keyword of the first set is added, type “-“ to finish the current set and start inputting the next set of field/keywords pairs, or “--” to finish entering all sets and start scanning the dataset.
   5. Note that for each set of field/keyword pairs, only dataset entries matching all keywords in the specified fields will be detected during browsing. On the other hand, the browsing process will compute the sum of unique dataset entries of all field/keyword sets. This means that dataset entries detected by more than one set of field/keywords will be counted only once.
   6. After the dataset scanning finishes, a screen summarizing the search process and its results will be displayed.
   7. Press any key to go back to the “Main Menu” screen.
9. Fetching Patric and Enterobase isolates’ (i.e., entries’) metadata with keywords:
   1. These dataset scanning options will result in the retrieval of isolates and metadata, consisting in the inclusion step of a sampling process exploring the dataset. The following files will be generated and stored in subfolders of the current working directory because of the process:
      * + “sampling report.txt” stored in the “samples metadata” folder, summarizing the entire sampling process from keywords used to isolates retrieved and their metadata;
        + “accession numbers.txt” stored in the “samples accession numbers” folder, consisting of the NCBI accession numbers to whole genomes of the retrieved isolates;
        + “isolates metadata.csv” and “…txt”, and “isolates metadata for statistics.csv” and “…txt” files stored in the “samples metadata” folder. These files contain the same contents stored in different file formats (i.e., csv and txt). Only “isolates metadata.csv” will be used by SampleMiner in later steps of the sampling process. The other three files may be explored by the user in downstream computation or data analysis. The txt files consist of lists of Python dictionaries containing the isolates metadata.
   2. Input option “4” or “12” at the “Please, enter your choice from above list:” prompt to begin a simple keyword search of the Patric or Enterobase datasets, respectively. This kind of keyword search does not limit the scanning process to specific fields of the datasets. So, it consists of a broad search.
   3. A screen informing the version of SampleMiner and the database in use will appear. This screen also presents instructions on how to enter keywords, conclude the entry of a set of keywords, and conclude the entry of all sets of keywords so the dataset search process may start.
   4. To enter the first set of keywords, type the keyword at the prompt “Please, enter keyword 1:” then press enter. Proceed adding the desired keywords in this manner for the current set. After the last keyword of the first set is added, type “-“ to finish the current set and start inputting the next set of keywords, or “--” to finish entering all sets and start scanning the dataset.
   5. Note that for each set of keywords, only dataset entries matching all keywords will be detected during the search. On the other hand, the search process will compute the sum of unique dataset entries of all keyword sets. This means that dataset entries detected by more than one set of keywords will be counted only once.
   6. After the dataset scanning finishes, a screen summarizing the search process and its results will be displayed, and the files mentioned in step “9.1” will be stored in their corresponding folders.
   7. Press any key to go back to the “Main Menu” screen.
10. Fetching Patric and Enterobase isolates’ (i.e., entries’) metadata with field-specific keywords:
    1. These dataset scanning options will result in the retrieval of isolates and metadata, consisting in the inclusion step of a sampling process exploring the dataset. The following files will be generated and stored in subfolders of the current working directory because of the process:
       * + “sampling report.txt” stored in the “samples metadata” folder, summarizing the entire sampling process from keywords used to isolates retrieved and their metadata;
         + “accession numbers.txt” stored in the “samples accession numbers” folder and consisting of the NCBI accession numbers to whole genomes of the retrieved isolates;
         + “isolates metadata.csv” and “…txt”, and “isolates metadata for statistics.csv” and “…txt” files stored in the “samples metadata” folder. These files contain the same contents stored in different file formats (i.e., csv and txt). Only “isolates metadata.csv” will be used by SampleMiner in later steps of the sampling process. The other three files may be explored by the user in downstream computation or data analysis. The txt files consist of lists of Python dictionaries containing the isolates metadata.
    2. Input option “5” or “13” at the “Please, enter your choice from above list:” prompt to begin a field-specific keyword search of the Patric or Enterobase datasets, respectively. This kind of keyword search limits the scanning process to specific fields of the datasets. So, instead of a broad search, it consists of a more precise search indicated to detect only dataset entries that fulfill the provided keywords at specific metadata fields.
    3. A screen informing the version of SampleMiner and the database in use will appear. This screen also presents the options of fields for the dataset, and instructions on how to enter keywords, conclude the entry of a set of keywords, and conclude the entry of all sets of keywords so the dataset search process may start.
    4. To enter the first field/keyword pair of the first set of keywords start by typing the number of the field to be scanned at the “Please, enter field 1:” prompt and pressing enter, then type the keyword at the prompt “Please, enter field 1's keyword:” and press enter. Proceed adding the desired pairs of fields/keywords in this manner for the current set. After the last pair of field/keyword of the first set is added, type “-“ to finish the current set and start inputting the next set of field/keywords pairs, or “--” to finish entering all sets and start scanning the dataset.
    5. Note that for each set of field/keyword pairs, only dataset entries matching all keywords in the specified fields will be detected during the search. On the other hand, the searching process will compute the sum of unique dataset entries of all field/keyword sets. This means that dataset entries detected by more than one set of field/keywords will be counted only once.
    6. After the dataset scanning finishes, a screen summarizing the search process and its results will be displayed, and the files mentioned in step “10.1” will be stored in their corresponding folders.
    7. Press any key to go back to the “Main Menu” screen.
11. Exclusion and re-inclusion of isolates that will have their whole genome sequences (WGS) downloaded from the National Center for Biotechnology Information (NCBI) in the study sample:
    1. The ‘isolates metadata.csv’ file has all metadata from each isolate in the sample plus the response to the question ‘keep entry?’. This response is ‘Y’ by default indicating the download of the isolate’s WGS from NCBI.
    2. To exclude undesired isolates:
       1. Open ‘isolates metadata.csv’ in Microsoft Excel (Microsoft Office, Microsoft Co.), Apple Numbers (Apple iWork, Apple Inc.), or another spreadsheet program;
       2. Change the response to the “keep entry?” question from ‘Y’ to ‘N’ in the undesired isolates;
       3. Save the file again in csv format (just overwrite the previous file);
       4. Input option “6” or “14” at the “Please, enter your choice from above list:” prompt to exclude the undesired isolates for WGS download from Patric or Enterobase accession numbers files, respectively;
       5. SampleMiner will generate a new ‘accession numbers.txt’ file, containing only the references to the sequences of the desired isolates in the “samples accession numbers” folder. Two other files will also be generated in this folder: “accession numbers\_RM.txt”, containing only the accession numbers of the isolates that will not have their WGS downloaded from NCBI; and “accession numbers\_BU.txt”, containing a backup of all accession numbers originally included during the sampling process.
    3. To reinclude isolates:
       1. Open ‘isolates metadata.csv’ in Microsoft Excel (Microsoft Office, Microsoft Co.), Apple Numbers (Apple iWork, Apple Inc.), or another spreadsheet program;
       2. Change the response to the “keep entry?” question from ‘N’ to ‘Y’ in the isolates to be reincluded in the sample that will have the WGS downloaded from NCBI;
       3. Save the file again in csv format (just overwrite the previous file);
       4. Input option “6” or “14” at the “Please, enter your choice from above list:” prompt to reinclude the isolates for WGS download in the Patric or Enterobase accession numbers files, respectively;
       5. SampleMiner will generate a new ‘accession numbers.txt’ file, containing only the references to the sequences of the desired isolates in the “samples accession numbers” folder. Two other files will also be generated in this folder: “accession numbers\_RM.txt”, containing only the accession numbers of the isolates that will not have their WGS downloaded from NCBI; and “accession numbers\_BU.txt”, containing a backup of all accession numbers originally included during the sampling process.
12. Producing a summary of the metadata of the enrolled sample of isolates:
    1. Input option “7” or “15” at the “Please, enter your choice from above list:” prompt to produce a written summary of the enrolled sample. The summary will scroll through the terminal screen, and a “isolates metadata summary.txt” file with exactly the same information will be stored in the “samples metadata” folder.
    2. Input option “8” or “16” at the “Please, enter your choice from above list:” prompt to produce a graphic summary of the enrolled sample. Each metadata summarized will be displayed in a different graphic. Each graphic will need to be separately saved to a directory of preference if the user wants to keep the information for later evaluation. The graphs will be displayed in succession with a new graphic popping up as a previous is closed by the user. To close the last graph, it will be necessary to click in the close red button at the left top of the window, then highlight the SampleMinder.py running terminal window and press any key.
13. Fetching isolates’ whole genome sequences (WGS) from the National Center for Biotechnology Information (NCBI):
    1. Input option “40” at the “Please, enter your choice from above list:” prompt to automatically bulk download from NCBI the WGS of the isolates represented in the ‘accession numbers.txt’ file to the “downloaded sequences” folder in the working directory.
    2. During the WGS download process, the information about the sequences being downloaded from NCBI will be displayed on the terminal screen.
    3. Two report files will also be saved to the ‘reports’ folder as a result of the WGS download process: “sequences download report.txt”, consisting of the same information that was displayed on the terminal screen during the WGS download process; “fasta headers.txt”, containing the FASTA header of each WGS downloaded from NCBI.
    4. Press any key to go back to the “Main Menu” screen.
    5. Notes:
       * + - The user’s email needs to be added to the statements ‘Entrez.email = "user@my.email"’ found in lines 1871, 1953 and 2132 of SampleMiner source code because NCBI Entrez requires this information in case it needs to contact the user;
           - Preferably perform bulk WGS download from NCBI at night or during holidays and weekends, as suggested by NCBI, particularly if long lists of WGS will be retrieved;
           - Bulk download of long WGS lists may seldomly fail during the execution process due to issues related to the NCBI servers or the internet route. If this occurs, try the download at a later time or on a different day.
14. Quitting SampleMiner.
    1. Input option “50” at the “Please, enter your choice from above list:” prompt to quit SampleMiner.
    2. When the program ends its execution, manually close the terminal window.