Gen2Epi GUI User Manual

How to use Gen2Epi to analyze dataset different from Test_DATA?

Case1: Complete analysis of *N.gonorrhoeae* WGS data starting from Data Cleaning to Strain typing.

1) Get your dataset ready for analysis

    a. Arrange your datasets in the following order:
        i. All *N. gonorrhoeae* plasmids save in a text file name "Plasmid.fasta".
        ii. One folder that contains all the .fastq files.
        iii. Second folder with *N. gonorrhoeae* genomes in it.
        iv. The third folder with *N. gonorrhoeae* genomes annotation. Make sure to use the annotation in "TXT" format.
    b. Datasets are now ready to use

2) Click on the "Gen2EpiGUI" icon on the Desktop to start the program.

3) Prepare Input

    a. Project -> Set Output Directory -> Browse the folder in which you would like to create the new output file. Write the name of your output directory in the search box e.g. "Output" and click confirm.
    b. Select Input directory that has all fastq files in it by clicking on "Project-> Set Input Directories->Fastq" and browse the folder that has fastq files in it.
    c. Click "Prepare Input" -> Enter a number e.g "5 or any number (depending upon how many characters you would like to have in the name of your file)" to define the length of the file name and click on Confirm. This will create a file name "Prepare_Input.txt" in your output directory folder (created in step 3a).
    Please Note: The number is basically the characters that you would like to use in your output directory name. For example, let's say you have a fastq files with name "WHO-F_S2_R1.fastq.gz" and the name that you would like to assign to your output folder is WHO-F. WHO-F has 5 characters in them so the length of the filename is 5. Instead of 5 if you will use 8 then the name of your output folder would be WHO-F_S2 and so on. Furthermore, this is the same number that you need to define during the "Prepare Input" step in Gen2Epi.

4) Set Input and Output Directories

    a. Select the input text file (tab separated file listing all paired-end fastq files - "Prepare_Input.txt" created in step 3c) by clicking on the "Project->Open Project".
    b. Set the output directory by clicking on "Set Output Directory" option. Write the name of your output directory in the search box and click confirm. If you have already set the output directory in step 3a then you need not to create the directory

again. Skip this step or if you would like to set the path again then simply select the output directory, leave the search box blank and click on confirm.

c. Set Input directories for Fastq files by clicking on Project-> Set Input Directories->Fastq. Browse the folder that you prepare in step 1a (ii).
d. Set Input directories for genome by clicking on Project-> Set Input Directories->Genome File. Browse the folder that you prepare in step 1a (iii).
e. Set Input directories for genome annotation by clicking on Project-> Set Input Directories->Annotation. Browse the folder that you prepare in step 1a (iv).

5) Data Cleaning

a. Find out the total percentage of *N.gonorrhoeae* reads in the WGS fastq files just by simply clicking on "Data Cleaning->Read Mapping".
b. Check the quality of your raw input fastq files by clicking on the "Data Cleaning->Quality Control" tab.
c. Next, to trim the raw reads click on "Data Cleaning->Trimming". Once you click on "Trimming" button, it will further ask for parameter confirmations. Users have the option to use either the default parameters or change them according to their requirements. After confirmation, the program will start trimming the reads.
Note: In case you would like to trim only a few fastq files (instead of all) then edit the "Prepare_Input.txt" file (make sure to keep a backup of the original file) by removing the entries that you do not want to trim. Upload the file again in Gen2Epi via Project->Open Project menu and proceed further with the trimming step.
d. Quality check and trimming of the raw fastq files can also be performed in a single step by clicking on the "Data Cleaning-> Both" menu.
e. Output generated from Data Cleaning step will be stored under the output directory created in step 3a.

6) Assembly

a. Next, perform the de novo assembly of raw reads or trimmed reads by clicking on "Assembly->De Novo". This step performs the de novo assembly for both chromosome and plasmid reads. **Please Note**: - This step may take a while depending on individual computer configurations.
b. To perform scaffolding of the assemble contigs click on "Assembly->Scaffolding->Chromosome->Multiple Genomes".
c. Now identify the *N.gonorrhoeae* plasmid type from assembled plasmid contigs using "Assembly->Scaffolding->Plasmid"
d. Output generated from this step will be stored under the output directory created in step 3a.

7) AMR & Strain

    a. NgMAST Typing

        i. Update the underlying MAST database by clicking “AMR & Strain->MAST->MAST Update”.
        ii. Perform the MAST typing by selecting “AMR & Strain->MAST->MAST Typing”.

    b. NgMLST Typing

        i. Update the underlying MLST database by clicking “AMR & Strain->MLST->MLST Update”.
        ii. Perform the MLST typing by selecting “AMR & Strain->MLST->MLST Typing”.

    c. NgSTAR

        i. Update the underlying MLST database by clicking “AMR & Strain->STAR->STAR Update->STARdbUpdate”
        ii. Update the underlying MLST database metadata by clicking “AMR & Strain->STAR->STAR Update->STARdbMetaDataUpdate”
        iii. Perform the STAR typing by selecting “AMR & Strain->STAR->STAR Typing

    d. Tetracycline Resistance

        i. Click on “AMR & Strain->TETRES-> Nucleotide” to extract the rpsJ nucleotide sequence alignment. These fasta sequences can be visualized under any multiple sequence aligner program.
        ii. Click on “AMR & Strain->TETRES->Protein” to extract the rpsJ protein sequence alignment. These fasta sequences can be visualized under any multiple sequence aligner program.

8) View Output Files: Finally, the text and HTML outputs generated from the above steps are accessible under “View Output File” tab

    a. Visualize the output results (Text files) in the main GUI window by clicking on “View Output Files->View Text File in Main Window” and browse the file.
    b. To visualize a text file in a new window click on “View Output Files-> View Text in New Window” and browse the file.

c. The HTML output files can be visualized using “View Output Files->View HTML in New Window”. Browse the file and you will see the results in the HTML browser window.

Transfer data between Gen2Epi VM image and Host machine

9) Create a Folder named “Gen2Epi_TestRun” on your host machine.

10) Share the folder between Gen2Epi VM and Host machine as shown in “**ReferenceGuide.pdf**” under the Help menu

11) Restart the Gen2Epi VM image and you will see a mounted folder named “sf_Gen2Epi_TestRun” on desktop.

Case2: Analyzing *N.gonorrhoeae* preassembled contigs using Gen2Epi.

1) Prepare Input:

    a. Project -> Set Output Directory -> Browse the folder in which you would like to create the new output file. Write the name of your output directory in the search box e.g. "Output" and click confirm.
    b. Select Input directory that has all fastq files in it by clicking on "Project-> Set Input Directories->Fastq" and browse the folder that has fastq files in it.
    c. Click "Prepare Input" -> Enter a number e.g "5 or any number (depending upon how many characters you would like to have in the name of your file)" to define the length of the file name and click on Confirm. This will create a file name "Prepare_Input.txt" in your output directory folder (created in step 3a).
    Please Note: The number is basically the characters that you would like to use in your output directory name. For example, let's say you have a fastq files with name "WHO-F_S2_R1.fastq.gz" and the name that you would like to assign to your output folder is WHO-F. WHO-F has 5 characters in them so the length of the filename is 5. Instead of 5 if you will use 8 then the name of your output folder would be WHO-F_S2 and so on. Furthermore, this is the same number that you need to define during the "Prepare Input" step in Gen2Epi.

2) Set Input and Output Directories:

    a. Select the input text file (tab separated file listing all paired-end fastq files - "Prepare_Input.txt" created in step 3c) by clicking on the "Project->Open Project".
    b. Set the output directory by clicking on "Set Output Directory" option. Write the name of your output directory in the search box and click confirm. If you have already set the output directory in step 3a then you need not to create the directory again. Skip this step or if you would like to set the path again then simply select the output directory, leave the search box blank and click on confirm.
    c. Set Input directories for Fastq files by clicking on Project-> Set Input Directories->Fastq. Browse the folder that you prepare in step 1a (ii).
    d. Set Input directories for genome by clicking on Project-> Set Input Directories->Genome File. Browse the folder that you prepare in step 1a (iii).
    e. Set Input directories for genome annotation by clicking on Project-> Set Input Directories->Annotation. Browse the folder that you prepare in step 1a (iv).

3) Preassembled genome:

    a. Copy the assembled contigs from different datasets in separate folders in a directory called "Chrom_AssemblyTrimmedReads". Next, make sure to copy the "Chrom_AssemblyTrimmedReads" folder that contains *de novo* assemblies for

each sample into the output directory created in step 1a. If you are not sure about the "Chrom_AssemblyTrimmedReads" folder then you can run Gen2Epi on Test_DATA and see the arrangement of files in this folder.

b. Now go to Assembly and click on Preassembled Genome tab. This step will start the scaffolding on your preassembled genomes
c. Once completed, you can access the results in "Chr_Scaffolds_Sameref" under your output directory created in step 1a.
d. If you have the assembled contigs from plasmid reads then copy them a folder named "Plasmid_AssemblyTrimmedReads" and identify the *N.gonorrhoeae* plasmid type from assembled plasmid contigs using "Assembly->Scaffolding->Plasmid".

4) AMR & Strain

a. NgMAST Typing

i. Update the underlying MAST database by clicking "AMR & Strain->MAST->MAST Update".
ii. Perform the MAST typing by selecting "AMR & Strain->MAST->MAST Typing".

b. NgMLST Typing

i. Update the underlying MLST database by clicking "AMR & Strain->MLST->MLST Update".
ii. Perform the MLST typing by selecting "AMR & Strain->MLST->MLST Typing".

c. NgSTAR

i. Update the underlying MLST database by clicking "AMR & Strain->STAR->STAR Update->STARdbUpdate"
ii. Update the underlying MLST database metadata by clicking "AMR & Strain->STAR->STAR Update->STARdbMetaDataUpdate"
iii. Perform the STAR typing by selecting "AMR & Strain->STAR->STAR Typing

d. Tetracycline Resistance

i. Click on "AMR & Strain->TETRES-> Nucleotide" to extract the rpsJ nucleotide sequence alignment. These fasta sequences can be visualized under any multiple sequence aligner program.

ii. Click on “AMR & Strain->TETRES->Protein” to extract the rpsJ protein sequence alignment. These fasta sequences can be visualized under any multiple sequence aligner program.

5) View Output Files: Finally, the text and HTML outputs generated from the above steps are accessible under “View Output File” tab

   a. Visualize the output results (Text files) in the main GUI window by clicking on “View Output Files->View Text File in Main Window” and browse the file.
   b. To visualize a text file in a new window click on “View Output Files-> View Text in New Window” and browse the file.
   c. The HTML output files can be visualized using “View Output Files->View HTML in New Window”. Browse the file and you will see the results in the HTML browser window.