Quick-Start Protocol for GMI Proficiency Test, 2017

HISTORY OF CHANGES; version 4

Under 3.1, specified the procedure of submitting sequences for the wet lab component

This document describes the basic instructions for both the dry and wet lab components of the PT. See the full Protocol document including appendices for more information: http://www.globalmicrobialidentifier.org/workgroups/about-the-gmi-proficiency-test-2017.

1 OVERVIEW

The proficiency test, 2017, consists of three parts:

1a. DNA extraction, purification, library-preparation, and whole-genome-sequencing (WGS) from **live cultures**

1b. Whole-genome-sequencing of **pre-prepared DNA**

2. Phylogenetic/clustering analysis of three **fastq datasets**

2 SHIPPING, RECEIPT AND STORAGE OF BACTERIAL STRAINS

All bacterial strains and DNA are shipped as UN3373, Biological substance category B. **Please confirm receipt of the parcel through the confirmation form enclosed in the shipment.**

3 PROCEDURE AND ANALYSIS OF TEST MATERIAL

3.1 Bacterial cultures and DNA

Subculture the bacterial strains on a relevant growth medium of the laboratory's own choice and incubate. Following incubation and assessment of purity of the bacterial cultures, perform DNA extraction and whole-genome-sequencing according to the laboratory's standard procedure.

For the purified PT-DNA received, perform whole-genome-sequencing according to the laboratory's standard procedure.

The metadata file and WGS data must be submitted as a batch-upload. The web-interface of the (https://cge.cbs.dtu.dk/services/ringtrials/) provides a possibility to upload several isolates in a single submission.

Step 1; Login to the server using provided username (gmi_xx) and password (PINK area).

Step 2; Download the Excel Metadata template to your computer (GREEN area).

Step 3; Fill in the required fields with all the relevant information (metadata) about the isolates, the associated WGS file names, sequencing platform and sequencing type used to generate the data, etc., all in one line for each sample. In the metadatasheet, each cell includes a brief description of the required metadata or provide drop-down lists with possible metadata. Note that **sample name** should be the **same as label-name of the sample**, e.g. **GMI17-003-BACT** or **GMI17-003-DNA**.

Preferably, the FASTQ-files should be renamed to match corresponding samples. For example, files for sample GMI17-003-BACT should be called GMI17-003-BACT_R1.fastq.gz and GMI17-003-BACT_R2.fastq.gz (if pair-end sequencing), and GMI17-003-BACT_R1.fastq.gz (if single-end sequencing). This is preferable, but not mandatory.

Step 4; When the spreadsheet is properly filled out, upload the metadata file as well as the individual WGS files, that were included as metadata in the spreadsheet, to the web-interface by dragging and dropping the files to the ‘Drop metadata and sequence files here’ (GREY area). After this, the spreadsheet will be validated to check if the metadata has the correct format.

If the spreadsheet contains invalid metadata, an error message will appear at the top of the uploader (between blue and green). To fix the errors, correct the errors in the original spreadsheet and upload the updated spreadsheet.

Step 5; Click on the Submit button to upload the files (BLUE area). The progress will be displayed in the Upload Progress bar (below the blue ‘Submit’ button). It is important to keep the window opened until the upload is completed. Pre-loading and validation will be displayed in blue and uploading will be displayed in green. After the files are successfully uploaded, click on the small, grey ‘submit’ button below the upload area.

Step 6; If the WGS files and metadata are successfully uploaded, you will be redirected to the next page saying ‘Your job is being processed’ or ‘Your job has been queued’. Here, you are suggested to provide an e-mail address. Processing of your submission might take up to 14 days, depending on the occupation of the server.

In addition, via the Internet-based survey (https://www.surveymonkey.com/r/PT_2017_bacterial_cultures_and_DNA; see also Appendix 2 in the full protocol document), answers should be submitted to the questions related to the analysed bacterial cultures and DNA.

3.2 Fastq dataset

The three fastq datasets should be downloaded from the ftp-site. They are organized into three different .zip archives appropriately labeled with the taxon they represent (st, ec, or sa). Within each archive the participant will find the paired-end reads.

The participant should perform variant detection and clustering (phylogenetics) of all files within each .zip archive according to participant’s standard procedure.

For each .zip dataset the following should be uploaded to the ftp-site:

1. The DNA sequence matrix used for clustering should be in fasta format (.fasta file) and the clusters should be in newick format (.tre file)
   - The matrix and tree file should contain ***only*** those samples provided through the ftp site (i.e., there should be only 14 st, 11 ec, and 11 sa samples in each file).
   - Syntax for the names of samples in each file should be ***only*** the prefix preceding the first underscore in the file name. For example, **st1_1.fastq** should be named **st1** in the matrix and tree files.
   - The file should be named as follows **GMILabID_Taxon.fasta** (e.g., **GMI01_st.FASTA**, **GMI01_ec.FASTA**, **GMI01_sa.FASTA**, **GMI01_ st.TRE**, **GMI01_ec.TRE**, or **GMI01_sa.TRE**).
2. The vcf (variant call format) files for each sample if a reference based approach was used and such files were produced.
   - The number of vcf files should match the number of samples found in the zipped archive from the ftp site.
   - Syntax for the names of the files should be *only* the prefix preceeding the first underscore in the file name. For example, **st1_1.fastq** should be named **st1** in the matrix.
   - The file should be named as follows **GMILabID_Taxon.tre** (e.g., **GMI01_st1.vcf**, **GMI01_ec1.vcf**, or **GMI01_sa1.vcf**).

If performing a reference based approach for variant detection, the reference applied for the analysis **must be** st_reference.fasta (*Salmonella*), ec_reference.fasta (*E. coli*) and sa_reference.fasta (*S. aureus*).

Via the Internet-based survey (https://www.surveymonkey.com/r/PT_2017_FASTQ_dataset; see also Appendix 2 in the full protocol document), answers should be submitted to the questions related to the analysed of the fastq dataset.

4 DISCUSSION FORUM

A web-based discussion forum (https://foros.isciii.es/viewforum.php?f=7) is available for participants in the GMI PT 2017. Appendix 4 in the full protocol document presents detailed information on the PT discussion forum.

5 CONTACT INFO

If you have any questions or concerns, please do not hesitate to contact us, preferably by using the web-based discussion forum (https://foros.isciii.es/viewforum.php?f=7).

PT organizer related to the dry-lab fastq datasets:

James Pettengill
U.S. Food and Drug Administration
Center for Food Safety and Applied Nutrition
CPK1 RM2D0195100 Paint Branch Parkway
College Park, MD 20740, US

Tel: +1 240-402-1992
E-mail: James.Pettengill@fda.hhs.gov

PT organizer in relation to other issues, e.g. organizational issues, please contact the GMI PT Coordinator:
Susanne Karlsmose Pedersen
National Food Institute, Technical University of Denmark
Kemitorvet, Building 204,
DK-2800 Kgs. Lyngby, DENMARK
Tel: +45 3588 6601
E-mail: suska@food.dtu.dk

--- --- ---

Note:
This document describes the basic instructions for both the dry and wet lab components of the PT. See the full protocol document including appendices for more information: http://www.globalmicrobialidentifier.org/workgroups/about-the-gmi-proficiency-test-2017.