Protocol for GCRSR Proficiency Test, 2021

1 Summary

2 Introduction 2

3 Outline of the GCRSR PT 3
- 3.1 FASTQ Datasets 3
- 3.2 Sequence download 3
- 3.3 Sequence Analysis 4

4 Discussion Forum 4

5 Reporting of Results and Evaluation 4
- 5.1 Procedure 5
- 5.2 Evaluation of Results 5
- 5.3 Deadline for Submission of Results 5
- 5.4 Analysis and Publication of Results 5

6 PT Coordinator 6

Appendix I: Using FTP to transfer files 7

Appendix II: Pipeline Questionnaire 8

Appendix III: Reporting Spreadsheet 9

Appendix IV: Discussion Forum **Error! Bookmark not defined.**

1 Summary

The application of whole-genome sequence (WGS) technology in regulatory food microbiology provides an unprecedented opportunity to produce highly informative laboratory analyses supporting risk assessment and risk management actions. The quality of WGS datasets will have a significant impact on downstream bioinformatics processes, with one critical element being the possible presence of adventitious DNA sequences due to contamination during sample handling and sequencing operations. This project addresses the need to assure WGS data quality by accounting for contamination events through determination of the impacts of sequencing data contamination events on downstream analyses, such as typing and marker discovery. The goal is to contribute to the development and implementation of harmonized quality protocols for the application of WGS technologies in the international regulatory food microbiology community.

2 Introduction

High quality WGS data is critical for regulatory and outbreak investigations of foodborne pathogens. Quality assessment of bacterial WGS data (usually short-read sequence data) involves examining measurements such as coverage, read-length, assembly metrics and quality score to ensure that it is fit-for-purpose. The purity of DNA samples is also an important consideration as cross-contamination during sample preparations for sequencing (e.g., during colony picking and growth of bacteria, gDNA extraction, and library preparation) and carryover from previous runs, a known source of impurities with the Illumina sequencing platform, may occur. The foci of genome sequencing in food safety laboratories are mainly *Escherichia coli*, *Listeria monocytogenes*, *Campylobacter jejuni/coli*, and *Salmonella enterica*. It is currently unknown, however, how contamination of WGS data for one of these pathogens by another (e.g., *E. coli* contamination of an *L. monocytogenes* sequencing run) influences downstream bioinformatic analyses and interpretation of results. Even less understood are the ramifications of contamination by the same species (e.g., *E. coli* contamination of an *E. coli* sequencing run) on analyses and final interpretations. Gaining an understanding of the impact of purity of samples and the quality of WGS data necessary for robust analyses and appropriate interpretation of results will allow laboratories to implement controls that will ensure the high level of confidence necessary for WGS analysis in a regulatory context.

In addition to data quality and purity, the impact of bioinformatic methodologies and parameters on the reproducibility of WGS analyses for pathogen identification and outbreak detection may affect the interpretation of results. For example, single-nucleotide polymorphism and multi-locus sequence subtyping methods may yield different results under certain conditions, such as variations in sequence quality, degrees of DNA sample purity, and different dataset compositions. In the case of phylogenetic analyses, these differences may manifest as variations in branching order and branch lengths that could change investigative conclusions. Conclusions regarding the identification of alleles in target loci, such as those that confer antimicrobial resistance or those that contribute to propensity to cause disease (virulence), may also be altered by data quality, purity, and methodology.

This project will undertake an inter-laboratory study to examine the impacts on accuracy, specificity, and sensitivity of WGS analyses performed with multiple bioinformatic approaches on real and simulated datasets of varying qualities, including different amounts of contamination with genomic DNA of various genetic distances. For this purpose, a network of user laboratories involved with the routine application of WGS technology in support of regulatory microbiological food safety will be recruited to receive proficiency verification materials (*in silico* datasets as well genomic DNA preparations) with incurred contamination and asked to perform their in-house clustering and gene discovery analyses.

The first round will consist of *in silico* datasets which the participating laboratories will analyze using their preferred bioinformatics workflows. The results will be recorded on specially designed reporting sheets and submitted to the project coordinator for determination of accuracy and impacts of contamination or other sequencing issues on quality metrics and error in the outputs. The main objective of this proficiency test is to quantify differences among laboratories in order to determine the influence that contamination (both inter- and intra-genus) and sequence quality have on assembly metrics, gene and/or allele determination, including virulence and antimicrobial profiling, and sequence typing.

3 Outline of the GCRSR PT

The proficiency test consists of three datasets. They are organized into mock Illumina MiSeq sequencing runs. The runs consist of the same 24 *Escherichia coli* samples. FASTQ files for these samples were created by simulating reads from MinION/PacBIO + Illumina MiSeq hybrid-assembly polished genomes or closed reference genomes downloaded from NCBI. Reads were simulated with ART [1]. One run was created using the empirical Illumina MiSeq profile, while the other runs were simulated using custom read profiles generated from over-clustered runs. The reference strains represent a diverse cross-section of serotypes, Shiga-toxin subtype, antimicrobial resistance (AMR) profiles, plasmid profiles.

3.1 FASTQ Datasets

There are three FASTQ datasets to be downloaded for analysis. Each dataset is a compressed archive consisting of 48 FASTQ files.

3.2 Sequence download

Zip archives containing FASTQ sequences, as well as other run-specific files from three simulated Illumina MiSeq sequencing runs are available to download separately from the FTP server (see Appendix I for additional information).
Runs are available at https://doi.org/10.5281/zenodo.4342461 as:

Run 1: GCRSR_run1.tar.gz
Run 2: GCRSR_run2.tar.gz
Run 3: GCRSR_run3.tar.gz

3.3 Sequence Analysis

Each of the sequencing runs should be processed by the participants with the *de novo* assembly pipeline routinely used in their laboratory. Though not a requirement, these pipelines should contain sequence quality assessments (FastQC, etc.), *de novo* assembly, quality assessments of assemblies (Quast, Qualimap, etc.), typing (MLST, rMLST, 16S, MASH, etc.), and gene predictions (serotype, Shiga toxin, AMR).

For each sample, a decision of whether the sample would pass whatever quality thresholds are implemented at your institution, and why or why not will be requested.

4 Discussion Forum

Separate from this document, you will have received an invitation to the GCRSR 2020 discussion forum. If you did not receive an invitation, or wish to use an alternative email address, please send a message to the PT organisers.

The discussion forum (https://groups.google.com/d/forum/gcrsr-2020) is available for participants in the GCRSR 2020 PT, allowing for individual sign-up and discussion with other PT-participants and the PT organizers in relation to issues relating to the analysis for the present PT.

Appendix IV presents detailed information on the PT discussion forum. We strongly encourage the use of this forum as both a resource among participants but also as a platform to ask questions of the organizers – which other participants might also be interested in hearing the answer.

5 Reporting of Results and Evaluation

Please complete the GCRSR questionnaire supplied at https://docs.google.com/forms/d/e/1FAIpQLSeHnqTyW-X1ydjaRPmZu9HYoqdhuZI3SYCa4Sh6vdHAL5zOmQ/viewform. The questionnaire focuses on the specifications of the pipelines used to process the samples (e.g. analyses performed, names, and versions of programs within the pipeline). Detailed instructions for completing the questionnaire can be found in Appendix II.

For each sample, the appropriate row in the reporting spreadsheet should be populated. Detailed instructions on how to populate the spreadsheet are included in Appendix III.

5.1 Procedure

Syntax for the names of the samples should be the prefix preceding the first underscore in the paired FASTQ file name (e.g. 2020-GCRSR-0001_S1_L001_R1_001.fastq.gz should be referred to as 2020-GCRSR-0001 in any reports, and assemblies should be named 2020-GCRSR-0001.fasta)

The completed spreadsheet, as well as all the assemblies in FASTA format should be compressed into a single zip archive with the following naming scheme: GCRSR_ORGANIZATION_NAME.zip (e.g. GCRSR_CFIA_Blais.zip), and uploaded to the FTP address below:

ftp://ftp.agr.gc.ca/incoming/cfia-ak/

5.2 Evaluation of Results

Submitted FASTA files will be processed with in-house typing pipelines to determine quality metrics, and gene profiles. In addition, the contigs of submitted assemblies will be mapped to the relevant closed genome to assess the sequence error rate and coverage of the scaffold.

Assessment of the submitted results from the analysis of the uploaded FASTA datasets is based on three criteria:

1. The concordance among laboratories in their answers to the questions in the supplied ‘GCRSR_reporting_table.xlsx’
2. The concordance between participants from the quality assessment of assemblies, and the outputs obtained from mapping assemblies against reference genomes.
3. The responses of the participants to the online survey

5.3 Deadline for Submission of Results

Results must be submitted electronically no later than **March 31, 2021** (the extended deadline). Immediately after this date, the PT will be closed and results submitted to the Internet-based survey, and to the ftp-site will be evaluated. Delayed submission of results
will not be accepted.

5.4 Analysis and Publication of Results

Individual results will be anonymized, and only the PT-organizers will have access to your laboratory’s results. Each participating laboratory will receive an individual summary of the obtained performance. An overall report summarizing the results will be published and subsequently in a peer-reviewed publication. Authors and co-authors of the publications will be those who have contributed to the preparation and execution of the proficiency test.

We are looking forward to receiving your results.

If you have any questions or concerns, please do not hesitate to contact us, preferably by using the web-based discussion forum (https://groups.google.com/d/forum/gcrsr-2020). In addition to receiving a response to your question, bringing up an issue via the forum allows other participants to also benefit from the discussions and the PT-organizers' response.

6 PT Coordinator

Burton Blais
Canadian Food Inspection Agency
Research and Development
960 Carling Ave
Building 22, Central Experimental Farm
K1A 0C6
Ottawa, Ontario, Canada
+1-613-759-1267
burton.blais@canada.ca

Appendix I: Using FTP to transfer files

The FTP server allows for anonymous connections, so logging in is not required.

Downloading files:

The simulated datasets can be downloaded at https://doi.org/10.5281/zenodo.4342461:
Run 1: GCRSR_run1.tar.gz
Run 2: GCRSR_run2.tar.gz
Run 3: GCRSR_run3.tar.gz

Uploading files:

From within a file browser (e.g. Linux Files, Windows Explorer, Mac Finder - NOTE that is not using a web browser), type ftp://ftp.agr.gc.ca/incoming/cfia-ak in the Address bar. You will connect to the FTP server. You will be able to copy the archive to the FTP server. Please send an email to the PT Coordinator once the upload is complete.

Appendix II: Pipeline Questionnaire

The questionnaire is available at https://docs.google.com/forms/d/e/1FAIpQLSeHnqTyW-X1ydjaRPmZu9HYoqdhuZI3SYCa4Sh6vdHAL5zOmQ/viewform

The questionnaire seeks to capture details regarding the pipelines and any *ad hoc* analyses used to process the FASTQ datasets.

If you have any questions or feedback for the submission of data via this survey, please create a post in the discussion forum (https://groups.google.com/g/gcrsr-2021), or contact the PT Coordinator.

Note: An asterisk (*) indicates a question that requires an answer.

Appendix III: Reporting Spreadsheet

The reporting spreadsheet is available at https://doi.org/10.5281/zenodo.4342461 and downloadable as GCRSR_reporting_table.xlsx

The reporting spreadsheet contains headers for the sample name (pre-populated) sample quality control, requested gene presence/absence or allele, as well as requested sequence types.

Enter whether the sequence data passes pipeline-specific quality controls (if any), provide free-text justification for any failed sequences, and indicate whether samples have been determined to be contaminated. Provide the profile number obtained for rMLST and MLST results, and the serotype obtained (e.g. O157:H7). For gene queries, either presence/absence (indicated by Y or N), or the subtype/allele can be supplied (e.g. either Y, *2a*, or *stx2a* would be an acceptable entry for the *stx2* column, while *aadA1*, 2b, or Y would be acceptable entries for the *aadA* column).

Due to the large size of the reporting spreadsheet, examples will be truncated.

Sequence Quality Control

SampleName	Genus	Species	Pass/ Fail	Fail Justification	Contamination Status	...
2021-GCRSR-9999	Escherichia	coli	Pass	NA	N	...

Sequence Typing

SampleName	...	rMLST_Profile	MLST_Profile	Serotype	...
2021-GCRSR-9999	...	2124	11	O157:H7	...

Virulence Typing

SampleName	...	*eae*	*hlyA*	*aggR*	*stx1*	*stx2*
2021-GCRSR-9999	...	N	N	N	N	stx2a;stx2c

Antimicrobial Resistance Typing (All results Y or N)

SampleName	...	*aac*	*aadA1*	*aph(3')*	...	*mdf(A)*	blaCTX-M-15
2021-GCRSR-9999	...	N	Y	N	...	Y	Y

Plasmid Typing (All results Y or N)

SampleName	...	*IncFIB*	*IncFIC*	*IncFII*	...	*IncX1*	IncX4
2021-GCRSR-9999	...	N	Y	N	**...**	N	N