FOOD SAFETY LAB / MILK QUALITY IMPROVEMENT PROGRAM

Standard Operating Procedure

Title: Submission of fastq.gz files to Sequence Read Archive (NCBI)			
SOP #:	Revision: 00	Revision Date: If changed	Effective Date: Date Upload
Author: Sarah Beno and Sophia Harrand		Approved by: Martin Wiedmann	

<u>Short Read Submission (SRA)</u>

FILE NAME: Short_read_submission_SRA_SOP.doc

TABLE OF CONTENTS

SECTION 1 INTRODUCTION

This is protocol #3 out of 4 protocols for whole genome sequencing:

#1 Library preparation for whole genome sequencing
#2 Quality control and assembly
#3 Short read submission (SRA)
#4 Assembly submission to NCBI WGS

1.1 Purpose

The purpose of this document is to set forth **standard** guidelines for submitting fastq.gz files from whole genome sequencing outputs to NCBI's Sequence Read Archive (SRA) database. This will ultimately lead to easy accession and allow for reproducible analysis.

1.2 Scope

This SOP applies to the Food Safety Lab and the Milk Quality Improvement Program.

1.3 Definitions

SRA: The Sequence Read Archive (SRA) stores sequence and quality data in aligned or unaligned formats from NextGen sequencing platforms.

1.4 Safety

This is a fully computational procedure.

SECTION 2 MATERIALS

- **Short read sequences in a fastq.gz format.**

SECTION 3 PROCEDURES

This protocol can be carried out using different web browsers in different operation systems, which will perform differently (see the troubleshooting section).

Based on the past experience the submission can be successfully performed using Safari on Mac OS, Firefox on Linux and Chrome on Windows operation system.

Important notes:

- Submit sequences with trimmed adapters (.trimmedP.fastq.gz).
- In the process of sequence submission to the SRA you will generate a BioProject and BioSamples; the Bioproject and BioSample number information needs to be submitted to the Food Microbe Tracker (see the relevant section of the #2 Quality control and assembly).
- Create a single BioProject for all related sequences (e.g., the same organism, the same paper).
- Create individual BioSamples for individual isolates, but submit both forward and reverse .trimmedP.fastq.gz files as part of the single SRA experiment linked to a single BioSample.

3.1. Accessing the Submission Portal of NCBI

(1) Log in to NCBI and go to: https://submit.ncbi.nlm.nih.gov/ to submit your fastq.gz files.
(2) On the Submission Portal page, click on “SRA”.
(3) Click the “New Submission” button .

3.2. Submitter Page

(1) Enter your name and email in the appropriate boxes.
(2) Under “Submitting organization,” type: “Cornell University”
(3) Under “Department,” type: “Department of Food Science”
(4) Use Martin’s office as the address: 347 Stocking Hall, Ithaca, New York, 14853, United States of America
(5) Click the “Continue” button at the bottom of the page. You can also check the box to update your contact information in profile

3.3. General Info.

(1) Creating a BioProject: In the first box, you will be asked “Do you want to create a new BioProject?” For each group of isolates that are part of a unique project, you will create a new

BioProject. If you are submitting additional isolates from an existing BioProject, you can select "no" and type in your existing BioProject title.
(2) Next, you will be asked if you would like to create new BioSamples for this submission. Because you have not previously submitted sequences for these isolates, you will select "Yes".
(3) You will then select a release date. We generally choose the option to release on a specified date or upon publication. For the specified date, choose one year later. Click "Continue" at the bottom of the page. The NCBI will inform you prior to the sequence release, so you will be given an opportunity to further postpone the publication of your data later on.
(4) If you created new BioProject in step 3.3.1, you will then need to enter the details of your project.

(a) First, enter a project title. This can be a description of the isolates (e.g., Bacillus cereus group isolates from dairy and dairy farm environments) or of the project (e.g., Paenibacillus psychrotolerance).
(b) Next, give a description of the project. This should be a short summary of the research goals with these isolates.
(c) Select the appropriate Relevance of your project.
(d) Select whether your project is part of a larger initiative already registered with NCBI.
(e) If external links to this project exist, enter a description and the URL to the external link.
(f) You can also enter information of the grant supporting this study. If you are not sure, ask your PI (i.e., Martin) for this information.
(5) Click the "Continue" button.
(6) Enter Publication information if available.
(7) Click the "Continue" button.

3.4. BioSample Type

(1) Here, you will select the package that best describes your samples. In our lab, we most commonly use "Pathogen affecting public health" and then accordingly select "Clinical" or "Host-associated pathogen" for isolates including *Listeria monocytogenes*, *Salmonella*, and *Bacillus cereus*. The "Microbe" package can be chosen for spoilage organisms, such as *Paenibacillus* spp.
(2) Click "Continue" at the bottom of the page

3.5. BioSample Attributes

(1) Here, you can download an Excel spreadsheet template. As made evident in the template, green fields are mandatory. If you do not submit information in a mandatory field, the submission will fail. Blue fields indicate that at least one of those fields is mandatory. For example, on the Microbe attributes sheet, you can enter strain OR isolate and host OR isolation source. Yellow fields are optional. We recommend putting in all of the information that you know for each isolate.
(2) BioSamples can be batch uploaded by entering all isolates' information into the spreadsheet.

(3) As indicated on the Excel spreadsheet template, the worksheet needs to be saved as a Text Tab-delimited file and uploaded under the Attributes tab of the Submission form. Click "Continue."

3.6. SRA Metadata

(1) Download and open the Excel template. Instructions are in the first tab, while what you will submit is on the second tab.
(2) If you have previously created a BioProject, enter the BioProject accession number in the first column. If you are creating your BioProject with this submission, leave this column blank.
(3) Under sample name, make sure to enter the same sample names that you entered in your BioSample attributes file.
(4) Enter the FSL under library ID. (e.g., FSL A5-0030)
(5) Title will be entered as [Library strategy] of [Genus species] isolate [FSL]. For example, WGS of *Paenibacillus* isolate FSL A5-0030 or WGS of *Bacillus cereus* group isolate FSL W8-0169.
(6) The columns highlighted in yellow have a drop-down menu. All of the drop-down menu columns will likely be the same for everything in your spreadsheet.
(7) Under "design_description" provide a brief summary of the methodology used. Provide minimum the library preparation kit (e.g., Nextera XT, TrueSeq).
(7) You will have two files to upload for each isolate. Be sure that you enter the file names accordingly.
(8) Save the worksheet as a Text Tab-delimited file and upload it under the SRA Metadata tab of the Submission form. Click "Continue."

3.7. Files

(1) Here, you will select all files that were mentioned in your SRA Metadata spreadsheet. Click "I will upload all the files now via HTTP/Aspera.
(2) Select all fastq.gz files that were entered into your SRA Metadata spreadsheet. You may be asked to allow connection with"upload.ncbi.nlm.nih.gov". Click "Allow". It will take a few minutes for the files to upload.
(2) Once all files are uploaded, click "Continue."
(3) We recommend not checking the "Autofinish submission" box at the bottom of the page. By not clicking the box you have the opportunity to review your submission one more time.

3.8. Overview

(1) Review all information. If everything is satisfactory, click "Submit" at the bottom of the page. You will receive an email with BioSample SRA accession numbers once they have been processed. (It will only take a few minutes).

SECTION 4 TROUBLESHOOTING

If you try to upload several .trimmedP.fastq.gz files for several BioSamples at once and it does not work, try submitting .trimmedP.fastq.gz files for each BiosSmple separately.

Based on the past experience the submission can be successfully performed using Safari on Mac OS, Firefox on Linux and Chrome on Windows operation system. The Firexox did not work well on Windows or Mac for some users.

SECTION 5 REFERENCES

SRA submission portal with additional information.
https://submit.ncbi.nlm.nih.gov/subs/sra/