FOOD SAFETY LAB / MILK QUALITY IMPROVEMENT PROGRAM

*Standard Operating Procedure*

| | | | |
|---|---|---|---|
| Title: #4 Submission of assembled contigs to NCBI Whole Genome Shotgun | | | |
| SOP #: | Revision: 00 | Revision Date: If changed | Effective Date: Date Upload |
| Author: Sarah Beno, Renato Orsi and Jingqiu Liao | | Approved by: Martin Wiedmann | |

# ***Submission of assembled contigs to NCBI Whole Genome Shotgun***

## **FILE NAME: Assembly_submission_NCBI_WGS.doc**

TABLE OF CONTENTS

# SECTION 1 INTRODUCTION

**This is protocol #4 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 SPAdes de novo assembly files from whole genome sequencing outputs to NCBI's Genomes (WGS & Complete) 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

**Genomes (WGS & Complete) database:** GenBank genomes accepts prokaryotic and eukaryotic genomes that are either draft/incomplete (WGS) or complete. This submission portal was updated on Feb. 3, 2014, and now accepts fasta sequences.

## 1.4 Safety

This protocol is computer-based only.

## SECTION 2 MATERIALS

- **Assembled draft genome.** Draft genome should be in a multifasta file, containing contigs larger than 200 bp, all confirmed to map to a sequenced organism either by Kraken or by BLAS. For details, please refer to the WGS_quality_control_SOP.docx.

- **Computer with OS, Windows or Linux operating system.**

- **Access to the internet.** Draft genomes should be submitted through NCBI Whole Genome Shotgun submission portal accessible on the following link: https://submit.ncbi.nlm.nih.gov/subs/wgs/

  - **NCBI's instructions on how to submit draft genomes are available on the following link: https://www.ncbi.nlm.nih.gov/genbank/wgs.submit/**

- **Web browser.** A submission whose files total more than 2G will need to use Chrome, Opera or Safari browsers, which do not have a size limit on what they can upload. You can also compress large files to decrease large files for the upload.

# SECTION 3 PROCEDURES

## 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 contig files.
(2) On the Submission Portal page, click on "Genomes (WGS & Complete)" on the left-hand side.
(3) Click the "New Submission" button at the top of the page.

## 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) In the first box, you will be asked if you have already registered a BioProject for this research. SRA reads should always be submitted prior to submitting WGS contigs. Select "Yes" and enter your existing BioProject.
(2) Next, you will be asked if you have already registered a BioSample for this sample. BioSamples should always be created prior to submitting WGS contigs (when you submit your SRA reads). Select "Yes" and enter your existing BioSample number.
(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.
(4) Next, you will enter Genome assembly metadata. Because we are submitting .fasta files, you will need to enter all information.

a. Assembly date: enter the date that the genomes were assembled. This is not required, but it is always good to provide as much information as possible.
b. Assembly method: This is a drop down menu. We generally use SPAdes. *Note: You can add another assembly method. This information can be found in the WGS_FMT spreadsheet in the folder for your given WGS run in the WGS folder on the server.
c. Version or date program was run: For examples, 3.8 or OCTOBER-2016, respectively. The version can be found in the WGS_FMT spreadsheet in the folder for your given WGS run in the WGS folder on the server.
d. Assembly name: This field is usually left blank.
e. Genome coverage: Enter the determined (not estimated) average coverage. The coverage for a given sample can be found in the WGS_FMT spreadsheet in the folder for your given WGS run in the WGS folder on the server.

f. Sequencing Technology: This is a drop down menu. You will select the relevant technology. For example, Illumina MiSeq. *Note: you can add another sequencing technology.

(5) You will be asked a few more questions:

a) Did your sample include the full genome? If you are submitting a draft genome select "Yes"

b. Is this the final version? Select "Yes" if you do not expect to do more sequencing or reassembly of this genome. You usually select "Yes".

c. Is it a de novo assembly? Select "Yes" unless you are comparing to a reference genome. We usually submit *de novo* assembled genomes.

d. Is it an update of existing submission? Select "No" unless you are updating an existing submission.

(6) There is an opportunity for additional comments at the bottom of the page. This is optional. When you have completed the General Info. section, click "Continue."

### 3.4. Source

(1) Here, you will enter the source the DNA is available from. We generally enter "Food Microbe Tracker, Cornell University"

(2) You will be asked if you would like the genome to be annotated in the NCBI Prokaryotic Annotation Pipeline. Click "Yes." It will generally take at least two weeks for your genome to be annotated in this pipeline. Failure to remove short (<200 bp) and contaminating (mapping to other organisms) contigs will result in substantial delays in annotation.

(3) Click "Continue" at the bottom of the page.

### 3.5. Files

(1) Here, you are asked which option describes your genome submission. Generally, we will select option 2. One or more chromosomes are still in multiple pieces and/or some sequences are not assembled into chromosomes. We select this because we are submitting multiple contigs.

(2) Select the file type ("FASTA")

(3) Upload your file.

a. Note that the maximum length of a sequence ID is 50 characters. You may need to shorten it. Otherwise, your sequence will not pass validation and you will be forced to resubmit.

(4) Do you have AGP files that assembled the individual contigs into scaffolds or chromosomes, OR assemble the submitted gapped sequences into chromosomes? Click "No".

(5) Click "Continue".

### 3.6. Assignment

(1) You will be asked two Yes or No questions. Answer accordingly.

a. Is any sequence a complete chromosome? If you are submitting a draft genome, the answer is "no".

b. Does any sequence belong to a plasmid? We usually do not have this information, and decide to answer “no”.

(2) Click “Continue.”

### 3.7. References

(1) Here, you will submit information for the reference.

a. Sequence authors: Enter your name. You can add multiple sequence authors if necessary.

b. Reference

1. Reference status: Select accordingly.

2. Enter a reference title (If it is a published paper, enter the paper title, if not, enter a descriptor).

Under reference authors, you can say “Same as sequence authors” or “Specify new authors”

(2) Click “Continue.”

### 3.8. Overview

(1) Review all information. If everything is satisfactory, click “Submit” at the bottom of the page. If there are problems with your submission, you will receive an email.

Note: You will likely receive an email re: contaminants. This is common. If the contaminated contigs are short (under 1000bp) you can delete them and re-upload your fasta file to WGS.

(2) You will receive an e-mail from NCBI with the subject “PGAP available…”. This e-mail has the link to two annotation files (*bgpipe.output.sqn and *bgpipe.output.gb). These files are made available for you in case you want to check their preliminary, automated, annotation. The annotation is reviewed manually by NCBI staff and may be edited by NCBI before it is made available. The e-mail asks you to REPLY to the e-mail stating whether you DO or DO NOT wish to modify the file they posted for your review. In the vast majority you should not make changes to the file and you should reply saying that you DO NOT wish to make changes. If you do need to make any change, make sure you make the changes in the .sqn file as stated in the e-mail.

## SECTION 4 TROUBLESHOOTING

Post-submission error message. Check if there are any contigs shorter than 200 bp or contigs that are mapping to other organisms than specified. Such contigs should be removed. If contigs >1000 bp are mapping to an unexpected source, discuss with Martin prior to deleting them and find a route cause.

If you the file is not uploading, try changing the internet browser. Google Chrome on Windows 10 seems to work well. Conversely, Edge on Windows 10 did not work as in November 2016.

## SECTION 5 REFERENCES

NCBI’s instructions on how to submit draft genomes are available on the following link:
https://www.ncbi.nlm.nih.gov/genbank/wgs.submit/