SFF and Library Metadata Files (July 12, 2010 snapshot of SRP002395)
Institute for Genome Sciences (IGS)

Author: Jonathan Crabtree
Version: 1.0
Effective Date:

1 Abstract

2 Introduction

This SOP describes the creation of .sff and .lmd files generated from the Human Microbiome Project 16S rRNA 454 Clinical Production Phase I (corresponding to SRA study accession number SRP002395).

3 Requirements

3.1 Program requirements

- *get_SRA_run_and_sample_xml.pl*
  This script is available for download at hmpdacc.org/HMR16S.

4 Procedure

4.1 SFF files

The 7518 .sff files (one for each of the 7518 SRA runs that are supposed to be associated with study SRP002395) in this directory were generated through the following process:

4.1.1. Download all 7518 runs in SRA native format from NCBI using Aspera client.

4.1.2. Convert all 7518 runs from SRA to SFF format using the "sffdump" utility from the NCBI SRA toolkit (the May 25, 2010 version, from http://www.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?cmd=show&f=software&m=software&s=software)

4.2 Library Metadata (.lmd) Files

The ".lmd" files are in an ad-hoc tab-delimited format and were generated by downloading and parsing the SRA XML files for all 7518 runs and the corresponding SRA samples using a custom Perl script, `get_SRA_run_and_sample_xml.pl`, available for download at hmpdacc.org/HMR16S.

This script takes as input a list of SRA Run accession numbers (i.e. ids of the form SRRxxxxxx) and does the following for each referenced Run:

4.2.1. Downloads the XML document containing the Run's metadata from the SRA into the current working directory.

4.2.2. Downloads the XML documents containing the metadata for any associated Samples from the SRA into the current working directory.

4.2.3. Parses and checks the Run and Sample XML files and produces a simple tab-delimited metadata file, which is written into the specified metadata directory (which must exist prior to running the script), with the file suffix ".lmd".

Steps 4.2.2 and 4.2.3 are not performed if the script is run with the retrieve_only option and the script will only download Run and Sample XML files that are not already present in the working directory. Also note that the parsing routines are specific to the SRA Studies that comprise the main HMP project and in some cases they may contain hard-coded workarounds that are not generally applicable to other studies and projects.

In addition to writing a series of .lmd metadata files into the named metadata directory the script also prints the following tab-delimited information to stdout:

1. SRA SRRxxxxxx accession number
2. Run Alias, as defined by the sequencing center
3. Instrument name (i.e., the sequencing machine used)
4. Center name (i.e., the sequencing center that generated the data)
5. dblock_filename (data block filename from the SRA metadata)
6. dblock member name (data block member name from the SRA metadata)

There should be exactly one .lmd file for each .sff file. Since most of the SFF files are already deconvoluted, these library metadata files contain quite a bit of duplicated information (see below).

Each tab-delimited row in one of the lmd files contains the following fields, in this order:

- SRA run accession (e.g., SRR012345)
- SRA experiment accession (e.g., SRX012345)
- run alias
- sequencing center
- experiment pool member_name (descriptor pulled from the XML that might serve as a library identifier)
- reverse barcode description
- reverse barcode sequence
- reverse primer description
- reverse primer sequence
- SRA sample accession (e.g., SRS012345)
- submitted anonymized subject id
- EMMES body site (with spelling errors corrected)
- submitted anonymized sample id

5 Implementation

6 Discussion

The crucial thing to note about the .lmd files is that each one typically contains a number of rows with the first field (SRA run accession) set to NULL and then one or more rows with the first field set to a non-NULL value. The rows with the initial NULLs enumerate all the samples for the specified **experiment** (typically corresponding to one or more 454 machine runs) and then the rows with the non-NULL SRA run accessions tell you which samples you should expect to see **in that particular SFF file**. So for most of the SFF files the corresponding .lmd file will start with a set of sample rows with NULL in the first column and then will have a single sample row with the accession of the SFF file in the first column. This is because the sequencing centers have already deconvoluted the data and so each SFF file downloaded from the SRA contains data from only one sample. However, for a subset of the SFF files, the SFF files in this directory are not fully deconvoluted (these used an alternate SRA submission format that encodes the deconvolution differently) and so the corresponding .lmd file will contain **multiple** rows with non-NULL initial fields, and these SFF files will have to be deconvoluted.

Caveats
Study SRP002395 has been in flux but at the time when this snapshot was made (July 12. 2010), it is believed to have the correct data for the Clinical Production Phase I 16S 454 sequencing. There was one spurious WGS Illumina run in the dataset on July 12, but that spurious run is NOT included in the set of files provided here.

7 Related Documents & References

Please e-mail jcrabtree@som.umaryland.edu with any questions and/or comments.

8 Revision History

Version	Author/Reviewer	Date	Change Made
1.0	Jonathan Crabtree	09/20/2011	Establish SOP