16S rRNA Processing
Institute for Genome Sciences

Author: Jonathan Crabtree

Version: 1.0
Effective Date: 06/16/2012

1 Abstract

Raw 16S rRNA sequence reads must be processed before they can be used to infer useful taxonomic information, something that typically entails either computing OTUs (Operational Taxonomic Units) or classifying reads against an existing taxonomy (or a combination of the two approaches.) A variety of algorithms can be used to infer taxonomic information and the processing and filtering requirements for the raw 16S sequences may depend on the particular choice of algorithm and/or approach. The processing and analysis procedures described in this SOP are intended to be a minimal baseline protocol that filters out as few 16S reads as possible and that also retains as much information as possible about each read in order to allow downstream analysis components to select the most useful subset of the available reads using whatever criteria are most appropriate.

2 Introduction

This SOP describes the baseline processing and analysis performed by IGS on the HMP 16S rRNA data. The Implementation section focuses on the processing of the Production Phase 1 (PP1) 16S rRNA data from the HMP Center “Healthy Cohort”, which is available at the NCBI Sequence Read Archive as Study SRP002395. However, the SOP has also been applied to the other 2 phases of the Healthy Cohort study: Production Phase 2 (PP2), SRP002860; and Pre-Production Pilot (PPS), SRP002012.

3 Requirements

Before applying the **Procedure** described in the next step, the 16S data to be processed--along with its accompanying metadata--must be available as a set of public (versus controlled-access) experiments from the SRA web site at http://www.ncbi.nlm.nih.gov/sra. For the HMP Center Healthy Cohort study, the approximate data flow between the sequencing centers and the SRA was as follows:

1. A sequencing center submits unfiltered sequence data and metadata to the SRA.
2. The sequence data appear initially on the SRA web site as a set of controlled-access Experiments. The sequence reads in these controlled-access Experiments may not have been completely filtered to remove human DNA sequences.
3. The SRA filters the sequence reads to remove human DNA sequences. For each controlled-access Experiment processed in this fashion a “shadow” public-access Experiment is created that links to the filtered read set (i.e., a copy of the original data in which all the sequence reads are still present, but possible human DNA sequences are masked by strings of Ns.)

4 Procedure

Please note that several of the following steps have had to be updated as the result of changes to the NCBI web services through which the SRA data and metadata can be accessed. The version of the procedure described herein is the one that was used to generate the original DACC/IGS analysis results for the Production Phase I 16S data in mid-late 2010 (i.e., the DACC-1.0 and DACC-1.1 SRP002395 result sets).

4.1 Download 16S data from the public SRA

1. An ad-hoc script (build_public_SRA_run_index.pl) was run to build a comprehensive index of the SRA's public FTP site and generate a mapping from SRA run accession number to the Aspera FASP URL of each corresponding directory on the SRA download site. (This step was necessitated by difficulties encountered retrieving FASP URLs directly from the SRA web site at the time the analysis was first started.)

2. The XML file describing the desired SRA study (SRP002395, "Human Microbiome Project 16S rRNA 454 Clinical Production Phase I") was manually downloaded from the appropriate SRA web service and then one-line Perl scripts were used to extract the "SRR" accession numbers of all the SRA runs belonging to the study, along with the SRA spot and base counts for those runs:

```
wget 'http://www.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?study=SRP002395&retmode=xml' \
-O SRP002395.xml

perl -ne 'while (/\<RUN acc=\"([^\"]+)\" spots=\"(\d+)\" bases=\"(\d+)\"/mg) \
{ print "$1\n"; }' < SRP002395.xml | sort > SRP002395-all-runs-from-XML.txt

perl -ne 'while (/\<RUN acc=\"([^\"]+)\" spots=\"(\d+)\" bases=\"(\d+)\"/mg)  \
{ print join("\t", $1, $2, $3) . "\n"; }' < SRP002395.xml \
> SRP002395-all-runs-from-XML-with-counts.txt
```

3. A wrapper script (get_SRA_runs.pl) was run to automatically download the SRA runs identified in step 2, using the command-line Aspera client (ascp) and the FASP URLs retrieved in step 1. The script automatically identified any missing runs (none, after correcting for some mislabeled data) and also automatically retried any failed Aspera transfers. 7518 runs were downloaded in this step.

4.2 Conversion from SRA to SFF format

1. The 1.0.0-b10 (May 25, 2010) release of the NCBI SRA Software Development Kit was downloaded from http://www.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?cmd=show&f=software&m=software&s=software and compiled and installed as per the included instructions.

2. A simple Perl wrapper script (sra2sff.pl) was used to run the NCBI-provided SRA to SFF conversion tool (sffdump) from the SDK on each of the 7518 SRA runs downloaded in the previous step.

3. Another script, check_SRA_spot_counts.pl, was used to cross-check the extracted spot counts reported by the sffdump tool against the spot counts parsed out of the study XML in a previous step, and to examine the stderr output of the sffdump tool for any other potential errors. 6 runs were found to have encountered problems in the sffdump step. 2 have since been corrected and reloaded by the SRA, although the status of the remaining 4 is still in question as of this writing (August 27, 2010). PLEASE NOTE THAT ALL DOWNSTREAM ANALYSES IN THE v1.0 DACC ANALYSIS HAVE THEREFORE TAKEN PLACE USING ONLY 7514 OF THE 7518 RUNS IN SRP002395. However, these 4 missing runs are supposed to contain only a relatively small number of spots/reads, as follows:

```
SRR043261 - 96 spot(s) expected

SRR043690 - 7367 spot(s) expected

SRR042823 - 1110 spot(s) expected

SRR042925 - 1820 spot(s) expected
```

As a consequence, only about 10,000 out of 72 million reads are missing from the inputs to the subsequent analyses (see below).

4.3 Conversion from SFF to FASTA and qual

1. The standard 454 utility sffinfo was run on each SFF file with the "-notrim" option to extract the complete FASTA sequence and quality score data for each run.

4.4 XML metadata download and flat file conversion

1. Another ad-hoc script (get_SRA_run_and_sample_xml.pl) was run to download all of the SRP002395 study's run and sample metadata in XML format from the SRA web site and parse out the crucial elements of the library construction metadata into an ad-hoc tab-delimited flat file format. The following fields are parsed out of the SRA Run and Sample XML files by this script into a set of tab-delimited ".lmd" files, one for each SRA Run:

- SRA run accession (e.g., SRR012345)
- SRA experiment accession (e.g., SRX012345)
- run alias
- sequencing center
- experiment pool member_name
- 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
  -note that the script removes the "G_DNA_" prefix and makes the
   following edits:
    Anterioir nares -> Anterior nares
    Attached/Keritinized gingivae -> Attached/Keratinized gingiva
-submitted anonymized sample id

4.5 Ad-hoc metadata correction

1. An initial attempt at verifying the deconvolution (by reverse barcode and primer) of the 7514 downloaded SRA runs revealed a number of discrepancies between the metadata parsed from the SRA XML files and the sequences actually present in the FASTA files extracted in a previous step. The deconvolution results were compared with the expected results from the SRA metadata files and--in consultation with the sequencing centers--the tab-delimited metadata files were updated accordingly.

4.6 Deconvolution [verification] and cs_nbp_rc subsequence extraction

1. A deconvolution script (get_SFF_clear_span_nbp.pl) was run on the FASTA file and tab-delimited .lmd metadata file corresponding to each of the 7514 SRA runs. In most cases the input FASTA file did not actually require deconvolution, as deconvolution by barcode had been performed by each sequencing center prior to SRA submission. In some cases, however, further deconvolution was necessary, either because:

   1. some submissions required looking at the primers in addition to the barcode to determine which window/variable region of the 16S gene had been sequenced (i.e., V13, V35, or V69) or
   2. some of the data were submitted to the SRA using an alternate metadata encoding, and we were not able to find an easy way to maintain the deconvolution information for these runs in the SRA -> SFF conversion step implemented using theNCBI sffdump tool. For each 16S sequence read the following(approximate) procedure was followed to extract the "cs_nbp_rc" sequence, if possible:

a. If the read does not start with an initial "TCAG" it is discarded.
b. If the next few bases of the read do not have an unambiguous needle(*) match to one of the expected reverse barcodes, allowing for at most one substitution or indel, the read is discarded.
c. If the next few bases of the read do not have an unambiguous needle(*) match to one of the 16S reverse primers that is supposed to accompany the barcode matched in step b., allowing for at most four substitutions and/or indels, the read is discarded.
d. The TCAG, reverse barcode, and reverse primer are removed from the sequence of the read. A search is also performed for the forward 16S primer and, if found, the forward primer and any sequence that follows it are likewise removed.

e. The remaining sequence span--from the end of the reverse primer to either the end of the sequence read or the beginning of the forward primer, whichever comes first--is intersected with the 454 clear range indicated by sffinfo.
f. If any sequence remains after step e. it is reverse complemented and output as the "cs_nbp_rc" sequence (cs = Clear Span, nbp = No Barcodes or Primers, and rc = Reverse Complemented.)
(*) needle = Needleman-Wunsch implementation from EMBOSS 6.2

Following this procedure, cs_nbp_rc sequences were generated for 99.4% (72043634) of the sequence reads in the 7514/7518 SRA runs for SRP002395. 0.1% of the sequence reads (73782) were filtered due to potential contamination from human DNA, and 0.4% of the reads (275856) lacked either a perfect match to the initial TCAG sequence or an unambiguous alignment with one of the expected barcode sequences.

4.7 [Partial] duplicate elimination

1. The resulting set of reverse-complemented clear_span_nbp sequences was split into 4 approximately equal subsets to help distribute the workload more evenly on the compute grid, and exact duplicates were eliminated from each of the 4 subsets, leaving around 65% of the sequence reads in each subset. Those unique reads were then passed along to the next step.

4.8 NAST-iEr, ChimeraSlayer, WigeoN, and RDP classifier processing

1. The unique sequence reads from the previous step were fed as input into an Ergatis (http://ergatis.sf.net) pipeline that ran the following analysis steps (plus some sequence file manipulations to break the input into appropriately-sized fragments):
   a. 16S reference alignment via the Broad's NAST-iEr alignment tool, using the clear_span_nbp sequences as input
   b. Chimera identification via the Broad's ChimeraSlayer, using the NAST-iEr alignments as input
   c. Aberrant sequence identification via the Broad's WigeoN, using the NAST-iEr alignments as input
   d. Taxonomic binning using the RDP classifier, using the clear_span_nbp sequences as input

Step a was run first and then steps b,c, and d were run in parallel (although d and a could also have been run in parallel.)

4.9 Duplicate restoration

1. All of the output files from the previous step were concatenated and run through a filter that "restored" the duplicate sequences by printing a duplicate line into the concatenated output file for each sequence that had been removed in the earlier duplicate elimination step. These concatenated output files with duplicates restored are what appear in this directory in gzipped form (see the "FILES" section of this document for a more detailed description of these output files.)

5 Implementation

5.1 Input Sequences

Please note the following caveats concerning the SRP002395 dataset used for this analysis:

1. 4 SRA runs containing a total of about 10,000 reads could not be successfully converted to SFF by the sffdump utility in the NCBI SRA SDK and have been excluded from this initial release of the analysis.
2. SRP002395 contains ALL of the WashU PPS reads (which are also present in the separate PPS study). To perform an analysis that ignores these reads you will need to filter out any and all reads with the following eleven 454 run ids (the 454 run id is a prefix of each individual sequence read identifier):
   F47LS8B02
   F48MJBB01
   F5MMO9001
   F5K51YR01
   F47USSH01
   F47USSH02
   F5K51YR02
   F47LS8B01
   F47543201
   F47543202
   F48MJBB02

5.2 Output Files

The following output files are produced by the SOP described herein:

SRP002395-7514-cs-nbp-rc.fsa.gz
Gzipped multi-FASTA file of reverse-complemented 454 clear ranges, with the following subsequences removed:
1. initial "TCAG" (must have been present in the original read),
2. reverse barcode sequence (must have been present in the original read),
3. reverse primer sequence (must have been present in the original read),
4. forward primer sequence (if present within the clear range.)

SRP002395-7514-rdp.results.gz
Concatenated and gzipped 'allrank' RDP classifier results for each of the sequences in SRP002395-7514-cs-nbp-rc.fsa.gz Version 2.2 of the RDP classifier was run using the default 032010 training set and taxonomy and the 'allrank' output format option.

SRP002395-7514-rdp.stdout.gz

Concatenated and gzipped stdout output from the RDP classifier runs; these files contain all of the warning messages about sequences that were too short to classify or, more generally, failed to have at least 42 (possibly overlapping) 8-mers that contain no Ns.

SRP002395-7514-nast-ier.NAST.gz
Concatenated and gzipped NAST-iEr results for each of the sequences in SRP002395-7514-cs-nbp-rc.fsa.gz run_NAST-iEr.pl from the 2010-04-29 release of the Broad microbiome utilities was run using the default parameters.

SRP002395-7514-cslayer.CPS.CPC.gz
Concatenated and gzipped .CPS.CPC ChimeraSlayer results for each of the sequence alignments in SRP002395-7514-nast-ier.NAST.gz ChimeraSlayer.pl from the 2010-04-29 release of the Broad microbiome utilities was run using the default parameters.

SRP002395-7514-wigeon.WigeoN.gz
Concatenated and gzipped WigeoN results for each of the sequence alignments in SRP002395-7514-nast-ier.NAST.gz run_WigeoN.pl from the 2010-04-29 release of the Broad microbiome utilities was run using the default parameters

6 Discussion

7 Related Documents & References

8 Revision History

Version	Author/Reviewer	Date	Change Made
1.0	Jonathan Crabtree	06/16/12	Establish SOP