**SOP# 04.001.01**

**Taxonomic identification of isolates (16S) - Greengenes**

**Author: Allison Perrotta**

**Date of Rev.: September 30, 2016**

**Page 1 of 3**

**Purpose**

This SOP describes the procedure used to identify the taxonomy of isolates using 16S Sanger sequences and the Greengenes database.

**Scope**

For exploratory purposes

**Regulatory References**

NA

**Responsibility**

* Responsibility of experimentalist: understanding and performing this procedure as described; reporting any deviations or problems to area supervisor; adequately documenting the procedures and results.
* Responsibility of area manager or supervisor: ensuring that the analyst performing this procedure is qualified; ensuring that the procedure is followed, and updating the procedure as necessary.

**Definitions/Abbreviations**

RDP – Ribosomal Database Projects

PCR – Polymerase Chain Reaction

**Related Documents:**

SOP# 02.355.01 – Sample preparation for GeneWiz Sanger sequencing – U515

SOP# 02.351.01 – Full Length 16S PCR preparation – 27F-1495R

SOP# 02.304.01 – DNA extraction from bacterial isolates – Alkaline PEG

**Required Equipment and Materials / Reagents**

* Computer capable of running Unix and bash commands from the command line
* Genegenes database reference file (version gg\_13\_5 or most recent version) containing the reference sequences as fasta files and the taxonomies of these sequences, downloaded at (<http://greengenes.secondgenome.com/downloads/database/13_5> or more recent version)
* Ncbi-blast+ module, available at  <ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/>

**Precautions**

N/A

**Procedure**

1. Once GeneWiz data is released, usually within 24 hours of submission:
   1. Download data from the GeneWiz website. Make sure to check any odd traces and exclude any samples that are true failures
   2. Copy data to desired directory on local computer
   3. Each reaction submitted, each reaction representing an isolate, will be provided as a separate sequence file.
   4. The vendor should provide information about the quality of these sequence files. The most common quality errors with this type of analysis are:
      1. Non-priming errors – This means that the primer did not hybridize to the template. This will result in very short sequence reads of less than 500 bases. In a successful Sanger sequencing reaction you can expect about 1,000 bases of sequence. To address this error, check the DNA concentrations of your extracted DNA and the PCR amplified 16S material you provided for sequencing, as one or both of these are most likely too dilute. Refer to the vendor’s submission requirements and re-submit (see SOP# 02.355.01 – Sample preparation for GeneWiz Sanger sequencing – U515).
      2. Mixed peaks – This occurs when there is more than one composition of sequence in the reaction. This is usually due to contamination but can also occur with mixed colony genetics. To address this error you must re- extract DNA from your stock (see SOP# 02.304.01 – DNA extraction from bacterial isolates – Alkaline PEG), prepare for sequencing (SOP# 02.351.01 – Full Length 16S PCR Preparation – 27F\_1492R), and re-submit (SOP# 02.355.01 – Sample preparation for GeneWiz Sanger sequencing – U515).
   5. Combine sequence files of good quality into one file using the following command from the command line:

cat \*.seq >> Name.fa

* 1. Align sequences to GreenGenes database to identify the taxonomy of each isolate:
     1. Load the NCBI BLAST module for aligning the isolate sequences to the database sequences
        1. module load ncbi-blast+
        2. makeblastdb -in gg\_13\_5/99\_otus.fasta -dbtype nucl -title gg99 -parse\_seqids
     2. Align the isolate sequences to the database sequences and track the best alignments or “top hits”
        1. $ blastn -db ${DB} -query sangerFile.fasta -out sangerFile.gg99.blastresults -evalue .000001 -max\_target\_seqs 1 -outfmt 6
        2. awk '!x[$1]++' sangerFile.gg99.blastresults > sangerFile.gg99.blastresults.tophit
     3. Parse the top hits to obtain only the best hit and the taxonomic name associated with that top hit in the database. This is done using a bash loop.
        1. cat ${OUT}/${NAME}.gg99.blastresults.tophit | while read line; do

SEQ=$(echo $line | awk '{print $1}')

GGHIT=$(echo $line | awk '{print $2}')

TAX=$(grep -w ${GGHIT} gg\_13\_5/99\_otu\_taxonomy\_sorted.txt

echo ${SEQ} ${TAX} >> sangerFile.gg99.blastresults.hits.names

done

* + 1. The output file sangerFile.gg99.blastresults.hits.names will contain information on the green genes reference identification number and taxonomy for the reference sequence that each isolate sequence aligned to best.

**Version History**

NA

**Worksheets**

NA

**Appendix**

NA