# README for Optional_Blast_Wrapper.pl

README generated by: Rishi R.Masalia, Burke Lab, Univ. of Georgia, August 2013
Corresponding email: jennifer.r.mandel@gmail.com or masalia@uga.edu

***Please read the General_Workflow_README file first.***

Goal: To automate the first steps in the Compositae-COS-Workflow. This is an optional wrapper and is here for convenience. This workflow can operate by simply performing the steps individually.

This wrapper will take raw sequence reads for a species or individual (fastq). Clean them, both reads 1 and 2 with prinseq-lite.pl (http://prinseq.sourceforge.net/manual.html). BLAST these reads against a database of your choice. Capture the top hit and create a file with the best hit and sequence, which will be assembled late in the Compositae-COS-Workflow.

Script Location:

Place this wrapper in the main directory

Output:

The output of this wrapper is TaxonName(R1 or R2).blasted_reads.fasta. These files will be located in the ./Output folder, and should be used when running Optional_Velvet_Prep_Wrapper.pl or when preparing your read files for Velvet Assembly.

Usage:

perl Optional_Blast_Wrapper.pl [parameters]

Parameters:

| Option/Flag | Description | Default |
|---|---|---|
| -name <file> | A names file to determine which files you want to run in this analysis. | - |
| -database <file> | Choose a database to BLAST against. | COS_sunf_lett_saff_all.fasta |
| -c <T/F> | Switch indicates that fastq file(s) is compressed via gzip. Toggling on, will have the program uncompress the file for you. If the file is compressed with something other than gzip, you have to uncompress the file yourself, through conventional methods. | –c F |
| -evalue <value> | Runs the BLAST with <value> cut off. | 0.00001 |

Name File Generation:

Generation of a "names" file signified by the, –name parameter. This is a separate file with a list of names (Taxon fastq files you wish to analyze). Be sure to make your actual file name ends in either *_R1 or *_R2 (for reads 1 and 2). Note that if your original fastq files are gzipped, the script will unzip them (if the –c T parameter is put). The .gz ending should be reflected in the "names" file. Also make sure your actual file names end in *.fq and NOT *.fastq

Example:

If your actual files are named: "HARG1.fastq" and "HARG2.fastq". They should be changed to "HARG_R1.fq" and "HARG_R2.fq". Additionally, make sure the filename ends in *.fq – however the names in the "name" file should not contain this 'fq' ending.

With the lines in the "names" file to read:
HARG_R1
HARG_R2

An example of this is provided on the GitHub, under
"Optional_Blast_Wrapper_Names_File_Example"

Choosing which BLAST database to use:

Please note that these are ALL raw fasta files. As such, these files DO need to be formatted via NCBI formatdb command as nucleotide databases (usage below).

There are three options here:

1. The default BLAST database, "COS_sunf_lett_saff_all.fasta", which is the source EST sequence data set that the probes were designed from.
2. The "COS_probes_blast.fasta" file which is a list of the 120mer probes/baits.
3. Finally, if you have your own BLAST database, you can choose to use that.

Formatdb Usage:

formatdb –i <database file> -p F

You specify which BLAST database through the –database <file> parameter. Note that if the parameter is not specified, our COS_sunf_lett_saff_all.fasta database will be used.

Programs Needed Prior to Running:

**Note: Blast and Prinseq-lite are ready-made programs we call in this workflow. As such, any questions regarding their operation or usage should be directed to their manuals, forums and help pages, or follow the necessary channels in contacting the authors of these programs.**

Note: All program executables should be placed in the ./programs folder, in the main working directory.

- Tophits.pl
  - Compositae-COS-workflow GitHub repository
  - This custom script can also be achieved by using any blast output parser followed by pairing the top hit information with the original cleaned fasta file.
- NCBI Blast 2.2.26
  - Download: http://www.ncbi.nlm.nih.gov/
  - Specifically the blastall and formatdb executables
- Prinseqlite.pl
  - Manual: http://prinseq.sourceforge.net/manual.html
  - Download: http://sourceforge.net/projects/prinseq/files/