Complete Genomes Information & Protocol

Instructions for the Pre-Processing & Guide Finder Scripts for Draft Genomes

The Guide Finder script is the only script necessary to design guides for complete genome annotations. The Guide Finder program is provided as an Rmarkdown file (.rmd). Open this file with RStudio. Before beginning, it is recommended that users read through these instructions and complete the example exercise (instructions located in Example folder). Line-by-line instructions are also provided in the pdf version of the Guide Finder Rmarkdown script. **Before beginning, download R and RStudio if you do not already have these programs.**

1. Identify Genbank Accession # for genome annotation.   
The GenBank number you will want to identify is listed as the INSDC number. For example, if I wanted to identify the genome for *S. mutans* UA159, I would use the INSDC number [AE014133.2](https://www.ncbi.nlm.nih.gov/nuccore/AE014133.2)

2. Download fasta file from Genbank.   
Download this file and store it in your working directory. Your working directory is a location where any input files should be held and is where any output flies created by the program will be saved to.

3. Within the Guide Finder script in RStudio, do the following:   
*Line-by-line instructions within the file*

A. Set working directory (where you FASTA file is saved and where files will output to) in “Set Working Directory”

B. Run the next three chunks of code (“Install Packages”, “Load Packages”, and “Functions Created”).

C. Set the path to makeblastdb for the MakeDatabase command in “Set BLAST Path & File”

D. After setting the path for makeblastdb, substitute the fasta file name for the name of your fasta file in the “Set BLAST Path & File” chunk

E. Set the path to blasn for the RunBlast command in “Set BLAST Path & File”

F. Input the fasta file in the “Input FASTA File” chunk. Input the fasta file by putting the name of your fasta file in the readDNAstringSet parentheses; file name should be in quotation

G. Define the putative promoter region by setting PromoterRegion to a number between 0-200. Set the desired putative promoter region, for example, setting PromoterRegion to 50 if you want the Guide Finder to search 50 bp upstream of the transcription start site for each gene to identify guides. If you do not want Guide Finder to make guides from the promoter region, set to 0. Do not set above 100 bp.

H. Set CompleteGenome option to TRUE in “Complete vs Draft Genomes”

I. In the “COMPLETE GENOMES ONLY” chunk, input the annotated GenBank file by putting accession # in GBAccession parentheses. Put the accession # in quotation.

J. Set parameters for guide design, including GC minimum, maximum distance from 5’ end, guides per gene, the minimum distance between guides (for paired guide design), bad-seed sequences, and restriction enzyme sequences in “Set Parameters for Guide Design”

K. To design guides for each gene in the genome, set Manual Selection to FALSE. To design guides for just a subset of genes, set Manual Selection to TRUE and add unique gene identifiers (for genes of interest) to GenesToSelect.

L. OPTIONAL: set lowered threshold parameters. Genes that did not produce guides with primary thresholds will be re-run through the program with these lowered thresholds. Also, set path for blastn for Re\_RunBlast and Re\_RunBlast\_Lowered.

M. Run the rest of the program!