**I. Guide Finder Utility:**

The Guide Finder program can be used to create any number of guides targeted to any number of genes for both complete and draft annotated microbial genomes.

**II. Inputs & Basic Workflow:**

The creation of guides for complete and draft genomes differs slightly (filtering and final selection steps remain the same):

**A.** Complete Genomes:

Inputs: FASTA file (single-sequence), GenBank Accession #

Workflow: For complete genomes, the GenBank Accession # is used to input the genome annotation information into the program. This information contains the start and end coordinates for each gene. These coordinates are then used to identify the coding sequence and putative promoter sequence (set by the user, up to 100 bp upstream of the start of the coding sequence) and select these sequences from the file. The promoter + coding sequence for each gene is then used to identify guides that can target within either the promoter or gene body. Within the gene body guides are designed to target the non-template strand while within the promoter guides are designed target either strand.

*In a handful of cases, users have reported an error when directly accessing genomes from GenBank with the readGenBank function; If you get an upload error, you can download the fasta file and coding sequence files from NCBI and treat them like a draft genome for the purposes of the script*

**B.** Draft Genomes:

Inputs: FASTA file (multi-sequence typically), Coding sequences file (coding sequence of each gene, 5’🡪 3’).   
Workflow: For draft genomes, the files must be pre-processed to put the information in the format required by the Guide Finder program, due to the variability of these types of annotations. Pre-processing can be done manually (if the user is working with a small subset of genes) or by the provided or custom scripts. For more information on the proper format of draft genomes required by the Guide Finder program, see the Draft Genomes Information document. The provided pre-processing script concatenates a multi-sequence fasta file into a single-sequence (adding N’s between contigs), parses the gene ID information from the coding sequence, identifies the gene coordinates using BLAST, and modifies the coordinates so that the smaller coordinate is always designated as the “start” coordinate (regardless of whether the gene is coded on the plus or minus strand). The output of the pre-processing script is a file that contains the gene ID, start and end coordinates, and notes whether the gene is coded on the plus or minus strand. This output file and the concatenated fasta file can then be read into the Guide Finder script for coordinate modification (inclusion of putative promoter sequence, up to 100 bp upstream), coding sequence identification, and guide design and filtering.

In summary, the Guide Finder program will create guides targeted to either the promoter or the gene body when complete or draft genome annotations are utilized. *Please refer to detailed use instructions for the type of annotated genome file you will be using (draft or complete)*

III. **User-defined Filtering Parameters**

Users can elect to filter based on the following parameters: Guide length, GC minimum, Distance from the TSS, # of Guide / Gene, Distance between guides (for paired guide design/simultaneous targeting), Bad seed sequences, and Restriction enzyme sequences

The following is a guide to set the guide design parameters (this information is also included in the Rmarkdown file):

**N:** Set minimum guide length (recommended 17-23); set as an integer

**GC\_Min:** GC minimum; set as a decimal (e.g. for GC min of 35%, set to .35)

**DistanceAsPercentageofCDS:** Maximum distance of guide from the transcription start site. As percentage of gene. For example, setting to .5 tell the program to look in the first 50% of the gene for guides

**GuidesPerGene:** Maximum number of guides/gene the program should output

**DistbtGuides:** Minimum distance between guides. For paired guide design. For example, using two guides to knockdown one gene in the same cell. In these cases, guides should be an appropriate distance apart. Overlapping guides have been shown to reduce extent of knockdown. *If only designing guides for single-guide knockdown, set DistbtGuides to 0*

**Test\_bad\_seed:** Set to TRUE or FALSE, if you want to turn this function on or off

**Bad\_seed:** If you want to remove sequences with a "bad seed", add the sequences here, in quotations

**Test\_RE:** Set to TRUE or FALSE, if you want to turn this function on or off

**RE\_site:** If you want to remove sequences with a "bad seed", add the sequences here

**IV. Output Files**

1. **CompleteGuidesList.csv.** Contains all potential guides for every gene, regardless of whether or not they meet set thresholds. Only use this list for reference if you want to see how many guides were created before filtering. Because these guides have not been filtered to remove those that don’t meet thresholds, *do not use this list to select guides to order.* These guides have also not yet been tested for off target effects.
2. **CompleteFilteredGuides.csv.** Contains all guides that have met thresholds. Since all of these guides have been filtered by thresholds, they all should be adequate guides. These guides also have been tested for off target effects. They have not yet been selected by distance/GC content. Can use this list as a reference or to manually select another guide if “best” guide created was not functional *in vivo.*
3. **TopHits\_Guides.csv.** A list of guides—the desired max number/gene set by the user—closest to the transcription start site.
4. **Pairwise\_Guides.csv:** A list of all possible guide pairs for each gene, for dual gene targeting. The minimum distance necessary between the guides is set by the user. Guides too close together (i.e. guide pairs that do not meet this minimum distance threshold) are filtered out.
5. **GenesWithoutGuide.csv.** A list of genes that did not produce guides with primary thresholds  
   Use this to identify genes that could not produce guides, given set thresholds.
6. **GuidesUsingLoweredThresholds:** A list of top hits guides for genes re-run with relaxed parameters.
7. **OverallGeneInfo:** Information about the number of genes that created guides (top hits versus paired guides) from primary thresholds and the number of genes that created guides with the lowered thresholds.
8. **Scrambles.csv –** A list of guides that do not have a match in the genome. Considered “scramble” guides that can be used as a negative control for CRISPR experiments.