Creating CRISPR sgRNA Libraries on CRISPOR

Introduction

This protocol gives instructions on how to generate a raw list of CRISPR sgRNAs, together with their respective scores, across a gene (or domain of interest in a gene) for saturating mutagenesis.

Materials

›

- .fasta file of gene

Procedure

Create a .fasta of your gene of interest

1. Find the **full** NCBI RefSeq ID of gene using Geneious and/or NCBI website

   Notes: you can also use other types of identifiers, as long as you change the track on UCSC Table Browser (more on this below). Please be sure to use the right ID for the **particular isoform** you are interested in—Uniprot usually has information on the canonical form of the protein. I also crosscheck at Ensembl.

   Avoid using gene names directly because they will generate a .fasta file with sequences of all of the isoforms

2. Go to USCS Table Browser to generate FASTA file of exonic sequences using the options as shown in the picture below:

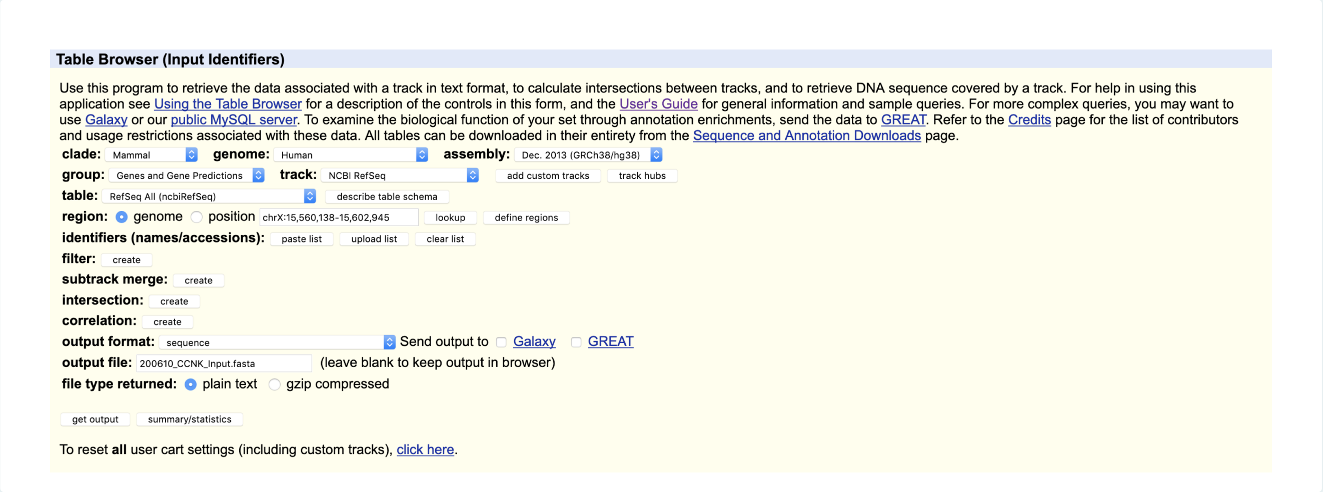


Table Browser (Input Identifiers)

Use this program to retrieve the data associated with a track in text format, to calculate intersections between tracks, and to retrieve DNA sequence covered by a track. For help in using this application see Using the Table Browser for a description of the controls in this form, and the User's Guide for general information and sample queries. For more complex queries, you may want to use Galaxy or our public MySQL server. To examine the biological function of your set through annotation enrichments, send the data to GREAT. Refer to the Credits page for the list of contributors and usage restrictions associated with these data. All tables can be downloaded in their entirety from the Sequence and Annotation Downloads page.

clade: Mammal **genome:** Human **assembly:** Dec. 2013 (GRCh38/hg38)

group: Genes and Gene Predictions **track:** NCBI RefSeq [add custom tracks] [track hubs]

table: RefSeq All (ncbiRefSeq) [describe table schema]

region: ● genome ○ position chrX:15,560,138-15,602,945 [lookup] [define regions]

identifiers (names/accessions): [paste list] [upload list] [clear list]

filter: [create]

subtrack merge: [create]

intersection: [create]

correlation: [create]

output format: sequence Send output to □ Galaxy □ GREAT

output file: 200610_CCNK_Input.fasta (leave blank to keep output in browser)

file type returned: ● plain text ○ gzip compressed

[get output] [summary/statistics]

To reset **all** user cart settings (including custom tracks), click here.

3. To specify your gene of interest, click 'paste list' and paste its **full** NCBI RefSeq ID into the text area.

   I highly suggest doing **1 gene at a time**—i.e. have 1 FASTA file for each gene

4. CRITICAL In the textbox next to 'output file', enter the input file name.

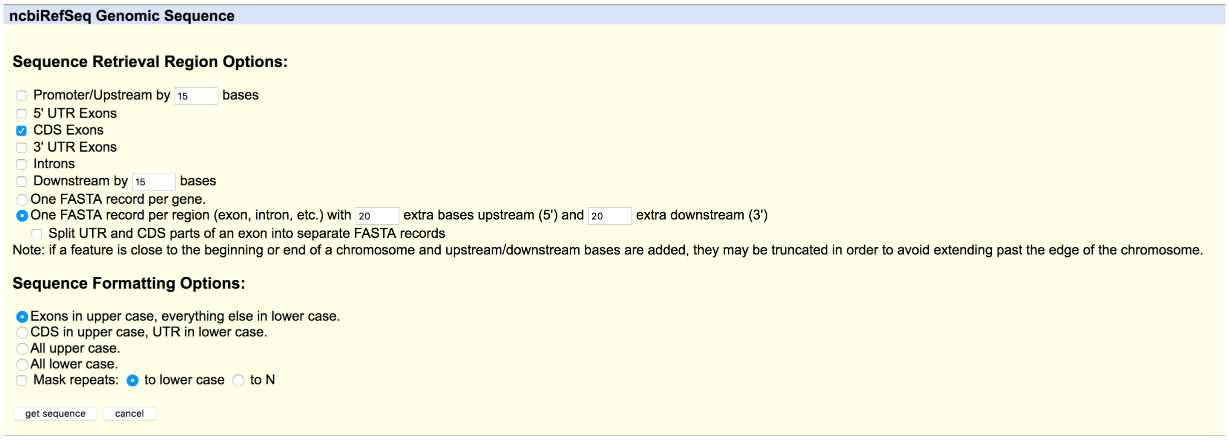
   Name your file in this format: YYMMDD_GENE_Input.fasta (case sensitive!!)
   e.g. 180807_DNMT1_Input.fasta or 181225_LSD1_Input.fasta

5. "Sequence type" for RefSeq All is: **Genomic**

   Note: there should only be one choice.

6. Fill out Sequence Retrieval Region Options (see picture below).

   Notes: I usually tick "CDS Exons" and select the option to have one FASTA record per region (this will make downstream analysis easier). I also do 20 extra bases upstream and downstream to facilitate generation of guides at the edge of the exons.



ncbiRefSeq Genomic Sequence

Sequence Retrieval Region Options:

□ Promoter/Upstream by 15 bases
□ 5' UTR Exons
☑ CDS Exons
□ 3' UTR Exons
□ Introns
□ Downstream by 15 bases
○ One FASTA record per gene.
● One FASTA record per region (exon, intron, etc.) with 20 extra bases upstream (5') and 20 extra downstream (3')
□ Split UTR and CDS parts of an exon into separate FASTA records
Note: if a feature is close to the beginning or end of a chromosome and upstream/downstream bases are added, they may be truncated in order to avoid extending past the edge of the chromosome.

Sequence Formatting Options:

● Exons in upper case, everything else in lower case.
○ CDS in upper case, UTR in lower case.
○ All upper case.
○ All lower case.
□ Mask repeats: ● to lower case ○ to N

7. Click 'get sequence' and save the .fasta file.

Generate library of gRNAs based on .fasta file submitted

8. Upload the .fasta input file into /n/holystore01/LABS/liau_lab/crisporPipeline/inputFolder

9. Edit the pipeline script (can be found in /crisporPipeline/) according to input file; memory request: #SBATCH --mem=300000 (units = MB); time request: #SBATCH -t 0-4:00 (D-HH:MM)

   Estimate ~1 GBs for each exon + 1 GB extra
   Estimate ~5 mins for each exon

10. Run the pipeline script: sbatch pipelineScript.sh

    test run: bash pipelineScript.sh

Installing environment (Source)

11. Download **crispor_python2_env.yaml** from crisporPipeline folder.

12. CRITICAL Open the .yml file. Change 'crispor_python2_env' in Line 1 to your name of choice (e.g. your_env)

    This is to prevent your created environment from accidentally overwriting the default one I have created.

13. Put the .yml file in your directory of choice.

    If you have changed the name, putting it in crisporPipeline works.

14. In your directory, type 'conda env create -f crispor_python2_env.yaml'

15. Change line 14 in pipelineScript.slum: source activate /your/directory/your_env

Running gRNA visualization tool

16. In the terminal, type *conda install bokeh*

    Before this step, you should have python 3 and anaconda/miniconda installed

17. Put main.py somewhere that is easy for you to navigate to

18. In the same directory, create a folder named 'data'

    For example, if you put main.py in ~/Users/your_name/, create ~/Users/your_name/data
    This folder is where all of the data files such as the fasta files and the tsv files will be placed

19. To run the tool, type in terminal: bokeh serve --show main.py

20. Note: I have also installed the tool in the group computer. See sticky note for instructions