Go Back

# CRISPR.sgRNA_create_ref_fasta, v2

Creates a multi-record FASTA file containing reference sgRNA sequences.

**Author:** Chet Birger;Broad Institute **Contact:** birger@broadinstitute.org **Algorithm Version:**

## Introduction

The CRISPR suite of GenePattern modules supports the computational processing of the data sets generated by CRISPR genome-scale functional screens.

In these screens:

- Cells are transduced with a library of lentiCRISPR vectors, each vector carrying the DNA sequence for one or a pair of sgRNAs.
- Within each infected cell, a Cas9:sgRNA complex generates a double stranded break (DSB) at the sgRNA's target locus and the cell's error-prone DSB repair mechanism will lead to a frame-shift indel and resulting loss-of-fuction mutation.
- Puromycin selection eliminates uninfected cells from the cell population. Following selection, DNA is extracted from the cell culture.
- The lentiCRISPR constructs integrated into infected cells' DNA are then amplified using PCR, and next generation sequencing produces FastQ files whose read records contain the read sequences associated with the transduced lentiCRISPR constructs.
- Through analysis of the read data, researchers can evaluate the representation of sgRNA or sgRNA pairings in the sequencing library, identifying selectively depleted or surviving sgRNAs in loss- or gain-of-function screens.

As indicated in the above, the CRISPR technology may be applied to both single-sgRNA and dual-sgRNA screens. Single-sgRNA screens are the most prevalent; see http://www.genome-engineering.org/crispr/. In dual-sgRNA screens, lentiCRISPR vectors carry two sgRNAs, and the functional screens can be used to study synthetic lethality and other forms of gene interaction. For both single-sgRNA and dual-sgRNA screens, the computational workflow begins with the fastq (or fastq.gz) files provided by the sequencing platform. In the case of single-sgRNA screens, there will be one fastq file per sample. In the case of dual-sgRNA screens, the sequencer platform delivers two fastq files from paired-end reads. The forward reads fastq file will contain read sequences for the first (in the 5' to 3' direction) sgRNA of the dual-sgRNA lentiCRISPR vector; the reverse reads fastq file will contain read sequences for the second (in the 5' to 3' direction) sgRNA.

Ultimately, the researcher wants to derive from these data files per-sample profiles of sgRNA (or sgRNA pairings in the case of a dual-sgRNA screen) depletion or survival. Starting with a csv file containing the list of sgRNA sequences and their IDs, and the fastq files, the following computational workflow will produce these profiles:

1. From a csv-formatted listing of sgRNA sequences represented in the lentiCRISPR library, create a reference FASTA file.
2. Trim FASTQ read records down to contain sgRNA sequence reads alone.
3. Align the trimmed reads to the reference FASTA using a short-read aligner (Bowtie1 or Bowtie2).
4. Tally the aligned, accumulating the read counts, and thus representation, of each reference sgRNA or sgRNA pair in the sequenced cell population.

We provide the following CRISPR GenePattern modules to support the above workflow:

- **CRISPR.sgRNA_create_ref_fasta** to create the reference FASTA (see step 1 above)
- **CRISPR.sgRNA_read_trimmer** to trim read records down to their sgRNA sequences (see step 2 above)
- **CRISPR.single_sgRNA_count** and **CRISPR.dual_sgRNA_count** to tally the aligned sgRNA read sequences (see step 4 above). CRISPR.single_sgRNA_count produces a two-column csv file, where the first column contains sgRNA identifiers, and the second column contains read counts for the respective reference sgRNAs. CRISPR.dual_sgRNA_count produces a three-column csv file, where the first two columns contain pairings of sgRNA identifiers, and the third column contains read counts for the respective pairings.
- **CRISPR.combine_csv_files** to combine csv-formated sgRNA counts from multiple samples into a single csv-formated dataset.

GenePattern supports several short read aligners. At the time of writing this documentation, GenePattern modules were available for BWA, Bowtie1, and Bowtie2. Any of these aligner modules may be used in step 3 above. Each aligner has its own companion indexer module, required to generate an index of the reference FASTA to which the trimmed reads will be aligned.

## Algorithm

The CRISPR.sgRNA_create_ref_fasta module takes as input a csv file, whose first column contains sgRNA sequences and second column contains sgRNA Sequence identifiers; e.g.,

```
CACCGTATATGCAATCGAAAGTGAC,ZNF217_4
CACCGGCTGCAGAACATGCAATCCA,ZNF217_3
CACCGCGGGGACTCGGAGACCGACC,YAP1_3
CACCGCATCAGATCGTGCACGTCCG,YAP1_1
CACCGATCAGATCGTGCACGTCCGC,YAF2_6
CACCGCGATCCACATTTTTCAATCT,YAF2_1
...
```

The resulting multi-record FASTA file generated by this module would be:

```
>ZNF217_4
CACCGTATATGCAATCGAAAGTGAC
>ZNF217_3
CACCGGCTGCAGAACATGCAATCCA
>YAP1_3
CACCGCGGGGACTCGGAGACCGACC
>YAP1_1
CACCGCATCAGATCGTGCACGTCCG
>YAF2_6
CACCGATCAGATCGTGCACGTCCGC
>YAF2_1
CACCGCGATCCACATTTTTCAATCT
...
```

## Parameters

| Name | Description |
|---|---|
| reference file * | CSV file containing sgRNA sequences and IDs. |
| basename * | Basename for the fasta file the module creates. |

* - required

This module is written in Python. The GenePattern server on which it is installed must have a custom configuration setting with name `python_2.7` whose value is set to the path of a python 2.7 interpreter. The module's python code imports tools from the Biopython package, which must be installed on the server's host system, along with the python 2.7.

## Platform Dependencies

| Task Type: | CPU Type: | Operating System: | Language: |
|---|---|---|---|
| CRISPR | any | any | python |

## Version Comments