CRISPR.combine_csv_files, v8

Combines per-sample sgRNA count files into a single csv file for downstream analysis.

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 a particular sgRNA, which guides the Cas9 nuclease to a specific genomic location. The Cas9:sgRNA complex will generate a double stranded break (DSB) at the targeted locus and the cell's error prone DSB repair mechanisms will lead to a frame-shift indel and resulting loss-of-fuction mutation. Puromycin selection eliminates uninfected cells from the 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 sequencers produce 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 each sgRNA in the sequencing library, identifying selectively depleted or surviving sgRNAs in loss- or gain-of-function screens.

Profiles of sgRNA depletion or survival can be obtained with the following computational workflow:

1. From a listing of sgRNA sequences represented in the lentiCRISPR library, create a reference FASTA file.
2. Sequencing data is provided as a collection of a FASTQ files, one (or one pair, in the case of paired sgRNA CRISPR screens) for each sample or time point. The FASTQ read records are trimmed down to contain sgRNA sequence reads alone.
3. The reads in the trimmed FASTQ file are aligned, using a short-read aligner like Bowtie or BWA, to the reference FASTA.
4. The aligned reads are tallied, accumulating the read counts, and thus representation, of each reference sgRNA 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 (step 1 above)
- CRISPR.sgRNA_read_trimmer to trim down read records to their sgRNA sequences (step 2 above)
- CRISPR.single_sgRNA_count and CRISPR.dual_sgRNA_count to tally the aligned sgRNA read sequences (step 4 above). CRISPR.dual_sgRNA_count supports CRISPR screens where the LentiCRISPR vector contains two sgRNAs, used in functional screens studying synthetic lethality and gene interaction. 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 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.

References

http://www.genome-engineering.org/crispr/

Parameters

Name	Description
csv files *	
output file basename *	

* - required

Input Files

1. csv files
   The list of csv files (one per sample) to be combined into a single csv-formatted dataset.

Output Files

1. <output file basename>.csv
   The csv-formatted dataset containing sgRNA profiles for multiple samples.

Requirements

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.

Platform Dependencies

Task Type:	CPU Type:	Operating System:	Language:
CRISPR	any	any	any

Version Comments

Version	Release Date	Description
8		
5		