Transposon-pipeline

User Guide

# Introduction

## Purpose

What is this program designed to do

## Requirements

POSIX system running Python 2.7, Bowtie2 and Samtools. Required Python packages can be installed via bioconda (a bioinformatics conda channel).

## Schematic

# MapFastq.py

## Purpose

To take concatenated FASTQ files, removes transposon and primer sequence (plus adaptor if set), and converts to a sorted BAM file. Program is written for use with the Berman lab transposon and primer, and for forward sequencing using the Illumina Universal adapter.

Add sequence info for both

## Default mode

Default settings require the input filename, assumes adapters are already removed, and will output to the current directory. (Note: Make sure you know whether or not adapters are present before running this program! Adapters can be identified using FastQC or other FASTQ file analysis programs.) Program is run from the terminal as follows:

python MapFastq.py -i ~/filepath/filename.fastq.gz

The program will output log information to the terminal giving the number of reads found in the FASTQ file, and the number (and percentage) of reads which contained the transposon with it’s sequence. Three files will be created in the output directory:

|  |  |
| --- | --- |
| filename.fastq.sorted.bam | Sorted BAM file |
| filename.fastq.sorted.bam.bai | BAM file index |
| log.txt | Logfile with same information seen in the terminal |

## Optional variables

|  |  |
| --- | --- |
| -o | Allows you to set the output directory. |
| -a | Cleans Illumina Universal adapter |
| -d | Deletes the original input FASTQ file |
| -k | Keeps the cleaned FASTQ files (files without transposon or adapter that are created and deleted during the program’s run) |
| -h | Shows help information |

# CreateHitFile.py

## Purpose

Explain all the info in the “hit file” (\*analysis.csv) and basics of other outputs . (Note: This hit file is the input for all later pipeline streams so keep track of it!)

## Default mode

Default settings will locate and use a sorted BAM file in the current directory, output the hit file to the current directory. Default minimum map quality threshold (the number of hits to parse form the BAM file for a hit to be counted) is 20. Default merge distance (merges two hits with *X* nt distance between them into a single hit at their midpoint) is 2 nt. Program is run from the terminal as follows:

python CreateHitFile.py

The program will output log information to the terminal giving the number of reads found in the FASTQ file, and the number (and percentage) of reads which contained the transposon with it’s sequence. Two files will be created in the output directory:

|  |  |
| --- | --- |
| filename.fsatq.sorted\_Hits.txt | Hit file |
| log.txt | Logfile with same information seen in the terminal. Will append to MapFastq.py’s logfile if it already exists in the output directory |

## Optional variables

|  |  |
| --- | --- |
| -i | Allows you to set the input directory. |
| -o | Sets the output directory |
| -q | Set minimum map quality |
| -k | Set maximum merge distance |
| -h | Shows help information |

# SummaryTable.py

## Purpose

Explain all the info in the “hit file”

## Default mode

Default settings will locate and use a hit file (or hit files) in the current directory, and output all results to the current directory. Default read depth below which insertion events will be ignored is 1. Program is run from the terminal as follows:

python SummaryTable.py

The program will output 27 files. Of these the most important are:

|  |  |
| --- | --- |
| filename.fastq.sorted\_analysis.csv | Analysis file with table for each *C. albicans* gene. Gives information detailed below. |
| hit\_map.filename.fastq.sorted.png | Image showing the hits within the *C. albicans* genome |
| log10\_read\_map.filename.fastq.sorted.png | Image showing log10 of the reads within the *C. albicans* genome |

### Analysis file

Explain all the columns here

### Additional file information

Explain other outputs here

## Optional variables

|  |  |
| --- | --- |
| Multiple hit files | Allows for statistical correlations to be found |
| -i | Sets the input directory |
| -o | Set the output directory |
| -f | Set the minimum read depth |
| -h | Shows help information |

# Classifier.py

## Purpose

Add info here

## Default mode

Explain default settings. Program is run from the terminal as follows:

python Classifier.py -???

Explain any terminal output. N files will be created in the output directory:

|  |  |
| --- | --- |
| File1 | Explanation |
| File2 | Explanation |

## Optional variables

|  |  |
| --- | --- |
| -o | Allows you to set the output directory. |
| -a | Cleans Illumina Universal adapter |
| -d | Deletes the original input FASTQ file |
| -k | Keeps the cleaned FASTQ files (files without transposon or adapter that are created and deleted during the program’s run) |
| -h | Shows help information |

# AnalyzeAssays.py

## Purpose

Add info here

## Default mode

Explain default settings. Program is run from the terminal as follows:

python Classifier.py -1st ~/filepath1 -2nd ~/filepath2

This program will create a single file in the output directory called “s\_score\_analysis.filename1\_vs\_filename2.csv”. This is an analysis file similar to that of SummaryTable.py but designed for comparison of key information between two hit files. There are two new column types in this table:

|  |  |
| --- | --- |
| S Value in filename | Explanation (Column appears twice: Once for each file being analyzed) |
| S score | Explanation |

## Optional variables

|  |  |
| --- | --- |
| -out | Allows you to set the output directory. |
| -h | Shows help information |

# DomainFigures.py

## Purpose

Add info here

## Default mode

Explain default settings and the region gene commands. Program is run from the terminal as follows:

python DomainFigures.py -???

Explain any terminal output. Figures will be created in the output directory for every ???. They follow the following naming scheme:

|  |  |
| --- | --- |
| Filename | Explanation |

## Optional variables

|  |  |
| --- | --- |
| ??? | Explain which ones are optional (and region gene commands?) |
| -h | Shows help information |