SigProfilerMatrixGenerator

Aug 28, 2018

INTRODUCTION

The purpose of this document is to provide a guide for using the sigProfilerMatrixGenerator framework to generate mutational matrices for a set of samples with associated mutational catalogues.

PREREQUISITES

The framework is written in PYTHON, however, it also requires the following additional software with the given versions (or newer) and access to BASH:

PYTHON	version 3.4 or newer
-PANDAS [MODULE]	version (any)
WGET	version 1.19

By default, the installation process will save the FASTA files for all chromosomes for the default genome assemblies (GRCh37, GRCh38, mm10). As a result, XXX Gb of storage must be available for the downloads.

QUICK START GUIDE

This section will guide you through the minimum steps required to create mutational matrices:

1. Run the install_catalogue_ref.py script in python3.

2. Once the installation is completed successfully, matrices can now be generated.

3. Create a new folder for each project/job that you run within the *references/vcf_files/* folder. Use a unique name for each project.

4. Place your *vcf* files within this new folder (*references/vcf_files/[project]/*).

5. From within the *scripts* folder, run the *sigProfilerMatrixGenerator.py* script in python3, specifying the reference genome, project name, indel (optional), and exome (optional) if desired:

   python3 sigProfilerMatrixGenerator.py –g GRCh37 –p BRCA –e -i

   *NOTE: only include –e if you would like to generate matrices using only mutations found within the exome, and only include –i if you would like to create the matrix for INDELS. (See available commands below).

6. The final matrices will be placed within the *references/matrix* folder. By default, the script will generate the matrices for 96, 192, 1536, 3072, and DINUC context.

COMMANDS

-g or --genome -> required: Followed by the reference genome (ex: GRCh37, GRCh38, mm10).
-p or --project -> required: Followed by the project name (ex: BRCA).
-e or --exome -> optional: Creates the matrix based solely on the exome.
-i or --indel -> optional: Creates the matrix for the limited list of INDELs (classifies insertions at micro-homologies as insertions at repeats of 0).
-ie or –extended_indel -> optional: Creates the matrix for the complete list of INDELs.

FOLDER STRUCTURE

The base framework consists of a *script*s folder, a *transcripts_original* folder, a *BRCA_example* folder, the *install_catalogue_ref.py* script, and this README file. The *scripts* folder contains all of the python scripts necessary to generate the reference files for the matrix generation as well as the *sigProfilerMatrixGenerator.py* script. The *transcripts_original* folder contains all protein-coding transcripts for each reference genome that is currently supported. The *BRCA_example* folder contains an example simple text file format (see EXAMPLE). These are necessary for generating the transcriptional strand bias matrices.

After running the *install_catalogue_ref.py* script, there will now be a *references* folder that contains all of the necessary reference files for creating the matrices. The user must place their input files within the *vcf_files* folder.

INPUT FILE FORMAT

This tool currently supports *maf, vcf,* and *simple text* file formats. The user must provide variant data adhering to one of these three formats. If the user's files are in *vcf* format, each sample must be saved as a separate file.

SIMULATING ADDITIONAL GENOMES

If the user desires to use a genome other than those currently supported (GRCh37, GRCh38, or mm10), they must:

1) Download the individual FASTA files for each chromosome
2) Place them into the *references/chromosomes/fasta* folder
3) Download the transcriptional data following the format presented in the *transcripts_original* folder.
4) Place this file within the *transcripts_original* folder
5) Download the exome ranges for the given genome and place in the *references/chromosomes/exome* folder.

EXAMPLE

Within the *references/vcf_files/BRCA_example/* folder, there is a simple text file (*BRCA_example_subs_simple.txt*) with four breast cancer sample and their associated mutational catalogues saved within a single file. To see an example of the output, run the following command from within the *scripts* folder:

python3 sigProfilerMatrixGenerator.py –g GRCh37 –t BRCA_example

The final matrices will be saved within the *references/matrix/* folder. To generate the INDEL matrix, place the INDEL *vcf* file within the *references/vcf_files/[project]/* folder and run the command:

python3 sigProfilerMatrixGenerator.py –g [genome] –t [project] -i

COPYRIGHT

CONTACT INFORMATION

Please address any queries or bug reports to Erik Bergstrom at ebergstr@eng.ucsd.edu