USER MANUAL

Version 1.0

(Genome Integration Sites Analysis Pipeline)

Saira Afzal

Molecular and Gene Therapy
Translational Oncology
German Cancer Research Centre (DKFZ)
National Centre for Tumor Diseases (NCT)

Supervisors: Raffaele Fronza
Manfred Schmidt

CONTENTS

1. Overview 3

1.1 Introduction 3

2. Installation 3

2.1 Availability & Implementation 3

2.2 Source code 3

2.3 Third-party tools 4

2.4 Perl modules 4

3. Input Files 4

3.1 Data files 4

3.2 Reference genome and index files 4

3.3 RefSeq table 5

3.4 LAM-PCR specific file 5

4. Configuration file 5

5. LAM-PCR mode 6

5.1 Testing 6

5.2 General Usage 8

5.2.1 Editing Tag table 8
5.2.2 Editing configuration file 8
5.2.3 Executing 10
5.2.4 Output 10

6. Targeted Sequencing mode 11

6.1 Testing 11

6.2 General Usage 14

6.2.1 Editing configuration file 14
6.2.2 Executing 15
6.2.3 Output 15

7. FAQ 17

8. References 19

GENE-IS

1. Overview

1.1 Introduction

Genome Integration Sites (GENE-IS) analysis tool has been designed to detect viral integration events in reference genome using high-throughput sequencing data. GENE-IS implements two basic analysis modes based on the experimental method used for data generation. These include 1) Linear Amplification Mediated (LAM) PCR data analysis mode 2) Targeted Sequencing (TES) data analysis mode. TES feature two further programs, which are a) Paired end TES mode and b) Single-end TES mode. GENE-IS has been tested for whole-genome sequencing data and targeted sequencing data. It can work with any available reference genome sequences.

The main focus behind GENE-IS development is to allow reliable and time-efficient characterization of intra-host viral junctions. It has been developed in a way to minimize the dependencies on the third-party tools and to allow easy and quick installation. Each module of GENE-IS is an independent program that performs a specific task.

It is implemented in Perl at Linux platform and is programmed in Perl, Python and Shell languages. There are a few third party tools that are required for executing pipeline and should be provided by user.

2. Installation

2.1 Availability & Implementation

GENE-IS is available at *https://bitbucket.org/gene-is1.0*
It has been tested at Ubuntu 10.10 and 12.04 LTES 64 bit operating system.
It requires following; Linux, Bash Shell, Perl v5.10.1 and Python 2.7.2.

2.2 Source code

Download GENE-IS by following these subsequent steps on your Linux operating system. (Time required ~ 3 min)

- Got to address *https://bitbucket.org*
- Provide username; *testergis* Password; *TesterGIS*
- Click on GENE-IS1.0
- Click on the left side at 2nd symbol from above
- Click on the clone option
- Copy the address that appears and paste into Linux terminal

hg clone https://TesterGIS@bitbucket.org/dkfzto/gene-is1.0

In the created "*GENE-IS*" directory, there are sub-directories including *lib, scripts and test*. The test directory contains *targetedSequencing* and *LAM-PCR* directories. GENE-IS contains three separate "configuration files" for LAM-PCR and TES paired end and single end integration sites analysis, respectively. User need to modify only these configuration files for analysis. GENE-IS provides options to adjust various explicit parameters according to user requirements.

2.3 Third-party tools

GENE-IS depends on several third party tools which are open source and are freely available. User needs to download and install them before running GENE-IS. Go to the links of individual tools, download and install these tools by following instructions in the related tool manual/web page. (For convenience these tools are also provided in GENE-IS_1.0, "ThirdPartyTools" folder. Please install them according to instructions mentioned on the below link)

Tool	Version	URL
BWA	0.7.4	*http://sourceforge.net/projects/bio-bwa/files/?source=navbar*
Bedtools	2.17.0	*https://code.google.com/p/bedtools/downloads/detail?name=BEDTools.v2.17.0.tar.gz&can=2&q=*
SAMtools	0.1.19	*http://samtools.sourceforge.net/*
BLAT	v.35	*http://users.soe.ucsc.edu/~kent/src/blatSrc35.zip*
Skewer	0.1.117	*http://sourceforge.net/projects/skewer/files/Binaries/*

2.4 Perl modules

The required Perl libraries are pre-packaged within the tool.

Two module, need to be installed by the user are; Bio::SeqIO and Bio::DB::Sam
These are usually available as a package on the Linux distributions.
However, can also be obtained from these links;
Bio::DB::Sam: http://search.cpan.org/~lds/Bio-SamTools/lib/Bio/DB/Sam.pm
Note: Please install SAMtools library before installing Bio::DB::Sam module.
Bio::SeqIO: http://search.cpan.org/~cjfields/BioPerl-1.6.924/Bio/SeqIO.pm

3. Input Files

3.1 Data files

GENE-IS takes as an input raw sequence reads in FASTQ format generated by LAM-PCR or TES experimental methods. In LAM-PCR and TES paired end mode, the inputs are forward and reverse raw read files. In case of single end TES mode, the input file is a single end forward raw read file.

3.2 Reference genome and index files

GENE-IS alignment module requires reference genome of the host as a single FASTA file (e.g., hg38), which also need to be indexed.

- For LAM-PCR mode user needs to create index files of reference genome for BWA aligner. First of all download FASTA sequence of your reference genome as a single FASTA file and run following command to create index files.

  /home/path-to-location/bwa-0.7.4/bwa index -a bwtsw /path-to-location/hg38.fa

  Create reference index files for BLAT by using this command;

/home/path-to-location/faToTwoBit referenceFileName.fa referenceFileName.fa.2bit

- For targeted sequencing mode of GENE-IS, create index files of reference genome plus viral genome together by BWA aligner and similarly by BLAT aligner as follows;
- First of all download FASTA sequence of reference genome and concatenate with FASTA sequence of one respective viral reference genome by using this command;
- *cat /home/path-to-location/referenceGenomeFileName.fa /home/path-to-location/viralGenomeFileName.fa > /home/path-to-location/referenceANDviral-FileName.fa*
- Run following command to create index files for BWA;
- */home/path-to-location/bwa-0.7.4/bwa index -a bwtsw /path-to-location/referenceANDviral-FileName.fa*
- Similarly, create reference index files for BLAT by using this command;
- */home/path-to-dir-blat/faToTwoBit referenceANDviral-FileName.fa referenceANDviral-FileName.fa.2bit*

3.3 RefSeq table

GENE-IS annotation module requires the refSeq table from UCSC for the related host reference genome. Download and provide a refSeq table for the similar version as of the reference genome (as used above in section 3.1)

- Go to the webpage of UCSC *http://genome.ucsc.edu/goldenpath/help/hgTablesHelp.html* and in *"Tools"* select *"Table Browser"* and for desired reference genome select the *"group; Gene and Gene Predictions"* and *"track, RefSeq Genes"* with *"table; refGene"* and *"output format; all fields from selected table"*. You can download it as gzip compressed file but unzip it before providing it as an input to GENE-IS.

3.4 LAM-PCR specific file

In LAM-PCR analysis, user needs to provide an additional table (tab separated file); we referred as experimental factor table or "tag table". This is a simple table that contains implicit parameters based on experimental data that are taken into account to perform analysis. Required parameters are fields/columns 6, 8, 11, 12, 17, 18, 21 in the same order as mentioned below;

Field1	***Field2***	***Field3***	***Field4***	***Field5***	***Field6***	***Field7***	***Field8***	***Field9***	***Field10***
NA	*NA*	*NA*	*SampleName*	*NA*	*5LAM/3LAM*	*NA*	*NA*	*NA*	*NA*

Field11	***Field12***	***Field13***	***Field14***	***Field15***	***Field16***	***Field17***
MiS_run	*ReverseFileBarcodeSequence*	*NA*	*NA*	*NA*	*NA*	*ForwardFileBarcodeSequence*

Field18	***Field19***	***Field20***	***Field21***
MegaPrimerSequence	*NA*	*NA*	*LinkerCassetteSequence*

4. Configuration file

GENE-IS has specific configuration files for each mode of analysis; LAM-PCR, TES paired and TES single end configuration files. Only the relevant configuration file should be modified for

particular analysis. User is required to provide path to the raw data files, reference genome index files, path to third-party tools and values for a number of parameters in the related configuration file. The template configuration files are provided and explained in the next sections 5.1, 5.2 and 6.1, 6.2 for LAM-PCR and TES analysis, respectively.

5. LAM-PCR mode

5.1 Testing

It is recommended that user first run GENE-IS analysis with test data set. To test LAM- PCR mode of GENE-IS we will use a dataset of 600 reads. Follow these instructions to complete testing process.

- Create a working/result directory for testing LAM by following this command.
  mkdir /home/path_to_location/gene-is1.0/test/LAM-PCR/results

- User need to modify certain lines in the configuration file. Therefore, open configuration file in GENE-IS directory "configFile_LAM-PCR_pairedEnd.txt" and modify following parameters accordingly. Here user only needs to change the path that is replace USER- PATH with their existing correct path.

```
#############################################################################
#######     Configuration File for LAM-PCR Sequencing                 ########
#######              Paired End Data Analysis                         ########
#############################################################################
#Path to script directory (script directory is in main folder of gene-is1.0)
scriptDir=/home/path_to_location/gene-is1.0/scripts
#Path to libraries containing directory (lib directory is in main folder of gene-is1.0)
libDir=/home/path_to_location/gene-is1.0/lib
#Number of possible parallel alignments
threads=8
#Number of reads in a split batch for parallel barcode sorting steps
group=2000000

#############################################################################
####### Experimental parameters (Barcodes sorting parameters)      ######
#############################################################################
#Choose one of these three options for lamType parameter "sorting|extracting|both"
#To perform only barcode sorting without IS analysis use "sorting"
#To perform only IS analysis without barcode sorting use "extracting"
#To perform barcode sorting and IS analysis use "both"
lamType = both
#Number of possible parallel barcode sorting
barThreads =8
#############################################################################
##                  Input data files and additional Parameters
#############################################################################
# Provide path to both forward and reverse FASTQ files and path to experimental factor/tag table
forward   = /home/path_to_location/gene-is1.0/test/LAM-PCR/r1.gz
reverse   = /home/path_to_location/gene-is1.0/test/LAM-PCR/r2.gz
tagTableName = /home/path_to_location/gene-is1.0/test/LAM-PCR/tagTable.txt

#provide run/sample name PREFIX that would be used as prefix for final result files
#Please do not include any space or strange characters
runName = testRunLAM
#Provide read ID prefix. Check in the raw fastq file and look at the header.
#Default is "@HWI". Maximum 4 characters
readHeader = @HWI
#############################################################################
#Quality filtration values (default value 30 or provide integer values only)
qual= 30
#Minimum length for reads
len=25
#Name used as prefix
lamPrefix=MiS
#User can choose alignment score threshold value, which is actually multiple hits score threshold.
```

```
#This is the value between primary alignment and secondary alignment for a sequence read.
alScore=0.95
#Minimum alignment identity percentage for re-alignment step with BLAT (default value 95)
minIden=95
#For topographical clustering that is genome IS position based clustering user can specify range of clustering
range=3
#User also has the option to provide value to "anchor" parameter, which is simply the LTR sequence that should be
#present and trimmed from the read
#anchor = CCACTCCCTCTCTGCGCGCT
#The mismatches allowed in anchor (frequency of error)
#anchorMM = 0.1

#This paramter allows user to decide maximum number of not-matched bases allowed before matching genomic part
#starts
#With default value this threshold is disabled (Recommended is 5 for vectors with no deletions)
#If  you are using this parameter then you must provide the complete LTR sequence in the above "anchor" paramter
notMatchThreshold=1000
##############################################################################
##      Reference fasta file and indexed files
##############################################################################
# The following parameter contain the alignment filename (DO NOT CHANGE)
alignmentOut = completAlignment

#Provide path to database with built-in index files. Choose path according to the aligner.
#(create indexed file as mentioned in section 3.1)
genomeVectorIndex=/home/path_to_location/hg38.fa
##############################################################################
##                  Third-party tools and files
##############################################################################
#Provide path to the BWA aligner
aligner     = /home/path_to_location/bwa

#Path to the secondary aligner. (BLAT)
blatAligner =  /home/path_to_location/blat
#Path to the BLAT indexed file of reference genome.
#(create indexed file as mentioned in section 3.1)
genomeVectorIndexBlat=/home/path_to_location/hg38.fa.2bit
##############################################################################
#Path to the trimming and filtering tool (Skewer)
skewer = /home/path_to_location/skewer
#Path to the Samtools
samtools= /home/path_to_location/samtools
#Path to the bedtools
bedTools= /home/path_to_location/bedtools
##############################################################################
#Files path required for annotation
#a complete UCSC refSeq table downloaded from UCSC (should be in text format)
#(download refSeq table as mentioned in section 3.2)
UCSCAnnoFile= /home/path_to_location/UCSC.anno.table_hg38.txt
##############################################################################
##############################################################################
```

- Save these modifications and close the "configFile_LAM-PCR_pairedEnd.txt" file.
- Type following command on terminal for changing directory to scripts

  cd /home/path_to_location/gene-is1.0/scripts
- Type following commands on terminal

  export GENIS=/home/path_to_location/gene-is1.0
- Run test suite by following command

  ./testGenis.sh

On the terminal will appear these options;

1) Targeted Sequencing Pair BWA 4) All
2) Targeted Sequencing Single 5) Clear
3) LAM-PCR 6) Quit

- Type at terminal *3* and press *enter*. Analysis will start and when it is finished type *6* at terminal to exit.

- Go to results directory by following command
  cd /home/path_to_location/gene-is1.0/test/LAM-PCR/results/

- Open "testRunLAM.GeneralStatistics.txt" file, if you see these following lines in your file, it means the analysis process completed successfully and LAM-PCR mode of GENE-IS is working properly.

##
GENERAL STATISTICS
##
Number of raw read pairs
600
Number of filtered and trimmed read pairs
500
Number of unclustered Integration Sites (sequence_count)
461
Number of clustered Integration Sites
461
##

5.2 General Usage

5.2.1 Editing Tag table

GENE-IS extracts information from the experimental factor table we refer as Tag Table for LAM-PCR data analysis. First to ensure that Tag Table is in correct format user need to run "buildTagTable.awk" program independently. This program takes a CSV Tag Table of 21 columns as an input and returns a tabular file of 6 columns that is used in order to perform a double barcode sorting (format of 21 column experimental factor/tag table as mentioned in section 3.3).

Execute this program by writing following command on terminal;

awk -f /home/path_to_location/gene-is1.0/scripts/buildTagTable.awk /home/path-to-location/TagTable21columns.txt

As an output, the generated file "tagTable.txt" contains LTR-barcode, linker barcode, linker barcode reverse complement, sample information, linker and megaprimer information. GENE-IS trimming module make use of linker and megaprimer information.

5.2.2 Editing configuration file

- For analyzing LAM-PCR based data user needs to modify the configuration file for the parameters mentioned above in section 5.1; these include path to scriptDir and libDir, forward file, reverse file and tagTableName. In addition, provide path to third party tools and create and provide reference genome index files specified for LAM-PCR mode along with UCSC refSeq table according to instructions in section 3.1 and 3.2, respectively.

- In addition to these, there are several other explicit parameters that user can adjust according to requirements or can continue with the default parameters. These include following;

- The threads parameter allows you t o choose number of possible parallel alignments. It has default value of 2. User should adjust it according to the available system memory etc.
  threads=2

- The group parameter allows you to choose number of reads in a split batch for barcoding. It has default value of 2000000. User should adjust it according to the available system memory etc. For example if system have 50 GB RAM, and if 3 people want to run barcode sorting step in parallel, then it would be better if they use 1 barThreads. As 2000000 batch takes 8 GB memory. So if we run with one thread, one individual will consume 8 GB, and 3 people will consume 24 GB. (If we use 2 barcode threads, then 3 people will consume 48 GB, that will be still ok, but for now we use 1 bar threads per person.)
  group= 2000000

- The "barThreads" parameter allows you to choose number of possible parallel sorting tasks for barcoding process. It has default value of 2. User should adjust it according to the requirements.
  barThreads=2

- User can choose one of these options for "lamType" parameter that is "sorting|extracting|both"
  To perform only barcode sorting without IS analysis use "sorting"
  To perform only IS analysis without barcode sorting use "extracting" To perform barcode sorting and IS analysis used "both"
  lamType = both

- There is also an option to choose minimum length for reads. Default length is 25.
  len=25

- User can choose alignment score threshold value, which is actually multiple hits score threshold. This is the value between primary alignment and secondary alignment for a sequence read with multiple hits. The default value of 0.9 means that all read sequences which have multiple hits with homology score of greater than 90 percent are discarded.
  alScore=0.9

- For topographical clustering that is clustering based on genome IS position clustering user can specify "range" of clustering. The default value of 10 means that all IS which are in +/- 10 range will be clustered together.
  range=10

Additional Parameters:

There are three additional parameters that are not required for analysis and by default are not taken into account for analysis. But if user wants then can mention these as well. As these are not required for analysis, so these parameters are commented if user needs to specify there values then user have to uncomment these in the configuration file.

- The parameter "anchor" is simply the LTR sequence that user wants that should be present and trimmed from the read.
  anchor = CCACTCCCTCTCTGCGCGCT (user specified LTR sequence)

- The mismatches allowed in anchor (frequency of error).
  anchorMM = 0.1

- The "notMatchThreshold" parameter allows user to decide maximum number of not-matched bases allowed before matching genomic part starts. With default value it is disabled. If this parameter is enabled, then user must provide the complete LTR sequence in the above "anchor" parameter (Recommended value is 5 for the vectors with no deletions).
  notMatchThreshold=1000

5.2.3 Executing

After modifying and saving changes in the configuration file run the following command on terminal to execute LAM-PCR mode of GENE-IS.

General Command:
perl -I <Perl library path> GENIS.pl -c <configuration file> -o <output directory>

Example:
perl -I /home/path_to_location/gene-is1.0/lib /home/path_to_location/gene-is1.0/scripts/GENIS.pl -c /home/path_to_location/gene-is1.0/configFile_LAM-PCR_pairedEnd.txt -o /home/path_to_location/resultsDir

5.2.4 Output

All intermediate files generated during the analysis process are stored in user specified working directory along with the result files.

There are seven basic result files generated in the user specified working directory at the end of analysis process.

- ✓ The first result file "ResultsClusteredAnnotated.csv" contains clustered insertion sites with orientation and sequence count information, nearby gene, refseq ID, gene name, orientation, distance to transcription start site, upstream or downstream distance, exon or intron details. The last column of this file contains sample name information and at the end of sample name there is the trimmed sequence related to the respective integration site.

  An example output:
  Chr , IS, Strand, Seq,_Coun,t RefSeq_ID, Gene_Strand, Gene_Name, Gene_Length, Dist_to_TSS, Upstream, Downstream, Intron_Exon
  chr1, 1318483, -, 1, NM_001256456, -, CPSF3L, 13103, 6204, , , NM_001256456_Intron6, SampleInfo
  chr1, 28918058, - ,1, NM_004437, +, EPB41, 232956, 30967, , ,NM_004437_Intron1, SampleInfo

- ✓ The result file "ResultsCompleteUnclustered.csv" contains un-clustered insertion sites with chromosome, genomic position, orientation, genomic span, sequence, sequence ID and sample name information.
  An example output:
  Chr Genomic_Position, Strand, Genomic_Span, Read_Sequence, Sequence_ID, Sample_Name
  chr1, 1318483, -, 72, , TTGGCTGAGGGTAGCT.., HWI-M00303:27:12.. , SampleInfo
  chr1, 28918058, - ,70, CCCGGCTGCAAATTAGCT.., HWI-M00303:27:12.. , SampleInfo

- ✓ Another result file "ResultsTenStrongestClones.csv" contains ten strongest clones.
  An example output:
  Chr IS Strand Seq_Count RefSeq_ID Gene_Strand Gene_Name Gene_Length Dist_to_TSS Upstream Downstream Intron_Exon
  chr1, 28421710, -, 99, NM_001048183, +, PHACTR4, 130788, 52128, , , NM_001048183_Intron2, SampleInfo
  chr1, 28918058, - ,85, NM_004437, +, EPB41, 232956, 30967, , ,NM_004437_Intron1, SampleInfo

- ✓ In addition, there is a complete file "wrongBC_.fastq" of all wrong barcode sequences detected during analysis process along with their sequence count.
  An example output:
  CTGCTGATCTGATCC CGTGGCACAGCAGTT 17
  CATCGCGACTGATCC CGTGGCACAGCAGTT 15
  CAGTGATCATGATCC CGTGGCACAGCAGTT 16
  TCTGATGAGTGATCC AGTGGCACAGCAGTT 11

- ✓ A general statistics file "GeneralStatistics.txt" provides basic statistics including number of raw reads, filtered-trimmed reads, detected integration sites, clustered integrations sites and there is additional information about the ten highest wrong barcode combinations detected.
  An example output:
  Number of raw read pairs
  10000
  Number of filtered and trimmed read pairs
  5974
- ✓ There are two additional result files that contain multiple aligned integration sites; "repeats.ResultsClusteredAnnotated.csv" file that contains clustered IS and another "repeats.ResultsCompleteUnclustered.csv" file with unclustered IS information.

6. Targeted Sequencing mode

6.1 Testing

To verify the download and installation process of TES mode of GENE-IS and third party-tools installation, there is TES specific test data sets provided in GENE-IS. It is recommended that user first run GENE-IS analysis with test data set.
To test TES paired end mode of GENE-IS we will use a one sample dataset of 500 raw reads. Follow these instructions to complete testing process.

- Create a working or result directory for testing TES by following this command.
  mkdir /home/path_to_location/gene-is1.0/test/targetedSequencing/results/pairedEnd/

- User need to modify certain lines in the configuration file. Therefore, open configuration file in GENE-IS directory "configFile_targetedSequencing_pairedEnd.txt" and modify following parameters accordingly. Here user only needs to change the path that is replace "/home/path_to_location/" with their existing correct path.

```
#############################################################################
#######    Configuration File for Targeted Sequencing        ########
#######  (SureSelect/Agilent) Paired End Data Analysis       ########
#############################################################################
#Path to script directory (script directory is in main folder of gene-is1.0)
scriptDir=/home/path_to_location/gene-is1.0/scripts
#Path to libraries containing directory (lib directory is in main folder of gene-is1.0)
```

```
libDir=/home/path_to_location/gene-is1.0/lib
#Data analysis type is Targeted Sequencing (SureSelect/AGILENT)(DO NOT CHANGE)
type=AGILENT
#Number of possible parallel
alignments threads=8
#############################################################################
##                  Input data files
#############################################################################
## Provide path to both forward and reverse FATESQ files
forward  = /home/path_to_location/gene-is1.0/test/targetedSequencing/r1.gz
reverse   =  /home/path_to_location/gene-is1.0/test/targetedSequencing/r2.gz

#Provide sample name PREFIX that would be used as prefix for final result files
#Please do not include any space or strange
characters sampleName = testDataTES
#############################################################################
##      Quality filtration and adapter trimming parameters
#############################################################################
#Quality filtration values (Use defaults values or provide integer values only)
qual= 20
#Use default Illumina adapters to remove from raw fastq forward and reverse files or provide your own
#Adapter to trim from forward file
adaptF = GATCGGAAGAGCACACGTCTGAACTCCAGTCAC
#Adapter to trim from reverse file
adaptR = AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCGTATCATT
#Use default output file names for forward and reverse files (DO NOT CHANGE)
suffOut = filtTrim
#############################################################################
##      Reference fasta file and indexed files
#############################################################################
# The following parameter contain the alignment filename (DO NOT CHANGE)
alignmentOut = completAlignment

##Provide path to the directory containing BWA aligner based index files and fasta file for refernce+vector genome
#(create indexed file as mentioned in section 3.1)
 genomeVector= /home/path_to_location/hg38_vectorSeq.fa

 #Specify the exact vector name that is mentioned in the reference/vector fasta sequence file
vectorString = vectorSeq
Minimum alignment identity percentage for re-alignment step with BLAT (default value 95)
minIden=95
#This is the value between primary alignment and secondary alignment for a sequence read (default value 0.9)
alScore=0.95
#For topographical clustering that is genome IS position based clustering user can specify range of clustering (default
#value 10)
range=3
#############################################################################
##                  Third-party tools and files
#############################################################################
#This program is in script directory of gene-is1.0. So provide path of gene-is1.0 script directory
 extractSClip = /home/path_to_location/gene-is1.0/scripts
#############################################################################
#Provide path to the BWA aligner
aligner     = /home/path_to_location/bwa
#Path to the secondary aligner. (BLAT)
blatAligner =  /home/path_to_location/blat
#Path to the BLAT indexed file of reference genome and vector.
#(create indexed file as mentioned in section 3.1)
genomeVectorIndexBlat=/home/path_to_location/hg38_vector.fa.2bit
#Path to the trimming and filtering tool (Skewer)
skewer = /home/path_to_location/skewer
#Path to the Samtools
samtools= /home/path_to_location/samtools
#Path to the bedtools
bedTools= /home/path_to_location/bedtools

#Files path required for annotation
#a complete UCSC refSeq table downloaded from UCSC (should be in text format)
#(download refSeq table as mentioned in section 3.2)
UCSCAnnoFile=/home/path_to_location/UCSC.anno.table_hg38.txt
#############################################################################
```

#For approximate IS extraction (TRUE/FALSE) (OPTIONAL)
approxIS=FASLE
#Provide path to individual separate BWA based indexed fasta files for vector and for reference genome respectively.
vectorIndexOut = /home/path_to_location/vector.fa
genomeIndexOut = /home/path_to_location/hg38.fa
##

#OPTIONAL
#For extra stringent filtering of IS reads TRUE/FALSE
#(Recommended only in cases where vector contains transgene which is highly homologues to reference genome region)
#(It can also cause loss of real IS)
extraFilt=FALSE
##
##

- Save these modifications and close the "configFile_targetedSequencing_pairedEnd.txt" file.
- Type following command on terminal for changing directory to scripts

  cd /home/path_to_location/gene-is1.0/scripts

- Type following commands on terminal

  export GENIS=/home/path_to_location/gene-is1.0

- Run test suite by following command

  ./testGenis.sh

  On the terminal will appear these options;

  1) Targeted Sequencing Pair BWA 4) All
  2) Targeted Sequencing Single 5) Clear
  3) LAM-PCR 6) Quit

- Type at terminal *1* and press *enter*. Analysis will start and when it is finished type *6* at terminal to exit.
- Go to results directory by following command

  cd /home/path_to_location/gene-is1.0/test/targetedSequencing/results/pairedEnd

- Open "testDataTES.GeneralStatistics.txt" file, if you see these following initial lines in your general-stats file, it means the analysis process completed successfully and TES mode of GENE-IS is working properly.

  ##
  GENERAL STATISTICS
  ##
  Number of raw read pairs
  500
  Number of filtered and trimmed read pairs
  500
  Number of correctly aligned vector-vector read pairs
  0
  Number of unclustered Integration Sites (sequence_count)
  490
  Number of clustered Integration Sites
  490

##

6.2 General Usage

6.2.1 Editing configuration file

- For analyzing TES based data modify the configuration file for the parameters mentioned above in section 6.1, these include path to scriptDir, libDir, and path to the third party tools. User needs to provide the complete path to the forward and reverse FASTQ files. In addition, create and provide correct and complete path to the index files of reference genome and UCSC refSeq table.
- The *genomeVectorIndex* file should be created as specified in section 3.1 for TES mode by BWA aligner. Provide exactly same value to the parameter *genomeVector* as of *genomeVectorIndex.* Similarly, create this *genomeVectorIndexBlat* index file as specified in same section 3.1 for TES mode by BLAT aligner. The *UCSCAnnoFile* parameters take *refSeq* table as input which is used for annotation. Therefore, the version of refSeq table should be the same as reference genome used for analysis. Download refSeq table as explained above in section 3.2.

  In addition, there are several other optional explicit parameters that user can adjust according to requirements or can continue with the default parameters. These include following;

- The threads parameter allows you to choose number of possible parallel alignments. It has default value of 2. User should adjust it according to the available system memory etc.
  threads=2

- User can choose values for quality filtration of their raw reads. Use default values or provide integer values only.
  qual=20
  (It is the minimum quality of bases required to keep. All bases less than this threshold of 20 are removed)

- Provide adapter sequences that are required to be trimmed from raw files. User should provide adapters for forward and reverse files for paired end mode of TES and only forward adapter for single end TES analysis. Otherwise, analysis will continue with default adapters.
  adaptF = GATCGGAAGAGCACACGTCTGAACTCCAGTCAC
  adaptR = AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCGTATCATT

- Specify the exact viral vector name that is mentioned in the header of vector sequence in the reference and vector FASTA sequence file
  vectorString = XXX

- User can choose alignment score threshold value, which is actually multiple hits score threshold. This is the value between primary alignment and secondary alignment for a sequence read with multiple hits. The default value of 0.9 means, that all read sequences which have multiple hits with homology score of greater than 90 percent are discarded.
  alScore=0.9

- For topographical clustering that is clustering based on genome IS position clustering user can specify "range" of clustering. The default value of 10 means that all IS which are in +/- 10 range will be clustered together.
  range=10

Optional extraction of approximate integration sites:
In case of paired end data, user can also detect approximate integration sites. These are integration sites where precise integration position cannot be mapped. One read of pair maps with the genome and the other read of pair maps with the vector.
approxIS=TRUE/FALSE
For approximate integration sites analysis, provide separately indexed BWA files for reference genome and vector genome (To create indexed files for BWA use the command mentioned in section 3.1).
vectorIndexOut = /home/path_to_location/path-to-viral-genome-indexed-file.fa
genomeIndexOut = /home/path_to_location/path-to-reference-genome-indexed-file.fa

Optional stringent filtering of integration sites:
User can choose the option "extraFilt" which is stringent filtering of integration sites. It is recommended only in cases where vector contains transgene which is highly homologues to reference genome region.
extraFilt=FALSE/TRUE

Specific requirement in case of Single End TES analysis:
In case of single end targeted sequencing analysis user also needs to provide value of these two parameters; path to the indexed FASTA file of vector genome and path to the indexed FASTA file of reference genome (To create indexed files for BWA use the command mentioned in section 3.1)
vectorIndexOut = /home/path_to_lcoation/path-to-viral-genome-indexed-file.fa
genomeIndexOut = /home/path_to_location/path-to-reference-genome-indexed-file.fa

6.2.2 Executing

After modifying and saving changes in the configuration file run the following command on terminal to execute paired end TES mode of GENE-IS. Similarly execute single end analysis by using configuration file for TES single end analysis.

General Command*:*
perl -I <Perl library path> GENIS.pl -c <configuration file> -o <output directory>

Example:
perl -I /home/path_to_location/gene-is1.0/lib /home/path_to_location/gene-is1.0/scripts/GENIS.pl -c /home/path_to_location/gene-is1.0/configFile_targetedSequencing_pairedEnd.txt –o /home/path_to_location/resultsDir

6.2.3 Output

All intermediate files generated during the analysis process are stored in user specified working directory along with the result files. The intermediate files generated during the analysis process can be useful to track analysis at various steps. However, it is recommended that user delete the intermediate files to avoid memory issues.

There are 7 basic result files generated in the user specified working directory at the end of analysis process.

- ✓ There is a complete file "ResultsCompleteUnclusterd.csv" that contains complete information about features of detected IS. These include read ID, chr, vector-name, genomic IS, vector IS, genomic and vector strand/orientation, soft-clip span (either vector or genome), soft-clip span identity percentage value (either vector or genome) and complete sequence of read that contains IS (vector+genome)

  An example output:

  SeqID Chr Vector Genomic_IS Vector_IS Genomic_Strand Vector_Strand SoftClip_PercentageIdentity SoftClip_Span Sequence

  HWI-ST486:10781:25275 chr9 XXX 4423457 3292 + - 100.00 33 CTAGAGCTCGCTGATCA…

  HWI-ST486:11038:38683 chr2 XXX 164874989 1090 + + 100.00 63 CGCCGACCAAAGAAG…..

- ✓ The second result file "ResultsClusteredAnnotated.csv" contains clustered integration sites with orientation and sequence count information, nearby gene, refseq ID, gene name, orientation, distance to transcription start site, upstream or downstream distance, exon or intron details.
  - An example output:

  Chr IS Strand Seq_Count RefSeq_ID Gene_Strand Gene_Name Gene_Length Dist_to_TESS Upstream Downstream Intron_Exon

  chr1, 12257159, +, 1, NM_015378, +, VPS13D, 282008, 27120, , ,
  NM_015378_Intron9 chr1, 43934993, +, 4, NM_001136215, +, ARTN, 3921, 1673, ,
  , NM_001136215_Intron1

- ✓ Another result file "ResultsTenStrongestClones.csv" contains ten strongest clones.

  An example output:

  Chr IS Strand Seq_Count RefSeq_ID Gene_Strand Gene_Name Gene_Length Dist_to_TESS Upstream Downstream Intron_Exon

  chr8, 19961021, +, 75, NM_000237, +, LPL, 28189, 21950, , ,
  NM_000237_Exon chr8, 19948221, +, 73, NM_000237, +, LPL, 28189, 9150,
  , , NM_000237_Exon2

- ✓ A general statistics file "GeneralStatistics.txt" provides basic statistics including number of raw reads, filtered-trimmed reads, vector-vector/vector reads, detected integration sites, clustered integrations sites. There is detailed information about the filtering and trimming process and basic alignment statistics are also provided.

  An example output:

  Number of raw read pairs;
  100000
  Number of filtered and trimmed read pairs;
  99958

- ✓ There are two result files that contain multiple aligned integration sites. These include "repeats.ResultsClusteredAnnotated.csv' and "repeats.ResultsCompleteUnclustered.csv".

- ✓ The optional approximate IS analysis leads to additional two result files including "approx.ResultsClusteredAnnotated.csv" and "approx.ResultsCompleteUnclustered.csv".

7. FAQ

a) What kind of data can be used as an input for GENE-IS?

GENE-IS can be used to analyze any kind of whole-genome sequencing data or targeted sequencing data for characterizing viral-host events.

b) What should be the format of raw files?

The input files for targeted sequencing (TES) mode should be gzipped "gz" fastq files paired end or single end. In case of linear amplification mediated (LAM) PCR, the input should be gzipped "gz" fastq file containing single/multiple samples for paired end analysis.

c) Can GENE-IS be used for paired and single end data?

TES mode of GENE-IS accepts paired and single end data files for analysis. There are already separate configuration files for each type of analysis. Whereas, LAM-PCR mode supports currently only paired end data files.

d) Can I use any reference/host genome?

You can use any reference genome and any viral sequence for IS detection. Just make sure to download reference genome in correct format. Create index files of corresponding reference sequence and ensure correct path in the configuration file also (For details see section 3.1).

In addition, download related refSeq table of same specie as of reference genome and the version of refSeq table and reference genome should be the same. (For details see section 3.2)

e) Can I use more than two viral vectors as reference?

GENE-IS works only with one viral vector at a time. If you want to analyze same data with more than one vector genome, then you have to re-run analysis with desired vector.

f) How I can be assure that GENE-IS and its dependencies has been installed properly?

Users are strongly recommended to test GENE-IS with provided datasets after obtaining source code and installation of third party tools. There are separate data sets for testing TES and LAM-PCR modes of analysis. Testing of both analysis types takes less than 30 minutes (for details see section 5.1 and 6.1)

g) How can I speed up analysis?

In order to speed up analysis user can adjust the "threads" parameters according to their requirements (see section 5.2.2 and 6.2.1). Please make sure you have sufficient space/memory available before starting your analysis.

h) Why I see the error "file does not exist"?

The most common reason for this error is the wrong path provided by user in the configuration file. You have to check all file names and paths in your configuration file. Linux is case sensitive and any file names with strange characters or spaces in between file names are most common cause of errors.

i) Why GENE-IS reported no integration sites in TES mode, but I am sure about integration sites in my data?

In the configuration file for TES, re-check the name of the parameter " vector_string ".

The value of this parameter is the name of the vector that you have used in your reference file to create index. Any case change or other errors in vector-string name would lead to failure of IS detection.

For example, if in your reference FASTA file of reference genome and vector sequence, the vector sequence name is *" >PVD191_AIP "* then in " vector_string " parameter you should provide " *PVD191_AIP " as value.*

In addition, make sure that that in reference genome file the chromosome names should have prefix " chr " and should appear in reference file as e.g., " >chr1 ".

j) **What should I do with intermediate files?**

Intermediate files are useful to track different steps of your analysis. But, we strongly recommend users to delete these files after analysis, as they consume huge amount of space.

8. References

1. Schmidt,M., Schwarzwaelder,K., Bartholomae,C., Zaoui,K., Ball,C., Pilz,I., Braun,S., Glimm,H. and von Kalle,C. (2007) High-resolution insertion-site analysis by linear amplification-mediated PCR (LAM-PCR*). Nat. Methods,* **4,** 1051–1057.
2. Ustek,D., Sirma,S., Gumus,E., Arikan,M., Cakiris,A., Abaci,N., Mathew,J., Emrence,Z., Azakli,H., Cosan,F. *et al.* (2012) A genome-wide analysis of lentivector integration sites using targeted sequence capture and next generation sequencing technology. *Infect. Genet. Evol.,* **12,** 1349-1354.
3. Li,H. and Durbin,R. (2009) Fast and accurate short Burrows Wheeler transform. *Bioinformatics,* **25,** 1754-1760.
4. Kent,W.J. (2002) BLAT-the BLAST-like alignment tool. *Genome Res.,* **12,** 656-664.
5. Karolchik,D., Hinrichs,A.S., Furey,T.S., Roskin,K.M., Sugnet,C.W., Haussler,D. and Kent,W.J. (2004) The UCSC Table Browser data retrieval tool. *Nucleic Acids Res.,* **32** (Database issue), D493–D496.
6. Quinlan,A.R. and Hall,I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26,** 841-842.
7. Jiang,H., Lei,R., Ding,S.W. and Zhu,S. (2014) Skewer: a fast and accurate adapter trimmer for next-generation sequencing paired-end reads. *BMC Bioinformatics*, **15,** 182.
8. Li,H., Handsaker,B., Wysoker,A., Fennell,T., Ruan.J., Homer.N., Marth,G., Abecasis,G., Durbin,R. and 1000 Genome Project Data Processing Subgroup. (2009) The Sequence alignment/map (SAM) format and SAMtools. *Bioinformatics,* **25,** 2078-2079.