Introduction

The Integration Analysis System (IAS) is a set of three software tools that are designed to take raw sequence data derived from cell-based phenotypic screens using the Sleeping Beauty transposon mutagenesis system, as described by Feddersen et al. (Feddersen et al. 2019). This study describes a ligation-mediated PCR approach that generates two independent libraries for each sample analyzed corresponding to the genomic/transposon junction derived from the left (IRL) and right (IRR) inverted repeats of the transposon. After multiplexing, these libraries are then sequenced using a paired-end approach on the Illumina platform, ultimately producing four FASTQ files for each sample (*i.e.* forward and reverse reads for both IRL and IRR). The IAS is designed to take these raw sequence files as input and produce a list of candidate genes that are associated with the phenotype of interest. This document provides an overview describing how each tool in IAS works in this process.

IAS_mapper.py

This tool takes a sample input file as the input (see IAS_user_manual for details). This file contains the necessary FASTQ filenames and inputs for each ligation-mediated PCR (LM-PCR) library to be analyzed. Consider the following example transposon inserted on chromosome 9 at position 9,687,810 (GRCh38):

Based on the structure of this insertion event, the sequences present in the LM-PCR libraries should appear as follows:

IRL forward read:

5′ **TGTAAACTTCCGACTTCAACTG**TACATGTATGCACACACAGAT**GTCCCTTAAGCGGAGCCCTA** 3′

IRL reverse read:

5′ **TAGGGCTCCGCTTAAGGGAC**ATCTGTGTGTGCATACATGTA**CAGTTGAAGTCGGAAGTTTACA** 3′

IRR forward read:

5′ **TGTAAACTTCCGACTTCAACTG**CTACTTATGAGATAATTATATT**GTCCCTTAAGCGGAGCCCTA**

IRR reverse read:

5′ **TAGGGCTCCGCTTAAGGGAC**AATATAATTATCTCATAAGTAG**CAGTTGAAGTCGGAAGTTTACA** 3′

transposon tag
adaptor tag
genomic sequence

These sequences will be processed in the following steps:

1) Removal of all transposon tags via cutadapt
2) Removal of all adaptor tags via cutadapt
3) Map genomic sequence to reference genome via HISAT2 → creates SAM file format
4) Removal of all transposon tags via cutadapt
5) Removal of all adaptor tags via cutadapt
6) Map genomic sequence to reference genome via HISAT2 → creates SAM file format
7) Parse IRL and IRR SAM files to create UNIQ file for the sample

These steps are repeated until all samples contained in the sample input file are processed.

UNIQtoGFF3.py

This tool filters data from UNIQ files to remove low abundance and low confidence insertion sites prior to analysis using the gCIS2 tool. This tool is based on prior work showing that the read depth for specific transposon insertion events is a semi-quantitative measure of the cell number that carries the transposon insertion event (Brett et al. 2011). In this sense, the UNIQtoGFF3 tool allows the user to restrict the downstream search for candidate genes to those transposon insertion events that are present in a high number of cells (*i.e.* clonally expanded cell populations).

This tool requires three parameters as input:

1) The normalized abundance* required to retain an insertion where reads were recovered from both the IRL and IRR ends of the transposon (suggested setting = 1)
2) The normalized abundance* required to retain an insertion where reads were recovered from only one end of the transposon (suggested setting = 5)
3) The minimum read number required to retain any insertion site (suggested setting = 10)

* The normalized abundance represents the proportion of the reads for any site relative to the most abundant site detected for the sample. For example, a normalized abundance value of 100 indicates that the insertion site is the most abundant site for the sample while a value of 1 indicates that the insertion site is present at 1% of the most abundant site in the sample.

gCIS2.py

This tool analyzes a combined GFF3 file to identify genes that have suffered transposon insertions at a rate significantly higher than expected given the characteristics of the input data. This tool requires a variety of prebuilt Python dictionaries that are included with the IAS within the GRCh38 directory. Currently, these dictionaries exist only for the GRCh38 build of the human reference genome.

The gCIS2 tool requires three parameter inputs:

1) Filename of the combined GFF3 to be analyzed
2) The promoter size to be considered during the analysis of each gene
3) Gene annotation set to be used (ex: refseq, ensemble, ucsc)

Here is the order of processes used by the gCIS2 tool to identify candidate genes:

1) Determines the number of insertion events and samples contained in the GFF3 input file
2) For each gene in the annotation set:
    a. determine gene boundary (+/- promoter)
    b. determine # of TA sites within the boundary (+/- promoter)
    c. determine # of insertions within the boundary (+/- promoter)
    d. determine probability of observed # of insertions within boundary (+/- promoter)
    e. determine kurtosis of insertions within boundary, including promoter
    f. determine skewness of insertions within boundary, including promoter
3) For each gene in the annotation set with at least 2 insertion events, performs chi-square test to determine significance *[expected number of insertions = (# of TA sites in gene (+/- promoter) / # of TA sites in genome) * # of total insertion events]*
4) Determines false discovery rate for genes with chi-square result
5) Evaluates each gene with five or more insertion events:
    a. If ≥ 75% of insertion events are in the same orientation as the gene AND the FDR for the gene is ≤ $1x10^{-5}$, prediction is "over-expression"
    b. If ≤ 75% of insertion events are in the same orientation as the gene AND ≥ 90% of insertion events are within the gene body AND the FDR for the gene is ≤ $1x10^{-5}$, prediction is "disruption"
    c. Otherwise the prediction is "false-positive"
6) Generates three output files:
    a. *_parameter_settings.txt = contains the input parameters used for the analysis
    b. *_all_genes.txt = contains results for all genes with at least 1 insertion event detected
    c. *_filtered_genes.txt = contains results for only genes with a prediction of "over-expression" or "disruption"

References

Brett BT, Berquam-Vrieze KE, Nannapaneni K, Huang J, Scheetz TE, Dupuy AJ. 2011. Novel molecular and computational methods improve the accuracy of insertion site analysis in Sleeping Beauty-induced tumors. *PLoS One* **6**: e24668.

Feddersen CR, Schillo JL, Varzavand A, Vaughn HR, Wadsworth LS, Voigt AP, Zhu EY, Jennings BM, Mullen SA, Bobera J et al. 2019. Src-dependent DBL family members drive resistance to vemurafenib in human melanoma. *bioRxiv* doi:https://doi.org/10.1101/561597.