The pipeline established filterers, aligns, and counts the raw reads to output potential Sleeping Beauty transposon insertion sites in the zebrafish genome. The output from the pipeline is informative as it makes a predictive and testable scenario.

**Methods**

**Filtering the raw data**

Since the Sleeping Beauty transposon only inserts itself in regions of the genome that contain the TA nucleotides, reads that contained the transposon arm with “TA” adjacent to them were selected for using “more file.fastq | grep –A1 –B2 “TGTATGTAAACTTCCGACTTCAACTGTA”. Reads where then sorted by barcode using the Barcode\_Splitter feature of FASTX. Each barcode was 6bp in length and there were a total of 24 barcodes. The transposon sequence was then trimmed off using Trimmomatic, leaving only the genomic sequence for alignment. Lastly, the reads had to be converted from fastq to fasta format in order to be compatible with the Bowtie alignment.

**Alignment to zebrafish reference genome and quantification.**

Bowtie was used for the filtered read alignment, since it is not a spliced aligner. The parameters used were based off the (Brett et al. 2011) study “-best –f –k2 –p7 –v3”. This gives the top two hits and only allows 3 mismatches in the read sequence. The alignment output was in SAM format and was filtered using the following criteria: the best match had to be at least 90% identical, including a perfect match to the ‘‘TA’’ at the start of the alignment, be at least 5% better than the second best match, and have 2 or less mismatches. The reads were counted and sorted using various unix commands to determine insertion frequency and location in the genome. The result of these commands outputs a 4 column file (Read\_Count | Insertion\_Location | Chromosome\_Number | strand aligned too). Files used to generate this data were “Transposon Mapping Script” and “Transposon Mapping companion script “.

Output:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Read Count** | **Location** | **Chromosome #** | **Alignment Orientation** |  |
| **186** | **37368708** | **1** | - |
| **86** | **40387948** | **5** | **+** |
| **85** | **22060064** | **20** | **+** |
| **62** | **40461110** | **4** | + |
| **48** | **37318502** | **18** | - |

**Annotation of alignment results.**

The “Java Annotation Script” takes in this 4 column file and adds in valuable information such as gene name, gene ID, gene feature (Exon, Exon number, Intron) and detects Intergenic insertion sites. It also adds some additional information such as nearest gene for intergenic insertions and orientation relative to a gene. In addition to this, the program lets the user know the insertion orientation relative to the coding strand of a gene. This software must also use 2 files to add all the necessary information; ZebraFishGene.gff and GeneFeature.gff

**Zebrafish gene information file**

This is a row from a 5 column file that contains gene information for every gene in the zebrafish genome (Gene Name, Gene\_ID, Chromosome number, start site, end site). This file was made from the zebrafish\_V9.gff file.

ZebraFishGene.gff:

si:ch73-252i11.3 ENSDARG00000104632 4 6733 52120

**Zebrafish gene feature information file**

This is a 6 column file that contains gene feature information for every gene in the zebrafish genome (Feature (Exon, CDS, UTR), start site, end site, gene\_ID, feature number (if coding exon), strand gene is on (+ or -)). This file was made from the zebrafish\_V9.gff file.

GeneFeature.gff

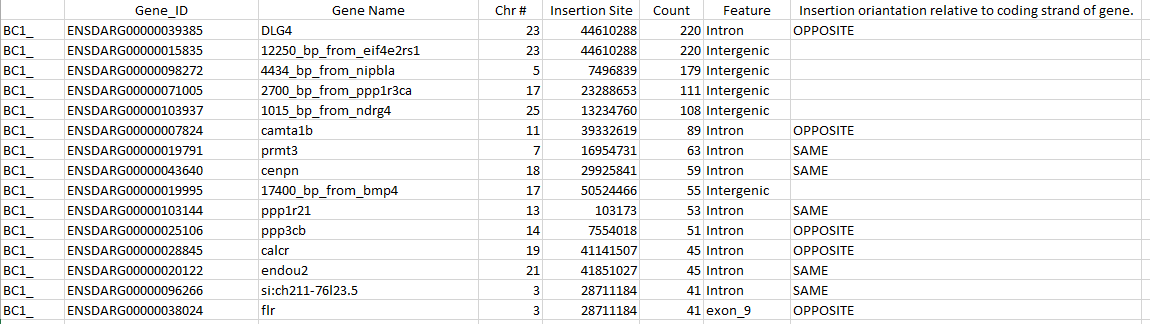
exon 30383 30614 ENSDARG00000098417 1 +

exon 33830 34259 ENSDARG00000098417 2 +

CDS 16269 16549 ENSDARG00000100660 1 +

UTR 33433 36772 ENSDARG00000098417 “ -

Output:

****

**Whole chromosome integration frequency plots**

Using a text file generated from the “Java Annotation Script”, insertion frequency plots for each chromosome can be generated. This is done in Microsoft excel by FREQUENCY function, plotting the number of unique insertions per 1000kb bin.

**Molecular genetic verification of insertion sites.**

When an insertion site of interest is detected, using the genomic location information, primers can be designed for detection of a true insertion. These primers simply flank the potential insertion site and a PCR followed by sequencing is done to confirm it.