1 Requirements

- Python and JAVA environment installed on the system.

- RNA-seq alignments in SAM format.
  (.fq files must be aligned using RNA-seq alignment program, and BAM files must be converted to SAM format)

- Reference DNA sequence in FASTA format.
  (FASTA files must be divided into separate chromosome files.
  ex)chr1.fa, chr2.fa ...)

2 Converting RNA-seq alignments

RNA-seq alignments in SAM format are filtered and merge. We create GFF format file after filtering. Use following command for converting SAM files to a GFF file.

```
python FilterSAMConvertToGFF.py [input_GFF_file_name]
[output_GFF_file_name] [SAM_file_name_list]
```

[input_GFF_file_name] is the previous GFF file that is generated using this program. This is specified to update and maintain created GFF files incrementally. If you are starting to build this GFF from a scratch, please create an empty file such as empty.gff.

[output_GFF_file_name] is the output GFF file name.

[SAM_file_name_list] is the file containing the list of SAM files to be converted. One should specify the full path. This file should look like following.

```
/home/SpliceGraph/Human/chr1.sam
/home/SpliceGraph/Human/chr2.sam
/home/SpliceGraph/Human/chr3.sam
/home/SpliceGraph/Human/chr4.sam
/home/SpliceGraph/Human/chr5.sam
...
```

3 Build Splice graph data structure

Before running the ConstructSpliceGraph.jar, reference DNA Fasta files should be converted to .index and .trie files using PrepDB.py.

```
python PrepDB.py FASTA [chr1_Fasta_file_name.fa]
```

Now, we create splice graph data structure files(.ms2db) using GFF files.

```
java -jar ConstructSpliceGraph.jar -r [GFF_file.gff] -t [input_trie_file.trie] -w
[splice_graph_file.ms2db]
```

4 Convert to FASTA

In order to use conventional MS/MS search tools, we convert splice graph data structure to a Fasta format.

```
python ACTG.py [splice_graph_file.ms2db] [output_fasta_file.fa]
[maximum_peptide_length_parameter(int)] [minimum_exon_len_parameter(int)]
```

[maximum_peptide_length_parameter(int)] is the maximum length of peptide expected to be found in MS/MS bottom-up spectra(we used 30 for our analysis).

[minimum_exon_len_parameter(int)] is the parameter to control the size of the Fasta file. You will get smaller Fasta file size as you increase this parameter. However, you may lose the combinations of some sequences that can be generated by the exons that have shorter length(in base pair) then this parameter(we used 20 for our analysis).

5 Build Sixframe database

Following command generates sixframe Fasta files from reference dna sequence.

```
java -jar SixFrameBuilder.jar -r [dna_chr1.fa] -w [output_sixframe_chr1.fa]
```

6 Search MS/MS spectra

Search MS/MS spectra against created splice graph ans sixframe Fasta databases.

7 Find Genomic locations of peptide spectrum matches

After removing the peptides that can be matched to the known protein database, one should find genomic locations(coordinates of DNA) of identified novel peptides. You need to provide sixframe and splice graph Fasta file created in previous stages. Following command should be used to search genomic locations of PSMs.

```
java -jar SearchLocation.jar
-r [/home/SpliceGraph/input_MSGFDB_result.txt]
-w [/home/SpliceGraph/output_location_chr1.txt]
-s [/home/SpliceGraph/splice_graph_chr1.fa]
-t [/home/SpliceGraph/sixframe_chr1.fa]
```

Currently, this program is set up as parsing the MSGFDB tsv results. Making this more flexible for other search tools is in our to do list.