User Guide for VirusSeq

Table of Contents

1. Introduction

1.1 Overview

VirusSeq comprises two PERL programs:

- VirusSeq_Detection.pl
- VirusSeq_Integration.pl

VirusSeq_Detection.pl detects the known viruses, annotated in the Genome Information Broker for Viruses database, in human cancer tissues using next-generation sequencing data. VirsuSeq_Integration.pl detects virus integration sites in the human genome using next-generation sequencing data. The paired-end (PE) reads in FASTQ format are used as input for both VirusSeq_Detection.pl and VirusSeq_Integration.pl. VirusSeq works with both whole-genome and whole-transcriptome sequencing data.

1.2 Reference genome and annotation files

The following reference genome and annotation files are required for VirusSeq:

- hg19.fa
- gibVirus.fa
- hg19Virus.fa
- hg19Virus_refGene_RIS.txt
- Spanner_anchor_hg19Virus.txt

Here, hg19.fa contains human genome reference sequences of version 19 in FASTA format that can be downloaded from the UCSC genome browser. gibVirus.fa contains the virus genome reference sequences in FASTA format from the Genome Information Broker for Viruses, which includes all known viruses. hg19Virus.fa is a reference genome in FASTA format that combines hg19.fa and chrVirus, which is a concatenation of selected virus genomes. The current version of chrVirus is a concatenation of the following genomes: EBV, CMV, HBV, HCV, HTLV, KSHV, MCV, XMRV, HPV16, HPV18, HPV33, HPV35, HPV45, HPV56, HPV35, HPV6, BKV, and JCV. hg19Virus_refGene_RIS.txt is the annotation file for the hg19Virus.fa genome. Specifically, hg19Virus_refGene_RIS.txt provides genomic locations of each refSeq gene of the human genome and each viral gene with recalculated coordinates for each virus in the refFlat format. Spanner_anchor_hg19Virus.txt specifies the length of each chromosome in the hg19Virus. More virus genomes can be concatenated into hg19Virus.fa with the annotation file of hg19Virus_refGene_RIS.txt updated accordingly.

The reference genome files, including hg19.fa, gibVirus.fa, and hg19Virus.fa, need to be converted into a “jump database” for the purpose of quick alignment/mapping by MOSAIK. The shell script for generating jump databases is provided in Section 4.

Please note that all genome reference sequences will be updated once the new versions are released.

2. Detection of virus by next-generation sequencing (NGS) data

2.1 Installation of tools, reference genome and annotation files

MOSAIK is used to perform alignment against the reference genome with paired-end (PE) reads in FASTQ format as input. MOSAIK implements both a hashing scheme and the Smith-Waterman algorithm to produce gapped optimal alignments. In order to explain the details of the installation and usage of VirusSeq, an example of RNA-Seq data with sampleID=L526401A is used in the rest of the text to illustrate the installation of VirusSeq and the procedures for detecting a virus and its integration sites.

Assume that the following directories are created:

1. /RIS/home/xsu1/VirusSeq
2. /RIS/home/xsu1/VirusSeq/Mosaik_bin
3. /RIS/home/xsu1/VirusSeq/Mosaik_JumpDb
4. /RIS/home/xsu1/VirusSeq/VirusSeq_Script
5. /RIS/home/xsu1/VirusSeq/L526401A
6. /RIS/home/xsu1/VirusSeq/L526401A/Gig
7. /RIS/home/xsu1/VirusSeq/L526401A/SV_gDNA

The directory */RIS/home/xsu1/VirusSeq/Mosaik_bin* contains the following executable tools:

1. MosaikBuild
2. MosaikAligner
3. MosaikSort
4. MosaikMerge
5. MosaikText
6. MosaikAssembler
7. Spanner

The directory */RIS/home/xsu1/VirusSeq/Mosaik_JumpDb* contains the following reference genome files and annotation files:

1. hg19.fa
2. gibVirus.fa
3. hg19Virus.fa
4. hg19Virus_refGene_RIS.txt
5. Spanner_anchor_hg19Virus.txt

6. Jump_file_builder.sh

The directory */RIS/home/xsu1/VirusSeq/Mosaik_Script* contains the following PERL scripts:

1. Virus_Detection.pl
2. Spanner_cross_converter.pl
3. Virus_Integration.pl

The directory */RIS/home/xsu1/VirusSeq/L526401A* contains the following files:

1. Gzipped fastq files L526401A_1.fq.gz and L526401A_2.fq.gz
2. Shell script L526401A_Virus_Detection.sh
3. Shell script L526401A.sh
4. Shell script L526401A_SE.sh

The directory */RIS/home/xsu1/VirusSeq/L526401A/SV_gDNA* contains the following file:

1. Shell script L526401A_Virus_Integration.sh

2.2 Virus detection by NGS data

For this example, the directory /RIS/home/xsu1/VirusSeq/L526401A is the working folder for the example with sampleID=L526401A. The shell script f L526401A_Virus_Detection.sh built for this example is used for virus detection. Specifically, the shell script has the following commands:

```
###start aligning the PE reads of the example against hg19 for human sequence subtraction

/RIS/home/xsu1/VirusSeq/Mosaik_bin/MosaikBuild
        -q /RIS/home/xsu1/VirusSeq/L526401A/L526401A_1.fq.gz
        -q2 /RIS/home/xsu1/VirusSeq/L526401A/L526401A_2.fq.gz
        -out L526401A_Virus.bin
        -st illumina
```

Here, MosaikBuild converts external read formats to the native MOSAIK format, -q specifies the FASTQ file for the first mate, -q2 specifies the FASTQ file for the second mate, -out specifies the binary output read file, and -st specifies the sequencing platform.

```
/RIS/home/xsu1/VirusSeq/Mosaik_bin/MosaikAligner
        -in L526401A_Virus.bin
        -ia /RIS/home/xsu1/VirusSeq/Mosaik_JumpDb/hg19.fa.bin
        -out L526401A_Virus.bin.aligned
        -hs 15 -mmp 0.1 -mmal -minp 0.5 -act 25 -mhp 100 -m unique
        -j /RIS/home/xsu1/VirusSeq/Mosaik_JumpDb/hg19.JumpDb -p 14 -km -pm
        -rur L526401A_Unalg.fq
```

Here, MosaikAligner performs alignment against human genome reference hg19, -in specifies the input read file generated by MosaikBuild, -ia specifies the input reference file, -out specifies the output alignment file, -j specifies the specified jump database, -p specifies the number of processors (cores), and -rur stores unaligned reads in a FASTQ file with filename L526401A_Unalg.fq.

```
##start aligning unmapped reads against virus genomes to detect the virus

/RIS/home/xsu1/VirusSeq/Mosaik_bin/MosaikBuild
        -q L526401A_Unalg.fq
        -out L526401A_Virus.bin
        -st illumina

/RIS/home/xsu1/VirusSeq/Mosaik_bin/MosaikAligner
        -in L526401A_Virus.bin
        -ia /RIS/home/xsu1/VirusSeq/Mosaik_JumpDb/gibVirus.fa.bin
        -out L526401A_Virus.bin.aligned
        -hs 15 -mmp  0.15 -act 25 -mhp 100 -m all
        –j /RIS/home/xsu1/VirusSeq/Mosaik_JumpDb/gibVirus.JumpDb -p 14 -km –pm

/RIS/home/xsu1/VirusSeq/Mosaik_bin/MosaikSort
        -in L526401A_Virus.bin.aligned
        -out L526401A_Virus.bin.aligned.sorted
```

Here, MosaikSort sorts the mapped reads in terms of their mapped genomic location.

```
/RIS/home/xsu1/VirusSeq/Mosaik_bin/MosaikAssembler
        -in L526401A_Virus.bin.aligned.sorted
        -ia /RIS/home/xsu1/VirusSeq/Mosaik_JumpDb/gibVirus.fa.bin
        -out Gig/L526401A_Virus.bin.aligned.sorted.assembled
        -f ace
        >L526401A_VirusLog.txt
```

Here, MosaikAssembler produces an assembly of the reads pileup for visualization purposes from a sorted MOSAIK alignment file. All ace files generated by MosaikAssembler are saved at directory "/RIS/home/xsu1/VirusSeq/L526401A/Gig." Meanwhile, MosaikAssembler also generates the overall count of the mapped reads in each viral genome. The log information from MosaikAssembler is piped/redirected to the file L526401A_VirusLog.txt.

```
###detect the virus in the sample by VirusSeq_Detection.pl

perl /RIS/home/xsu1/VirusSeq/VirusSeq_Script/VirusSeq_Detection.pl
        L526401A_VirusLog.txt
        1000
        L526401A_VirusName.txt
```

VirusSeq_Detection.pl takes the log file L526401A_VirusLog.txt generated by MosaikAssembler as input for virus detection. The parameter 1000, which is defined as the overall count of the mapped reads within a viral genome, is the empirical cutoff to detect the virus. The filename L26401A_VirusName.txt is specified by the user to save virus detection results.

3. Detection of virus integration sites by NGS data

3.1 PE reads FASTQ file mapping against hybrid genome hg19Virus

The directory /RIS/home/xsu1/VirusSeq/L526401A is the working folder for the example with sampleID=L526401A. First, the FATSQ files of PE reads without human genome subtraction are mapped/aligned to the hybrid reference genome hg19Virus by MOSAIK. The shell script L526401A.sh is built for this example to perform the alignment f. Specifically, the shell script has the following commands:

```
###start alignment against hg19Virus

/RIS/home/xsu1/VirusSeq/Mosaik_bin/MosaikBuild
        -q /RIS/home/xsu1/VirusSeq/L526401A/L526401A_1.fq.gz
        -q2 /RIS/home/xsu1/VirusSeq/L526401A/L526401A_2.fq.gz
        -out L526401A.bin
        -st illumina

/RIS/home/xsu1/VirusSeq/Mosaik_bin/MosaikAligner
        -in L526401A.bin
        -ia /RIS/home/xsu1/VirusSeq/Mosaik_JumpDb/hg19Virus.fa.bin
        -out L526401A.bin.aligned
        -hs 15 -mmp 0.06 -mmal -minp 0.5 -act 25 -mhp 100 -m unique -a all
        -j /RIS/home/xsu1/VirusSeq/Mosaik_JumpDb/hg19Virus.JumpDb -km -pm -p 14


###Spanner is used to generate the list of discordant reads across different chromosomes

/RIS/home/xsu1/VirusSeq/Mosaik_bin/Spanner
        --scan --infile L526401A.bin.aligned
        --outdir SV_gDNA
```

Here, --scan specifies that the function of this step is to check the integrity of the MOSAIK alignment file. --infile specifies the input alignment file from MOSAIK, and --outdir specifies the directory where a summary statistics file (MSK.stats) from Spanner will be saved.

```
/RIS/home/xsu1/VirusSeq/Mosaik_bin/Spanner
        --build --infile L526401A.bin.aligned
        --outdir SV_gDNA
        -f SV_gDNA/MSK.stats
        -a /RIS/home/xsu1/VirusSeq/Mosaik_JumpDb/Spanner_anchor_hg19Virus.txt -t
```

Here, --build specifies that the function of this step is to generate the files with all discordant reads for each chromosome, -f specifies the summary statistics file generated by scan step, -a specifies the relationship between the chromosome name, chromosome number, and the length of each chromosome, and -t specifies the output that will be in text format. Spanner generates the files with all discordant reads for each chromosome, which only specifies the genomic location of the discordant reads. But, the files with all the discordant reads for each chromosome by Spanner do not contain information such as readID, its corresponding sequences, and read mapping orientation. This information will be retrieved in the next step, described in Section 3.2.

3.2 Single-end reads FASTQ file mapping against hybrid genome hg19Virus

This step generates the genomic location for each uniquely-mapped read with all the information, including readID, read sequence, and mapping orientation, etc. The file from this step will be used to annotate the discordant reads generated by the previous step. Running the shell script L526401A_SE.sh will generate the file in AXT format. Specifically, the shell script has the following commands:

```
##align the first mate against hg19Virus

/RIS/home/xsu1/VirusSeq/Mosaik_bin/MosaikBuild
        -q /RIS/home/xsu1/VirusSeq/L526401A/L526401A_1.fq.gz
        -out L526401A_1.bin
        -st illumina

/RIS/home/xsu1/VirusSeq/Mosaik_bin/MosaikAligner
        -in L526401A_1.bin
        -ia /RIS/home/xsu1/VirusSeq/Mosaik_JumpDb/hg19Virus.fa.bin
        -out L526401A_1.bin.aligned
        -hs 15 -mmp 0.06 -mmal -minp 0.5 -act 25 -mhp 100 -m unique
        -j /RIS/home/xsu1/VirusSeq/Mosaik_JumpDb/hg19Virus.JumpDb -p 14 -km –pm

/RIS/home/xsu1/VirusSeq/Mosaik_bin/MosaikSort
        -in L526401A_1.bin.aligned
        -out L526401A_1.bin.aligned.sorted
        -u
```

Here, -u specifies that only uniquely-aligned reads are retained in the sorted file.

```
##align the second mate against hg19Virus

/RIS/home/xsu1/VirusSeq/Mosaik_bin/MosaikBuild
        -q /RIS/home/xsu1/VirusSeq/L526401A/L526401A_2.fq.gz
        -out L526401A_2.bin
        -st illumina
```

```
/RIS/home/xsu1/VirusSeq/Mosaik_bin/MosaikAligner
        -in L526401A_2.bin
        -ia /RIS/home/xsu1/VirusSeq/Mosaik_JumpDb/hg19Virus.fa.bin
        -out L526401A_2.bin.aligned
        -hs 15 -mmp 0.06 -mmal -minp 0.5 -act 25 -mhp 100 -m unique
        -j /RIS/home/xsu1/VirusSeq/Mosaik_JumpDb/hg19Virus.JumpDb -p 14 -km –pm


/RIS/home/xsu1/VirusSeq/Mosaik_bin/MosaikSort
        -in L526401A_2.bin.aligned
        -out L526401A_2.bin.aligned.sorted
        -u
```

##merge two sorted files

```
/RIS/home/xsu1/VirusSeq/Mosaik_bin/MosaikMerge
        -in L526401A_1.bin.aligned.sorted
        -in L526401A_2.bin.aligned.sorted
        -out L526401A_SE.bin.aligned.sorted

/RIS/home/xsu1/VirusSeq/Mosaik_bin/MosaikText
        -in L526401A_SE.bin.aligned.sorted
        -axt L526401A_SE.bin.aligned.sorted.axt
```

Here, -axt specifies the output file of sorted alignment file in AXT format.

3.3 Detection of virus integration site

In this step, the working directory is "/RIS/home/xsu1/VirusSeq/L526401A/SV_gDNA". Running the shell script L526401A_Virus_Integration.sh will generate the result for the virus integration sites. Specifically, the shell script has the following commands:

##Spanner cross reads converter

```
perl /RIS/home/xsu1/VirusSeq/VirusSeq_Script/Spanner_cross_converter.pl
        /RIS/home/xsu1/VirusSeq/Mosaik_JumpDb/hg19Virus_refGene_RIS.txt
        /RIS/home/xsu1/VirusSeq/L526401A/L526401A_SE.bin.aligned.sorted.axt
        L526401A_CrossReads.txt
```

Here, the PERL script "Spanner_cross_converter.pl" annotates the discordant reads in the file for chrVirus, with readID, read sequence, and mapping orientation, etc., for the next step of virus integration site detection. The file hg19Virus_refGene_RIS.txt is the annotation file for all refSeq genes and viral genes concatenated in hg19Virus. The file L526401A_SE.bin.aligned.sorted.axt contains the mapping information for all uniquely aligned reads. The file L526401A_CrossReads.txt, which is specified by the user, contains the annotated discordant reads.

```
##Virus integration site detection

perl /RIS/home/xsu1/VirusSeq/VirusSeq_Script/VirusSeq_Integration.pl
        L526401A_CrossReads.txt
        /RIS/home/xsu1/VirusSeq/Mosaik_JumpDb/hg19Virus_refGene_RIS.txt hg19
        192 95 50
        L526401A_Integration.txt
```

Here, the PERL program “VirusSeq_integration.pl” performs the detection of the virus integration site. The file L526401A_CrossReads.txt is the input file that contains the annotated discordant reads. The file hg19Virus_refGene_RIS.txt is the annotation file for all refSeq genes and viral genes concatenated in hg19Virus. 192 is the parameter that specifies the length of the library fragment, 95 is the parameter for the standard deviation of the library fragment length, and 50 is the parameter for the length of the PE read. The values of all three parameters can be obtained from the Illumina Cores. The file L526401A_Integration.txt, which is specified by the user, contains the output of the virus integration site detection.

4. Shell command for MOSAIK “jump database” generation

MosaikJump is a tool that converts a reference sequence archive into a “jump database.” The jump database is a custom data structure that separates seeds and hash positions in a manner that is fast and collision-free. Using a Linux cluster with a large memory, the jump database will make MOSAIK run much faster.

Please make sure that the /tmp directory is big enough (at 50GB) on the hard disk, or specify another directory by setting the MOSAIK_TMP environment variables in hash:

```
export MOSAIK_TMP = /RIS//home/xsu1/VirusSeq/tmp
```

Set the directory “/RIS/home/xsu1/VirusSeq/Mosaik_JumpDb” as the working folder. Then, running the shell script Jump_file_builder.sh will generate all related jump databases required. If there is still an administrative issue when generating the jump database, please feel free to contact me at xsu1@mdanderson.org.

Please be informed that you have to replace the master directory name “/RIS/home/xsu1VirusSeq/” in all the shell scripts once you create your own master directory in your own Linux machine. Please gunzip all three reference genome fasta files before running the jump shell script “Jump_file_builder.sh” under the directory “RIS/home/xsu1/VirusSeq/Mosaik_JumpDb/”. Please don’t gunzip fastq files for the example PE reads files since MOSAIK can directly deal with gzipped fastq files.