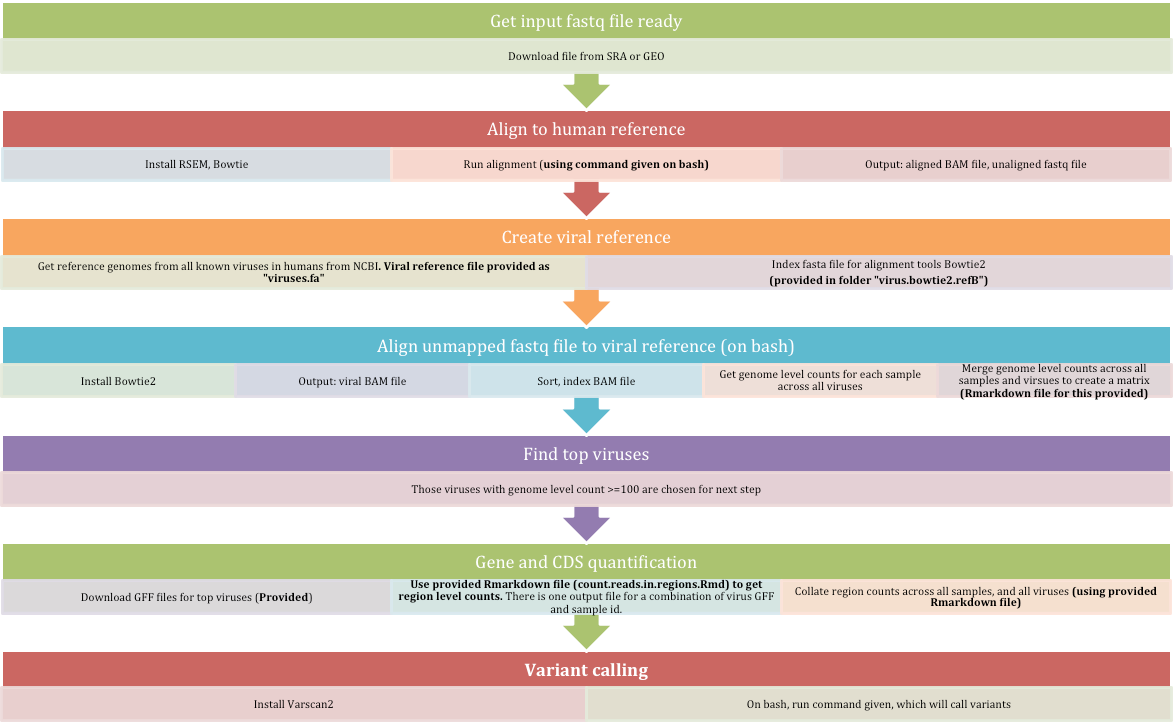
# Tutorial for viGEN workflow

### **Steps in general**



### **Steps in detail**

* For this tutorial, we used a sample RNA-seq file from liver cancer from public study SRA <http://www.ncbi.nlm.nih.gov/bioproject/PRJNA279878> and GEO <http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE65485> (also available at EBI SRA here: <http://www.ebi.ac.uk/ena/data/view/SRR1946637>). We downloaded the raw reads in the form of paired end Fastq file from these SRA links for one sample SRR1946637, and unzip them.
* About the dataset: Data consists of Liver cancer patients in China. The authors look for HBV-human integration sites in the dataset of 50 Liver cancer patients in China and 5 adjacent normal tissues. This particular sample is an adjacent normal liver tissue.
* Alignment to human reference
  + In this tutorial, we use tool RSEM for alignment
  + Install RSEM as shown in <https://github.com/bli25ucb/RSEM_tutorial>
  + RSEM uses aligner Bowtie1, Bowtie2 or STAR. We chose to use Bowtie1 which is available as pre-compiled executable. For Bowtie1 installation, follow instructions in this manual: <http://bowtie-bio.sourceforge.net/manual.shtml#obtaining-bowtie>
  + Test RSEM successful installation

./rsem-calculate-expression --version

./rsem-calculate-expression --help

* + Generate transcriptome reference for RSEM

RSEM uses a transcriptome reference for alignment. RSEM has a command called “rsem-prepare-reference” where it uses the genome fasta reference file and the GTF file to create the transcriptome reference. We have generated and provided these files for human hg19 reference genome in the “**Rsem\_ref\_hg19**” folder

* + - If you are interested in using another reference, please follow steps provided here to generate transcriptom reference for your file: <https://github.com/bli25ucb/RSEM_tutorial>
  + Once RSEM is installed, and the transcriptome reference is ready, run alignment using RSEM on SRA fastq files using following command

./rsem-calculate-expression --paired-end -p 8 --output-genome-bam \

--keep-intermediate-files \

--bowtie-path /Users/username/Documents/software/bowtie/bowtie-1.1.2 \

--append-names \

--estimate-rspd \

--time \

input/SRR1946637\_1.fastq input/SRR1946637\_2.fastq \

Rsem\_ref\_hg19/Rsem\_ref\_hg19 \

output/SRR1946637

In addition to the aligned BAM file (genome level and transcriptome level), this will generate the unaligned (unmapped) fastq files named **SRR1946637\_un\_1.fq** and **SRR1946637\_un\_2.fq**. They consist of the reads that did not align to the human reference.

This workflow is part of “Module 1 (filtered human input files)” in Figure 1 in the manuscript.

* Create viral reference
  + We were interested in exploring all viruses existing in humans. So we first obtained reference genomes of all known and sequenced human viruses obtained from NCBI (as of Sep 2015), and merged them into one file (referred to as the ‘viral reference file’) in fasta file format.
  + Merge all virus fasta file into one big fasta file called “**viruses.fa**”. For this tutorial, we have provided this file.

*We plan to provide code that does automatic download of fasta files from NCBI FTP site through our Bioconductor package.*

* + NCBI also allows to download information/annotation about these viruses from: http://www.ncbi.nlm.nih.gov/genomes/GenomesGroup.cgi?taxid=10239. This information has been provided as “**Complete\_Sequence\_info.csv**”
  + We have also indexed the viral reference files, so that these files are ready for alignment tools Bowtie2 (folder name: **virus.bowtie2.refB**) or BWA (folder name: **viral.bwa.ref)**

In case user is interested in indexing the reference file on their own, this is the command to use: ./bowtie2-build /PathToLocation/viruses.fa virus.bowtie2.refB

* Align these unmapped fastq files to the viral reference. In this tutorial, we use alignment tool Bowtie2.
  + Install Bowtie2 pre-compiled binary file based on instructions in this manual: <http://bowtie-bio.sourceforge.net/bowtie2/manual.shtml#obtaining-bowtie-2>
  + Alignment generates output in the form of a SAM file. This SAM file contains the reads from various viruses aligned to the viral reference.
  + Convert SAM to BAM using Samtools
  + Sort the BAM file coordinate wise
  + Use “Samtools idx” tool. This produces a tab delimited file, with each line consisting of a virus sequence name, sequence length, # mapped reads and # unmapped reads. The # of mapped reads shows the genome level counts.

**We have provided a small shell script “viral.pipeline\_public\_final.sh” that encompasses all of these above steps.**

* Get genome level matrix file
  + If you have more than one sample, then it might be useful to concatenate the number of mapped reads (genome level counts) from each sample into a matrix format. This file can then be used for any further analysis
  + We have provided an Rmarkdown file that has code and output for how to merge all the “Samtools idx” files. See file “**merge.idx.stats.pdf**” created from “**merge.idx.stats.Rmd**”
  + Add annotation using information from “Complete\_Sequence\_info.csv” file.
* The genome level counts from the previous step will help us determine the top virus genomes to use for the next step (Gene and CDS quantitation). We use a cutoff genome level count to shortlist the viruses.
  + For this tutorial, we decided to use a threshold genome level count of 100. So all viral genomes that have “genome level count” >= 100 are short listed. This gives us 43 viral genomes used in the next step.
  + For this tutorial, we only use one sample, so finding the top viruses was easy.
  + If you have multiple samples belonging to case and control sub-groups, we recommend:
    - averaging the genome level counts in each in each sub-group,
    - Find the top viruses in each sub-group

*We plan to provide code for this in our Bioconductor package under preparation.*

* Gene and CDS quantification
  + In this step, we use our in-house pipeline to generate gene and CDS counts is for every input BAM file.
  + The region level information is extracted from Gene Feature Format files (GFF) files which are available for most viral genomes from NCBI.
  + Download GFF files for the top viruses.
    - For this tutorial, we have provided the GFF files in folder “**un\_bowtie\_topGff**”. *We plan to provide code that does automatic download of GFF files from NCBI FTP site in our Bioconductor package.*
    - Note: Not all fasta files will have GFF files. In this tutorial, out of 43 top viruses, we obtained GFF files for 40 viruses.
  + Input to our in-house pipeline for Gene & CDS quantification:
    - The viral bam files (the BAM file output from previous Bowtie2 step)
  + We have provided R markdown file “**count.regions.in.regions.Rmd**” that generates these Gene and CDS counts. *We plan to integrate this code into our Bioconductor package.*
  + Output :
    - This code produces counts for each region in the GFF file. One file is created for each virus (GFF file) for each sample. Three types of counts are calculated - “in region”, “on boundary” and “in gaps”. This output is available as a “.csv” file and “Rdata” files.
    - Collate all read counts across all samples and across all top viruses to create a matrix. This code (provided as **collate.output.files.Rmd**) calculates total of “in region” and “on boundary” (referred to as “sum”) when collating.
    - Add virus annotation using information from “Complete\_Sequence\_info.csv” file.

The viruses that highest the highest region counts are shown in table below. You can see various gene/CDS regions of Hepatitis B virus showing up on top with highest counts. This is a verification of the HBV status of the sample.

|  |  |  |  |
| --- | --- | --- | --- |
| **Region Name** | **SRR1946637** | **ncId** | **annot.name** |
| NC\_003977.1\_region\_1\_3215 | 96329 | NC\_003977.1 | Hepatitis B virus |
| NC\_018464.1\_region\_1\_927 | 82509 | NC\_018464.1 | Shamonda virus |
| NC\_018476.1\_region\_1\_6895 | 49784 | NC\_018476.1 | Simbu virus |
| NC\_003977.1\_gene\_1374\_1838 | 43011 | NC\_003977.1 | Hepatitis B virus |
| NC\_003977.1\_CDS\_1374\_1838 | 43011 | NC\_003977.1 | Hepatitis B virus |
| NC\_009823.1\_region\_1\_9711 | 31470 | NC\_009823.1 | Hepatitis C virus genotype 2 |
| NC\_003977.1\_CDS\_155\_835 | 21591 | NC\_003977.1 | Hepatitis B virus |
| NC\_004102.1\_region\_1\_9646 | 17922 | NC\_004102.1 | Hepatitis C virus |
| NC\_004102.1\_three\_prime\_UTR\_9378\_9646 | 17922 | NC\_004102.1 | Hepatitis C virus |
| NC\_001672.1\_region\_1\_11141 | 16336 | NC\_001672.1 | Tick-borne encephalitis virus |
| NC\_018711.1\_region\_1\_7195 | 14251 | NC\_018711.1 | Lunk virus NKS-1 |
| NC\_018478.1\_region\_1\_4417 | 13745 | NC\_018478.1 | Simbu virus |
| NC\_015521.1\_region\_1\_7310 | 10122 | NC\_015521.1 | Cutthroat trout virus |
| NC\_015521.1\_three\_prime\_UTR\_7191\_7310 | 10122 | NC\_015521.1 | Cutthroat trout virus |
| NC\_022518.1\_region\_1\_9472 | 9028 | NC\_022518.1 | Human endogenous retrovirus K113 |
| NC\_009827.1\_region\_1\_9628 | 8438 | NC\_009827.1 | Hepatitis C virus genotype 6 |
| NC\_009827.1\_three\_prime\_UTR\_9400\_9628 | 8438 | NC\_009827.1 | Hepatitis C virus genotype 6 |
| NC\_009827.1\_region\_9433\_9495 | 8438 | NC\_009827.1 | Hepatitis C virus genotype 6 |
| NC\_001798.1\_region\_1\_154746 | 6388 | NC\_001798.1 | NA |
| NC\_018382.1\_region\_1\_6796 | 6232 | NC\_018382.1 | Bat hepevirus |
| NC\_018382.1\_three\_prime\_UTR\_6691\_6796 | 6232 | NC\_018382.1 | Bat hepevirus |
| NC\_003977.1\_gene\_1814\_2452 | 5980 | NC\_003977.1 | Hepatitis B virus |
| NC\_003977.1\_CDS\_1814\_2452 | 5980 | NC\_003977.1 | Hepatitis B virus |
| NC\_003977.1\_gene\_2307\_4838 | 5327 | NC\_003977.1 | Hepatitis B virus |
| NC\_003977.1\_CDS\_2307\_4838 | 5327 | NC\_003977.1 | Hepatitis B virus |
| NC\_002645.1\_region\_1\_27317 | 5080 | NC\_002645.1 | Human coronavirus 229E |
| NC\_003977.1\_CDS\_1901\_2452 | 4776 | NC\_003977.1 | Hepatitis B virus |
| NC\_022518.1\_long\_terminal\_repeat\_1\_968 | 4558 | NC\_022518.1 | Human endogenous retrovirus K113 |
| NC\_003977.1\_gene\_2848\_4050 | 3876 | NC\_003977.1 | Hepatitis B virus |
| NC\_003977.1\_CDS\_2848\_4050 | 3876 | NC\_003977.1 | Hepatitis B virus |

* + Application: These gene and CDS count files can be used to compare case and control groups of interest using popular tools like EdgeR (<http://bioconductor.org/packages/edgeR/>) or DESeq2 in Bioconductor that can accept raw counts.
* Variant calling
  + Install Varscan2 from here: <https://sourceforge.net/projects/varscan/files/>
  + Run Varscan2 on command line using the following command:

samtools mpileup -B -f /Users/ls483/Documents/SRA.GEO/viral.reference/viruses.fa -d 9999 -Q 17 -q 17 SRR1946637\_un.bowtie2.sorted.bam| java -Xmx2g -jar /Users/ls483/Documents/software/varscan2/VarScan.v2.3.9.jar mpileup2cns --output-vcf 1 --min-var-freq 0.01 | awk '/^#/ || $7=="PASS"' > /Users/ls483/Documents/SRA.GEO/output\_varscan2/SRR1946637\_un.vcf

* Output: vcf file with viral-variants called