RNA-seq Workshop, ACECR, October 2021

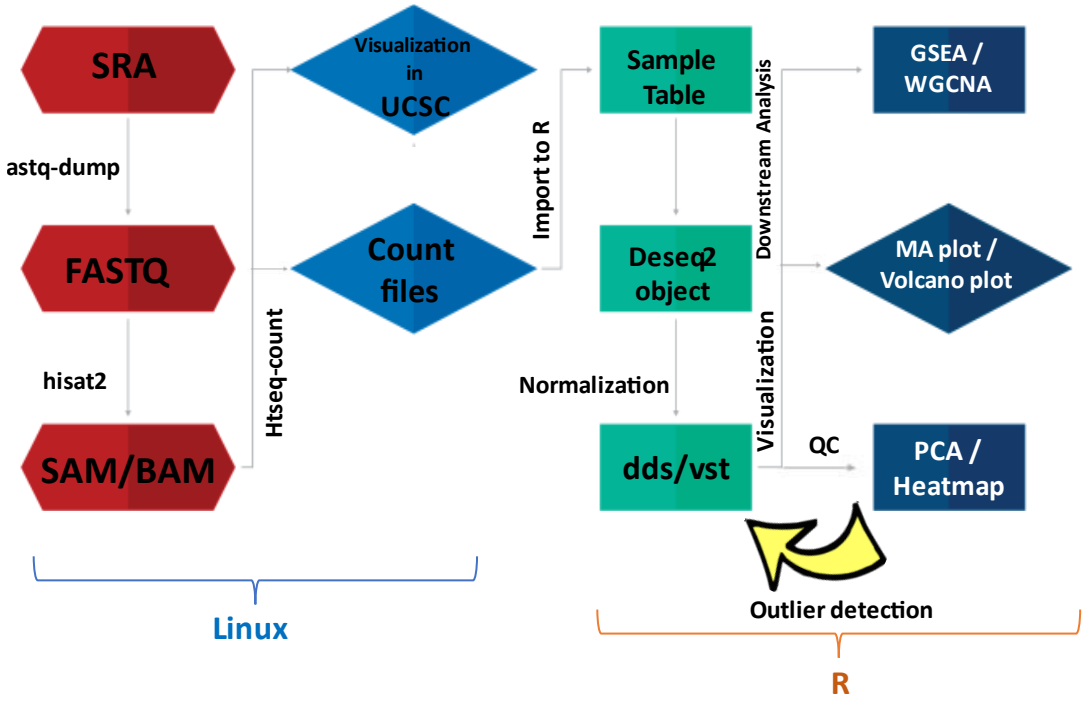
Linux

Basic Ubuntu Commands for RNA-seq Data Analysis	Function
`ls`	List your working directory content
`ls -lh`	List your working directory content • `-L` use a long listing format • `-h` with -l and/or -s, print human readable . Sizes (e.g., 1k 234m 2g)
`cd <dir_name>`	Change your current working directory
`cd /`	Go to the root directory
`cd ..`	Go up one directory level
`mkdir <dir_name>`	Create a new directory
`touch <file_name>`	Create a new file
`pwd`	Display the full path of the current directory
`whoami`	Display the username of the current user
`cat <file_name>`	View content of a file
`zcat <file_name>`	View content of a compressed file
`head <file_name>`	Display the first ten lines of a file
`head -n <number> <file_name>`	Display a specific number of lines
`tail <file_name>`	Displays the last ten lines of a file
`tail -n <number> <file_name>`	Display a specific number of lines
`history`	Check command history
`tree`	See directory tree structure
`mv <file_name>`	Cut and paste a file or rename a file
`grep`	Search for pattern in each file
`wc -l <file_name>`	Calculate the number of lines in a file
`nano <file_name>`	Open a file in a text editor
`gunzip <zipfile_name>`	Decompress files

`rm <file_name>`	Remove a file
`rm -r <dir_name>`	Remove a directory
`cp <file_name or file_path> <dir_path>`	Copy and paste a file
`wget <URL>`	Download a file from internet

Ubuntu Terminal Shortcuts	Function
Ctrl + Shift + C	Copy the highlighted command to the clipboard
Ctrl + Shift + V	Paste the contents of the clipboard
Ctrl + L	Clear the terminal
Ctrl + C	Kill the current process

RNA-Seq Data Analysis Pipeline



Downloading files from SRA

```
fastq-dump --split-files --gzip SRR1234567
```

`--split-files`: split spots into individual reads

`--gzip`: compress output using gzip

Quality control of fastq files

```
fastqc SRR123456.fastq
```

Indexing reference genomes

```
hisat2-build -p 12 GRCh38.d1.vd1.fa GRCh38_index
```

`-p`: number of threads

Aligning reads using HISAT2 (generating SAM file)

```
hisat2 -p 12 -x /home/Human_genome/GRCh38_index -1
/home/RNA_seq_leukemia/1.fastq -2 /home/RNA_seq_leukemia/2.fastq -S N_new.sam
```

`-p`: number of threads

`-x`: index filename prefix

`-1`: Files with #1 mates, paired with files in -2

`-2`: files with #2 mates, paired with files in -1

`-S`: file for sam output

Converting SAM to BAM

```
samtools sort -@ 6 -o N_new_sorted.bam N_new.sam
```

-@: number of additional threads to use

-o: write final output to FILE rather than standard output

Indexing BAM file using SAMTOOLS

```
samtools index N_new_sorted.bam
```

Generating bigwig

```
bamCoverage -p 40 -b N_new_sorted.bam -o N_new_bam.bw
```

-p: number of threads

-b: bamsorted filename

Read counting

```
htseq-count --stranded=yes --idattr=gene_name --mode=union --format=bam
N_new_sorted.bam ~/Human.GTF> N_new.counts
```

--stranded {yes,no,reverse}

Whether the data is from a strand-specific assay.

Specify 'yes', 'no', or 'reverse' (default: yes).

'reverse' means 'yes' with reversed strand

interpretation

--idattr IDATTR

GTF attribute to be used as feature ID (default,

suitable for Ensembl GTF files: gene_id)

--mode {union,intersection-strict,intersection-nonempty}

Mode to handle reads overlapping more than one feature

(choices: union, intersection-strict, intersection-

nonempty; default: union)

--format {sam,bam}

Type of <alignment_file> data, either 'sam' or 'bam'

(default: sam)