Reads Alignment (HISAT2)

HISAT2 is a fast and sensitive alignment program for mapping next-generation sequencing reads to reference genome(s) [6]. We are going to use this tool to align the reads to hg19 genome, HHV8 genome, and mm10 genome, respectively.

Install HISAT2

```
# Download and extract the latest version
$ wget ftp://ftp.ccb.jhu.edu/pub/infphilo/hisat2/downloads/hisat2-2.0.4-
Linux_x86_64.zip
$ unzip hisat2-2.0.4-Linux_x86_64.zip

# Append to PATH environment variable
$ export PATH=$PATH:/path/to/hisat2-2.0.4

# Verify installation
$ hisat2 --help
$ hisat2 --version
```

Install Samtools

Samtools is a suite of programs for interacting with high-throughput sequencing data [7]. SAM files produced by HISAT2 must be sorted and converted to BAM using samtools before running StringTie.

```
# Download and extract samtools
$ wget https://github.com/samtools/samtools/releases/download/1.11/samtools-
1.11.tar.bz2 -O samtools-1.11.tar.bz2
$ tar -xjvf samtools-1.11.tar.bz2

# Append to PATH environment variable
$ export PATH=$PATH:/path/to/samtools-1.11

# Verify installation
$ samtools --version
```

Read Alignment

1. Build indexes

You can either download HISAT2 indexes from its website

```
$ wget https://genome-idx.s3.amazonaws.com/hisat/hg19_genome.tar.gz
$ tar -zxvf hg19_genome.tar.gz
```

or download reference sequence and gene annotation from Illumina iGenome before building index by `hisat2-build` command.

```
$ cd /path/to/homo/
$ hisat2-build -p 20 hg19_genome.fa genome

$ cd /path/to/HHV8/
$ hisat2-build -p 20 hhv8_sequence.fasta genome

$ cd /path/to/mm10/
$ hisat2-build -p 20 mm10_genome.fasta genome
```

`hisat2-build` generates eight `.ht2` files, from `genome.1.ht2` to `genome.8.ht2`, which we will use for alignment in the next step.

2. Run HISAT2 with Samtools

Getting sorted BAM and index:

```
#!/bin/bash
cd /path/to/homo/ # where storing index
hisat_samtool(){
hisat2 -x genome --summary-file "$1".m6A.align_summary -p 5 -U
/path/to/trim_galore_result/"$1"_trimmed.fq | samtools view -Su |samtools sort -
o /path/to/homo_result/"$1"_sorted.bam
samtools index /path/to/homo_result/"$1"_sorted.bam
}
export -f hisat_samtool

for s in SRR5978827 SRR5978828 SRR5978829 SRR5978834 SRR5978835 SRR5978836
SRR5978869 SRR5978870 SRR5978871 SRR5179446 SRR5179447 SRR5179448
do
hisat_samtool ${s}
done
mkdir alignment_summary
mv *.align_summary alignment_summary/
```

```
#!/bin/bash
cd /path/to/mm10/ # where storing index
hisat_samtool(){
hisat2 -x genome --summary-file "$1".m6A.align_summary -p 5 -U
/path/to/trim_galore_result/"$1"_trimmed.fq | samtools view -Su |samtools sort -
o /path/to/mm10_result/"$1"_sorted.bam
samtools index /path/to/mm10_result/"$1"_sorted.bam
}
export -f hisat_samtool

for s in SRR866997 SRR866998 SRR866999 SRR867000 SRR867001 SRR867002 SRR866991
SRR866992 SRR866993 SRR866994 SRR866995 SRR866996
do
hisat_samtool ${s}
done
mkdir alignment_summary
mv *.align_summary alignment_summary/
```

```
#!/bin/bash
cd /path/to/hhv8/ # where storing index
hisat_samtool(){
```

```
hisat2 -x genome --summary-file “$1”.m6A.align_summary -p 5 -U
/path/to/trim_galore_result/“$1”_trimmed.fq | samtools view -Su |samtools sort -
o /path/to/hhv8_result/“$1”_sorted.bam
samtools index /path/to/hhv8_result/“$1”_sorted.bam
}
export -f hisat_samtool

for s in SRR5978827 SRR5978828 SRR5978829 SRR5978834 SRR5978835 SRR5978836
SRR5978869 SRR5978870 SRR5978871 SRR5179446 SRR5179447 SRR5179448
do
hisat_samtool ${s}
done
mkdir alignment_summary
mv *.align_summary alignment_summary/
```

- Note that if the aligned results are going to assemble transcript with StringTie, a `--dta` option is necessary to include the tag `XS` to indicate the genomic strand that produced the RNA from which the read was sequenced. This is required by StringTie. The `hisat2-samtools` command should be modified as

  ```
  hisat2 -x genome --summary-file SRR5978827.m6A.align_summary -p 5 -U
  /path/to/trim_galore_result/SRR5978827_trimmed.fq --dta | samtools view -Su
  |samtools sort -o /path/to/homo_result/SRR5978827_sorted.bam
  ```

> `--dta/--downstream-transcriptome-assembly`
> Report alignments tailored for transcript assemblers including StringTie. With this option, HISAT2 requires longer anchor lengths for de novo discovery of splice sites. This leads to fewer alignments with short-anchors, which helps transcript assemblers improve significantly in computation and memory usage.

- Also note that for paired end data, you need to modify the `hisat2` command as follows to get SAM file

```
hisat2 -x genome --summary-file $s.m6A.align_summary -p 5 -1
/path/to/trim_galore_result/SRR5978827_trimmed_1.fq -2
/path/to/trim_galore_result/SRR5978827_trimmed_2.fq -S
/path/to/homo_result/SRR5978827.sam
```

or modify the `hisat2-samtools` combined command to directly get sorted BAM file

```
hisat2 -x genome --summary-file SRR5978827.m6A.align_summary -p 5 -1
/path/to/trim_galore_result/SRR5978827_trimmed_1.fq -2
/path/to/trim_galore_result/SRR5978827_trimmed_2.fq | samtools view -Su
|samtools sort -o /path/to/homo_result/SRR5978827_sorted.bam
```