Hisat2_Htseq_Pipeline

In [3]:
```
from IPython.display import FileLink,FileLinks, Image, IFrame, HTML, display
```

Fastqc

In [ ]:
```
%%bash

for file in $(ls *gz)
do
    fastqc $file
    echo "$file is running"
    echo "QC is finiched for $file"
done
```

Run Multiqc Fastq Files

In [4]:
```
!multiqc *
```

```
/Users/pbanerjee/anaconda3/lib/python3.7/site-packages/multiqc/utils/config.py:45:
YAMLLoadWarning: calling yaml.load() without Loader=... is deprecated, as the defa
ult Loader is unsafe. Please read https://msg.pyyaml.org/load for full details.
  configs = yaml.load(f)
/Users/pbanerjee/anaconda3/lib/python3.7/site-packages/multiqc/utils/config.py:51:
YAMLLoadWarning: calling yaml.load() without Loader=... is deprecated, as the defa
ult Loader is unsafe. Please read https://msg.pyyaml.org/load for full details.
  sp = yaml.load(f)
/Users/pbanerjee/anaconda3/bin/multiqc:229: DeprecationWarning: The 'warn' method
is deprecated, use 'warning' instead
  logger.warn('MultiQC Version {} now available!'.format(remote_version))
[WARNING]         multiqc : MultiQC Version v1.8 now available!
[INFO   ]         multiqc : This is MultiQC v1.7
[INFO   ]         multiqc : Template    : default
[INFO   ]         multiqc : Searching 'Deseq2.ipynb'
[INFO   ]         multiqc : Searching 'Hisat2_Htseq_pipeline.ipynb'
/Users/pbanerjee/anaconda3/bin/multiqc:495: DeprecationWarning: The 'warn' method
is deprecated, use 'warning' instead
  logger.warn("No analysis results found. Cleaning up..")
[WARNING]         multiqc : No analysis results found. Cleaning up..
[INFO   ]         multiqc : MultiQC complete
```

Display Multiqc Report

In [4]:
```
HTML("/Users/pbanerjee/Documents/Test_Data/multiqc_report.html")
```

Out[4]:

v1.7

Loading report..

General Stats

FastQC

- Sequence Counts
- Sequence Quality Histograms
- Per Sequence Quality Scores
- Per Base Sequence Content
- Per Sequence GC Content
- Per Base N Content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Adapter Content

Hisat2

In [19]:
```
%%bash

for file in $(ls *read1.fastq.gz)
do
  r1=$file;
  r2=${file/read1.fastq.gz/}read2.fastq.gz
  echo "My file name is $file"
  echo "Running ----HISAT2-----"
  hisat2 --threads 8 --time -x /Users/pbanerjee/Documents/Test_Data/test_genome/chr2
2_ERCC92 -1 $r1 -2 $r2 -S aligned_$file.sam --summary-file summary_$file.txt
  echo "The output is aligned_$file.sam"
done
echo "------------------------------Finished Running Hisat2
------------------------------"
```

```
My file name is HBR_Rep1_ERCC-Mix2_Build37-ErccTranscripts-chr22.read1.fastq.gz
Running ----HISAT2-----
The output is aligned_HBR_Rep1_ERCC-Mix2_Build37-ErccTranscripts-chr22.read1.fast
q.gz.sam
------------------------------Finished Running Hisat2
------------------------------
HBR_Rep1_ERCC-Mix2_Build37-ErccTranscripts-chr22.read1.fastq.gz
HBR_Rep1_ERCC-Mix2_Build37-ErccTranscripts-chr22.read1_fastqc.html
HBR_Rep1_ERCC-Mix2_Build37-ErccTranscripts-chr22.read1_fastqc.zip
HBR_Rep1_ERCC-Mix2_Build37-ErccTranscripts-chr22.read2.fastq.gz
HBR_Rep1_ERCC-Mix2_Build37-ErccTranscripts-chr22.read2_fastqc.html
HBR_Rep1_ERCC-Mix2_Build37-ErccTranscripts-chr22.read2_fastqc.zip
aligned_HBR_Rep1_ERCC-Mix2_Build37-ErccTranscripts-chr22.read1.fastq.gz.sam
build.log
multiqc_data
multiqc_data_1
multiqc_report.html
multiqc_report_1.html
read1
read2
summary_HBR_Rep1_ERCC-Mix2_Build37-ErccTranscripts-chr22.read1.fastq.gz.txt
test_genome

Time loading forward index: 00:00:00
Time loading reference: 00:00:00
Multiseed full-index search: 00:00:02
237372 reads; of these:
  237372 (100.00%) were paired; of these:
    727 (0.31%) aligned concordantly 0 times
    235628 (99.27%) aligned concordantly exactly 1 time
    1017 (0.43%) aligned concordantly >1 times
    ----
    727 pairs aligned concordantly 0 times; of these:
      626 (86.11%) aligned discordantly 1 time
    ----
    101 pairs aligned 0 times concordantly or discordantly; of these:
      202 mates make up the pairs; of these:
        93 (46.04%) aligned 0 times
        101 (50.00%) aligned exactly 1 time
        8 (3.96%) aligned >1 times
99.98% overall alignment rate
Time searching: 00:00:02
Overall time: 00:00:02
```

Samtools Sort and Convert BAM

In [23]:
```
%%bash

for file in $(ls aligned_*)
do
  echo "Running ----SAMTOOLS SORT AND CONVERSION TO BAM----"
  echo "My file name is $file"
  samtools sort -o sorted_$file.bam -O bam $file
  echo "The output is sorted_$file"

done
echo "-------------------------------Finished sorting and converting files with SAMTO
OLS-------------------------------"
```

```
Running ----SAMTOOLS SORT AND CONVERSION TO BAM----
My file name is aligned_HBR_Rep1_ERCC-Mix2_Build37-ErccTranscripts-chr22.read1.fas
tq.gz.sam
The output is sorted_aligned_HBR_Rep1_ERCC-Mix2_Build37-ErccTranscripts-chr22.read
1.fastq.gz.sam
-------------------------------Finished sorting and converting files with SAMTOOLS
-------------------------------
HBR_Rep1_ERCC-Mix2_Build37-ErccTranscripts-chr22.read1.fastq.gz
HBR_Rep1_ERCC-Mix2_Build37-ErccTranscripts-chr22.read1_fastqc.html
HBR_Rep1_ERCC-Mix2_Build37-ErccTranscripts-chr22.read1_fastqc.zip
HBR_Rep1_ERCC-Mix2_Build37-ErccTranscripts-chr22.read2.fastq.gz
HBR_Rep1_ERCC-Mix2_Build37-ErccTranscripts-chr22.read2_fastqc.html
HBR_Rep1_ERCC-Mix2_Build37-ErccTranscripts-chr22.read2_fastqc.zip
aligned_HBR_Rep1_ERCC-Mix2_Build37-ErccTranscripts-chr22.read1.fastq.gz.sam
build.log
multiqc_data
multiqc_data_1
multiqc_report.html
multiqc_report_1.html
read1
read2
sorted_aligned_HBR_Rep1_ERCC-Mix2_Build37-ErccTranscripts-chr22.read1.fastq.gz.sam
sorted_aligned_HBR_Rep1_ERCC-Mix2_Build37-ErccTranscripts-chr22.read1.fastq.gz.sa
m.bam
summary_HBR_Rep1_ERCC-Mix2_Build37-ErccTranscripts-chr22.read1.fastq.gz.txt
test_genome
```

Htseq Counts

In [24]:
```
%%bash

for file in $(ls *bam)
do
  echo "Running ----HTSEQ COUNTS----"
  echo "My file name is $file"
  htseq-count --format bam --order pos -t exon $file /Users/pbanerjee/Documents/Test
_Data/griffithlab_brain_vs_uhr/GRCh38_Ens87_chr22_ERCC/genes_chr22_ERCC92.gtf > coun
ts_$file.txt
  echo "The output is counts_$file.txt"
done
echo "------------------------------Finished HTSEQ Counts
------------------------------"
echo "------------------------------Pipeline Over------------------------------"
```

```
Running ----HTSEQ COUNTS----
My file name is sorted_aligned_HBR_Rep1_ERCC-Mix2_Build37-ErccTranscripts-chr22.re
ad1.fastq.gz.sam.bam
The output is counts_sorted_aligned_HBR_Rep1_ERCC-Mix2_Build37-ErccTranscripts-chr
22.read1.fastq.gz.sam.bam.txt
------------------------------Finished HTSEQ Counts------------------------------
------------------------------Pipeline Over------------------------------
HBR_Rep1_ERCC-Mix2_Build37-ErccTranscripts-chr22.read1.fastq.gz
HBR_Rep1_ERCC-Mix2_Build37-ErccTranscripts-chr22.read1_fastqc.html
HBR_Rep1_ERCC-Mix2_Build37-ErccTranscripts-chr22.read1_fastqc.zip
HBR_Rep1_ERCC-Mix2_Build37-ErccTranscripts-chr22.read2.fastq.gz
HBR_Rep1_ERCC-Mix2_Build37-ErccTranscripts-chr22.read2_fastqc.html
HBR_Rep1_ERCC-Mix2_Build37-ErccTranscripts-chr22.read2_fastqc.zip
aligned_HBR_Rep1_ERCC-Mix2_Build37-ErccTranscripts-chr22.read1.fastq.gz.sam
counts_sorted_aligned_HBR_Rep1_ERCC-Mix2_Build37-ErccTranscripts-chr22.read1.fast
q.gz.sam.bam.txt
griffithlab_brain_vs_uhr
griffithlab_brain_vs_uhr.tar.gz
multiqc_data
multiqc_data_1
multiqc_report.html
multiqc_report_1.html
read1
read2
sorted_aligned_HBR_Rep1_ERCC-Mix2_Build37-ErccTranscripts-chr22.read1.fastq.gz.sam
sorted_aligned_HBR_Rep1_ERCC-Mix2_Build37-ErccTranscripts-chr22.read1.fastq.gz.sa
m.bam
summary_HBR_Rep1_ERCC-Mix2_Build37-ErccTranscripts-chr22.read1.fastq.gz.txt
test_genome

56934 GFF lines processed.
100000 SAM alignment record pairs processed.
200000 SAM alignment record pairs processed.
Warning: Mate records missing for 1402 records; first such record: <SAM_Alignment
object: Paired-end read 'HWI-ST718_146963544:7:2214:6503:36074' aligned to 22:[175
92413,17592513)/+>.
238073 SAM alignment pairs processed.
```

Run Multiqc Again on all Files

In [38]:
```
!multiqc *
```

```
/Users/pbanerjee/anaconda3/lib/python3.7/site-packages/multiqc/utils/config.py:45:
YAMLLoadWarning: calling yaml.load() without Loader=... is deprecated, as the defa
ult Loader is unsafe. Please read https://msg.pyyaml.org/load for full details.
  configs = yaml.load(f)
/Users/pbanerjee/anaconda3/lib/python3.7/site-packages/multiqc/utils/config.py:51:
YAMLLoadWarning: calling yaml.load() without Loader=... is deprecated, as the defa
ult Loader is unsafe. Please read https://msg.pyyaml.org/load for full details.
  sp = yaml.load(f)
/Users/pbanerjee/anaconda3/bin/multiqc:229: DeprecationWarning: The 'warn' method
is deprecated, use 'warning' instead
  logger.warn('MultiQC Version {} now available!'.format(remote_version))
[WARNING]         multiqc : MultiQC Version v1.8 now available!
[INFO   ]         multiqc : This is MultiQC v1.7
[INFO   ]         multiqc : Template    : default
[INFO   ]         multiqc : Searching 'HBR_Rep1_ERCC-Mix2_Build37-ErccTranscripts-
chr22.read1.fastq.gz'
[INFO   ]         multiqc : Searching 'HBR_Rep1_ERCC-Mix2_Build37-ErccTranscripts-
chr22.read1_fastqc.html'
[INFO   ]         multiqc : Searching 'HBR_Rep1_ERCC-Mix2_Build37-ErccTranscripts-
chr22.read1_fastqc.zip'
[INFO   ]         multiqc : Searching 'HBR_Rep1_ERCC-Mix2_Build37-ErccTranscripts-
chr22.read2.fastq.gz'
[INFO   ]         multiqc : Searching 'HBR_Rep1_ERCC-Mix2_Build37-ErccTranscripts-
chr22.read2_fastqc.html'
[INFO   ]         multiqc : Searching 'HBR_Rep1_ERCC-Mix2_Build37-ErccTranscripts-
chr22.read2_fastqc.zip'
[INFO   ]         multiqc : Searching 'aligned_HBR_Rep1_ERCC-Mix2_Build37-ErccTran
scripts-chr22.read1.fastq.gz.sam'
[INFO   ]         multiqc : Searching 'counts_sorted_aligned_HBR_Rep1_ERCC-Mix2_Bu
ild37-ErccTranscripts-chr22.read1.fastq.gz.sam.bam.txt'
[INFO   ]         multiqc : Searching 'griffithlab_brain_vs_uhr'
[INFO   ]         multiqc : Searching 'griffithlab_brain_vs_uhr.tar.gz'
[INFO   ]         multiqc : Searching 'multiqc_data'
[INFO   ]         multiqc : Searching 'multiqc_data_1'
[INFO   ]         multiqc : Searching 'multiqc_data_2'
[INFO   ]         multiqc : Searching 'multiqc_data_3'
[INFO   ]         multiqc : Searching 'multiqc_data_4'
[INFO   ]         multiqc : Searching 'multiqc_report.html'
[INFO   ]         multiqc : Searching 'multiqc_report_1.html'
[INFO   ]         multiqc : Searching 'multiqc_report_2.html'
[INFO   ]         multiqc : Searching 'multiqc_report_3.html'
[INFO   ]         multiqc : Searching 'multiqc_report_4.html'
[INFO   ]         multiqc : Searching 'read1'
[INFO   ]         multiqc : Searching 'read2'
[INFO   ]         multiqc : Searching 'sorted_aligned_HBR_Rep1_ERCC-Mix2_Build37-E
rccTranscripts-chr22.read1.fastq.gz.sam'
[INFO   ]         multiqc : Searching 'sorted_aligned_HBR_Rep1_ERCC-Mix2_Build37-E
rccTranscripts-chr22.read1.fastq.gz.sam.bam'
[INFO   ]         multiqc : Searching 'summary_HBR_Rep1_ERCC-Mix2_Build37-ErccTran
scripts-chr22.read1.fastq.gz.txt'
[INFO   ]         multiqc : Searching 'test_genome'
Searching 70 files..  [####################################]  100%
[INFO   ]           htseq : Found 1 reports
[INFO   ]         bowtie2 : Found 1 reports
[INFO   ]          fastqc : Found 2 reports
[INFO   ]         multiqc : Compressing plot data
[WARNING]         multiqc : Previous MultiQC output found! Adjusting filenames..
[WARNING]         multiqc : Use -f or --force to overwrite existing reports instea
d
[INFO   ]         multiqc : Report      : multiqc_report_5.html
```

Display Summary Multiqc Report

In [5]:
```
HTML('/Users/pbanerjee/Documents/Test_Data/multiqc_report_4.html')
```

Out[5]:

v1.7

Loading report..

- General Stats
- HTSeq Count
- Bowtie 2
- FastQC
  - Sequence Counts
  - Sequence Quality Histograms
  - Per Sequence Quality Scores
  - Per Base Sequence Content
  - Per Sequence GC Content
  - Per Base N Content
  - Sequence Length Distribution
  - Sequence Duplication Levels
  - Overrepresented sequences
  - Adapter Content

In [ ]: