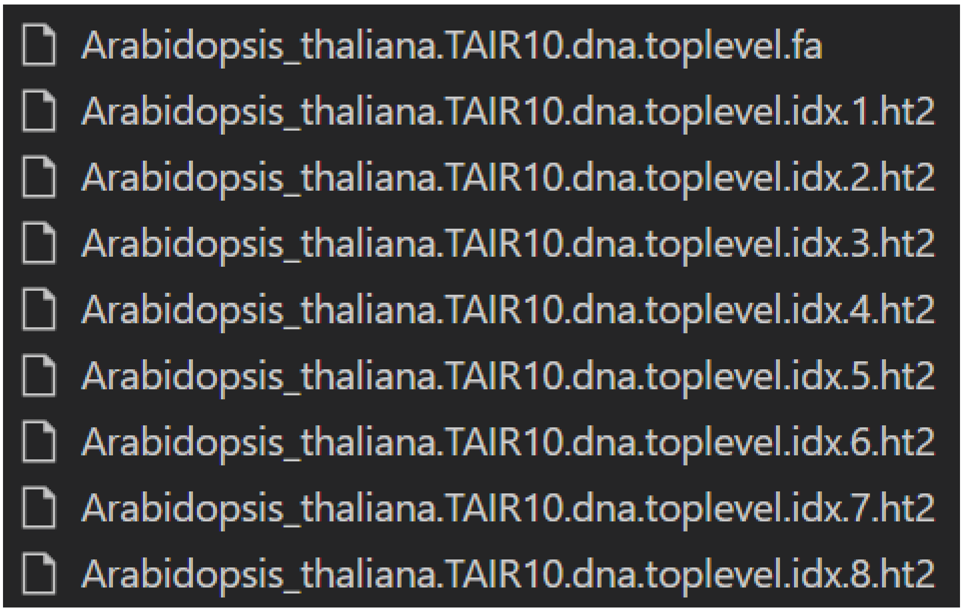
Download data

- download .sra file in my own computer
  - use FileZilla to transfer them to the Nebula
  - convert .sra to .fastq
- download Arabidopsis_thaliana.TAIR10.dna.toplevel.fa in my own computer
  - use FileZilla to transfer them to the Nebula

1.1 build HISAT2 index from .fa file

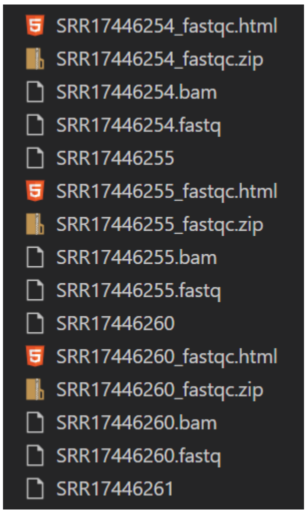
`hisat2-build Arabidopsis_thaliana.TAIR10.dna.toplevel.fa Arabidopsis_thaliana.TAIR10.dna.toplevel.idx`



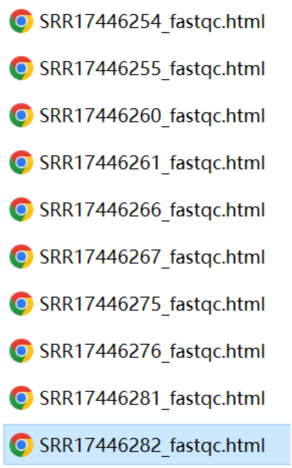
```
Arabidopsis_thaliana.TAIR10.dna.toplevel.fa
Arabidopsis_thaliana.TAIR10.dna.toplevel.idx.1.ht2
Arabidopsis_thaliana.TAIR10.dna.toplevel.idx.2.ht2
Arabidopsis_thaliana.TAIR10.dna.toplevel.idx.3.ht2
Arabidopsis_thaliana.TAIR10.dna.toplevel.idx.4.ht2
Arabidopsis_thaliana.TAIR10.dna.toplevel.idx.5.ht2
Arabidopsis_thaliana.TAIR10.dna.toplevel.idx.6.ht2
Arabidopsis_thaliana.TAIR10.dna.toplevel.idx.7.ht2
Arabidopsis_thaliana.TAIR10.dna.toplevel.idx.8.ht2
```

1.2 fastqc

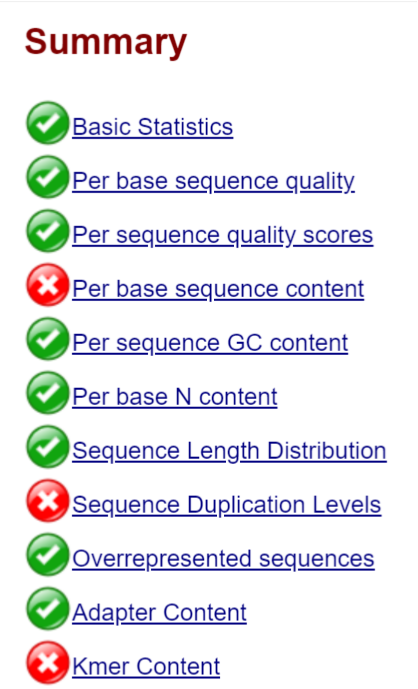
- 1.2.1 use sra toolkit to convert .sra to .fastq
- 1.2.2 use fastqc to do quality control
  - `module load fastqc`
  - `fastqc -t 8 RNA_seq/SRR17446254.fastq -o RNA_seq/`
- 1.2.3: omit, because all the .html files show that the quality is acceptable



SRR17446254_fastqc.html
SRR17446254_fastqc.zip
SRR17446254.bam
SRR17446254.fastq
SRR17446255
SRR17446255_fastqc.html
SRR17446255_fastqc.zip
SRR17446255.bam
SRR17446255.fastq
SRR17446260
SRR17446260_fastqc.html
SRR17446260_fastqc.zip
SRR17446260.bam
SRR17446260.fastq
SRR17446261



SRR17446254_fastqc.html
SRR17446255_fastqc.html
SRR17446260_fastqc.html
SRR17446261_fastqc.html
SRR17446266_fastqc.html
SRR17446267_fastqc.html
SRR17446275_fastqc.html
SRR17446276_fastqc.html
SRR17446281_fastqc.html
SRR17446282_fastqc.html

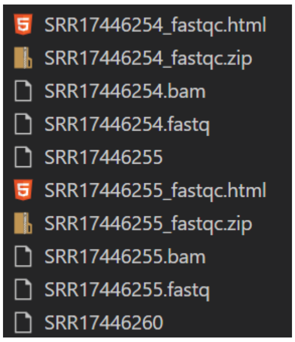


Summary

✓ Basic Statistics
✓ Per base sequence quality
✓ 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
✗ Kmer Content

1.3 use HISAT2 to compare

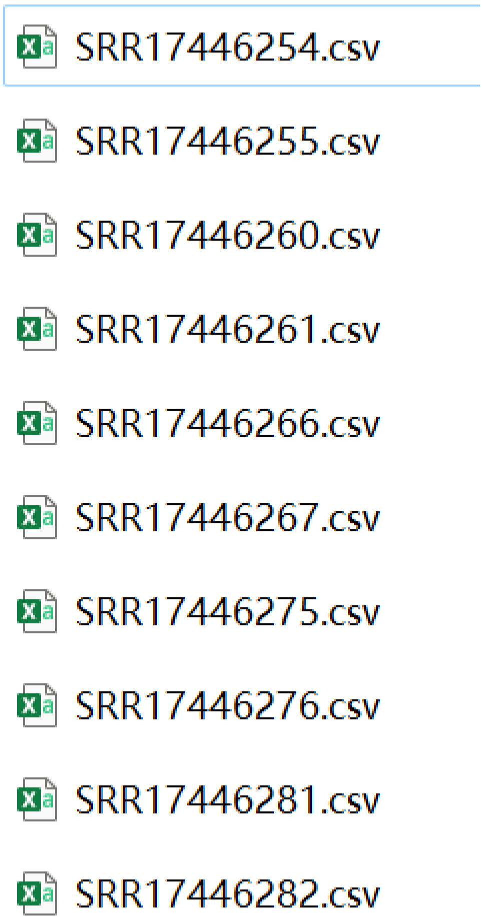
```
hisat2 -q --rna-strandness R -x Arabidopsis_thaliana.TAIR10.dna.toplevel.idx -U
$i.fastq | samtools sort -o $i.bam
```



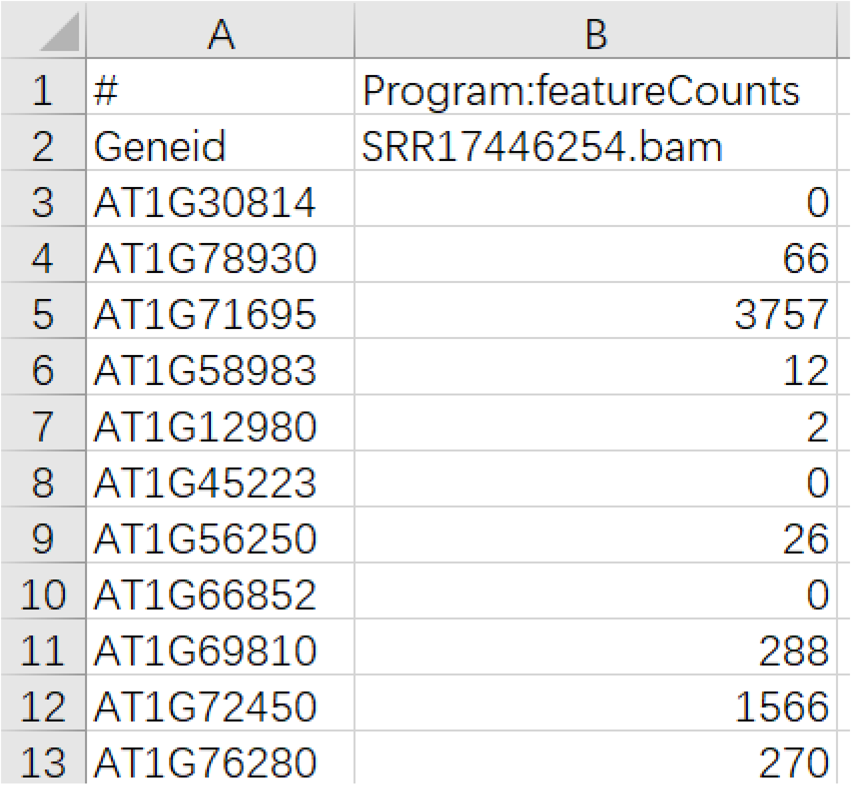
SRR17446254_fastqc.html
SRR17446254_fastqc.zip
SRR17446254.bam
SRR17446254.fastq
SRR17446255
SRR17446255_fastqc.html
SRR17446255_fastqc.zip
SRR17446255.bam
SRR17446255.fastq
SRR17446260

1.4 use featureCounts to count

```
featureCounts -a Arabidopsis_thaliana.TAIR10.52.gtf -o SRR17446255.txt
SRR17446255.bam
cut -f 1,7 $i.txt | awk 'BEGIN {OFS=","} {print $1, $2}' > $i.csv
```



SRR17446254.csv
SRR17446255.csv
SRR17446260.csv
SRR17446261.csv
SRR17446266.csv
SRR17446267.csv
SRR17446275.csv
SRR17446276.csv
SRR17446281.csv
SRR17446282.csv



	A	B
1	#	Program:featureCounts
2	Geneid	SRR17446254.bam
3	AT1G30814	0
4	AT1G78930	66
5	AT1G71695	3757
6	AT1G58983	12
7	AT1G12980	2
8	AT1G45223	0
9	AT1G56250	26
10	AT1G66852	0
11	AT1G69810	288
12	AT1G72450	1566
13	AT1G76280	270