**Cyverse shared folder: “2017\_Arabidopsis”**

“Fastq” subfolder (raw fastq files):

2CL1

2CL2

4CL1

4CL2

2WL1

2WL2

2WL4

4WL3

2CL3

2CL4

4CL3

4CL4

“Reference\_Seqs” subfolder:

Athaliana\_167\_TAIR9.fa.gz

Athaliana\_167\_TAIR10.gene\_exons.gff3.gz

(assembly and annotation files, downloaded from Phytozome V12 on 3/8/17 4pm; <http://genome.jgi.doe.gov/pages/dynamicOrganismDownload.jsf?organism=Phytozome>)

Athaliana\_167\_TAIR10.gene\_exons.gtf

(gff3 file converted to gtf format, for making the “spliceSites” file)

Athaliana\_167\_TAIR10\_spliceSites.txt

(splice sites used by HISAT2 mapper)

Athaliana\_167\_TAIR9.#.ht2

(HISAT2 indices)

ERCC92.fa

ERCC92.gtf

ERCC92.#.ht2

(HISAT2 ERCC indices)

**Cyverse Atmosphere Image with tools needed to run the pipeline: “**Ploidy-Seq”

**Programs used in pipeline:**

Trimmomatic: <http://www.usadellab.org/cms/?page=trimmomatic>

Fastqc: <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>

HISAT2: <https://ccb.jhu.edu/software/hisat2/index.shtml>

Cufflinks gffread: <http://cole-trapnell-lab.github.io/cufflinks/file_formats/#the-gffread-utility>

HTSeq: <http://www-huber.embl.de/HTSeq/doc/overview.html>

**Pipeline:**

1. Quality Filter and remove adapters w/ Trimmomatic (note that barcodes have already removed from Fastq files)

**java -jar usr/local/bin/trimmomatic-0.35.jar SE -phred33 <libname.fastq.gz> <libname\_Trimmo.fq.gz> ILLUMINACLIP:usr/local/bin/Trimmomatic-0.35/adapters/TruSeq3-SE.fa:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36**

1. Run FastQC on all trimmed, clipped, & filtered files

**Fastqc <libname.fq.gz>**

1. Map reads onto Athaliana reference genome using Hisat2

**hisat2 -p 8 -k 10 -x Athaliana\_167\_TAIR9 -U <libname\_Trimmo.fq> --known-splicesite-infile Athaliana\_167\_TAIR10\_spliceSites.txt --no-unal -t -S <libname\_At\_HS2.sam>**

[-p 8] use 4 cpus in parallel

[-k 10] find up to 10 distinct alignments

[-x] base name of indexed reference

[-U] input fastq file

[-S] name of output SAM file

[--known-splicesite-infile] provide file of known splice sites to use for alignment

[--no-unal] suppress SAM records for reads that fail to align

1. Map reads onto ERCC reference genome using Hisat2

**hisat2 -p 8 -k 10 -x ERCC92 -U <filename\_Trimmo.fq> --no-unal -t -S <libname\_ERCC\_HS2.sam>**

1. Count reads by gene using HTSeq

**python -m HTSeq.scripts.count -m intersection-nonempty -s no -t gene -i ID <libname\_At\_HS2.sam> Athaliana\_167\_TAIR10.gene\_exons.gff3 > <libname\_At\_HS2\_HTSeq.txt>**

**python -m HTSeq.scripts.count -m intersection-nonempty -s no -t gene -i ID <libname\_ERCC\_HS2.sam> ERCC92.gtf > <libname\_ERCC\_HS2\_HTSeq.txt>**

1. Remaining steps in R