Scratch notes for Hesperis

**Collection notes:**

Note from Art 7/2019

Here are the flower photos for Maddy samples (Link to images of samples [here](https://umich.box.com/s/8q7nl5at1a5z2snslxfcqsw3ft1wwadi)).

\*Several of these individuals may be the mother for some of the seeds I sent you last week.

\*These are all from a single population near the main entrance at KSR that includes all four colour morphs, plus a number of variagated forms.

\*The photos were taken in late June, and the seeds collected in mid-September.

Samples sent to Gina from Art 7/2019:

All samples [for transcriptome sequencing] were collected at KSR (same population as the seeds I sent). And all were photographed. I will have to check for sure, but Maddy May have used the parents of the seeds I sent you. [Seeds sent from Art below]



**Transcriptome** **sample notes:**

**The below samples are from floral tissues and are saved under**

**/scratch/rsbaucom\_flux/rsbaucom/Hesperis/floral**

And backed up raw data as:

/nfs/turbo/rsbaucom/lab/Hesperis/floral \*\*need to transfer data from scratch to turbo

Sample Name # Reads # Bases Avg. Qual.

32-W 76,695,868 15,339,173,600 38

37-W 57,442,152 11,488,430,400 38

8-Pu 92,863,672 18,572,734,400 38

9-Pu 71,812,575 14,362,515,000 38

2-Pu 75,407,578 15,081,515,600 38

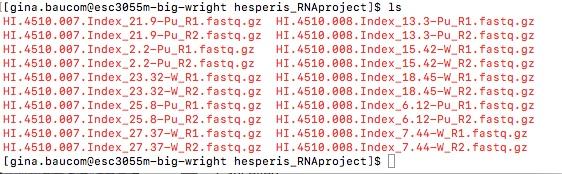
45-W 77,327,644 15,465,528,800 38

3-Pu 68,716,732 13,743,346,400 38

12-Pu 76,133,539 15,226,707,800 38

44-W 78,251,413 15,650,282,600 38

42-W 70,344,396 14,068,879,200 38



After adaptor trimming, changing file names for next steps --IMP

Hesp1=HI.4510.007.Index\_21.9-Pu

Hesp2=HI.4510.007.Index\_2.2-Pu

Hesp3=HI.4510.007.Index\_23.32-W

Hesp4=HI.4510.007.Index\_25.8-Pu

Hesp5=HI.4510.007.Index\_27.37-W \*Gina notes: Hesp 5 clusters with purples, not whites

Hesp6=HI.4510.008.Index\_13.3-Pu

Hesp7=HI.4510.008.Index\_15.42-W

Hesp8=HI.4510.008.Index\_18.45-W

Hesp9=HI.4510.008.Index\_6.12-Pu \*Gina notes: Hesp 9 clusters with whites, not purples

Hesp10=HI.4510.008.Index\_7.44-W

**Hesp1-10F = floral tissue**

**Current steps:**

1. Downloading/backing up, done 7/25
2. Adaptor trimming -- 5% of reads still have adaptors, going to run as script, running 7/25
   1. Cutadapt
   2. Trimmomatic: 7/26
3. Assembly
   1. Trinity -- started 7/27 using 4 samples worth of reads -- Hesp1, 2, 3, and 5
      1. Using fluxm, general rule of thumb is 1gb per m reads!!
      2. Output ~350K contigs
   2. Velvet --
      1. run steps of 2 from 25,27,29,31 (started 8/1)
      2. After figure out best step size, make transcriptome using one white and one purple (Hesp 6 and 7)
      3. Hesp6--multi-kmers (25, 27, 29, 31) and Hesp7--multi-kmers (25, 29)
      4. Combine >350 transcripts from trinity with all from velvet/oases.
4. Merge all via evigene pipeline
   1. Rough goal: 30-60K contigs, N50 ~1200-1500 (not sure about this), other metrics?
   2. Evigene returned ~58K contigs
   3. Ran transrate -- said ~39K were ‘good’
5. Busco to determine which transcriptome is better -- 58K from evigene or 39K from transrate
   1. Busco run suggests ~58 contigs better, 88% singletons id’d
6. Annotation via blast, assignment to GOs? Running blastx, 8/10, but cluster going off-line 8/12, so possibly re-start at new position
7. Mapping via BWA: Mapping in progress, almost done 8/10
8. Counts via samtools: Job sent to cluster and queuing 8/10
9. Expression analysis via edgeR
   1. Ran analysis, 185/183 genes down/up regulated. Currently annotating these DEGs but did not find anything super interesting from initial cursory look.
   2. Removed 2 libraries, #5 and #9, and re-ran analysis. The gene expression of transcripts from these libraries clustered in the wrong direction, so I wondered if their labels were accidentally switched. Decided to drop them and see how this changed expression analysis. Now have 700 Down and 1000 Up transcripts! Haven’t annotated them yet tho.
   3. Annotation id’s usual suspects involved in the production of anthocyanins, plus a host of disease resistance genes and other stuff
   4. Final 58K transcriptome plus results of DESeq run and annotation of ~1700 DEGs can be accessed [here](https://umich.box.com/s/3xovo8xjdx0zezglckn7ct2gnx05eicv).
   5. Annotation of whole 58K set not finished.
10. Might consider another way to reduce transcript number in future
11. No movement with leaf transcriptome.

**INFO ON LEAF TISSUE READS AND ASSEMBLY**

**\*Most work performed by Maddy in Wright lab**

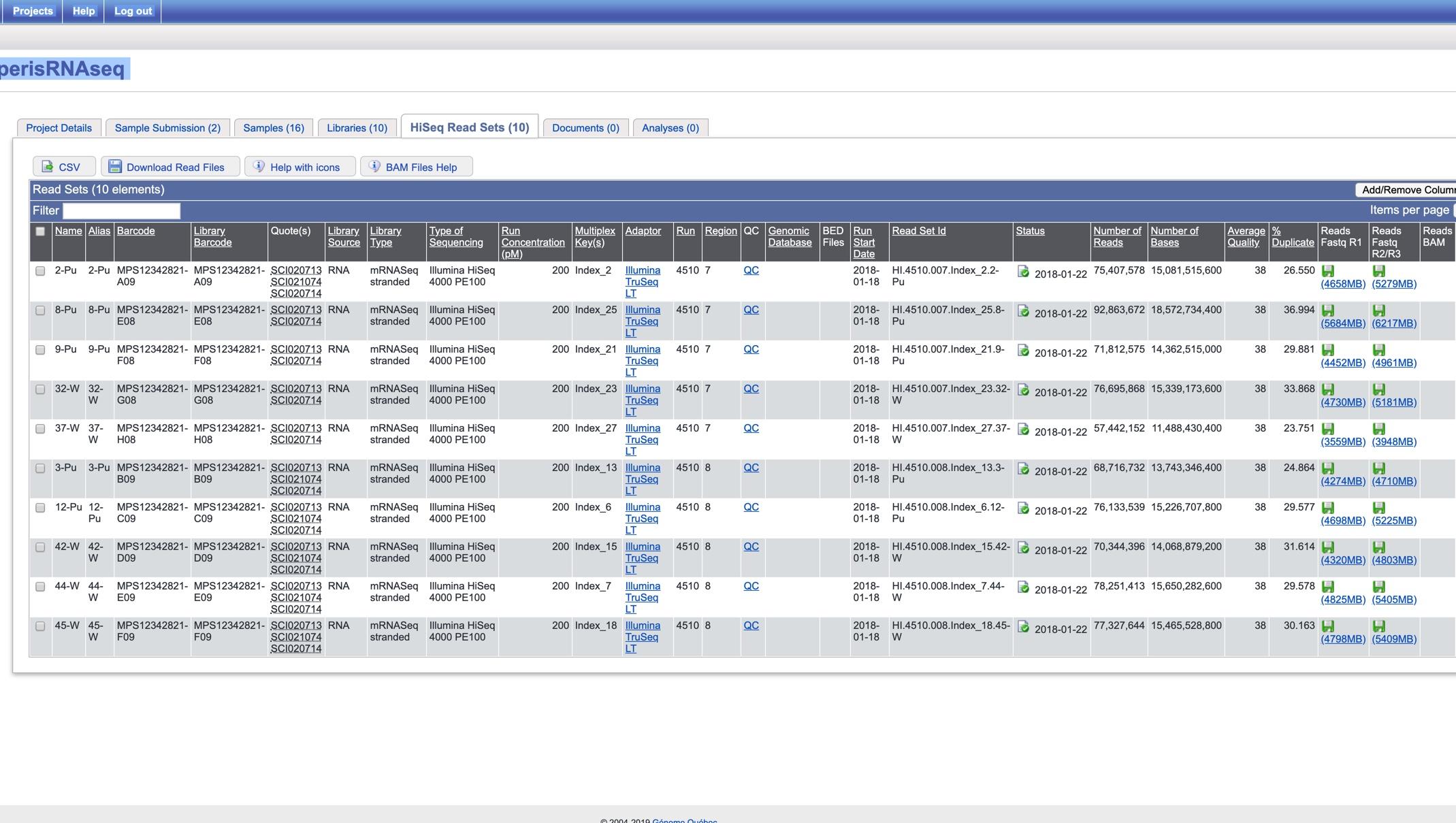
**The below samples are from leaf tissues and are saved under**

**/scratch/rsbaucom\_flux/rsbaucom/Hesperis/leaf**

And backed up raw data as:

/nfs/turbo/rsbaucom/lab/Hesperis/leaf

--These leaf samples are random from a place at KSR, original sequencing job. May be same population as ind used for floral samples but unknown.

****

**\*Maddy used all samples to make a single transcriptome, saved in trinity\_output\_dir**

Info on run saved in Trinity.timing

Trinity Version: Trinity-v2.4.0

Compiler: GCC

Trinity Parameters: --seqType fq --SS\_lib\_type RF

--left /ohta/maddy.chiang/HI.0286.001.Index\_13.Hesp1\_R1.fastq.gz,/ohta/maddy.chiang/HI.0286.001.Index\_15.Hesp10\_R1.fastq.gz,/ohta/maddy.chiang/HI.0286.001.Index\_16.Hesp2\_R1.fastq.gz,/ohta/maddy.chiang/HI.0286.001.Index\_18.Hesp3\_R1.fastq.gz,/ohta/maddy.chiang/HI.0286.001.Index\_19.Hesp4\_R1.fastq.gz,/ohta/maddy.chiang/HI.0286.001.Index\_21.Hesp5\_R1.fastq.gz,/ohta/maddy.chiang/HI.0286.001.Index\_22.Hesp6\_R1.fastq.gz,/ohta/maddy.chiang/HI.0286.001.Index\_23.Hesp7\_R1.fastq.gz,/ohta/maddy.chiang/HI.0286.001.Index\_25.Hesp8\_R1.fastq.gz,/ohta/maddy.chiang/HI.0286.001.Index\_27.Hesp9\_R1.fastq.gz

--right /ohta/maddy.chiang/HI.0286.001.Index\_13.Hesp1\_R2.fastq.gz,/ohta/maddy.chiang/HI.0286.001.Index\_15.Hesp10\_R2.fastq.gz,/ohta/maddy.chiang/HI.0286.001.Index\_16.Hesp2\_R2.fastq.gz,/ohta/maddy.chiang/HI.0286.001.Index\_18.Hesp3\_R2.fastq.gz,/ohta/maddy.chiang/HI.0286.001.Index\_19.Hesp4\_R2.fastq.gz,/ohta/maddy.chiang/HI.0286.001.Index\_21.Hesp5\_R2.fastq.gz,/ohta/maddy.chiang/HI.0286.001.Index\_22.Hesp6\_R2.fastq.gz,/ohta/maddy.chiang/HI.0286.001.Index\_23.Hesp7\_R2.fastq.gz,/ohta/maddy.chiang/HI.0286.001.Index\_25.Hesp8\_R2.fastq.gz,/ohta/maddy.chiang/HI.0286.001.Index\_27.Hesp9\_R2.fastq.gz --CPU 30

--max\_memory 50G --no\_bowtie --verbose --output /ohta/maddy.chiang/trinity\_output\_dir

**Stats from Maddy for trinity\_output\_dir**

################################

## Counts of transcripts, etc.

################################

Total trinity 'genes': 107169

Total trinity transcripts: 366896

Percent GC: 43.11

########################################

Stats based on ALL transcript contigs:

########################################

Contig N10: 3428

Contig N20: 2558

Contig N30: 2033

Contig N40: 1657

Contig N50: 1350

Median contig length: 581

Average contig: 890.53

Total assembled bases: 326733279

#####################################################

## Stats based on ONLY LONGEST ISOFORM per 'GENE':

#####################################################

Contig N10: 3520

Contig N20: 2558

Contig N30: 1968

Contig N40: 1540

Contig N50: 1193

Median contig length: 422

Average contig: 747.17

Total assembled bases: 80073290

**Current steps:**

1. **Downloading/backing up**
2. **Adaptor trimming -- needed**
   1. **Cutadapt -- running 7/25/19**
   2. **Trimmomatic: 7/26**
3. **Correct reads -- skipped!**
   1. **Rcorrector**
4. **Assembly**
   1. **Trinity -- started 7/27**
   2. **Velvet**
5. **Merge via evigene pipeline**