ACROPORA assembly steps:

A. cervicornis

- Pooled together 8 samples, run on NextSeq500, Mid Output Flowcell, across 4 lanes. This was done for paired-end 2x75 bp (which will have 260-400 million reads passing filter). Looks like for RNA, insert size is around 200 bp. So fragment size is 200 + 75 + 75 + adapter. Unknown fragment sequence will be 200-150 = 50?
    - "Illumina sequencing technology uses cluster generation and sequencing by synthesis (SBS) chemistry to sequence millions or billions of clusters on a flow cell, depending on the sequencing platform. During SBS chemistry, for each cluster, base calls are made and stored for every cycle of sequencing by the Real-Time Analysis (RTA) software on the instrument. RTA stores the base call data in the form of individual base call (or BCL) files. When sequencing completes, the base calls in the BCL files must be converted into sequence data. This process is called BCL to FASTQ conversion. A FASTQ file is a text file that contains the sequence data from the clusters that pass filter on a flow cell (for more information on clusters passing filter, see the "additional information" section of this bulletin). If samples were multiplexed, the first step in FASTQ file generation is *demultiplexing*. Demultiplexing assigns clusters to a sample, based on the cluster's index sequence(s). After demultiplexing, the assembled sequences are written to FASTQ files per sample. If samples were not multiplexed, the demultiplexing step does not occur, and, for each flow cell lane, all clusters are assigned to a single sample. For a single-read run, one Read 1 (R1) FASTQ file is created for each sample per flow cell lane. For a paired-end run, one R1 and one Read 2 (R2) FASTQ file is created for each sample for each lane. FASTQ files are compressed and created with the extension **.fastq.gz*."

/gpfs/group/ibb3/default/genome_annotation/Acerv_assemblies
/storage/home/cco38/scratch/217cerv
Hybrid 217cerv

1. Need to concatenate all R1 files and all R2 files to make 2 fastq files to feed into STAR. With the genome index already built by Shiela, I just need to call upon the indexes and align my 2 fastq files to it, inputting both with an *.fastq —> Concatenate with cat *_R1_001.fastq > 217CervR1.fastq

2. Map reads to Cerv genome with:

```
#!/bin/bash
#PBS -l nodes=1:ppn=10:himem
#PBS -l walltime=48:00:00
#PBS -A open
#PBS -l mem=100gb
#PBS -j oe
```

```
#PBS -N s2b

cd $PBS_O_WORKDIR

#export PERL5LIB=/gpfs/group/ibb3/default/tools/perl/lib/perl5
#export AUGUSTUS_CONFIG_PATH=/gpfs/group/ibb3/default/tools/augustus-3.2.3/config
#export GENEMARK_PATH=/gpfs/group/ibb3/default/tools/gm_et_linux_64/gmes_petap
export SAMTOOLS_PATH=/gpfs/group/ibb3/default/tools/samtools-1.9

module purge
module load gcc/5.3.1
module load bamtools/2.4.1
module load samtools/1.5

# Create genome index for cerv
#/gpfs/group/ibb3/default/tools/STAR/bin/Linux_x86_64/STAR --runThreadN 10 --runMode
genomeGenerate \
#--genomeDir /gpfs/group/ibb3/default/genome_annotation/Acerv_assemblies/ \
#--genomeFastaFiles /gpfs/group/ibb3/default/genome_annotation/Acerv_assemblies/
Acerv_assembly_v1.0_171209.masked.fa\

# Map the reads from 217cerv to cerv genome
/gpfs/group/ibb3/default/tools/STAR/bin/Linux_x86_64/STAR --runThreadN 40 \
--genomeDir /gpfs/group/ibb3/default/genome_annotation/Acerv_assemblies --seedPerReadNmax 50000 \
--outSAMtype BAM Unsorted --twopassMode Basic --outFileNamePrefix ./AC217bc_STAR \
--readFilesIn /storage/home/cco38/scratch/217cerv/217CervR1.fastq /storage/home/cco38/scratch/
217cerv/217CervR2.fastq \
--seedSearchStartLmax 0 \
--outFilterScoreMinOverLread 0 \
--outFilterMatchNminOverLread 0 \
--outFilterMatchNmin 0 \
--outFilterMultimapNmax 20

#sort the bam files
#/gpfs/group/ibb3/default/tools/samtools-1.9/samtools sort \
#-m 7G -o /storage/work/khs18/Braker2/ASTR_cut29_sorted.bam \
#-T /storage/work/khs18/Braker2/ASTR_cut29_temp \
#--threads 40 /storage/work/khs18/Braker2/ASTR_cut29_ssslm0Aligned.out.bam
```

Results:

Version 1 (AC217a_STARLog.final.out)

```
Started job on | Sep 25 14:28:00
Started mapping on |   Sep 25 16:05:22
Finished on |    Sep 25 17:40:41
Mapping speed, Million of reads per hour |    100.87
Number of input reads |160246102
Average input read length |  150

UNIQUE READS:
Uniquely mapped reads number |     79511609
Uniquely mapped reads % |    49.62%
Average mapped length |148.35
Number of splices: Total |   29906217
Number of splices: Annotated (sjdb) |    29767016
Number of splices: GT/AG |   29521984
Number of splices: GC/AG |   211566
Number of splices: AT/AC |   17853
Number of splices: Non-canonical | 154814
Mismatch rate per base, % |  0.63%
Deletion rate per base |     0.03%
Deletion average length |    1.68
Insertion rate per base |    0.02%
Insertion average length |   1.71

MULTI-MAPPING READS:
Number of reads mapped to multiple loci |     5782961
% of reads mapped to multiple loci |     3.61%
Number of reads mapped to too many loci |     149863
% of reads mapped to too many loci |     0.09%

UNMAPPED READS:
% of reads unmapped: too many mismatches |     0.00%
% of reads unmapped: too short |   46.61%
% of reads unmapped: other | 0.07%
```

```
CHIMERIC READS:
Number of chimeric reads |   0
% of chimeric reads |  0.00%
```

Edited, version 2 (AC217b0_STARLog.final.out)

```
Started job on |    Sep 26 12:07:09
Started mapping on |      Sep 26 12:46:41
Finished on |Sep 26 13:29:53
Mapping speed, Million of reads per hour |    222.56
Number of input reads |    160246102
Average input read length |      150

UNIQUE READS:
Uniquely mapped reads number |    111521331
Uniquely mapped reads % |  69.59%
Average mapped length |    117.31
Number of splices: Total |30297336
Number of splices: Annotated (sjdb) |   30146935
Number of splices: GT/AG |29898870
Number of splices: GC/AG |212296
Number of splices: AT/AC |18642
Number of splices: Non-canonical |      167528
Mismatch rate per base, % |      1.03%
Deletion rate per base |  0.03%
Deletion average length | 1.72
Insertion rate per base | 0.03%
Insertion average length |1.71

MULTI-MAPPING READS:
Number of reads mapped to multiple loci |      45617626
% of reads mapped to multiple loci |    28.47%
Number of reads mapped to too many loci |      3012954
% of reads mapped to too many loci |    1.88%

UNMAPPED READS:
% of reads unmapped: too many mismatches |     0.00%
% of reads unmapped: too short | 0.00%
% of reads unmapped: other |     0.06%

CHIMERIC READS:
Number of chimeric reads |0
% of chimeric reads |     0.00%
```

Edited, version 3 (AC217bc_STARLog.final.out)

```
Started job on |    Oct 07 15:50:20
Started mapping on |      Oct 07 16:29:48
Finished on |Oct 07 17:13:20
Mapping speed, Million of reads per hour |    220.86

Number of input reads |    160246102
Average input read length |      150
UNIQUE READS:
Uniquely mapped reads number |    111526776
Uniquely mapped reads % |  69.60%
Average mapped length |    117.30
Number of splices: Total |30295694
Number of splices: Annotated (sjdb) |   30145340
Number of splices: GT/AG |29897244
Number of splices: GC/AG |212272
Number of splices: AT/AC |18615
Number of splices: Non-canonical |      167563
Mismatch rate per base, % |      1.03%
Deletion rate per base |  0.03%
Deletion average length | 1.72
Insertion rate per base | 0.03%
Insertion average length |1.72
MULTI-MAPPING READS:
Number of reads mapped to multiple loci |      48239751
% of reads mapped to multiple loci |    30.10%
Number of reads mapped to too many loci |      363645
% of reads mapped to too many loci |    0.23%
UNMAPPED READS:
% of reads unmapped: too many mismatches |     0.00%
% of reads unmapped: too short | 0.00%
% of reads unmapped: other |     0.07%
CHIMERIC READS:
Number of chimeric reads |0
% of chimeric reads |     0.00%
```

3. Sort bam files with (set mem to 400gb):

```
#sort the bam files
```

```
/gpfs/group/ibb3/default/tools/samtools-1.9/samtools sort \
-m 7G -o /gpfs/group/ibb3/default/UTGSAF_Hybrid/217/217Cerv/AC217bc_sorted.bam \
-T /gpfs/group/ibb3/default/UTGSAF_Hybrid/217/217Cerv/AC217bc_temp \
--threads 40 /gpfs/group/ibb3/default/UTGSAF_Hybrid/217/217Cerv/AC217bc_STARAligned.out.bam
```

Output file = AC217bc_sorted.bam

4. Trinity genome-guided assembly

Trinity RNA-Seq de novo transcriptome assembly: trinity/2.2.0

Need to load module (GNU Compiler Collection) gcc/5.3.1 (or gcc/7.3.1): module load gcc/7.3.1

```
#!/bin/bash
#PBS -l nodes=1:ppn=40:himem
#PBS -l walltime=48:00:00
#PBS -A open
#PBS -l mem=600gb
#PBS -j oe
#PBS -N s2b

cd $PBS_O_WORKDIR

#export PERL5LIB=/gpfs/group/ibb3/default/tools/perl/lib/perl5
#export AUGUSTUS_CONFIG_PATH=/gpfs/group/ibb3/default/tools/augustus-3.2.3/conf$
#export GENEMARK_PATH=/gpfs/group/ibb3/default/tools/gm_et_linux_64/gmes_petap
#export SAMTOOLS_PATH=/gpfs/group/ibb3/default/tools/samtools-1.9

export BOWTIE2_PATH=/gpfs/group/ibb3/default/tools/bowtie2
module purge
module load gcc/5.3.1
module load trinity/2.2.0

#Run genome-guided Trinity
Trinity --genome_guided_bam /gpfs/group/ibb3/default/UTGSAF_Hybrid/217/217Cerv/AC217bc_sorted.bam \
--genome_guided_max_intron 10000 \
--max_memory 600G --CPU 40
```

Output file = trinity_out_dir

= took many tries to run successfully since memory and thread was not optimized. This led to hitting the time wall and failing. Also needed to match the server command at the top to the trinity command at the bottom (memory and CPU)

5. BUSCO quality check

```
#!/bin/bash
#PBS -l nodes=1:ppn=20:himem
#PBS -l walltime=48:00:00
#PBS -A open
#PBS -l mem=100gb
#PBS -j oe
#PBS -N bwa

cd $PBS_O_WORKDIR

conda activate base

/storage/work/khs18/anaconda3/bin/run_BUSCO.py -i /gpfs/group/ibb3/default/UTGSAF_Hybrid/trinity_out_dir/Trini$
-l /storage/home/khs18/scratch/BUSCO/metazoa_odb9 -m tran
```

Results:

C:92.9%[S:57.6%,D:35.3%],F:3.7%,M:3.4%,n:978

908	Complete BUSCOs (C)
563	Complete and single-copy BUSCOs (S)
345	Complete and duplicated BUSCOs (D)
36	Fragmented BUSCOs (F)

34 Missing BUSCOs (M)
978 Total BUSCO groups searched

6. Trinity quality check:

```
/opt/aci/sw/trinity/2.2.0_gcc-5.3.1/util/TrinityStats.pl Trinity.fasta
```

```
################################
## Counts of transcripts, etc.
################################
Total trinity 'genes':   784479
Total trinity transcripts:   837465
Percent GC: 48.85

########################################
Stats based on ALL transcript contigs:
########################################

Contig N10: 4035
Contig N20: 2294
Contig N30: 1122
Contig N40: 566
Contig N50: 393

Median contig length: 260
Average contig: 415.86
Total assembled bases: 348272093


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

Contig N10: 2800
Contig N20: 1052
Contig N30: 546
Contig N40: 397
Contig N50: 327

Median contig length: 256
Average contig: 360.50
Total assembled bases: 282801155
```

7.

average gene size is 1000 basepairs, so average transcript length will be shorter
Genes needs to be less than 40,000
How I sit calculating gene number in trinity.
LOOK AT FILTERING STEPS IN OTHER SYMBIOTIC ORGANSIMS.

Next step: get rid of sym transcriptomes reads
So compare my transcriptome to the ones in genome resources older
Or microbial datasets.

- make a blast database of concatenation of all the coral hosts I would think it would have hits do. Do the same to the symbiont/microbial database
-blast my assembly against that, filter out what matches symbiont database.
Use sheila’s method in 2015 https://www.g3journal.org/content/ggg/5/11/2441.full.pdf
Can use genome databases for palm and cerv as host database.
Add in other transcriptome datas