TAMU HPRC Short Courses – NGS Assembly – Noushin Ghaffari, PhD

Source/additional information:

https://github.com/trinityrnaseq/BerlinTrinityWorkshop2017/wiki/trinity_assembly

==========

Set up:

```
echo $SCRATCH
cd $SCRATCH
mkdir NGS_assembly_Oct18
mkdir NGS_assembly_Oct18/Data
mkdir NGS_assembly_Oct18/Scripts
mkdir NGS_assembly_Oct18/Outputs
cp /scratch/training/NGS_assembly/Data/Fastq_files/*.fastq $SCRATCH/NGS_assembly_Oct18/Data
cd $SCRATCH/NGS_assembly_Oct18/Data
ls -l
head -n 16 GSNO_rep1_1.fastq
```

==========

Running Trinity on TAMU HPRC Ada System:

https://hprc.tamu.edu/wiki/Ada:NGS:RNA-seq#Sample_Trinity_Paired_End_Assembly_job_Scripts

```
module spider Trinity
module load Trinity/2.8.3-GCCcore-6.3.0-Python-2.7.12-bare
```

++

Running Trinity for SMALL input:

```
head -n400000 GSNO_rep1_1.fastq > left_GSNO.100k.fastq
head -n400000 GSNO_rep1_2.fastq > right_GSNO.100k.fastq

head -n400000 ph8_rep1_1.fastq > left_ph8.100k.fastq
head -n400000 ph8_rep1_2.fastq > right_ph8.100k.fastq

cp /scratch/training/NGS_assembly/Scripts/Trinity_GSNO_ph8_100K.sh $SCRATCH/NGS_assembly_Oct18/Scripts

cd $SCRATCH/NGS_assembly_Oct18/Scripts
```

```
cat Trinity_GSNO_ph8_100K.sh

bsub < Trinity_GSNO_ph8_100K.sh

bjobs

++++++++++++++++++++++++++++++++++++++++

Running Trinity for ALL the input reads:

cp /scratch/training/NGS_assembly/Scripts/Trinity_All.sh $SCRATCH/
NGS_assembly_Oct18/Scripts/

cd $SCRATCH/NGS_assembly_Oct18/Scripts

cat Trinity_All.sh

bsub < Trinity_All.sh

bjobs


================================================================================
==============


Assembly QC:


- Counting the output transcripts:

grep '>' /scratch/user/noushin/NGS_assembly_Oct18/Outputs/
Trinity_Output_GSNO_ph8_100K/Trinity.fasta | wc -l


- Looking at the STAT file:

cat /path_to_stat_file/Trinity_stats_GSNO_ph8.txt

cat /path_to_stat_file/Trinity_all.txt

++++++++++++++++++++++++++++++++++++++++

QC for trinity.fasta based on ALL Data (probably still running, so
let's use the previously assembled output file):

cd $SCRATCH/NGS_assembly_Oct18/Outputs
mkdir All_Data
cd All_Data
```

```
cp /scratch/training/NGS_assembly/Data/workshop_shared/shared/
Trinity.fasta  .


grep '>' $SCRATCH/NGS_assembly_Oct18/Outputs/All_Data/Trinity.fasta |
wc -l
```

- Looking at the STAT:

```
$TRINITY_HOME/util/TrinityStats.pl $SCRATCH/NGS_assembly_Oct18/
Outputs/All_Data/Trinity.fasta > $SCRATCH/NGS_assembly_Oct18/Outputs/
All_Data/All_Data_Stats.txt

cat $SCRATCH/NGS_assembly_Oct18/Outputs/All_Data/All_Data_Stats.txt
```

++

- Mapping reads back to the contigs:

```
cp /scratch/training/NGS_assembly/Scripts/Mapping_GSNO_ph8_100K.sh
$SCRATCH/NGS_assembly_Oct18/Scripts/

bsub < $SCRATCH/NGS_assembly_Oct18/Scripts/Mapping_GSNO_ph8_100K.sh
```

- Visualizing the mapped reads to the contigs, using IGV

Method 1)

```
ssh -X username@ada.tamu.edu

module load IGV

igv.sh

-- load the 'Trinity.fasta' file as a 'genome' via the IGV 'Genomes'-
>'Load Genome from File' menu

-- load in the 'bowtie2.coordSorted.bam' file via the IGV 'File'-
>'Load from File' menu
```

Method 2)

Connect to Open OnDemand in here: portal.hprc.tamu.edu

++

- Estimate the expression levels of the transcripts, creating count

tables, and Making ExN50

```
cp /scratch/training/NGS_assembly/Scripts/Bowtie_RSEM.sh $SCRATCH/
NGS_assembly_Oct18/Scripts/

bsub < $SCRATCH/NGS_assembly_Oct18/Scripts/Bowtie_RSEM.sh

head -n20 /scratch/user/noushin/NGS_assembly_Oct18/Outputs/
Trinity_Output_GSNO_ph8_100K/Trinity_trans.TMM.EXPR.matrix
```

++

- Primary annotation/QC

```
cp /scratch/training/NGS_assembly/Scripts/BlastX_GSNO_ph8.sh $SCRATCH/
NGS_assembly_Oct18/Scripts/

bsub < $SCRATCH/NGS_assembly_Oct18/Scripts/BlastX_GSNO_ph8.sh

cd $SCRATCH/NGS_assembly_Oct18/Outputs/Trinity_Output_GSNO_ph8_100K

head blastx.outfmt6
```