#TRINITY ASSEMBLY GUIDE
#Last edited by Ava Hoffman 9/27/17

#This is a Trinity guide to accompany Hoffman and Smith 'Gene expression differs in co-dominant prairie grasses under drought'. While the guide file names and specific examples are linked to this paper, the methodology is similar across experiments. Please use Trinity's guide for best practices on command parameters.

#Trinity and many other programs will need to be run on a computing cluster where jobs are typically submitted as shell scripts ending in '.sh'. Providing paths within these scripts will ensure your commands find the necessary code. The code below is primarily raw commands. Please see accompanying scripts for submitted commands.

```
export PATH=$PATH:~/rnaseq/trinityrnaseq-2.1.1
export PATH=$PATH:~/rnaseq/trinityrnaseq-2.1.1/Trinity
export PATH=$PATH:~/rnaseq/bowtie-1.1.2
export PATH=$PATH:/usr/local/trinity
export PATH=$PATH:/home/avahoffman/bowtie2-2.2.7
export PATH=$PATH:/home/avahoffman/ncbi-blast-2.3.0+/bin
```

 # 3. Build Trinity

#Download Trinity. Check the github page to see if there are new versions.

```
wget
https://github.com/trinityrnaseq/trinityrnaseq/archive/v2.1.1.zip
unzip v2.1.1
cd trinityrnaseq-2.1.1
# not sure what the following does, but it updates something that is necessary..
sudo apt-get install libncurses5-dev
# and compile and wait a few minutes! At this point, Trinity's github page is very helpful..
https://github.com/trinityrnaseq/trinityrnaseq/wiki/Installing%20Trinity
make
make plugins
#need bowtie; make sure you add it to path (above)
wget https://sourceforge.net/projects/bowtie-bio/files/bowtie/1.1.2/bowtie-1.1.2-linux-x86_64.zip
unzip bowtie-1.1.2-linux-x86_64.zip
# run sample data to make sure Trinity is correctly installed
cd sample_data/test_Trinity_Assembly/
./runMe.sh
# should see "All commands completed successfully. :-)" if it worked.
```

5. Run Trinity

 # trying different trimming parameters. I did trimming using iPlant's discovery environment. For andro 1, I used parameters HEADCROP:12 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:25 MINLEN:30, except SDR8 which had a quality cutoff of 28. Below for andro 2, I made all samples have a quality cutoff of 25, plus a minimum length of 40. I also added strand specificity (--SS_lib_type RF) and minimum coverage (--min_kmer_cov 3).. this does not seem to improve the quality of the assembly. Tried just --SS_lib_type RF option and quality was less than non-strand specificity, with 78% improper pairs for ADR7 later on. Will try --SS_lib_type FR. This is not any better in terms of N50. Will see what the proper/improper looks like. Does not look better. Also does not look better when --SS_lib_type option remove from Bowtie alignment.

#Look at excel file in supplementary data for final params for trimming
 # left = 1 = reverse, right = 2 =forward

#final assembly code will look like this for andro and sorgh

```
~/trinityrnaseq-2.1.1/Trinity --seqType fq --max_memory 60G --
left
TrmPr1_A_DR7_s_5_1_sequence.txt,TrmPr1_A_DR8_s_5_1_sequence.txt,T
rmPr1_A_WW7_s_5_1_sequence.txt,TrmPr1_A_WW8_s_5_1_sequence.txt --
right
TrmPr2_A_DR7_s_5_2_sequence.txt,TrmPr2_A_DR8_s_5_2_sequence.txt,T
rmPr2_A_WW7_s_5_2_sequence.txt,TrmPr2_A_WW8_s_5_2_sequence.txt --
CPU 8 --min_contig_length 300 --output Trinity_andro_out --
full_cleanup
~/trinityrnaseq-2.1.1/Trinity --seqType fq --max_memory 32G --
left
TrmPr1_S_DR7_s_4_1_sequence.txt,TrmPr1_S_DR8_s_4_1_sequence.txt,T
rmPr1_S_WW8_s_4_1_sequence.txt,TrmPr1_A_WW9_s_4_1_sequence.txt --
right
TrmPr2_S_DR7_s_4_2_sequence.txt,TrmPr2_S_DR8_s_4_2_sequence.txt,T
rmPr2_S_WW8_s_4_2_sequence.txt,TrmPr2_S_WW9_s_4_2_sequence.txt --
CPU 16 --min_contig_length 300 --output Trinity_sorgh_out --
full_cleanup
```

 # if you get something like this: "Error, cannot locate file: at /home/avahoffman/trinityrnaseq-2.1.1/Trinity line 2150.
main::create_full_path('ARRAY(0x291a458)', 1) called at /home/avahoffman/trinityrnaseq-2.1.1/Trinity line 1106" I think it means you need more memory ☹, I made it only one file and it started running with no issues.
 # save your fasta files to a local machine for safekeeping

6. Post-Assembly quality checks

 # will give you Nx statistics, total assembled bases, contig stats etc.

```
~/trinityrnaseq-2.1.1/util/TrinityStats.pl Trinity.fasta
```

 # Not all samples achieved the typical percent mapping back to the assembly using Bowtie. Bowtie 2 may help with species with a lot of redundancy (eg, polyploids), although I think the newest version of Trinity is better able to cope with overlaps. http://bowtie-bio.sourceforge.net/bowtie2/manual.shtml. First, build the reference:

```
bowtie2-build Trinity_andro_out.Trinity.02122016.fasta andro-
bowtie2ref
```

 # Dovetail allows ends of reads to hang off each other, but still be considered paired. -p indicates number of threads to use. Very sensitive option makes sure that as reads align to the reference they realign more often (fewer gaps.. check the bowtie2 guide, this is not necessarily intuitive). Realign the paired end data to the reference:

```
#local option means some read ends can be clipped if they don't
align perfectly
bowtie2 --very-sensitive-local --dovetail -p 800 -x andro-
       bowtie2ref -1 ~/andro/TrmPr1_A_DR7_s_5_1_sequence.txt -2
       ~/andro/TrmPr2_A_DR7_s_5_2_sequence.txt -S andro_dr7.sam
bowtie2 --very-sensitive-local --dovetail -p 800 -x andro-
bowtie2ref -1 ~/andro/TrmPr1_A_DR8_s_5_1_sequence.txt -2
~/andro/TrmPr2_A_DR8_s_5_2_sequence.txt -S andro_dr8.sam
bowtie2 --very-sensitive-local --dovetail -p 800 -x andro-
       bowtie2ref -1 ~/andro/TrmPr1_A_WW7_s_5_1_sequence.txt -2
       ~/andro/TrmPr2_A_WW7_s_5_2_sequence.txt -S andro_ww7.sam
bowtie2 --very-sensitive-local --dovetail -p 800 -x andro-
       bowtie2ref -1 ~/andro/TrmPr1_A_WW8_s_5_1_sequence.txt -2
       ~/andro/TrmPr2_A_WW8_s_5_2_sequence.txt -S andro_ww8.sam

#without local option
bowtie2 --very-sensitive --dovetail -p 800 -x
       ~/bowtie_output/andro-bowtie2ref -1
       ~/andro/TrmPr1_A_DR7_s_5_1_sequence.txt -2
       ~/andro/TrmPr2_A_DR7_s_5_2_sequence.txt -S andro_dr7.sam
bowtie2 --very-sensitive --dovetail -p 800 -x
~/bowtie_output/andro-bowtie2ref -1
~/andro/TrmPr1_A_DR8_s_5_1_sequence.txt -2
~/andro/TrmPr2_A_DR8_s_5_2_sequence.txt -S andro_dr8.sam
bowtie2 --very-sensitive --dovetail -p 800 -x
       ~/bowtie_output/andro-bowtie2ref -1
       ~/andro/TrmPr1_A_WW7_s_5_1_sequence.txt -2
       ~/andro/TrmPr2_A_WW7_s_5_2_sequence.txt -S andro_ww7.sam
bowtie2 --very-sensitive --dovetail -p 800 -x
       ~/bowtie_output/andro-bowtie2ref -1
       ~/andro/TrmPr1_A_WW8_s_5_1_sequence.txt -2
       ~/andro/TrmPr2_A_WW8_s_5_2_sequence.txt -S andro_ww8.sam
```

 # Convert sam to bam:

```
samtools view -bS andro_dr7.sam > andro_dr7.bam
samtools view -bS andro_dr8.sam > andro_dr8.bam
samtools view -bS andro_ww7.sam > andro_ww7.bam
samtools view -bS andro_ww8.sam > andro_ww8.bam

# And finally, get your statistics.
http://www.htslib.org/doc/samtools.html

# Flagstat is fast..
# Stats is fast too.

samtools flagstat andro_dr7.bam
samtools flagstat andro_dr8.bam
samtools flagstat andro_ww7.bam
samtools flagstat andro_ww8.bam

samtools stats andro_dr7.bam
samtools stats andro_dr8.bam
samtools stats andro_ww7.bam
samtools stats andro_ww8.bam

# Compare to existing proteins. DL'd on FEB 16 2016
# Possible to run on local machine as well.

wget ftp://ftp.ncbi.nlm.nih.gov/blast/executables/LATEST/ncbi-
blast-2.3.0+-x64-linux.tar.gz
tar zxvpf ncbi-blast-2.3.0+-x64-linux.tar.gz

wget
ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowl
edgebase/complete/uniprot_sprot.fasta.gz
gunzip uniprot_sprot.fasta.gz
makeblastdb -in uniprot_sprot.fasta -dbtype prot

wget
ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowl
edgebase/complete/uniprot_trembl.fasta.gz
gunzip uniprot_sprot.fasta.gz
makeblastdb -in uniprot_trembl.fasta -dbtype prot

# FYI: e-value of -20 means looking for semi-closely related
species. More distant homology should use ~ -10. Note, the next
step takes a long time!!

blastx -query
       /Users/avahoffman/Documents/CSU/Research/RNASEQDATA/Trinit
       y_andro_out.Trinity.02122016.fasta -db uniprot_sprot.fasta
       -out blastx.outfmt6 -evalue 1e-20 -num_threads 6 -
       max_target_seqs 1 -outfmt 6
```

```
#always a good idea to try a subset first to make sure your code
       works
~/trinityrnaseq-2.1.1/util/analyze_blastPlus_topHit_coverage.pl
       blastx.outfmt6 Subset_for_blast.txt uniprot_sprot.fasta

~/trinityrnaseq-2.1.1/util/analyze_blastPlus_topHit_coverage.pl
blastx.outfmt6

~/Atrimdata/trinity_out_dir/Trinity.fasta uniprot_sprot.fasta

blastx -query ~/Strimdata/trinity_out_dir/Trinity.fasta -db
uniprot_sprot.fasta -out blastx.outsorgh -evalue 1e-20 -
num_threads 8 -max_target_seqs 1 -outfmt 6

~/trinityrnaseq-2.1.1/util/analyze_blastPlus_topHit_coverage.pl
blastx.outsorgh

~/Strimdata/trinity_out_dir/Trinity.fasta uniprot_sprot.fasta
```

7. Estimate abundance

```
# need to install RSEM. Use wget command, and run 'make'.
# ensure that PATH points to RSEM and express. Will need to redo
every time you restart terminal.
# can be done on local machine if you have enough memory

wget https://github.com/deweylab/RSEM/archive/v1.2.28.tar.gz
tar -xzf RSEM-1.2.28.tar.gz

export PATH=$PATH:/<local machine path>/RSEM-1.2.28
export PATH=$PATH:/<local machine path>/bowtie2-2.2.7

/<local machine path>/RSEM-1.2.28/rsem-prepare-reference --
bowtie2
/<local machine path>/Trinity_andro_out.Trinity.02122016.fasta
andro_ref

# alternatively, shifting back to cluster
export PATH=$PATH:/home/avahoffman/RSEM-1.2.28
export PATH=$PATH:/home/avahoffman/bowtie2-2.2.7

#prepare reference (doesn't take too long)

~/RSEM-1.2.28/rsem-prepare-reference --bowtie2
~/Trinity_andro_out.Trinity.02122016.fasta andro_ref

# est abundance. RSEM currently does not support partial
alignments!

~/RSEM-1.2.28/rsem-calculate-expression -p 8 --paired-end --
bowtie2 --bowtie2-sensitivity-level very_sensitive --estimate-
```

```
rspd --append-names --calc-ci
~/andro/TrmPr1_A_DR7_s_5_1_sequence.txt
~/andro/TrmPr2_A_DR7_s_5_2_sequence.txt andro_ref exp/adr7

~/RSEM-1.2.28/rsem-calculate-expression -p 8 --paired-end --
bowtie2 --bowtie2-sensitivity-level very_sensitive --estimate-
rspd --append-names --calc-ci
~/andro/TrmPr1_A_DR8_s_5_1_sequence.txt
~/andro/TrmPr2_A_DR8_s_5_2_sequence.txt andro_ref exp/adr8

~/RSEM-1.2.28/rsem-calculate-expression -p 8 --paired-end --
bowtie2 --bowtie2-sensitivity-level very_sensitive --estimate-
rspd --append-names --calc-ci
~/andro/TrmPr1_A_WW7_s_5_1_sequence.txt
~/andro/TrmPr2_A_WW7_s_5_2_sequence.txt andro_ref exp/aww7

~/RSEM-1.2.28/rsem-calculate-expression -p 8 --paired-end --
bowtie2 --bowtie2-sensitivity-level very_sensitive --estimate-
rspd --append-names --calc-ci
~/andro/TrmPr1_A_WW8_s_5_1_sequence.txt
~/andro/TrmPr2_A_WW8_s_5_2_sequence.txt andro_ref exp/aww8
```

 # This step will export as a matrix, which is useful for downstream analysis. Runs in a few seconds. Need to install edgeR as an R package first, pretty straightforward directions on their website..

```
R
source("http://bioconductor.org/biocLite.R")
biocLite("edgeR")
q()

~/trinityrnaseq-2.1.1/util/abundance_estimates_to_matrix.pl --
est_method RSEM adr7.isoforms.results adr8.isoforms.results
aww7.isoforms.results aww8.isoforms.results

~/trinityrnaseq-2.1.1/util/abundance_estimates_to_matrix.pl --
est_method RSEM --name_sample_by_basedir
~/Strimdata/trinity_out_dir/SDR7_bowtie_out/RSEM.isoforms.results
~/Strimdata/trinity_out_dir/SDR8_bowtie_out/RSEM.isoforms.results
~/Strimdata/trinity_out_dir/SWW8_bowtie_out/RSEM.isoforms.results
~/Strimdata/trinity_out_dir/SWW9_bowtie_out/RSEM.isoforms.results
```

 # now you can calculate N90 statistics! Tells you that 90% of transcripts are found in a contig of "x" length. Run same script below for both species, but make sure you are in the species specific trinity_out_dir directory. Trinity web page says it's good to graph this.

```
export PATH=$PATH:/home/avahoffman/trinityrnaseq-2.1.1/util/misc/
```

```
~/trinityrnaseq-2.1.1/util/misc/contig_ExN50_statistic.pl
matrix.TMM.EXPR.matrix ~/Trinity_andro_out.Trinity.02122016.fasta
| tee ExN50.stats

# output TPM values.

~/trinityrnaseq-
2.1.1/util/misc/count_matrix_features_given_MIN_TPM_threshold.pl
matrix.TPM.not_cross_norm | tee
matrix.TPM.not_cross_norm.counts_by_min_TPM
```

 # Quality check samples and replicates

 # want to make sure your replicates are correlated. Make sure you have uploaded a .txt file that describes your samples to the datastore (this file must have a row for each "condition", e.g., well watered samples on one row and drought samples on the next). May need to install some R packages. I ran this next bit of code in the trinity_out_dir area. Need xvfb so we can get this to run on a VM. Make sure the "samples" file has a separate sample on each line, eg:

```
# WW AWW7_bowtie_out
# WW AWW8_bowtie_out
# DR ADR7_bowtie_out
# DR ADR8_bowtie_out

apt-get install Xvfb

R
source("http://bioconductor.org/biocLite.R")
biocLite("qvalue")
biocLite('Biobase')
quit(save = "default", status = 0, runLast = TRUE)

xvfb-run ~/trinityrnaseq-
2.1.1/Analysis/DifferentialExpression/PtR --matrix
~/RSEM_output/exp/matrix.counts.matrix --samples
~/andro/samples_described.txt --CPM --log2 --compare_replicates
xvfb-run ~/trinityrnaseq-
2.1.1/Analysis/DifferentialExpression/PtR --matrix
matrix.counts.matrix --samples ~/Strimdata/samples_sorgh.txt --
CPM --log2 --compare_replicates

xvfb-run ~/trinityrnaseq-
2.1.1/Analysis/DifferentialExpression/PtR --matrix
~/RSEM_output/exp/matrix.counts.matrix -s
~/andro/samples_described.txt --log2 --sample_cor_matrix
xvfb-run ~/trinityrnaseq-
2.1.1/Analysis/DifferentialExpression/PtR --matrix
matrix.counts.matrix -s ~/Strimdata/samples_sorgh.txt --log2 --
sample_cor_matrix
```

```
xvfb-run ~/trinityrnaseq-
2.1.1/Analysis/DifferentialExpression/PtR --matrix
~/RSEM_output/exp/matrix.counts.matrix -s
~/andro/samples_described.txt --log2 --prin_comp 3

#can also run on local machine

/Users/avahoffman/trinityrnaseq/Analysis/DifferentialExpression/P
tR --matrix matrix.counts.matrix -s samples_described.txt --log2
--prin_comp 3
```

 # Differential expression (DE) analysis

 # Run Trinotate beforehand so that DE genes can be identified. Already have the blast tools ☺
 # Run TransDecoder. This identifies long ORFs and likely coding regions.

```
wget
https://github.com/TransDecoder/TransDecoder/archive/2.0.1.tar.gz
tar xvfz 2.0.1.tar.gz
make
export PATH=$PATH:/home/avahoffman/TransDecoder-2.0.1
~/TransDecoder-2.0.1/TransDecoder.LongOrfs -t
~/Atrimdata/trinity_out_dir/Trinity.fasta
~/TransDecoder-2.0.1/TransDecoder.LongOrfs -t
~/Strimdata/trinity_out_dir/Trinity.fasta
~/TransDecoder-2.0.1/TransDecoder.Predict -t
~/Atrimdata/trinity_out_dir/Trinity.fasta
~/TransDecoder-2.0.1/TransDecoder.Predict -t
~/Strimdata/trinity_out_dir/Trinity.fasta

wget https://github.com/Trinotate/Trinotate/archive/v2.0.2.tar.gz
tar xvfz v2.0.2.tar.gz
make?
wget http://www.sqlite.org/2015/sqlite-shell-linux-x86-
3090200.zip
unzip sqlite-shell-linux-x86-3090200.zip
make?
wget http://selab.janelia.org/software/hmmer3/3.1b2/hmmer-3.1b2-
linux-intel-x86_64.tar.gz
tar xvfz hmmer-3.1b2-linux-intel-x86_64.tar.gz
make
wget
https://data.broadinstitute.org/Trinity/Trinotate_v2.0_RESOURCES/
uniprot_uniref90.trinotate_v2.0.pep.gz
```

```
mv uniprot_uniref90.trinotate_v2.0.pep.gz
uniprot_uniref90.trinotate.pep.gz
gunzip uniprot_uniref90.trinotate.pep.gz
makeblastdb -in uniprot_uniref90.trinotate.pep -dbtype prot
wget
https://data.broadinstitute.org/Trinity/Trinotate_v2.0_RESOURCES/
Pfam-A.hmm.gz
gunzip Pfam-A.hmm.gz
hmmpress Pfam-A.hmm



# First, will need to install Bioconductor stuff in R. You should
already have edgeR installed.

R
source("http://bioconductor.org/biocLite.R")
biocLite('edgeR')
biocLite('limma')
biocLite('DESeq2')
biocLite('ctc')
biocLite('Biobase')
install.packages('gplots')
install.packages('ape')

# Install ROTS package

wget
http://www.btk.fi/fileadmin/Page_files/Research/compbiomed/ROTS_1
.1.2.tar.gz
R CMD INSTALL ROTS_1.1.2.tar.gz

#had to force update xml ..

sudo apt-get install r-cran-xml


#try switching out different methods (edgeR, voom, DESeq2, ROTS)

xvfb-run ~/trinityrnaseq-
2.1.1/Analysis/DifferentialExpression/run_DE_analysis.pl --matrix
~/andro/matrix.counts.matrix --method DESeq2 --samples_file
~/andro/samples_described.txt --contrasts contrasts.txt

xvfb-run ~/trinityrnaseq-
2.1.1/Analysis/DifferentialExpression/run_DE_analysis.pl --matrix
~/andro/matrix.counts.matrix --method ROTS --samples_file
~/andro/samples_described.txt --contrasts contrasts.txt

#if there are problems with original, use the new version, as per
trinity users site.
```

```
wget http://groups.google.com/group/trinityrnaseq-
users/attach/1b3a837ad8297d0b/analyze_diff_expr.pl

#set directory
~/trinityrnaseq-2.1.1/Analysis/DifferentialExpression/

sudo analyze_diff_expr.pl --matrix ~/andro/matrix.TMM.EXPR.matrix
--samples ~/andro/samples_described.txt
```