Differential Expression Analysis

Transcriptome Analysis Overview

1) Download reads from sequencing facility or SRA
2) Check the quality of raw sequence files (FastQC)
3) Remove adapter sequence (e.g. CutAdapt)
4) Align RNAseq reads to the genome (e.g. STAR, HISAT2)
5) Get transcript abundances (e.g. HTSeq2)
6) QC individual count files
7) Make an expression matrix combining multiple samples into 1 file
8) QC matrix
9) Conduct biological analysis (e.g. determine differentially expressed genes with DESeq2)

Maize Gene Atlas

the plant journal

SEB Society for Experimental Biology

The Plant Journal (2011) doi: 10.1111/j.1365-313X.2011.04527.x

Genome-wide atlas of transcription during maize development

Rajandeep S. Sekhon[1,2,†], Haining Lin[3,4,†], Kevin L. Childs[3,4], Candice N. Hansey[3,4], C. Robin Buell[3,4], Natalia de Leon[1,2] and Shawn M. Kaeppler[1,2,*]

[1]*Department of Energy Great Lakes Bioenergy Research Center, University of Wisconsin-Madison, 1575 Linden Drive, Madison, WI 53706, USA,*

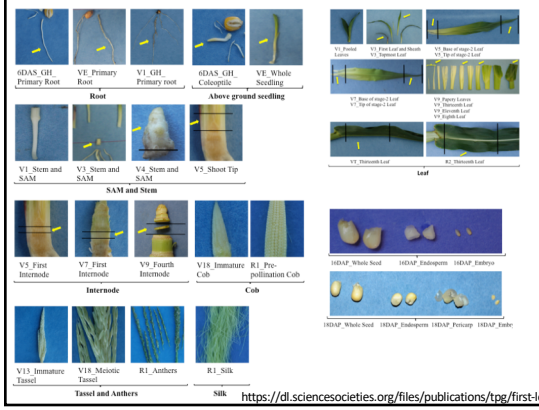
[2]*Department of Agronomy, University of Wisconsin-Madison, 1575 Linden Drive, Madison, WI 53706, USA,*

[3]*Department of Energy Great Lakes Bioenergy Research Center, Michigan State University, East Lansing, MI 48824, USA, and*

[4]*Department of Plant Biology, Michigan State University, East Lansing, MI 48824, USA*

Maize Gene Atlas

Catalog of gene expression throughout tissues throughout development



Tissues
2 Cob
8 Internode
18 Leaf
18 Root
5 SAM/Stem
25 Seed
1 Silk
3 Tassel/Anther

(Sekhon et al., PLoS ONE 2013)
https://dl.sciencesocieties.org/files/publications/tpg/first-look/plantgenome-tpg-2015-04-0025.pdf

1) Download reads from SRA to MSI

```
#!/bin/bash

#SBATCH --time=3:00:00
#SBATCH --ntasks=8
#SBATCH --mem=16g
#SBATCH --tmp=16g
#SBATCH --mail-type=ALL
#SBATCH --mail-user=cnhirsch@umn.edu


echo "Step 0: Setting up directory structure..."
mkdir /home/agro5431/cnhirsch/transcriptome_analysis/
cd /home/agro5431/cnhirsch/transcriptome_analysis/

echo "Step 1: Downloading the file from the SRA..."
module load sratoolkit/2.8.2
fastq-dump --split-files -F -Q 33 SRR940252

echo "Step 2: Check the quality of the raw reads..."
module load fastqc/0.11.7
fastqc -f fastq  SRR940252_1.fastq

echo "Step 3: Remove the adapter sequences..."
module load cutadapt/1.18
cutadapt -a GATCGGAAGAGCACACGTCTGAACTCCAGTCACCAGATCATCTCGTATGCCGTCTTCTGCTTG -a AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT
-f fastq -m 30 --quality-base=33 -o SRR940252_1_cutadapt.fastq SRR940252_1.fastq > SRR940252_cutadapt.log

echo "Step 4: Check the quality of the cleaned reads..."
fastqc -f fastq  SRR940252_1_cutadapt.fastq

echo "Step 5: Align RNAseq reads to the genome..."
module load star/2.5.3a
STAR --genomeDir /home/agro5431/shared/B73_STAR \
     --runThreadN 8 \
     --readFilesIn SRR940252_1_cutadapt.fastq \
     --outFileNamePrefix SRR940252_1_cutadapt_STAR \
     --outSAMtype BAM SortedByCoordinate \
     --outSAMunmapped Within \ --outSAMattributes Standard

echo "Step 6: Get transcript abundance..."
module load htseq/0.7.2
htseq-count -f bam -s no -t gene -i ID -m union -a 20 SRR940252_1_cutadapt_STARAligned.sortedByCoord.out.bam /home/agro5431/shared/Zea_may
s.AGPv4.33.gff3 > htseq_SRR940252.txt
```

% fastq-dump --split-files -F -Q 33 XXXXXXXXX.sra

2) Check the quality of raw sequence files

```
#!/bin/bash

#SBATCH --time=3:00:00
#SBATCH --ntasks=8
#SBATCH --mem=16g
#SBATCH --tmp=16g
#SBATCH --mail-type=ALL
#SBATCH --mail-user=cnhirsch@umn.edu


echo "Step 0: Setting up directory structure..."
mkdir /home/agro5431/cnhirsch/transcriptome_analysis/
cd /home/agro5431/cnhirsch/transcriptome_analysis/

echo "Step 1: Downloading the file from the SRA..."
module load sratoolkit/2.8.2
fastq-dump --split-files -F -Q 33 SRR940252

echo "Step 2: Check the quality of the raw reads..."
module load fastqc/0.11.7
fastqc -f fastq  SRR940252_1.fastq

echo "Step 3: Remove the adapter sequences..."
module load cutadapt/1.18
cutadapt -a GATCGGAAGAGCACACGTCTGAACTCCAGTCACCAGATCATCTCGTATGCCGTCTTCTGCTTG -a AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT
-f fastq -m 30 --quality-base=33 -o SRR940252_1_cutadapt.fastq SRR940252_1.fastq > SRR940252_cutadapt.log

echo "Step 4: Check the quality of the cleaned reads..."
fastqc -f fastq  SRR940252_1_cutadapt.fastq

echo "Step 5: Align RNAseq reads to the genome..."
module load star/2.5.3a
STAR --genomeDir /home/agro5431/shared/B73_STAR \
     --runThreadN 8 \
     --readFilesIn SRR940252_1_cutadapt.fastq \
     --outFileNamePrefix SRR940252_1_cutadapt_STAR \
     --outSAMtype BAM SortedByCoordinate \
     --outSAMunmapped Within \ --outSAMattributes Standard

echo "Step 6: Get transcript abundance..."
module load htseq/0.7.2
htseq-count -f bam -s no -t gene -i ID -m union -a 20 SRR940252_1_cutadapt_STARAligned.sortedByCoord.out.bam /home/agro5431/shared/Zea_may
s.AGPv4.33.gff3 > htseq_SRR940252.txt
```

% fastqc -f fastq XXXXXXXXX_1.fastq

3) Remove adapter sequence

```
#!/bin/bash

#SBATCH --time=3:00:00
#SBATCH --ntasks=8
#SBATCH --mem=16g
#SBATCH --tmp=16g
#SBATCH --mail-type=ALL
#SBATCH --mail-user=cnhirsch@umn.edu


echo "Step 0: Setting up directory structure..."
mkdir /home/agro5431/cnhirsch/transcriptome_analysis/
cd /home/agro5431/cnhirsch/transcriptome_analysis/

echo "Step 1: Downloading the file from the SRA..."
module load sratoolkit/2.8.2
fastq-dump --split-files -F -Q 33 SRR940252

echo "Step 2: Check the quality of the raw reads..."
module load fastqc/0.11.7
fastqc -f fastq  SRR940252_1.fastq

echo "Step 3: Remove the adapter sequences..."
module load cutadapt/1.18
cutadapt -a GATCGGAAGAGCACACGTCTGAACTCCAGTCACCAGATCATCTCGTATGCCGTCTTCTGCTTG -a AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT
-f fastq -m 30 --quality-base=33 -o SRR940252_1_cutadapt.fastq SRR940252_1.fastq > SRR940252_cutadapt.log

echo "Step 4: Check the quality of the cleaned reads..."
fastqc -f fastq  SRR940252_1_cutadapt.fastq

echo "Step 5: Align RNAseq reads to the genome..."
module load star/2.5.3a
STAR --genomeDir /home/agro5431/shared/B73_STAR \
     --runThreadN 8 \
     --readFilesIn SRR940252_1_cutadapt.fastq \
     --outFileNamePrefix SRR940252_1_cutadapt_STAR \
     --outSAMtype BAM SortedByCoordinate \
     --outSAMunmapped Within \ --outSAMattributes Standard

echo "Step 6: Get transcript abundance..."
module load htseq/0.7.2
htseq-count -f bam -s no -t gene -i ID -m union -a 20 SRR940252_1_cutadapt_STARAligned.sortedByCoord.out.bam /home/agro5431/shared/Zea_may
s.AGPv4.33.gff3 > htseq_SRR940252.txt
```

% cutadapt -a <TruSeq_Adapter> -a <Universal_Adapter> -f fastq -m 30 --quality-base=33 -o XXXXXXXXX_1_cutadapt.fastq XXXXXXXXX_1.fastq > XXXXXXXXX_cutadapt.log

4) Align RNAseq reads to the genome

HISAT2
graph-based alignment of next generation sequencing reads to a population of genomes

Johns Hopkins University Center for Computational Biology CCB

HISAT2 is a fast and sensitive alignment program for mapping next-generation sequencing reads (both DNA and RNA) to a population of human genomes (as well as to a single reference genome). Based on an extension of BWT for graphs [Sirén et al. 2014], we designed and implemented a graph FM index (GFM), an original approach and its first implementation to the best of our knowledge. In addition to using one global GFM index that represents a population of human genomes, HISAT2 uses a large set of small GFM indexes that collectively cover the whole genome (each index representing a genomic region of 56 Kbp, with 55,000 indexes needed to cover the human population). These small indexes (called local indexes), combined with several alignment strategies, enable rapid and accurate alignment of sequencing reads. This new indexing scheme is called a Hierarchical Graph FM index (HGFM).

OSI certified

BIOINFORMATICS **ORIGINAL PAPER**

Vol. 29 no. 1 2013, pages 15–21
doi:10.1093/bioinformatics/bts635

Sequence analysis

Advance Access publication October 25, 2012

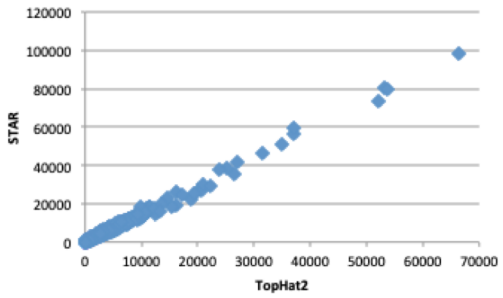
STAR: ultrafast universal RNA-seq aligner

Alexander Dobin[1,*], Carrie A. Davis[1], Felix Schlesinger[1], Jorg Drenkow[1], Chris Zaleski[1], Sonali Jha[1], Philippe Batut[1], Mark Chaisson[2] and Thomas R. Gingeras[1]

[1]Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, USA and [2]Pacific Biosciences, Menlo Park, CA, USA

Associate Editor: Inanc Birol

Comparison of Counts from Different Aligners



4) Align RNAseq reads to the genome

```
module load star

STAR  --runMode genomeGenerate \
      --genomeDir /home/hirschc1/shared/projects/te-expr/data/databases/star \
      --genomeFastaFiles /home/maize/shared/databases/genomes/Zea_mays/PH207/ZmaysPH207_443_v1.0.fa \
      --sjdbGTFfile /home/maize/shared/databases/genomes/Zea_mays/PH207/ZmaysPH207_443_v1.1.gene_exons.gtf \
      --sjdbOverhang 49 \
      --runThreadN 1 \
      --limitGenomeGenerateRAM 35773245482
```

4) Align RNAseq reads to the genome

```
#!/bin/bash

#SBATCH --time=3:00:00
#SBATCH --ntasks=8
#SBATCH --mem=16g
#SBATCH --tmp=16g
#SBATCH --mail-type=ALL
#SBATCH --mail-user=cnhirsch@umn.edu

echo "Step 0: Setting up directory s
mkdir /home/agro5431/cnhirsch/transc
cd /home/agro5431/cnhirsch/transcrip

echo "Step 1: Downloading the file (
module load sratoolkit/2.8.2
fastq-dump --split-files -F -Q 33 SRR940252

echo "Step 2: Check the quality of the raw reads..."
module load fastqc/0.11.7
fastqc -f fastq  SRR940252_1.fastq

echo "Step 3: Remove the adapter sequences..."
module load cutadapt/1.18
cutadapt -a GATCGGAAGAGCACACGTCTGAACTCCAGTCACCAGATCATCTCGTATGCCGTCTTCTGCTTG -a AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT
-f fastq -m 30 --quality-base=33 -o SRR940252_1_cutadapt.fastq SRR940252_1.fastq > SRR940252_cutadapt.log

echo "Step 4: Check the quality of the cleaned reads..."
fastqc -f fastq  SRR940252_1_cutadapt.fastq

echo "Step 5: Align RNAseq reads to the genome..."
module load star/2.5.3a
STAR --genomeDir /home/agro5431/shared/B73_STAR \
     --runThreadN 8 \
     --readFilesIn SRR940252_1_cutadapt.fastq \
     --outFileNamePrefix SRR940252_1_cutadapt_STAR \
     --outSAMtype BAM SortedByCoordinate \
     --outSAMunmapped Within \ --outSAMattributes Standard

echo "Step 6: Get transcript abundance..."
module load htseq/0.7.2
htseq-count -f bam -s no -t gene -i ID -m union -a 20 SRR940252_1_cutadapt_STARAligned.sortedByCoord.out.bam /home/agro5431/shared/Zea_mays
s.AGPv4.33.gff3 > htseq_SRR940252.txt
```

STAR --genomeDir /home/agro5431/shared/B73_STAR \
--runThreadN 8 \
--readFilesIn ../cutadapt_cleaned_reads/XXXXXXXXX_1_cutadapt.fastq \
--outFileNamePrefix XXXXXXXXX_1_cutadapt_STAR \
--outSAMtype BAM SortedByCoordinate \
--outSAMunmapped Within \ --outSAMattributes Standard

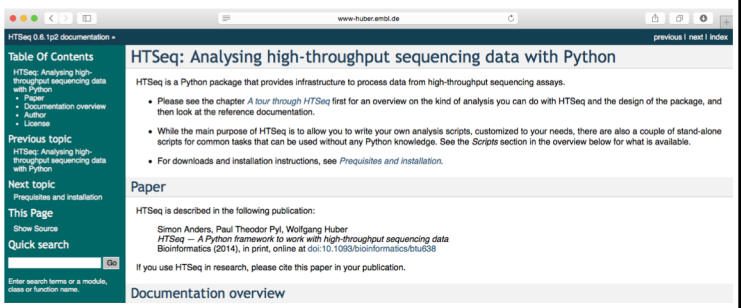
4) Align RNAseq reads to the genome

% more XXXXXXX_1_cutadapt_STARLog.final.out

```
                                 Started job on |       Apr 01 10:55:10
                             Started mapping on |       Apr 01 10:56:24
                                    Finished on |       Apr 01 10:58:56
       Mapping speed, Million of reads per hour |       292.59

                          Number of input reads |       12353839
                      Average input read length |       99
                                    UNIQUE READS:
                   Uniquely mapped reads number |       10636220
                        Uniquely mapped reads % |       86.10%
                          Average mapped length |       99.10
                       Number of splices: Total |       2571479
            Number of splices: Annotated (sjdb) |       2463999
                       Number of splices: GT/AG |       2530253
                       Number of splices: GC/AG |       34963
                       Number of splices: AT/AC |       1520
           Number of splices: Non-canonical |       4743
                  Mismatch rate per base, % |       1.55%
                     Deletion rate per base |       0.00%
                    Deletion average length |       1.86
                    Insertion rate per base |       0.00%
                   Insertion average length |       1.38
                                 MULTI-MAPPING READS:
      Number of reads mapped to multiple loci |       545476
           % of reads mapped to multiple loci |       4.42%
      Number of reads mapped to too many loci |       279225
           % of reads mapped to too many loci |       2.26%
                                 UNMAPPED READS:
% of reads unmapped: too many mismatches |       0.00%
     % of reads unmapped: too short |       7.02%
         % of reads unmapped: other |       0.21%
                                 CHIMERIC READS:
              Number of chimeric reads |       0
                   % of chimeric reads |       0.00%
```

5) Get transcript abundances



5) Get transcript abundances

```
#!/bin/bash

#SBATCH --time=3:00:00
#SBATCH --ntasks=8
#SBATCH --mem=16g
#SBATCH --tmp=16g
#SBATCH --mail-type=AL
#SBATCH --mail-user=cn

echo "Step 0: Setting
mkdir /home/agro5431/c
cd /home/agro5431/cnhirsch/transcriptome_analysis/

echo "Step 1: Downloading the file from the SRA..."
module load sratoolkit/2.8.2
fastq-dump --split-files -F -Q 33 SRR940252

echo "Step 2: Check the quality of the raw reads..."
module load fastqc/0.11.7
fastqc -f fastq  SRR940252_1.fastq

echo "Step 3: Remove the adapter sequences..."
module load cutadapt/1.18
cutadapt -a GATCGGAAGAGCACACGTCTGAACTCCAGTCACCAGATCATCTCGTATGCCGTCTTCTGCTTG -a AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT
-f fastq -m 30 --quality-base=33 -o SRR940252_1_cutadapt.fastq SRR940252_1.fastq > SRR940252_cutadapt.log

echo "Step 4: Check the quality of the cleaned reads..."
fastqc -f fastq  SRR940252_1_cutadapt.fastq

echo "Step 5: Align RNAseq reads to the genome..."
module load star/2.5.3a
STAR --genomeDir /home/agro5431/shared/B73_STAR \
     --runThreadN 8 \
     --readFilesIn SRR940252_1_cutadapt.fastq \
     --outFileNamePrefix SRR940252_1_cutadapt_STAR \
     --outSAMtype BAM SortedByCoordinate \
     --outSAMunmapped Within \ --outSAMattributes Standard

echo "Step 6: Get transcript abundance..."
module load htseq/0.7.2
htseq-count -f bam -s no -t gene -i ID -m union -a 20 SRR940252_1_cutadapt_STARAligned.sortedByCoord.out.bam /home/agro5431/shared/Zea_may
s.AGPv4.33.gff3 > htseq_SRR940252.txt
```

% htseq-count -f bam -s no -t gene -i ID -m union -a 20 XXXXXXXXX_1_cutadapt_STARAligned.sortedByCoord.out.bam /home/agro5431/shared/Zea_mays.AGPv4.33.gff3 > htseq_XXXXXXXXX.txt

6) QC Individual Count Files

- Check that genes have counts
- Check the number of genes printed out is correct (n=39,498)
- use unix 'tail' to look at metrics at bottom of file

```
__no_feature    590886
__ambiguous     132898
__too_low_aQual 0
__not_aligned   745432
__alignment_not_unique  2166383
```

- Look at the annotation of the highest expressed genes
  % awk '$1 !~ /__/' htseq_XXXXXXXX.txt | sort -k 2 -n -r | head -n 3
  % grep <gene_name> /home/agro5431/shared/B73v4.gene_function.txt
- Does this make sense?

What we are doing today

1) Evaluating QC of raw data
2) Evaluating QC of read cleaning
3) Evaluating QC of read mapping
4) Evaluating QC of our matrix
5) Determining differentially expressed genes between two tissues

What we are doing today

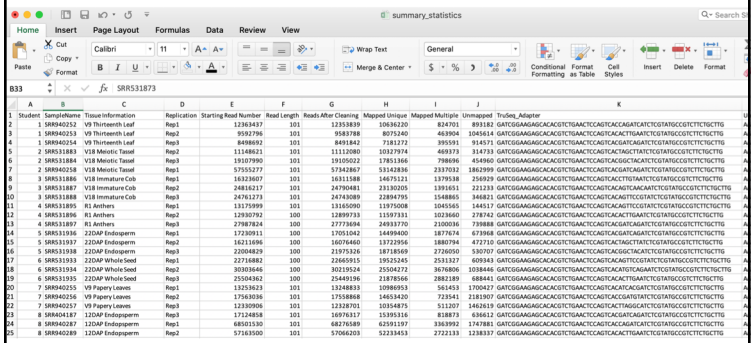
1) Evaluating QC of raw data

Aggregating Results for QC

What we are doing today

1) Evaluating QC of raw data
2) Evaluating QC of read cleaning
3) Evaluating QC of read mapping

Summarized Statistics from Steps 1-5

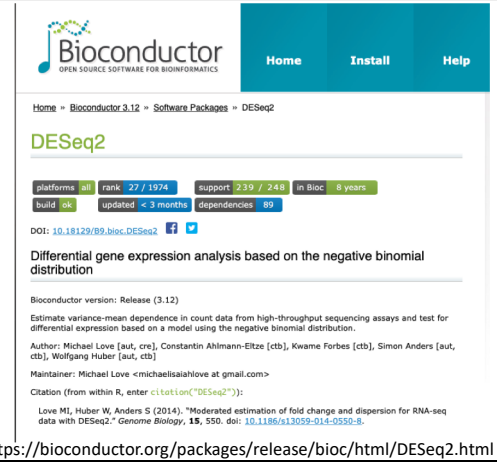


Statistics in summary_statistics.xlsx file on GitHub

What we are doing today

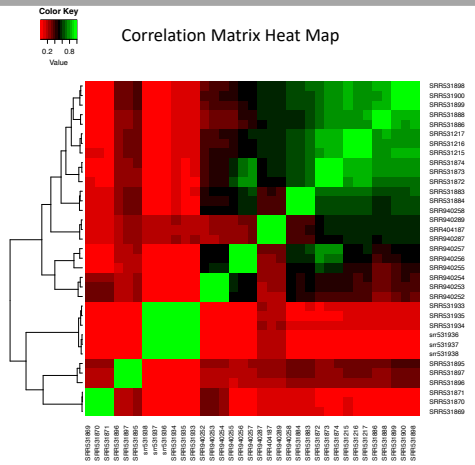
1) Evaluating QC of raw data
2) Evaluating QC of read cleaning
3) Evaluating QC of read mapping
4) Evaluating QC of our matrix

8) QC Normalized Matrix



https://bioconductor.org/packages/release/bioc/html/DESeq2.html

8) QC Normalized Matrix



What we are doing today

1) Evaluating QC of raw data
2) Evaluating QC of read cleaning
3) Evaluating QC of read mapping
4) Evaluating QC of our matrix
5) Determining differentially expressed genes between two tissues

9) Determine DE Genes

DESeq2.R on GitHub

	baseMean	log2FoldChang	lfcSE	stat	pvalue	padj
GRMZM5G859187	0	NA	NA	NA	NA	NA
Zm00001d005057	0.62632282	-0.057532155	3.845326996	-0.014961577	0.98806284	NA
GRMZM5G872747	0	NA	NA	NA	NA	NA
Zm00001d024009	3640.805735	0.48561575	0.113026272	4.296485611	1.74E-05	0.00011502
Zm00001d049223	877.2138353	0.699905569	0.106420918	6.576766891	4.81E-11	8.57E-10
Zm00001d025661	13.95323308	-0.410223971	0.481732533	-0.851559616	0.394458562	0.565817239
Zm00001d019605	867.4093887	-0.660120136	0.835199737	-0.790373975	0.429309397	0.602943056
Zm00001d002425	18.34313278	-0.178667099	0.464514568	-0.384631852	0.700510192	0.842170001
Zm00001d033005	114.6465514	-2.154534698	1.306818596	-1.648686898	0.099211796	0.198444747
Zm00001d047463	717.7405169	-0.185373023	0.404502517	-0.458274091	0.646755536	0.797913196
Zm00001d021685	1.962652374	-2.881742848	1.624449927	-1.773980718	0.07606634	NA
Zm00001d051863	1.47159629	0	2.54340852	0	1	NA
Zm00001d046861	0.062965346	0	9.04767805	0	1	NA
Zm00001d037010	684.7157625	-1.248200245	0.161193094	-7.74350946	9.67E-15	2.58E-13
Zm00001d036900	4.9922491	-0.05753222	3.538881941	-0.016257174	0.987029223	1
GRMZM5G828883	0.101158303	0	9.04767805	0	1	NA
Zm00001d039012	105.2065898	0.19877112	0.14791903	1.343783281	0.179018478	0.314184329
Zm00001d010827	2.355157857	-1.539465249	1.114040501	-1.381875478	0.167009939	NA
Zm00001d021346	141.9781733	-0.967311889	0.190682259	-5.07289925	3.92E-07	3.64E-06
Zm00001d007364	0.00703322	0	9.04767805	0	1	NA
Zm00001d031958	333.7757076	-1.446922213	0.198506185	-7.28905355	3.12E-13	7.26E-12
Zm00001d013540	450.6146234	0.146613916	0.100803339	1.454454957	0.145820213	0.268048064
Zm00001d010950	798.1235901	0.22187154	0.086007788	2.579668013	0.009889534	0.029931584
Zm00001d048087	1.295122992	0	3.255323401	0	1	NA
Zm00001d014170	0.168154698	0	9.04767805	0	1	NA
Zm00001d014690	553.9177661	-0.449766813	0.168724872	-2.665681762	0.007683239	0.024211298
Zm00001d012811	316.6732233	0.081869405	0.114448767	0.715336713	0.474401008	0.648614656
Zm00001d040304	14.30249056	0	1.718964003	0	1	1
Zm00001d049195	0.009648685	0	9.04767805	0	1	NA
Zm00001d028813	378.8982744	0.132692969	0.107224654	1.237522943	0.215893002	0.362522545
Zm00001d018159	38.44747157	0.27293919	0.34187274	0.798364882	0.424658776	0.598085857

Discussion Questions

- What if any samples are problematic?
- What evidence is there for this?
- How many differentially expressed genes did you find?
- Is this more/less than you expected?