Supplementary Documentation

In Unix Terminal:

Getting fastq files using SRR numbers:

- fastq-dump SRR2517972 -N 1 -X 1000000 --split-files -readids --defline-qual + --gzip
- fastq-dump SRR2517975 -N 1 -X 1000000 --split-files -readids --defline-qual + --gzip

Building the hisat2 index:

- hisat2_build xenlae9.fa xl

Aligning the fastq files for each stage to the genome:

- hisat2 --no-unal --phred33 -x xl -U SRR2517972_1.fastq -S egg_aligned.sam --summary-file egg_alignment_summary.txt
- hisat2 --no-unal --phred33 -x xl -U SRR2517975_1.fastq -S st10_aligned.sam --summary-file st10_alignment_summary.txt

Converting .sam to .bam:

- samtools view -hSb egg_aligned.sam -o egg_aligned.bam
- samtools view -hSb st10_aligned.sam -o st10_aligned.bam

Sorting .bam files:

- samtools sort egg_aligned.bam -o egg_aligned_final.bam
- samtools sort st10_aligned.bam -o st10_aligned_final.bam

Indexing .bam files

- samtools index egg_aligned_final.bam
- samtools index st10_aligned_final.bam

Running featureCounts:

- featureCounts -s 0 -a xbGene.9.1.gtf -g gene_id -o st10_aligned_final_featureCounts.txt st10_aligned_final.bam

- featureCounts -s 0 -a xbGene.9.1.gtf -g gene_id -o egg_aligned_final_featureCounts.txt egg_aligned_final.bam

Converting to Scaffold-less files:

- grep -v "Scaffold" egg_aligned_final_featureCounts.txt > egg_R.txt
- grep -v “Scaffold” st10_aligned_final_featureCounts.txt >st10_R.txt

Transferring files:

- Use WinSCP (Windows) or Fetch (Mac) to take files off of the cluster or remote server and bring it onto one’s hard drive.

Using R or RStudio:

Reading in featureCounts output files into R:

- >egg = read.table(“Desktop/egg_R.txt”, sep = “\t”, header = T)
- > st10 = read.table("Desktop/st10_R.txt", sep = "\t", header = T)

Making Histogram:

- >egg_reads = egg$egg_aligned_final.bam
- >st10_reads = st10$st10_aligned_final.bam

- >st10_rpm = (1000000*st10_reads)/sum(st10_reads)
- >egg_rpm = (1000000 * egg_reads)/sum(st10_reads)

- >lg_egg_rpm = log2(egg_rpm +0.5)
- >lg_st10_rpm = log2(st10_rpm +0.5)

- >max(lg_egg_rpm)
- >max(lg_st10_rpm)

- >hist(lg_egg_rpm, 10, xlim = c(0,12))

- >hist(lg_st10_rpm, 10, xlim = c(0,12))

Creating subsets of only Geneids and the corresponding read frequencies:

- > egg_sub = data.frame(egg$Geneid,egg$egg_aligned_final.bam)
- > st10_sub = data.frame(st10$Geneid,st10$st10_aligned_final.bam)

Renaming columns:

- > colnames(st10_sub)[1] = "Geneid"
- > colnames(st10_sub)[2] = "read count"
- > colnames(egg_sub)[2] = "read count"
- > colnames(egg_sub)[1] = "Geneid"

Example of the subsets at this point:

- > head(egg_sub)

	Geneid	read count
1	Xelaev18004772m.g	1
2	Xelaev18004775m.g	0
3	Xelaev18004811m.g	0
4	Xelaev18004812m.g	0
5	Xelaev18004815m.g	115
6	Xelaev18004816m.g	4

Creating the combined dataframe:

- > egg_vs_st10 = egg_sub
- > colnames(egg_vs_st10)[2] = "egg_reads"
- > colnames(st10_sub)[2] = "st10_reads"
- > egg_vs_st10$st10_reads = st10_sub$st10_reads

Normalizing and Smoothing the data for comparisons:

- > rpm_egg = 1000000 * egg_vs_st10$egg_reads / sum(egg_vs_st10$egg_reads)

- > lg_rpm_egg = log2(0.5 + rpm_egg)
- > rpm_st10 = 1000000 * egg_vs_st10$st10_reads / sum(egg_vs_st10$st10_reads)
- > lg_rpm_st10 = log2(0.5 + rpm_st10)
- > egg_vs_st10$lg_rpm_egg = lg_rpm_egg
- > egg_vs_st10$lg_rpm_st10 = lg_rpm_st10

Visualization:

- > plot(egg_vs_st10lg_rpm_egg, egg_vs_st10lg_rpm_st10, main = 'Egg vs St10', xlab ='lg_rpm_egg', ylab = 'lg_rpm_st10')
- > diff_exp = subset(egg_vs_st10, (lg_rpm_egg - lg_rpm_st10 > 2))
- > points(diff_exp[,c('lg_rpm_egg', 'lg_rpm_st10')], col='red')
- > abline(0,1, col = 'green')

Final dataframe - containing all the genes that experience maternal clearance:

- > head(diff_exp)

	Geneid	egg_reads	st10_reads	lg_rpm_egg	lg_rpm_st10
1	Xelaev18004772m.g	1	0	1.027626	-1.000000
5	Xelaev18004815m.g	115	12	7.471242	4.336891
14	Xelaev18004832m.g	1	0	1.027626	-1.000000
23	Xelaev18004847m.g	25	4	5.284169	2.821609
50	Xelaev18004886m.g	1	0	1.027626	-1.000000
56	Xelaev18004896m.g	164	4	7.982089	2.821609

In Unix Terminal:

Searching for GCACTT in 3'UTR regions of egg data results:

- grep -i -B 1 'GCACTT' xl9_3utr.fa > check.txt
- Search the 3'utr fasta file for entries containing the target sequence and print line denoting position
- grep '>' check.txt > check1.txt
- wc check1.txt

- Count how many entries contain the target sequence
- sed ‘s/^.//’ check1. > checkcut.txt
- remove ‘>’ for further use
- grep -f checkcut.txt xbGene.9.1.gtf > grep.txt
- find matching gtf entries, including scaffolds
- So, entries in grep.txt are ones that have a 3’UTR with a target sequence.
- cut -c1-18 maternal_cleared.txt > maternalname.txt
- This gives only the gene names.
- grep -f maternalname.txt grep.txt > match.txt
- These will be R determined maternally cleared genes that also have the target 3’ UTR sequence.

Fisher tests

- cut -f9 match.txt > test.txt to remove only identifiers from all entries that have both target sequence and decrease with time
- cut -c33-59 test.xt > test2.txt to remove only geneid
- uniq -c test2.txt | wc to count unique geneid identifiers; this gives the number of genes that have both the target sequence and decrease with time: 550
- grep “>” x19_3utr.fa | wc to count all genes expressed: 29508
- grep “GCACTT” x19_3utr.fa | wc gives 8754 genes with sequence
- grep -v “GCACTT” xl9_3utr.fa > test.txt
- grep -v “>” test.txt | wc to give 20754 genes that do not have the target sequence
- wc maternal_cleared_fix.txt gives 1584 total decreased genes

Easy Fisher Exact Test Calculator

Success! The Fisher exact test statistic and statement of significance appear beneath the table. Blue means you're dealing with dependent variables; red, independent.

Results			
	has seq	not seq	*Marginal Row Totals*
is decrease	550	1033	1583
no decrease	8204	19721	27925
Marginal Column Totals	8754	20754	29508 (Grand Total)

The Fisher exact test statistic value is 8E-06. The result is significant at p < .05.

L vs. S

>grep “L” match.txt > matchL.txt

>cut -f9 matchL.txt

>cut -c33-70 matchL.txt > matchL2.txt

>uniq -c matchL2.txt | wc

316 genes on L chromosomes that are decreased and have sequence.

Repeat for “S”

234 genes on S chromosomes that are decreased and have sequence.

Results			
	L	S	*Marginal Row Totals*
dec/seq	316	234	550
no dec/seq	14438	14520	28958
Marginal Column Totals	14754	14754	29508 (Grand Total)

The Fisher exact test statistic value is 0.000479. The result is significant at p < .05.