How to properly use the RNA-Seq data of *Manduca* transcriptomes

1. Acknowledgements

We would like to thank all the people who have contributed to the Manduca RNA-Seq datasets. Among them, Dr. Gary Blissard is the most prominent one and, his postdoc Yunru Chen performed all the dissections and prepared most of the cDNA libraries for sequence determination. Costs for the Illumina sequencing, mostly done at Baylor and Cornell, were covered by a DARPA grant to Dr. Blissard. The assemblies of Cufflinks 1.0b, Trinity 3.0, Oasis 3.0, and the analysis of expression profiles were done by Xiaolong Cao (xiaolong.cao@okstate.edu) in our laboratory at Oklahoma State University, which is supported by an NIH Grant (GM58634 to H. Jiang). The computation was performed at OSU High Performance Computing Center supported in part through an NSF grant (OCI-1126330). Note: since the cDNA libraries were built for the purposes of discovering as many transcripts as possible and validating MAKER gene models, no biological replicates were set up for most libraries. Please be cautious if you want to focus on quantitative aspect of the expression data.

2. Libraries

2.1. Thirty three paired end libraries

Details can be found in the file "RNAseq-M.sextaTissures20140507-B-gb.xlsx"
Read length: 100 bp.
Sequenced at Baylor College of Medicine sequencing.

2.2. Nineteen single end libraries

Details can be found in the file "RNAseq-M.sextaTissures20140507-B-gb.xlsx"
Read length: 50 bp.
Sequenced at Cornell

2.3. CIFH libraries [C: naïve 5th instar larvae as control, I: induced or bacteria-injected 5th instar larvae, F: fat body, H: hemocytes. For details, please read Zhang et al., 2011, Pyrosequencing-based expression profiling and identification of differentially regulated genes from *Manduca sexta*, a lepidopteran model insect. Insect Biochem Mol Biol, 41, 733-746; http://www.sciencedirect.com/science/article/pii/S096517481100107X]

2.4. Eight strand-specific libraries (Smith et al. 2014. Complete dosage compensation and sex-biased gene expression in the moth *Manduca sexta*. Genome Biol Evo.6, 526-537; http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3971586/) (The sequences seem to be reverse complement of mRNAs). Details can be found in the file "RNAseq-M.sextaTissures20140507-B-gb.xlsx"

3. Methods for data processing

RSEM version1.2.12 is used.

For 33 paired end libraries, paired reads are trimmed to 50 bp, and used as single end reads to align to the reference sequences at the same time. (If used as paired reads, the effective length for many genes will be too small which makes prediction inaccurate. Additionally, treating them as single end reads makes it consistent with results of the single end libraries).

The 19 single end libraries were used directly.

For CIFH libraries. Trim reads to 50bp.

For strand-specific reads, use them as strand-specific reads.

Below is an example of the code for RSEM:

```
/home/ks2073/p/rsem-1.2.12/rsem-calculate-expression --calc-ci --ci-memory 32000 -p 12 --no-bam-output /scratch/ks2073/RNAlab/RNApairtosingle/50bp/G2-4f.txt,/scratch/ks2073/RNAlab/RNApairtosingle/50bp/G2-4r.txt /scratch/ks2073/RSEM/RSEMTemp/DNAseq/ProDB /scratch/ks2073/RSEM/RSEMTemp/out2-4/2-4
```

The output includes many information, based on transcript_id or gene_id, like:

transcript_id	Msex2.00001-RA
gene_id	Msex2.00001
length	4917
effective_length	4868
expected_count	1045.01
TPM	12.75
FPKM	11.65
IsoPct	100
pme_expected_count	1045.02
pme_TPM	12.74
pme_FPKM	11.65
IsoPct_from_pme_TPM	100
TPM_ci_lower_bound	11.9732
TPM_ci_upper_bound	13.5182
FPKM_ci_lower_bound	10.9478
FPKM_ci_upper_bound	12.3611

gene_id	Msex2.00011
transcript_id(s)	Msex2.00011-RA, Msex2.00011-RB, Msex2.00011-RC, Msex2.00011-RD
length	2382.89
effective_length	2333.89
expected_count	236.03
TPM	6.01
FPKM	5.49
pme_expected_count	236.03
pme_TPM	6.04
pme_FPKM	5.52
TPM_ci_lower_bound	5.25252
TPM_ci_upper_bound	6.86477
FPKM_ci_lower_bound	4.8025
FPKM_ci_upper_bound	6.27644

TPM, FPKM, and expected count are pulled out to separate files.

FPKM/RPKM stands for Fragments (Reads) Per Kilobase of transcript per Million mapped reads.

TPM stands for Transcripts Per Million.

$$\mathrm{FPKM} = \frac{r_g \times 10^9}{fl_g \times R}$$

$$\mathrm{TPM} = \frac{r_g \times r_l \times 10^6}{fl_g \times T} \qquad \mathrm{T} = \sum_{g \in G} \frac{r_g \times r_l}{fl_g}$$

r_g: number of reads mapped to a particular gene region.

fl_g: number of nucleotide in a mapable region of a gene.

R: total reads number.

r_l: reads length

T: total transcript/mRNA number

$\frac{r_g \times r_l}{fl_g}$: target mRNA number

$$\mathrm{TPM} = \frac{r_g \times 10^9}{fl_g \times R} \times \frac{R \times r_l}{T} \times \frac{1}{1000} = FPKM \times \frac{Z}{1000} \qquad Z = \frac{R \times r_l}{T}$$

Z: average transcript/mRNA length

Thus, TPM stands for the number of transcripts from the target gene out of 1 million transcripts. FPKM stands for read numbers mapped to one kilobase of the target gene out of 1 million mapped reads. Here, TPM seems to be more reasonable compared to FPKM, as it reflects the ratio of mRNA numbers. You can compare TPM value of different genes across different libraries, if TPM is calculated from the same gene set (e.g., Cufflinks 1.0b). However, compared to FPKM, TPM has a Z, which is dependent on the gene set used, meaning that, if the gene set changes, TPM value will also change. You cannot compare TPM from OGS2.0 to TPM from OGS1.0, or to TPM from Cufflinks1.0b. However, you can compare FPKM from those model sets. As most of the future work will be based on OGS2.0, you may consider using TPM.

TPM/FPKM_ci_lower/upper_bound is the 95% confident interval (CI) provided by RSEM. Based on the paper, RSEM use a Bayesian version of its model to produce a posterior mean estimate and 95% CI for the abundance of each gene and isoforms. If you want to compare expression of very similar genes or isoforms based on the RNA-seq data, you may consider to take a look at the CI to get more confident conclusion as in these cases the difference between upper and lower bound of CI will be big.

4. How to use the data

4.1. General using and pulling out groups of data

If the genes you are interested are in the OGS2 list, go and find their IDs and their expression level. You may use the file based on transcript ID or gene ID, depending on whether you are interesting in differentiates alternative splicing.

If you have a list of genes you are interested and it might be too slow to find them one by one. You do the following steps:

4.1.1. Add a new column after the transcript ID as shown

transcript_id	New column	gene_id	length	effective_length	brain_2nd_1d_FPKM 9-1
Msex2.00001-RA		Msex2.00001-RA	4917	4868	11.29
Msex2.00002-RA		Msex2.00002-RA	909	860	4.04
Msex2.00003-RA		Msex2.00003-RA	4511	4462	3.86
Msex2.00004-RA		Msex2.00004-RA	1434	1385	3.1
Msex2.00005-RA		Msex2.00005-RA	5227	5178	6.51
Msex2.00005-RB		Msex2.00005-RB	1866	1817	7.92
Msex2.00005-RC		Msex2.00005-RC	2284	2235	0.48
Msex2.00005-RD		Msex2.00005-RD	2108	2059	0

4.1.2. In a new sheet, add ids you are interested like:

Msex2.00007-RB	1
Msex2.00007-RC	1
Msex2.00007-RD	1
Msex2.00008-RA	1
Msex2.00009-RA	1

Msex2.00009-RB	1
Msex2.00009-RC	1

4.1.3. Select all the “New column”, then type in “=vlookup(A2,secondpart,2,0)” in the formula box of excel (second part stands for the two columns you have in the “new sheet”, you need to select them, it should be like “Sheet2!A:B” in the equation), and then press ctrl+enter.

4.1.4. After this step, the genes you are interested will be marked “1” in New column. Then you can simply do data filter to pull out their expression data.

transcript_id		gene_id	length	effective_length	brain_2nd_1d_FPKM 9-1
Msex2.00006-RA	#N/A	Msex2.00006-RA	11026	10977	1.51
Msex2.00007-RA	#N/A	Msex2.00007-RA	1646	1597	0.06
Msex2.00007-RB	1	Msex2.00007-RB	1255	1206	0
Msex2.00007-RC	1	Msex2.00007-RC	1102	1053	0
Msex2.00007-RD	1	Msex2.00007-RD	1062	1013	0
Msex2.00008-RA	1	Msex2.00008-RA	300	251	20.95
Msex2.00009-RA	1	Msex2.00009-RA	1954	1905	13.8
Msex2.00009-RB	1	Msex2.00009-RB	2221	2172	14.7
Msex2.00009-RC	1	Msex2.00009-RC	3254	3205	0.07
Msex2.00010-RA	#N/A	Msex2.00010-RA	1715	1666	0
Msex2.00010-RB	#N/A	Msex2.00010-RB	2552	2503	10.53

4.2. Specific calculation and further analyses

4.2.1. If the sequences are not there, or they are incomplete or bad in your opinion, you can send us your sequences. We will replace the inaccurate models in OGS2 with yours and then do the calculation.

4.2.2. Genes can be clustered based on the expression pattern. If you would like to know what other genes share the expression patterns with the ones you are interested (suggestive of co-regulation), please let Xiaolong (xiaolong.cao@okstate.edu) know so he can pull out the information for you. Cluster analysis of expression data and tissue/stage-specific gene lists are available upon request. Typically, we use log2(FPKM+1) values to do cluster analysis.