Assignment #1: RNA-seq for Gene Expression Analyses (P/BIO 381)

Dung beetles (*Onthophagus taurus*), a model organism for polyphenism, have been employed in animal farms to reduce fly pests and diseases associated with the livestock. They were deliberately introduced to Australia from Mediterranean, and accidentally introduced to the eastern United States from an unknown origin around six decades ago. Here, we are using RNAseq data of 72 beetles sampled from Western Australia and North Carolina of both sexes and different developmental stages (Larval, Prepupae, Pupa and Adult), to examine gene expression patterns and how they differ between two populations.

The raw RNAseq data (100 bp paired ended reads) obtained from Illumina HiSeq2500 were cleaned and trimmed using Trimmomatic v- 0.33, and mapped against the reference genome of *Onthophagus taurus* using bwa aln function of Bwa-0.6. The counts were then extracted using custom written python script and exported to R v 3.4.3 for differential gene expression using DESeq2 v-1.18.1 package. Results of DESeq2 were used by various packages like ggplot2, Pheatmap, and RColorBrewer to create MAplot, PCA plot and Heat maps. The detailed of the scripts used can be found in my github notebook, page 11.

Table *1* shows that over 16,000 unique contigs were mapped into the reference genome of which slightly more than 200 were differentially expressed when compared among Western Australia (WA) and North Carolina (NC) populations overall. Differential expression of genes was further marginal when compared between WA and NC males having 25 up regulated and 12 down regulated genes. Between females of two populations, around double the number of genes were differentially expressed than in males. From figure *1*, it seems that the number of genes differentially present in these populations is rather small.

However, it's worthwhile to see what might be the functions of these genes which are differentially expressed. This was done by performing Gene Ontology-Mann-Whitney U (GO-MWU) analysis. The detailed process can be found in my GitHub notebook page. *Figure 2* shows that the GO categories related to ion transport and metabolic processes are significantly enriched between the WA and NC populations. Whereas, GO categories related to post-trascriptional regulation of gene expression and macromolecular complex subunit organization were enriched between WA and NC female populations.

The transcriptomic profiling of these two populations of beetles show that there are few differences in their transcriptome. Next, we could study the genetic diversity within and among these populations. Clearly, these populations have been separated physically for over 50 years and experience different geographical and environmental conditions. It would be interesting to see how their population genomics parameters vary; without largely altering the gene expression pattern.

Table 1: Summary of DESeq2 analysis showing number of up and down regulated genes in different models with p value < 0.1.

Model	Total Read Counts	Up regulated	Down Regulated	Outliers	Low counts
WAvsNC (Overall)	16848	131	88	0	657
WAvsNC (Males)	16589	25	12	0	966
WAvsNC (Females)	16460	54	41	0	2238

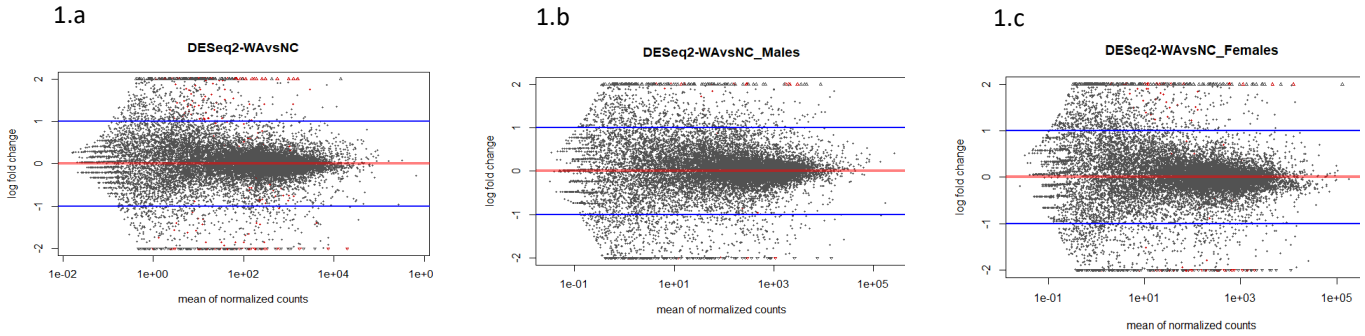

Figure 1: MA plot showing contigs identified in beetle populations. Grey dots represent each contig and red dots represent contigs those are significantly differentially abundant (p<0.1, and FDR cut-off of 5%) between; *a*) Western Australia (WA) and North Carolina (NC) overall; *b)* males of WA and NC; *c)* females of WA and NC populations. Blue lines delimit the twofold interval.

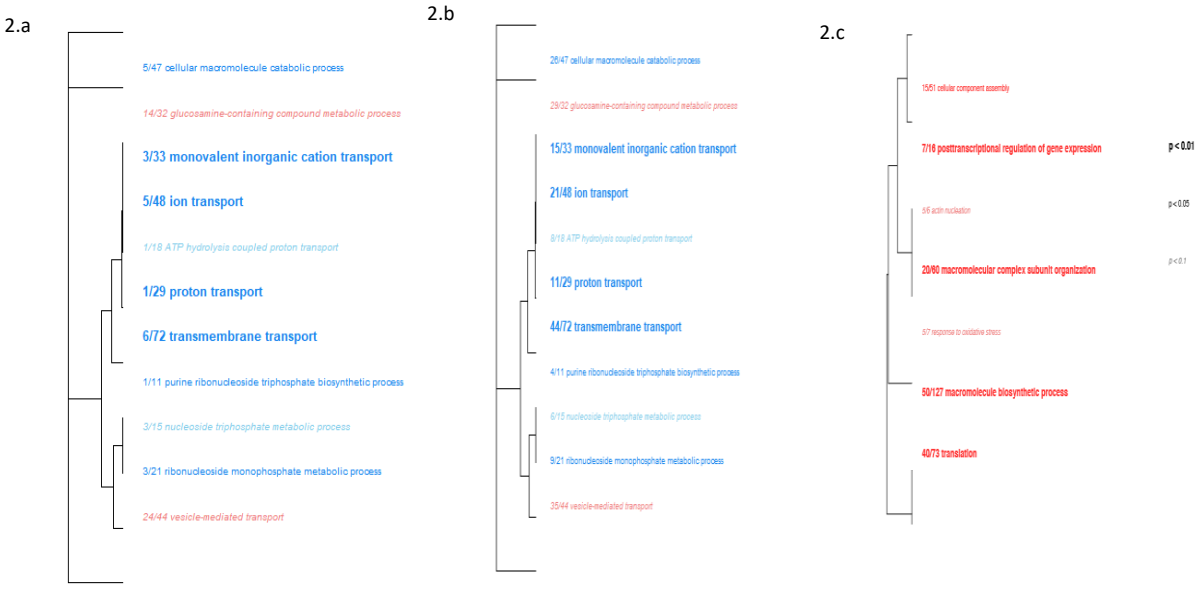

Figure 2: Gene Ontology categories significantly enriched with genes with either upregulated (red) downregulated *(a,b)* between WA and NC populations overall *c)* Between females of WA and NC. The font depicts the FDR-adjusted p-value derived from MWU test. The dendrogram represents the sharing of genes between categories; the categories with no branch length between them are subsets of each other. The fractions in *figure 2.a* represent the number of genes with an unadjusted p<0.05 relative to the total number of genes within the category. In *figure b*, the value was changed to 0.5. Note the increase in number of genes.