Weighted Gene Co-Expression Network Analysis of bull-headed dung beetle (*Onthophagus taurus*) transcriptomes across instars, populations and sex

Durham, Samuel M.; Dury, Guillaume J; Muralidhar, Anirudh K.
Indiana University, Bloomington, IN

Abstract:

Insects, in particular those with holometabolous development, are the most successful animals in terms of number of species. Beetles in particular are the largest single order of animals. Dung beetles, especially of the genus Onthophagus, have become model organisms for polyphenism. The horned and hornless males of the bull-headed dung beetle (*Onthophagus taurus*) have been studied extensively. In this study, we used weighted-gene co-expression network analysis (WGCNA) in order to investigate gene expression patterns in *Onthophagus taurus* across instars, populations, and sex. The diversity of developmental processes throughout the various life stages of *Onthophagus taurus* are the result of differential expression of genes. In order to investigate these differential gene expression patterns, one must group the genes into modules that have similar expression patterns. After determining well-defined modules, additional experiments and analysis can elucidate the biological and regulatory pathways in which the genes of a given module interact. In addition to investigating the effects of lifestage on gene expression, the effects of additional traits such as sex and sample provenance can provide a more thorough understanding of gene expression patterns. A thorough understanding of these gene modules and their regulatory/biological pathways will provide insight into diversification mechanisms of *Onthophagus taurus* across instars, populations, and sex. An understanding of these mechanisms may be helpful for characterizing similar pathways in other species.

Introduction:

Insects are very successful organisms; in number of species for example, estimates range from 2.6 to 7.8 million species with a mean of 5.5 million species–more than any other class of animals [1]. Insects are more numerous and accessible than larger animals, they are easy to rear, manipulate and measure; their short generation times allow consequences of manipulations to be assessed more easily than long lived species. Additionally, they exhibit a wide range of behaviours, allowing comparisons which are often impossible with vertebrates.

One group of insects that is particularly successful is the holometabolous insects–these have a complete life cycle consisting of an egg stage, usually several larval instars, a pupal instar and then adult insect. This contrasts with hemimetabolous insects which do not have larval and pupal stages, but have a series of

nymph instars [2]. Beetles are holometabolous, and the order Coleoptera has more species than any other animal order [1].

Dung beetles are a particularly important group of beetles in the family Scarabaeidae. Economically, the American Institute of Biological Sciences has estimated the value of dung beetles in the United States alone to be US$380 million annually through burying above-ground livestock feces [3]. For this reason, several species, especially of genus *Onthophagus*, have been voluntarily and involuntarily introduced in many places including Hawaii, continental Unites States, Australia and New Zealand [4-6].

Many aspects of the development of dung beetles have been studied in the using beetles of the genus *Onthophagus*, especially *Onthophagus taurus*. One particularly interesting aspect of the biology of this species is a polyphenism in which males develop horns only when they are well fed, whereas when they are not well fed, the males are hornless and look superficially more like females [7]. *Onthophagus taurus* was the first and is still the only dung beetle with a published genome. Studies of gene expression in this beetle have been conducted with microarrays [8,9], but no studies of whole-body RNA expression exist to date. We propose to use Weighted Gene Co-Expression Network Analysis (WGCNA) to study the patterns of gene expression of the bull-headed dung beetle across life-stages, sex and populations.

Methods:

Estimation of gene expression levels for *Onthophagus taurus* was done via *in silico* analyses of high-throughput RNAseq datasets generated in the context of an as yet unpublished study. Data was obtained from three males and three females of four life stages: late third instar larvae, early pre-pupae, first day pupae and young adult individuals; whole bodies were flash-frozen in liquid N_2, ground in a Geno/Grinder tissue homogenizer (SPEX SamplePrep, Metutchen, NJ) and total RNA was extracted using RNeasy Mini spin columns (Qiagen). This was done for beetles from Western Australia and from North Carolina for all life stages and for adult beetles from Italy (see Table 1).

	LifeStage			
Provenance	Late 3rd Instar Larvae	Early Pre-Pupae	1st-day Pupae	Young Adult
Italy				X
North Carolina	X	X	X	X
Western Australia	X	X	X	X

Table 1. Specimens sampled; four life stages were sampled from North Carolina and Western Australia, only adults for Italy. Three males and three females were sampled for each of those provenance/life-stage combinations.

Total RNA was quality checked using RNA ScreenTape TapeStation System (Agilent) and quantified with a Quant-iT RiboGreen Assay Kit (Thermo Fisher). RNA-stranded RNA sequencing libraries were constructed using the TruSeq Stranded mRNA Sample Preparation Kit (Illumina, San Diego, CA) according to manufacturer's instructions. Libraries were quantified using a Quant-iT DNA Assay Kit (Thermo Fisher), pooled in equal molar amounts, and sequenced either on HiSeq2000 (WDS libraries) or on NextSeq500 (TSP libraries) instruments (Illumina, San Diego, CA). Resulting read sequences were cleaned using Trimmomatic version 0.32 [10] to remove adapter sequences and perform quality trimming. Trimmomatic was run with the following parameters, “2:20:5 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:18”. The trimmed reads where then reversed complemented using FASTX-Toolkit version 0.0.13.2 [11].

A reference transcriptome set was generated from samples of various locations, lifestages, and sex (unpublished Moczek laboratory at Indiana University). The trimmed, complemented reads were then mapped to either the *Onthophagus* genome [12] or this reference transcriptome set using Tophat [13] and the parameters “-p 8 --b2-very-sensitive --read-edit-dist 2.” These mapped reads then served as input to HTSeq [14] in order to count the number of times that each gene was expressed for each sample. These read counts were then normalized with edgeR [15]. Normalization involved excluding genes with low counts (< 3 counts / Mbps) within a sample and low variation across samples. Additionally, HTSeq utilizes TMM (trimmed mean of M-values) normalization in order to correct the bias that results from highly-expressed genes having a disproportionately high read count.

After obtaining the normalized read counts for all samples and genes, WGCNA [16] was used to infer relationships among genes. Additionally, this tool was used to hypothesize which traits (i.e. sex, provenance, lifestage) may be responsible for the gene expression patterns that defined the gene modules. Within WGCNA, average-based hierarchical clustering (“hclust” method in WGCNA) was used to generate dendrograms of overall gene expressions (all genes considered together) across all samples and within

each provenance. A soft threshold value of 6 was used with the “blockwiseModules” method in order to calculate the gene modules.

We used WGCNA for the following reasons

1. It can be used to describe the correlation pattern among the genes of the samples.
2. It can be used find clusters of highly correlated genes.

Apart from these functions, WGCNA package helps in visualization of the data and interfacing that with the external software too. In this study, SampleVisualization.R was used to calculate topological overlap among the samples, which provided visual comparisons among the samples.

The two genes with highest expression were identified in the genome, their respective sequences were translated in all reading frames, the correct frame was identified and run through protein BLAST.

Results:

Graphical representation of the results

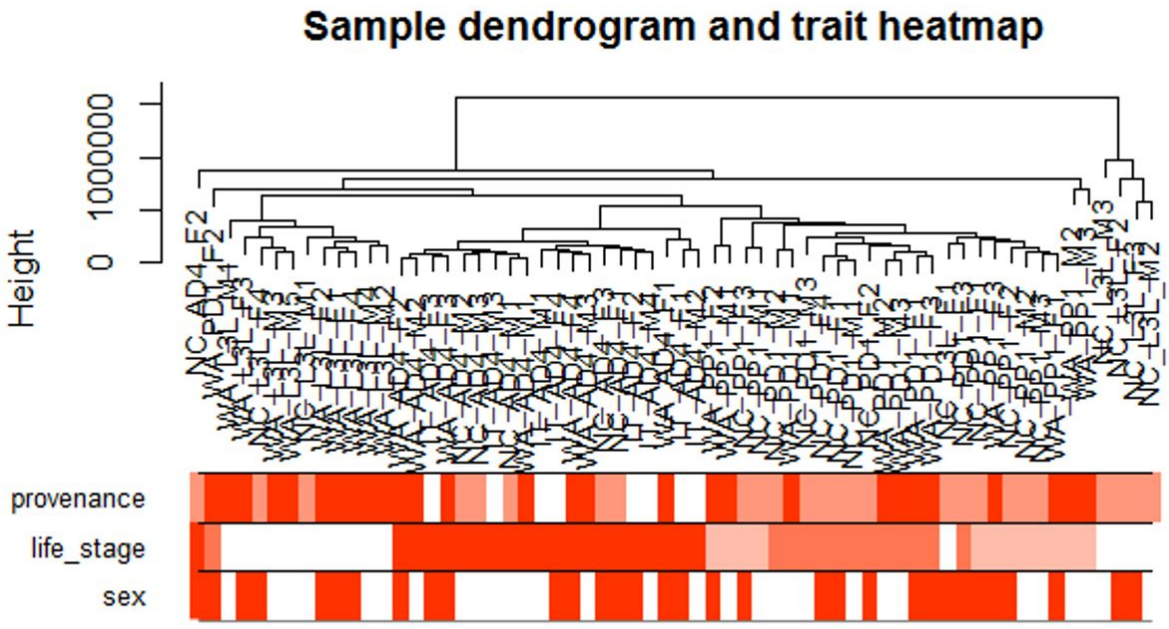

Figure 1. WGCNA dendrogram of samples using Unweighted Pair Group with Arithmetic Mean, showing how bull-headed dung beetle transcriptomes group in terms of gene expression profiles.

When the samples are clustered with the help of hierarchical clustering using Unweighted Pair Group with Arithmetic Mean, then mapped according to trait (life-stage, sex, provenance), we can see in Figure 1 that the samples mainly group by instar (as indicated by the large bars representing the same instar). On the other hand, there is a discontinuity in the provenance and sex traits.

Visualization of sample networks based on various locations

Western Australia

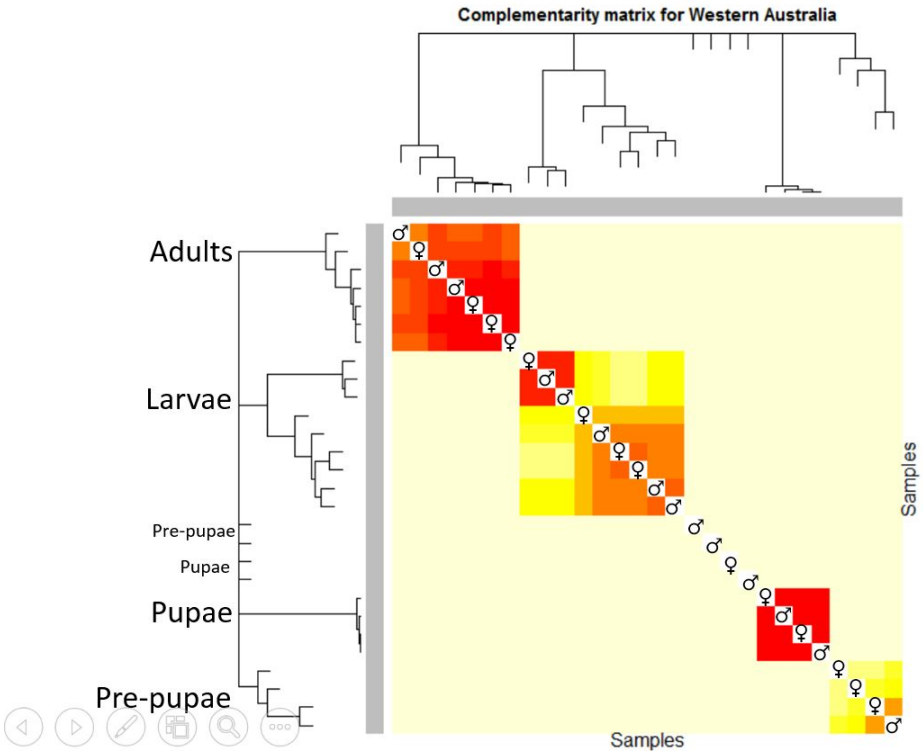

Figure 2. Complementarity matrix of transcriptomes generated using WGCNA for samples from Western Australia with light color denoting low adjacency overlap and dark color denoting higher adjacency.

From the graph we can see that in the Western Australia region, the adults, larvae, pupae and pre-pupae match within each other compared to others.

Of all the life stages, we can see that the adults and pupae samples have high similarity within each other irrespective of sex.

Also we could observe two subgroups in larvae. This might be because the dung beetle might be in growing stage, so we could observe different gene expression among various samples. The same might be the case with pre-pupae.

The next observation that we can make via this graph is each life stage matches within them only, not with other life stages.

Also we could see few bad samples in the graph, this might be because of some experimental mistakes while extracting the sequences, or the sample itself should some rare behaviour. But for our analysis we could neglect these bad samples are their count is very less.

North Carolina

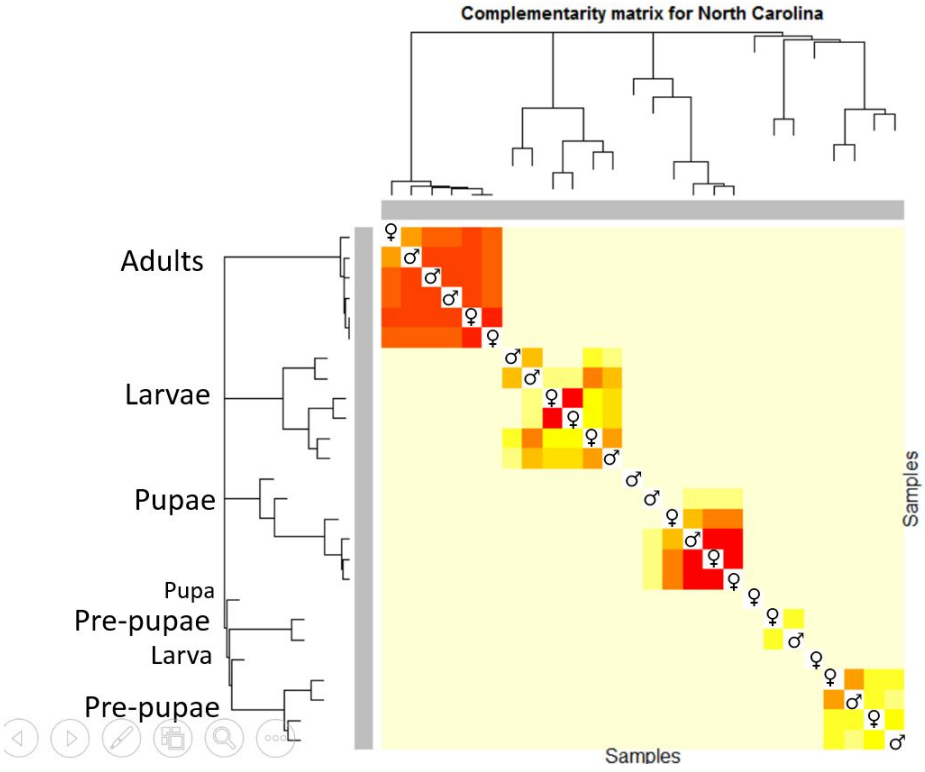

Figure 3. Complementarity matrix of transcriptomes generated using WGCNA for samples from North Carolina with light color denoting low adjacency overlap and dark color denoting higher adjacency.

In the case of North Carolina we could see similar kind of results like in the case of Western Australia irrespective of the sex of the samples.

Italy

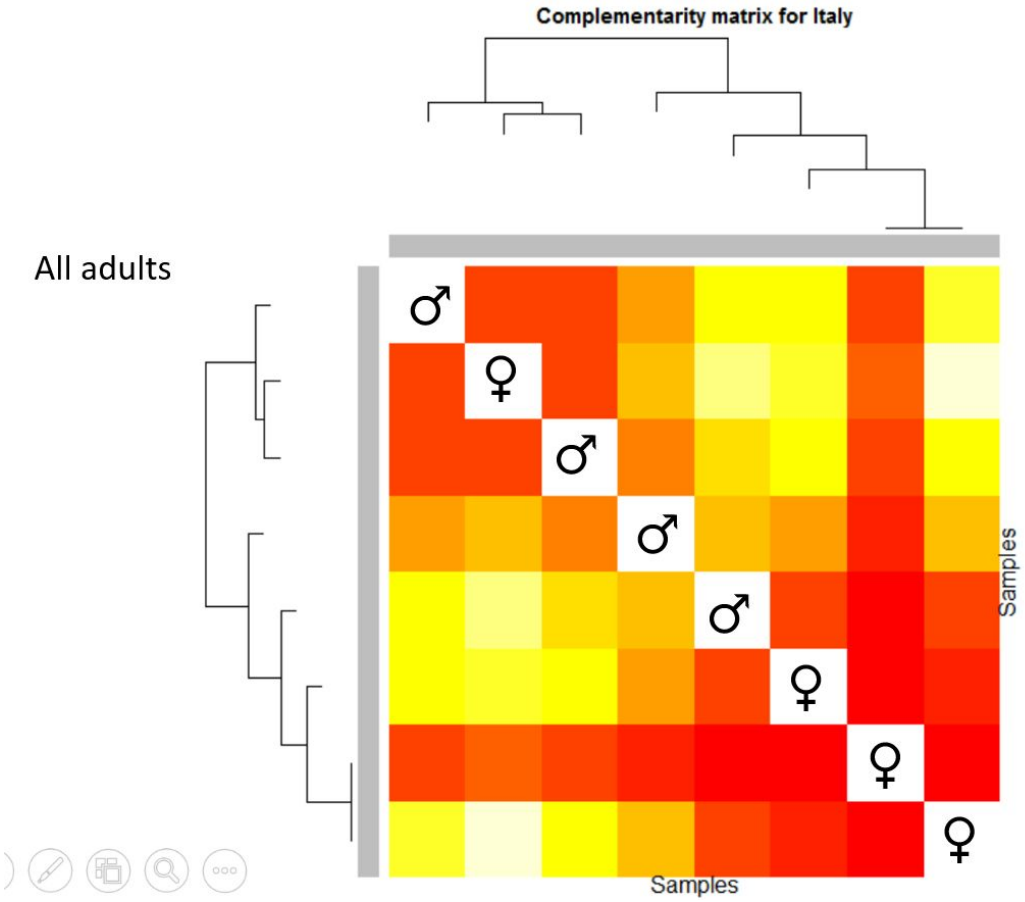

Figure 4. Complementarity matrix of transcriptomes generated using WGCNA for samples from Italy where the life stages for all samples is adult with light color denoting low adjacency overlap and dark color denoting higher adjacency.

Like North Carolina and Western Australia, we could see high correlation within samples in the case of Italy as well irrespective of sex trait.

Graphical visualization of the genes

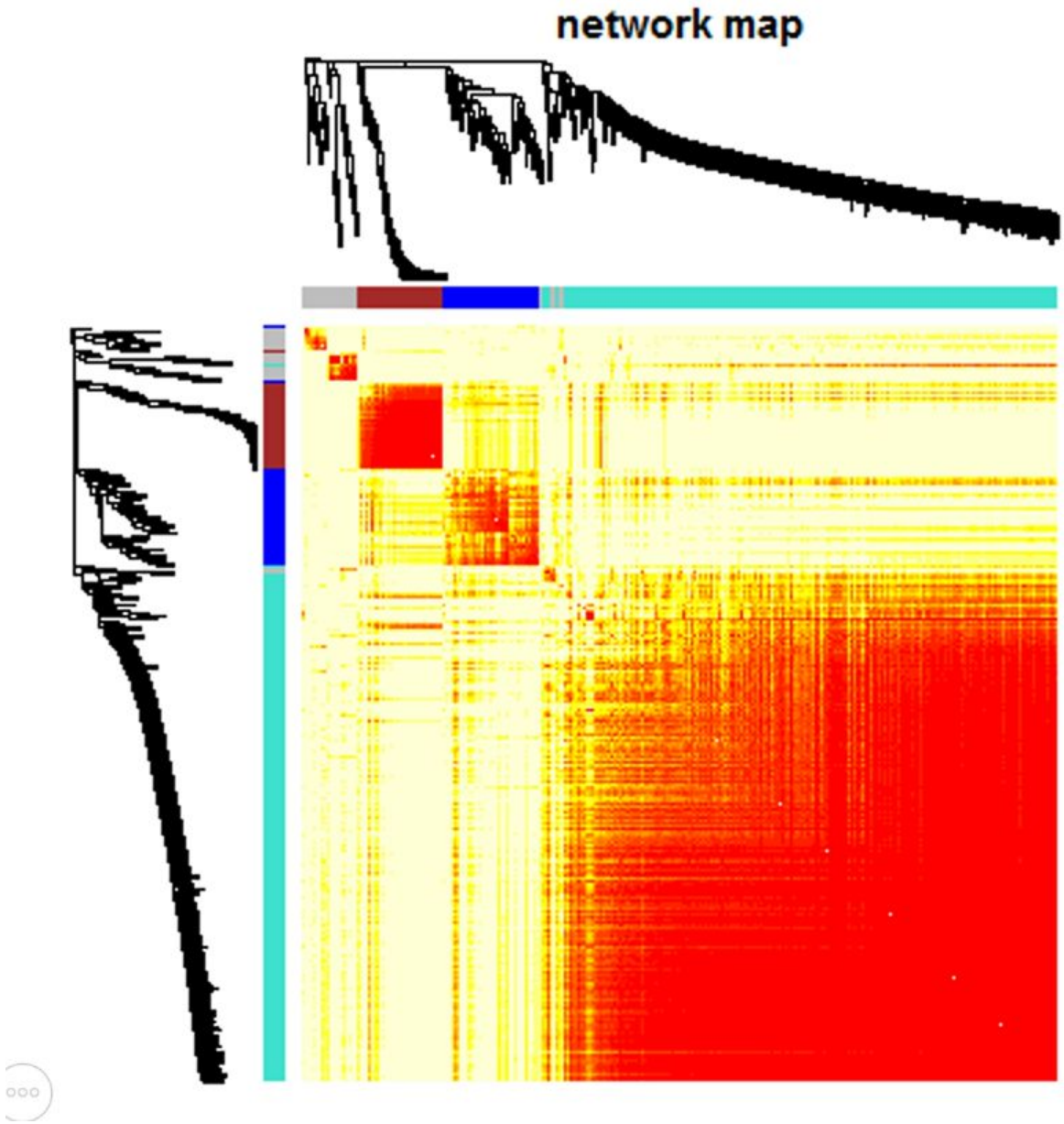

Figure 5. Correlation between various genes which is grouped into various modules with light color denoting low adjacency overlap and dark color denoting higher adjacency.

This graphs shows us the various modules that are being grouped together based on genes. From this we can see that there is a long light blue module. This shows us that most of the genes in the data have similar expression levels.

Identification of genes through protein BLAST (blastp):

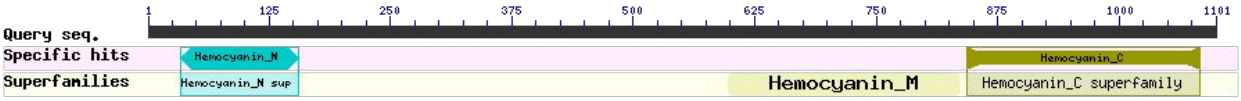

Figure 6. Putative conserved domains of the gene with highest gene expression, OTAU001920-RA. See appendix for DNA and protein sequences of this gene.

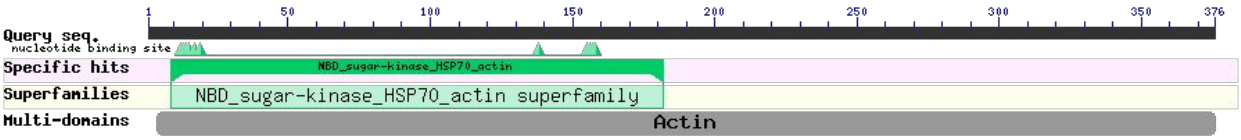

Figure 7. Putative conserved domains of the gene with second highest gene expression, OTAU012316-RA. See appendix for DNA and protein sequences of this gene.

A protein BLAST search of the most and second-most expressed genes identified them respectively as an hemocyanin and an actin (see Fig. 6 and 7). The best match for OTAU001920-RA was “hypothetical protein AMK59_6997” of the rhinoceros beetle *Oryctes borbonicus*, at 47%. The best match for OTAU012316-RA was “beta-actin” of the grasshopper *Diabolocatantops pinguis* at 100%.

Discussion:

From Figure 1, we can infer that the samples are grouped based on life stages predominantly. This was not surprising, given the diversity of biological processes that are occurring among the various life stages. Additionally, although we used species from three different locations, the two non-native populations (Western Australia and North Carolina) were introduced relatively recently, which may have not allowed significant time for location-based changes to become significant, especially in comparison with the diverse life stages.

Looking at Figures 2 and 3 of the results, the adult samples seemed to group better than the other life stages. This can be explained by the relative stability of the specimens in the adult life stage. Upon reaching adulthood, the beetles are experiencing relatively minor structural changes. This is in contrast to the significant changes that are occurring during the transformative development of the larval and pupal phases, during which the organisms are morphing into their “final” form. Within these transformative life stages, natural variation in the onset of the substages could easily explain the decreased level of grouping that was observed in the results.

Using protein BLAST, the two most expressed genes were identified as an actin and a hemocyanin. It makes sense that these be very highly expressed in *O. taurus*; actin is a protein with multiple functions and is present in essentially all eukaryotic cells. It is a component of the cytoskeleton of cells [16]. From the blastp output for gene OTAU001920-RA, it seems like the *O. taurus* equivalent protein has a large section

which other beetles like the rhinoceros beetle *Oryctes borbonicus* do not possess, which could explain the match of 47%. The second most expressed gene (OTAU012316-RA) was one for hemocyanin, it is also an ubiquitous protein that transports oxygen in arthropods and molluscs, analogous to vertebrate hemoglobin [17].

Conclusions & future directions:

From our study, we can draw two major conclusions: firstly, with the settings we used, one module (pale blue) was disproportionately larger, meaning many genes had similar expression profiles; secondly, we conclude that the most important of the traits we studied trait in explaining the gene expression of bull-headed dung beetles is life stage. Samples did not really cluster in terms of sex or provenance.

In future studies, the module-calculation parameters could be adjusted in order to split the relative low number of modules (~5) into more modules. Separating the genes into more modules may result in module sizes that can be more easily labeled as being responsible for specific functions or biological pathways. In order to obtain a higher number of modules, one could decrease the “minModuleSize” or “mergeCutHeight” parameters of the “blockwiseModules” method in WGCNA.

As mentioned in the “Discussion” above, the larval and pupal life stages did not group as well as the adult samples (see Figures 2 and 3). Future studies may involve sampling at multiple time points throughout these two life stages, rather than using a single time point for sampling. This could determine whether the diminished grouping was due to natural variation in the onset of the subphases of these two life stages, or whether the differences are due to differential expression among samples in the same subphase. In either case, sampling at multiple time points may provide a better representation of the gene expression patterns throughout a given life stage, especially with regard to transformative life stages.

This research raises many interesting questions that could serve as next steps. Specific differences within the species could be studied; for example, why what are the gene expression profiles of male dung beetle with and without horns. By analysing only how groups of samples differ, it may be possible to identify the differences between local populations. The line of inquiry started by using BLAST to attribute function could be extended to many more genes, and combined with the gene modules, in order to help understand the observed patterns and functions of gene expression. More experiments could be conducted using specific tissues and environmental conditions for the beetles.

Author's contributions:

This project is the result of equal contribution by A.K.M., S.M.D. and G.J.D. Some authors contributed more to certain parts; for example A.K.M. and S.M.D. did more of the analysis in WGCNA and G.J.D. contributed more to the PowerPoint presentation and normalization of the data in DESeq and edegR. We thank Dr. Eduardo Zattara for generous advice, mainly on mapping and read counting.

References:

1. Stork, Nigel E., James McBroom, Claire Gely, and Andrew J. Hamilton (2015) New Approaches Narrow Global Species Estimates for Beetles, Insects, and Terrestrial Arthropods. *Proceedings of the National Academy of Sciences* 112(24): 7519–7523.
2. Capinera, John L (2008) *Encyclopedia of Entomology*, vol.4. Springer Science & Business Media.
3. Losey, John E., and Mace Vaughan (2006) The Economic Value of Ecological Services Provided by Insects. *BioScience* 56(4): 311–323.
4. Kirk, AA, JP Lumaret, RH Groves, and F di Castri (1991) The Importation of Mediterranean-Adapted Dung Beetles (Coleoptera: Scarabaeidae) from the Northern Hemisphere to Other Parts of the World. *Biogeography of Mediterranean Invasions*. Cambridge University Press, Cambridge, UK: 413–424.
5. Dymock, J. J. (1993) A Case for the Introduction of Additional Dung-Burying Beetles (Coleoptera: Scarabaeidae) into New Zealand. *New Zealand Journal of Agricultural Research* 36(1): 163–171.
6. MacRae, Ted C, and Stephen R Penn (2001) Additional Records of Adventive Onthophagus Latreille(Coleoptera: Scarabaeidae) in North America. *The Coleopterists Bulletin* 55(1): 49–50.
7. Moczek, Armin P. (1998) Horn Polyphenism in the Beetle Onthophagus Taurus: Larval Diet Quality and Plasticity in Parental Investment Determine Adult Body Size and Male Horn Morphology. *Behavioral Ecology* 9(6): 636–641.
8. Kijimoto, Teiya, James Costello, Zuojian Tang, Armin Moczek, and Justen Andrews. (2009) EST and Microarray Analysis of Horn Development in Onthophagus Beetles. *BMC Genomics* 10(1): 504.
9. Snell-Rood, EC, A Cash, MV Han, et al. (2011) Developmental Decoupling of Alternative Phenotypes: Insights from the Transcriptomes of Horn-Polyphenic Beetles. Evolution; *International Journal of Organic Evolution* 65(1): 231–245.
10. Bolger A. M., Lohse M., and Usadel B. (2014). Trimmomatic: A flexible trimmer for Illumina Sequence Data. *Bioinformatics*, btu170.
11. FASTX-Toolkit: FASTQ/A short-reads pre-processing tools. http://hannonlab.cshl.edu/fastx_toolkit/index.html
12. i5k Workspace@NAL. Onthophagus taurus Genome. https://i5k.nal.usda.gov/Onthophagus_taurus
13. Kim D., Pertea G., Trapnell C., Pimentel H., Kelley R. and Salzberg S.L. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol* 14:R36
14. Anders S., Pyl T. P. and Huber W. (2014) HTSeq — A Python framework to work with high-throughput sequencing data. *Bioinformatics* 31(2): 166-169.
15. edgeR: Empirical Analysis of Digital Gene Expression Data in R. Bioconductor. https://www.bioconductor.org/packages/3.3/bioc/html/edgeR.html

16. Langfelder P. and Horvath S. (2008) WGCNA: an R package for weighted correlation network analysis. BMC *Bioinformatics*, 9:559
17. Karlin, K. D., and Z. Tyeklar (2012) *Bioinorganic Chemistry of Copper*. Springer Science & Business Media.
18. Lewin, Benjamin (200) *Cells*. Jones & Bartlett Learning.

Appendix:

DNA sequence of the most expressed gene:

>OTAU001920-RA

ATGAAGTCTCTAGTTTTCTTGGGGCTTTTGGCCCTTGCATGGGCAACCACTCTCAAAAAACCCGTCTAC
AATACACAAGAACCGGTGTACCCAACTGAAGAGTTCTTGCAACAGCAGAAGATTGTTTATCAATTATTCC
AATATGTGTACCAACCCAACTATTTGCCTGAATACGAAAAGATTGGTTTAACTTACAACCTCGAGGAGAA
TTTGCAAGGATACAGCAACCCGAAATATGTACTTGATTTCTTGAGTAACTACCAGGTTGGTAGTTTACCG
AAAGGACAAGTATTTTCCATCTTCTACGACAAGCACCAAAATGAAGCGATCGCTTTCTTCAGGGTTTTAT
ACAGCGCTGTTGATTTCGAAACTTTCTACAAAACCGCCGTTTGGGGAAGGATGCACTTGAATGAGGGTC
TTTACATCTACTCTCTTTCGGTTGCTTTGGTCCACAGACCCGATACCTGGAATCTCGAGTTACCACCGAT
CTACGAAATCTACCCATTCTATTTCTACAACTTCAACGTGTTCAGATTAGCCGAACAAAAGACTGGATAT
TATTTTGGAAACTACCCAACAACTTCTAAATACCAATACTACTTTGGAAAGTATTTGCAAAACCAACAATA
CCCATATGGGCAATTGTATGAAAGATTCGGTCAACAACCATGGTTCCAAATGAGCAAAACCACCCCGTT
CACTCAACAATATTTACAGAAATACGGACAATACCCGTGGTTCTCCAAATACGAGCAATCCAATCCAGTT
TATGGAAGATATATGGAATCGATTGAAACTCCCCAGCCATGGTCGTTCAAATACGAACAATCTTACCCAT
ACCAATATGCCAAAAAACAATTCCCAACTCAATACGTTCCATACAACCAATACCAACAAAGTTACCCATA
CGTACCACAACCCGAACAATACTATCAACAATACAGCAAACAATATGGAAAATCATACCCATGGAATCAA
GTACCAACTCAATACACTCCATACAGCCAACAATACTACCAAAAATACAACCAATATCCAACTCCATACG
GACCACAAATTGAGCAATACCAAAAATTGAGCAAACTCCCATGGTTCTCCTTCGAAAAATACGGAAAGT
CATACCCATGGAATCAGGTTCCATCTCAATACATCCCATACAGTCAACAATACTACCAAAAATACAACCA
ATATCCATCTTATCAATACCCATACGAACCACAAAGCGAACAATATTACCAACAATCCACCAAGCTCCCA
TGGTTGTTATCCGAGAAATATGAAAAATCTTACCCAATGAACCAAGTACCAACCCAATACACTCCATACA
ATCAACAATATTTCCAAAGATACAACCAATTCCCATCTCAATATATGCCTCAATCTAAATATATGCCCGAA
CCTGAATACAGGCCATATTCTCCCGAATACATCTCAACCTACGAGAAATACCCATATTTCCAACAAAGAT
ACCCATACATGGAACAAAAATCGACTCCATTCCCCCCATCTTACTACCAAAAATACGTAGAAGAACCGAT
GCCCGAATCTTTCAGGCCATATTCTCAACAATATTACCAAAAATACCAACAATACCCATATTTCCAAATG
AACAAGCAATTCACCAAACCGATGCCGTTCGTGAACAGCCCACTCTACCAAAAATCAATCTTCAAATACA
CCCCATCTCAATCTGAGATATTCCAAAAAATCCAGTATTCTCAAGTACCACAAGAAGTTCCGGAATTCTA
CCAGGAGAGAGAAAAGATACCACTTGAATACTATCAACAAGGAATTTTAATCACAGCTAATTTCACCAAC
TTCAACGCCCAATTCAACCCGGAATACCAATTATCTTATTTCACCCAAGATATTGGAATGAACGCTTTCT
ACTATTACTACCACATTTATTTCCCATTCTGGATGAAGAACGAAGAAATCGCCTTTAATACTCAAAGACG
TGGTGAACAATTCTATTACATTTACCAACAACTCTTGGCTAGATACAACCTCGAACGTATTTCTAACGGA
CTCGGTGAAATCCCACTTTTGCAATATGAACAAGGAGTTCCAGTTGGTTTCCGCACCCCAATGATGTAC
CCCAATGGAATGCCATTCCCAGAAAGGCCGGATCATTCTCCATTTGAATACTATCAAGGACCAGCTTAT
TATGATAACGATACTTTGAGTTTGTTAAACATTAAATTGTACGAAAGGAGGCTTTTGGATGCTATCGATT
CTGGATTTGTTTTAACTAGGGATGGAAGATTTATTTCTATTTTTGAACCGCAAGGTTTTGAAATTTTGGGC
AATTTGATTCAAGCTAACCCCGAATCTCCCAATGCAAGGTATTACGGACCTATTCTAGTCTTCGCTAGAT
ACCTTCTTGGACACTCAAATTATCCTCTCAATCAATACTACATCAACCCATCTGTGATGGAACAATACGA
AACTTCCCTCCATGATCCAGCATTCTATCAACTCTACAAAGTTATGATCTACTATTTCCAAAGATTCAAGC
AATACATGCCGATGTACACCGAACAAGAACTCTTATTACCAGGTGTTAAAATCCAAGATGTTAAAGTTGA
CAAATTGACCACTTATTTCGACTATTTCTACACTGACATCTCCAACATCATGCACAGCAAGGGACCAGTT
GAAGATATCCCAATTCAAGTCAGACAATATAGATTGAACCACAAACCATTCAACTATTACATGAATGTTAT
GAGTGAGAAACCATATGAAGCCGTTGTCAGAGTTTTCTTGGGACCAAAATACAGCGTTAATGGAGAACC
TATTGATATCAACAAGGATAGAATGAGTTTCGTTGAATTAGACAAATTCCAATACACCTTAGTACCAGGA
CTTAACACCATCGAAAGAAACTCGAGGGAATTCTACGGTTACGTTCCAGAACGTAGCACATACAAGGAA
CTTTTCATGCAATTGGTTAACTACCAAACTCCAATTCAATATAGCCCAAGAGTTTACTGGGGATTACCAC
AAAGAATGATGCTTCCTAAGGGTACTCAACAAGGCCAAGAATACCAAATGTACTTCTTGATCACCCCATA

CCAACCAGTTGAGAAGGTTTTGCCAAGAAGCTACCTTGGAAACAGACCACACTTGTACTGGGACCGCA
CCTTCGACAAACTCCCATACGGGTATCCATTCGACAGACCCATTGATGAAAGCGTTTTCTACGTTCCAA
ACTCTTACTTCAAGGATGTCGTCATCTACAACGAATTCACCCCCGAATTCTCCGGGGAATACCCATACA
CCCCATTCTATGAGAACTAT

Corresponding protein sequence:
MKSLVFLGLLALAWATTLKKPVYNTQEPVYPTEEFLQQQKIVYQLFQYVYQPNYLPEYEKIGLTYNLEENLQ
GYSNPKYVLDFLSNYQVGSLPKGQVFSIFYDKHQNEAIAFFRVLYSAVDFETFYKTAVWGRMHLNEGLYIYS
LSVALVHRPDTWNLELPPIYEIYPFYFYNFNVFRLAEQKTGYYFGNYPTTSKYQYYFGKYLQNQQYPYGQLY
ERFGQQPWFQMSKTTPFTQQYLQKYGQYPWFSKYEQSNPVYGRYMESIETPQPWSFKYEQSYPYQYAKK
QFPTQYVPYNQYQQSYPYVPQPEQYYQQYSKQYGKSYPWNQVPTQYTPYSQQYYQKYNQYPTPYGPQIE
QYQKLSKLPWFSFEKYGKSYPWNQVPSQYIPYSQQYYQKYNQYPSYQYPYEPQSEQYYQQSTKLPWLLS
EKYEKSYPMNQVPTQYTPYNQQYFQRYNQFPSQYMPQSKYMPEPEYRPYSPEYISTYEKYPYFQQRYPY
MEQKSTPFPPSYYQKYVEEPMPESFRPYSQQYYQKYQQYPYFQMNKQFTKPMPFVNSPLYQKSIFKYTPS
QSEIFQKIQYSQVPQEVPEFYQEREKIPLEYYQQGILITANFTNFNAQFNPEYQLSYFTQDIGMNAFYYYYHIY
FPFWMKNEEIAFNTQRRGEQFYYIYQQLLARYNLERISNGLGEIPLLQYEQGVPVGFRTPMMYPNGMPFPE
RPDHSPFEYYQGPAYYDNDTLSLLNIKLYERRLLDAIDSGFVLTRDGRFISIFEPQGFEILGNLIQANPESPNA
RYYGPILVFARYLLGHSNYPLNQYYINPSVMEQYETSLHDPAFYQLYKVMIYYFQRFKQYMPMYTEQELLLP
GVKIQDVKVDKLTTYFDYFYTDISNIMHSKGPVEDIPIQVRQYRLNHKPFNYYMNVMSEKPYEAVVRVFLGPK
YSVNGEPIDINKDRMSFVELDKFQYTLVPGLNTIERNSREFYGYVPERSTYKELFMQLVNYQTPIQYSPRVY
WGLPQRMMLPKGTQQGQEYQMYFLITPYQPVEKVLPRSYLGNRPHLYWDRTFDKLPYGYPFDRPIDESVF
YVPNSYFKDVVIYNEFTPEFSGEYPYTPFYENY

DNA sequence of the second-most expressed gene:
>OTAU012316-RA
ATGTGTGACGACGATGTTGCCGCGTTGGTCGTTGATAACGGATCCGGTATGTGCAAGGCCGGTTTCGC
CGGAGATGACGCTCCCCGCGCCGTCTTCCCATCCATCGTTGGTCGCCCACGTCATCAGGGTGTCATGG
TTGGTATGGGACAAAAAGACTCCTACGTAGGCGATGAGGCCCAAAGCAAGAGAGGTATCCTTACCTTG
AAATACCCAATTGAACATGGTATCATCACCAACTGGGATGACATGGAAAAGATCTGGCATCACACCTTC
TACAATGAACTCCGTGTTGCCCCAGAAGAACACCCAATCCTCCTTACTGAAGCTCCACTCAACCCTAAA
GCCAACAGGGAAAAGATGACCCAAATCATGTTTGAAACCTTCAACGCCCCAGCTATGTATGTTGCCATC
CAAGCTGTACTCTCCTTGTACGCCTCCGGTCGTACCACTGGTATCGTCTTGGACTCTGGAGATGGTGTA
TCCCACACCGTACCAATCTACGAAGGTTACGCCCTTCCCCATGCCATCCTTCGTCTTGACTTGGCCGGT
CGCGATTTGACCGACTACTTGATGAAGATCCTCACCGAAAGAGGCTACTCATTCACCACCACGGCTGAA
CGTGAAATCGTTAGGGACATTAAGGAAAAGCTTTGCTATGTCGCCCTGGACTTCGAACAAGAAATGGCC
ACCGCCGCCGCCTCCACCTCTTTGGAAAAAAGCTACGAATTGCCTGATGGTCAAGTTATCACCATCGGT
AACGAAAGATTCCGTTGCCCAGAAGCCCTCTTCCAACCTTCCTTCTTGGGTATGGAATCTTGCGGTATC
CATGAAACCGTATACAACTCCATCATGAAGTGCGACGTTGATATCCGTAAGGACCTCTACGCCAACAAC
GTACTCTCCGGTGGTACCACCATGTACCCTGGTATTGCTGATCGTATGCAAAAGGAAATCACCGCCTTG
GCCCCATCGACCATCAAGATCAAGATCATCGCTCCCCCAGAAAGGAAATACTCCGTATGGATCGGTGG
ATCCATCTTGGCTTCCCTCTCCACCTTCCAACAGATGTGGATCTCCAAACAGGAATACGACGAATCCGG
CCCAGGCATTGTCCACCGCAAGTGCTTCTAA

Corresponding protein sequence:
MCDDDVAALVVDNGSGMCKAGFAGDDAPRAVFPSIVGRPRHQGVMVGMGQKDSYVGDEAQSKRGILTLK
YPIEHGIITNWDDMEKIWHHTFYNELRVAPEEHPILLTEAPLNPKANREKMTQIMFETFNAPAMYVAIQAVLSL
YASGRTTGIVLDSGDGVSHTVPIYEGYALPHAILRLDLAGRDLTDYLMKILTERGYSFTTTAEREIVRDIKEKLC
YVALDFEQEMATAAASTSLEKSYELPDGQVITIGNERFRCPEALFQPSFLGMESCGIHETVYNSIMKCDVDIR
KDLYANNVLSGGTTMYPGIADRMQKEITALAPSTIKIKIIAPPERKYSVWIGGSILASLSTFQQMWISKQEYDE
SGPGIVHRKCF